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

1. Introduction and Background
   - Aims
   - Smoothed Bootstrap iid case
   - Bootstrap methods for dependent data

2. Already existing smoothing parameter selectors

3. Bootstrap bandwidth selector under independence

4. Smooth Stationary Bootstrap under dependence

5. Smooth Moving Blocks Bootstrap under dependence

6. Simulations

7. Real data application

8. Conclusions

9. References

10. Contact info
General dependent data, \( \{X_t\}_{t \in \mathbb{Z}} \): stationary, \( \alpha \)-mixing, \( \phi \)-mixing, \ldots

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
   
   $X^{\text{NAIVE}*} = (X^{\text{NAIVE}*}_1, \ldots, X^{\text{NAIVE}*}_n)$ from $F_n$

2. Draw a sample $U = (U_1, \ldots, U_n)$ from the density $K$

3. Obtain the smoothed bootstrap resample $X^* = (X^*_1, \ldots, X^*_n)$, where
   
   $X^*_i = X^{\text{NAIVE}*}_i + hU_i$
Motivation & Background

Dependent data: bootstrap methods

Moving Blocks Bootstrap (MBB)

**MBB algorithm**
Künsch (1989), Liu and Singh (1992)

1. Fix the block length, \( b \in \mathbb{N} \), 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

\[
J_n(u, F) = \mathbb{P}(\tau_n(T_n - \theta) \leq u)
\]

3. Fix some \(b \in \mathbb{N}\) such that \(b < n\) and define:

\[
S_{n,i} = T_b(B_{i,b}), \quad i = 1, 2, \ldots, N, \text{ where } N = n - b + 1.
\]

4. Use:

\[
L_n(x) = \frac{1}{N} \sum_{i=1}^{N} 1\{\tau_b(S_{n,i} - T_n) \leq x\}
\]

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:

$$AMISE(h) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2^2R(f'') - h^6\frac{1}{24}\mu_2\mu_4R(f''') + \frac{1}{n}\left(2\sum_{i=1}^{n-1}\left(1 - \frac{i}{n}\right)\int g_i(x, x)dx - R(f)\right).$$

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 $(X_j, X_{i+j})$,

$J_1 = \frac{R(K)}{\mu_2^2R(f'')} \quad$ and $\quad J_2 = \frac{\mu_4R(f''')}{20\mu_2R(f''')}.$

- 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 $J_2$. 
Already existing bandwidth selectors

Plug-in method under dependence (PI)

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

\[
\hat{I}_k = 2\hat{\theta}_{1k} - \hat{\theta}_{2k}, \quad k = 2, 3,
\]

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

\[
\hat{\theta}_{2k} = 2 \left( n(n - 1)h_1^{2(k+1)} \right)^{-1} \sum_{1 \leq i < j \leq n} \sum \int K_{1}^{(k)} \left( \frac{x - X_i}{h_1} \right) K_{1}^{(k)} \left( \frac{x - X_j}{h_1} \right) dx.
\]
Leave-$(2l + 1)$-out cross validation ($CV_l$)

Hart and Vieu (1990)

Define

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

where

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

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 selector is

\[ h_{PCV} = h_{CVl} + \bar{\lambda}. \]

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

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

where \( \hat{\rho} \) is the estimated autocorrelation
Stute (1992) proposed it for independent data

■ Define

$$SMCV(h) = \frac{1}{nh} \int K^2(t)dt + \frac{1}{n(n-1)h} \sum_{i \neq j} \left[ \frac{1}{h} \int K \left( \frac{x - X_i}{h} \right) K \left( \frac{x - X_j}{h} \right) dx \right]$$

$$- \frac{1}{nn_lh} \sum_{j=1}^{n} \sum_{i: |j-i| > l} \left[ K \left( \frac{X_i - X_j}{h} \right) - dK'' \left( \frac{X_i - X_j}{h} \right) \right].$$

■ The $SMCV$ bandwidth selector is

$$h_{SMCV} = \arg \min_h SMCV(h)$$
Exact MISE expression for the iid case

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

where

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

\[ V(h) = \int \text{Var} \left( \hat{f}_h(x) \right) \, dx. \]

Exact expression for \( MISE(h) \):

\[ B(h) = \int (K_h \ast f(x) - f(x))^2 \, dx, \quad \text{and} \]

\[ V(h) = n^{-1} h^{-1} R(K) - n^{-1} \int (K_h \ast f(x))^2 \, dx. \]
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:

\[
h_{MISE}^* = \arg \min_{h > 0} MISE^*(h).
\]
Closed expression for the bootstrap MISE

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

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

where $\ast$ denotes the convolution operator: $f \ast g(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy$.

