Smoothed stationary bootstrap bandwidth selection for density estimation with dependent data

Ricardo Cao
MODES group, University of A Coruña (Spain)
Joint work with Inés Barbeito Cal

Galician Seminar of Nonparametric Statistical Inference,
June 8, 2016

Index

Setup and aims

- General dependent data, $\{X_t\}_{t\in\mathbb{Z}}$: stationary, α-mixing, ϕ-mixing, ...
- Nonparametric Parzen-Rosenblatt kernel density estimation

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

- Smooth bootstrap methods
- Bandwidth (h) selection

Smoothed bootstrap for independent data

Consider some statistic of interest: $R\left(\vec{X}, F\right)$

Smoothed bootstrap algorithm

1. Using the sample $(X_1, \ldots, X_n)$ and the bandwidth $h > 0$, compute $\hat{f}_h$
2. Draw bootstrap resamples $\vec{X}^* = (X_1^*, \ldots, X_n^*)$ from $\hat{f}_h$
3. Obtain the bootstrap version of the statistic: $R^* = R\left(\vec{X}^*, \hat{F}_h\right)$
4. Repeat Steps 1-3, B times to obtain $R^{*(1)}, \ldots, R^{*(B)}$
5. Use the values $R^{*(1)}, \ldots, R^{*(B)}$ to approximate the sampling distribution of R.

How to draw from $\hat{f}_h$?

Considering two independent random variables: $Y \sim F_n$ and U with density K, it is easy to prove that $Y + hU$ has density $\hat{f}_h$

Drawing resamples from $\hat{f}_h$

1. Draw naive bootstrap resamples $\overrightarrow{X}^{NAIVE*} = (X_1^{NAIVE*}, \ldots, X_n^{NAIVE*})$ from F_n
2. Draw a sample $\overrightarrow{U} = (U_1, \ldots, U_n)$ from the density K
3. Obtain the smoothed bootstrap resample $\overrightarrow{X}^{*} = (X_1^*, \ldots, X_n^*)$, where $X_i^* = X_i^{NAIVE*} + hU_i$

Moving Blocks Bootstrap (MBB)

MBB algorithm

Künsch (1989), Liu and Singh (1992)

1. Fix the block lenght, $b \in \mathbb{N}$, and define $k = \min_{\ell \in \mathbb{N}} \ell \geq \frac{n}{b}$
2. Define:

$$B_{i,b} = (X_i, X_{i+1}, \ldots, X_{i+b-1})$$

3. Draw $\xi_1, \xi_2, \ldots, \xi_k$ with uniform discrete distribution on $\{B_1, B_2, \ldots, B_q\}$, with $q = n - b + 1$
4. Define $\vec{X}^*$ as the vector formed by the first n components of

$$(\xi_{1,1}, \xi_{1,2}, \ldots, \xi_{1,b}, \xi_{2,1}, \xi_{2,2} \ldots, \xi_{2,b}, \ldots, \xi_{k,1}, \xi_{k,2}, \ldots, \xi_{k,b})$$

Stationary Bootstrap (SB)

SB algorithm

Politis and Romano (1994a)

1. Draw X_1^* from F_n
2. Once obtained $X_i^* = X_j$, for some $j \in \{1, 2, \ldots, n-1\}$, $i < n$, define X_{i+1}^* as follows:

$$X_{i+1}^* = X_{j+1} \text{ (if } j = n,\ X_{j+1} = X_1\text{), with probability } 1-p$$
$$X_{i+1}^* \text{ is drawn from } F_n \text{ with probability } p$$

Subsampling

Subsampling algorithm (for dependent data)

Politis and Romano (1994b)

1. Consider a dependent data sample $(X_1, \ldots, X_n)$ with marginal distribution F and $\theta = \theta(F)$
2. An estimator $T_n = T_n(X_1, \ldots, X_n)$ of $\theta = \theta(F)$ is considered and
$$J_n(u, F) = \mathbb{P}\left(\tau_n(T_n - \theta) \leq u\right)$$
3. Fix some $b \in \mathbb{N}$ such that $b < n$ and define:
$$S_{n,i} = T_b(B_{i,b}), i = 1, 2, \ldots, N, \text{ where } N = n - b + 1.$$
4. Use:
$$L_n(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{\tau_b(S_{n,i} - T_n) \leq x\}}$$
to approximate the sampling distribution of $\tau_n(T_n - \theta)$.

Plug-in method under dependence (PI)

Hall, Lahiri and Truong (1995)

- Minimizing in h the asymptotic MISE:

$$AMISE(h) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2^2 R(f'') - h^6\frac{1}{24}\mu_2\mu_4 R(f''') + \frac{1}{n}\left(2\sum_{i=1}^{n-1}\left(1-\frac{i}{n}\right)\int g_i(x,x)dx - R(f)\right).$$

  results in $h_{AMISE} = \left(\frac{J_1}{n}\right)^{1/5} + J_2\left(\frac{J_1}{n}\right)^{3/5}$, with $g_i(x_1,x_2) = f_i(x_1,x_2) - f(x_1)f(x_2)$, f_i the density of (X_j, X_{i+j}), $J_1 = \frac{R(K)}{\mu_2^2 R(f'')}$ and $J_2 = \frac{\mu_4 R(f''')}{20\mu_2 R(f'')}$.

- Now $h_{PI} = \left(\frac{\hat{J}_1}{n}\right)^{1/5} + \hat{J}_2\left(\frac{\hat{J}_1}{n}\right)^{3/5}$, with $\hat{J}_1$ and $\hat{J}_2$ some estimators of J_1 and J_2.

Plug-in method under dependence (PI)

- Replace $R(f'')$ by $\hat{I}_2$ and $R(f''')$ by $\hat{I}_3$, where:

$$\hat{I}_k = 2\hat{\theta}_{1k} - \hat{\theta}_{2k}, k = 2, 3,$$

$$\hat{\theta}_{1k} = 2\left(n(n-1)h_1^{2k+1}\right)^{-1} \sum\sum_{1\le i<j\le n} K_1^{(2k)}\left(\frac{X_i - X_j}{h_1}\right),$$

$$\hat{\theta}_{2k} = 2\left(n(n-1)h_1^{2(k+1)}\right)^{-1} \sum\sum_{1\le i<j\le n} \int K_1^{(k)}\left(\frac{x - X_i}{h_1}\right) K_1^{(k)}\left(\frac{x - X_j}{h_1}\right) dx.$$

Leave-$(2l+1)$-out cross validation (CV_l)

Hart and Vieu (1990)

- Define

$$CV_l(h) = \int \hat{f}^2(x)dx - \frac{2}{n}\sum_{j=1}^{n} \hat{f}_l^j(X_j),$$

where

$$\hat{f}_l^j(x) = \frac{1}{n_l}\sum_{i:|j-i|>l}^{n} \frac{1}{h}K\left(\frac{x-X_i}{h}\right).$$

- Choose n_l such that:

$$nn_l = \#\{(i,j) : |i-j| > l\}.$$

- The CV_l bandwidth selector is

$$h_{CV_l} = \arg\min_h CV_l(h).$$

Penalized cross validation (PCV)

Estévez, Quintela and Vieu (2002) proposed it for hazard rate estimation

- The PCV bandwidth selecctor is

$$h_{PCV} = h_{CV_l} + \bar{\lambda}.$$

- $\bar{\lambda}$ is chosen empirically as follows:

$$\lambda_n = \left(0.8e^{7.9\hat{\rho}-1}\right) n^{-3/10} \frac{h_{CV_l}}{100},$$

where $\hat{\rho}$ is the estimated autocorrelation

Modified cross validation under dependence ($SMCV$)

Stute (1992) proposed it for independent data

- Define

$$\begin{aligned} SMCV(h) &= \frac{1}{nh}\int K^2(t)dt \\ &+\frac{1}{n(n-1)h}\sum_{i\neq j}\left[\frac{1}{h}\int K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)dx\right] \\ &-\frac{1}{nn_l h}\sum_{j=1}^{n}\sum_{i:|j-i|>l}^{n}\left[K\left(\frac{X_i-X_j}{h}\right)-dK''\left(\frac{X_i-X_j}{h}\right)\right]. \end{aligned}$$

- The $SMCV$ bandwidth selector is

$$h_{SMCV}=\arg\min_h SMCV(h)$$

Exact MISE expression for the iid case

$$MISE\left(h\right)=\mathbb{E}\left[\int\left(\hat{f}_h\left(x\right)-f\left(x\right)\right)^2dx\right]=B\left(h\right)+V\left(h\right),$$

where

$$\begin{aligned} B\left(h\right) &= \int\left[\mathbb{E}\left(\hat{f}_h\left(x\right)\right)-f\left(x\right)\right]^2dx, \text{ e} \\ V\left(h\right) &= \int Var\left(\hat{f}_h\left(x\right)\right)dx \end{aligned}$$

Exact expression for $MISE(h)$:

$$\begin{aligned} B\left(h\right) &= \int\left(K_h*f\left(x\right)-f\left(x\right)\right)^2dx, \text{ and} \\ V\left(h\right) &= n^{-1}h^{-1}R\left(K\right)-n^{-1}\int\left(K_h*f\left(x\right)\right)^2dx. \end{aligned}$$

Smoothed bootstrap for the iid case

Smooth bootstrap algorithm for bandwidth selection Cao (1993)

