

UNIFORM CONVERGENCE OF KERNEL ESTIMATORS
WITH RANDOM BANDWIDTH

Ryszard Zieliński

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

R.Zielinski@impan.gov.pl

Abstract

Standard kernel estimators do not converge to the true distribution uniformly. A consequence is that no inequality like Dvoretzky-Kiefer-Wolfowitz one can be constructed, and as a result it is impossible to answer the question how many observations are needed to guarantee a prescribed level of accuracy of the estimator. A remedy is to adapt the bandwidth to the sample at hand.

Key Words: kernel estimators, asymptotics, Glivenko-Cantelli theorem, Dvoretzky-Kiefer-Wolfowitz inequality, bandwidth, adaptive estimators, uniform limit laws

AMS classification: 62G20, 62G30, 62G07

1. GLIVENKO-CANTELLI THEOREM AND DVORETZKY-KIEFER-WOLFOWITZ INEQUALITY. Let X_1, X_2, \dots, X_n be a sample from an (unknown) distribution $F \in \mathcal{F}$ where \mathcal{F} is the class of all continuous distribution functions.

The version of the Glivenko-Cantelli theorem in the form to be exploited below states that

$$(GCT) \quad (\forall \varepsilon)(\forall \eta)(\exists N)(\forall n \geq N)(\forall F \in \mathcal{F}) \quad P\left\{ \sup_{x \in \mathbf{R}^1} |F_n(x) - F(x)| \geq \varepsilon \right\} \leq \eta$$

where

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

The theorem is effective in the sense that for every $\varepsilon > 0$ and for every $\eta > 0$ one can effectively calculate $N = N(\varepsilon, \eta)$. That can be done by the following version of Dvoretzky-Kiefer-Wolfowitz inequality (Massart 1990)

$$(*) \quad P\left\{ \sup_{x \in \mathbf{R}^1} |F_n(x) - F(x)| \geq \varepsilon \right\} \leq 2e^{-2n\varepsilon^2}.$$

Due to the above, *GCT* together with $(*)$ give us a genuinely statistical tool; if all that a statistician knows is that an unknown distribution F belongs to \mathcal{F} , he is able to make a precise inference about F (testing hypotheses or constructing confidence intervals).

2. KERNEL ESTIMATORS. The standard kernel density estimator is of the form (e.g. Wegman 2006)

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

with appropriate $h_n, n = 1, 2, \dots$. We shall consider kernel distribution estimator in its classical form

$$\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)$$

where $K(x) = \int_{-\infty}^x k(t)dt$, and we shall show that (GCT) does not hold if F_n is replaced by \hat{F}_n , i.e. that the following is true

$$(\exists \varepsilon)(\exists \eta)(\forall N)(\exists n \geq N)(\exists F \in \mathcal{F}) \quad P\left\{\sup_{x \in \mathbf{R}^1} |\hat{F}_n(x) - F(x)| \geq \varepsilon\right\} \geq \eta.$$

Obviously it is enough to demonstrate that

$$(\dagger) \quad (\exists \varepsilon)(\exists \eta)(\forall n)(\exists F \in \mathcal{F}) \quad P\{\hat{F}_n(0) > F(0) + \varepsilon\} \geq \eta.$$

Concerning the kernel K , only the following assumptions are relevant: 1) $0 < K(0) < 1$ and 2) $K^{-1}(t) < 0$ for some $t \in (0, K(0))$. Concerning the sequence $(h_n, n = 1, 2, \dots)$ we assume that $h_n > 0, n = 1, 2, \dots$

