.

SMOOTHING EMPIRICAL DISTRIBUTION FUNCTION AND DVORETZKY-KIEFER-WOLFOWITZ INEQUALITY

Ryszard Zieliński
Institute of Mathematics Polish Acad. Sc., Warszawa, Poland

Presented to

XXVIII International Seminar
on Stability Problems for Stochastic Models
31 May - 5 June, 2009
Zakopane, Poland

Summary

In Nahariya (*International Seminar on Stability Problems for Stochastic Models, Oct 22–26,2007, Nahariya, Israel*) I showed that standard kernel estimators do not converge to the true distribution UNIFORMLY over the space $\mathcal{F}$ of all continuous and strictly increasing distribution functions. A consequence was that no inequality like Dvoretzky-Kiefer-Wolfowitz (DKW)

$$P_F\{\sup_{x\in\mathbf{R}} |F_n(x) - F(x)| \geq \epsilon\} \leq 2e^{-2n\epsilon^2}$$

can be constructed, and as a result it was IMPOSSIBLE TO ANSWER THE QUESTION HOW MANY OBSERVATIONS ARE NEEDED TO GUARANTEE A PRESCRIBED LEVEL OF ACCURACY OF THE ESTIMATOR of an unknown distribution function $F \in \mathcal{F}$. A remedy was to modify the estimator adapting the bandwidth to the sample at hand. It appears that polynomial and spline estimators share the disadvantage. It is however possible to construct some subspaces of $\mathcal{F}$ on which the estimators converge uniformly and in consequence DKW holds.

Dvoretzky-Kiefer-Wolfowitz inequality (Massart 1990)

$$P_F\{\sup_{x\in\mathbf{R}} |F_n(x) - F(x)| \geq \epsilon\} \leq 2e^{-2n\epsilon^2}$$

By the inequality, given $\epsilon > 0$ and $\eta > 0$ one can easily find the smallest $N = N(\epsilon, \eta)$ such that if $n \geq N(\epsilon, \eta)$ then

$$(\forall F \in \mathcal{F}) \quad P_F\{\sup_{x\in\mathbf{R}} |F_n(x) - F(x)| \geq \epsilon\} \leq \eta$$

E.g. $N(0.1, 0.1) = 150$ and $N(0.01, 0.01) = 26\,492$

Glivenko-Cantelli theorem

$$(\forall \epsilon)(\forall \eta)(\exists N)(\forall n \geq N)(\forall F \in \mathcal{F}) \quad P_F\{\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| \geq \epsilon\} \leq \eta$$

where

$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1_{(-\infty, x]}(X_j)$$

Here $N = N(\epsilon, \eta)$ does not depend on $F \in \mathcal{F}$!

Standard kernel density estimator

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h_n} k\Big(\frac{x - X_j}{h_n}\Big)$$

Kernel distribution estimator

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} K\Big(\frac{x - X_j}{h_n}\Big), \qquad K(x) = \int_{-\infty}^{x} k(t)dt$$

GLIVENKO-CANTELLI THEOREM DOES NOT HOLD:

$$(\exists\epsilon)(\exists\eta)(\forall N)(\exists n \geq N)(\exists F \in \mathcal{F}) \quad P_F\{\sup_{x\in\mathbf{R}} |\widehat{F}_n(x) – F(x)| \geq \epsilon\} \geq \eta$$

The statement is true under assumptions:

Concerning the kernel K:

1) $0 < K(0) < 1$ and
2) $K^{-1}(t) < 0$ for some $t \in (0, F(0))$

Concerning the sequence $(h_n, n = 1, 2, \ldots)$ the only assumption is that $h_n > 0, n = 1, 2, \ldots$

It follows that standard kernel estimators are useless for statistical applications!

