A new method of nonparametric density estimation

Andrey Pepelyshev

Institut für Statistik & Wirtschafts-mathematik
RWTH AACHEN

Cardiff
December 7, 2011

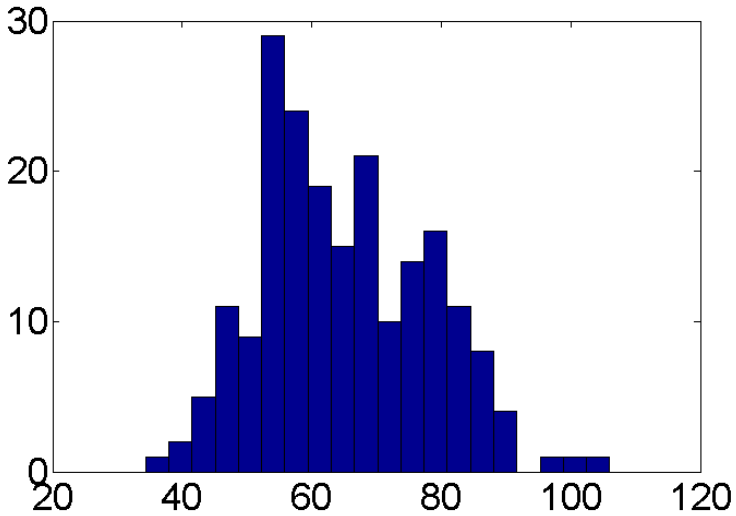
Contents

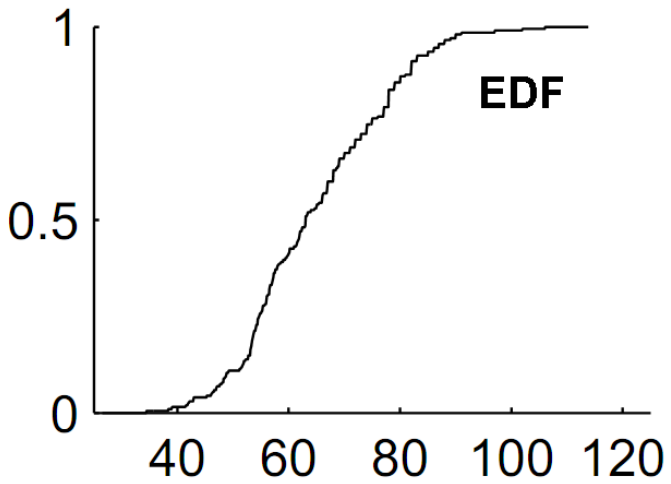
Introduction

Let $(x_1, \ldots, x_m)$ be a sample from a distribution $F(x)$ with density $p(x) = F'(x)$.

The problem is to estimate $p(x)$ and $F(x)$.

Empirical distribution function





The EDF $F_m(x)$ is the minimum-variance unbiased nonparametric estimate of $F(x)$.

$$F_m(x) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{[x_i,\infty)}(x)$$

Kernel estimation

The kernel density estimate is

$$\hat{p}_{m,h}(x) = \frac{1}{mh} \sum_{i=1}^{m} K(\frac{x - x_i}{h})$$

where h is the bandwidth and the kernel K is a continuous function,

$$\int K(x)dx = 1.$$

The popular choice of $K(x)$ is the Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Kernel estimation

The corresponding estimate of $F(x)$ have the form

$$\hat{F}(x) = (K * F_m)(x) = \int K(x-z)F_m(z)dz$$

since the equality

$$\int_{-\infty}^{x} K(y-z)dy = \int_{z}^{\infty} K(y-x)dy$$

holds for all x and z if $K(x)$ is a symmetric kernel.

Bandwidth selection

An optimal bandwidth minimizes a risk function, e.g.

