# An Improved Estimator for Removing Boundary Bias in Kernel CDF Estimation

Jan Koláček

Department of mathematics and statistics
Faculty of Science
Masaryk University
Brno, Czech Republic
www.muni.cz

# Contents

- Introduction
- Kernel distribution estimators
- Boundary effects
- Proposed estimator
- Examples
- References

# Kernel function

Let $\nu$, $k$ be nonnegative integers, $0 \le \nu \le k-2$, $k \le k_0$, $\nu + k$ even integer. Let $K$ be a real valued function continuous on $\mathbb{R}$ and satisfying conditions

$$K \in Lip\,[-1,1], support(K) = [-1,1]$$

$$\int_{-1}^{1} x^j K(x)dx = \begin{cases} 0, \quad 0 \le j < k,\ j \ne \nu \\ (-1)^\nu\ \nu!,\ j = \nu \\ \beta_k \ne 0, \quad j = k\,. \end{cases}$$

Such a function $K$ is called a *kernel of order $k$* and a class of such functions is denoted by $S_{\nu,k}$.

## Table of kernels

| $\nu$ | $k$ | Kernel (on $[-1,1]$) |
|---|---|---|
| 0 | 2 | $K_{0,2}(x) = \frac{3}{4}(1-x^2)$ |
| 0 | 2 | $K_{0,2}(x) = \frac{15}{16}(1-x^2)^2$ |
| 0 | 2 | $K_{0,2}(x) = \frac{35}{32}(1-x^2)^3$ |
| 0 | 4 | $K_{0,4}(x) = \frac{15}{32}(x^2-1)(7x^2-3)$ |
| 2 | 4 | $K_{2,4}(x) = \frac{105}{16}(1-x^2)(5x^2-1)$ |
| 1 | 3 | $K_{1,3}(x) = \frac{15}{4}x(1-x^2)$ |

# Kernel distribution estimators

Let $X_1,\ldots,X_n$ be independent real random variables each having the same cumulative distribution $F$. Our model is defined by the assumption $F \in C^{k_0}$, where $k_0$ is a positive integer.

For the given data set the corresponding kernel estimate of a distribution function $F$ is

$$\widehat{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right), \qquad W(x) = \int_{-1}^{x} K(t)dt \tag{1}$$

where $h$ is a smoothing parameter called *bandwidth* ($h = h(n)$ is a non-random sequence of positive numbers) and $K \in S_{0,2}$, $K(x) \geq 0$ on $[-1, 1]$.

# Optimal bandwidth

Under additional assumptions $\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh = \infty$ it can be shown (e.g. Bowman, A., Hall, P., Prvan, T. [2]) that the leading term of MISE (Mean Integrated Square Error) takes the form

$$\overline{\mathrm{MISE}}(\widehat{F}_{h,K}) = \underbrace{\frac{1}{n}\int F(x)(1-F(x))dx - q_1\frac{h}{n}}_{\overline{var}(\widehat{F}_{h,K})} + \underbrace{q_2 h^4}_{\overline{bias}^2(\widehat{F}_{h,K})},$$

$$q_1 = \int_{-1}^{1} W(x)(1-W(x))dx > 0, \qquad q_2 = \frac{\beta_2^2}{4}\int (F^{(2)}(x))^2 dx.$$

Hence, the optimal bandwidth $h^F_{opt,0,2}$ minimizing $\overline{\mathrm{MISE}}$ with respect to $h$ is

$$h^F_{opt,0,2} = n^{-1/3}\left(\frac{q_1}{4q_2}\right)^{1/3}. \tag{2}$$

# Boundary Effects

**Assumptions:**

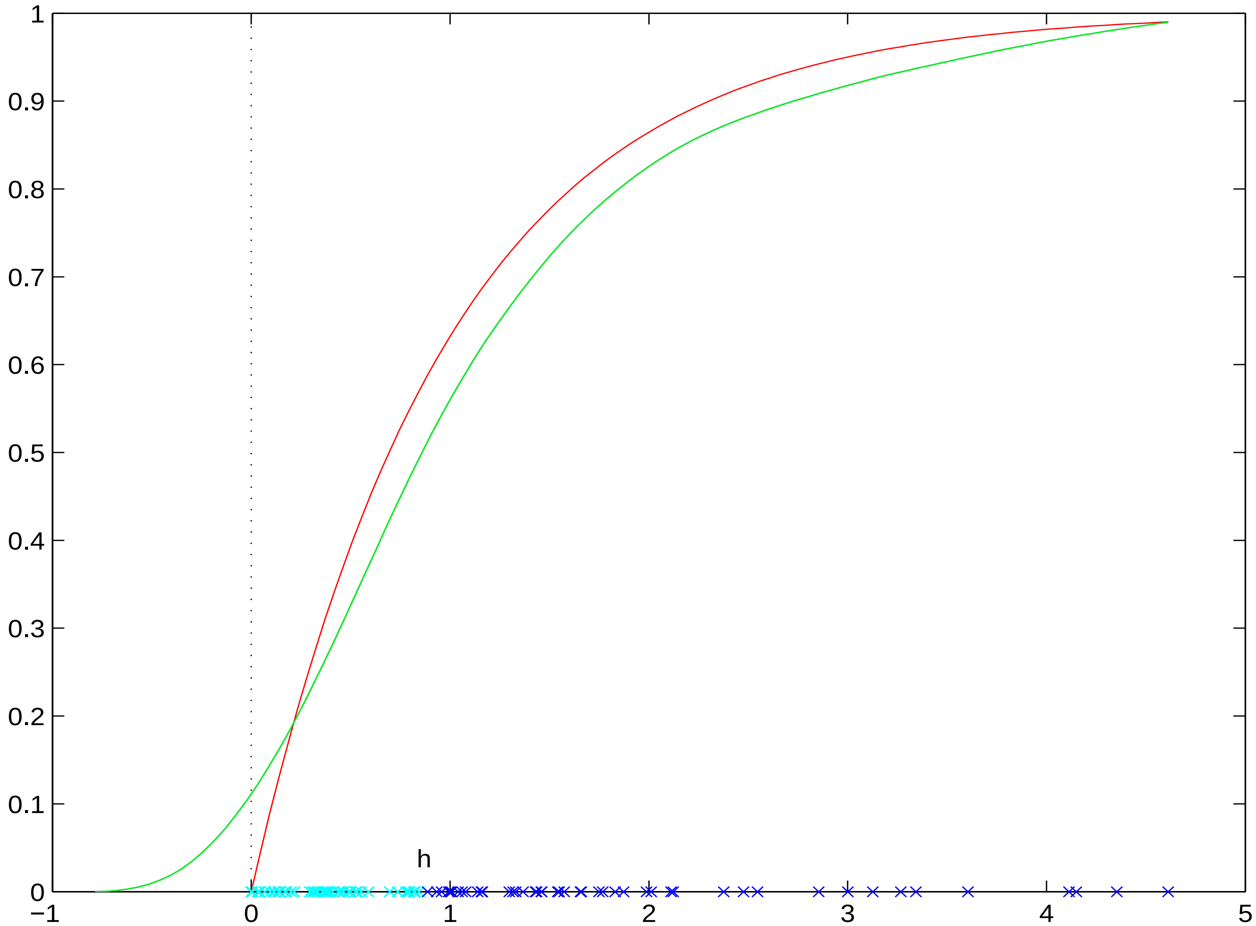
- $X_i,\ i = 1, \ldots, n$ are nonnegative
- the distribution function $F$ has a support $[0, \infty)$
- $f(0) \neq 0$