Take $\varepsilon \in (0, t)$ and $\eta \in (t - \varepsilon, 1)$. Given ε, η , and n , take F such that $F(0) = t - \varepsilon$ and $F(-h_n K^{-1}(t)) = P\{X_j < -h_n K^{-1}(t)\} > \eta^{1/n}$. Then

$$P\left\{K\left(-\frac{X_j}{h_n}\right) > t\right\} > \eta^{1/n}$$

and due to the fact that

$$\bigcap_{j=1}^n \left\{K\left(-\frac{X_j}{h_n}\right) > F(0) + \varepsilon\right\} \subset \left\{\frac{1}{n} \sum_{j=1}^n K\left(-\frac{X_j}{h_n}\right) > F(0) + \varepsilon\right\}$$

we have

$$P\left\{\frac{1}{n} \sum_{j=1}^n K\left(-\frac{X_j}{h_n}\right) > t\right\} = P\left\{\frac{1}{n} \sum_{j=1}^n K\left(-\frac{X_j}{h_n}\right) > F(0) + \varepsilon\right\} > \eta;$$

hence (\dagger) .

It follows that for classical kernel estimators no inequality like (*) can be obtained which makes the estimators of a doubtful usefulness for statistical applications.

3. RANDOM BANDWIDTH. A remedy is as follows. Let $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ be order statistics from the sample X_1, X_2, \dots, X_n . Define

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

Define the kernel estimator

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

where for K we assume:

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

$K(0) = \frac{1}{2}$, $K(t)$ continuous and nondecreasing in $(-\frac{1}{2}, \frac{1}{2})$.

Now, for $k = 1, 2, \dots, n$ we have $|\tilde{F}_n(X_{k:n}) - F_n(X_{k:n})| \leq \frac{1}{2n}$. Kernel estimator $\tilde{F}_n(x)$ is continuous and increasing, empirical distribution function $F_n(x)$ is a step function, and in consequence $|\tilde{F}_n(x) - F_n(x)| \leq \frac{1}{2n}$ for all $x \in (-\infty, \infty)$. By the triangle inequality

$$|\tilde{F}_n(x) - F(x)| \leq |F_n(x) - F(x)| + \frac{1}{2n}$$

we obtain

$$P\left\{\sup_{x \in \mathbf{R}^1} |\tilde{F}_n(x) - F(x)| \geq \varepsilon\right\} \leq P\left\{\sup_{x \in \mathbf{R}^1} |F_n(x) - F(x)| + \frac{1}{2n} \geq \varepsilon\right\}$$

and hence, by (*) we have

$$(**) \quad P\left\{\sup_{x \in \mathbf{R}^1} |\tilde{F}_n(x) - F(x)| \geq \varepsilon\right\} \leq 2e^{-2n(\varepsilon - 1/2n)^2}, \quad n > \frac{1}{2\varepsilon}$$

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

A COMMENT. Observe that the smallest $N = N(\varepsilon, \eta)$ that guarantees the prescribed accuracy is somewhat greater for kernel estimator \tilde{F}_n than that for crude empirical step function F_n . For example, $N(0.1, 0.1) = 150$ for F_n and $= 160$ for \tilde{F}_n ; $N(0.01, 0.01) = 26,492$ for F_n and $= 26,592$ for \tilde{F}_n . Another disadvantage of kernel smoothing has been discovered by Hjort and Walker (2001): "*kernel density estimator with optimal bandwidth lies outside any confidence interval, around the empirical distribution function, with probability tending to 1 as the sample size increases*". Perhaps a reason is that smoothing adds to observations something which is rather arbitrarily chosen and which may spoil the inference.

A GENERALIZATION. Inequality (**) holds for every smoothed nondecreasing distribution function $\tilde{F}_n(x)$ that satisfies $|\tilde{F}_n(X_{k:n}) - F_n(X_{k:n})| \leq \frac{1}{2n}$, $k = 1, 2, \dots, n$.

References

- Hjort, N.L., and Walker, S.G. (2001). A note on kernel density estimators with optimal bandwidths. *Statistics & Probability Letters* 54, 153-159
- Massart, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Annals of Probability*, 18: 1269–1283
- Wegman, E.J. (2006). Kernel estimators. In *Encyclopedia of statistical sciences*. Second Edition, Vol. 6, Wiley–Interscience