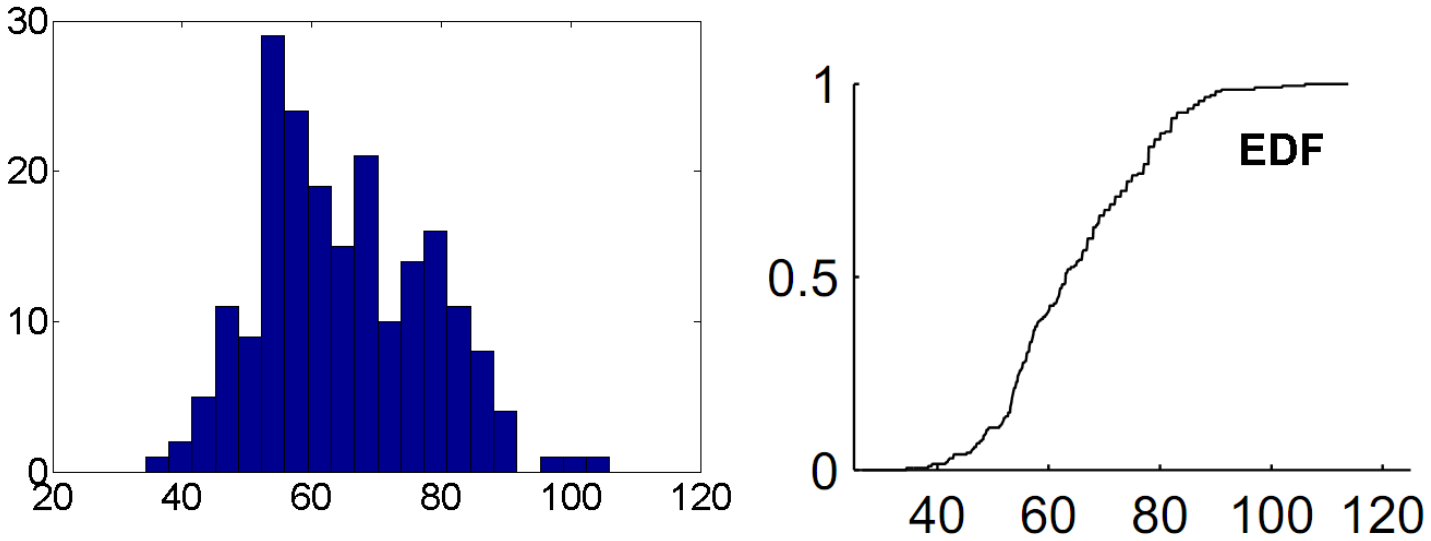
$$\mathrm{MISE}(h) = E \int (p(x) - \hat{p}_{m,h}(x))^2 dx.$$

The LSCV-bandwidth minimizes

$$\Psi(h) = \int \hat{p}_{m,h}(x)^2 dx - \frac{2}{m} \sum_{i=1}^{m} \hat{p}_{m-1,h}^{[x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n]}(x_i)$$

A new estimate

$$p(x) = \frac{d}{dx} F(x)$$



The aim is to obtain a smoothed EDF which can be differentiated.

A new estimate

- Consider values of the EDF at equidistant points as a time series.
- Apply Singular Spectrum Analysis (SSA) for its smoothing.

Golyandina N., Nekrutkin V., Zhigljavsky A. (2001) Analysis of Time Series Structure: SSA and Related Techniques. London: Chapman & Hall/CRC.

The SSA procedure for a CDF series

- Transform a series $(f_1, \ldots, f_N)$ to a matrix $\mathbb{X}$ as $\mathbb{X} = (x_{i,j}) = (f_{i+j-1})$, $i = 1, \ldots, L$, $j = 1, \ldots, N-L+1$.
- Make the SVD decomposition $\mathbb{X} = \sum_{i=1}^{L} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$, $\lambda_1 \geq \lambda_2 \geq \ldots$.

$$\mathbb{X} = \begin{array}{|ccccc|} \hline f_1 & f_2 & f_3 & \cdots & f_{N-L+1} \\ f_2 & f_3 & f_4 & \cdots & f_{N-L+2} \\ f_3 & f_4 & f_5 & \cdots & f_{N-L+3} \\ \vdots & \vdots & \vdots & & \vdots \\ f_L & f_{L+1} & f_{L+2} & \cdots & f_N \\ \hline \end{array}$$

