No.562

On the asymptotic equivalence between Hellinger distance and Kullback-Leibler loss

by

Yuichiro Kanazawa Atuyuki Kogure

December 1993

On the asymptotic equivalence between Hellinger distance and Kullback-Leibler loss

Yuichiro Kanazawa
University of Tsukuba

Atsuyuki Kogure
University of Chiba

December 13, 1993

Abstract

Hellinger distance and Kullback-Leibler distance are shown to be asymptotically equivalent up to a constant. Implications of the result are discussed.

AMS 1980 Subject Classifications: Primary 62G05; Secondary 62E20.
Key words: Akaike's information criterion; Hellinger distance; histogram; integrated absolute error; integrated square error; kernel estimators; Kullback-Leibler loss; least-squares cross-validation; likelihood cross-validation.

1 Introduction

Suppose we wish to construct the kernel density estimator or histogram $\hat{f}$ with the width of its smoothing parameter being h based on a random sample $X_1, \ldots, X_n$ of size n from a density f. The kernel estimator $\hat{f}$ is defined by

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where $K(x)$ is a kernel function and h is a window width. The histogram is determined by the counts of observations in the cells $[a + (j-1)h, a + jh)$, $j = 1, \ldots, k$, where a is an origin, h is a cell width, and k is the number of cells. They are closed on the left and open on the right for definiteness.

Since the parameter h primarily controls the amount of smoothing and global discrepancy between the estimator and density, it is of interest to obtain the theoretically optimal width of the h. To do so, one first selects a global measure of discrepancy between the estimator and density. Then one minimizes, usually asymptotically, the expected value of the measure.

A few measures of discrepancy were utilized in density estimation. As a measure based on distance, Parzen(1962) used mean integrated squared error(MISE) $E\left\{\int (\hat{f} - f)^2\right\}$ to obtain the asymptotically optimal window width

$$h_{MISE}^{window} = \left[\frac{\int K(t)^2 dt}{\{\int t^2 K(t) dt\}^2}\right]^{1/5} \left\{\int f^{(2)}(x)^2 \, dx\right\}^{-1/5} n^{-1/5} \tag{2}$$

for the kernel estimator. Scott(1979) and Freedman and Diaconis(1981) used the MISE to obtain the asymptotically optimal cell width

$$h_{MISE}^{cell} = \left[\frac{6}{\int f^{(1)}(x)^2 \, dx}\right]^{1/3} n^{-1/3} \tag{3}$$

for the histogram.

Devroye and Györfi(1985) employed the asymptotic upper bound of mean integrated absolute error(MIAE) $E\left\{\int|\hat{f}-f|\right\}$ to obtain the close-to-optimal width

$$h_{MIAE}^{window}=(2\pi)^{-1/5}\left[\frac{\int K(t)^2dt}{\{\int t^2K(t)dt\}^2}\right]^{1/5}\left[\left\{\frac{\int|f^{(2)}(x)|\,dx}{\int f(x)^{1/2}\,dx}\right\}^2\right]^{-1/5}n^{-1/5}$$

for the kernel estimator. Similarly they obtained the close-to-optimal width

$$h_{MIAE}^{cell}=\left[\frac{8\left\{\int f(x)^{1/2}\,dx\right\}^2}{\pi\left\{\int|f^{(1)}(x)|\,dx\right\}^2}\right]^{1/3}n^{-1/3}$$

for the histogram.

As a measure founded on information-theoretic concepts, expected value of Kullback-Leibler loss(EKLL) $E\int f\log(f/\hat{f})$ was carefully examined by Hall(1987) to obtain the asymptotically optimal width

$$h_{EKLL}^{window}=(4\cdot|I|)^{1/5}\left[\frac{\int K(t)^2dt}{\{\int t^2K(t)dt\}^2}\right]^{1/5}\left\{\int_I\frac{f^{(2)}(x)^2}{f(x)}\,dx\right\}^{-1/5}n^{-1/5}\qquad(4)$$