Boundary effects arise by estimates in points "near" the left boundary, it is for $x \in [0, h]$.

In next, we will write

$$x = ch,\ 0 \leq c \leq 1.$$

$X \sim Exp(1)$ – the kernel estimate of $F$ $(n = 100,\ h^F_{opt,0,2} = 0.8479)$



The *Bias* of $\widehat{F}_{h,K}(x)$ in $x = ch$,

- "near" the left boundary ($0 \le c < 1$):

$$\mathrm{E}(\widehat{F}_{h,K}(x)) - F(x) = h f(0) \int_{-1}^{-c} W(t)dt + h^2 f^{(1)}(0) \left\{ \frac{c^2}{2} + c \int_{-1}^{-c} W(t)dt - \int_{-1}^{c} tW(t)dt \right\} + o(h^2)$$

- interior points ($c \ge 1$):

$$\mathrm{E}(\widehat{F}_{h,K}(x)) - F(x) = \frac{h^2}{2} f^{(1)}(0) \int_{-1}^{1} tW(t)dt + o(h^2)$$

# Possible solutions

- *boundary kernels* – estimators could be negative, some remedies have been proposed
- *pseudo-data* – generating some extra data nearby the boundary and then combining them with the original data
- *data transformation*
  - (a) a transformation is selected from a parametric family,
  - (b) a kernel estimator is applied to transformed data,
  - (c) estimated values are converted by an inverse formula
- *reflection method* – reflecting the data and applying the classical kernel estimator

$$\widehat{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{W\left(\frac{x-X_i}{h}\right) - W\left(-\frac{x+X_i}{h}\right)\right\} \qquad (3)$$

# Proposed estimator

“Generalized” reflection method

(Zhang et al. [10], Karunamuni and Alberts [5] – the density case)

$$\widetilde{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right)\right\}$$

$$g_1 = g_2 \Rightarrow \widetilde{F}_{h,K}(0) = 0$$

Set $g := g_1 = g_2$

- $g$ is nonnegative, continuous and monotonically increasing function defined on $[0,\infty)$
- $g^{-1}$ exists
- $g(0) = 0$
- $g^{(1)}(0) = 1$
- $g^{(2)}$ exists and is continuous on $[0,\infty)$.

The bias of $\widetilde{F}_{h,K}(x)$ at $x = ch$, $\boxed{0 \le c < 1}$

$$\mathrm{E}(\widetilde{F}_{h,K}(x)) - F(x) = h^2 \Big\{ f^{(1)}(0)[c^2/2 + 2cI_1 - I_2] - f(0) g^{(2)}(0)[c^2 + 2cI_1 - I_2] \Big\} + O(h^3),$$

where $I_1 = \int_{-1}^{-c} W(t)dt,\ I_2 = \int_{-c}^{c} tW(t)dt$

The bias of $\widetilde{F}_{h,K}(x)$ at $x = ch$, $\boxed{c \ge 1}$

$$\mathrm{E}(\widetilde{F}_{h,K}(x)) - F(x) = \frac{1}{2} h^2 \Big\{ f^{(1)}(0)\beta_2 - f(0) g^{(2)}(0)[c^2 + \beta_2] \Big\} + O(h^3)$$

Set

$$g^{(2)}(0) = \begin{cases} d_1 \frac{\frac{c^2}{2}+2cI_1-I_2}{c^2+2cI_1-I_2}, & \text{for } 0 \le c < 1 \\ d_1 \frac{\beta_2}{c^2+\beta_2}, & \text{for } c \ge 1 \end{cases} \qquad (= A_c)$$

where

$$d_1 = \frac{f^{(1)}(0)}{f(0)}.$$

# A construction of $g(y)$

**An estimate of $d_1$**

$$d_1 = \frac{f^{(1)}(0)}{f(0)} = (\ln f(x))^{(1)}_{x=0} \approx \hat{d}_1 = \frac{\ln f^*(h_1) - \ln f^*(0)}{h_1}, \; h_1 \approx n^{-\frac{1}{6}}$$

(see Zhang et al. [10],
Karunamuni R.J., Alberts T. [5])

Hence $\hat{d}_1 \Rightarrow \widehat{A}_c$

$$\widehat{g}_c(y) = \lambda \widehat{A}_c^2 y^3 + \frac{1}{2}\widehat{A}_c y^2 + y,$$

where $\lambda$ is a positive constant such that $\lambda > \frac{1}{12}$.

(our experience: $\lambda = 0.1$)

