The Canadian Journal of Statistics
Vol.6 (1978), No.2, 193-200.
La Revue Canadienne de Statistique

INVALIDITY OF AVERAGE SQUARED ERROR CRITERION IN DENSITY ESTIMATION

by

J. Michael Steele*
Stanford University

Key words and phrases: Density estimation, mean integrated squared error, average squared error. *AMS 1970 subject classifications:* Primary 62G20; secondary 62F20.

ABSTRACT

The average squared error has been suggested earlier as an appropriate estimate of the integrated squared error, but an example is given which shows their ratio can tend to infinity. The results of a Monte Carlo study are also presented which suggest the average squared error can seriously underestimate the errors inherent in even the simplest density estimations.

1. INTRODUCTION

In almost all theoretical work on density estimation the quality of the estimator $\hat{f}_n(x)$ has been measured by the mean integrated squared error (MISE) which is given by

$$\text{MISE}(f,\hat{f}_n) = \mathcal{E}_f[\int_{-\infty}^{\infty}(f(x)-\hat{f}_n(x))^2 w(x)dx] \ . \tag{1}$$

Here $w(x)$ is a fixed weight function often taken to be 1, and $\mathcal{E}_f$ is the expectation under the true density $f(x)$.

While this criterion is very convenient in mathematical analyses, the presence of the n - fold integral represented succinctly by $\mathcal{E}_f$ presents considerable difficulty when the MISE is to be determined numerically. In order to circumvent this difficulty Wegman (1972) proposed the use of the average squared error (ASE) which is defined by

$$\text{ASE}(f,\hat{f}_n) = (1/n)\sum_{i=1}^{n}(\hat{f}_n(x_i)-f(x_i))^2 \ , \tag{2}$$

where $x_1, \ldots, x_n$ are the same observations used in the construction of $\hat{f}_n(x)$.

* Research supported in part by the National Research Council Canada while the author was at The University of British Columbia.

This is certainly a more convenient measure of quality than the MISE, but, this convenience is paid for in several ways. In the first place, how does one relate the ASE to the MISE? In the second, how much effect is there on the ASE due to the use of the same data in constructing $\hat{f}_n(x)$ as in testing $\hat{f}_n(x)$?

The original motivation for the ASE given by Wegman (1972, p.228) was that it should approximate the integrated squared error with a weight function $f(x)$. This interpretation was carried on by Fryer (1977) who refers to the ASE as "an experimental MISE". The first objective of this paper is to scrutinize this motivation further and to suggest the difficulties it poses.

The second objective is to report the results of simulations which were conducted to determine the effects of using the observations both for constructing and assessing $\hat{f}_n(x)$. Since such a procedure is analogous to testing a discriminator on the data used to construct it, one could expect the ASE to seriously underestimate the errors of estimation. Although the study reported here is modest in scope, the results clearly support this expectation. In the Conclusion, some problems are mentioned which would be of interest for the foundations of the theory of density estimation.

2. HOW CLOSE ARE THE ASE AND THE MISE?

The ASE can be written suggestively as

$$\int_{-\infty}^{\infty}(\hat{f}_n(x)-f(x))^2 dF_n(x) , \tag{3}$$

where $F_n(x)$ is the empirical distribution function. As Wegman notes, when n becomes large $dF_n(x)$ approximates $f(x)dx$, hence (3) approximates

$$\int_{-\infty}^{\infty}(\hat{f}_n(x)-f(x))^2 f(x)dx \quad . \tag{4}$$

This last integral is naturally the integrated squared error (ISE) with weight $f(x)$.

One intrinsic difficulty with this line of reasoning is that both (3) and (4) are approximately zero for large n. Hence, in using ASE as a stand-in for (4) which is in turn used as a stand-in for the MISE, the relevant comparison would be provided by the ratio of (3) to (4). The theorem of this section shows that even in the favorable case of estimation of smooth densities this ratio can be disappointingly large. This is one means of showing that the ASE is an inappropriate substitute for the MISE.

Before stating the theorem we recall that a density estimator $\hat{f}_n$ is consistent for a family of densities F provided that for any $f \in F$ one has

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} (\hat{f}_n(x)-f(x))^2 f(x)\,dx = 0 \ .$$

If the convergence above is almost sure, $\hat{f}_n$ is called strongly consistent, and if the convergence is in probability, $\hat{f}_n$ is simply said to be consistent.

