International Statistical Review (1989), **57**, 3, pp. 233–247. Printed in Great Britain

E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation

Commentary on Fix and Hodges (1951)

B.W. Silverman and M.C. Jones

School of Mathematical Sciences, University of Bath, Bath, BA2 7AY, U.K.

Summary

In 1951, Evelyn Fix and J.L. Hodges, Jr. wrote a technical report which contained prophetic work on nonparametric discriminant analysis and probability density estimation, and which was never published by the authors. The report introduced several important concepts for the first time. It is of interest not only for historical reasons but also because it contains much material that is still of contemporary relevance. Here, the report is printed in full together with a commentary placing the paper in context and interpreting its ideas in the light of more modern developments.

Key words: Binary regression; Density estimation; History of Statistics; Kernel; Nearest neighbour; Nonparametric discrimination; Smoothing.

1 Introduction

In recent years, nonparametric methods for function estimation generally, and in particular for probability density functions (Silverman, 1986), have become important subjects for research and application. These smoothing methods are of considerable interest both in their own right and as constituent parts of techniques for tackling other problems of which one particularly important example is that of nonparametric discriminant analysis (Hand, 1981). As early as February 1951, however, Evelyn Fix and J.L. Hodges, Jr. introduced a number of ideas which have proved to be basic to the development of nonparametric density estimation and nonparametric discrimination. Fix and Hodges' paper entitled 'Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties'—henceforth F&H—appeared at the time only as a technical report, Report Number 4, Project Number 21-49-004 of the USAF School of Aviation Medicine, Randolph Field, Texas, and was never published in the mainstream statistical literature. This technical report is remarkable in that it anticipated a great deal of later published work. It is of particular interest that this first proposed use of density estimates was not for data presentation—without modern computer graphics such use of density estimates would have been well-nigh impossible—but rather for the problem of discrimination or classification.

The reasons for publishing this paper in this journal nearly forty years after it was written are not just historical. The paper contains much clear and sensible thinking about nonparametric function estimation. If anything this has become more relevant over the intervening years, and it is unfortunate that it has not been more accessible to the general statistical public. Thus it is our hope that the publication of the paper here will make an important contribution to current practice, understanding and research. These introductory notes are intended to assist this object by placing the paper in context and interpreting its ideas in the light of more modern developments.

As F&H's title suggests, the main thrust of their work was directed at the discrimination problem. Theirs is the very first paper to establish the consistency of the approach consisting in nonparametrically estimating the likelihood ratio. Fix & Hodges (1951) are also responsible for introducing the widely used class of nearest neighbour allocation rules for nonparametric discrimination. A perusal of the literature reveals that the importance of these contributions is widely recognised. Much less well realised, however, is the fact that F&H also first introduced two popular methods for nonparametric density estimation: the kernel density estimate and, rather briefly, the nearest neighbour density estimate. Although some authors do recognise F&H's contribution to this latter topic, only in das Gupta (1973) have we found a fairly full appreciation of the breadth of ideas expressed in this paper.

Apart from F&H itself, we shall give a few other early references, but to avoid launching into a complete review of what has become an enormous area of interest, citations of recent work are limited to appropriate review books and articles. Fix and Hodges followed up the work presented here with a second report (Fix & Hodges, 1952) which we shall call F&H2. This was an early attempt at exploring the small sample performance of the nearest neighbour discrimination rules proposed in F&H, and is briefly summarised in the final section of our commentary. Both reports were printed in Agrawala (1977).

2 The Motivation of the Paper: Nonparametric Discrimination

Fix & Hodges' (1951) § 1 is a remarkably lucid introduction to various possible approaches to the discriminant analysis problem. Suppose we wish to discriminate between two populations using training samples $X_1, X_2, \ldots, X_m$ from distribution F (with density f) and $Y_1, Y_2, \ldots, Y_n$ from distribution G (density g). If F and G were known, best allocation of a new sample point at z depends only on the likelihood ratio

$$l(z) = f(z)/g(z). \tag{1}$$

If F and G are unknown, it is natural to plug estimates $\hat{f}$ and $\hat{g}$ of f and g into (1) to yield an empirical version of the likelihood ratio and to proceed as if this estimate of l were the true l. Fix & Hodges (1951) point out that Fisher's (1936) linear discriminant is equivalent to a procedure of this form, with $\hat{f}$ and $\hat{g}$ being chosen as estimates of f and g within a parametric family; of course, Fisher's original motivation was quite different. They then make the visionary step of pointing out the need for *nonparametric* procedures, where it is no longer assumed that f and g fall in any particular parametric family, but only that f and g satisfy certain smoothness assumptions. This need leads at once to nonparametric density estimation, as discussed below.