# A simulation study

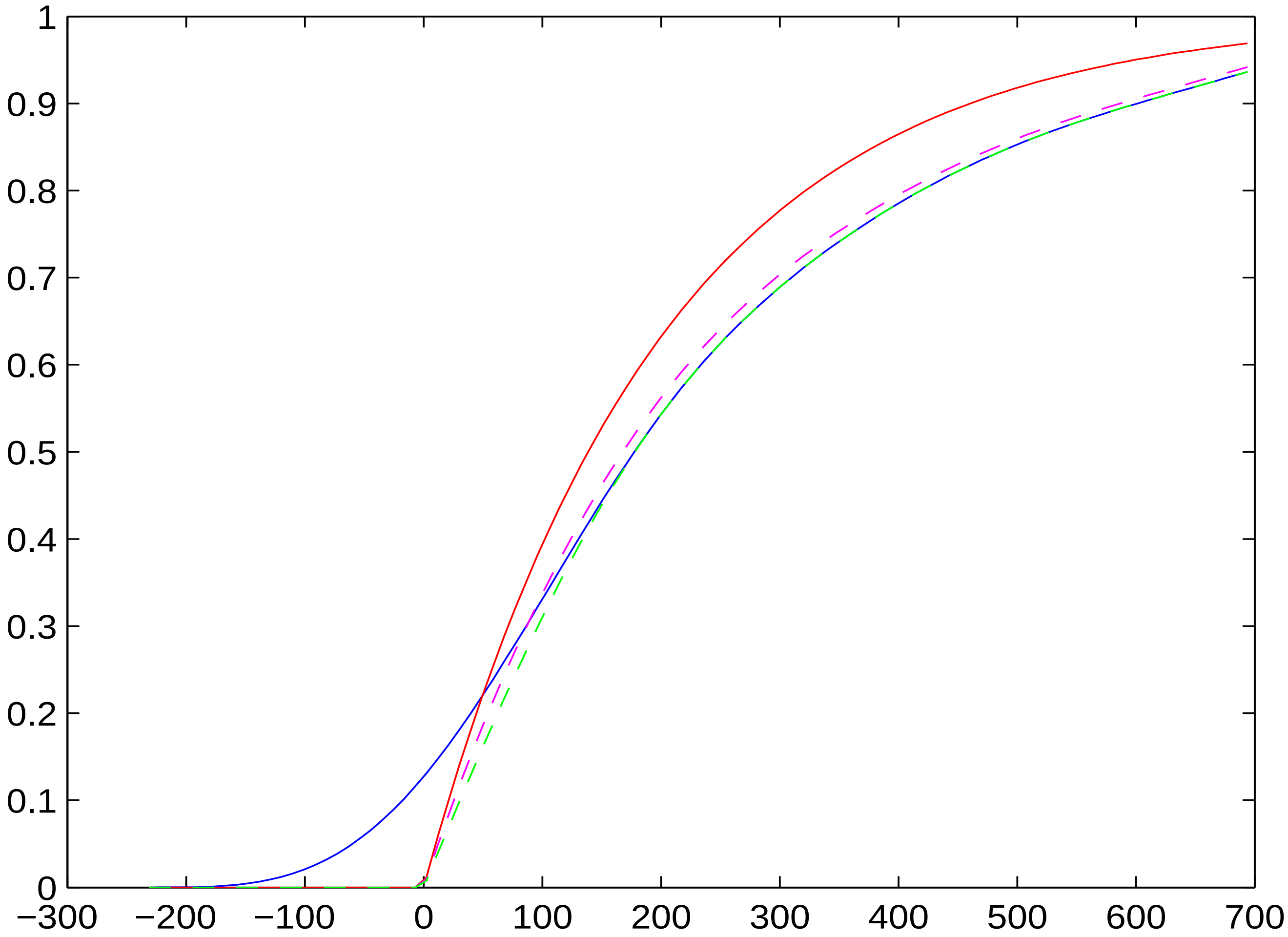
- $X \sim Exp(0.005)$, $n = 100$ (Dette, H., Weissbach, R. [3])
- 1 000 replications
- We used the quartic kernel

$$K_{0,2}(x) = \frac{15}{16}(1 - x^2)^2 I_{[-1,1]},$$

where $I_A$ is the indicator function on the set $A$.

- The optimal bandwidth was computed from (2)
- The results were compared with classical estimator (1) and the reflection method (3)

$X \sim Exp(0.005)$ – the kernel estimate of $F$
$(n = 100,\ h^F_{opt,0,2} = 231.35)$

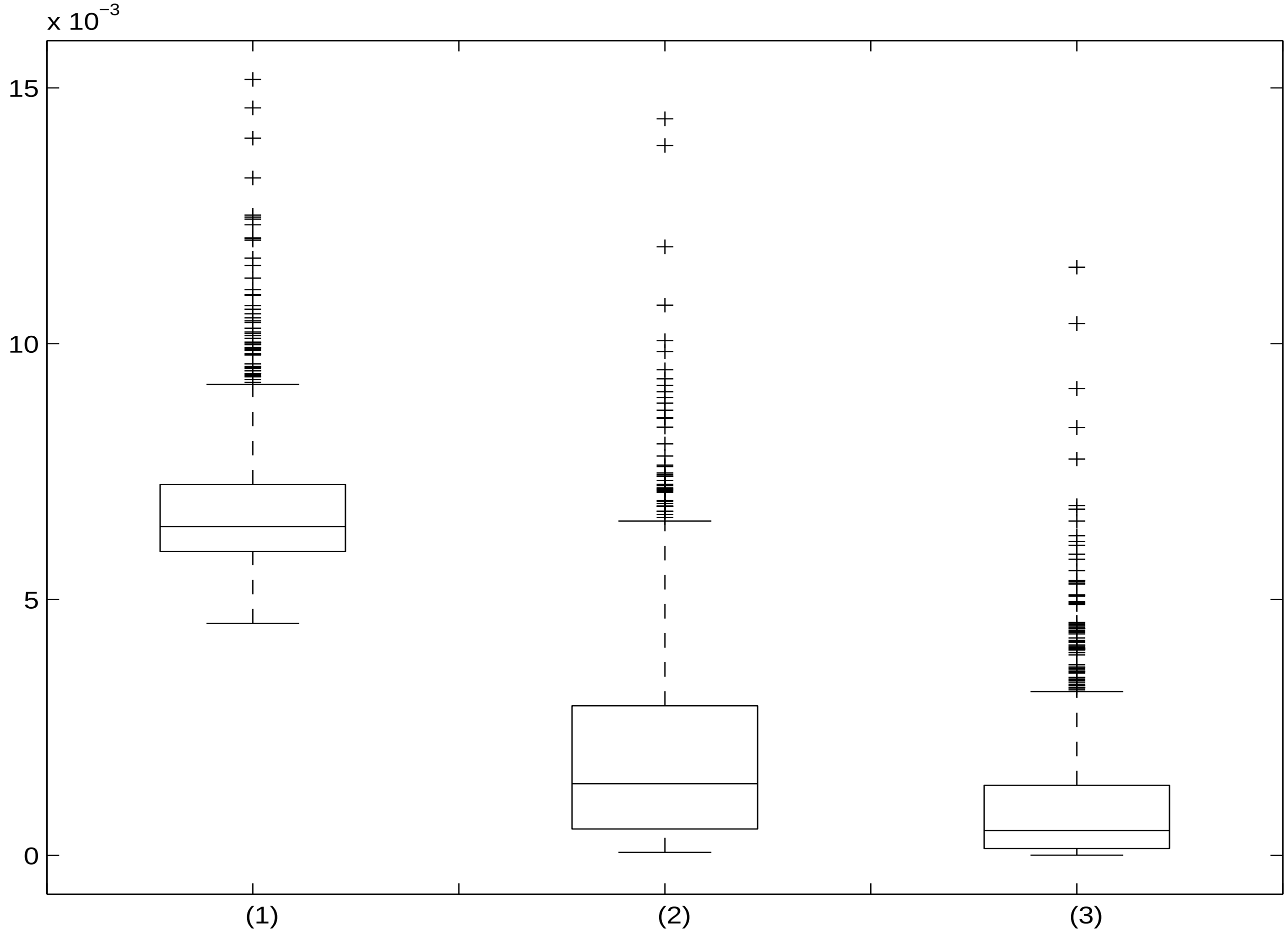


# A comparison

MISE – Mean Integrated Square Error on the interval $[0, h^F_{opt,0,2}]$

| *Method* | *Mean* | *STD* |
|---|---|---|
| Classical | 0.0068 | 0.0014 |
| Reflection | 0.0020 | 0.0020 |
| Proposed | 0.0010 | 0.0014 |