THEOREM. *There is a density estimator which is strongly consistent for the class of differentiable squared integrable densities, but for which*

$$\frac{\mathrm{ISE}(\hat{f}_n,\phi)}{\mathrm{ASE}(\hat{f}_n,\phi)} \to \infty \quad a.s. \quad \text{as} \quad n\to\infty \ , \tag{5}$$

where ϕ *is the unit normal density.*

Proof. Let $\hat{g}_n$ be any strongly consistent density estimator. We will obtain $\hat{f}_n$ as a modification of $\hat{g}_n$. Now for any subset S of $\mathcal{R}$, $I_S(x)$ will be the indicator of S. Letting $A_i = [x_i - 1/n^2, x_i + 1/n^2]$ we set

$$r_n(x) = (1+n^{-2})(2\pi)^{-\frac{1}{2}} \sum_{i=1}^{n} e^{-\frac{1}{2}x_i^2} I_{A_i}(x) \ .$$

Next choose a sequence ν_n of reals which tends to infinity and which satisfies

$$\int_{\nu_n+\frac{1}{2}}^{\nu_n+1} \phi(x)\,dx \geq 1/n \ . \tag{6}$$

We then let $B_n = [\nu_n, \nu_n+1] \setminus \bigcup_{i=1}^{n} A_i$ and set

$$s_n(x) = I_{B_n}(x) \ .$$

Finally set $C_n = ((\bigcup_{i=1}^{n} A_i) \cup B_n)^c$ and define $\hat{f}_n(x)$ by the sum

$$\hat{f}_n(x) = \hat{g}_n(x) I_{C_n}(x) + r_n(x) + s_n(x) \ .$$

Write $\|\psi\|_2$ for the weighted L^2 norm of ψ which is given by

$$\left(\int_{-\infty}^{\infty} \psi^2 \phi\,dx\right)^{\frac{1}{2}} \ .$$

To prove $\hat{f}_n$ is consistent we apply Minkowski's inequality to obtain

$$\|f-\hat{f}_n\|_2 \leq \|f-\hat{g}_n\|_2 + \|\hat{g}_n - I_{C_n}\cdot \hat{g}_n\|_2 + \|r_n\|_2 + \|s_n\|_2 \ . \tag{7}$$

One easily sees that $\|f-\hat{g}_n\|_2$, $\|r_n\|_2$, and $\|s_n\|_2$ go *a.s.* to 0 as $n\to\infty$, so we consider the more subtle second term. By Schwarz' inequality,

$$\|\hat{g}_n - I_{c_n}\hat{g}_n\|_2 \le \|\hat{g}_n\|_2 \|I_{c_n^c}\|_2$$

$$\le \|\hat{g}_n\|_2 \Big(\int_{B_n \cup (\bigcup_{i=1}^n A_i)} \phi_n(x)\,dx\Big)^{\frac{1}{2}} \le \|\hat{g}_n\|_2 \Big(\int_{\nu_n}^{\nu_{n+1}} \phi(x)\,dx + \int_{\bigcup_{i=1}^n A_i} \phi(x)\,dx\Big)^{\frac{1}{2}}.$$

Since $\hat{g}_n$ is consistent $\|\hat{g}_n\|_2$ is bounded. Further since the measure of $\bigcup_{i=1}^n A_i$ is at most $2/n$ and by the boundedness of ϕ we see that

$$n\int_{\bigcup_{i=1}^n A_i} \phi(x)\,dx$$

goes to 0 as $n\to\infty$. Finally, since $\nu_n\to\infty$ we have

$$\int_{\nu_n}^{\nu_{n+1}} \phi(x)\,dx \to 0$$

as $n\to\infty$; hence (7) shows $\hat{f}_n$ is in fact strongly consistent. To prove the key condition (5) we note that

$$\mathrm{ASE}(f,\hat{f}_n) = \sum_{i=1}^n \left(\frac{n^{-2}e^{-\frac{1}{2}x_i^2}}{\sqrt{2\pi}}\right)^2 = \frac{1}{2\pi n^4}\sum_{i=1}^n e^{-\frac{1}{2}x_i^2} \le 1/n^3 .$$

Next we have

$$\mathrm{ISE}(f,\hat{f}_n) = \int_{-\infty}^{\infty}(f_n(x) - \phi(x))^2\phi(x)\,dx \ge \int_{B_n}(1-\phi(x))^2\phi(x)\,dx . \qquad (8)$$

Since $\bigcup_{i=1}^n A_i$ has measure at most $2/n$ we have

$$\mathrm{ISE}(f,\hat{f}_n) \ge \int_{\nu_n+2/n}^{\nu_n+1}(1-\phi(x))^2\phi(x)\,dx \ge (1-\phi(\nu_n))^2 \int_{\nu_n+\frac{1}{2}}^{\nu_n+1}\phi(x)\,dx . \qquad (9)$$

Finally, inequalities (6), (8), (9) show that the ratio of $\mathrm{ISE}(\hat{f}_n,\phi)$ to $\mathrm{ASE}(\hat{f}_n,\phi)$ increases like n^2 as n tends to infinity. Q.E.D.