for the kernel estimator for sufficiently smooth and compactly supported densities on $|I|$. Let ν_i be the number of data in the i-th cell, h the cell width, and J the total number of histogram cells. Taylor(1987) used expected value of Akaike's information criterion(EAIC) $J-\log\left[\Pi_{i=1}^{\#\text{ of cells}}\{\nu_i/(nh)\}^{\nu_i}\right]$ to obtain the asymptotically optimal cell width

$$h_{EAIC}^{cell}=\left[\frac{12|I|}{\int\{f^{(1)}(x)^2/f(x)\}\,dx}\right]^{1/3}n^{-1/3}\qquad(5)$$

for the histogram where $|I|$ represents the width of the support of the data. Hall(1990) employed Kullback-Leibler loss(KLL) to obtain the cell width

h^{cell}_{KLL} that asymptotically maximizes the KLL and that is *equivalent* to the h^{cell}_{EAIC} for sufficiently smooth and compactly supported densities. Here he had to work with raw loss KLL instead of the expected loss EKLL because there is a nonzero probability that the histogram vanishes in any given cell. The equivalence between the h^{cell}_{EAIC} and h^{cell}_{KLL} should not surprise us because AIC is essentially a model selection rule based on KLL as Akaike(1973) wrote: "...this [AIC] is equivalent to maximizing an information theoretic quantity which is given by the definition

$$E \log \frac{f(x|\hat{\theta})}{f(x|\theta)} = E \int f(x|\theta) \log \frac{f(x|\hat{\theta})}{f(x|\theta)} dx.$$

The integral in the right-hand side of the above equation gives the Kullback-Leibler's mean information for discrimination between $f(x|\hat{\theta})$ and $f(x|\theta)$...."

As for the relationship between distance-based and information-theoretic measures of discrepancy, we know the following. For the kernel estimator, the window width h^{window}_{MHD} that minimizes mean Hellinger distance(MHD)

$$MHD(\hat{f}, f) = E \int \left\{ \hat{f}(x)^{1/2} - f(x)^{1/2} \right\}^2 dx$$

was found asymptotically equivalent in Kanazawa(1993b) to the h^{window}_{EKLL} in (4). For the histogram, the cell width h^{cell}_{MHD} that minimizes the MHD was recognized asymptotically equivalent in Kanazawa(1993a) to the h^{cell}_{EAIC} in (5) as well. The results strongly suggest that information-theoretic Kullback-Leibler loss itself be equivalent to distance-based Hellinger distance(HD) in some sense. We shall investigate the asymptotic relationship in section 2. We shall discuss its implications in density estimation in section 3. The proof is given in section 4. For Hellinger distance applied to the histogram with varying cell widths, see Kanazawa(1988,1992).

2 Asymptotic result

Let $\hat{f}$ be the kernel estimator or histogram. For the former, assume:

K1. $K(t)$ is a compactly supported function on I symmetric about 0.

K2. $\int_I K(t)dt = 1$.

K3. $\int_I t^2 K(t)dt \neq 0$.

Remark 1 These assumptions are commonly made in density estimation and satisfied by such standard compactly supported kernels as Epanechnikov, biweight, triangular, and rectangular. See Silverman(1986, p38–43).

Define variance-stabilizing integrated square error(VSISE) as

$$VSISE(\hat{f}, f) \equiv \int \frac{\left\{\hat{f}(x) - f(x)\right\}^2}{f(x)}\, dx. \tag{6}$$

We shall explain what variance-stabilizing means in section 3.

Theorem 1 *Let the following conditions on the density f be satisfied:*

D1. *f is compactly supported on I.*

D2. *f is twice continuously differentiable on I.*

D3. $\int_I f^{(j)}(x)^2/f(x)\, dx < \infty$, $j = 1, 2$.

Then as $n \to \infty$, the following relationships hold.

$$VSISE(\hat{f}, f) = 4 \cdot HD(\hat{f}, f) = 2 \cdot KLL(\hat{f}, f).$$

Remark 2 Notice from the kernel estimators in (2) and (4) that, the MISE-based bias $(f^{(2)})^2$ is divided by $f(x)$ to yield the MHD-based $f^{(2)}(x)^2/f(x)$, as is the MISE-based variance f which results in the MHD-based $\int dx$. Similar

in (3) and (5) for the histogram, except that $f^{(2)}$ is now replaced by $f^{(1)}$. The smoothing parameters h are exactly those that would be obtained if we started with expected value of the VSISE in place of the MHD or EKLL.