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 $MISE(h)$:

$$MISE(h) = B(h) + V(h),$$

where

$$B(h) = \int (K_h \ast f(x) - f(x))^2 \, dx,$$

and

$$V(h) = n^{-1} h^{-1} R(K) - \int (K_h \ast f(x))^2 \, dx$$

$$+ 2n^{-2} \sum_{\ell=1}^{n-1} (n - \ell) \int \int K_h(x-y)f(y)(K_h \ast f_\ell(\bullet|y))(x) \, dx \, dy,$$

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

$$\mathbb{P}^* (I_{i+1}^* = 1) = 1 - p,$$

$$\mathbb{P}^* (I_{i+1}^* = 0) = p.$$

Assign $X_i^{*(SB)} |_{I_{i+1}^* = 1} = X_{(j \mod n) + 1}$ and draw $X_i^{*(SB)} |_{I_{i+1}^* = 0}$ 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:

$$MISE^*(h) = n^{-1} h^{-1} R(K)$$

$$+ \left[ \frac{n-1}{n^3} - 2 \frac{1-p - (1-p)^n}{pn^3} + 2 \frac{(n-1)(1-p)^{n+1} - n(1-p)^n + 1-p}{p^2n^4} \right]$$

$$\cdot \sum_{i,j=1}^{n} [(K_h \ast K_g) \ast (K_h \ast K_g)] (X_i - X_j)$$

$$- 2n^{-2} \sum_{i,j=1}^{n} (K_h \ast K_g \ast K_g) (X_i - X_j)$$

$$+ n^{-2} \sum_{i,j=1}^{n} (K_g \ast K_g) (X_i - X_j) + 2n^{-3} \sum_{\ell=1}^{n-1} (n-\ell)(1-p)^\ell$$

$$\cdot \sum_{k=1}^{n} [(K_h \ast K_g) \ast (K_h \ast K_g)] \left( X_k - X_{\lceil(k+\ell-1) \mod n \rceil + 1} \right).$$
**SMBB resampling plan**

1. Fix the block length, $b \in \mathbb{N}$, 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 $X_1^{*(MBB)}, \ldots, X_n^{*(MBB)}$ as the first $n$ components of

   $$(\xi_1,1, \xi_1,2, \ldots, \xi_1,b, \xi_2,1, \xi_2,2 \ldots, \xi_2,b, \ldots, \xi_k,1, \xi_k,2, \ldots, \xi_k,b)$$

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 $i = 1, 2, \ldots, n$
MISE closed expression for SMBB

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

If $b = n$,

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

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

If $b < n$, 

\[
MISE^*(h) = \frac{R(K)}{nh} 
+ \sum_{i=1}^{n} a_i \sum_{j=1}^{n} a_j \cdot \psi(X_i - X_j) 
- \frac{2}{n} \sum_{i=1}^{n} a_i \sum_{j=1}^{n} [(K_h * K_g) * K_g] (X_i - X_j) 
+ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} [K_g * K_g] (X_i - X_j) 
- \frac{b - 1}{n(n - b + 1)^2} \sum_{i=b-1}^{n-b+1} \sum_{j=b}^{n-b+2} \psi(X_i - X_j) 
- \frac{1}{nb \cdot (n - b + 1)^2} \left[ \sum_{i=1}^{b-1} \sum_{j=1}^{b-1} (\min\{i, j\}) \psi(X_i - X_j) \right]
\]
MISE closed expression for SMBB

\[
\begin{align*}
&+ \sum_{i=1}^{b-1} \sum_{j=b}^{n-b+1} \psi(X_i - X_j) + \sum_{i=1}^{b-1} \sum_{j=n-b+2}^{n} (\min\{(n - b + i - j + 1), i\}) \psi(X_i - X_j) \\
&+ \sum_{i=b}^{n-b+1} \sum_{j=1}^{b-1} j \cdot \psi(X_i - X_j) + \sum_{i=n-b+2}^{n} (\min\{(n - i + 1), b\}) \sum_{j=b}^{n-b+1} \psi(X_i - X_j) \\
&+ \sum_{i=b}^{n-b+1} \sum_{j=n-b+2}^{n} (\min\{(n - j + 1), b\}) \cdot \psi(X_i - X_j) \\
&+ \sum_{i=n-b+2}^{n} \sum_{j=1}^{b-1} (\min\{(n - b + j - i + 1), j\}) \psi(X_i - X_j) + b \sum_{i=b}^{n-b+1} \sum_{j=b}^{n-b+1} \psi(X_i - X_j) \\
&+ \sum_{i=n-b+2}^{n} \sum_{j=n-b+2}^{n} (n + 1 - \max\{i, j\}) \psi(X_i - X_j)
\end{align*}
\]
MISE closed expression for SMBB