1. Starting from $(X_1, \ldots, X_n)$ (iid), and using a pilot bandwidth, g, compute $\hat{f}_g$
2. Draw bootstrap resamples $(X_1^*, \ldots, X_n^*)$ from $\hat{f}_g$
3. For every $h > 0$, obtain

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

4. Construct the bootstrap version of $MISE$:

$$MISE^*(h) = \int \mathbb{E}^* \left[\left(\hat{f}_h^*(x) - \hat{f}_g(x) \right)^2 \right] dx$$

5. Obtain the bootstrap selector:

$$h_{MISE}^* = \arg\min_{h>0} MISE^*(h).$$

Smoothed bootstrap for the iid case

Closed expression for the bootstrap MISE

An exact expression for $MISE^*(h)$ can be found:

$$\begin{aligned} MISE^*(h) &= \frac{1}{n^2}\sum_{i,j=1}^{n}\left[(K_h * K_g - K_g) * (K_h * K_g - K_g)\right](X_i - X_j) \\ &+ \frac{R(K)}{nh} - \frac{1}{n^3}\sum_{i,j=1}^{n}\left[(K_h * K_g) * (K_g * K_g)\right](X_i - X_j), \end{aligned}$$

where $*$ denotes the convolution operator: $f * g(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy$. Consequently, there is no need to draw bootstrap resamples by Monte Carlo to approximate $MISE^*(h)$.

Exact MISE expression under dependence and stationarity

Exact expression for $MISE(h)$:

$$MISE\,(h) = B\,(h) + V\,(h)\text{, where}$$
$$B\,(h) = \int \left(K_h * f\,(x) - f\,(x)\right)^2 dx\text{, and}$$
$$V\,(h) = n^{-1}h^{-1}R\,(K) - \int \left(K_h * f\,(x)\right)^2 dx$$
$$+ \quad 2n^{-2}\sum_{\ell=1}^{n-1}(n-\ell)\int\int K_h\,(x-y)\,f\,(y)\,(K_h * f_\ell\,(\bullet|y))\,(x)\;dx\,dy,$$

where $f_\ell\,(\bullet|y)$ is the conditional density function of $X_{t+\ell}$ given $X_t = y$.

Smooth Stationary Bootstrap

SSB resampling plan Barbeito and Cao (2016)

1. Draw $X_1^{*(SB)}$ from F_n.
2. Draw U_1^* with density K and independently of $X_1^{*(SB)}$ and define
$$X_1^* = X_1^{*(SB)} + gU_1^*$$
3. Assume we have drawn $X_1^*, \ldots, X_i^*$ and consider the index $j/X_i^{*(SB)} = X_j$. Define I_{i+1}^*, such that
$$\begin{aligned} \mathbb{P}^* \left(I_{i+1}^* = 1 \right) &= 1 - p, \\ \mathbb{P}^* \left(I_{i+1}^* = 0 \right) &= p. \end{aligned}$$
Assign $X_{i+1}^{*(SB)} \mid_{I_{i+1}^*=1} = X_{(j \bmod n)+1}$ and draw $X_{i+1}^{*(SB)} \mid_{I_{i+1}^*=0}$ from the empirical distribution function
4. Define $X_{i+1}^* = X_{i+1}^{*(SB)} + gU_{i+1}^*$ (where U_{i+1}^* has density K). Go to the previous step if $i + 1 < n$.

MISE closed expression for SSB

An explicit expression for $MISE^*(h)$ can be obtained:

$$
\begin{aligned}
MISE^*(h) &= n^{-1}h^{-1}R(K) \\
&+ \left[\frac{n-1}{n^3} - 2\frac{1-p-(1-p)^n}{pn^3} + 2\frac{(n-1)(1-p)^{n+1} - n(1-p)^n + 1 - p}{p^2n^4}\right] \\
&\cdot \sum_{i,j=1}^{n} \left[(K_h * K_g) * (K_h * K_g)\right](X_i - X_j) \\
&-2n^{-2} \sum_{i,j=1}^{n} (K_h * K_g * K_g)(X_i - X_j) \\
&+n^{-2} \sum_{i,j=1}^{n} (K_g * K_g)(X_i - X_j) + 2n^{-3} \sum_{\ell=1}^{n-1} (n-\ell)(1-p)^{\ell} \\
&\cdot \sum_{k=1}^{n} \left[(K_h * K_g) * (K_h * K_g)\right]\left(X_k - X_{\lceil (k+\ell-1) \bmod n \rceil + 1}\right).
\end{aligned}
$$

Smooth Moving Blocks Bootstrap

SMBB resampling plan

1. Fix the block lenght, $b \in \mathbb{N}$, and define $k = \min_{\ell \in \mathbb{N}} \ell \geq \frac{n}{b}$
2. Define:

$$B_{i,b} = (X_i, X_{i+1}, \dots, X_{i+b-1})$$

3. Draw $\xi_1, \xi_2, \dots, \xi_k$ with uniform discrete distribution on $\{B_1, B_2, \dots, B_q\}$, with $q = n - b + 1$
4. Define $X_1^{*(MBB)}, \dots, X_n^{*(MBB)}$ as the first n components of

$$(\xi_{1,1}, \xi_{1,2}, \dots, \xi_{1,b}, \xi_{2,1}, \xi_{2,2} \dots, \xi_{2,b}, \dots, \xi_{k,1}, \xi_{k,2}, \dots, \xi_{k,b})$$

5. Define $X_i^* = X_i^{*(MBB)} + gU_i^*$, where U_i^* has been drawn with density K and independently from $X_i^{*(MBB)}$, for all $i = 1, 2, \dots, n$

MISE closed expression for SMBB

An explicit expression for $MISE^*(h)$ can be obtained, considering n an entire multiple of b.

- If $b = n$,

$$\begin{aligned} MISE^*(h) &= \frac{R(K)}{nh} \\ &+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi(X_i - X_j) \\ &- \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(K_h * K_g) * K_g\right](X_i - X_j) \\ &+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[K_g * K_g\right](X_i - X_j) \\ &+ \frac{\psi(0)}{n}, \end{aligned}$$

where $\psi(X_i - X_j) = \left[(K_h * K_g) * (K_h * K_g)\right](X_i - X_j)$.

MISE closed expression for SMBB

- If $b < n$,

$$
\begin{aligned}
MISE^*(h) \quad = \quad & \frac{R(K)}{nh} \\
& + \sum_{i=1}^{n} a_i \sum_{j=1}^{n} a_j \cdot \psi(X_i - X_j) \\
& - \frac{2}{n} \sum_{i=1}^{n} a_i \sum_{j=1}^{n} \left[(K_h * K_g) * K_g \right] (X_i - X_j) \\
& + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[K_g * K_g \right] (X_i - X_j) \\
& - \frac{b-1}{n(n-b+1)^2} \sum_{i=b-1}^{n-b+1} \sum_{j=b}^{n-b+2} \psi(X_i - X_j) \\
& - \frac{1}{nb \cdot (n-b+1)^2} \left[\sum_{i=1}^{b-1} \sum_{j=1}^{b-1} (\min\{i,j\}) \, \psi(X_i - X_j) \right.
\end{aligned}
$$

MISE closed expression for SMBB

$$+\sum_{i=1}^{b-1} i \sum_{j=b}^{n-b+1} \psi(X_i - X_j) + \sum_{i=1}^{b-1} \sum_{j=n-b+2}^{n} (\min\{(n-b+i-j+1), i\})\, \psi(X_i - X_j)$$

$$+\sum_{i=b}^{n-b+1} \sum_{j=1}^{b-1} j \cdot \psi(X_i - X_j) + \sum_{i=n-b+2}^{n} (\min\{(n-i+1), b\}) \sum_{j=b}^{n-b+1} \psi(X_i - X_j)$$

$$+\sum_{i=b}^{n-b+1} \sum_{j=n-b+2}^{n} (\min\{(n-j+1), b\}) \cdot \psi(X_i - X_j)$$

$$+\sum_{i=n-b+2}^{n} \sum_{j=1}^{b-1} (\min\{(n-b+j-i+1), j\})\, \psi(X_i - X_j) + b \sum_{i=b}^{n-b+1} \sum_{j=b}^{n-b+1} \psi(X_i - X_j)$$

$$+\sum_{i=n-b+2}^{n} \sum_{j=n-b+2}^{n} (n+1-\max\{i,j\})\, \psi(X_i - X_j) \Bigg]$$

MISE closed expression for SMBB

$$+\frac{2}{nb(n-b+1)}\sum_{s=1}^{b-1}\sum_{j=1}^{n-s}\left(\min\{j,b-s\}-\max\{1,j+b-n\}+1\right)\psi(X_{j+s}-X_j)$$

$$-\frac{2}{nb(n-b+1)^2}\left[\sum_{\substack{k,\ell=1\\k<\ell}}^{b}\left[\sum_{i=k}^{b-2}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)+\sum_{i=n-b+2}^{n-b+k}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)\right.\right.$$

$$\left.+\sum_{i=k}^{b-2}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)+\sum_{i=n-b+2}^{n-b+k}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)\right]$$

$$+\sum_{k=1}^{b-1}(b-k)\sum_{i=k}^{b-2}\sum_{j=b}^{n-b+2}\psi(X_i-X_j)+\sum_{\ell=2}^{b}(\ell-1)\sum_{i=b-1}^{n-b+1}\sum_{j=\ell}^{b-1}\psi(X_i-X_j)$$

$$\left.+\sum_{\ell=2}^{b}(\ell-1)\sum_{i=b-1}^{n-b+1}\sum_{j=n-b+3}^{n-b+\ell}\psi(X_i-X_j)+\sum_{k=1}^{b-1}(b-k)\sum_{i=n-b+2}^{n-b+k}\sum_{j=b}^{n-b+2}\psi(X_i-X_j)\right],$$

