SMOOTHING AND ORDERING
IN
DISCRIMINANT ANALYSIS

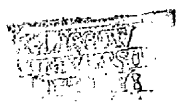
by

ALISON JANE GRAY

A dissertation submitted to the
UNIVERSITY OF GLASGOW
for the degree of
DOCTOR OF PHILOSOPHY,

Department of Statistics, 1990.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346



CONTENTS

Continued.

Continued.

CONTENTS cont'd.

ACKNOWLEDGEMENTS

I wish to thank my supervisor, Dr. G.D. Murray for his considerable help, guidance and encouragement, and Dr. Murray and Professor D.M. Titterington both for their patience and for arranging funding for my final 9 months of study. I would also like to thank Dr. A.W. Bowman for some very helpful discussions. Funding for the first 27 months was provided by the Science and Engineering Research Council.

Finally, thanks are also due to my husband and family for their patience and moral support.

Part of the material in Section 3.5.2 has been published, as indicated in the text, as Murray and Wilson (1986) and (1987).

ADDENDUM

The following paragraph should be inserted immediately before the paragraph beginning 'Table 2.7 shows the Brier score' on p. 152:-

'For consistency, we have rescaled the Brier score as $1-\frac{1}{2}S_B$ where S_B is given by (1.13), so that the rescaled score ranges from 0 to 1, with high values indicating good performance as for the log and modified log scores.'

ERRATA

Chapter 1

p.15 line 8 up. For 'qualitative ordinal discrete variables' read 'quantitative ordinal discrete variables'.

p.27 line 9. For '$\sum_{\underline{x}} \phi_r(\underline{x})\, f(\underline{x})$' read '$\sum_{\underline{x}} \phi_j(\underline{x})\, f(\underline{x})$'.

p.43 line 19. For 'n-V cases' read 'n-s_v cases, where s_v is the size of the vth subset.

Chapter 2

p.151 line 11. For 'most notably relative to method 2' read 'most notably method 2'.

Chapter 3

p.176 Formula (3.7) should read

$$K(\underline{x}|\underline{X}, \underline{\lambda}) = \prod_{j=1}^{d} \lambda_j^{\,1-|x_j-(\underline{X})_j|}\, (1-\lambda_j)^{\,|x_j-(\underline{X})_j|}.$$

p.184 line 6. For '$\prod_{j=1}^{n} \hat{p}(\underline{X}_j)$' read '$\prod_{i=1}^{n} \hat{p}(\underline{X}_i)$'.

p.193 line 4 up. For '$S^{-1} = \text{diag}\ (w_1^{-1}, w_2^{-1}, \ldots\ldots, w_n^{-1})$' read '$S = \text{diag}\ (w_1^{-1}, w_2^{-1}, \ldots\ldots, w_n^{-1})$'.

Chapter 4

p.199 line 3 up. For '$\underline{\lambda}$' read 'λ'.

p.228 line 6. For 'Once more MV XVAL is superior' read 'Once more STD XVAL is superior'.

Chapter 5

p.282 line 8 up. For 'most notably relative to method 2' read 'most notably method 2'.

p.284 line 15 up. For '$g(x) \equiv f_1(\underline{x})\theta_1 - f_2(\underline{x})\theta_2$' read '$g(\underline{x}) \equiv f_1(\underline{x})\theta_1 - f_2(\underline{x})\theta_2$'.

SUMMARY

This thesis addresses the question of how to achieve reliable estimation of the posterior probability function in discriminant analysis, both for continuous and ordered discrete feature variables. In the latter instance we are also concerned with the estimation of a posterior, which, regarded as a function of the feature variables, is ordered with respect to one or more independent variable.

Chapter 1 introduces the discrimination problem, establishes notation and describes the possible approaches. Methods of density estimation, for use in discriminant analysis, are described, including the kernel method, as are some more direct approaches to discrimination and classification. Some comparative studies and their conclusions are reviewed. Means of assessing the performance of a discriminant rule are described with emphasis on measures of reliability rather than separation. The final section mentions briefly the important problem of variable selection, although this is not addressed elsewhere in the thesis.

Chapter 2 addresses the problem of choosing smoothing parameters in kernel density estimation with continuous variables when this is to be used in the discrimination context. It is natural to suspect that the optimal degree of smoothing for marginal density estimates may not be that which will produce an optimal density ratio or posterior probability function when two such estimates are combined.

A simulation study confirms that some popular methods for choosing the smoothing parameter can produce an estimated density ratio which is poor in terms of mean square error. Some alternatives are proposed based on direct assessment measures of reliability, not of the marginal estimates but of the predicted probabilities. These are compared to the marginal approaches. To a more limited extent, the optimal (minimum mean square error) kernel method is compared to an optimal spline estimate of the density ratio.

Both the marginal and direct methods are then applied to a real data set and the resulting estimates compared with a spline estimate.

Chapter 3 discusses ordered variables, from qualitative orderings to grouped continuous variables, ways in which ordering

can affect a data set and suitable models in each case. Particular emphasis is given to discrete kernel estimators and isotonic regression techniques. Some problems in applying existing algorithms for the latter are described and suggestions made for overcoming these.

Chapter 4 applies ordered kernels and isotonic regression to 1- and 2-dimensional problems using the data of Titterington et al. (1981), concluding that the kernel methods are unable to recover the type of ordering manifested by the data and that a diagnostic approach is required. The results are compared in the univariate case to those in Chapter 2, Section 2.6 which used continuous kernels. The use of isotonic regression is then compared with 2 logistic models and an independence model using the same data set but with 3 variables. Suggestions are made for further smoothing of the isotonic estimator, 2 of which are implemented.

Finally, Chapter 5 draws some conclusions and makes suggestions for further work. In particular, isotonic splines may be worthy of investigation.

CHAPTER 1 DISCRIMINANT ANALYSIS

1.1 INTRODUCTION

Discriminant analysis is a term applied to a set of statistical techniques used to allocate an object to one of a number of disjoint categories on the basis of data measurable on any such object. It is assumed that an effective means of allocation does exist, but for reasons of cost, time or convenience, for instance, that we wish to develop an alternative method. An obvious application would be in medical diagnosis, or, less commonly, prognosis, where a patient is allocated to a particular disease category on the basis of signs, symptoms and the results of laboratory tests. However, the potential applications are much more numerous than this and examples occur in almost any field of research e.g. classification of archaeological discoveries to a particular culture or era, species identification in botany or anthropology, problems of disputed authorship, and image processing (see, for instance, Ripley and Taylor, 1987).

Classic references are Anderson (1958, Chapter 6), and Lachenbruch (1975a), with Hand (1981a) providing a more recent and more general treatment. A recent collection of papers on advances in the theory and applications can be found in Choi (1986).

1.2 NOTATION

No matter what the practical application, the basic elements of the problem are formally identifiable as the following :-

Individuals/objects/cases are assumed to belong to one of k disjoint, exhaustive populations $\pi_1, \ldots, \pi_k$.

These populations have associated prior probabilities or incidence rates $p(\pi_i) = \theta_i$, $i = 1, \ldots, k$ such that $\Sigma\theta_i = 1$.

Feature variables or indicants with realisations $x_1, \ldots, x_d$ are measurable, giving a feature vector $\underline{x}$ for each case. The distribution of $\underline{x}$ within the ith class $p(\underline{x}|\pi_i)$ is denoted by $f_i(\underline{x})$, $\{f_i(\underline{x}),\ i = 1, \ldots, k\}$ being known as the class conditional densities.

A set of design or training cases is available, from past experience, whose feature vectors are known and populations of origin have been confirmed. The training set will be used to derive a discriminant or classification rule with which to classify a future case with feature vector $\underline{x}$. Often a further set of

confirmed cases is available, known as a test set, with which to assess the discriminant rule, by means of performance measures or scores. (A test set is not strictly necessary, and means of assessment are discussed further in Section 1.6)

1.3 APPROACHES

1.3.1 Introduction

The aim of discriminant analysis is to assign future cases or objects of unknown class, on the basis of their feature vectors, to one of $\pi_1, .., \pi_k$, usually in order to minimise the error rate, or, more generally, a loss function. In an automated setting, such as industrial quality control procedures, an allocation rule may be all that is required. However in more complex situations, such as arise in a medical context, other factors (e.g. expert knowledge) will enter into the decision making process and it may be more useful just to provide the estimated probabilities, $\{\hat{p}(\pi_i|\underline{x}),\ i = 1,\ldots,\ k\}$, which may then be used in conjunction with other sources of diagnostic information. Equivalently, the estimated posterior odds ratios

$$\left\{\frac{\hat{p}(\pi_i|\underline{x})}{\hat{p}(\pi_j|\underline{x})}\right\}$$

may be specified, either as point or interval estimates. For recent work on confidence intervals for the log-odds ratio see, for instance, Rigby (1982), Critchley and Ford (1985) and Davis (1987).

1.3.2 Classification

The classification procedure involves partitioning the sample space of the feature variables into decision regions $R_1,\ldots,\ R_k$ such that a case with feature vector $\underline{x}$ is allocated to π_i if $\underline{x} \in R_i$. Welch (1939) suggested choosing the partition in order to minimise the probability of misclassification. More generally if cost $C(j|i)$ can be associated with misallocation of an observation arising from π_i to π_j, the class of Bayes' procedures are those which minimise expected loss,

$$\sum_{i=1}^{k} \theta_i \sum_{\substack{j=1 \\ j\neq i}}^{k} C(j|i)\ P(j|i,R)$$

where $P(j|i,R) = \int_{R_j} f_i(\underline{x})\, d\underline{x}$.

It can be shown (e.g. Anderson, 1958, Chapter 6) that this is achieved by the partition $(R_1, \ldots, R_k)$ such that

$$R_j = \left\{ \underline{x} : \sum_{\substack{i=1 \\ i\neq j}}^{k} \theta_i f_i(\underline{x}) C(j|i) < \sum_{\substack{i=1 \\ i\neq l}}^{k} \theta_i f_i(\underline{x}) C(l|i),\ l = 1,\ldots,k;\ l \neq j \right\}$$

where $\sum_{\substack{i=1 \\ i\neq j}}^{k} \theta_i f_i(\underline{x}) C(j|i)$ is proportional to the expected loss associated with assigning an object to π_j.

