On Nonparametric Discrimination Using Density Differences
Author(s): Peter Hall and Matthew P. Wand
Source: *Biometrika*, Vol. 75, No. 3 (Sep., 1988), pp. 541-547

Biometrika (1988), **75**, 3, *pp.* 541–7
Printed in Great Britain

On nonparametric discrimination using density differences

BY PETER HALL AND MATTHEW P. WAND
Department of Statistics, Australian National University, Canberra, ACT 2601, Australia

SUMMARY

We propose a technique for nonparametric discrimination in which smoothing parameters are chosen jointly, according to a criterion based on the difference between two densities. The approach is suitable for categorical, continuous and mixed data, and uses information from both populations to determine the smoothing parameter for any one population. In the case of categorical data, optimal performance is sometimes achieved using negative smoothing parameters, a property which does not emerge if the smoothing parameters are chosen individually.

Some key words: Categorical data; Continuous data; Cross-validation; Density difference; Density estimate; Discrimination; Kernel estimate; Likelihood ratio.

1. INTRODUCTION

The common approach to nonparametric discrimination, using either continuous or categorical data, is to construct nonparametric density estimators for individual populations and combine them using a likelihood ratio rule. For example, suppose we wish to discriminate between X and Y populations. Let f_X and f_Y be the respective true probability densities, and let p be the prior probability that an unclassified observation z is from population X. The 'ideal' rule is to assign z to X if and only if

$$f_X(z)/f_Y(z) \geqslant (1-p)/p. \tag{1·1}$$

If f_X and f_Y are unknown then they are usually replaced by estimates $\hat{f}_X$ and $\hat{f}_Y$, whose construction depends crucially on respective smoothing parameters h_X and h_Y. The value of h_X is chosen to minimize the distance between $\hat{f}_X$ and f_X, and h_Y is chosen to minimize the distance between $\hat{f}_Y$ and f_Y. We classify z as coming from X if and only if

$$\hat{f}_X(z)/\hat{f}_Y(z) \geqslant (1-p)/p. \tag{1·2}$$

In the present paper we suggest the following alternative approach. Notice that inequalities (1·1) and (1·2) are equivalent to $g(z) \geqslant 0$ and $\hat{g}(z) \geqslant 0$, respectively, where

$$g \equiv pf_X - (1-p)f_Y, \quad \hat{g} \equiv p\hat{f}_X - (1-p)\hat{f}_Y.$$

One can select h_X and h_Y jointly, rather than separately, to minimize the distance between $\hat{g}$ and g. Following that, a natural discrimination rule is to assign z to population X if and only if $\hat{g}(z) \geqslant 0$.

This procedure uses data in training samples from both populations to determine the smoothing parameter in a density estimate for any one population. That is an attractive feature, since data from each population provides information which is helpful in discriminating against the other, and which should be incorporated into both density estimates.

We use the L_2 metric to measure the distance between g and $\hat{g}$, and employ a variant of squared-error cross-validation as a tool for minimizing this distance. The criterion which we minimize numerically is a continuous function of the smoothing parameters. In this respect our approach has an advantage over rules which minimize an estimate of error rate (Tutz, 1986). The estimate of error rate is a discontinuous function of smoothing parameters, and itself requires smoothing for effective implementation. Nevertheless, Tutz's criterion does have a direct and very important meaning in terms of the discrimination problem.

In principle, a version of our procedure may be used to select smoothing parameters which minimize mean squared error of $\hat{f}_X/\hat{f}_Y$ or $\hat{f}_Y/\hat{f}_X$. However, such a technique is hardly practicable, because of serious problems when the denominator is close to zero. It is influenced far too greatly by cells with low counts, in the discrete case, or by tail properties of f_X and f_Y, in the continuous case.

Although our density estimators are constructed according to a criterion based on the difference between two densities, they are consistent and have smoothing parameters which are not radically different from those which would normally be employed to compute estimators of individual densities. For example, if we are using nonnegative kernel estimators with continuous univariate data, then smoothing parameters which are 'optimal' for individual densities and for density differences are all asymptotic to constant multiples of $n^{-1/5}$, only the constants being different. Section 2 discusses our approach in the case of categorical data, and § 3 treats continuous data. There is no difficulty using our method in other cases, for example with categorical data estimates (Wang & Van Ryzin, 1981), orthogonal series estimates for discrete data (Ott & Kronmal, 1976), and estimates for mixed data (Krzanowski, 1980; Vlachonikolis & Marriott, 1982; Hall, 1983).

