The Annals of Statistics
1987, Vol. 15, No. 3, 1317–1320

AN APPLICATION OF THE EFRON–STEIN INEQUALITY IN DENSITY ESTIMATION[1]

BY LUC DEVROYE

McGill University

The Efron–Stein inequality is applied to prove that the kernel density estimate f_n, with an arbitrary nonnegative kernel and an arbitrary smoothing factor, satisfies the inequality $\text{var}(\int |f_n - f|) \leq 4/n$ for all densities f. Similar inequalities are obtained for other estimates.

The main result. Let $X_1, \ldots, X_n, X_{n+1}$ be iid random vectors and let $S(x_1, \ldots, x_n)$ be a symmetric function of its arguments. Define

$$S_i = S(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{n+1}), \qquad i = 1, \ldots, n+1,$$

and $\bar{S} = (n+1)^{-1} \sum_{i=1}^{n+1} S_i$. The Efron–Stein inequality [Efron and Stein (1981)] states that

$$\text{var}(S(X_1, \ldots, X_n)) \leq E\left(\sum_{i=1}^{n+1} (S_i - \bar{S})^2 \right).$$

The inequality originates from studies of the jackknife estimate of variance. Several interesting proofs [see, e.g., Vitale (1984)] and applications [see, e.g., Steele (1981, 1982)] highlight the depth and usefulness of the Efron–Stein inequality. The purpose of this short note is to point out an application in density estimation.

We consider in particular a density estimate f_n of a density f on R^d with the following properties.

A. f_n is a symmetric function of the data $X_1, \ldots, X_n$.
B. f_n is absolutely integrable.
C. $\int |f_{ni} - f_{nj}| \leq \delta$ for all $1 \leq i, j \leq n+1$, where f_{ni} is the density estimate based upon $X_1, \ldots, X_{n+1}$, with X_i deleted.

Most well-known density estimates satisfy properties A–C. It is well known that symmetric functions of the data make the best density estimates [Wertz (1976); Devroye and Gyorfi (1985), page 283], so that A is not restrictive. The kernel and histogram estimates satisfy A and B. The constant δ in condition C can always be taken equal to 2 when f_n is a density itself. Unfortunately, the

Received September 1986.
[1]Supported by NSERC Grant A3456 and FCAR Grant EQ-1679.
AMS 1980 *subject classifications.* 60E15, 62G05.
Key words and phrases. Efron–Stein inequality, density estimation, kernel estimate, distribution-free confidence interval.

results presented below are only useful when δ is small and decreasing in n. Consider for example the *kernel estimate*

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i),$$

where K is a given absolutely integrable function (the kernel) $\int K = 1$, $K_h(x) = (1/h^d)K(x/h)$, and the smoothing factor h is a positive number [Parzen (1962), Rosenblatt (1956)]. It is clear that if h is not a function of the data,

$$\int |f_{ni} - f_{nj}| \le \frac{1}{n}\left(\int |K_h(x - X_i)|\,dx + \int |K_h(x - X_j)|\,dx\right) = \frac{2\int |K|}{n}.$$

Similarly, for a histogram estimate based upon a partition of the space that is independent of the data,

$$\int |f_{ni} - f_{nj}| \le \frac{2}{n}.$$

The main result is

THEOREM 1. *Let f_n be a density estimate satisfying* A–C *for some constant $\delta > 0$. Then*

$$\operatorname{var}\left(\int |f_n - f|\right) \le n\delta^2.$$

It is interesting to observe that this inequality is valid for *all* densities f, and that it has virtually no relationship with the consistency of the density estimate or the closeness of f_n to f. It merely states that for estimates on which deletions of one data point have little impact (i.e., estimates with small δ), $\int |f_n - f|$ cannot oscillate wildly.

For symmetric density estimates, we have

$$\operatorname{var}\left(\int |f_n - f|\right) \le nE\left(\int |f_{n1} - f_{n2}|\right).$$

This inequality is obtained very easily by generalizing the proof of Theorem 1.

THEOREM 2. *For the kernel estimate with kernel K, we have*

$$\operatorname{var}\left(\int |f_n - f|\right) \le \frac{4\int^2 |K|}{n}.$$

For the histogram estimate and for the kernel estimate with nonnegative kernel, we have

$$\operatorname{var}\left(\int |f_n - f|\right) \le \frac{4}{n}.$$

Relative stability of the kernel estimate. Chebyshev's inequality implies that an estimate f_n is *relatively stable*, i.e.,

$$\frac{\int |f_n - f|}{E(\int |f_n - f|)} \to 1 \quad \text{in probability}$$

when $\sqrt{\text{var}(\int|f_n - f|)} = o(E(\int|f_n - f|)$. Let us recall that for the kernel estimate with data-independent h, $E(\int|f_n - f|) \geq 1/\sqrt{528n}$ [Devroye (1986b)]. Thus, a kernel estimate is relatively stable at a density f when

$$\lim_{n\to\infty} \sqrt{n}E\left(\int|f_n - f|\right) = \infty.$$

It is well known that this is the case for most combinations of K and h. In particular, relative stability follows for any f and any sequence h when $K \geq 0$ [in view of a universal lower bound of the order of $n^{-2/5}$ obtained by Devroye and Penrod (1984)]. Relative stability also follows for all f and K when $h \to 0$ [in view of a result found on page 136 of Devroye and Gyorfi (1985)]. But there are combinations of K, f and h for which $E(\int|f_n - f|) = O(1/\sqrt{n})$, so that we cannot make the statement as general as we would like it to be. An example of this includes a density with bounded spectrum, combined with a kernel whose characteristic function is one in an open neighborhood of the origin, and a small enough constant smoothing factor h.