3 Discussions

In density estimation, integrated square error(ISE) has generally come to be recognized as the *standard* measure of discrepancy between the density and estimator. There are several solid reasons. First, the ISE was the original measure of discrepancy used in Rosenblatt(1956), is mathematically tractable, and mean square error is familiar to us. Second, the asymptotically mean ISE-optimal window width in (2) for the kernel estimator and cell width in (3) for the histogram show that the smoothing parameters do not contain the width of support for the density. This makes the ISE especially attractive because asymptotically optimal size of smoothing parameter can be computed for the Gaussian density, a density considered to be a reference by most statisticians. Third, a remarkable result of Stone(1984a) showed least-squares cross-validation(LSCV) achieves asymptotically the best choice of window width for the kernel estimator in the sense of minimizing the ISE, assuming only f is bounded, and under very mild conditions on the kernel. A similar result in Stone(1984b) showed, however, that asymptotically the best choice of cell width for the histogram is attained only if the LSCV was applied to the compactly supported densities f.

There seem to be evidences, both theoretically and in practice, mounting against the use of the information-theoretic KLL and its accompanying data-

based methods of likelihood cross-validation(LCV) for the kernel estimator and of AIC for the histogram. First, the KLL, though proposed relatively early in Kullback(1959), was purpose-built for discrimination and is argued as an inappropriate measure of discrepancy because $\hat{f}$ and f are not interchangeable. Second, the asymptotically mean KLL-optimal window width in (4) for the kernel estimator and cell width in (5) for the histogram show the smoothing parameters include the width $|I|$ of support for the density. The KLL should therefore be used only for the densities with compact support, which excludes the Gaussian density. Third, tail sensitive nature of the KLL was often mentioned. Schuster and Gregory(1981) found that the LCV will not select consistent estimates of the density for long tailed distributions such as the double exponential and Cauchy distributions. Hall(1987) showed that the KLL's asymptotic properties are profoundly influenced by tail properties of the kernel and of the density. Broniatowski, Deheuvels, and Devroye(1989) found that the LCV is non-consistent whenever the tails of f decrease as an exponential or slower.

In spite of the seemingly insurmountable evidences, we would argue that the HD, KLL, or their asymptotic analog, the VSISE in (6), is a reasonable alternative to the more conventional ISE for the following reasons.

First the argument that neither the KLL nor VSISE is not mathematical distance and thus not suitable for density estimation is somewhat weakened, at least asymptotically, because $\hat{f}$ and f are interchangeable in the HD.

The second reason, far more significant in our judgment, has to do with whether one can afford to have the density estimator fluctuate more in one region than in the other. We shall make our case using the following least-

squares estimation situation in which the variances in error terms vary with observations. It is of accepted practice to adjust the residual sum of squares in that situation by multiplying the inverse of its heteroscedastic, and thereby non-identity, variance-covariance matrix to obtain the best linear unbiased estimator. The same principle, we believe, needs to be employed to density estimation, resulting the use as our measure of the discrepancy of

$$\int \frac{\left\{\hat{f}(x) - f(x)\right\}^2}{V(\hat{f}(x))}\, dx. \tag{7}$$

For the window width h, variance of the kernel estimator $\widehat{f_{kernel}}(x)$ is

$$V(\widehat{f_{kernel}}(x)) = \frac{f(x)}{nh} \int K(t)^2\, dt + O(n^{-1}).$$

For the cell width h, variance of the histogram $\widehat{f_{histogram}}(x)$ is

$$V(\widehat{f_{histogram}}(x)) = \frac{f(x)}{nh} + O(n^{-1}).$$

If we assume the smoothing parameter does not vary with x, the variance $V(\hat{f}(x))$ in (7), whether the $\hat{f}(x)$ is the kernel estimator or histogram, is proportional to $f(x)$. We obtain (6) by substituting $f(x)$ for $V(\hat{f}(x))$.