For equal costs i.e. to minimise probability of misclassification

$$R_j = \{\underline{x} : \theta_l f_l(\underline{x}) < \theta_j f_j(\underline{x}),\ \forall l \neq j\}$$

so that allocation is on the basis of the highest posterior probability. The expression $\theta_j f_j(\underline{x})$ is called the jth discriminant score.

When k = 2, the 2 population case, to minimise expected loss

$$C(2|1)\, p(2|1,R)\, \theta_1 + C(1|2)\, p(1|2,R)\, \theta_2$$

the optimal partition becomes

$$R_1 = \left\{ \underline{x} : \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \frac{\theta_2}{\theta_1} \cdot \frac{C(1|2)}{C(2|1)} \right\},\ R_2 = \bar{R}_1,$$

or, for equal costs,

$$R_1 = \left\{ \underline{x} : \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \frac{\theta_2}{\theta_1} \right\},$$ (Welch, 1939; Hoel and Peterson, 1949),

which minimises the probability of misclassification

$$\theta_1 \int_{R_2} f_1(\underline{x})\, d\underline{x} + \theta_2 \int_{R_1} f_2(\underline{x})\, d\underline{x}.$$

Where priors $\{\theta_i\}$ are not known, the decision rule will be based on the likelihood ratio, with

$R_1 = \left\{ \underline{x} : \dfrac{f_1(\underline{x})}{f_2(\underline{x})} > c \right\}$, where c would be chosen to minimise conditional expected loss given the population of origin,

$C(2|1)\ p(2|1, R) = r(1, R)$ (risk) if $\underline{x} \in \pi_1$,

$C(1|2)\ p(1|2, R) = r(2, R)$ if $\underline{x} \in \pi_2$.

In this case Anderson (1958, Chapter 6) shows that the Bayes' rule generates the class of admissible procedures. (In practice, in the absence of known priors, they are commonly assumed equal, or, if the training data has been sampled from a mixture, estimated from the sample proportions, n_i/n, the maximum likelihood estimates (MLEs)).

In practice, eliciting costs, or even cost ratios, can be difficult, though Anderson (1969) proposed a method of optimal allocation subject to constraints imposing upper bounds on the probabilities of misallocation, which to a certain extent avoids the need to do so. However, costs are usually assumed to be equal.

1.3.3 Estimation

Modelling effort may be focussed directly on the $\{p(\pi_i|\underline{x})\}$ or indirectly, on the class conditional densities, reformulating

$$p(\pi_i|\underline{x}) \text{ as } \frac{f_i(\underline{x})\ \theta_i}{\sum_{j=1}^{k} f_j(\underline{x})\ \theta_j},\ i = 1, .., k, \qquad (1.1)$$

and hence

$$\frac{p(\pi_i|\underline{x})}{p(\pi_k|\underline{x})} = \frac{f_i(\underline{x})\ \theta_i}{f_k(\underline{x})\ \theta_k}, \text{ via Bayes' Theorem.}$$

Where the $\{\theta_i\}$ are unknown, the sample relative frequencies are used as plug-in estimates. In either case, a parametric or nonparametric approach may be taken. Dawid (1976) termed the direct and indirect approaches the "diagnostic" and "sampling" paradigms respectively, and advocated use of the diagnostic method in that it is less sensitive to changes in the target populations. However the sampling method has in its favour a much richer class of models. The sampling and diagnostic approaches are discussed in the next two sections, on density estimation and odds ratio

estimation respectively.

1.4 DENSITY ESTIMATION

1.4.1 Introduction

Using the sampling method one estimates $\{p(\pi_i|\underline{x})\}$ indirectly by modelling the class conditional densities $\{f_i(\underline{x})\}$, making available the wide class of density estimation methods, both parametric and nonparametric. Nonparametric models, discussed in Section 1.4.3, rely on weaker assumptions than the former, such as smoothness properties (see the kernel and orthogonal series methods) or unimodality (maximum likelihood estimators). The data are allowed to "speak for themselves" to a greater extent than by imposing a particular parametric model upon them. Certain nonparametric methods extend readily to higher dimensions, while multivariate parametric models are less abundant, especially for continuous data. The type of parametric model which is appropriate also depends more upon the type of feature variables, which in some fields, e.g. in medical applications, tend increasingly to be discrete rather than continuous. Discrete variables may be qualitative, or ordered, in some sense reflecting an underlying continuous scale whether or not this has been measured (as in grouped continuous data) or is not directly measurable (as, for instance, in specifying degree of pain). It is important with respect to model choice to distinguish between genuinely discrete and ordered variables, and this is discussed further in Chapters 3 and 4.

1.4.2 Parametric estimators

Continuous variables

For continuous feature variables the most common practice has been to assume that their joint distribution within each population is Multivariate Normal. For 2 populations with $f_i(\underline{x}) \sim N(\underline{\mu}_i, \Sigma_i)$, $i = 1,2$, assuming equal costs and equality of covariance matrices, $\Sigma_1 = \Sigma_2 = \Sigma$, the Bayes' procedure gives the decision regions

$$R_1 = \left\{ \underline{x} : \underline{x}^T\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) - \tfrac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)^T\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \geq \log \frac{\Theta_2}{\Theta_1} \right\}, \quad (1.2)$$

$R_2 = \bar{R}_1$, (Welch, 1939), while for k regions we have

$$R_j = \left\{ \underline{x} : u_{jm}(\underline{x}) \geq \log \frac{\theta_m}{\theta_j},\ m = 1,\ldots,k;\ j \neq m \right\}, \text{ where}$$

$$u_{jm} = \left[\left[\underline{x} - \frac{1}{2}(\underline{\mu}_j + \underline{\mu}_m) \right]^T \Sigma^{-1} (\underline{\mu}_j - \underline{\mu}_m) \right].$$

If priors are assumed equal (1.2) becomes

$$R_1 = \left\{ \underline{x} : \underline{x}^T \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2) \geq \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)^T \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2) \right\}, \quad R_2 = \bar{R}_1. \tag{1.3}$$

In practice the parameters ($\underline{\mu}_1$, $\underline{\mu}_2$, Σ) are not usually known and are most commonly replaced by their sample equivalents

$$\hat{\underline{\mu}}_i = \bar{\underline{x}}_i,\ i=1,2 \text{ where } \bar{\underline{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \underline{x}_{ij} \text{ and}$$

$$\hat{\Sigma} = S = \frac{1}{n_1+n_2-2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\underline{x}_{ij} - \bar{\underline{x}}_i)(\underline{x}_{ij} - \bar{\underline{x}}_i)^T,$$

the pooled sample covariance matrix, assuming we have samples $(\underline{x}_{11}, \ldots, \underline{x}_{1n_1})$ and $(\underline{x}_{21}, \ldots, \underline{x}_{2n_2})$ from populations π_1 and π_2 respectively. This yields from (1.3)

$$R_1 = \left\{ \underline{x} : \left[\underline{x} - \frac{1}{2}(\bar{\underline{x}}_1 + \bar{\underline{x}}_2) \right]^T S^{-1} (\bar{\underline{x}}_1 - \bar{\underline{x}}_2) > c \right\}, \tag{1.4}$$

c a constant, where the term on the left-hand side is the Wald-Anderson classification statistic after Wald (1944) and Anderson (1951). For c = 0 (equal priors and equal costs) this is the procedure of Fisher (1936) and $\underline{x}^T S^{-1} (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$ is Fisher's Linear Discriminant Function (LDF) which he showed to be the linear combination, $\underline{b}^T \underline{x}$, of the feature variables which maximises the ratio of between-sample separation $(\underline{b}^T \underline{\mu}_1 - \underline{b}^T \underline{\mu}_2)^2$ to within-sample variance $\underline{b}^T \Sigma \underline{b}$. (1.4) is referred to as the (sample) linear discriminant rule.

If the assumption that $\Sigma_1 = \Sigma_2$ is relaxed the optimal procedure is the Quadratic Discrimination Function rule (QDF) (Smith, 1947),

$$R_1 = \left\{ \underline{x} : \underline{x}^T(\Sigma_1^{-1} - \Sigma_2^{-1})\underline{x} + 2(\underline{\mu}_2^T\Sigma_2^{-1} - \underline{\mu}_1^T\Sigma_1^{-1})\underline{x} \right.$$

$$\left. - \underline{\mu}_2^T\Sigma_2^{-1}\underline{\mu}_2 + \underline{\mu}_1\Sigma_1^{-1}\underline{\mu}_1 - \log \frac{|\Sigma_2|}{|\Sigma_1|} \leq -2 \ln \frac{\theta_2}{\theta_1} \right\},$$

although best linear discriminants have been proposed for use in this context (Clunies-Ross and Riffenburgh, 1960; Anderson and Bahadur, 1962). The QDF will be appropriate for separation of 2 populations with a common mean which will therefore be dependent on exploiting differences in covariance structure (Bartlett and Please, 1963) though Lachenbruch (1975b) used the LDF on the absolute value of deviations from the mean.

Due to sampling variability neither sample-based rule will be Bayes' optimal even when the distributional assumptions are satisfied i.e. they will not minimise expected loss, although results of Glick (1972) show that they are consistent in the sense that expected loss tends to that of the optimal rule.

Rather than replace unknown parameters, $\underline{\Phi}$ say, in the likelihood ratio by their estimates, the "predictive" method of Geisser (1964) and Aitchison and Dunsmore (1975) assigns to the parameters a prior probability density $p(\underline{\Phi})$, based on ignorance, and instead of $r(\underline{x}|\pi_i, \underline{z}) = p(\underline{x}|\pi_i, \hat{\Phi}(\underline{z}))$ where $\underline{z}$ is the available data, uses the "predictive" density function

$$q(\underline{x}|\pi_i, \underline{z}) = \int_{\underline{\Phi}} p(\underline{x}|\pi_i, \underline{\Phi})\, p(\underline{\Phi}|\underline{z}, \pi_i)\, d\underline{\Phi} \text{ , where}$$

$$p(\underline{\Phi}|\underline{z}, \pi_i) \propto p(\underline{z}|\underline{\Phi}, \pi_i)\, p(\underline{\Phi})$$

and hence averages out the effect of sampling variability.