The SSA procedure for a CDF series

- Transform a series $(f_1, \ldots, f_N)$ to a matrix $\mathbb{X}$ as $\mathbb{X} = (x_{i,j}) = (f_{i+j-1})$, $i = 1, \ldots, L$, $j = 1, \ldots, N - L + 1$.
- Make the SVD decomposition $\mathbb{X} = \sum_{i=1}^{L} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$, $\lambda_1 \geq \lambda_2 \geq \ldots$.
- Extract the r leading components $\mathbb{X}^{(r)} = \left(x_{i,j}^{(r)}\right) = \sum_{i=1}^{r} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$.
- Transform $\mathbb{X}^{(r)}$ back to the series

$$\hat{f}_j = \hat{f}_j(L, r) = \begin{cases} 0 & 1 \leq j < L, \\ \dfrac{1}{L} \displaystyle\sum_{k=1}^{L} x_{k,j-k+1}^{(r)} & L \leq j \leq K, \\ 1 & K < j \leq N, \end{cases}$$

Representation of the SSA procedure

$$\hat{f}_j = \sum_{i=1}^{L} \sum_{l=1}^{L} (u_{1,i}u_{1,l} + \ldots + u_{r,i}u_{r,l}) f_{j+i-l}/L$$

for $j = L, \ldots, K$, where $(u_{l,1}, \ldots, u_{l,L})^{\mathrm{T}} = U_l$.

- This is a data-adaptive filter.
- L is half of the length of the filter.
- r is a complexity of the filter.
- L controls the smoothness.

Selection of points t_j for time series

$$f_j = F_m(t_j),\ t_{j+1} - t_j = \delta$$

Choose δ such that

$$\delta \approx \frac{1}{m \max_x p(x)}$$

and define t_j as

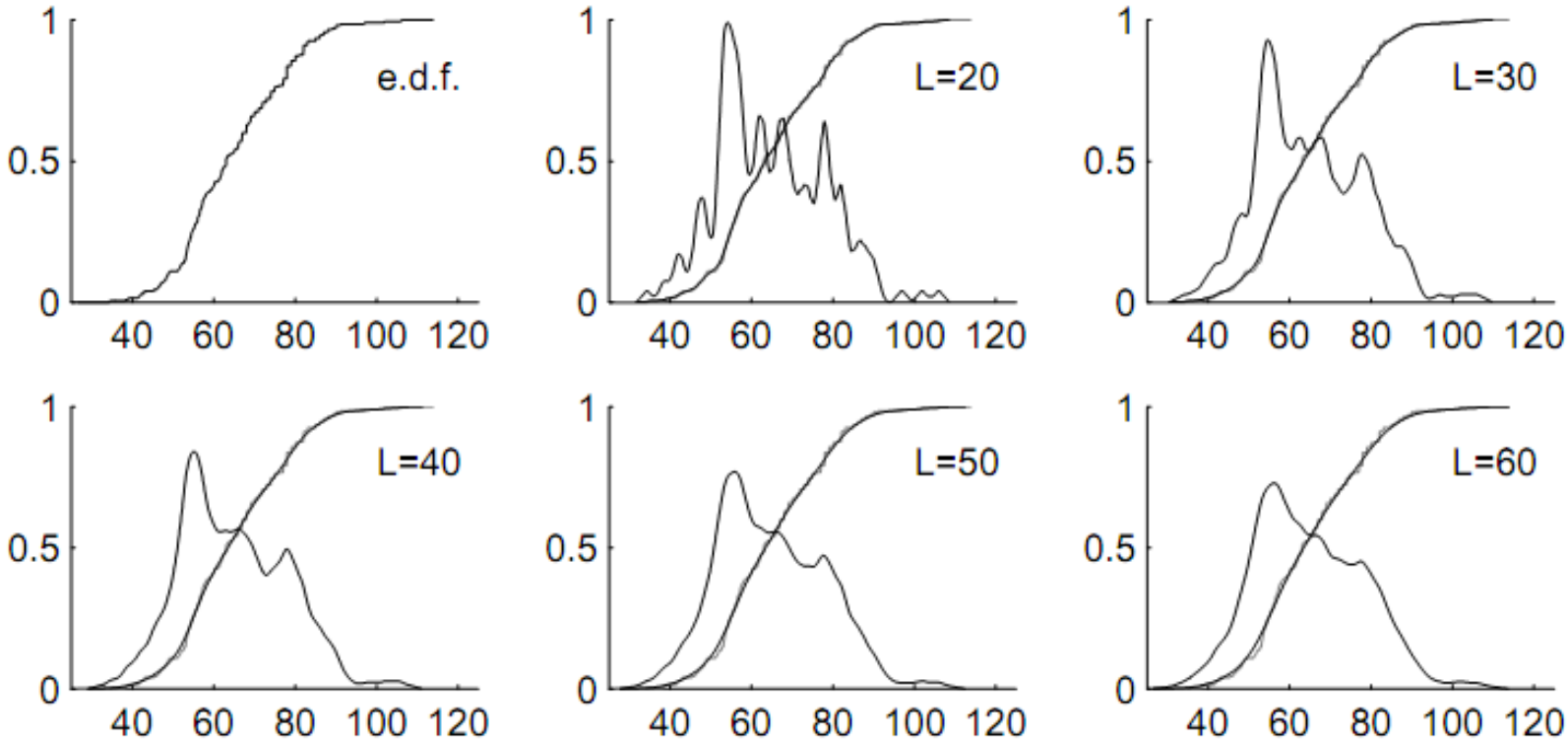
$$t_j = \delta(j - 2L) + \min_{i=1,\dots,m} x_i,$$