Sections 2 and 3 of the paper are concerned with defining notions of consistency for discriminant procedures and exploring the consistency of parametric methods; in the first part of § 4 these results are extended to the nonparametric case. This results in a full discussion of the consistency properties of plug-in versions of the likelihood ratio discriminant rule. In both cases, consistency of the discrimination procedures is obtained essentially as a consequence of consistency properties of estimates of f and g. Fix & Hodges (1951) is one of the first places where this is recognised in the parametric situation (Theorem 1 of § 3) although slightly earlier work of Hoel & Peterson (1949, not cited in

F&H) is also relevant. Theorem 2 of § 4, where similar requirements in the nonparametric case are spelled out, appears to be a seminal result in the nonparametric context. For more recent developments on asymptotic properties of this type of allocation rule, see Prakasa Rao (1983, Ch. 8). It is interesting to note that F&H consider general sufficient conditions for consistency of nonparametric discriminant procedures before introducing any specific methods of nonparametric density estimation.

3 Nonparametric Density Estimation

The paragraph following Theorem 2 of the paper gives the definition of, and the reasoning behind, a general version of what is now often called the naive kernel density estimate. Consider the problem of estimating an unknown p-dimensional density f at a point z, given a random sample $X_1, X_2, \ldots, X_m$ from F. The estimator suggested by F&H is of the form

$$\hat{f}(z) = m^{-1} \sum_{i=1}^{m} K_m(X_i - z), \tag{2}$$

where the kernel K_m is the uniform probability density over a given neighbourhood Δ_m^0 of zero. Note that, because they are considering the point z as fixed, F&H work in terms of the corresponding neighbourhood $\Delta_m = z + \Delta_m^0$ of z. They prove consistency of $\hat{f}$ under the conditions that f is continuous at z, that the Lebesgue measure of Δ_m^0 tends to zero more slowly than m^{-1} as $m \to \infty$, and that $\sup\{\|x\| : x \in \Delta_m^0\} \to 0$ as $m \to \infty$. If, for example, Δ_m^0 is taken to be the hypercube $[-h_m, h_m]^p$, then (2) becomes the kernel estimate with bandwidth h_m and kernel equal to the uniform density over $[-1, 1]^p$; the conditions for consistency then become the familiar ones, $h_m \to 0$ and $mh_m^p \to \infty$ as $m \to \infty$.

Since $\hat{f}$ inherits the smoothness properties of K, it is now more common for exploratory and presentational purposes to use a smooth kernel in (2). When plotted out the naive estimator has a locally irregular appearance; see Silverman (1986; Fig. 2.3). However, as far as behaviour at individual points is concerned, little is lost by using the uniform kernel rather than other positive kernels, and F&H's estimator could be found just by counting the number of points lying in the neighbourhood Δ_m of z; to compute a more general kernel estimate without more modern computing facilities would have been a daunting task.

Fix & Hodges (1951) recognise that the more critical choice in (2) is that of an appropriate value for the degree of smoothing applied to the data. Early in F&H's § 5 there is an intuitive discussion of what can be interpreted as the tradeoff between variance and bias governed by a smoothing parameter. The hard problem of selecting how much to smooth in practice to 'steer a middle course' remains the subject of much research (e.g. Marron (1989)).

The first published paper on kernel density estimation is Rosenblatt (1956), which describes properties of the naive estimator and recognises the potential of more general kernels in the univariate case ($p = 1$). Indeed, there has been an emphasis on the univariate situation in the literature ever since. The first look at the multivariate case appears to be taken by Cacoullos (1966). The popularity of the kernel density estimate in the more recent literature is due to its easy-to-understand explicit definition, its theoretical tractability and its good practical performance. When the basic kernel estimate proves inadequate, more sophisticated competitors include extensions of the same idea. For a recent introduction to the theory and application of kernel density estimates, see Silverman (1986).

Immediately after discussing the compromise required in selecting h_m for the naive kernel estimator, F&H briefly mention the complementary approach of nearest neighbour density estimation. Rather than fixing Δ_m and counting the number, M, of data points falling in Δ_m, fix the value of M and find Δ_m just big enough to include the M nearest points to z. The first mention of this approach in the published literature is by Loftsgaarden & Quesenberry (1965). See Silverman (1986, §§ 2.5, 5.2) for further discussion. Fix & Hodges (1951) do not consider this nearest neighbour estimate in any detail, but pass on quickly to a two-sample approach more directly tailored to the discriminant analysis problem.