MISE closed expression for SMBB

considering a_j such that:

$$a_j = \begin{cases} \dfrac{j}{b(n-b+1)} & \text{, if } j = 1, \ldots, b-1 \\ \dfrac{1}{n-b+1} & \text{, if } j = b, \ldots, n-b+1 \\ \dfrac{n-j+1}{b(n-b+1)} & \text{, if } j = n-b+2, \ldots, n \end{cases} .$$

Simulated models

Six time series models have been considered

- **Model 1:**

$$X_t = -0.9X_{t-1} - 0.2X_{t-2} + a_t,$$

where the $a_t \stackrel{d}{=} N(0,1)$ are independent. Thus $X_t \stackrel{d}{=} N(0,0.42)$

- **Model 2:**

$$X_t = a_t - 0.9a_{t-1} + 0.2a_{t-2},$$

where $a_t \stackrel{d}{=} N(0,1)$ are independent. Thus $X_t \stackrel{d}{=} N(0,1.85)$.

Simulated models

- **Model 3:**

$$X_t = \phi X_{t-1} + (1 - \phi^2)^{1/2} a_t,$$

with $a_t \stackrel{d}{=} N(0,1)$, $\phi = 0, \pm 0.3, \pm 0.6, \pm 0.9$. Thus $X_t \stackrel{d}{=} N(0,1)$.

- **Model 4:**

$$X_t = \phi X_{t-1} + a_t,$$

where the distribution of a_t is given by $\mathbb{P}(I_t = 1) = \phi$, $\mathbb{P}(I_t = 2) = 1 - \phi$, with $a_t|_{I_t=1} \stackrel{d}{=} 0$ (constant), $a_t|_{I_t=2} \stackrel{d}{=} \exp(1)$, and $\phi = 0, 0.3, 0.6, 0.9$. We have $X_t \stackrel{d}{=} \exp(1)$

Simulated models

- **Model 5:**

$$X_t = \phi X_{t-1} + a_t,$$

where the distribution of a_t is $\mathbb{P}(I_t = 1) = \phi^2$, $\mathbb{P}(I_t = 2) = 1 - \phi^2$, with $a_t|_{I_t=1} \stackrel{d}{=} 0$ (constant), $a_t|_{I_t=2} \stackrel{d}{=}$ Dexp(1), and $\phi = 0, \pm 0.3, \pm 0.6, \pm 0.9$. Thus $X_t \stackrel{d}{=}$ Dexp(1).

- **Model 6:**

$$X_t = \begin{cases} X_t^{(1)} & \text{with probability } 1/2 \\ X_t^{(2)} & \text{with probability } 1/2 \end{cases},$$

where $X_t^{(j)} = (-1)^{j+1} + 0.5X_{t-1}^{(j)} + a_t^{(j)}$ with $j = 1, 2$, $\forall t \in \mathbb{Z}$, $a_t^{(j)} \stackrel{d}{=} N(0, 0.6)$ independent and $X_t \stackrel{d}{=} \frac{1}{2}N(2, 0.8) + \frac{1}{2}N(-2, 0.8)$

Performance measures

The following results will be shown for the six models considered in the simulations

$$\log\left(\frac{\hat{h}}{h_{MISE}}\right)$$

$$\log\left(\frac{MISE(\hat{h})}{MISE(h_{MISE})}\right),$$

where $\hat{h} = h_{CV_l}, h_{SMCV}, h_{PCV}, h_{PI}, h^*_{SSB}, h^*_{SMBB}$.

Approximating the optimal bandwidth

Consider some criterion function $\Psi(h)$ (e.g. $MISE^*(h)$ under SSB or SMBB; $CV_l(h)$ for Hart and Vieu's CV, Stute's MCV or Estévez, Quintela and Vieu PCV).

1. Consider a set of five equispaced bandwidths, $\mathcal{H}_1$ between 0.01 and 10
2. Obtain $h_{OPT_1} = \arg\min_{h\in\mathcal{H}_1} \Psi(h)$
3. Consider h_a the previous value of h_{OPT_1} within $\mathcal{H}_1$ and h_b the following value to h_{OPT_1} within $\mathcal{H}_1$
4. Construct a new set, $\mathcal{H}_2$, of equispaced bandwidths between h_a and h_b
5. Repeat Steps 2-4 10 times
6. The approximated optimal bandwidth is the value obtained in the 10th repetition

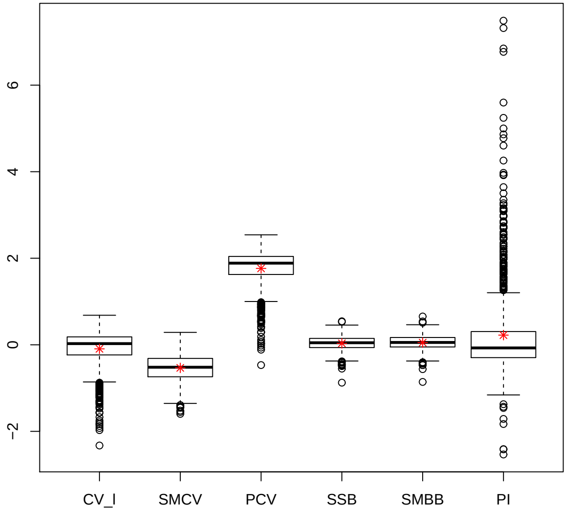
Technical aspects

- $l = 5$ for CV_l
- h_{SMCV} is considered as the smallest h for which $SMCV(h)$ attains a local minimum, not its global one
- Pilot bandwidth for PI: $h_1 = 1$
- Pilot bandwidth for h^*_{SSB} and h^*_{SMBB} as in the iid case: some normal reference estimator of

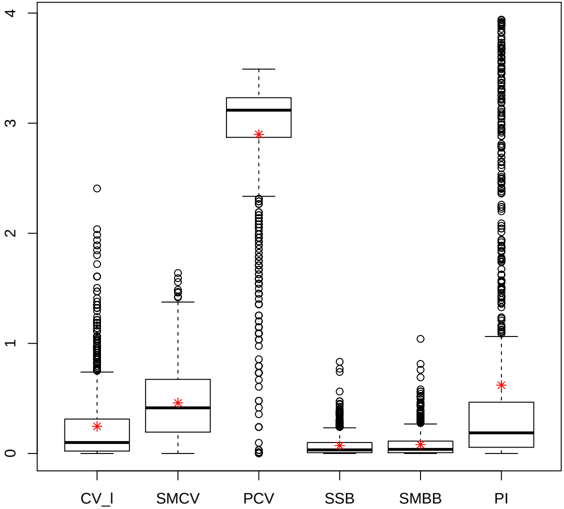
$$g_0 = \left(\frac{\int K''(t)^2 dt}{n d_K \int f^{(3)}(x)^2 dx} \right)^{1/7}$$

- $p = 0.05$ for SSB
- $b = 20$ for SMBB
- For every model, 1000 random samples of size $n = 100$ were drawn

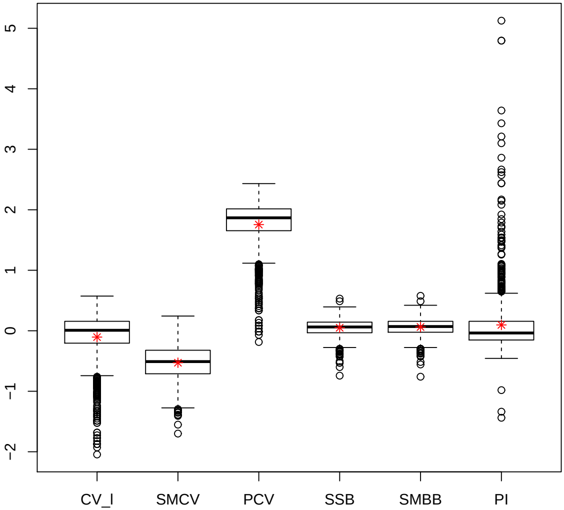
$\log(\hat{h}/h_{MISE})$. Model 1



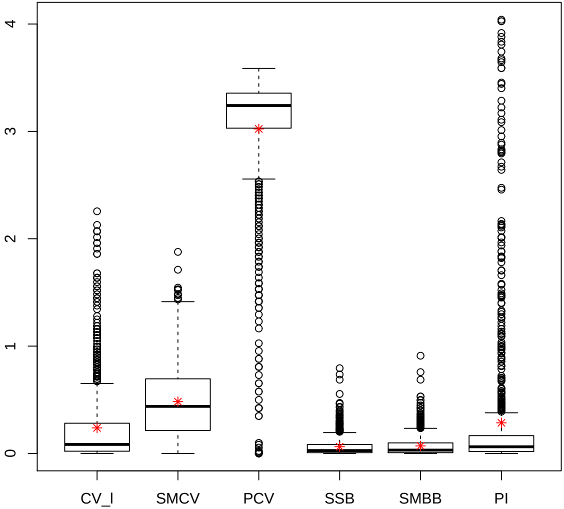
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 1