Assuming multivariate normality, if a vague prior (Aitchison and Dunsmore, 1975, p. 21) is used, while $r(\underline{x}|\pi_i, \underline{z}) \sim N_d(\bar{\underline{x}}_i, S_i)$, $q(\underline{x}|\pi_i, \underline{z}) \sim St_d\left[n_i-1, \bar{\underline{x}}_i, \left[\frac{1+1}{n_i}\right]S_i\right]$, a d-dimensional Student density, both centred on the same mean vector but the latter being a less concentrated distribution. Asymptotically the two approaches are seen to be identical.

Aitchison, Habbema and Kay (1977) compared the "estimative" and "predictive" approaches in this situation and showed that the ratio q/r could vary enormously, while Aitchison (1975) showed that viewed as a function of $\underline{x}$, in terms of a measure based on the Kullback-Leibler directed divergence (Kullback, 1959, pp. 6-7), overall q estimates the true density more closely than r and therefore the predictive method more reliably estimates the ratio of densities. Murray (1977b) strengthened this result to show that the predictive estimator is also the optimal estimator minimising this distance function which is invariant to translation and non-singular linear transformation of the sample space. Aitchison et al. (1977) carried out a simulation study using various parameter configurations, a range of dimensions and moderate to small sample sizes, to compare the 2 approaches, both for equal and unequal covariance matrices, in terms of mean absolute deviation of assessed log-odds from true log-odds, averaged over simulations. The results indicated that the QDF was by far the worst procedure even when $\Sigma_1 \neq \Sigma_2$, presumably due to the difficulties in reliably estimating so many parameters, and that in each case the appropriate predictive method was superior to either the LDF or alternative predictive method, markedly so when $\Sigma_1 = \Sigma_2$. They also found that the estimative method tends to produce over-extravagant log-odds while the predictive approach was much more conservative. Moran and Murphy (1979) showed that for the case $\Sigma_1 = \Sigma_2$ the mean bias of the estimative log-odds is positive and increases with dimension while that of predictive log-odds is much smaller, negative and does not alter with dimension. Extending the simulations of Aitchison et al. to cover LDF and QDF rules corrected for bias, they found that, while the advantage remained with the predictive methods, the bias-corrected methods were comparable for estimation of log-odds and that there was little difference in terms of percentage of test cases misclassified, whether or not the covariance structures were assumed equal. Murphy and Moran (1986) found the predictive LDF to be clearly superior to the standard and bias-corrected LDFs for estimation of log-odds, and concluded that while the uncorrected procedure should never be used for this purpose, the predictive approach was still better than the corrected LDF, especially as the number of variables increased relative to sample size.

Its widespread availability, simplicity and nonparametric justifications have led to the LDF rule commonly being used even when the assumptions required for its optimality in the likelihood ratio sense are clearly violated.

Gilbert (1969) studied the performance of the LDF relative to the QDF, assuming known parameters and $\Sigma_2 = c\ \Sigma_1$, where c is a constant, and all correlations assumed equal, and also concluded that LDF may often be useful for classification but not for estimation of log-odds. Error rates were similar only for a moderate range of c (not too far from 1) and with some amount of linear separation of the populations. LDF became worse as the dimensionality increased.

Marks and Dunn (1974) considered a slightly wider variety of situations and also considered the best linear discriminant such that $R_1 = \{\underline{x} : \underline{x}^T\underline{b} < \underline{b}^T\underline{\mu}_1 + t_1\ \underline{b}^T\Sigma_1\underline{b}\}$ where $\underline{b} = (t_1\Sigma_1 + t_2\Sigma_2)^{-1}(\underline{\mu}_2 - \underline{\mu}_1)$ and (t_1, t_2) is chosen to minimise the misclassification probability. In each case parameters were estimated. For large samples and Σ_1 quite distinct from Σ_2, QDF was much better in terms of misclassification, provided the dimension was not too large. For similar $\{\Sigma_i\}$ QDF was only slightly better. QDF was increasingly poor as sample size decreased especially as the number of variables increased. Small sample sizes affected the LDF and best linear methods less. Where $\{\Sigma_i\}$ were similar the best linear and LDF methods were similar also and, for bigger differences in covariance, while the best linear method could be much better this was often when QDF was better still.

Lachenbruch, Sneeringer and Revo (1973) studied robustness of LDF and QDF rules to non-normality using 3 non-linear transformations of normal data with independent variables and found that both were seriously affected. The quadratic rule was very poor especially for heavy-tailed distributions while the LDF could be badly misleading in that error rates were distorted, one being higher than optimal and the other lower. Their sum also increased for some distributions. Consequently transformation to approximate normality was recommended before using the LDF.

In short, for estimation purposes, estimative methods may be erratic and an appropriate predictive method should be used, or else the unbiased LDF or QDF of Moran and Murphy (1979). For

allocation purposes the estimative, predictive and unbiased methods are equivalent for equal-sized samples. For unequal sample sizes, provided $\{\Sigma_i\}$ are reasonably similar, the LDF may be useful in low dimensions. If $\{\Sigma_i\}$ are quite distinct, a QDF is to be preferred unless samples are small relative to dimension, in which case a best linear method should be used. In any event, prior transformation to approximate normality is advisable. In practice it may be difficult to find blanket transformations to transform each feature variable to approximate normality in all populations simultaneously, and hence nonparametric rules may be preferred.

Discrete variables

The most general model that can be assumed for discrete explanatory variables is that their joint distribution is Multinomial. This makes very few assumptions about the form of the distribution. In practice the maximum likelihood estimates of the cell probabilities $\{p(\underline{x}|\pi_i)\}$, the observed sample relative frequencies $n_i(\underline{x})/n_i$, would be used as plug-in estimates, where $n_i(\underline{x})$ is the number of observations arising from the ith population with the feature vector $\underline{x}$ and

$$n_i = \sum_{\underline{x}} n_i(\underline{x}).$$

Unfortunately as the number of variables, d, increases, the number of cells $\prod_{j=1}^{d} s_j$ where s_j is the number of states assumed by the jth variable, increases exponentially and hence the number of parameters to be estimated, $\prod_{j=1}^{d} s_j - 1$, can quickly become prohibitive. Even for large samples empty cells or cells containing only 1 or 2 observations are common, giving at best unreliable estimates and at worst none. This is a particular problem in discriminant analysis as future cases may arise displaying patterns of feature variables not encountered in one or more of the design sets, with consequent difficulty in classification. Especially for disproportionate sample sizes, when interest is often in the less commonly occurring population, a zero

estimate for a given state in π_1 may represent a different, or higher, probability from a zero, or non-zero estimate for the same state in π_2 with the result that error rates can be misleading (Goldstein and Dillon, 1978, pp. 14, 45).

Two main approaches have been taken to remedy this :

1) smoothed relative frequency estimators, discussed in Section 3.4.2, and

2) imposition of some degree of parametric structure on the multinomial probabilities so as to reduce the number of parameters to be estimated. Saturated reparametrised models will generally have the same number of parameters as the full Multinomial but simpler models can be sought in the hope that a reasonably close fit will be obtained and comparable classification performance achieved but problems of sparse data more easily dealt with. Where the parameters are known the likelihood ratio rule based on the full Multinomial distribution would incur a smaller misclassification rate than a reduced model but due to unreliable parameter estimates a sample-based reduced model may well achieve smaller error rates than the full maximum likelihood rule. In some cases formal statistical model fitting is feasible as, for instance, with log-linear models. Where this is not possible a more arbitrary choice will be made, as with latent class models.

The most restrictive model is to assume complete independence between feature variables within each class (Warner et al., 1961). In practice this will rarely be justified but independence models have proved surprisingly robust to lack of independence (de Dombal et al., 1972, and Horrocks et al., 1972; Titterington et al., 1981). In fact, Hilden (1984) shows that while conditional independence of the feature variables is a sufficient condition for the independence-based Bayes' rule to hold, it is not necessary, in that the model is still valid under certain types of interdependence. Independence models also have the advantage of being very simple to apply and coping easily with missing values. As parameter estimation relies only on the marginal distributions it does not suffer to the same extent from data sparseness.

For binary variables Bahadur models provide a wide range of models incorporating interaction terms. Bahadur (1961) showed that the Multinomial model for a binary feature vector $\underline{x}$ can be represented as

$$p(\underline{x}|\pi_i) = p_I(\underline{x}|\pi_i)\left\{1 + \sum_{j<k} p_{jk}\, z_j z_k + \sum_{j<k<l} p_{jkl}\, z_j z_k z_l + \ldots . \right. \left. p_{12.....d}\, z_1 \cdot\cdot\cdot\cdot\cdot z_d \right\} \qquad (1.5)$$

where $p_I(\underline{x}|\pi_i)$ fits the independence model, $\prod_{j=1}^{d} \theta_j^{x_j}(1-\theta_j)^{1-x_j}$, $\theta_j = E(X_j)$, $X_j \sim Bi(1,\theta_j)$, and the term in brackets models any interactions. $Z_j \equiv (X_j-\theta_j)/\sqrt{\theta_j(1-\theta_j)}$,

$$\left.\begin{aligned} p_{jk} &= E(Z_j Z_k) \\ &\vdots \\ p_{12\ldots d} &= E(Z_1 \ldots . Z_d) \end{aligned}\right]$$

so that $(p_{1,2,\cdots}, p_{d-1,d\cdots\cdots}, p_{12....d})$ are correlation coefficients. Setting $\underline{p} = \underline{0}$ gives the independence model or 1st order Bahadur model, while a 2nd order model sets interactions of 2nd order and above to zero, utilising 1st and 2nd order marginal distributions to estimate $\{\theta_j\}$, $\{Z_j\}$ and $\{p_{jk}\}$, maximum likelihood estimates being given by

$$\hat{\theta}_j = \sum_{w_j} \frac{n(\underline{x})}{n}, \quad z_j = \frac{x_j - \hat{\theta}_j}{\sqrt{\hat{\theta}_j(1-\hat{\theta}_j)}} \quad \text{and} \quad \hat{p}_{jk} = \frac{\sum_{w_{jk}} \frac{n(\underline{x})}{n} - \hat{\theta}_j\,\hat{\theta}_k}{\sqrt{\hat{\theta}_j(1-\hat{\theta}_j)\hat{\theta}_k(1-\hat{\theta}_k)}}$$

where $w_j = \{\underline{x} : x_j = 1\}$ and $w_{jk} = \{\underline{x} : x_j = 1 \text{ and } x_k = 1\}$. The model has the disadvantage that it can lead to negative estimates when sample correlations are used to estimate population correlations (Moore, 1973).