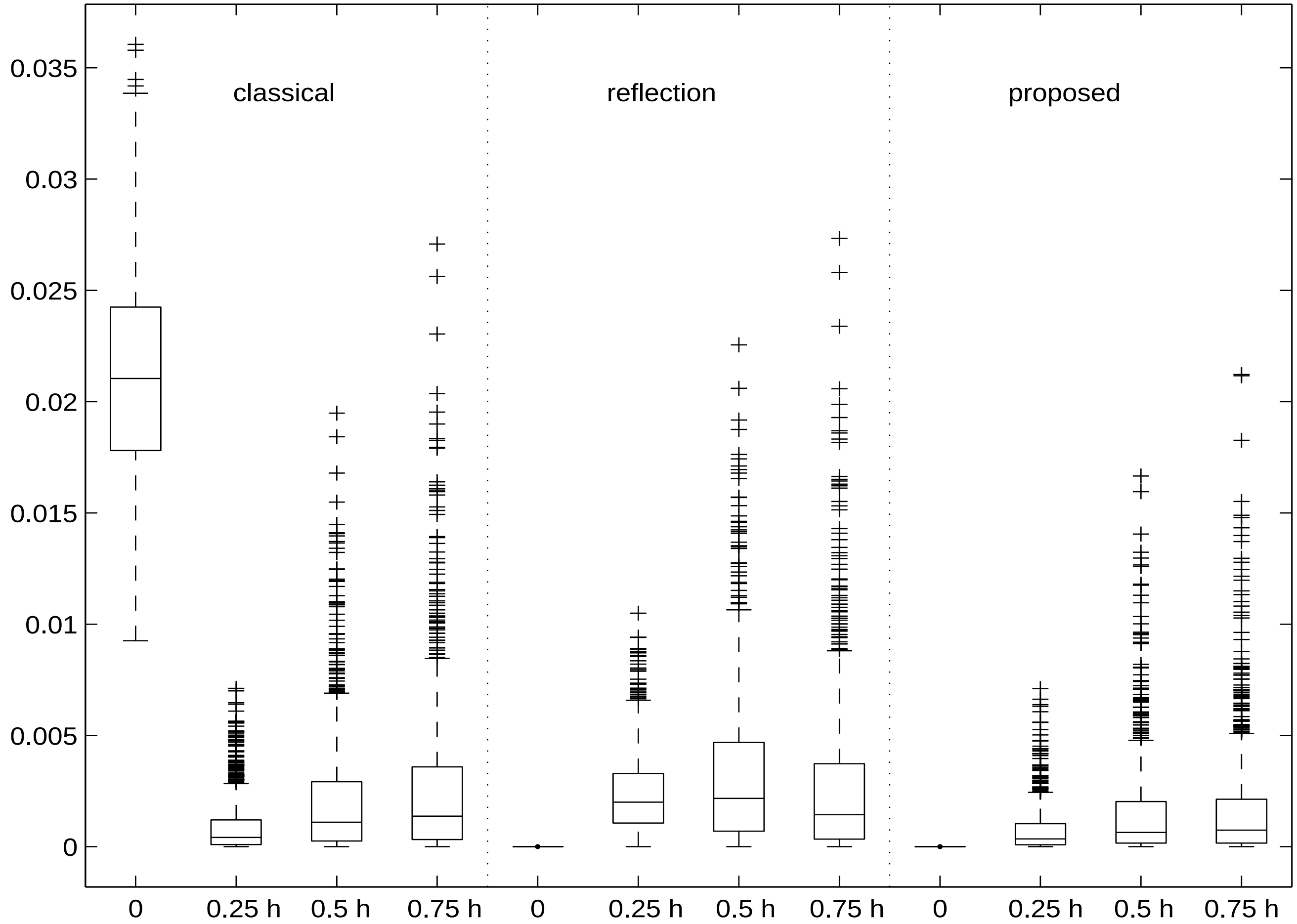
**Table 1.** Means and STD's for MISE




MISE for estimates of CDF for the classical estimator with boundary effects (1), the reflection method (2) and for our proposed method (3).

| | *Classical* | | *Reflection* | | *Proposed* | |
|---|---|---|---|---|---|---|
| $c$ | *Mean* | *STD* | *Mean* | *STD* | *Mean* | *STD* |
| 0.00 | 0.0215 | 0.0048 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.25 | 0.0009 | 0.0013 | 0.0023 | 0.0017 | 0.0008 | 0.0010 |
| 0.50 | 0.0021 | 0.0025 | 0.0032 | 0.0032 | 0.0016 | 0.0021 |
| 0.75 | 0.0026 | 0.0033 | 0.0027 | 0.0034 | 0.0017 | 0.0024 |

**Table 2.** Means and STD's for MSE at $x = ch^F_{opt,0,2}$.




MSE at points $x = ch^F_{opt,0,2}$, $c = 0,\ 0.25,\ 0.5,\ 0.75$ for the classical estimator, the reflection method and for our proposed method.

# Practical usage

## ROC

- The Receiver Operating Characteristic (ROC) describes the performance of a diagnostic test which classifies subjects into either group without condition $\mathcal{G}_0$ or group with condition $\mathcal{G}_1$ by means of a continuous discriminant score $X$, i.e. subject is classified as $\mathcal{G}_1$ if $X \geq d$ and $\mathcal{G}_0$ otherwise for the given cutoff point $d \in \mathbb{R}$.
- Let $\boxed{F_0}$ and $\boxed{F_1}$ be the distribution functions of $X$ in the $\mathcal{G}_0$ and $\mathcal{G}_1$.