The traditional approach to nonparametric discrimination, in which smoothing parameters are determined separately rather than jointly, is surveyed and analysed by Hand (1981, 1982). More recent work includes contributions by Titterington et al. (1981), Hand (1983), Krzanowski (1983), Lauder (1983) and Butler & Kronmal (1985).

2. CATEGORICAL DATA

In this section we assume that the sample space is the binary space $\{0, 1\}^d$, that is the set of all d-tuples of zeros and ones. It may be thought of as representing the set of all possible responses to d questions, for each of which the answer is 'yes' or 'no'. Our techniques also apply to other types of categorical data, such as unstructured multinomial data.

Assume that we have training samples $X_1, \ldots, X_m$ from the X-population and $Y_1, \ldots, Y_n$ from the Y-population, and that these samples are independent. Let f_X, f_Y, $\hat{f}_X$, $\hat{f}_Y$, g, $\hat{g}$ and p be as in § 1. Given a vector $x \equiv (x^{(1)}, \ldots, x^{(d)})$ of zeros and ±ones, put $|x| = \Sigma|x^{(i)}|$. Let

$$f_{X,1}(z) \equiv \sum_{x:|x-z|=1} f_X(x)$$

denote the sum of f_X-probabilities in cells distant one unit from cell z, and define $f_{Y,1}(z)$

analogously. Density estimates based on X and Y training samples are

$$\hat{f}_X(z\,|\,h_X) \equiv m^{-1} \sum_{i=1}^{m} h_X^{|z-X_i|}(1-h_X)^{d-|z-X_i|},$$
$$\hat{f}_Y(z\,|\,h_Y) \equiv n^{-1} \sum_{i=1}^{n} h_Y^{|z-Y_i|}(1-h_Y)^{d-|z-Y_i|}, \tag{2·1}$$

respectively (Aitchison & Aitken, 1976). Taking $h_X = 0$ reproduces relative frequencies of the X-sample, and $h_Y = 0$ has the same effect for the Y-sample.

A little algebra along the lines described by Hall (1981) and Bowman, Hall & Tittington (1984) shows that, as m and n increase, the values of h_X and h_Y which minimize $\Sigma_z E\{\hat{g}(z) - g(z)\}^2$ converge to zero and satisfy

$$h_X \sim h_{X,o} \equiv (T_{XX}T_{YY} - T_{XY}^2)^{-1}(T_{YY}m^{-1}S_X + \rho T_{XY}n^{-1}S_Y), \tag{2·2}$$
$$h_Y \sim h_{Y,o} \equiv (T_{XX}T_{YY} - T_{XY}^2)^{-1}(T_{XX}n^{-1}S_Y + \rho^{-1}T_{XY}m^{-1}S_X), \tag{2·3}$$

provided $T_{XX}T_{YY} - T_{XY}^2 \neq 0$, where $\rho \equiv p^{-1} - 1$,

$$T_{XX} \equiv \sum (f_{X,1} - df_X)^2, \quad T_{YY} \equiv \sum (f_{Y,1} - df_Y)^2,$$
$$T_{XY} \equiv \sum (f_{X,1} - df_X)(f_{Y,1} - df_Y),$$
$$S_X \equiv d + \sum (f_{X,1} - df_X)f_X, \quad S_Y \equiv d + \sum (f_{Y,1} - df_Y)f_Y.$$

An intriguing aspect of formulae (2·2) and (2·3) is that one or other of the 'optimal' smoothing parameters $h_{X,o}$, $h_{Y,o}$ can be negative. For example, take $d = 2$ and let f_X, f_Y be given by

$$f_X(0,0) = f_Y(1,0) = 0{\cdot}1, \quad f_X(0,1) = f_Y(1,1) = 0{\cdot}2,$$
$$f_X(1,0) = f_Y(0,0) = 0{\cdot}3, \quad f_X(1,1) = f_Y(0,1) = 0{\cdot}4.$$

Then

$$h_{X,o} = \tfrac{45}{32}(5m^{-1} - 3\rho n^{-1}), \quad h_{Y,o} = \tfrac{45}{32}(5n^{-1} - 3\rho^{-1}m^{-1}).$$

If $5n < 3\rho m$ then $h_{X,o} < 0$, and if $5\rho m < 3n$ then $h_{Y,o} < 0$.