Lancaster models (Victor, Trampisch, and Zentgraf, 1974) provide an alternative for general discrete variables and reduce to the corresponding order of Bahadur model for binary variables. Based on the definition of mth order interaction of Lancaster (1969, pp. 252-256), for d random variables, Zentgraf (1975) showed that if interactions of order m and higher disappear, that

$$P(X_1 = s_1, \ldots, X_d = s_d) =$$

$$\sum_{j=2}^{m}(-1)^{m-j}\begin{bmatrix}d-1-j\\ m-j\end{bmatrix}\cdot\sum_{C_d^j} P(X_{k_1}=s_{k_1},\ldots,X_{k_j}=s_{k_j})$$

$$\cdot P(X_{i_1}=s_{i_1})\ldots\ldots P(X_{i_{d-j}}=s_{i_{d-j}})$$

$$+(-1)^m\left[\begin{bmatrix}d-1\\ m\end{bmatrix}-d\begin{bmatrix}d-2\\ m-1\end{bmatrix}\right]\cdot\prod_{i=1}^{d}P(X_i=s_i)$$

where s_i is an outcome of variable i and C^j_d is the set of all combinations of j elements out of $\{1,\ldots,d\}$ and $(i_1,\ldots,i_{d-j}) = (1,\ldots,d) \setminus (k_1,\ldots,k_j)$. Again cell probabilities are estimated by use of marginals of orders 1 to m and relative frequency estimates should be stable for m sufficiently small even for moderate sample sizes, as explains the robustness of the independence model in the presence of associations compared to more complex models (Victor et al., 1974; Titterington et al., 1981). It is easily fitted sequentially, increasing the order m. Like the Bahadur model, the 1st and dth order models correspond to the independence and full Multinomial models, and in certain cases the model may yield negative estimates.

For m = 2 the model becomes

$$P(X_1=s_1,..,X_d=s_d) =$$

$$\sum_{C_d^2} P(X_{k_1}=s_{k_1}, X_{k_2}=s_{k_2})\cdot P(X_{i_1}=s_{i_1})\ldots P(X_{i_{d-2}}=s_{i_{d-2}})$$

$$-\left[\begin{bmatrix}d\\ 2\end{bmatrix}-1\right]\cdot\prod_{i=1}^{d}P(X_i=s_i).$$

Trampisch (1978) compared the performance of the independence, multinomial, linear and quadratic discriminant rules and the 2nd order Lancaster model in terms of error rate, for 2 populations where equal priors were assumed. Five binary variables were used and a wide range of sample sizes. In all cases the error rates of the independence model and LDF were very close as were those of the 2nd order model and QDF, despite the continuous methods being inappropriate for discrete data. Where the variables were nearly independent the former were to be preferred whereas these were poor for data of unknown structure (generated randomly) where the Multinomial was always superior even for very small sample sizes. Where the data followed a 2nd order Lancaster model the full Multinomial would only be recommended for very large sample sizes but there was little to choose between any of the other methods.

Log-linear models, used in contingency table analysis, provide

a wide class of models for representing the joint distribution of discrete variables such that all states have non-zero probability, parametrising $p(\underline{x}|\pi_i)$ in terms of main effects and interactions of all orders $\leq d$,

$$\ln p(\underline{X} = (x_1, \ldots, x_d)|\pi_i) = \alpha_0^{(i)} + \sum_l \alpha_l^{(i)}(x_l) + \sum_{l<j} \sum \alpha_{lj}^{(i)}(x_l, x_j)$$

$$+ \ldots\ldots\ldots + \alpha_{1\ldots d}^{(i)}(x_1, \ldots, x_d)$$

subject to the identifiability constraints

$\sum_{x_l=1}^{s_l} \alpha_l^{(i)}(x_l) = 0$, $\forall l = 1,..,d$ where s_l is the number of states of the lth variable, $\sum_{x_l} \alpha_{lj}^{(i)}(x_l, x_j) = 0 = \sum_{x_j} \alpha_{lj}^{(i)}(x_l, x_j)$, $\forall\ l,j$.

$$\vdots$$

$$\sum_{x_k} \alpha_{1\ldots d}^{(i)}(x_1, \ldots, x_d) = 0, \ \forall\ k = 1,.., d.$$

The full representation is again equivalent to the Multinomial model. Here reduced models may be fitted by a hierarchical goodness-of-fit approach based on Pearson's x^2 or the Generalised Likelihood Ratio Test statistic, testing each model within the saturated one. Maximum likelihood parameter estimates are obtained iteratively by iterative scaling (iterative proportional fitting) (Bishop, Fienberg and Holland, 1975, pp. 83-96) or iteratively reweighted least squares (Finney, 1971, pp. 52-57; Nelder, 1974).

Log-linear models extend to ordinal variables and a survey of suitable models is provided by Agresti (1983). They are also closely related to logistic models, considered in Section 1.5, and are equivalent for the special case of binary variables.

The study of Gilbert (1968) used data generated from a log-linear model though her parameterisation was slightly different.

A less used class of models is latent class models (see for instance Skene (1978), while Everitt (1984) provides a general introduction to such models). These are based on assuming the existence of an additional discrete latent variable y, taking values $n = 1,\ldots, m$ which contains all the information provided by the d-dimensional feature vector and defines "latent classes" of similar feature vectors. Within each class the feature variables are generally assumed to be independently distributed so that the

number of parameters is automatically reduced. This gives

$$p(\underline{x}|\pi_i) = \sum_{n=1}^{m} \prod_{j=1}^{d} p_j(x_j|\underline{\theta}_{jn})\ p(y=n|\pi_i),$$

a mixture, where only the mixing weights depend on the outcome of interest, π_i, and $\underline{\theta}_{jn}$ is a vector of parameters. The choice of the number of levels m by statistical means is non-trivial. Latent class models are examples of mixture models (Titterington, Smith and Makov, 1985, pp. 25-27) and choosing the order m is equivalent to estimating the number of component densities, the problem of which is well known. Titterington et al. (1985), pp. 148-167, discuss formal statistical procedures, of which p. 157 relates specifically to latent structure models. Maximum likelihood estimates of the parameters $\{\underline{\theta}_{jn}, p(y=n|\pi_i)\}$ are obtained using the EM algorthm (Dempster et al., 1977) and the classification procedure is then based on $p(\pi_i|\underline{x})$ as usual.

While the assumptions of the model may not be justified, various features make it useful in that, due to the assumption of independence, incomplete training cases are easily incorporated, different forms may be used for data of mixed type, and the estimated mixing weights characterise the outcome classes. When the numbers of latent and outcome classes are equal and there exists a 1 to 1 correspondence between the latent classes and outcomes, $p(\underline{x}|\pi_i)$ reduces to the independence model.

Skene (1978) gives an example using the data of Jennett et al. (1976) (considered further in Chapter 4) using 2 variables treated as normal, 1 Binomial and a nominal variable, with 500 training cases of 2 outcome types. Of 500 test cases, 382 were correctly classified but the method was not compared with any other.

Discrete variables and linear models.

Categorical variables range from categorised continuous variables and qualitative ordinal discrete variables such as number of symptoms displayed, to qualitative ordinal ones, such as degree of pain, and qualitative nominal variables with no intrinsic ordering at all e.g. race. Where qualitative variables are encountered, numerical scores are commonly assigned to them, hence inducing a linearity which may not exist. For nominal variables the problem is removed by replacing an s-category variable with s-1 linearly independent binary design variables. For ordinal

variables scoring is more sensible but the problem remains that the categories do not necessarily reflect an interval scale, while use of design variables would obscure the ordering completely. Whether or not scores have been assigned, in practice standard procedures such as the LDF have commonly been applied to discrete data as if it were continuous. Various authors have studied the robustness of the LDF when used in this way. Some representative results of these studies are now given.

Gilbert (1968) studied the performance of the LDF, the full Multinomial model, the independence model and 2 linear logistic models, on multivariate binary data generated from a 1st order (log-linear) interaction model, using 6 variables and 15 pairs of populations, comparing them in terms of misclassification rates and correlation between estimated and true log likelihood ratio. Over 100 simulations for mixed sample sizes of 100 and 500 there was little to choose between the 4 linear methods, which all outperformed the Multinomial, and the LDF was recommended as it should remain stable as the number of variables increases (due to only using 2nd order marginals in parameter estimation), and is readily adaptable for use with mixed data.

Moore (1973) also studied the Multinomial, independence, and LDF models, comparing them to the QDF and 2nd order Bahadur models ((1.5) et seq.) on data generated from the latter, on the grounds that the Bahadur models represent a wider class of distributions than those used by Gilbert and conclusions can be drawn in terms of correlation structure. Again 6 binary variables were used, and now 19 pairs of populations such that within each population variable means were equal, as were non-zero correlations, for 50 samples each of size 50 and 100. In general, the LDF and independence models were superior in populations with zero correlations, and were comparable, while 2nd order procedures were poorer but better than the full Multinomial, as was also true for populations with only 1 non-zero correlation. Where all correlations were positive, 2nd order models were at least as good as any other procedure. Where log likelihood ratios featured reversals i.e. were not monotonic in the number of positive x_j's, no linear model would do well and LDF and independence methods were poorer than the full Multinomial while the 2nd order Bahadur model was best. However the superiority of the latter did not maintain in further sampling from higher order models. In such situations the full Multinomial

was recommended. The quadratic rule rarely performed as well as the LDF and its use was not recommended for binary variables.