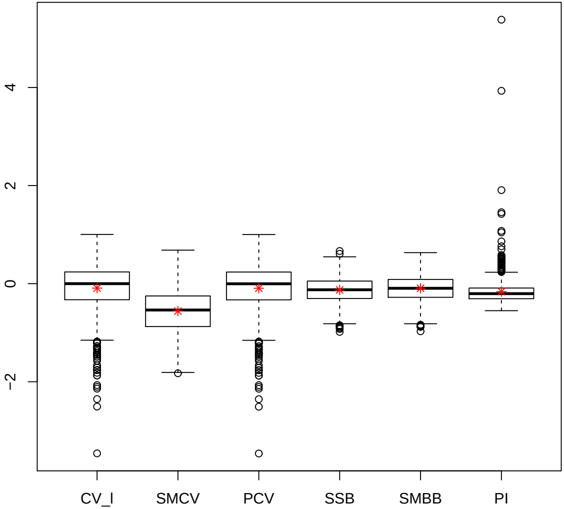
$\log(\hat{h}/h_{MISE})$. Model 2



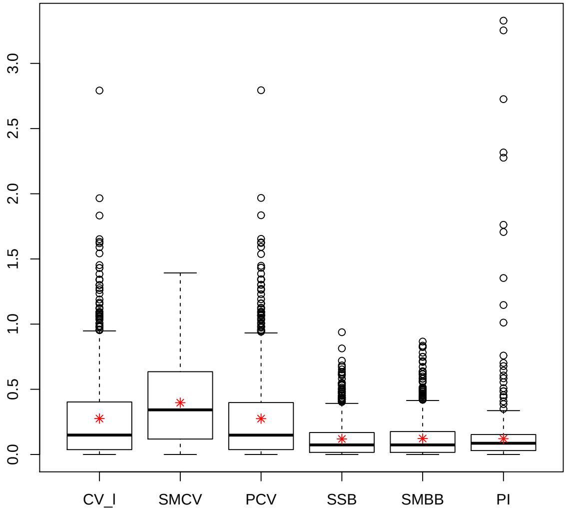
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 2



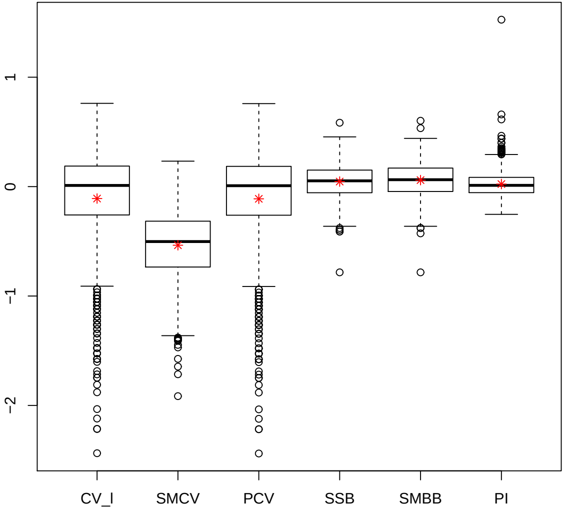
$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = -0.9$



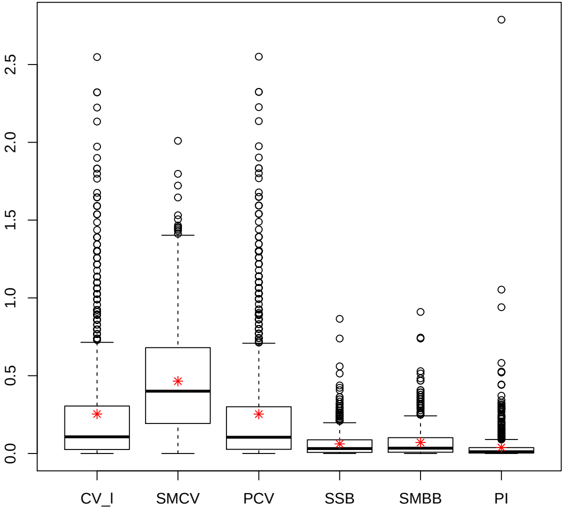
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 3, $\phi = -0.9$



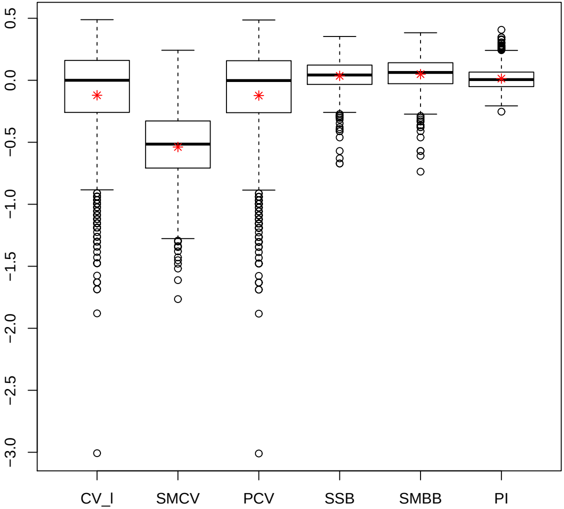
$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = -0.6$



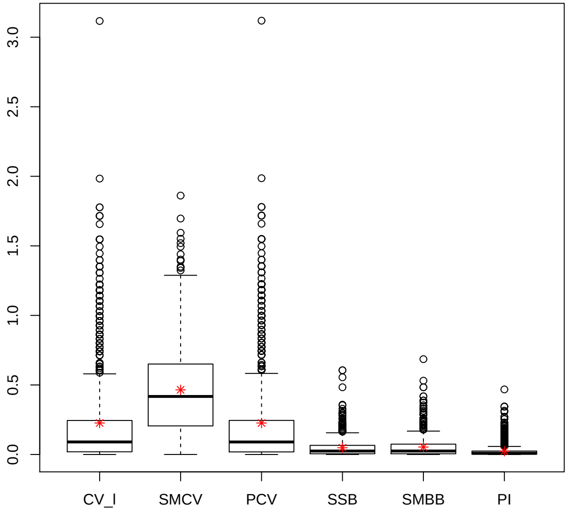
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 3, $\phi = -0.6$



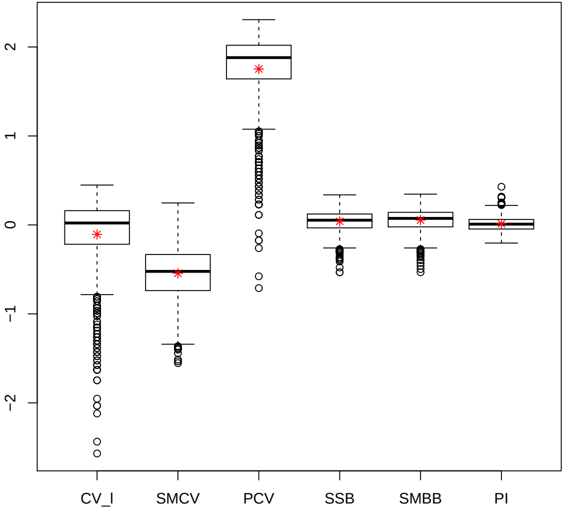
$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = -0.3$



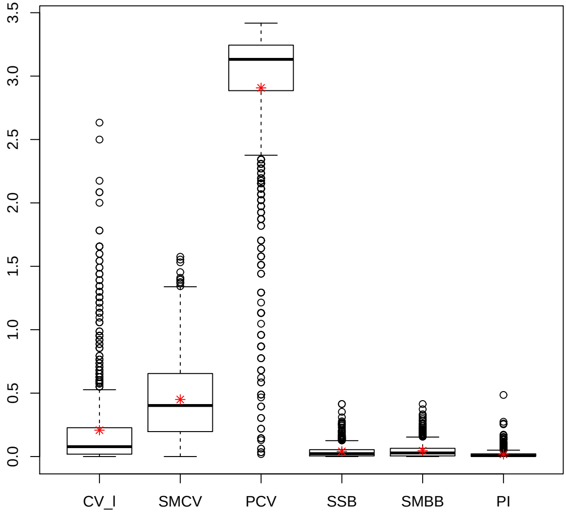
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 3, $\phi = -0.3$



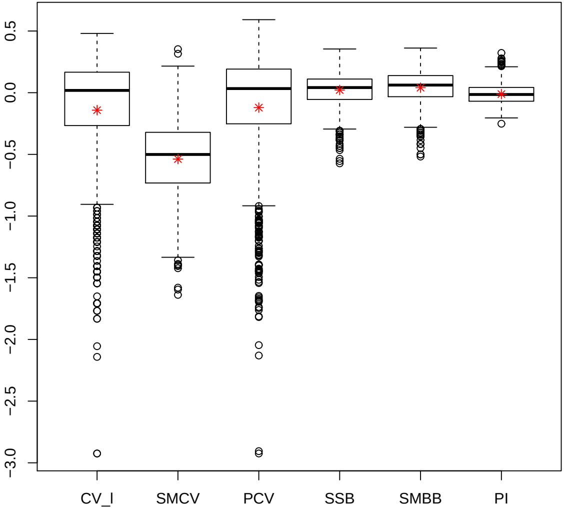
$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = 0$