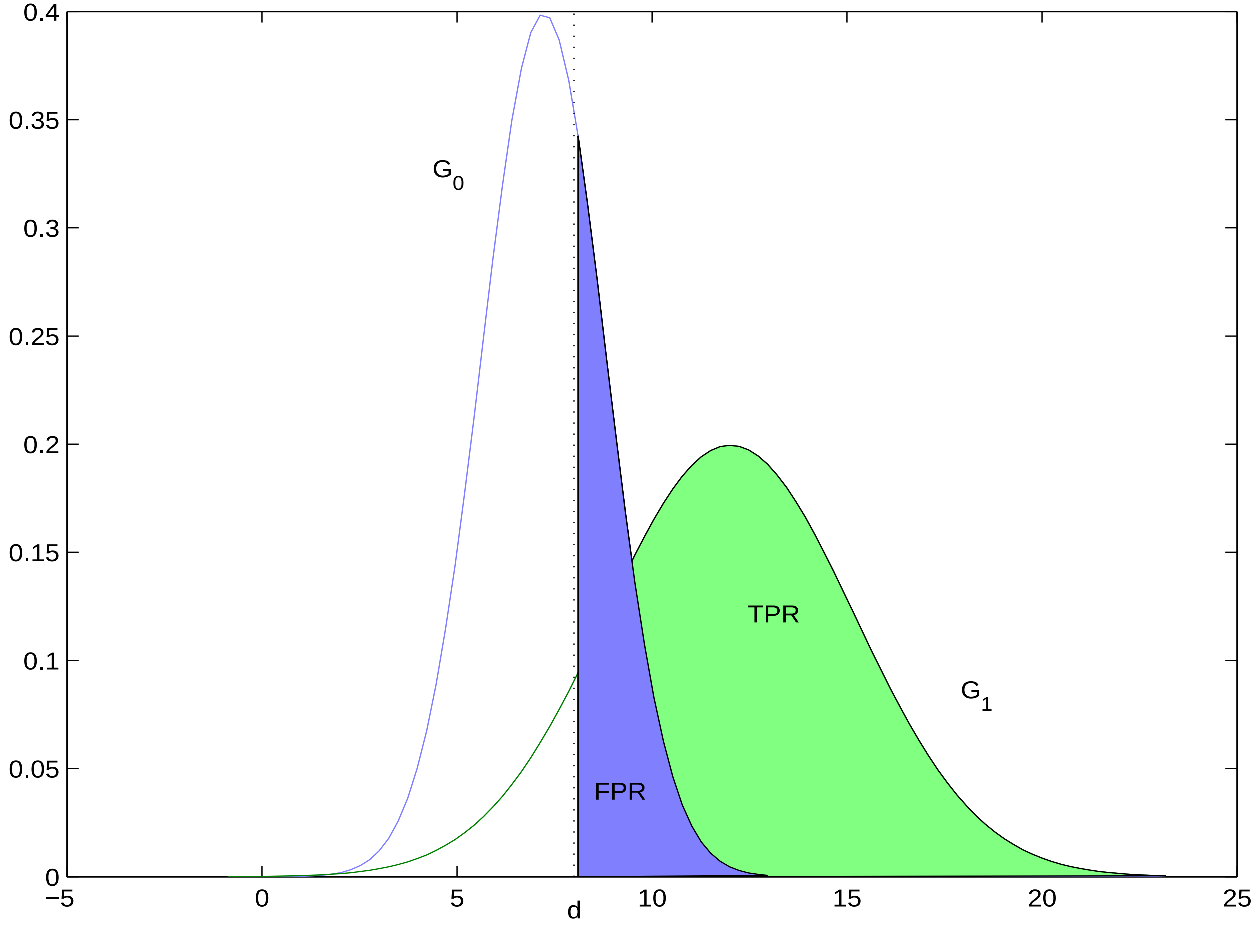
- The ROC is defined as a plot of probability of **false classification of subjects from $\mathcal{G}_1$** versus the probability of **true classification of subjects from $\mathcal{G}_0$** across all possible cutoff point values of $X$.

- ROC curve can be written as

$$R(p) = 1 - F_1(F_0^{-1}(1-p)), \quad 0 < p < 1$$

where $p$ is the false positive rate in $(0,1)$ as the corresponding cut-off point $d$ ranges from $-\infty$ to $+\infty$.

# ROC



# Real data

*Consumer loans data*

The use of some (not specified) scoring function for predicting the solidity of a client.

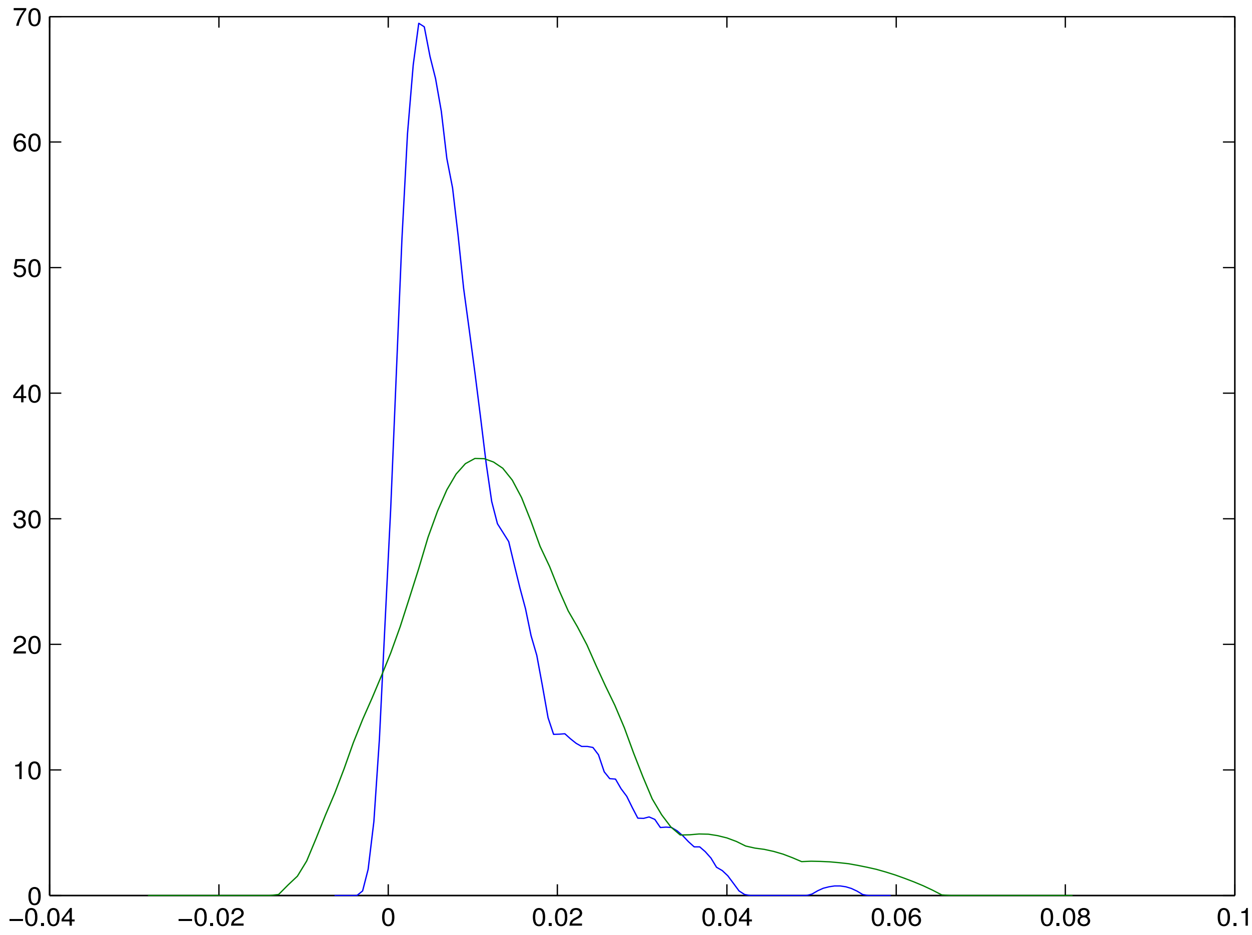
We are interested in determining which clients are able to pay their loans.