Dillon and Goldstein (1978), again using 6 binary variables, considered a wider variety of situations than Moore (1973), and also considered a distance method and 2 Martin-Bradley models (see Section 1.4.3) - the main effects model and 1st order interaction model. For equal sample sizes the distance method would behave identically to the Multinomial model in classification and therefore disproportionate sample sizes were used. For 3 types of correlation structure and comparing methods in terms of mean error rate over 100 simulations the distance method was almost always better than the Multinomial, especially for large differences in correlations between populations, or large correlations where a common correlation structure was assumed. These situations also caused poor performance of the linear models, as did the case of similar mean vectors. For the LDF, large negative correlations appeared less detrimental than large positive ones. The results confirmed those of Moore (1973) that linear models can be poor in the presence of reversals in log likelihood ratios. In general the 2nd order Bahadur model was better than other procedures though Martin-Bradley main effects and 2nd order models were reasonable in situations of small differences in population means, and the latter superior to the 2nd order Bahadur when correlations of large magnitude were present.

Lachenbruch (1975a, pp. 44-45) quotes Revo (1970) as having studied the performance of the LDF and full Multinomial procedures as well as a nearest-neighbour method, relative to the sample likelihood ratio rule based on the appropriate model for ordered data generated from discretised Normal distributions partitioned into 6 categories, Poisson and Negative Binomial distributions in 1 and 2 dimensions. Equal samples of sizes 10, 20 and 50 were used in 1 dimension and 20 and 50 in the bivariate case. In terms of various error rates, the LDF compared favourably to the optimal rule in most cases and in general the likelihood ratio rule did as well as the LDF. Especially for small samples, the nearest-neighbour and multinomial rules were poorer due to sampling error in parameter estimation. The discriminatory ability of the LDF decreased as correlation between variables increased, confirming the results of the studies above, but the performance of the multinomial and nearest-neighbour methods improved.

In a comprehensive study of discrimination techniques on a real data set, Titterington et al. (1981) found the LDF and independence methods to be consistently robust. Lancaster models also performed well and were comparable to the LDF. The QDF, latent class and nonparametric kernels were poor by comparison. Further details are given in Chapter 4.

Some general conclusions of these studies are therefore that the LDF and independence models will perform well when used with many real data sets, but though robust to non-normality and lack of independence respectively, will perform less well if interactions, reversals in the log likelihood ratio, or large differences in dispersion matrices/population correlation structure are present. Given sufficient data relative to dimension to enable reliable parameter estimation, where these features are present a higher order model is preferable - the QDF for continuous data transformed to approximate normality, or, say, a 2nd order Bahadur or Lancaster model for discrete variables. Reliable fitting of the full Multinomial will only be feasible for very large samples. In general, model fit is less crucial when classification rather than estimation of odds is the aim.

Mixed variables

Where feature variables are a mixture of continuous and binary or discrete variables it is common to treat the discrete variables as continuous and use, say, the LDF as discussed above, or less commonly, to convert the continuous variables to categorical ones and use discrete procedures. Cochran and Hopkins (1961) advocated the latter on the grounds that discrete procedures are simpler since distinct discriminants are required for each multivariate state defined by the discrete variables (Linhart, 1959) if mixed data are retained. For large numbers of independent normal random variables and 2 populations little discriminatory power was lost by partitioning into 5 or 6 states. For fewer variables relative efficiencies were a little lower.

Mixed models are available however. As indicated above, the latent class model readily adapts. The "location model" of Krzanowski (1975) for q binary variables $\underline{X}$ and p continuous variables $\underline{Y}$, assumes that for each realisation of $\underline{X}$, $\underline{Y}|\underline{X} = \underline{x}$ is Multivariate Normal so that the unconditional distribution of $\underline{Y}$ is mixed Normal. The Bayes' classification procedure is then to

allocate on the basis of 2^q separate LDFs, one for each cell $\underline{X} = \underline{x}$. Assuming common dispersion matrices, Σ, in all cells and writing $\underline{X} = (X_1, .., X_q)$ as multinomial $\underline{Z} = (Z_1, .., Z_s)$, $s = 2^q$ and p_{im} denoting

$$\Pr(\underline{X} = Z_m | \pi_i), \quad m = 1 + \sum_{i=1}^{q} x_i \, 2^{(i-1)},$$

we have the rule :

$$\text{allocate to } \pi_1 \text{ if } \quad (\underline{\mu}_1^m - \underline{\mu}_2^m)^T \Sigma^{-1} (\underline{y} - \tfrac{1}{2}(\underline{\mu}_1^m + \underline{\mu}_2^m)) + \ln \frac{p_{1m}}{p_{2m}} \geq r,$$

where $\underline{\mu}_i^m$ is the mean in population i and cell m, and r is the usual cut-off depending on costs and prior probabilities. This generalises the work of Chang and Afifi (1974) for q = 1, although they allowed distinct covariance matrices for each cell.

If $\underline{Y}$ and $\underline{X}$ are independent so that $\underline{X}$ does not affect classification the location rule reduces to the standard LDF based only on the p continuous variables. As for the LDF, the rule is Bayes' risk consistent in that the probability of correct classification converges uniformly to the optimal rate (van Ryzin, 1966; Glick, 1972). In practice the procedure requires estimation of more parameters and therefore may give less reliable error rates than the LDF especially for multiple binary variables, or equivalently a single polychotomous variable, to which the method generalises (using 2^q LDFs for q binary variables or 1 2^q-cell multinomial). In view of the difficulty of parameter estimation due to small n_i relative to s, causing some cell frequencies to be close to zero, Krzanowski suggested fitting reduced models for estimation purposes, such as a 2nd order log-linear model for the $\{p_{im}\}$ and an additive linear model for the means $\{\hat{\underline{\mu}}_i\}$, deriving $\hat{\Sigma}$ in the usual way from $\{\hat{\underline{\mu}}_i\}$. These estimates would then be substituted into the optimal rule. Goldstein and Dillon (1978), pp. 94-95, note that the location model extends immediately to q polychotomous variables with $\prod_{j=1}^{q} s_j$ states but that state sparseness becomes more problematic, (see also Krzanowski, 1980), and, further, that it can be generalised from linearity in the obvious way by taking $\text{cov}(\underline{Y}|\underline{X}) = \Sigma_i$ under π_i but common to all states.

A suitably small number of binary/discrete variables may be chosen by a step-down variable selection procedure of Krzanowski (1983b), based on a measure of distance between 2 populations in terms of mixed variables, where the location model holds (Krzanowski, 1983a).

Comparing the LDF (treating all variables as continuous) and the location model, for 1 continuous variable and q independent binary variables with common mean p_i in π_i, i = 1, 2, and unit Mahalanobis distance between populations in each cell, assuming the location model to hold, Krzanowski (1975) found that optimum error rates were very similar for a range of p_i and q = 2, 3 and 4 where there were no interactions, but the LDF error rate was up to 50% higher in all cases where there was interaction between binary variables and the populations, since the hyperplanes separating the populations in each cell are then no longer parallel and averaging them reduces discriminatory power. In a simulation-based comparison for 2 continuous variables, 2 to 5 binary variables, and moderate p_i, average actual error rates over 10 simulations of 50 observations from each population were seen to increase with the number of variables for the location method, due to poorer estimation, whereas the LDF was more stable. Consequently, an upper limit of 6 or 7 binary variables were recommended, or fewer for smaller samples.

Five medical data sets were also used to compare the location model, the LDF, logistic discrimination (Anderson, 1972) and a discrete nonparametric rule of Hills (1967) (see Section 3.4.2), dichotomising continuous variables, in terms of leaving-one-out error rate. In 3 out of 5 cases the LDF and location model were comparable while the latter was superior in the 2 larger data sets due to better estimation, although cases misclassified differed considerably. The logistic method was almost identical to the LDF in both respects while Hills' procedure was consistently poorest. The results therefore indicated that treating discrete variables as continuous is less serious than the reverse (although dichotomisation represents an extreme loss of information).

Krzanowski (1977) also compared use of the LDF and location model when the latter holds, assuming population parameters to be known, in a wide variety of situations. The joint distribution of the binary variables was represented as a 2nd order Bahadur model with equal correlations between all binary variables and common

binary means in a given population so that the within category discriminant functions are parallel. Like Lachenbruch et al. (1973) he found that error rates for the LDF were almost always higher than optimal in one population and lower in the other, and confirmed the results of Moore (1973) that the LDF fared worse with binary variables when correlation terms in the Bahadur representation were non-zero except where $p_{1m} = p_{2m}$, i.e. where binary variables contributed no information. The largest differences in performance of the LDF between independent and correlated binary variables occurred for widely separated populations, especially as dimensionality increased. Again the increase in mean error rate over optimal was much higher where the relationship between the continuous variable means in the 2 populations varied from cell to cell, causing the optimal discriminants not just to be non-parallel but also change direction from cell to cell. On the grounds of Moore's results for all binary variables, he recommended that if the sample correlation matrices of the variables reveal numerous moderate positive correlations that the multinomial rule or another discrete procedure be used and that the same correlation structure indicates for mixed data that the location model is more appropriate than the LDF, as it will also be for high correlation between a binary and continuous variable or between-population differences in magnitude and/or sign of such correlations.

Vlachonikolis and Marriott (1982) proposed a slightly different procedure, the modified LDF. The q binary variables are replaced by $s = 2^q-1$ design variables $z_1,\ldots, z_s$ representing a particular pattern of $(x_1, \ldots, x_q)$, so that the LDF on $(z_1,\ldots, z_s, y_1,\ldots, y_p)$ requires estimation of $2^q + p$ coefficients rather than $p + q + 1$ and the LDFs dividing the continuous variables at each location will be parallel but their positions determined separately for each state. More generally, including products $\{z_iy_j\}$ would involve $2^q(1 + p)$ coefficients, as does the location model, and also fits 2^q separate LDFs but is more general in not assuming common dispersion matrices at each location. Including the $\{z_i\}$ and $\{z_iy_j\}$ allows for any interaction due to binary variables. Again unless q is small the number of parameters is large. Since $(z_1,\ldots, z_{s+1})$ is equivalently represented by $(\ldots x_i,\ldots, x_ix_j,\ldots, x_ix_jx_k,\ldots, x_1x_2x_3\ldots x_q)$, i.e. the full Multinomial, reduced models may be fitted, a 2nd order model, for

instance, involving a LDF using $\{x_i\}$, $\{x_ix_j\}$, $\{y_k\}$, $\{x_iy_k\}$, and $\{x_ix_jy_k\}$, $i \neq j = 1,\ldots, q$ and $k = 1,\ldots, p$. Alternatively, the location model approach to parameter reduction could be used.