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

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

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

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

\[ + \sum_{\ell=2}^{b-1} (\ell - 1) \sum_{i=b-1}^{n-b+1} \sum_{j=n-b+3}^{n-b+\ell} \psi(X_i - X_j) + \sum_{k=1}^{b-1} (b - k) \sum_{i=n-b+2}^{n-b+k} \sum_{j=b}^{n-b+2} \psi(X_i - X_j) \]
MISE closed expression for SMBB

considering $a_j$ such that:

$$
\begin{align*}
    a_j &= \begin{cases} 
        \frac{j}{b(n - b + 1)} , & \text{if } j = 1, \ldots, b - 1 \\
        \frac{1}{n - b + 1} , & \text{if } j = b, \ldots, n - b + 1 \\
        \frac{n - j + 1}{b(n - b + 1)} , & \text{if } j = n - b + 2, \ldots, n 
    \end{cases}
\end{align*}
$$
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 \( a_t \sim N(0, 1) \) are independent. Thus \( X_t \sim N(0, 0.42) \)

- **Model 2:**
  \[ X_t = a_t - 0.9a_{t-1} + 0.2a_{t-2}, \]
  where \( a_t \sim N(0, 1) \) are independent. Thus \( X_t \sim N(0, 1.85) \).
Simulated models

- **Model 3:**
  \[ X_t = \phi X_{t-1} + (1 - \phi^2)^{1/2} a_t, \]
  with \( a_t \overset{d}{=} N(0, 1), \phi = 0, \pm 0.3, \pm 0.6, \pm 0.9. \) Thus \( X_t \overset{d}{=} N(0, 1). \)

- **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 \( a_t|_{I_t=1} \overset{d}{=} 0 \) (constant), \( a_t|_{I_t=2} \overset{d}{=} \exp(1), \)
  and \( \phi = 0, 0.3, 0.6, 0.9. \) We have \( X_t \overset{d}{=} \exp(1) \)
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 \( a_t \mid I_t=1 \overset{d}{=} 0 \) (constant), \( a_t \mid I_t=2 \overset{d}{=} \text{Dexp}(1) \), and \( \phi = 0, \pm 0.3, \pm 0.6, \pm 0.9. \) Thus \( X_t \overset{d}{=} \text{Dexp}(1). \)

- **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 \( j = 1, 2, \ \forall t \in \mathbb{Z}, \)
  \( a_t^{(j)} \overset{d}{=} N(0, 0.6) \) independent and \( X_t \overset{d}{=} \frac{1}{2}N(2, 0.8) + \frac{1}{2}N(-2, 0.8) \)
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 \( \hat{h} = h_{CVl}, h_{SMCV}, h_{PCV}, h_{PI}, h_{SSB}^*, h_{SMBB}^* \).
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
Simulations

\[ \log(\text{MISE}(\hat{h})/\text{MISE}(h_{\text{MISE}})). \] Model 1
Simulations

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

![Box plot comparison of different bandwidth selection methods: CV, SMCV, PCV, SSB, SMBB, PI. The x-axis represents different selection criteria, and the y-axis shows the log of the ratio of estimated to optimal bandwidth. Each box plot shows the distribution of the ratios across different datasets or conditions.]
Simulations

\[ \log\left( \frac{MISE(\hat{h})}{MISE(h_{MISE})} \right) \]. Model 2

R. Cao
SSB bandwidth selection
GSNSI, June 8, 2016 35 / 81
$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = -0.9$
Simulations

\[ \log \left( \frac{MISE(\hat{h})}{MISE(h_{MISE})} \right). \] Model 3, \( \phi = -0.9 \)
Simulations

\[ \log(\frac{\hat{h}}{h_{MISE}}) \]. Model 3, \( \phi = -0.6 \)
Simulations

$$\log\left( \frac{MISE(\hat{h})}{MISE(h_{MISE})} \right) \text{. Model 3, } \phi = -0.6$$
$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = -0.3$
\log(\text{MISE}(\hat{h})/\text{MISE}(h_{\text{MISE}})). \text{ Model 3, } \phi = -0.3
$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = 0$
\[ \log(\text{MISE}(\hat{h})/\text{MISE}(h_{\text{MISE}})). \] Model 3, $\phi = 0$
$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = 0.3$
\[ \log\left( \frac{MISE(\hat{h})}{MISE(h_{MISE})} \right). \text{ Model 3, } \phi = 0.3 \]
Simulations

$log(h/h_{MISE})$. Model 3, $\phi = 0.6$
\[ \log\left( \frac{MISE(\hat{h})}{MISE(h_{MISE})} \right). \] Model 3, \( \phi = 0.6 \)
Simulations