$$j = 1, \dots, N,\ N = 2L_{\max} + 4L,$$

$$L_{\max} = \left\lfloor \frac{\max_i x_i - \min_i x_i}{2\delta} \right\rfloor.$$

Influence of the parameter L

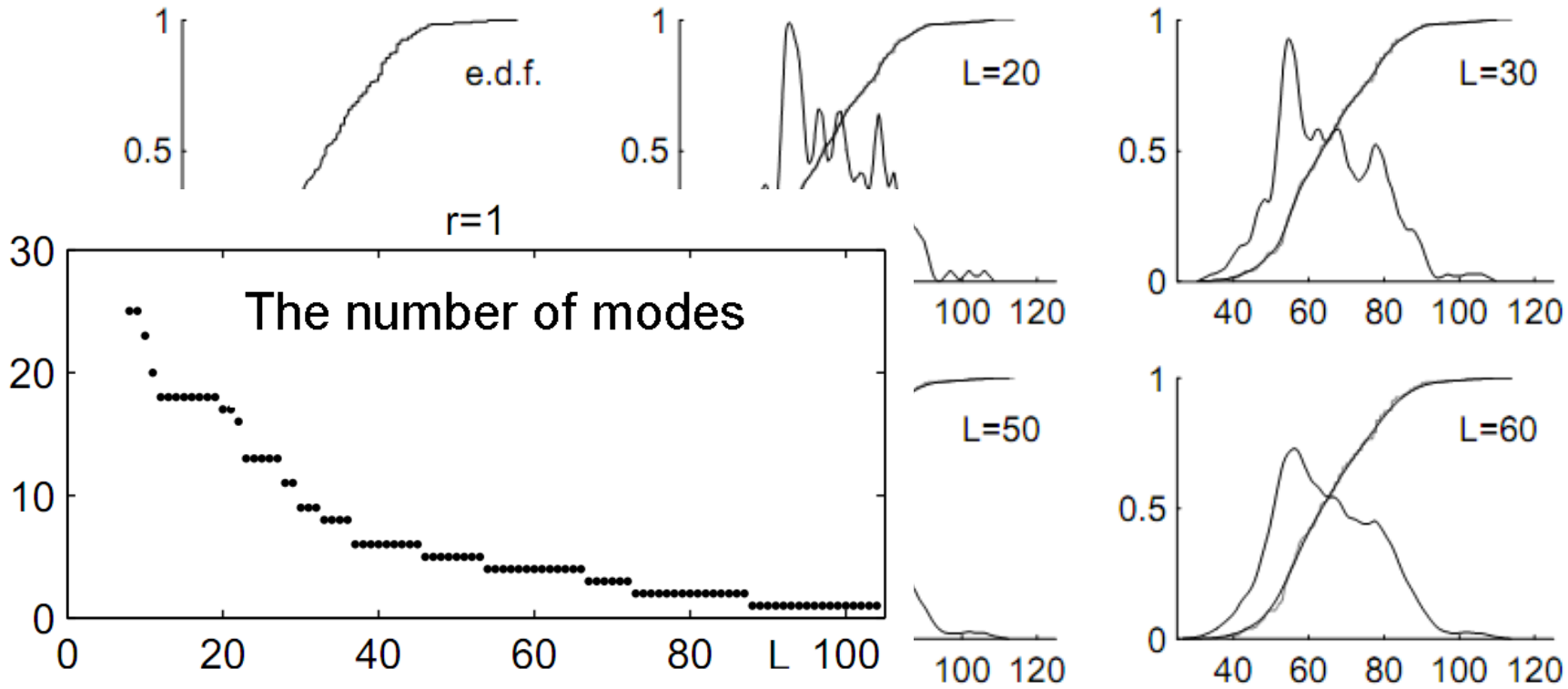
Measurements of Lean Body Mass from the Australian Institute of Sport Data