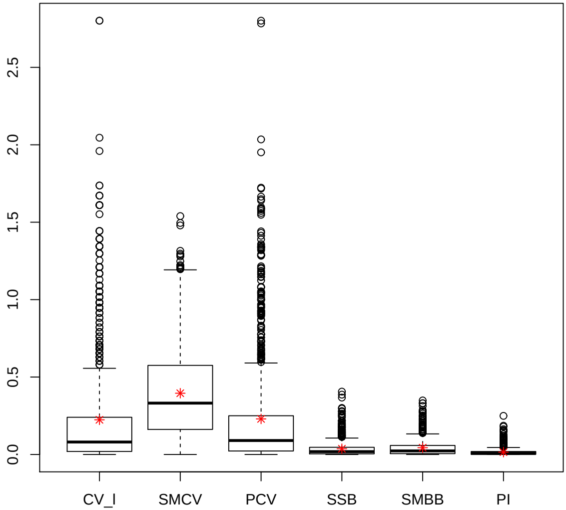
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 3, $\phi = 0$



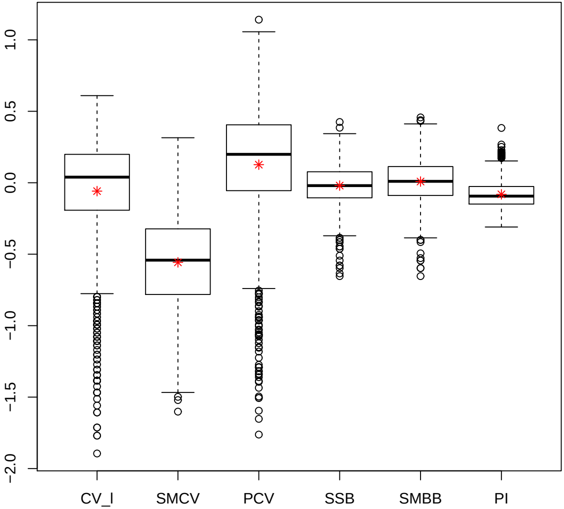
$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = 0.3$



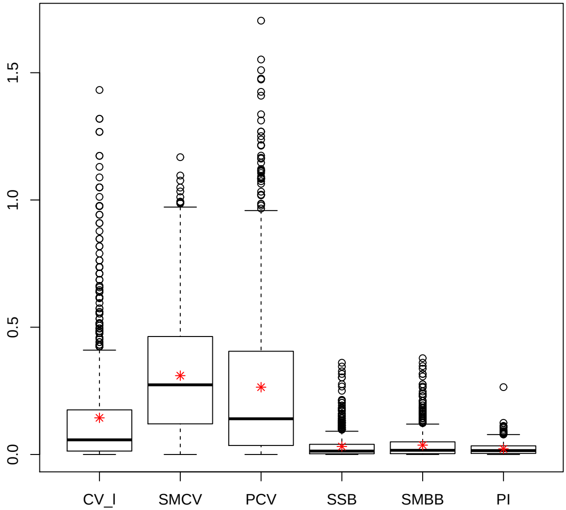
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 3, $\phi = 0.3$



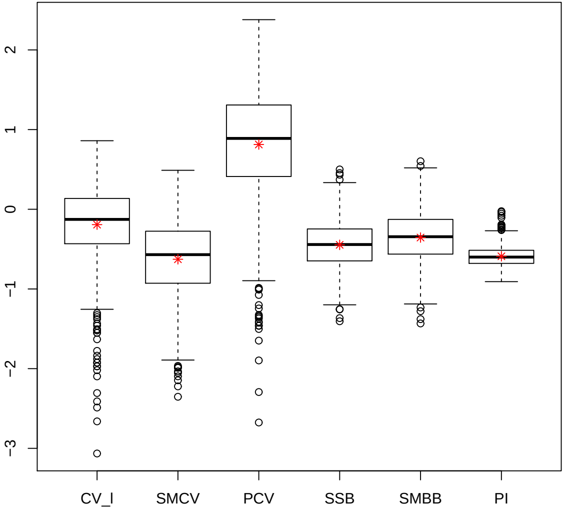
$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = 0.6$



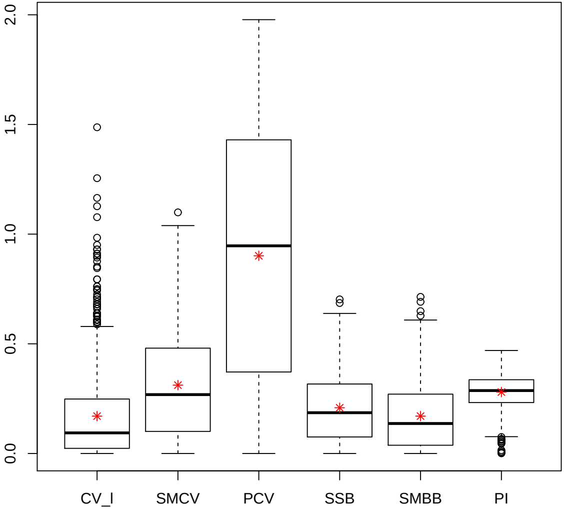
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 3, $\phi = 0.6$



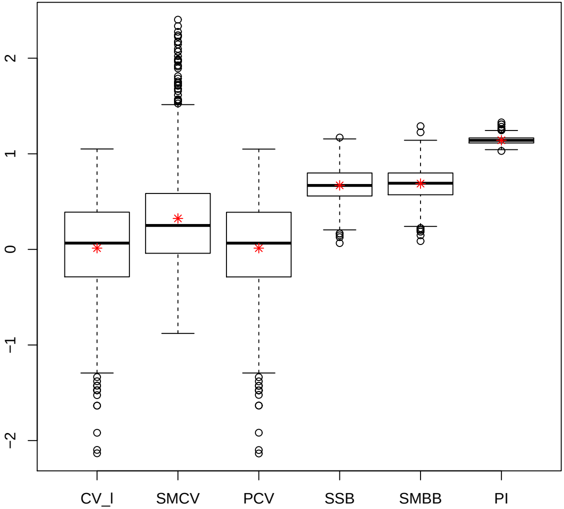
$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = 0.9$



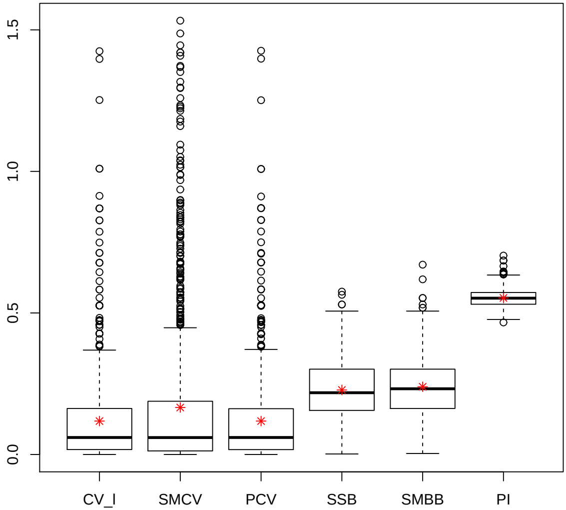
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 3, $\phi = 0.9$



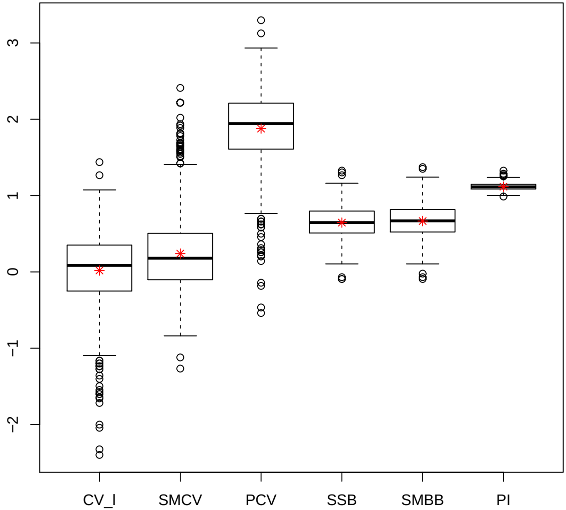
$\log(\hat{h}/h_{MISE})$. Model 4, $\phi = 0$



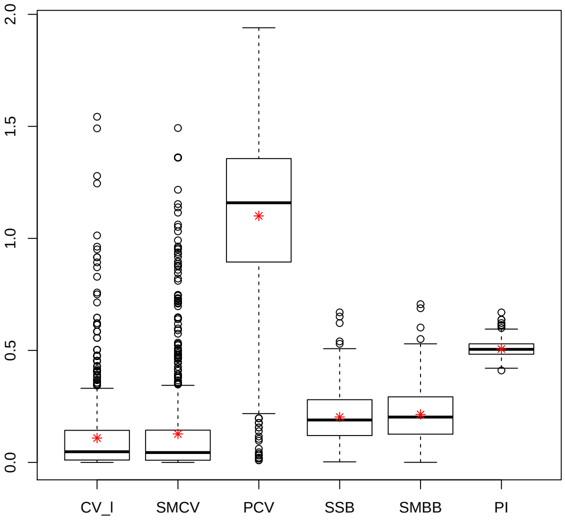
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 4, $\phi = 0$