$\log(\hat{h}/h_{MISE})$. Model 3, $\phi = 0.9$
\[ \log\left( \frac{MISE(\hat{h})}{MISE(h_{MISE})} \right) \]. Model 3, \( \phi = 0.9 \)
Simulations

\[ \log(\hat{h}/h_{MISE}). \text{ Model 4, } \phi = 0 \]
\[
\log(\text{MISE}(\hat{h})/\text{MISE}(h_{\text{MISE}})). \quad \text{Model 4, } \phi = 0
\]
$\log(\hat{h}/h_{MISE})$. Model 4, $\phi = 0.3$
$\log(MISE(\hat{h})/MISE(h_{MISE}))$. Model 4, $\phi = 0.3$
Simulations

\[ \log \left( \frac{\hat{h}}{h_{MISE}} \right) . \] Model 4, \( \phi = 0.6 \)
\log\left(\frac{\text{MISE}(\hat{h})}{\text{MISE}(h_{\text{MISE}})}\right).\text{ Model 4, } \phi = 0.6
\log(\hat{h}/h_{\text{MISE}})$. Model 4, $\phi = 0.9$.
Simulations

\[ \log(\text{MISE}(\hat{h})/\text{MISE}(h_{\text{MISE}})) \]. Model 4, \( \phi = 0.9 \)
$\log(\hat{h}/h_{MISE})$. Model 5, $\phi = -0.9$
\[
\log \left( \frac{MISE(\hat{h})}{MISE(h_{MISE})} \right). \quad \text{Model 5, } \phi = -0.9
\]
\[ \log(\hat{h}/h_{MISE}) \]. Model 5, \( \phi = -0.6 \)
\log(\text{MISE}(\hat{h})/\text{MISE}(h_{\text{MISE}})). \text{ Model 5, } \phi = -0.6
Simulations

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

$\log(\hat{h}/h_{MISE})$. Model 5, $\phi = 0$
\[ \log \left( \frac{MISE(\hat{h})}{MISE(h_{MISE})} \right). \text{ Model 5, } \phi = 0 \]
\[ \log(\frac{\hat{h}}{h_{MISE}}) \]. Model 5, \( \phi = 0.3 \)
Simulations

\[
\log\left( \frac{MISE(\hat{h})}{MISE(h_{MISE})} \right). \text{ Model 5, } \phi = 0.3
\]
$\log(\hat{h}/h_{MISE})$. Model 5, $\phi = 0.6$
Simulations

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

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

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

Model 5, $\phi = 0.9$
$\log(\hat{h}/h_{MISE})$. Model 6
\[ \log\left( \frac{MISE(\hat{h})}{MISE(h_{MISE})} \right) \]. Model 6
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).

\[
(1 - \phi_1 B - \phi_2 B_2 - \phi_2 B^3 - \phi_4 B^4)(1 - B)(1 - B^{12})Y_t = \\
\quad c + (1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3 + \theta_4 B^4) \cdot (1 + B^{12}\Theta_1)a_t.
\]
Real data application: lynx data set
Real data application: sunspot.year data set

The graph compares different bandwidth selection methods for smoothing the density estimate of the sunspot.year dataset. The methods represented include:

- **h_SSB**: Standard Scott's Rule
- **h_SMBB**: Scott-Muguerza Bandwidth
- **h_CV_I**: Cross-validation with modified Scott's Rule
- **h_PCV**: Plug-in method
- **h_SMCV**: Semi-automatic bias-corrected cross-validation
- **h_PI**: Plug-in method

The x-axis represents the year, and the y-axis shows the density of sunspot activity. Each line corresponds to a different bandwidth selection method, allowing for comparison of how they affect the smoothness and interpretability of the density estimate.
Real data application: Bandwidth parameters

<table>
<thead>
<tr>
<th>$h^*_{SSB}$</th>
<th>$h^*_{SMBB}$</th>
<th>$h_{CV_l}$</th>
<th>$h_{PCV}$</th>
<th>$h_{SMCV}$</th>
<th>$h_{PI}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4345</td>
<td>0.4246</td>
<td>0.3173</td>
<td>0.6194</td>
<td>0.2585</td>
<td>0.4152</td>
</tr>
</tbody>
</table>

**Table:** Bandwidth parameters for lynx data set.

<table>
<thead>
<tr>
<th>$h^*_{SSB}$</th>
<th>$h^*_{SMBB}$</th>
<th>$h_{CV_l}$</th>
<th>$h_{PCV}$</th>
<th>$h_{SMCV}$</th>
<th>$h_{PI}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3173</td>
<td>0.3295</td>
<td>0.3002</td>
<td>0.5065</td>
<td>0.196</td>
<td>0.3392</td>
</tr>
</tbody>
</table>

**Table:** Bandwidth parameters for sunspot.year data set.
Main conclusions

- New SSB and SMBB bootstrap resampling plans under dependence.
- Closed expressions for $\text{MISE}^*$ 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 bandwidths
- $h^*_{SSB}$ and $h^*_{SMBB}$ display the overall best performance.


Thank you for your attention!

You can contact me at rcao@udc.es