Using a data set of Krzanowski (1975) where the LDF had been seen to be far poorer than the location model, 1st and 2nd order modified LDFs were also used as was the kernel method used by Habbema, Hermans and Remme (1978) with both (transformations of) a single and separate smoothing parameters for the discrete and continuous variables. The 2nd data set involved a large number of variables and locations and the location method was not used. Stepwise methods were used for a logistic model and the LDF methods. In terms of leaving-one-out error rate, on both data sets the kernel methods were poor, though 2 smoothing parameters were preferable to 1. The modified LDFs were better than the simple LDF on both data sets, but the location model still much superior on the 1st data set. The logistic model was slightly better than the simple LDF but did not identify interactions. The conclusions were that despite the LDF and independence models being robust against non-normality and dependence respectively, neither is robust against interaction, and hence the location and modified methods can remove the disadvantages of the simple LDF.

It therefore seems that the LDF (or the logistic model) is comparable to the location model in terms of error rate except where there are interactions between discrete variables or between discrete and continuous variables, when the LDF loses power. The location model may be superior for larger samples but where samples are small relative to dimension it will suffer due to poorer estimation. The number of discrete states with which it will cope well is relatively small, and in higher dimensions modified LDFs may provide an alternative.

1.4.3. Nonparametric methods

While parametric methods are the most powerful when the model assumed is appropriate, often weaker assumptions may be preferred. Various nonparametric means of density estimation are described below, which may be used for discrimination by substitution in (1.1). The final section discusses some nonparametric procedures which bypass density estimation and classify directly.

Kernel density estimators

The kernel method replaces the assumptions of a specific parametric form for a density with weaker assumptions of smoothness.

Given a random sample $X_1, \ldots, X_n$, a kernel density estimator takes the form

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

where K(t), the kernel function, is a smooth function, usually a probability density function (p.d.f.), and h, the smoothing parameter, determines the width of K(y) and hence the degree of smoothness of $\hat{f}(x)$. While the choice of $K(\cdot)$ does not greatly affect the appearance of the estimate, the size of h is critical. Many methods have been proposed for its choice, some of which are discussed in Section 2.3. While Rosenblatt (1956) showed that the estimator is biased, Parzen (1962) showed that subject to the conditions

$$\left.\begin{array}{ll} 1) & \int K(t)\, dt = 1 \\ 2) & \int |K(t)| dt < \infty \\ 3) & \sup |K(t)| < \infty \\ 4) & \lim_{t \to \infty} |t\, K(t)| = 0 \\ 5) & \lim_{n \to \infty} h(n) = 0 \\ 6) & \lim_{n \to \infty} nh(n) = \infty \end{array}\right\} \qquad (1.6)$$

$\hat{f}(x)$ is consistent, asymptotically unbiased and asymptotically normal. Much of the literature has been concerned with establishing suitable conditions on the sequence $\{h(n)\}$ to ensure desirable asymptotic behaviour. Cacoullos (1966) and Epanechnikov (1969) extended the method to multivariate $\underline{x}$. It also extends readily to discrete variables, the method of Aitchison and Aitken (1976) being the most commonly used, estimating p(x) by

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x|X_i, \lambda)$$

where λ now plays the role of the smoothing parameter. For a single binary variable

$$K(x|X_i, \lambda) = \lambda^{1-|x-X_i|}(1-\lambda)^{|x-X_i|} = \left.\begin{array}{rl} \lambda, & X_i = x \\ 1-\lambda, & X_i \neq x \end{array}\right\}, \quad \tfrac{1}{k} \leqslant \lambda \leqslant 1.$$

It can be extended to multiple dimensions and/or general nominal or ordinal discrete variables. For mixed data, $\underline{z} = (\underline{x}, \underline{y})$ Aitchison and Aitken (1976) proposed

$$\hat{p}(\underline{z}) = \frac{1}{n} \sum_{j=1}^{n} K_1(\underline{x}|\underline{X}_j, \lambda)\, K_2(\underline{y}|\underline{Y}_j, h) \qquad (1.7)$$

where $\underline{x}$ and $\underline{y}$ denote discrete and continuous random vectors, K_1 is a discrete kernel and K_2 a continuous one and λ, h are corresponding smoothing parameters. Alternatively, Goldstein and Dillon (1978, pp. 95-96) suggest generalising a location model by relaxing the assumption of conditional normality and allowing a more general joint distribution of the continuous variables $\underline{Y}|\underline{X}$, possibly distinct within each state and each population (though this is probably only feasible for a very few binary variables $\underline{X}$ and fairly high frequencies in each state). If the continuous density of $\underline{Y}|\underline{X}$ in state m is $f_{im}(\underline{y})$ under π_i, the discriminant rule would be to allocate an observation $(\underline{x}, \underline{y})$ to π_1 if

$$\frac{f_{1m}(\underline{y}) \cdot p_{1m}}{f_{2m}(\underline{y}) \cdot p_{2m}} \geqslant r,$$

where $\{p_{im}\}$, the probabilities of state m in π_i, are estimated from n_{im} sample observations, and f_{im} by any suitable nonparametric density estimation technique. This approach allows use of standard continuous kernels rather than mixed ones.

Thus the kernel method is readily applicable to continuous, discrete or mixed data, either univariate or multivariate, making it a flexible and widely used approach. Discrete kernels are discussed in more detail in Section 3.4.2 and continuous kernels in Chapter 2, as is their use in classification.

Nearest-neighbour methods

The kernel method described above determines the width spanned by each kernel by its choice of h, the smoothing parameter. An

alternative, proposed by Loftsgaarden and Quesenberry (1965) is to determine the number of neighbouring points, k, to be spanned by the kernel. If $r_k(\underline{x})$ is the kth largest Euclidean distance of $\underline{x}$ from each of the n sample points, then the kth-nearest-neighbour estimator is given by

$$\hat{f}(\underline{x}) = \frac{k}{nV_k(\underline{x})},$$

where $V_k(\underline{x})$ is the volume of the d-dimensional hypersphere of radius $r_k(\underline{x})$ centred on $\underline{x}$. It provides a consistent estimator, and is closely related to the kernel method, but has the disadvantage that it is not a p.d.f.. It leads naturally to a discriminant rule based on the ratio of the density estimates : if the order k is common to both samples, and sample relative frequencies are used to estimate $p(\pi_i)$, classification is on the basis of the smallest volume $V^{(i)}(\underline{x})$. Alternatively, if the sample points from each class are pooled and amongst the k nearest neighbours contained in the hypersphere of volume V are k_i originating from class i and $\hat{p}(\underline{x}|\pi_i)$ is now given by $k_i/(n_iV)$, the resulting Bayes' rule would be :

allocate to class i if $k_i = \max_j (k_j)$.

This is known as the kth-nearest-neighbour classification rule and is considered again below.

Hand (1981a, p. 103) shows that the method can be used for categorical variables also. For instance, for multiple binary variables where the distance between any 2 points is the number of components on which they differ, the volume V is given by

$$\sum_{r=0}^{i-1} \begin{bmatrix} d \\ r \end{bmatrix} + \left[\frac{k-n_{i-1}}{n_i-n_{i-1}} \right] \begin{bmatrix} d \\ i \end{bmatrix},$$

where the kth-nearest neighbour is at distance i from $\underline{x}$, $n_{i-1} < k \leq n_i$ where n_i is the number of points at distance i from $\underline{x}$, and a proportion of the $\begin{bmatrix} d \\ i \end{bmatrix}$ cells at distance i are included.

Orthogonal series estimators

Another type of nonparametric estimator, due to Cencov (1962),

views a univariate density as a waveform and expands it in terms of a series of orthonormal basis functions, e.g. as a Fourier series, or using certain sets of polynomials. If $\{\phi_i\}$ is a set of basis functions such that

$$\int \phi_i(x)\ \phi_j(x)\ dx = \delta_{ij},$$

where δ_{ij} is the Kronecker delta, and

$p(x|\pi_i) = \sum_{i=1}^{\infty} a_i\ \phi_i(x)$ where the $\{a_i\}$ are coefficients, a suitable estimator is obtained by truncating the series to a finite number of terms, m, either arbitrarily or selecting a model on the basis of goodness-of-fit. Including too few terms corresponds to oversmoothing, while too many terms will overfit, corresponding to undersmoothing. Tarter and Kronmal (1976), who provide a useful introduction to series methods, give an example. The coefficients are estimated by minimisation of an error criterion such as integrated square error,

$$\int \left(p(x|\pi_i) - \sum_{i=1}^{m} a_i\phi_i(x) \right)^2 dx,$$

from which $a_j = E\{\phi_j(x)\}$ and hence

$$\hat{p}(x|\pi_i) = \sum_{j=1}^{m} \left\{ \frac{1}{n_i} \sum_{k=1}^{n_i} \phi_j(x_k) \right\} \phi_j(x).$$

A disadvantage is that where the sample points occupy a small region of the sample space a large number of terms may be required for an adequate fit (Hand, 1981a, p. 41).

In principle the method extends to vector-valued $\underline{x}$. However, as the number of dimensions increases, the number of terms in the series increases exponentially, limiting its applicability.