It has already been pointed out that F&H state an early version of the conditions for pointwise consistency of the kernel estimate given, for example, by Parzen (1962). Theoretical aspects of nonparametric density estimation have been the source of much subsequent research. Most results concern consistency and, to distinguish more finely between competing prescriptions, rates of convergence to the true density. For a lead into the vast theoretical literature, see Prakasa Rao (1983) and Devroye & Györfi (1985).

4 Specific Methods for Nonparametric Discrimination

In their final section, F&H make two specific suggestions for nonparametric discrimination using density estimates. The first of these is to construct, independently of one another, consistent kernel estimates of f and g of the form (2), and then to plug these into (1). This approach, now called kernel discriminant analysis, has attracted a great deal of interest, much of it from a practical viewpoint. See, for example, the recent monographs of Hand (1982) and Coomans & Broeckaert (1986) devoted entirely to this topic. Modern developments mirror those in kernel density estimation itself; the use of variable kernel estimates of f and g, wherein the single smoothing parameter h_m is replaced by a set of different parameters, one for each datapoint, and the automatic selection of appropriate values for smoothing parameters are two of the major problems of interest.

The second specific suggestion is based on a variant of nearest neighbour density estimation. The idea is to use an estimate of the form (2) for f, choosing Δ_m to be just large enough to contain a given number, k, of points in the *combined* sample. The same neighbourhood, now called $\Delta_{m,n}$, is used to construct an estimate of g. Consistency of such estimates is proved in Theorem 4 of F&H. Though this method is motivated via density estimation, it also has connections with other ideas in the discriminant analysis literature. The last equation in Fix & Hodges' paper shows that, if $k = 1$, their method is precisely the rule assigning a new point to the class of its nearest neighbour in the training set; Cover & Hart's (1967) famous bound on its classification error rate was instrumental in popularising this method. For $k > 1$, their approach incorporates a variety of so-called k-NN discrimination rules including simple majority vote amongst the classes of a point's k nearest neighbours and modifications which take differing sample sizes into account. For a review of nearest neighbour discriminant analysis, including many of the leading modern variants on the basic theme, see, for example, Devijver & Kittler (1982, Ch. 3).

There is also a close implicit link with nonparametric binary regression estimation. Combine the X and Y samples into a single sample $Z_1, Z_2, \ldots, Z_{m+n}$ and attach to each an indicator function of the form

$$\delta_i = \begin{cases} 1 & \text{if } Z_i \text{ is one of the } X\text{'s}, \\ 0 & \text{if } Z_i \text{ is one of the } Y\text{'s}. \end{cases}$$

Write $r(z) = \text{Prob}\,(\delta = 1 \mid Z = z)$ for the regression of δ on Z. Then it is easy to see that

$$r(z)/\{1 - r(z)\} = mf(z)/\{ng(z)\}. \tag{3}$$

Now r can be estimated in analogous ways to f and g. A nearest neighbour regression estimate might be written

$$\hat{r}(z) = \sum_{j=1}^{m+n} \delta_i K\{d_k^{-1}(z)(z - Z_j)\} \Big/ \sum_{j=1}^{m+n} K\{d_k^{-1}(z)(z - Z_j)\}, \tag{4}$$

where $d_k(z)$ is the distance from z to the kth nearest Z_j. Replacing r by $\hat{r}$ in (3) yields an estimate of $l = f/g$; when K is the uniform density in (4), this results precisely in the k-NN allocation rule. Such a binary regression approach to two sample discrimination has the advantage that a single estimate only, that of r, need be evaluated rather than two separate estimates, those of f and g, whose ratio is then taken in the density estimation approach. Finally, we note that there is no distinction between the two approaches when kernel density/regression estimates are used, provided the same smoothing parameter is used in estimating both f and g.

5 The Second Report

Fix & Hodges' (1951) investigation of the small sample performance of nonparametric discriminant analysis reported in F&H2 was attempted before powerful computational resources were readily available. Discrimination procedures are naturally assessed by considering probabilities of misclassification but the theoretical intractability of such quantities in most situations means that F&H2's investigation is necessarily limited. Indeed, interest is confined to the case of equal sized samples from two p-variate normal distributions with equal covariance matrices. This is, of course, the situation in which the linear discriminant approach is appropriate. Fix & Hodges (1952) consider the k-NN nonparametric method, mostly for $k = 1$, and assesses its performance in this situation. Unfortunately, only for $p = 1$ was it possible to determine misclassification probabilities for the linear discriminant rule and thus to compare with the 1-NN rule whose misclassification probabilities proved a little easier to handle (even the latter involved numerical integration and/or approximation steps). Tables and graphs of performance are given for a variety of mean separations and sample sizes. Brief consideration is given in later sections of F&H2 to nearest neighbour rules in $p > 1$ dimensions and to extensions to k-NN rules for odd $k > 1$ and alternative distance functions.