A negative smoothing parameter means that, when large numbers of observations in neighbouring cells appear to suggest that the probability in the present cell should be weighted up, it is actually weighted down. This is not so absurd as might at first appear. We are estimating the difference between two densities, and not an individual density. If a cell z, distant one unit from cell z_0, makes a positive contribution to an optimally constructed estimator of $f_Y(z_0)$, then it will make a negative contribution to that estimator of $pf_X(z_0) - (1-p)f_Y(z_0)$ which is obtained by simply subtracting estimators of $f_X(z_0)$ and $f_Y(z_0)$. This may well be suboptimal, particularly if an accurate estimator of $f_X(z_0)$ assigns a large positive weight to data from cell z. However, note that a negative smoothing parameter leads to kernel weights which oscillate in sign as the distance from the cell at which the estimator is evaluated increases.

Minimizing $\Sigma_z E\{\hat{g}(z) - g(z)\}^2$ is equivalent to minimizing

$$\begin{aligned}\Delta(h_X, h_Y) \equiv{}& \textstyle\sum_z E\{p\hat{f}_X(z\,|\,h_X) - (1-p)\hat{f}_Y(z\,|\,h_Y)\}^2 \\ &- 2\textstyle\sum_z [p^2 f_X(z) E\hat{f}_X(z\,|\,h_X) + (1-p)^2 f_Y(z) E\hat{f}_Y(z\,|\,h_Y) \\ &\quad - p(1-p)\{f_X(z) E\hat{f}_Y(z\,|\,h_Y) + f_Y(z) E\hat{f}_X(z\,|\,h_X)\}].\end{aligned}$$

An unbiased estimate of $\Delta(h_X, h_Y)$ is

$$\hat{\Delta}(h_X, h_Y) \equiv \sum_z \{p\hat{f}_X(z \mid h_X) - (1-p)\hat{f}_Y(z \mid h_Y)\}^2 - 2\Bigg[p^2 m^{-1} \sum_{i=1}^{m} \hat{f}_{X,i}(X_i \mid h_X) + (1-p)^2 n^{-1} \sum_{i=1}^{n} \hat{f}_{Y,i}(Y_i \mid h_Y) - p(1-p)\Bigg\{m^{-1} \sum_{i=1}^{m} \hat{f}_Y(X_i \mid h_Y) + n^{-1} \sum_{i=1}^{n} \hat{f}_X(Y_i \mid h_X)\Bigg\}\Bigg], \tag{2·4}$$

where

$$\hat{f}_{X,i}(z \mid h_X) \equiv (m-1)^{-1} \sum_{j \neq i} h_X^{|z-X_j|} (1-h_X)^{d-|z-X_j|},$$

$$\hat{f}_{Y,i}(z \mid h_Y) \equiv (n-1)^{-1} \sum_{j \neq i} h_Y^{|z-Y_j|} (1-h_Y)^{d-|z-Y_j|}.$$

A practical procedure is to choose $(\hat{h}_X, \hat{h}_Y)$ to minimize $\hat{\Delta}(h_X, h_Y)$. An unclassified observation z may then be assigned to population X or Y according as

$$p\hat{f}_X(z \mid \hat{h}_X) - (1-p)\hat{f}_Y(z \mid \hat{h}_Y) \geqslant 0 \text{ or } < 0.$$

A slight variant of the criterion $\hat{\Delta}(h_X, h_Y)$ may be arrived at by following a prescription based on cross-validation, much as by Titterington (1978, 1980) or Bowman (1984). We have settled on $\hat{\Delta}$ because it is a little simpler than its cross-validatory counterpart, although the two are asymptotically equivalent. The argument which postulates $\hat{\Delta}$ because it is an unbiased estimator of Δ, is close to ones given by Rudemo (1982) and Brown & Rundell (1985) in related settings.