Strong relative stability was studied by Devroye (1986a). He also obtained distribution-free exponential bounds for the deviation $J_n - E(J_n)$, where $J_n = \int|f_n - f|$ and f_n is the kernel estimate. For example,

$$P(|J_n - E(J_n)| \geq \varepsilon) \leq \exp(-c\sqrt{n\varepsilon^2})$$

for some constant c depending upon K only, and for all ε smaller than a given constant. This result is stronger than Theorem 2 in the sense that a distribution-free $O(1/n)$ bound for $\text{var}(J_n)$ can be obtained from it. Unfortunately, the constant is worse than in Theorem 2; the proof is much more involved; and some restrictions have to be placed on K (K needs to be bounded, and of compact support). We can obtain a quadratic bound

$$P(|J_n - E(J_n)| \geq \varepsilon) \leq \frac{4\int^2|K|}{n\varepsilon^2}$$

by Chebyshev's inequality. In contrast, the exponential inequality is better for large values of $n\varepsilon^2$. It is more useful for some kinds of confidence intervals and for studies of strong convergence.

Confidence intervals for estimating the L1 error. In a simulation study, we may wish to estimate $E(J_n) = E(\int|f_n - f|)$, where both f and f_n are known. This is a common problem in the testing stage of density estimators. If we estimate the quantity by $\int|f_n - f|$, i.e., without averaging over any runs, we nevertheless see that

$$P\left(\left|\frac{J_n}{E(J_n)} - 1\right| > \varepsilon\right) \leq \frac{\text{var}(J_n)}{\varepsilon^2 E^2(J_n)},$$

which, in the case of a kernel estimate with $K \geq 0$, or a histogram estimate, is further bounded by

$$\frac{4}{n\varepsilon^2 E^2(J_n)}.$$

Since $E(J_n) \geq (0.86 + o(1))n^{-2/5}$, we see that the bound is $O(n^{-1/5})$ for all f and for constant ε. The fact that all the constants are explicitly known makes bounds of this type very useful in practice.

Needless to say, the performance of the estimate can be improved by averaging over more than one run. Also, $\int|f_n - f|$ is sometimes difficult to compute accurately, especially when f has infinite tails or infinite peaks. In those cases, one could use Monte Carlo methods based upon an independent sample drawn from f.

PROOF OF THEOREM 1. By the Efron–Stein inequality

$$\operatorname{var}\left(\int|f_n - f|\right) \leq (n+1)E\left(\left(\frac{1}{n+1}\sum_{j=1}^{n+1}\left(\int|f_{n1} - f| - \int|f_{nj} - f|\right)\right)^2\right)$$

(by the symmetry in the problem)

$$\leq (n+1)E\left(\left(\frac{1}{n+1}\sum_{j=1}^{n+1}\int|f_{n1} - f_{nj}|\right)^2\right)$$

(by the triangle inequality)

$$\leq (n+1)\left(\frac{n\delta}{n+1}\right)^2 \qquad \text{(by assumption)}$$

$$\leq n\delta^2. \qquad \square$$

REFERENCES

DEVROYE, L. (1986a). The kernel estimate is relatively stable. Technical Report, School of Computer Science, McGill Univ.

DEVROYE, L. (1986b). A universal lower bound for the kernel estimate. Technical Report, School of Computer Science, McGill Univ.

DEVROYE, L. and GYORFI, L. (1985). *Nonparametric Density Estimation: The L1 View*. Wiley, New York.

DEVROYE, L. and PENROD, C. S. (1984). Distribution-free lower bounds in density estimation. *Ann. Statist.* **12** 1250–1262.

EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596.

PARZEN, E. (1962). On the estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

STEELE, J. M. (1981). Complete convergence of short paths and Karp's algorithm for the tsp. *Math. Oper. Res.* **6** 374–378.

STEELE, J. M. (1982). Optimal triangulation on random samples in the plane. *Ann. Probab.* **10** 548–553.

VITALE, R. A. (1984). An expansion for symmetric statistics and the Efron–Stein inequality. In *Inequalities in Statistics and Probability* (Y. L. Tong, ed.) 112–114. IMS, Hayward, Calif.

WERTZ, W. (1976). Invariant density estimation. *Monatsh. Math.* **81** 315–324.

SCHOOL OF COMPUTER SCIENCE
MCGILL UNIVERSITY
805 SHERBROOKE STREET WEST
MONTREAL, QUEBEC H3A 2K6
CANADA