The SSA^{1c} estimate (i.e. with $r = 1$) of the distribution function and the density for $L = 20, \ldots, 60$

Influence of the parameter L

Measurements of Lean Body Mass from the Australian Institute of Sport Data



The SSA^{1c} estimate (i.e. with $r = 1$) of the distribution function and the density for $L = 20, \ldots, 60$

Specific SSA estimates

- SSA^{1c}: the SSA estimate with $r = 1$
- SSA^{2c}: the SSA estimate with $r = 2$
- SSA^{b}: the SSA estimate with $r = 1$ and bias correction

Validity of the SSA procedure

The estimated series $(\hat{f}_1, \ldots, \hat{f}_N)$ can be transformed to the SSA estimate of $F(x)$ as follows

$$\hat{F}_m(x) = \sum_{j=2}^{N} \left(\hat{f}_{j-1} + (\hat{f}_j - \hat{f}_{j-1}) \frac{x - t_{j-1}}{t_j - t_{j-1}} \right) \mathbf{1}_{[t_{j-1}, t_j)}(x)$$

$$+ \mathbf{1}_{[t_N, \infty)}(x),$$

Lemma.
The SSA^{1c} estimate is a valid distribution function.

The SSA^b estimate

Let $S(F)$ be a filter (weighted moving average).
Our case: $S(F)$ is the SSA^{1c} estimate.
The problem: this filter has a bias.

The idea is to correct $S(F)$ as an estimate F by use of $S(S(F))$ and $S(S(S(F)))$.

$$S^b(\mathcal{F}) = 3S(\mathcal{F}) - 3S^2(\mathcal{F}) + S^3(\mathcal{F})$$

Simulation results for different L

The performance of SSA estimates for samples of size 100, the replication number is 10000.

	$\mathbf{E}D_{\mathrm{ISE}}$	$\mathbf{E}D_{\mathrm{KS}}$	$\mathbf{E}D_{\mathrm{H}}$
model $0.4N(0,1)+0.6N(5,2^2)$			
Kernel est. with h_{LSCV}	0.0058	0.0617	0.0231
SSA^{1c} est. with $L=50$	0.0044	0.0595	0.0172
SSA^{1c} est. with $L=60$	0.0044	0.0593	0.0179
SSA^{1c} est. with $L=70$	0.0048	0.0603	0.0204
SSA^{2c} est. with $L=80$	0.0045	0.0598	0.0190
SSA^{2c} est. with $L=90$	0.0042	0.0591	0.0179
SSA^{2c} est. with $L=100$	0.0042	0.0588	0.0174
SSA^{b} est. with $L=70$	0.0044	0.0608	0.0193
SSA^{b} est. with $L=80$	0.0042	0.0602	0.0182
SSA^{b} est. with $L=90$	0.0044	0.0607	0.0185

How to choose L automatically

A law of the iterated logarithm.