It is true that data-based methods of LCV for the kernel estimator and of AIC for the histogram based on the HD, KLL, or VSISE must be used only for compactly supported densities. It is also true that the density estimators constructed by LCV or AIC must be more tail-sensitive than those based on the ISE because tail areas have relatively more weight in the HD, KLL, or VSISE than in the ISE. What we gain in return in the density estimators constructed by LCV or AIC is that they properly discounts the contributions

from high density regions where the variabilities are also high by giving there the weight $1/f(x)$ to have the variance stabilized over the support, a property from which the name "variance-stabilizing"in (6) is originated.

The more conventional ISE, neglecting this essential weighting operation, fails in this respect. Actually data-based methods of LSCV for the kernel estimator and histogram based on the ISE can be applied to infinitely supported densities without being tail-sensitive, precisely because the ISE neglects this weighting operation and concentrate on the high density region.

We believe that seemingly limited applicability of data-based methods of LCV for the kernel estimator and of AIC for the histogram to compactly supported densities should not discourage us from using them as tools for exploring data because, as Chambers, Cleveland, Kleiner, and Tukey(1983) pointed out: "...Real data can never come from a genuine normal distribution, for that would require data...with the possibility of including arbitrarily large and small values. In practical terms,...they are bounded above and below...."

4 Proofs

Proof. Since $(1+x)^{1/2} = 1 + x/2 - x^2/8 + O(x^3)$, we have

$$\left\{\hat{f}(x)^{1/2} - f(x)^{1/2}\right\}^2 = \hat{f}(x) + f(x) - 2\left\{1 + \frac{\hat{f}(x) - f(x)}{f(x)}\right\}^{1/2} f(x)$$

$$= \hat{f}(x) + f(x) - 2\left[1 + \frac{1}{2}\left\{\frac{\hat{f}(x) - f(x)}{f(x)}\right\} - \frac{1}{8}\left\{\frac{\hat{f}(x) - f(x)}{f(x)}\right\}^2 + O_p\left(\left\{\frac{\hat{f}(x) - f(x)}{f(x)}\right\}^2\right)\right] f(x)$$

$$= \frac{1}{4}\frac{\left\{\hat{f}(x)-f(x)\right\}^2}{f(x)} + O_p\left(\frac{\left\{\hat{f}(x)-f(x)\right\}^3}{f(x)^2}\right).$$

Thus Hellinger distance has become

$$\begin{aligned} HD(\hat{f},f) &\equiv \int\left\{\hat{f}(x)^{1/2}-f(x)^{1/2}\right\}^2 dx \\ &= \frac{1}{4}\int\frac{\left\{\hat{f}(x)-f(x)\right\}^2}{f(x)}dx + O_p\left(\int\frac{\left\{\hat{f}(x)-f(x)\right\}^3}{f(x)^2}dx\right). \quad (8) \end{aligned}$$

Since $\log(1+x) = x - x^2/2 + O(x^3)$, we have

$$\begin{aligned} \log\frac{f(x)}{\hat{f}(x)} &= -\log\frac{\hat{f}(x)}{f(x)} = -\log\left\{1+\frac{\hat{f}(x)-f(x)}{f(x)}\right\} \\ &= -\frac{\hat{f}(x)-f(x)}{f(x)} + \frac{1}{2}\left\{\frac{\hat{f}(x)-f(x)}{f(x)}\right\}^2 + O_p\left(\left\{\frac{\hat{f}(x)-f(x)}{f(x)}\right\}^3\right). \end{aligned}$$

Thus Kullback-Leibler loss has become

$$\begin{aligned} KLL(\hat{f},f) &\equiv \int f(x)\log\frac{f(x)}{\hat{f}(x)}dx \\ &= \frac{1}{2}\int\frac{\left\{\hat{f}(x)-f(x)\right\}^2}{f(x)}dx + O_p\left(\int\frac{\left\{\hat{f}(x)-f(x)\right\}^3}{f(x)^2}dx\right). \quad (9) \end{aligned}$$