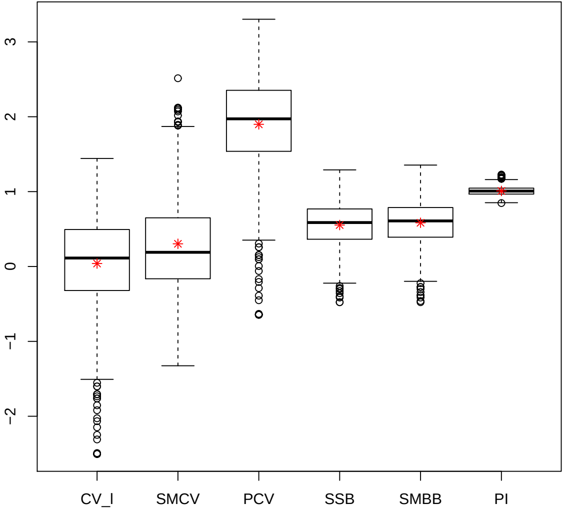
$\log(\hat{h}/h_{MISE})$. Model 4, $\phi = 0.3$



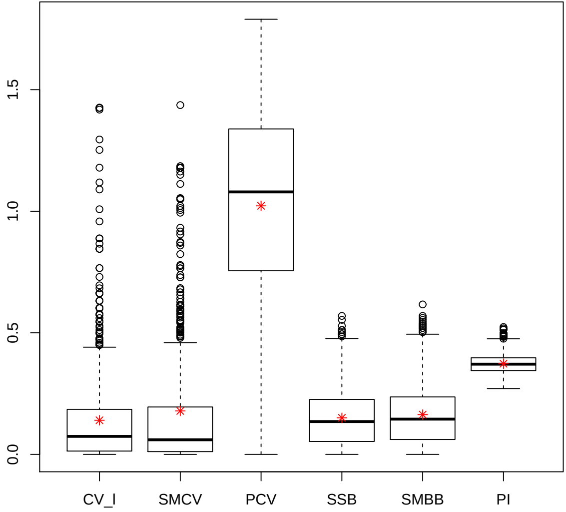
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 4, $\phi = 0.3$



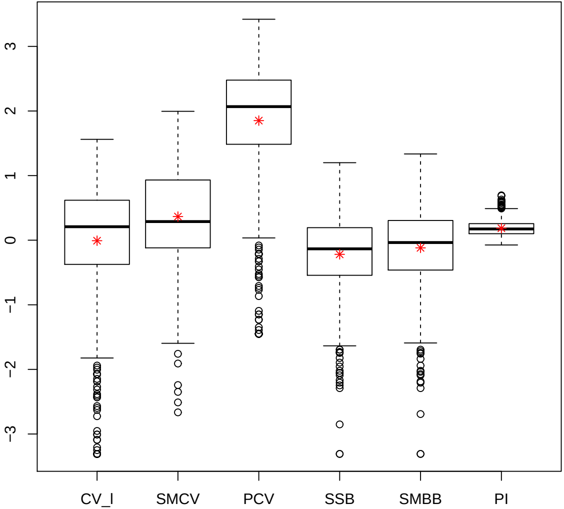
$\log(\hat{h}/h_{MISE})$. Model 4, $\phi = 0.6$



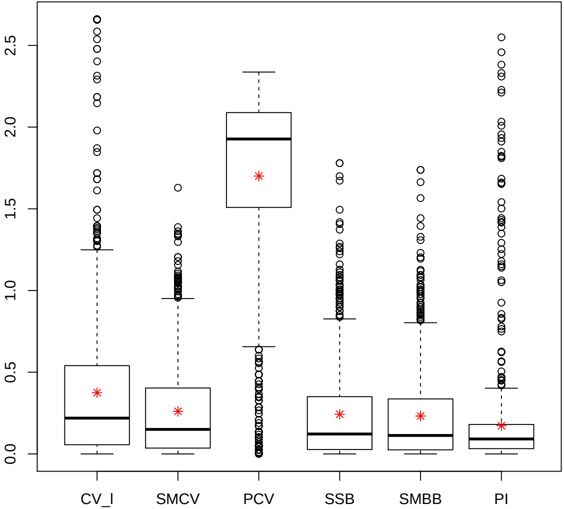
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 4, $\phi = 0.6$



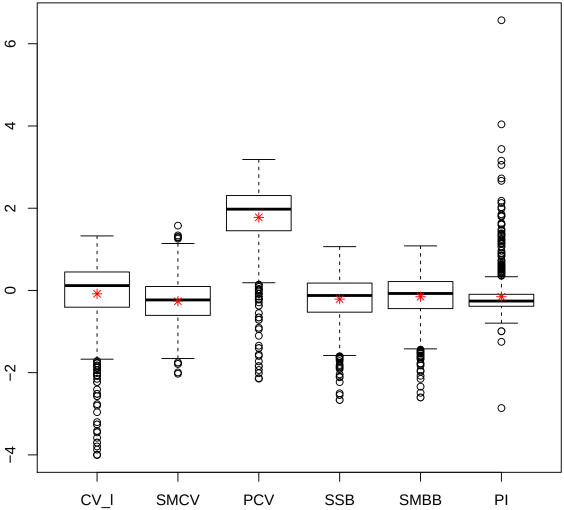
$\log(\hat{h}/h_{MISE})$. Model 4, $\phi = 0.9$



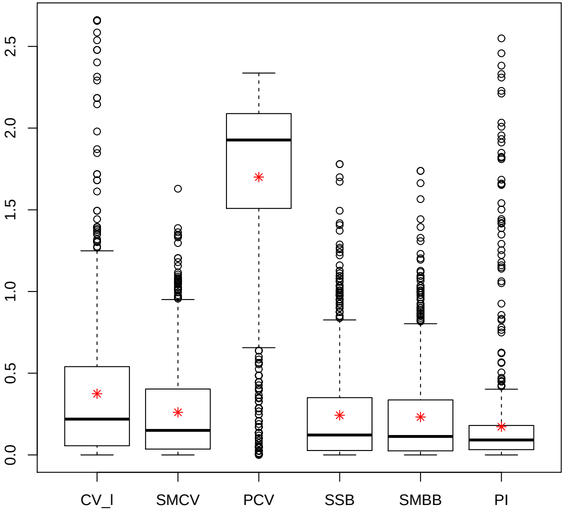
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 4, $\phi = 0.9$



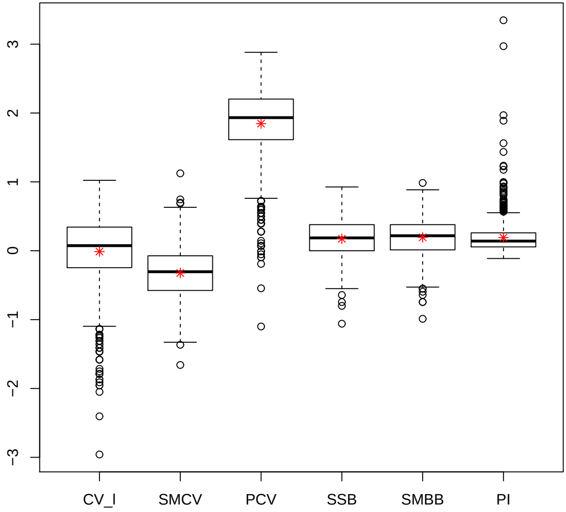
$\log(\hat{h}/h_{MISE})$. Model 5, $\phi = -0.9$



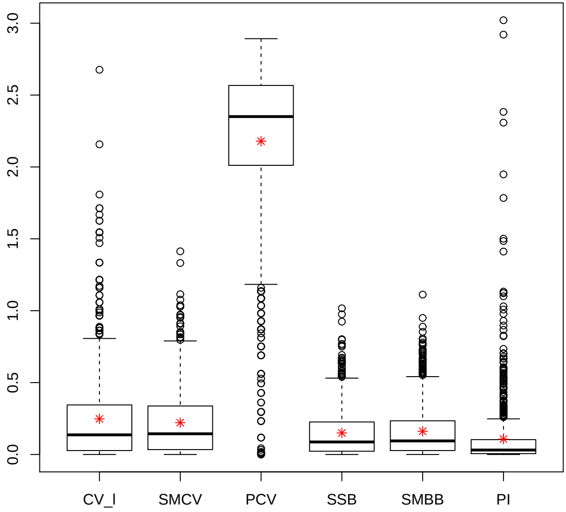
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 5, $\phi = -0.9$



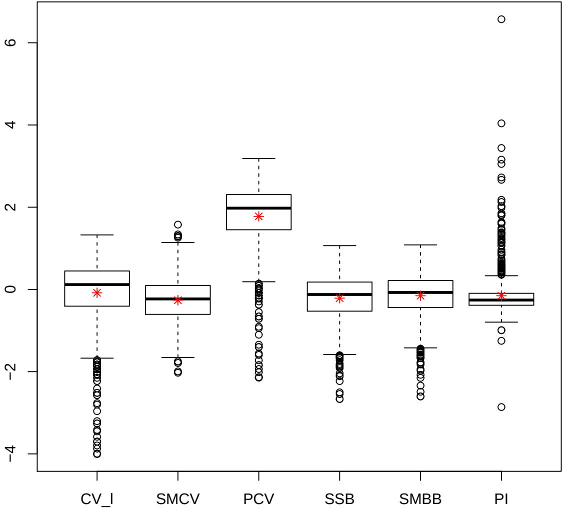
$\log(\hat{h}/h_{MISE})$. Model 5, $\phi = -0.6$