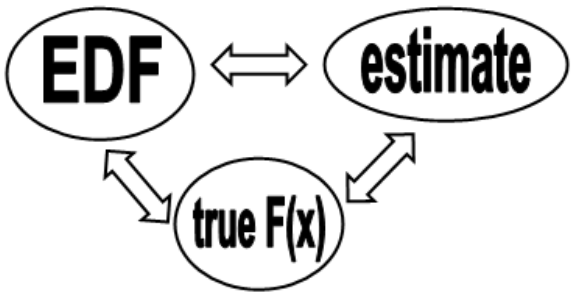
Glivenko-Cantelli result

The empirical distribution function $F_m(x)$ satisfies

$$\limsup_{m\to\infty} R_m \left\| F_m(x) - F(x) \right\|_\infty \le 1$$

almost surely, where $R_m = \dfrac{2\sqrt{m}}{\sqrt{2\ln\ln m}}$.

If an estimate is far from the EDF,
then this estimate is not good.



How to choose L automatically

1. Compute

$$\bar{L} = \max\left\{L : \left\|\mathcal{F}_m - \hat{\mathcal{F}}_m(l)\right\|_\infty \le 1/R_m \;\forall\, l \in \{1,\dots,L\}\right\}.$$

2. Compute the sequence $M_1,\dots,M_{\bar{L}}$, where M_j is the number of modes of estimated density for $L = j$. This sequence has decrasing tendency.

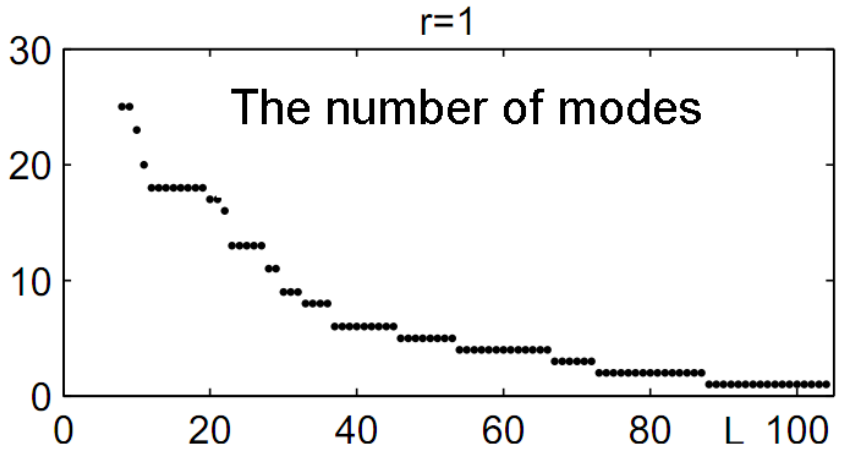
3. Compute $\check{M}_j = \min\{M_1, M_2, \dots, M_j\}$.

How to choose L automatically

4. Divide the set $\{1, 2, \ldots, \bar{L}\}$ into groups such that the values $\check{M}_j$ are equal to each other within each group, i.e. define $a_1, b_1, \ldots, a_k, b_k$ and k such that

$$1 = a_1 \leq b_1 < \ldots < a_k \leq b_k = \bar{L},$$

$a_{i+1} = b_i + 1$ and $\check{M}_j = \check{M}_j$ for all $i, j \in \{a_l, \ldots, b_l\}$, $l \in \{1, \ldots, k\}$.