Hand (1981a, p. 42) notes that the series estimator is of kernel form, writing for the general multivariate case

$$K\left[\frac{\underline{x}-\underline{x}_k}{h}\right] = h \sum_{j=1}^{m} \phi_j\ (\underline{x}_k)\ \phi_j(\underline{x}).$$

For multivariate binary variables, Ott and Kronmal (1976) represented $f(\underline{x})$ as

$$f(\underline{x}) = 2^{-d} \sum_{r} \alpha_r \, \phi_r(\underline{x}),$$

a linear combination of the orthogonal polynomials

$$\{\phi_r(\underline{x}) = (-1)^{\underline{x}^T \underline{r}}\}$$

where $\underline{r}$ is a binary index vector labelling the 2^d points in the sample space. The orthogonality is such that

$$\sum_{\underline{x}} \phi_r(\underline{x}) \, \phi_j(\underline{x}) = \left.\begin{array}{ll} 2^d, & r = j \\ 0, & r \neq j \end{array}\right\}. \qquad (1.8)$$

The parameters $\{\alpha_r\}$ are such that $\alpha_j = E\{\phi_j(\underline{x})\}$, since

$$\sum_{\underline{x}} \phi_r(\underline{x}) \, f(\underline{x}) = E\{\phi_j(\underline{x})\},$$

with maximum likelihood estimator

$$\hat{f}(\underline{x}) = 2^{-d} \sum_{r} \hat{\alpha}_r \, \phi_r(\underline{x}), \qquad \hat{\alpha}_r = \sum_{\underline{x}} \frac{\phi_r(\underline{x}) \, n(\underline{x})}{n}.$$

Here $f(\underline{x})$ involves $2^d - 1$ independent parameters and is equivalent to the full Multinomial model. Kronmal and Tarter (1968) used the change in mean summed square error

$$E \sum_{\underline{x}} (f(\underline{x}) - \hat{f}(\underline{x}))^2$$

as a goodness-of-fit criterion for inclusion of terms, while Goldstein and Dillon (1978, p. 38) suggest in the 2 population discrimination problem use of $E[(\hat{f}_1(\underline{x}) - \hat{f}_2(\underline{x})) - (f_1(\underline{x}) - f_2(\underline{x}))]^2$. Ott and Kronmal (1976) used these and 2 other criteria with these models, in the 2 population case, comparing them with the independence, Multinomial and a logistic model in terms of mean error rate and also assessed density estimators in terms of mean sum of squares. For 11 multinomial populations and 5 variables they found that the independence method did best for 1st order interaction models of the type studied by Gilbert (1968) and the logistic model was nearly as good. Both also did reasonably well in general except where optimal error rate was high. Three out of 4 of the series methods behaved very similarly, while a weighted mean summed squared error method was generally better. Both for classification and density estimation, a basic mean summed square error method almost always outperformed the Multinomial. However,

the Kronmal-Tarter models suffer from the disadvantage that density estimates can be negative.

A different class of polynomials was used by Martin and Bradley (1972), again for multivariate binary data, who fitted models of the form

$$f_i(\underline{x}) = f(\underline{x})\,[1 + h(\underline{a}^{(i)}, \underline{x})],$$

in a manner similar to Bahadur models (1.5), where $f_i(\underline{x})$ is the ith class conditional density, and $h(\underline{a}^{(i)}, \underline{x})$ is a polynomial in $\underline{x}$ with coefficients $\underline{a}^{(i)}$ in π_i,

$$h(\underline{a}^{(i)}, \underline{x}) = \sum_{j=0}^{2^d} a_j^{(i)}\, \Phi_j(\underline{x}).$$

The polynomials used were

$$\Phi_0(\underline{x}) = 1$$

$$\Phi_i(\underline{x}) = 2x_i - 1, \quad i = 1, \ldots, d.$$

$$\Phi_{d+1}(\underline{x}) = (2x_1-1)(2x_2-1)$$

$$\Phi_{d+\binom{d}{2}}(\underline{x}) = (2x_{d-1}-1)(2x_d-1)$$

$$= (2x_i-1)(2x_j-1)(2x_k-1), \quad i < j < k \leq d$$

$$\Phi_{2^d-1}(\underline{x}) = \prod_{i=1}^{d} (2x_i-1)$$

with orthogonality as in (1.8).

An equivalent representation of the model is

$$h(\underline{a}^{(i)}, \underline{x}) = \frac{f_i(\underline{x}) - f(\underline{x})}{f(\underline{x})}, \quad f(\underline{x}) \neq 0, \quad \text{whence}$$

$$a_j^{(i)} = 2^{-d} \sum_{\underline{x}} \Phi_j(\underline{x}) \left\{ \frac{f_i(\underline{x}) - f(\underline{x})}{f(\underline{x})} \right\}.$$

As in the previous series estimators, the coefficients $\{a_j^{(i)}\}$ may be interpreted as main effects and interactions so that $a_j^{(i)}$ measures the ability of the jth binary variable to indicate the ith

population and $a_{l,j}^{(i)}$ the joint ability of the lth and jth variables to do so. Therefore variables which do not contribute singly to classification may still contribute jointly with other variables. Including all 2^d polynomials again gives the full Multinomial model and a reduced model will be used involving $h_m(\underline{a}^{(i)}, \underline{x})$, m denoting a subset of polynomials corresponding to main effects and low order interactions. Iterative maximum likelihood is required for parameter estimation. As with Bahadur and Tarter-Kronmal models, $\hat{f}_i(\underline{x})$ is not necessarily positive, nor will the probabilities necessarily sum to 1, and additional restrictions are imposed to avoid these problems. Other similar models are those of Brunk (1978), who expanded

$\frac{p(\underline{x})}{p_0(\underline{x})}$, where $p_0(\underline{x})$ may be thought of as a prior estimate of $p(\underline{x})$,

but then used a Bayesian approach, putting a prior on the coefficients of the terms in the orthogonal series in order to estimate them from a posterior mode. Brunk and Pierce (1974) did the same but expanded

$\log\left[\frac{p(\underline{x})}{p_I(\underline{x})}\right]$, where $p_I(\underline{x})$ is the independence model.

Martin-Bradley models were used in a comparative study by Dillon and Goldstein (1978) (see Section 1.4.2). Butler and Kronmal (1985) extend Bahadur, Martin-Bradley, and Ott-Kronmal models from binary to polychotomous predictors, while Hall (1983b) describes orthogonal series methods for the general case of multivariate mixed data. The former compared the models using discrete Fourier functions, and Kronmal and Tarter's (1968) mean square error rule for inclusion of terms, in terms of the increase in average actual error over the optimal error rate. For 4 and 5 trichotomous variables, samples of equal sizes of 50 to 400, with "small", "medium", and "large" differences between the populations, were simulated from 2 populations represented by 2nd order log-linear models. They found the Martin-Bradley model to be superior, improving with repect to other methods as the difference between populations increased, especially for sparse data. The Ott-Kronmal model also improved over the full Multinomial for relatively small sample sizes and low interaction differences. The

Bahadur model was generally poor, contradicting the results of Dillon and Goldstein (1978), and rarely improved over the independence model. This was attributed to the inclusion rule used. The logistic model, when it was estimable, was very poor.

Spline estimators

A polynomial spline of degree 2m-1 is a piecewise polynomial of order 2m-1 with smooth joins such that the first 2m-2 derivatives of the spline are continuous. The (abscissae of the) join points are known as "knots". Commonly m = 2, giving a piecewise cubic.

Let $L_2(a, b)$ denote the set of measurable square integrable functions in (a, b) and $W_m(a, b)$ the set of functions on (a, b) such that $D^j f$, $j \leq m-1$, is absolutely continuous and $D^m f \in L_2$, where D is the differential operator. Then mathematically, given data (x_i, y_i) $i = 1,.., n$, the "cubic interpolating spline" is the solution s(x) of the problem:

minimise over $W_2(-\infty, \infty)$ $\int_{-\infty}^{\infty} \{ D^2 f(x) \}^2 dx$ such that

1) $D^j f \in L_2 (-\infty, \infty)$, $j = 0,.., 2$

2) $f(x_i) = y_i$, $i = 1,..., n$.

More generally, replacing D^2 and W_2 by D^m and W_m respectively, so that the first m-1 derivatives are absolutely continuous and the mth is square integrable, the solution s(x) is such that $D^{2m} s(x) = 0$, and is an interpolating spline of order 2m-1, with knots at the sample points. More generally still, D^m may be replaced by a general linear differential operator, L, of order m, with constant coefficients, $L \equiv \sum_{i=1}^{m} a_i D_i$. The solution is found by solving systems of linear equations imposed by the constraints. The polynomial character of the solution is a consequence of the operator L and other forms of spline are possible.

Implicitly such splines are appropriate in the absence of noise. Smoothing splines provide a more general class of estimators, of which there are 3 main classes. (Wegman and Wright, 1983).

1) Penalised least squares splines minimise, over $W_m(a, b)$,

$$\sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \int_a^b (Lf(x))^2 \, dx,$$

where λ is a constant, $\lambda > 0$, so that the first term replaces the interpolating constraints and again the second term is a measure of curvature, now penalising for lack of smoothness. As $\lambda \to 0$ the interpolating spline results while as $\lambda \to \infty$ faithfulness to the data is smoothed out, and, as for kernel estimators, as $n \to \infty$, $\lambda \to 0$. Various authors have suggested cross-validation as a means of choosing a suitable value of λ e.g. Wahba and Wold (1975), Silverman (1984a). This is an important technique, described in the kernel density estimation context in Section 2.3.2. With $L = D^m$ the solution is again a polynomial spline of order 2m-1 with possible knots at the data points (Kimeldorf and Wahba, 1970). Kimeldorf and Wahba (1971), and Wahba (1978) approach computation by quadratic programming algorithms.

2) The 100% confidence interval approach is an alternative, appropriate when error terms are bounded, which relaxes the constraints rather than generalising the objective function. If (α_i, β_i) is a 100% confidence interval for $f(x_i)$, $\alpha_i < \beta_i$, then we minimise

$$\int_a^b [L(f(x))]^2 \, dx, \text{ such that } f \in W_m \text{ and } \alpha_i \leq f(x_i) \leq \beta_i, \; i = 1, \ldots, n.$$

Again the solution has the same order and polynomial form.

3) The third approach is to assume the piecewise polynomial form, choosing the degree m and number of knots N (usually rather less than n) whose positions may also be specified. Satisfying continuity of the spline and of its first m-1 derivatives leaves $m + N + 1$ coefficients to be determined. For fixed knots the spline function may then be uniquely determined by ordinary least squares. This is the regression spline approach.

Splines were first used in density estimation rather than nonparametric regression by Boneva, Kendall and Stevanov (1971), Wahba (1971), and Good and Gaskins (1971), the last of which is discussed in the next section. Boneva et al. defined a class of estimators called "histosplines", which essentially fit an interpolating spline to the usual histogram, finding the unique solution which minimises

$\int_a^b [f'(x)]^2 \, dx$ over W_1.

such that the areas under the function between each pair of ordinates, defined by bins of width h, are equal to the observed relative frequencies represented by the histogram.