A way to improve the situation is modification of the kernel estimator

In Nahariya a kernel estimator with RANDOM BANDWIDTH was presented:

$X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ – order statistics

$$H_n = \min\{X_{j:n} - X_{j-1:n},\ j = 2, 3, \ldots, n\}$$

$$\widetilde{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} K\Big(\frac{x - X_j}{H_n}\Big)$$

where for K we assume:

$$K(t) = \begin{cases} 0, & \text{for } t \leq -1/2 \\ 1/2, & \text{for } t = 0 \\ 1, & \text{for } t \geq 1/2 \end{cases}$$

$K(t)$ continuous and increasing in $(-1/2, 1/2)$

Dvoretzky-Kiefer-Wolfowitz inequality takes on the form:

$$P_F\{\sup_{x\in\mathbf{R}} |\widetilde{F}_n(x) - F(x)| \geq \epsilon\} \leq 2e^{-2n(\epsilon-1/2n)^2}, \quad n > \frac{1}{2\epsilon}$$

which enables us to calculate $N = N(\epsilon, \eta)$ that guarantees the prescribed accuracy of the kernel estimator $\widetilde{F}_n(x)$.

Another way is to restrict the statistical model $\mathcal{F}$ to a smaller class.

That is what I want to present now.

The results which follow come from a joint paper by Zbigniew Ciesielski and myself:

Polynomial and Spline Estimators of the Distribution Function with Prescribed Accuracy. Applicationes Mathematicae 36, 1(2009), pp. 1-12

POLYNOMIAL ESTIMATORS on $[0,1]$

Basic polynomials on $[0,1]$:

$$N_{i,m}(x) = \binom{m}{i} x^i (1-x)^{m-i}$$

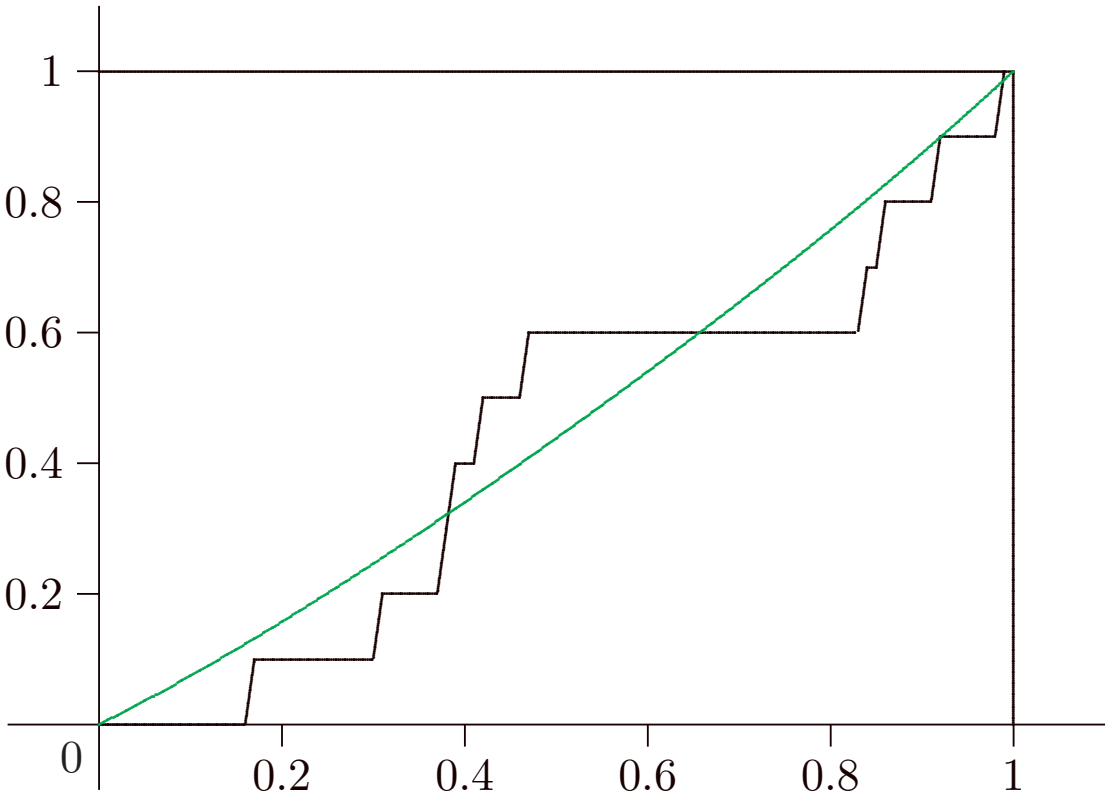
$$0 \le x \le 1, \quad i = 0, 1, \ldots, m; \quad m \ge 1$$