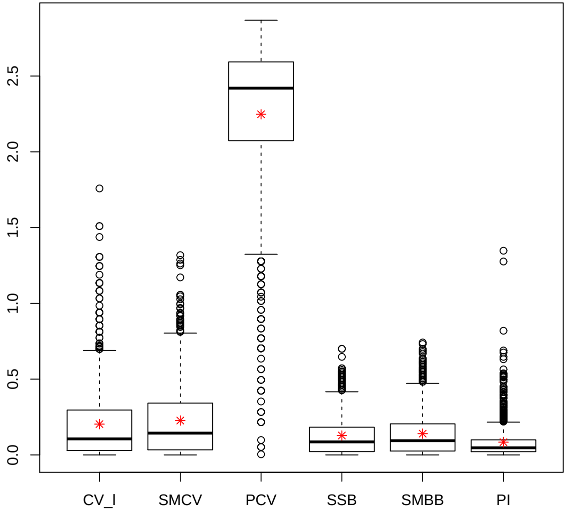
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 5, $\phi = -0.6$



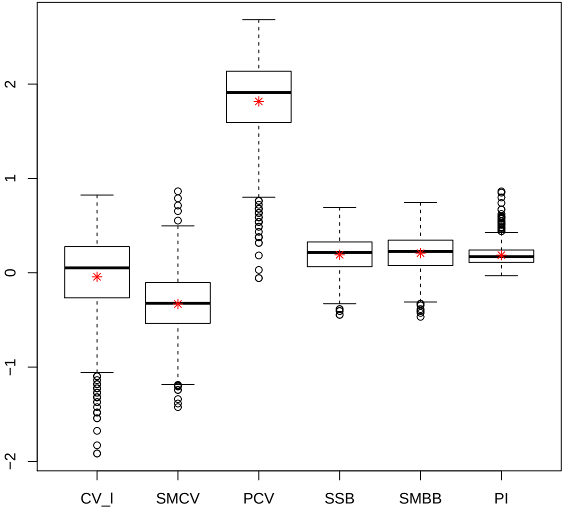
$\log(\hat{h}/h_{MISE})$. Model 5, $\phi = -0.3$



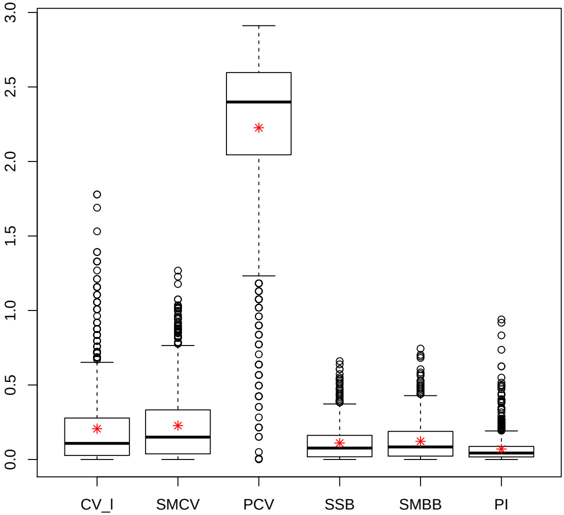
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 5, $\phi = -0.3$



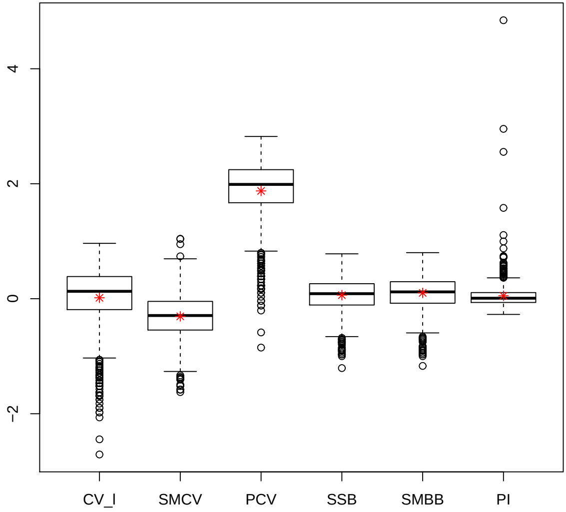
$\log(\hat{h}/h_{MISE})$. Model 5, $\phi = 0$



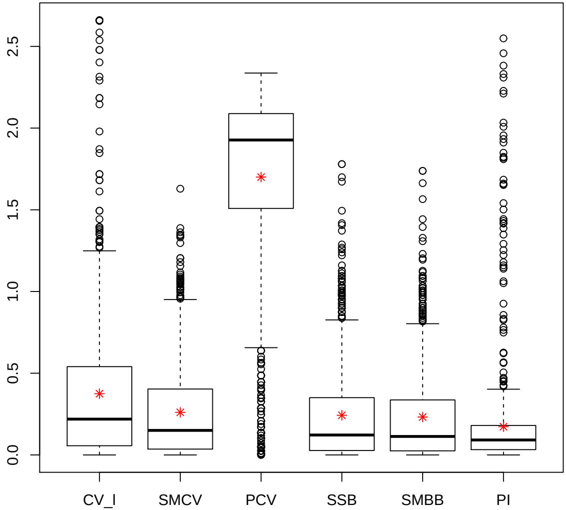
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 5, $\phi = 0$



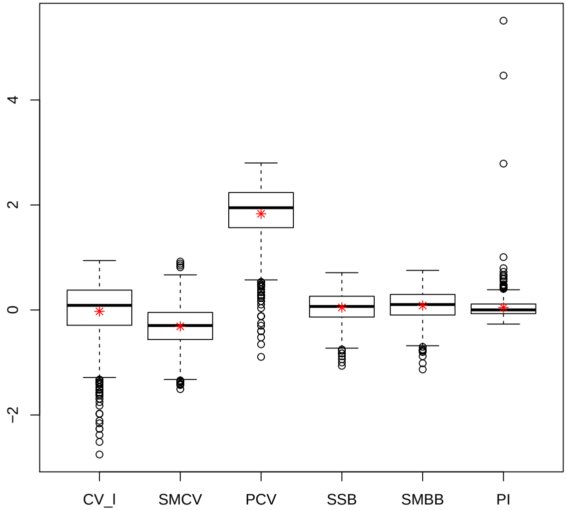
$\log(\hat{h}/h_{MISE})$. Model 5, $\phi = 0.3$



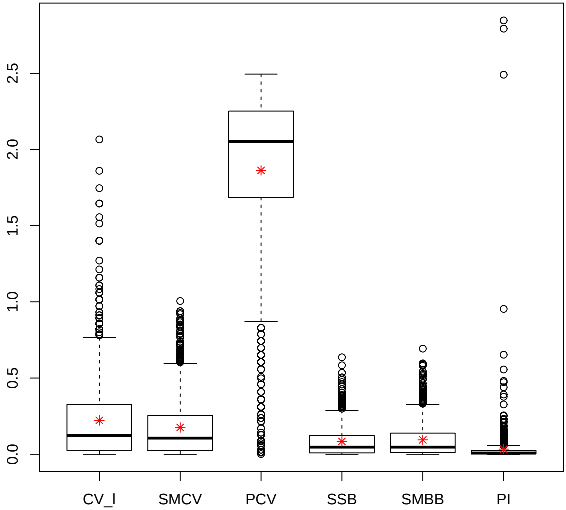
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 5, $\phi = 0.3$



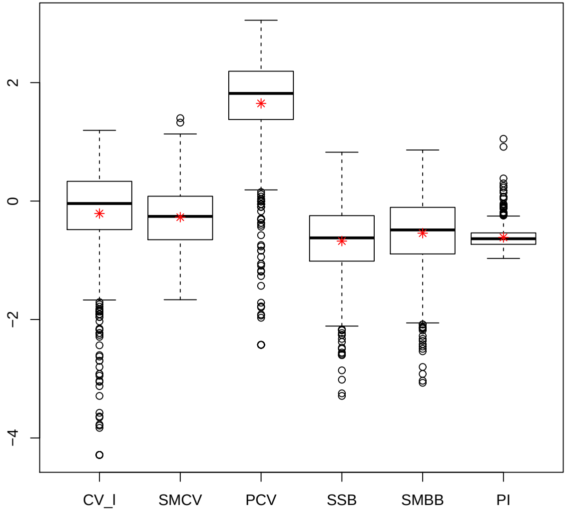
$\log(\hat{h}/h_{MISE})$. Model 5, $\phi = 0.6$