A test set: 332 clients – 309 have paid back their loans (group $\mathcal{G}_0$) and 22 had problems with payments or did not pay (group $\mathcal{G}_1$).

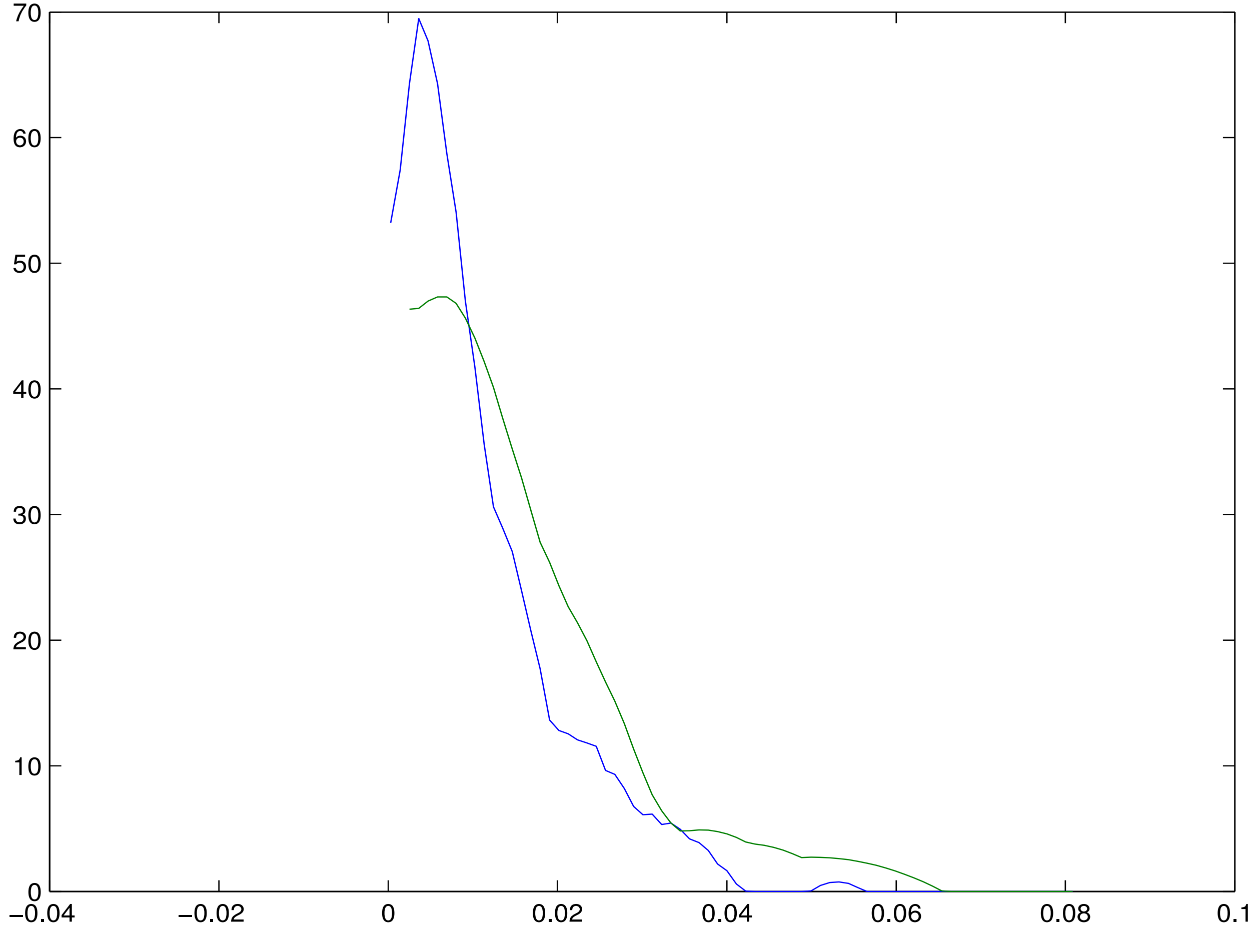
We use the ROC curve to assess the discrimination between clients with and without a good solidity.

We want to know if our scoring function is a good predictor of the solidity.