Operator T_m (Ciesielski 1988):

$$T_m F(x) = \sum_{i=0}^{m} \int_0^1 (m+1) N_{i,m}(y) dF(y) \int_0^x N_{i,m}(z) dz$$

If F is a distribution function on $[0,1]$, continuous or not, then $T_m F$ is a polynomial distribution function on $[0,1]$ and

$$F_{m,n} = T_m F_n \text{ is an estimator of } F$$



Simple formulas for $F_{m,n}$:

$$F_{m,n}(x) = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{m} N_{i,m}(X_j)\int_0^x (m+1)N_{i,m}(z)dz$$

$$N_{i,m}(x) = \binom{m}{i}x^i(1-x)^{m-i} = b(i,m,x)$$

$$I_x(p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\int_0^x t^{p-1}(1-t)^{q-1}dt.$$

$$F_{m,n}(x) = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{m} b(i,m,X_j)I_x(i+1,m-i+1).$$

THEOREM

$$(\exists \epsilon > 0)(\exists \eta > 0)(\forall m)(\forall n)(\exists F \in \mathcal{F})$$

$$P_F\{\sup_{x \in \mathbf{R}} |F_{m,n}(x) - F(x)| > \epsilon\} > \eta$$

THEOREM

$$(\forall \epsilon > 0)(\forall \eta > 0)(\forall M > 0)(\exists m)(\exists n)(\forall F \in W_M)$$

$$P_F\{\sup_{x \in \mathbf{R}} |F_{m,n}(x) - F(x)| > \epsilon\} < \eta$$

where W_M is a subclass of $\mathcal{F}$ such that $F \in W_M$ if and only if the density $f = F'$ is absolutely continuous and

$$\int_0^1 |f'(x)|^2 dx \leq M$$

In practical applications:

If a statistician knows the constant M such that

$$\int_0^1 |f'(x)|^2 dx \leq M$$

then to have

$$P_F\{\sup_{x\in\mathbf{R}} |F_{m,n}(x) - F(x)| > \epsilon\} < \eta$$

it is enough to choose the degree of the approximating polynomial m and the sample size n such that

$$\frac{2M}{m^{1/4}} < \epsilon \quad and \quad 2\exp\Big(-2\frac{nM^2}{m^{1/2}}\Big) < \eta$$

$B^{(r)}(x)$ is a symmetric cardinal B-spline of order r if:

$B^{(r)}(x) \geq 0, \ x \in R$,

$supp \ B^{(r)} = [-r/2, r/2]$,

$B^{(r)}$ is a polynomial of order $r-1$ on each interval $[j - r/2, j + 1 - r/2], \quad j = 0, 1, \ldots, r-1,$

$B^{(r)} \in C^{(r-2)}(R)$ (step function if $r = 1$)

Probabilistic interpretation:

$B^{(r)}$ is the density function of the distribution of the sum of r i.i.d. random variables distributed as $U(-1/2, 1/2)$