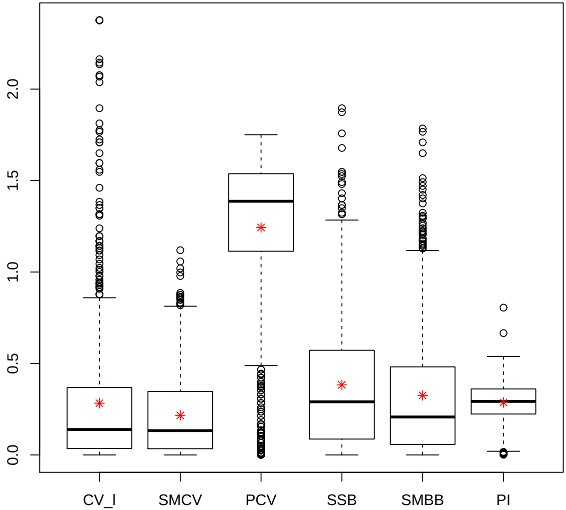
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 5, $\phi = 0.6$



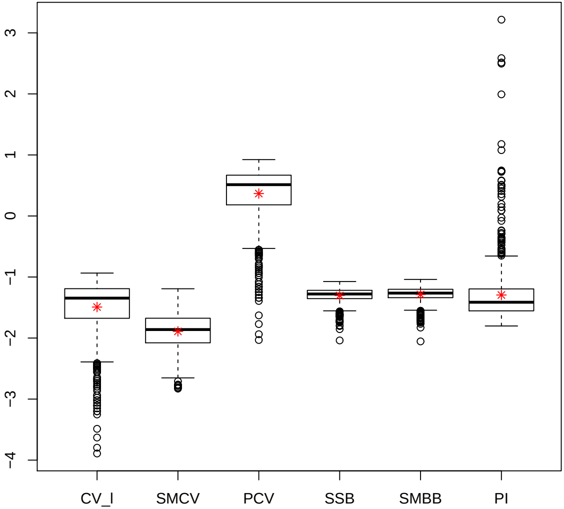
$\log(\hat{h}/h_{MISE})$. Model 5, $\phi = 0.9$



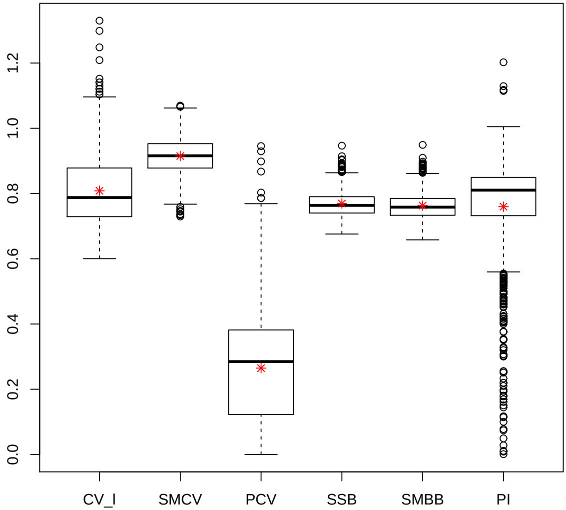
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 5, $\phi = 0.9$



$\log(\hat{h}/h_{MISE})$. Model 6

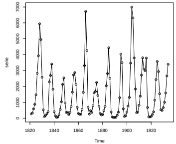


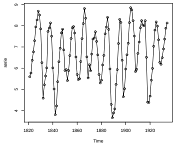
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 6



Real data application: Data sets considered

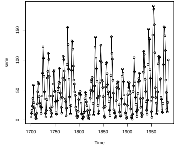
1. **`lynx` data set**: Number of Canadian lynxes trapped (114 observations).

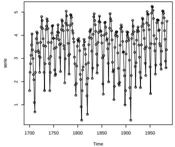




$$(1 - \phi_1 B - \phi_2 B^2) Y_t = c + (1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3)(1 + \Theta_1 B^{12}) a_t.$$

2. **`sunspot.year` data set**: Yearly number of sunspots from 1700 to 1988 (289 observations).

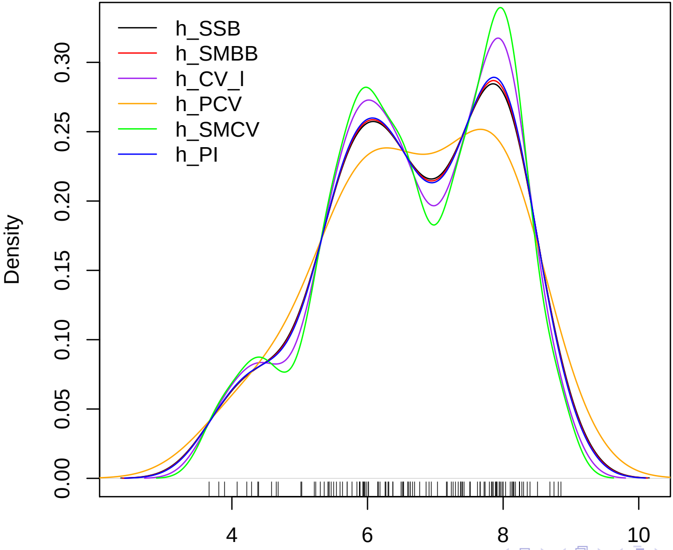




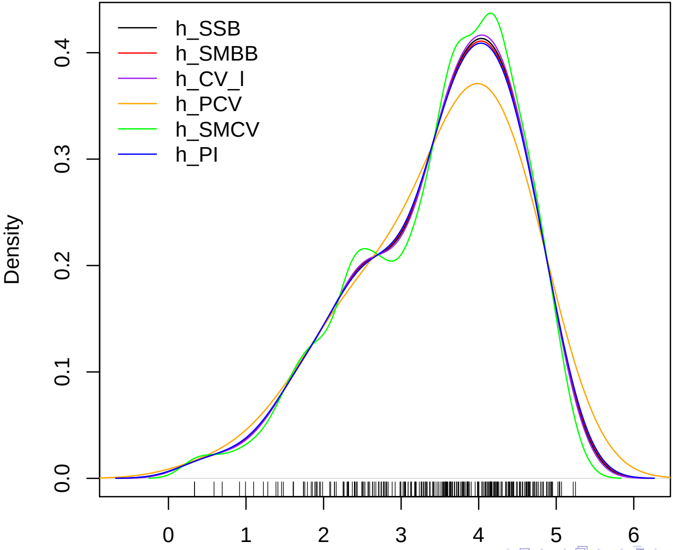
$$(1 - \phi_1 B - \phi_2 B_2 - \phi_2 B^3 - \phi_4 B^4)(1 - B)(1 - B^{12}) Y_t =$$

$$c + (1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3 + \theta_4 B^4) \cdot (1 + B^{12} \Theta_1) a_t.$$

Real data application: `lynx` data set



Real data application: sunspot.year data set



Real data application: Bandwidth parameters

h^*_{SSB}	h^*_{SMBB}	h_{CV_l}	h_{PCV}	h_{SMCV}	h_{PI}
0.4345	0.4246	0.3173	0.6194	0.2585	0.4152

Table: Bandwidth parameters for `lynx` data set.

h^*_{SSB}	h^*_{SMBB}	h_{CV_l}	h_{PCV}	h_{SMCV}	h_{PI}
0.3173	0.3295	0.3002	0.5065	0.196	0.3392

Table: Bandwidth parameters for `sunspot.year` data set.

Main conclusions

- New SSB and SMBB bootstrap resampling plans under dependence.
- Closed expressions for $MISE^*$ under SSB and SMBB. Monte Carlo is not needed.
- Bandwidth selection for the KDE with dependent data:
  - Plug-in
  - Leave-$(2l+1)$-out cross validation
  - Penalized cross validation
  - Modified cross validation
  - Smooth Stationary Bootstrap
  - Smooth Moving Blocks Bootstrap
- Good empirical behaviour of h_{PI}, but sometimes it produces extremely large banwdidths
- h^*_{SSB} and h^*_{SMBB} display the overall best performance.

References

- Barbeito, I. and Cao, R. (2016). Smoothed stationary bootstrap bandwidth selection for density estimation with dependent data. Unpublished manuscript.
- Cao, R. (1993). Bootstrapping the mean integrated squared error. *J. Mult. Anal.* 45, 137-160.
- Efron, B. and Tibishirani, R. (1986). *An Introduction to the Bootstrap*. Chapman and Hall.
- Estévez-Pérez, G., Quintela-del-Río, A. and Vieu, P. (2002). Convergence rate for cross-validatory bandwidth in kernel hazard estimation from dependent samples. *J. Statist. Planning and Inference*, 104, 1-30.
- Hall, P., Lahiri, S.N. and Truong, Y.K.(1995). On bandwidth choice for density estimation with dependent data. *Ann. Statist.*, 23, 6, 2241-2263.

References (continued)

- Hart, J. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.*, 18, 873-890.
- Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17, 1217-1241.
- Liu, R.Y., and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence, in *Exploring the Limits of the Bootstrap*, ed. by R. LePage and L. Billiard. New York: Wiley.
- Politis, D.N. and Romano, J.R. (1994a). The stationary bootstrap. *J. Amer. Statist. Assoc.*, 89, 1303-1313.
- Politis, D.N. and Romano, J.R. (1994b). Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions. *Ann. Statist.*, 22, 2031-2050 .
- Stute, W. (1992). Modified cross-validation in density estimation. *J. Statist. Planning and Inference*, 3, 293-305.

Contact info

Thank you for your attention!

You can contact me at rcao@udc.es

2017
International Workshop
on Functional
and Operational Statistics
IWFOS