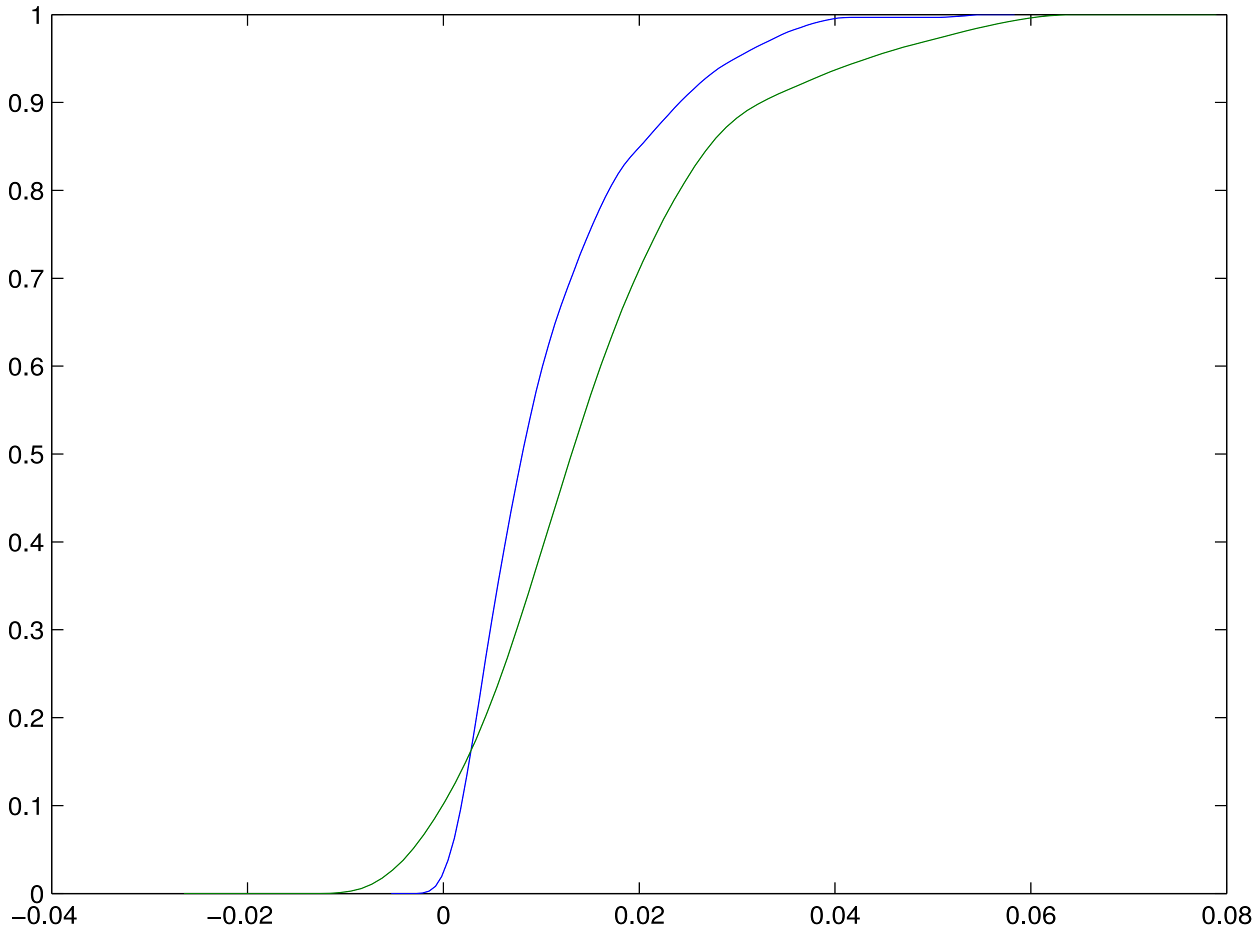
The estimate of $f_0(x)$ ($\hat{h}^{f0}_{opt,0,2} = 0.0032$) and $f_1(x)$ ($\hat{h}^{f1}_{opt,0,2} = 0.0153$) with boundary effects



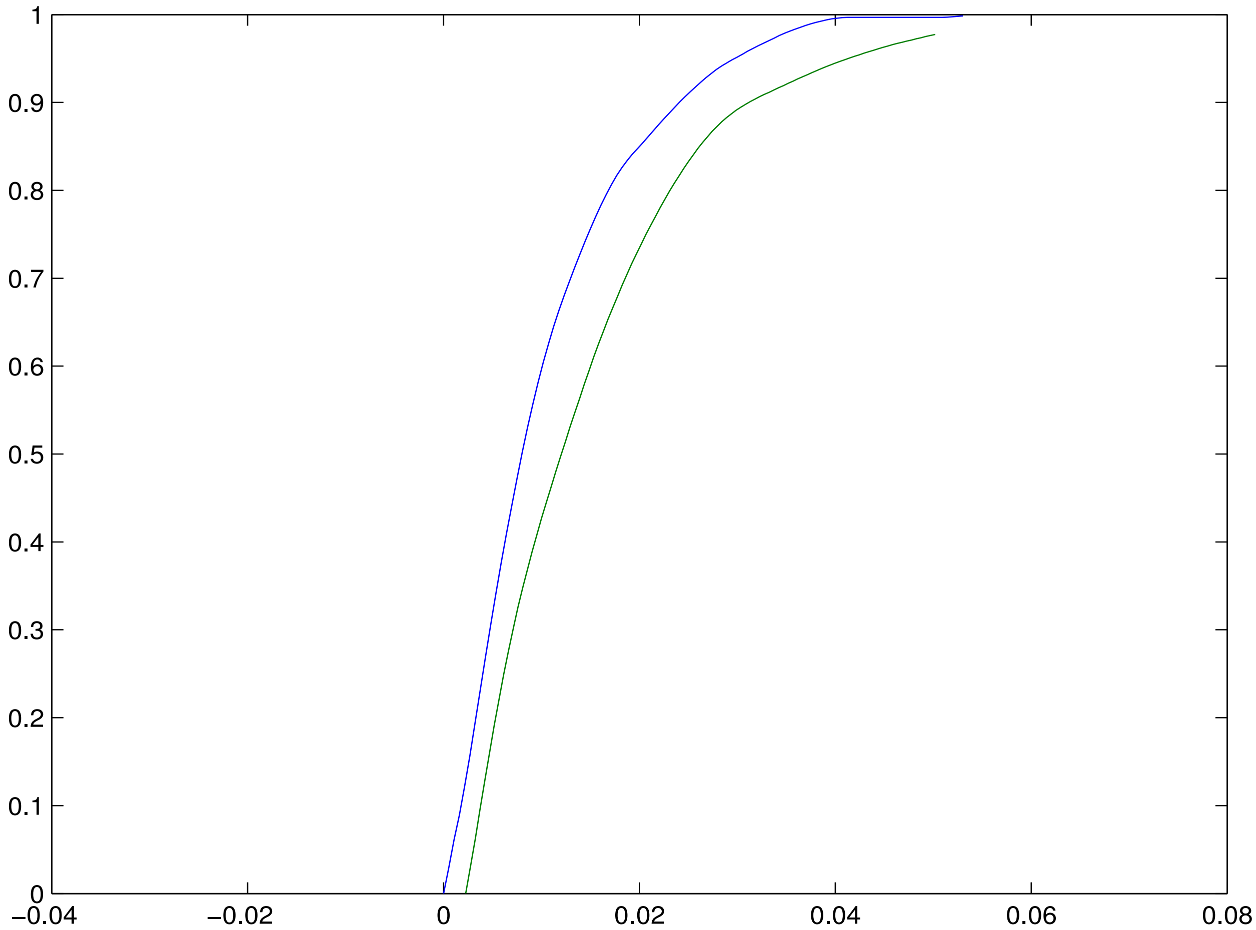
The estimate of $f_0(x)$ ($\hat{h}^{f0}_{opt,0,2} = 0.0032$) and $f_1(x)$ ($\hat{h}^{f1}_{opt,0,2} = 0.0153$) with NO boundary effects