Nice formulas:

$$B^{(r)}(x)=\begin{cases} 0, & \text{if } x<\dfrac{1}{2}, \\ \dfrac{1}{(r-1)!}\displaystyle\sum_{i=0}^{[x+r/2]}(-1)^i\binom{r}{i}(x+\frac{r}{2}-i)^{r-1}, & \text{if } -\dfrac{1}{2}\le x\le\dfrac{1}{2} \\ 0, & \text{if } x>\dfrac{1}{2} \end{cases}$$

$$\mathcal{B}^{(r)}(x) = \int_{-\infty}^{x} B^{(r)}(t)dt$$

$$\mathcal{B}^{(r)}(x) = \begin{cases} 0, & \text{if } x < -\dfrac{1}{2}, \\ \dfrac{1}{r!}\displaystyle\sum_{i=0}^{[x+r/2]} (-1)^i \binom{r}{i} (x+\frac{r}{2}-i)^r, & \text{if } -\dfrac{1}{2} \le x \le \dfrac{1}{2} \\ 1, & \text{if } x > \dfrac{1}{2} \end{cases}$$

Given $r \geq 1$, $h > 0$, $i \in \mathcal{Z}$ define

$$B_{h,i}^{(r)}(x) = B^{(r)}\left(\frac{x}{h} - i\right)$$

Given $r \geq 1$, $1 \leq k \leq r$, $r - k = 2\nu$, $\nu -$ integer, $i \in \mathcal{Z}$, and $h > 0$ define the operator (Ciesielski 1988, 1991)

$$T_h^{(k,r)} F(x) = \frac{1}{h} \sum_{i \in \mathcal{Z}} \int_R B_{h,i+\nu}^{(k)}(y) dF(y) \int_{-\infty}^{x} B_{h,i}^{(r)}(y) dy$$

Operator $T_h^{(k,r)}$ transforms distribution functions (continuous or not) in distributions functions which are splines of order r.

$T_h^{(k,r)} F_n$ is a spline estimator of F:

$$T_h^{(k,r)} F_n(x) = \sum_{i \in \mathcal{Z}} \left[\frac{1}{n} \sum_{j=1}^{n} B^{(k)} \left(\frac{X_j}{h} - \left(i + \frac{r-k}{2} \right) \right) \right] \mathcal{B}^{(r)} \left(\frac{x}{h} - i \right)$$

CLASSES OF DISTRIBUTION FUNCTIONS FOR WHICH DKW HOLDS

Define

$$\omega_1(F,\delta) = \sup_{|t|<\delta} \sup_x |F(x+t) - F(x)|$$

$$\omega_2(F,\delta) = \sup_{|t|<\delta} \sup_x |F(x+2t) - 2F(x+t) + F(x)|$$

and for a modulus of smoothness ω (bounded, continuous, vanishing at 0, non-decreasing and subadditive function) define two Hölder classes of distribution functions:

$$H_{\omega,1}^{(k,r)} = \{F \in \mathcal{F} : \omega_1\big(F, \frac{r+k}{2}h\big) \leq \omega(h)\}$$

$$H_{\omega,2}^{(k,r)} = \{F \in \mathcal{F} : \big(2(4+(r+k)^2\big)\omega_2(F,h) \leq \omega(h)\}$$

In *Zbigniew Ciesielski and Ryszard Zieliński: Polynomial and Spline Estimators of the Distribution Function with Prescribed Accuracy. Applicationes Mathematicae 36, 1(2009), pp. 1-12* one can find the proof of the following theorem:

THEOREM. Let $i = 1, 2$, $1 \leq k \leq r$ and let $r - k$ be even. Then for each $\epsilon > 0$ and for each $\eta > 0$ there are $h > 0$ and $n \geq 1$ such that

$$P_F\{\|F_{h,n} - F\|_\infty > \epsilon\} < \eta \quad for \quad all \quad F \in H^{(k,r)}_{\omega,i}.$$

The parameters n and h can be realized by choosing them so that

$$\omega(h) < \frac{\epsilon}{2} \quad and \quad 2\exp\Big(-\frac{n\epsilon^2}{2}\Big) < \eta.$$