More recent small sample investigations of nonparametric discriminators suggest that their performance does not fall far short of that of the parametric (linear) rule when the assumptions on which the latter is based are met and that they outperform linear discriminant analysis in other situations. For detailed references see Silverman (1986, § 6.1.3). Fix & Hodges (1952, § 4) quantifies the former point precisely in one particular case; F&H2 did not have the opportunity to investigate the latter. Rather, F&H2 dwells on the point that selecting the better of the two approaches depends on the 'judgment[al]' assessment of whether or not the parametric assumptions hold true.

Acknowledgements

We are most grateful to J.L. Hodges for giving us his kind permission to prepare the paper for publication; to E.L. Lehmann for his invaluable assistance; and for support from the U.S. Army European Research Office and the Science and Engineering Research Council. Any opinions or intentions imputed in this commentary to the authors of the paper are based entirely on the evidence of the paper itself.

References

Agrawala, A.K. (Ed). (1977). *Machine Recognition of Patterns*. New York: IEEE Press.
Cacoullos, T. (1966). Estimation of a multivariate density. *Ann. Inst. Statist. Math.* **18,** 179–189.
Coomans, D. & Broeckaert, I. (1986). *Potential Pattern Recognition in Chemical and Medical Decision Making*. Letchworth: Research Studies Press.
Cover, T.M. & Hart, P.E. (1967). Nearest neighbour pattern classification. *IEEE Trans. Info. Theory* **IT-13,** 21–27.
das Gupta, S. (1973). Theories and methods in classification: a review. In *Discriminant Analysis and Applications,* Ed. T. Cacoullos, pp. 77–137. New York: Academic Press.
Devijver, P.A. & Kittler, J. (1982). *Pattern Recognition: a Statistical Approach*. London: Prentice-Hall.
Devroye, L. & Györfi, L. (1985). *Nonparametric Density Estimation: the L_1 View*. New York: Wiley.
Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7,** 179–188.
Fix, E. & Hodges, J.L. (1951). Discriminatory analysis. Nonparametric discrimination; consistency properties. Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas. (Reprinted as pp 261–279 of Agrawala, 1977).
Fix, E. & Hodges, J.L. (1952). Discriminatory analysis. Nonparametric discrimination: small sample performance. Report Number 11, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas. (Reprinted as pp 280–322 of Agrawala, 1977).
Hand, D.J. (1981). *Discrimination and Classification*. Chichester: Wiley.
Hand, D.J. (1982). *Kernel Discriminant Analysis*. Chichester: Research Studies Press.
Hoel, P.G. & Peterson, R.P. (1949). A solution to the problem of optimum classification. *Ann. Math. Statist.* **20,** 433–438.
Loftsgaarden, D.O. & Quesenberry, C.P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.* **36,** 1049–1051.
Marron, J.S. (1989). Automatic smoothing parameter selection: a survey. *Empirical Econ.* **13,** 187–208.
Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33,** 1065–1076.
Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*. Orlando, Florida: Academic Press.
Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27,** 832–837.
Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Résumé

En 1951, E. Fix et J.L. Hodges, Jr. ont écrit un rapport technique prophétique sur l'analyse non-paramétrique de discrimination et l'estimation de la densité de probabilité, mais celui-ci ne fut jamais publié par ses auteurs. Ce rapport introduit plusieurs idées nouvelles et importantes. Il nous intéresse non seulement pour des raisons historiques, mais aussi parce qu'il contient des concepts qui sont encore importants de nos jours. Nous le publions ici en entier, accompagné d'un commentaire qui l'interprète d'un point de vue plus moderne.

[*Received June* 1988, *accepted May* 1989]

Discriminatory Analysis—Nonparametric Discrimination: Consistency Properties

Evelyn Fix† and J.L. Hodges, Jr.

University of California, Berkeley

This paper originally appeared as Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, in February 1951. Enquiries should be addressed to the authors of the accompanying commentary, B.W. Silverman and M.C.

† Evelyn Fix died on 30 December 1965.