How to choose L automatically

5. Finally, compute L_a (a 'best' value of L) as an average with weight coefficients, which are proportional to the sizes of these k groups, namely

$$L_a = \left\lfloor \sum_{i=1}^{k} c_i w_i \right\rfloor,$$

where

$$c_i = \gamma_i a_i + (1-\gamma_i) b_i, \;\; w_i = \frac{b_i - a_i}{\sum_{j=1}^{k} b_j - a_j},$$

$\gamma_i = 1/2$ if $\check{M}_{a_i} = 1$ and $\gamma_i = 0.9$ otherwise.

How to choose the bandwidth

1. Compute

$$\bar{h} = \max\left\{h : \left\|\mathcal{F}_m - \hat{\mathcal{F}}_m(\hbar)\right\|_\infty \le 1/R_m \;\forall\, \hbar \in (0,h)\right\}.$$

2. Define a dense set $\{h_1, \ldots, h_n\} \in (0, \bar{h})$ and compute the sequence $M_1, \ldots, M_n$, where M_j be the number of modes of estimated density for $h = h_j$.

3. Compute $\check{M}_j = \min\{M_1, M_2, \ldots, M_j\}$.

4. Divide the set $\{h_1, \ldots, h_n\}$ into groups as earlier.

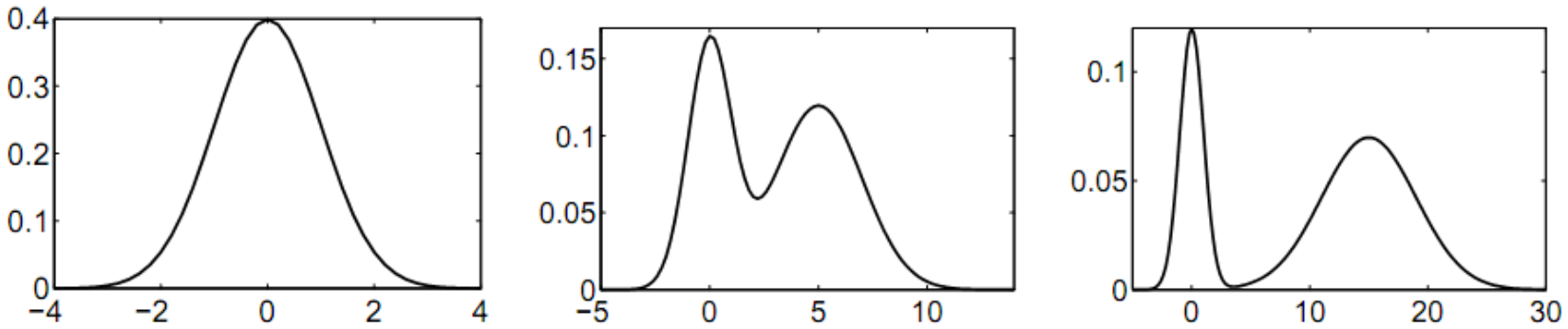
5. Compute $h_a = \sum_{i=1}^{k} c_i w_i$, where c_i and w_i are defined in the same manner.

Consistency

Lemma. The kernel estimate with h_a and any SSA estimate with L_a of the distribution function are consistent.

Simulation study

Consider several models of the density $p(x)$



and simulate 10000 samples of size 100.

$$D_{\mathrm{ISE}}(\hat{p}, p) = \int \big(\hat{p}(x) - p(x)\big)^2 dx$$

$$D_{\mathrm{KS}}(\hat{F}, F) = \big\|\hat{F} - F\big\|_\infty = \max_x \big|\hat{F}(x) - F(x)\big|$$

$$D_{\mathrm{H}}(\hat{p}, p) = \int \left(\sqrt{\hat{p}(x)} - \sqrt{p(x)}\right)^2 dx$$