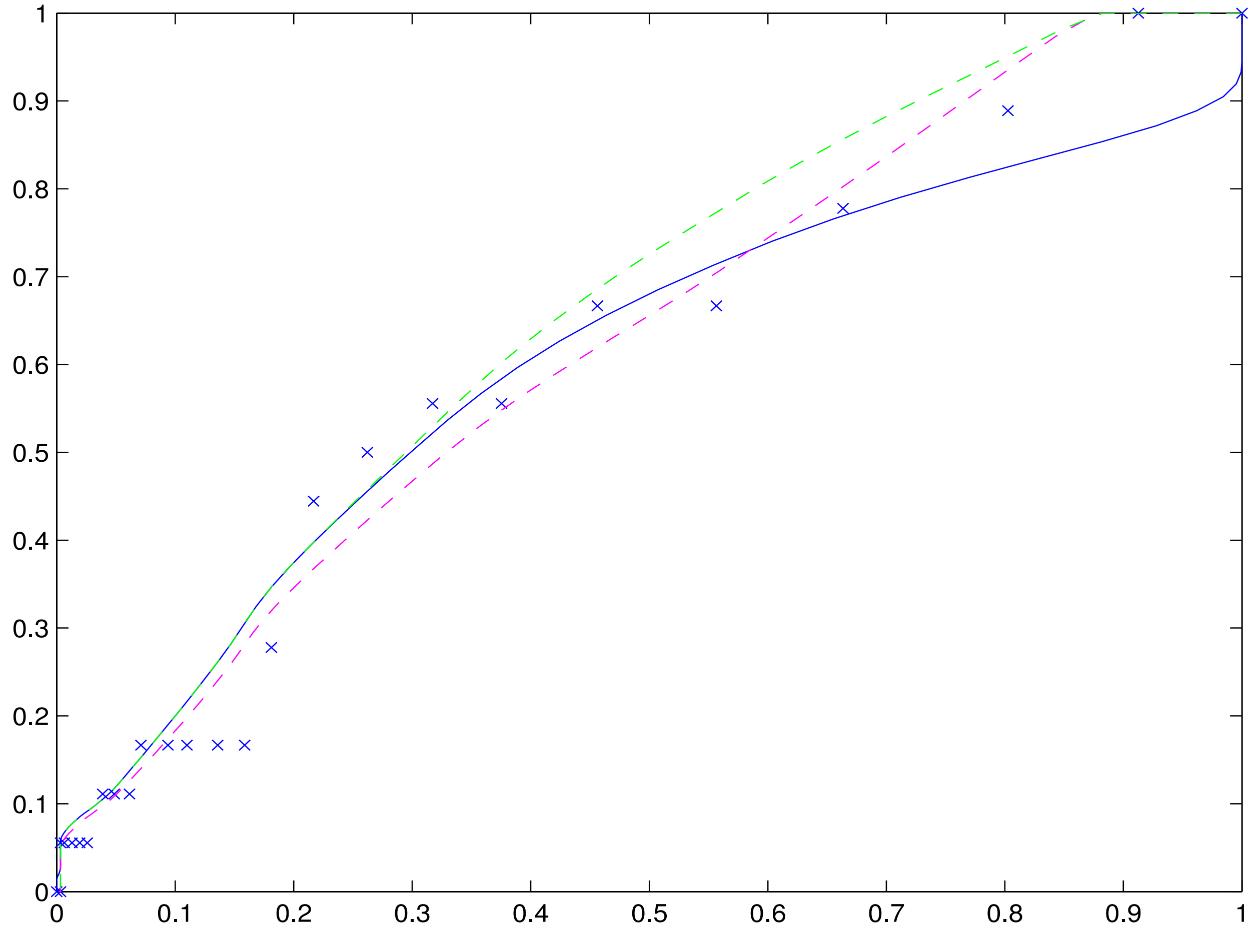
The estimate of $F_0(x)$ ($\hat{h}^{F0}_{opt,0,2} = 0.0068$) and $F_1(x)$ ($\hat{h}^{F1}_{opt,0,2} = 0.0286$) with boundary effects



The estimate of $F_0(x)$ ($\hat{h}^{F0}_{opt,0,2} = 0.0068$) and $F_1(x)$ ($\hat{h}^{F1}_{opt,0,2} = 0.0286$) with NO boundary effects



# The estimate of ROC



# References

[1] Azzalini, A.: *A note on the estimation of a distribution function and quantiles by a kernel method*. Biometrika, 68, No 1, pp. 326–328, 1981.

[2] Bowman, A., Hall, P., Prvan, T.: *Bandwidth selection for the smoothing of distribution functions*. Biometrika, 85, No 4, pp. 799–808, 1998.

[3] Dette, H., Weissbach, R.: *Kolmogorov-Smirnov-type testing for the partial homogeneity of Markov processes – with application to credit risk*. Applied Stochastic Models in Business and Industry, Vol. 23, No. 3, pp. 223–234, 2007.

[4] Horová, I., Zelinka, J.: *Different approaches to ROC curve fitting for a continuous diagnostic test*. CSDA, submitted, 2007.

[5] Karunamuni, R.J., Alberts T.: *On boundary correction in kernel density estimation*. Statistical Methodology 2, pp. 191–212, 2005.

[6] Lloyd, C.J., Zhou Yong: *Kernel estimators of the ROC curve are better than empirical*. Statistics and Prob. Letters 44, pp. 221–228, 1999.

[7] Silverman, B.W.: *Density estimation for statistics and Data Analysis*. Chapman and Hall, New York, 1986.

[8] Terrell, G. R.: *The maximal smoothing principle in density estimation*. Journal of the American Statistical Association. Vol. 85, No. 410, pp. 440-447, 1990.

[9] Wand, I.P. and Jones, M.C.: *Kernel smoothing*. Chapman & Hall, London, 1995.

[10] Zhang, S., Karunamuni, R.J., Jones, M.C.: *An improved estimator of the density function at the boundary*. Journal of the Amer. Stat. Assoc., 448, pp. 1231–1241, 1999.