Remarks. Much of the detail of this construction is caused by the necessity of making $\mathrm{ISE}(\hat{f}_n,\phi)$ non-zero and the desirability of making $\mathrm{ASE}(\hat{f}_n,\phi)$ also non-zero. The estimator f_n constructed here is not itself a density since it need not have integral 1. While it is not unusual to consider such estimators (e.g., orthogonal-series methods give non-density density estimators)

one should note that $\hat{f}_n$ can be adjusted easily to make it a proper density. Also, we note that $\hat{f}_n$ is not smooth, but it can be made smooth by routine modifications.

There is no claim that the example constructed above is "natural" but it serves well enough to pinpoint the possible extreme pathologies of the ASE. The more natural pathologies are taken up in the next section.

3. DOES THE ASE UNDERESTIMATE?

The fact that in some cases the ISE can be many times larger than the ASE might not deter a practical person's desire to use the ASE in assessing the quality of an estimator. On the other hand, if the ASE were seen to consistently underestimate the error of an estimator in very simple situations, then almost any application of the ASE would be dubious.

In order to detect this possibility we consider the new average sum of squared errors (NASE). If $\hat{f}_n$ is constructed on the basis of a sample $X_1, X_2, \ldots, X_n$ from a population with density f, then a second independent sample $X_1', X_2', \ldots, X_n'$ is drawn and we set

$$\text{NASE}(f, \hat{f}_n) = (1/n) \sum_{i=1}^{n} (\hat{f}_n(X_i') - f(X_i'))^2 .$$

We can now check that the motivations given to support the ASE's case for approximating ISE can be repeated verbatim on behalf of the NASE. First we write $\tilde{F}_n$ for the empirical distribution function of $X_1', X_2', \ldots, X_n'$. Since the X_i' have density $f(x)$ we see $d\tilde{F}_n$ approximates $f(x)dx$ just as dF_n did earlier. Since $\text{NASE}(f, \hat{f}_n)$ is precisely equal to

$$\int_{-\infty}^{\infty} (\hat{f}_n(x) - f(x))^2 d\tilde{F}_n(x) ,$$

the $\text{NASE}(f, \hat{f}_n)$ is thus an estimate of $\text{ISE}(f, \hat{f}_n)$ with just the same pedigree as the $\text{ASE}(f, \hat{f}_n)$.

The point to be made is that (a) if the a priori arguments in favour of NASE and ASE are the *same* and (b) if NASE and ASE *differ* significantly, then one must conclude that the a priori arguments do not constitute a significant motivation for the ASE.

It remains to be seen if the NASE and ASE are significantly different. To this end a modest Monte Carlo study was undertaken of the ratio

$$\text{RHO}(f, \hat{f}_n) = \text{NASE}(f, \hat{f}_n)/\text{ASE}(f, \hat{f}_n) . \qquad (10)$$

In order not to confound the difficulties of interpreting Monte Carlo evaluations of (10), RHO was calculated for very simple estimators of an essentially parametric type. The cases considered were the following:

I. The unit normal was estimated by $f_n(x) = \phi(x + \bar{x})$.

II. The triangular density with base $[0,1]$ was estimated by $\hat{f}_n(x)$, the triangular density of base $[0,2\bar{x}]$.

III. The rectangular density with base $[0,1]$ was estimated by $\hat{f}_n(x)$, the rectangular density with base $[0, \frac{n+1}{n} \max_{1 \le i \le n}(X_i)]$.

TABLE 1: Numbers of Values of RHO out of 1000 for Three Different Density Estimators.