An alternative approach is to insist from the outset that $h_X = h_Y$. Then the 'optimal' window is asymptotic to

$$h_0 \equiv \frac{m^{-1}\{d + \sum f_X(f_{X,1} - df_X)\} + n^{-1}\rho^2\{d + \sum f_Y(f_{Y,1} - df_Y)\}}{\sum \{(f_{X,1} - df_X) - \rho(f_{Y,1} - df_Y)\}^2},$$

as m and n increase. A practical procedure for choosing the smoothing parameter in this circumstance is to minimize $\hat{\Delta}(h, h)$, given by (2·4) with $h_X = h_Y = h$. Another variant is to minimize $\sum_z E\{\hat{g}(z) - g(z)\}^2 w(z)$, for a weight function w. A slight modification of the function $\Delta(h_X, h_Y)$ produces an adaptive version of this criterion.

An argument similar to that given by Bowman et al. (1984) shows that, as $m, n \to \infty$, the smoothing parameters $(\hat{h}_X, \hat{h}_Y)$ which minimize $\hat{\Delta}(h_X, h_Y)$ satisfy $\hat{h}_X / h_X \to 1$ and $\hat{h}_Y / h_Y \to 1$ in probability, where h_X and h_Y are given by (2·2) and (2·3). In this sense, our adaptive criterion produces asymptotically optimal smothing parameters. Analogous results hold under the restriction $h_X = h_Y$.

We applied the binary kernel estimates, defined at (2·1), to data from Anderson et al. (1972) on diagnosis of keratoconjunctivitis sicca, KCS. The X-sample comprises $m = 40$ KCS patients, and the Y-sample contains $n = 37$ non-KCS patients. We took $p = \frac{1}{2}$. Table 3·1(a) lists smoothing parameters (h_X, h_Y) obtained by: (i) minimizing $\hat{\Delta}(h_X, h_Y)$, defined at (2·4); (ii) applying squared-error cross-validation to X- and Y-samples individually; and (iii) applying likelihood cross-validation to X- and Y-samples individually, as Aitchison & Aitken (1976).

Each method was then tested by omitting one observation from the training set in turn and using the reduced set to classify the omitted observation. Table 3·1(b) lists the errors incurred by each method. When viewed in this light, for this data set, method (i) performs in between methods (ii) and (iii).

Table 3·1. (*a*) *Smoothing parameters*, (*b*) *total misclassifications obtained by* '*leaving-one-out*' *method*

	(a)			(b)	
	h_X	h_Y		No. misclass. from X	No. misclass. from Y
(i)	0·2161	0·0124	(i)	4	2
(ii)	0·1950	0·0083	(ii)	4	3
(iii)	0·1570	0·0400	(iii)	4	1

Method (i) minimizes $\hat{\Delta}(h_X, h_Y)$, defined at (2·4); method (ii), squared-error cross-validation applied to X- and Y-samples individually; method (iii), likelihood cross-validation applied to X- and Y-samples individually.

3. Continuous data

Again suppose that we have training samples $X_1, \ldots, X_m$ from the X-population and $Y_1, \ldots, Y_n$ from the Y-population, and that these samples are independent. Assume that the data are d-variate, and let K be a d-variate kernel function. Write $h = (h_1, \ldots, h_d)$ for a d-variate smoothing parameter, define x/h to be $(x_1/h_1, \ldots, x_d/h_d)$ for vectors $x = (x_1, \ldots, x_d)$, and put

$$\hat{f}_X(z\,|\,h) \equiv \left(m \prod_{i=1}^{d} h_i\right)^{-1} \sum_{i=1}^{m} K\{(z - X_i)/h\},$$

$$\hat{f}_Y(z\,|\,h) \equiv \left(n \prod_{i=1}^{d} h_i\right)^{-1} \sum_{i=1}^{n} K\{(z - Y_i)/h\},$$

$$\hat{f}_{X,i}(z\,|\,h) \equiv \left\{(m-1) \prod_{j=1}^{d} h_j\right\}^{-1} \sum_{j \neq i} K\{(z - X_j)/h\},$$

$$\hat{f}_{Y,i}(z\,|\,h) \equiv \left\{(n-1) \prod_{j=1}^{d} h_j\right\}^{-1} \sum_{j \neq i} K\{(z - Y_j)/h\}.$$

See Prakasa Rao (1983, Ch. 2, 3) for details of nonparametric density estimation with continuous data. A very important special case of these estimates is that in which each component h_i is identical. There it is usual to standardize each component for scale.