Hence

$$\begin{aligned} VSISE(\hat{f},f) &= 4\cdot HD(\hat{f},f) + O_p\left(\int\frac{\left\{\hat{f}(x)-f(x)\right\}^3}{f(x)^2}dx\right) \\ &= 2\cdot KLL(\hat{f},f) + O_p\left(\int\frac{\left\{\hat{f}(x)-f(x)\right\}^3}{f(x)^2}dx\right) \end{aligned}$$

Under **K1–K3** and **D1–D3**, routine calculations show that

$$\hat{f} = f + O(h^2) + O_p(n^{-1/2}h^{-1/2})$$

for the kernel estimator. Similarly under **D1–D3**, we have

$$\hat{f} = f + O(h) + O_p(n^{-1/2}h^{-1/2})$$

for the histogram. The terms of order $O_p\left(\int \frac{\{\hat{f}(x)-f(x)\}^3}{f(x)^2}\,dx\right)$ in (8) and (9) therefore are asymptotically negligible. □

Acknowledgments This research was supported by the Grant-in-Aid for the Scientific Research from the Japanese Ministry of Education and the University of Tsukuba Project Research. A part of this research was presented at the Japanese Statistical Society annual conference in July, 1993. The authors thank Dr. M.C.Jones at The Open University for his interest in our work and for his encouragement to us communicated through a letter.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In Petrov B.N. and Cźaki P., editors, *2nd International Symposium on Information Theory*, pages 267–281, Akademiai Kiadó, Budapest, Hungary, 1973.

[2] M. Broniatowski, P. Deheuvels, and L. Devroye On the relationship between stability of extreme order statistics and convergence of the maximum likelihood kernel density estimate. *The Annals of Statistics*, 17:1070–1086, 1989.

[3] J.M. Chambers, W.S. Cleveland, B. Kleiner, and P.A. Tukey. *Graphical methods for data analysis.* Wadsworth International Group & Duxbury Press, Belmont, California & Boston, Massachusetts, 1983.

[4] L. Devroye and L. Györfi. *Nonparamtric Density Estimation The L_1 view.* John Wiley & Sons, New York, 1985.

[5] D. Freedman and P. Diaconis. On the histogram as a density estimator: l_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57:453–476, 1981.

[6] P. Hall. Akaike's information criterion and Kullback-Leibler loss for histogram density estimation. *Probability Theory and Related Fields*, 85:449–467, 1990.

[7] P. Hall. On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15:1491–1519, 1987.

[8] M.C. Jones. Personal communication. 1993.

[9] Y. Kanazawa. Hellinger distance and Akaike's information criterion for the histogram. *Statistics & Probability Letters*, 17(4):293–298, 1993.

[10] Y. Kanazawa. Hellinger distance and Kullback-Leibler loss for the kernel density estimator. *Statistics & Probability Letters*, 18(4):315–321, 1993.

[11] Y. Kanazawa. An optimal variable cell histogram. *Communications in Statistics, Part A: Theory and Methods*, 17:1401–1422, 1988a.

[12] Y. Kanazawa. An optimal variable cell histogram based on the sample spacings. *The Annals of Statistics*, 20:291–304, 1992.

[13] S. Kullback. *Information Theory and Statistics.* John Wiley & Sons, New York, 1959.

[14] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[15] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27:832–837, 1956.

[16] E.F. Schuster and G.G. Gregory On the nonconsistency of maximum likelihood nonparametric density estimators. In Eddy W.F., editor, *Computer Science and Statistics, Proceedings of the 13th Symposium on the Interface*, pages 295–298, Springer-Verlag, New York, N.Y., 1981.

[17] D.W. Scott. On optimal and data-based histograms. *Biometrika*, 66:605–610, 1979.

[18] B.W. Silverman Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, 1986.

[19] C.J. Stone. An asymptotically optimal histogram selection rule. In Le Cam L.M. and Olshen R.A., editors, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, vol.II*, pages 513–520, Wadsworth, Monterey, CA, 1984.

[20] C.J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12:1285–1297, 1984.

[21] C.C. Taylor. Akaike's information criterion and the histogram. *Biometrika*, 74:636–639, 1987.