Range of RHO	I - Normal n = 10	20	50	II - Triangular 10	20	50	III - Rectangular 10	20	50
0 - 1	36	1	0	55	11	0	1	0	0
1 - 2	85	15	0	138	71	7	532	512	506
2 - 3	121	45	2	127	116	44	44	5	0
3 - 4	110	60	8	86	107	54	33	17	1
4 - 5	83	56	8	85	81	77	34	16	1
5 - 6	77	62	11	60	75	67	31	15	3
6 - 7	54	56	27	46	55	69	24	19	3
7 - 8	39	49	37	26	29	45	23	21	9
8 - 9	36	37	19	12	32	44	12	24	4
9 - 10	26	38	30	4	31	42	13	9	4
10 - 11	27	32	23	2	19	34	11	19	8
11 - 12	20	31	29	0	29	27	12	15	8
12 - 13	18	40	23	1	21	28	13	17	6
13 - 14	16	21	32	0	9	36	8	13	7
14 - 15	13	14	14	5	6	16	10	4	4
15 - 16	10	19	31	2	5	9	14	6	7
16 - 17	12	22	13	1	2	7	7	8	3
17 - 18	9	11	20	2	0	13	8	13	9
18 - 19	12	17	17	0	0	3	3	10	8
19 - 20	12	15	18	3	0	10	8	9	11
20 - 30	61	102	154	30	6	126	34	56	53
30 - ∞	133	249	484	315	295	242	125	197	345
Mean	25	78	138	4700	55000	37000	34	89	118
St.Dev.	141	313	1100	41200	105000	10^6	316	1140	548
Largest	4000	30000	25000	93000	29×10^5	29×10^6	7900	26000	9300

For sample sizes n=10, n=20, and n=50 the value of RHO$(f,\hat{f}_n)$ was calculated 1000 times in each of the cases. For example, from Table 1 we see that in 1000 calculations of RHO by procedure II with a sample of size n=20 there were 29 times that RHO was between 7 and 8. Also, the mean, standard deviation and largest reported at the bottom of Table 1 refer to the set of all 1000 values generated for each column.

There are a number of conclusions which can be drawn from the calculations. The most naïve but most basic is that RHO is frequently very large and consequently the ASE may frequently give a serious under-estimation of the error inherent in estimating $f(x)$ by $\hat{f}_n(x)$. There are qualifications to be made and these are taken up in the Conclusion, but there is a lot of validity in the naïve observation.

4. CONCLUSION

The Theorem of Section 2 gives a theoretical reason why the ASE might be very small compared to the ISE, and the simulations of Section 3 show how ASE can be very small compared to the even more natural measure NASE. The clear conclusion is that the ASE can seriously under-estimate the errors of density estimation.

There are numerous possible criticisms of the procedures which have been applied here, yet none of these seem to seriously inveigh against the conclusion. One may observe that the example given in Section 2 is not natural since $\hat{f}_n$ was constructed precisely to have a small ASE when used to estimate ϕ. This may indeed be trickery, but one should not be prepared to take as a standard a measure which can be so easily tricked. Moreover, as it is always the case with counter-examples, once one has been produced it suggests the possibility of more natural ones existing all around us.

The criticisms of the second section are essentially the generic criticisms of any Monte Carlo study. Since all possible care was taken with the random number generation, one must conclude that the huge range of values of RHO given in Table 1 are true reflections of the ratio of the NASE and ASE. Since only three estimation problems were considered it is possible that there are density estimation problems in which ISE, NASE, and ASE are all comparable. While it would be interesting to know if such estimators exist, it seems already a sufficient indictment of the ASE that it is shown deficient by the three simplest estimators.

Beyond the basic conclusions to be drawn from Table 1, the data given there suggest several problems. Can one prove that under most circumstances

$$\lim_{n\to\infty} \text{NASE}(f,\hat{f}_n)/\text{ASE}(f,\hat{f}_n) = \infty \ a.s. \quad ?$$

This is hinted at by Table 1, and has been proved in some special cases but it would be interesting to know how generally it holds.

The main problem suggested by the preceding analysis is naturally the following: Is there a numerically expedient measure which accurately reflects the error of a density estimation? The incentives which lead to Wegman's original introduction of the ASE remain as valid as before, but the deficiencies of the ASE serve as an indication of the problems to be overcome.

RÉSUMÉ

L'erreur carrée approachée (ASE) a été introduite comme un bon estimateur de l'erreur carrée integrée (ISE). On présente un exemple où le quotient de ces deux erreurs tend vers l'infini. Cet article contient également les resultats d'une étude de Monte Carlo qui suggère que l'erreur carrée approché (ASE) sous-estime les erreurs encourues, même dans des cas très simples.

REFERENCES

Fryer, M.J. (1977). A review of some non-parametric methods of density estimation, *J. Inst. Math. Appl.*, 20, 335-354.

Wegman, E.J. (1972). Nonparametric probability density estimation: a comparison of density estimation methods. *J. Statist. Comp. and Simulation*, 1, 225-245.

Received 9 August 1978
Revised 11 October 1978

Department of Statistics
Stanford University
Stanford, California
94305 U.S.A.