Following the argument leading to (2·4) we see that on the present occasion our adaptive criterion is

$$\begin{aligned}\hat{\Delta}(h_X, h_Y) \equiv \int \{p\hat{f}_X(z\,|\,h_X) - (1-p)\hat{f}_Y(z\,|\,h_Y)\}^2\, dz \\ -2\Bigg[p^2 m^{-1} \sum_{i=1}^{m} \hat{f}_{X,i}(X_i\,|\,h_X) + (1-p)^2 n^{-1} \sum_{i=1}^{n} \hat{f}_{Y,i}(Y_i\,|\,h_Y) \\ -p(1-p)\left\{m^{-1} \sum_{i=1}^{m} \hat{f}_Y(X_i\,|\,h_Y) + n^{-1} \sum_{i=1}^{n} \hat{f}_X(Y_i\,|\,h_X)\right\}\Bigg].\end{aligned}$$

The integral on the right-hand side may be calculated explicitly if the standard normal kernel is used. The pair $(\hat{h}_X; \hat{h}_Y)$ which minimizes $\hat{\Delta}(h_X, h_Y)$ is asymptotically optimal in the sense of minimizing the L_2 distance between $\hat{g}$ and g. That is,

$$\frac{D(\hat{h}_X, \hat{h}_Y)}{\inf D(h_X, h_Y)} \to 1$$

almost surely as $m, n \to \infty$, where

$$D(h_X, h_Y) \equiv \int \{\hat{g}(z) - g(z)\}^2 \, dz$$

$$= \int [p\hat{f}_X(z \mid h_X) - (1-p)\hat{f}_Y(z \mid h_Y) - \{pf_X(z) - (1-p)f_Y(z)\}]^2 \, dz.$$

This result may be proved essentially by the arguments of Stone (1984). Regularity conditions are that $\int g^2 < \infty$, that f_X and f_Y and their one-dimensional marginals are bounded, and that the kernel is compactly supported, symmetric and Hölder continuous. Conditions on the kernel may be relaxed by following Hall (1985).

By focusing on the difference between two density estimates rather than on their ratio, we circumvent difficulties caused by negative density estimates. It is well known that estimates with particularly fast rates of convergence are sometimes negative (Prakasa Rao, 1983, p. 42 ff), but such estimates can be difficult to interpret if discrimination is based on a ratio rule.

We applied our kernel estimates to two discrimination problems. The first involved synthetic data. The X-population had the standard normal density while the standard Cauchy density was chosen for the Y-population. We simulated 50 training samples, each with $m = n = 50$, assumed equal prior probabilities, and used the standard normal kernel. For each training sample we calculated estimates of (h_X, h_Y) using (i) minimization of $\hat{\Delta}(h_X, h_Y)$, and (ii) squared-error cross-validation applied to the X- and Y-samples separately. Kullback–Leibler, or likelihood, cross-validation is not a viable alternative here (Schuster & Gregory, 1981; Bowman, 1984, 1985). For each of the 50 replications the error or misclassification rate for each rule was then estimated by generating 10 000 new observations from each population and counting the number of misclassifications. In 30 out of 50 of the cases the rule based on method (i) produced a lower estimated error-rate than the rule based on method (ii). The average error rates were 38·99% and 39·75%, with standard errors 0·18 and 0·29.

The second problem used trivariate data on skull measurements, respectively basilar length, occipotonasal length and palatilar length, of the kangaroo species *M. giganteus*. See Table 53.1 of Andrews & Herzberg (1985). There were 25 males and 25 females. Each component of each data triple was rescaled so that it had unit sample standard deviation, and we analysed the data by taking each component of h_X to be identical and each component of h_Y to be identical. To test the 'joint smoothing' method against 'individual smoothing' we made 50 passes of 50 data triples, leaving out one observation each time. We discriminated the omitted observation using the remaining 49 triples. The methods performed similarly, with joint smoothing misclassifying 18 kangaroos out of 50 and individual smoothing misclassifying 19.