Results on bandwidth selection

	$\mathbf{E}D_{\mathrm{ISE}}$	$\mathbf{E}D_{\mathrm{KS}}$	$\mathbf{E}D_{\mathrm{H}}$
model $N(0,1)$			
Kernel est. with h_{LSCV}	0.0071	0.0551	0.0143
Kernel est. with h_{SJPI}	0.0066	0.0536	0.0131
Kernel est. with h_{ICV}	0.0075	0.0546	0.0146
Kernel est. with h_a	0.0063	0.0546	0.0131
model $0.4N(0,1)+0.6N(5,2^2)$			
Kernel est. with h_{LSCV}	0.0058	0.0617	0.0231
Kernel est. with h_{SJPI}	0.0052	0.0609	0.0210
Kernel est. with h_{ICV}	0.0055	0.0614	0.0229
Kernel est. with h_a	0.0053	0.0611	0.0236
model $0.3N(0,1)+0.7N(15,4^2)$			
Kernel est. with h_{LSCV}	0.0048	0.0672	0.0394
Kernel est. with h_{SJPI}	0.0069	0.0733	0.0602
Kernel est. with h_{ICV}	0.0049	0.0670	0.0396
Kernel est. with h_a	0.0050	0.0674	0.0440

Results on the SSA estimates

	$\mathbf{E}D_{\mathrm{ISE}}$	$\mathbf{E}D_{\mathrm{KS}}$	$\mathbf{E}D_{\mathrm{H}}$	$\mathbf{E}L_a$
model $N(0,1)$				
Kernel est. with h_a	0.0063	0.0546	0.0131	
SSA^{1c} est.	0.0061	0.0537	0.0128	107.6
SSA^{2c} est.	0.0060	0.0503	0.0142	145.7
SSA^{b} est.	0.0052	0.0488	0.0141	141.7
model $0.4N(0,1)+0.6N(5,2^2)$				
Kernel est. with h_a	0.0053	0.0611	0.0236	
SSA^{1c} est.	0.0051	0.0607	0.0215	67.6
SSA^{2c} est.	0.0047	0.0610	0.0195	86.4
SSA^{b} est.	0.0052	0.0617	0.0206	89.7
model $0.3N(0,1)+0.7N(15,4^2)$				
Kernel est. with h_a	0.0050	0.0674	0.0440	
SSA^{1c} est.	0.0053	0.0679	0.0451	40.9
SSA^{2c} est.	0.0046	0.0660	0.0380	60.0
SSA^{3c} est.	0.0043	0.0643	0.0349	73.8
SSA^{b} est.	0.0047	0.0670	0.0391	48.9

Application

In quality control, acceptance sampling is used to determine whether to accept or reject a large production lot by checking a small number of items.

A sampling plan consists of

- the number n of items to be measured and
- the critical value c such that a lot is rejected if an appropriate statistic is larger than c.

Sampling plans for quality control

Let measurements be distributed according to a distribution $G(x)$ with mean a and variance σ^2. The asymptotically optimal sampling plan (n, c) is

$$n = \left\lceil \frac{\left(\Phi^{-1}(\alpha) - \Phi^{-1}(1-\beta)\right)^2}{\left(F^{-1}(\mathrm{AQL}) - F^{-1}(\mathrm{RQL})\right)^2} \right\rceil,$$

$$c = -\frac{\sqrt{n}}{2}\left(F^{-1}(\mathrm{AQL}) + F^{-1}(\mathrm{RQL})\right),$$

where $F(x) = G((x-a)/\sigma)$, AQL is the acceptable quality level, RQL is the rejectable quality level, α is the producer risk and β is the consumer risk.

Comparison of sampling plans

Means and standard deviations of sampling plan size using the empirical distribution function, the kernel density estimate and the SSA^{1c}, SSA^{b} and SSA^{2c} estimates for samples of size m from $0.4N(0,1)+0.6N(5,2^2)$.
The true sampling plan size is 392.

	$m=250$	$m=500$	$m=1000$
EDF	469.7(468.9)	462.7(278.9)	422.0(160.1)
Kernel h_{LSCV}	322.3(102.8)	324.0(79.7)	337.8(59.2)
Kernel h_a	302.1(77.2)	321.2(65.2)	339.1(52.8)
SSA^{1c} est.	301.4(71.8)	316.3(61.3)	329.1(46.9)
SSA^{b} est.	409.6(112.9)	408.9(88.8)	403.6(67.7)
SSA^{2c} est.	390.4(107.8)	393.1(81.4)	392.8(62.1)

Thank you for your attention!