ACKNOWLEDGEMENT

The referee's detailed suggestions have led to many improvements in presentation.

REFERENCES

AITCHISON, J. & AITKEN, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413–20.

ANDERSON, J. A., WHALEY, K., WILLIAMSON, J. & BUCHANAN, W. W. (1972). A statistical aid to the diagnosis of keratoconjunctivitis sicca. *Quart. J. Med.* **41**, 175–89.

ANDREWS, D. F. & HERZBERG, A. M. (1985). *Data.* New York: Springer.

BOWMAN, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353–60.

BOWMAN, A. W. (1985). A comparative study of some kernel-based non-parametric density estimators. *J. Statist. Comput. Simul.* **21**, 313–27.

BOWMAN, A. W., HALL, P. & TITTERINGTON, D. M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika* **71**, 341–51.

BROWN, P. J. & RUNDELL, P. W. K. (1985). Kernel estimates for categorical data. *Technometrics* **27**, 293–9.

BUTLER, W. J. & KRONMAL, R. A. (1985). Discrimination with polychotomous predictor variables using orthogonal functions. *J. Am. Statist. Assoc.* **80**, 443–8.

HALL, P. (1981). On nonparametric multivariate binary discrimination. *Biometrika* **68**, 287–94.

HALL, P. (1983). Orthogonal series methods for both qualitative and quantitative data. *Ann. Statist.* **11**, 1004–7.

HALL, P. (1985). Asymptotic theory of minimum integrated square error for multivariate density estimation. In *Proc. of the Sixth International Symposium on Multivariate Analysis*, Ed. P. R. Krishnaiah, pp. 289–309. Amsterdam: North-Holland.

HAND, D. J. (1981). *Discrimination and Classification.* Chichester: Wiley.

HAND, D. J. (1982). *Kernel Discriminant Analysis.* Chichester: Research Studies Press.

HAND, D. J. (1983). A comparison of two methods of discriminant analysis applied to binary data. *Biometrika* **39**, 683–94.

KRZANOWSKI, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics* **36**, 493–9.

KRZANOWSKI, W. J. (1983). Stepwise location model choice in mixed-variable discrimination. *Appl. Statist.* **32**, 260–6.

LAUDER, I. J. (1983). Direct kernel assessment of diagnostic probabilities. *Biometrika* **70**, 251–6.

OTT, J. & KRONMAL, R. A. (1976). Some classification procedures for multivariate binary data using orthogonal functions. *J. Am. Statist. Assoc.* **71**, 391–9.

PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation.* New York: Academic Press.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9**, 65–78.

SCHUSTER, E. F. & GREGORY, G. C. (1981). On the non-consistency of maximum likelihood nonparametric density estimators. In *13th Annual Symposium on the Interface of Computer Science and Statistics*, Ed. W. F. Eddy, pp. 295–8. New York: Springer.

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12**, 1285–97.

TITTERINGTON, D. M. (1978). Contribution to discussion of paper by T. Leonard. *J. R. Statist. Soc.* B **40**, 139–40.

TITTERINGTON, D. M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics* **22**, 259–68.

TITTERINGTON, D. M., MURRAY, G. D., MURRAY, L. S., SPIEGELHALTER, D. J., SKENE, A. M. HABBEMA, J. D. F. & GELPKE, G. J. (1981). Comparison of discrimination techniques applied to a computer data set of head injured patients. *J. R. Statist. Soc.* A **144**, 145–75.

TUTZ, G. (1986). An alternative choice of smoothing for kernel-based estimates in discrete discriminant analysis. *Biometrika* **73**, 405–11.

VLACHONIKOLIS, I. G. & MARRIOTT, F. H. C. (1982). Discrimination with mixed binary and continuous data. *Appl. Statist.* **31**, 23–31.

WANG, M. C. & VAN RYZIN, J. (1981). A class of smooth estimators for discrete distributions. *Biometrika* **68**, 301–9.

[*Received March* 1987. *Revised November* 1987]