Rather than use a fixed bin width, Wahba (1971, 1976) interpolated the empirical cumulative distribution function (c.d.f.) evaluated at m + 1 order statistics by an mth degree polynomial and estimated the density by the derivative of the fitted polynomial. This is equivalent to the histospline method with bin widths determined by order statistics.

Wegman and Wright (1983), in a survey paper on the use of splines in statistics, comment that since they are based on interpolating rather than smoothing splines, histosplines cannot be expected to perform much more adequately as density estimates than the histogram itself in the presence of noise. Wahba (1975) confirmed this, showing that both variants of the histospline have the same order of magnitude of mean squared error as the kernel method, certain orthogonal series estimators and the ordinary histogram, when smoothing parameters and bin widths are optimally chosen.

Unlike most other methods of density estimation, confidence regions are readily obtainable for splines, by taking a Bayesian view of curve fitting (Silverman, 1985; Wahba, 1983; Wecker and Ansley, 1983).

Splines extend to higher dimensions, where the problem becomes one of surface fitting. Suitable penalty functions become more unwieldy and computation more difficult (Silverman, 1985). Wegman and Wright (1983) discuss relevant work of Wahba and others.

Penalised maximimum likelihood estimators

In general maximisation of the likelihood function

$\prod_{i=1}^{n} f(x_i)$ subject to the constraints

$$\left.\begin{aligned} &\int_a^b f(x)\, dx = 1 \\ &f(x) \geq 0, \ \forall\, x \in [a, b] \end{aligned}\right\}$$

does not have a smooth solution over most sets of functions of interest. Good and Gaskins (1971) suggested instead maximising

$$\prod_{i=1}^{n} f(x_i) \exp \{-\beta\Phi(f)\}$$

or equivalently

$$\sum_{i} \log f(x_i) - \beta\Phi(f),$$

subject to the same conditions, where $\Phi(\cdot)$ is a functional of f, $\Phi(\cdot) \geq 0$, which penalises for roughness, and β is a constant smoothing parameter. They demonstrated that if

$$\int_a^b (\log f(x))^2 f(x)\, dx < \infty$$

then $\hat{f}(x)$ is consistent in that

$$\int_a^b \hat{f}(x)\, dx \xrightarrow{p} \int_a^b f(x)\, dx \ , \ a < b.$$

The solution $\hat{f}(x)$ is called the maximum penalised likelihood estimator (MPLE) of f(x). The roughness penalty Φ is often a function involving a second derivative, and commonly, to avoid the non-negativity constraint, $\Phi(v)$ is used instead where v is the root density, $v = f^{\frac{1}{2}}$, or log f (Silverman, 1982a). De Montricher, Tapia and Thompson (1975) proved the existence and uniqueness of $\hat{f}(x)$ and characterised it, for suitable choice of penalty function, as a polynomial spline with knots at the data points. Klonias (1984) constructed a general class of MPLEs which includes those of Good and Gaskins (1971) and De Montricher et al.. The method has a Bayesian interpretation (Good and Gaskins, 1971), regarding $e^{-\Phi}$ as proportional to a prior density on the function space, so that the MPLE maximises the posterior density over this space. The related estimators of Leonard (1978) use the Bayesian approach more directly, enforcing smoothing via a prior on the derivative of the logistic transform of f, rather than by means of a roughness penalty. MPLEs are closely related to the smoothing spline approach to density estimation and Silverman (1984b, 1985) shows that both the penalised least squares smoothing spline with penalty function

$$\int \{ f''(x)\}^2 \, dx$$

and the MPLE with the same penalty imposed on the log density are approximately equivalent to an adaptive kernel estimator (see Section 2.2.2) with local bandwidth $\lambda^{\frac{1}{4}} n^{-\frac{1}{4}} f(x)^{-\frac{1}{4}}$ where λ is the coefficient of the penalty function.

Titterington and Bowman (1985) and Simonoff (1983) consider MPLEs for discrete probability functions and, as in Good and Gaskins (1980), estimate the roughness penalty coefficient from the data. Simonoff (1983) demonstrated with simulations from the uniform probability vector that in terms of mean squared error his MPLE could considerably improve over a simple smoothed relative frequency estimator of Fienberg and Holland (1973) (see Section 3.4.2). He also suggested suitable higher-dimensional penalty functions.

In principle, MPLEs also extend to multivariate continuous densities and Good and Gaskins (1971) discuss suitable penalty functions.

While "nonparametric" methods typically involve rather fewer "parameters" than "parametric" estimators, once a given method has been selected in general there still remain 2 choices, one of which is more crucial to the success of the model. These are, for the series methods the order of the model/ number of terms to be included, while for each of the other methods the critical factor is the degree of smoothing or order of nearest neighbour. (See also Silverman, 1978c). Less important is the kernel function, set of polynomials chosen (series methods), penalty function (MPLEs), and degree of polynomial and, possibly, number and position of knots (smoothing splines).

The kernel method is simple, flexible, widely used and conceptually appealing, and is the one on which we concentrate henceforth.

Distance based and clustering procedures

Assuming equal costs and equal priors, the LDF is easily seen to assign an object $\underline{x}$ to the population which is nearest in terms of the Mahalanobis distance, $(\underline{x} - \underline{u}_i)^T \Sigma^{-1} (\underline{x} - \underline{u}_i)$, $i = 1,..,k$ (see e.g. Mardia, Kent and Bibby, 1979, pp. 303-304).

Numerous nonparametric procedures dispense with the likelihood ratio criterion but also use distance measures as a means of classification.

For discrete data Matusita (1955) proposed the distance measure $|| F_1-F_2 ||^2 = \Sigma_j (\sqrt{p_1(\underline{x}_j)} - \sqrt{p_2(\underline{x}_j)})^2$, where $F_1 = \{p_1(\underline{x}_j)\}$ and $F_2 = \{p_2(\underline{x}_j)\}$ are 2 discrete distributions defined on the same sample space, which also equals $2(1-\rho)$ where $\rho = \Sigma_j \sqrt{p_1(\underline{x}_j)}\sqrt{p_2(\underline{x}_j)}$, a measure of affinity. If n and m observations are sampled from each population, with empirical distributions S_n and S_m, Dillon and Goldstein (1978) suggested the classification rule :

$$\left.\begin{array}{l}\text{assign to } \pi_1 \text{ if } || S_{n+1} - S_m || \geq || S_n - S_{m+1} || \\ \text{assign to } \pi_2 \text{ otherwise}\end{array}\right\}$$

where S_{n+1} and S_{m+1} are respectively the empirical distributions of the n and m observations augmented by the observation to be classified, i.e. assign to π_1 if the inter-sample-based distributional distance is greater than if it were assigned to π_2. For equal sample sizes n = m this classifies as the multinomial rule would. However, where a zero is observed in one sample, often the smaller of 2 disproportionately sized samples, the multinomial rule automatically classifies to the other population, whereas the distance-based rule is more sensitive and allows the possibility of classification to the smaller population, usually when the non-zero observation is small relative to other counts in that sample. A related point for disproportionate sample sizes is that while total error rate/misclassification probability may be low, often rules tend to misclassify almost all observations from the smaller sample and therefore may be poor in practice (Goldstein and Dillon, 1978, pp. 45-47). Goldstein and Dillon (1978), pp. 47-50 give an example comparing the distance rule with the LDF and multinomial rules where the former was far superior in the smaller group and only slightly poorer overall than the LDF which, like the Multinomial, displayed the behaviour described but to a lesser degree. The multinomial rule had the lowest overall error rate. In practice of course it is often the rare population which is of most interest, as in medical screening for instance.

The 1-nearest-neighbour (1-NN) procedure is one of the simplest distance methods and simply classifies a point $\underline{x}$ to the population of its nearest neighbour in the training sample, while more

generally the kth order procedure assigns $\underline{x}$ to the population most heavily represented in the set of k closest points, $V_k(\underline{x})$. The number of points, n, in the training sample must be sufficiently large relative to k for the number of points in each population in the design set to be greater than or equal to k. Fix and Hodges (1951) showed that it is consistent if $k_n \to \infty$ such that $k_n/n \to 0$. Cover and Hart (1967) showed that in terms of error rate the 1-NN method is admissible amongst the class of k_n-NN rules for the n-sample case and that asymptotically the error rate is bounded above by twice the Bayes' optimal rate. Hellman (1970) discussed error bounds when a reject option or classification of doubt is permitted.

Loftsgaarden and Quesenberry's (1965) modification of the procedure provides a means of consistent density estimation.

The nearest-neighbour method is also a clustering procedure, known as single linkage. Other clustering procedures (see, for instance, Cormack, 1971) may also be used in classification. Fisher and Van Ness (1973) compared the k-NN method, LDF and QDF and the Bayes' procedure using kernels, in terms of 7 admissibility criteria, with the furthest neighbour or complete linkage method (if $\underline{z}_j$ is the farthest point from $\underline{x}$ in group j and $d(\cdot, \cdot)$ is a distance measure,

$$\text{classify to group k} \quad \text{s.t. } d(\underline{x}, \underline{z}_k) = \min_j d(\underline{x}, \underline{z}_j)),$$

the centroid method which chooses the group with closest centroid to $\underline{x}$, average linkage (classifying on the basis of the smallest average sample distance) and the least squares method. The latter is used for batch classification by partitioning it into sets such that the new within-cluster sum of squares is minimised, and is only feasible, without use of iterative approximation, for very small batch sizes. For single point classification it is similar to the centroid method. The conclusions were that for the criteria considered, the nearest-neighbour method satisfied more than any other procedure while the furthest-neighbour and kernel methods were next best. Of the clustering methods only the centroid method satisfied none at all.

An alternative nonparametric approach to classification is provided by binary classification trees (Breiman et al., 1984). These recursively partition subsets of the sample space S,

beginning with S itself, into 2 disjoint sets on the basis of a "splitting rule" such that descendant "nodes" or subsets become progressively "purer" in the sense that they contain fewer classes of training cases than parent nodes. "Terminal" nodes are encountered when further splitting produces no appreciable improvement in purity. The resulting tree structure is used for classification by associating with each terminal node a particular diagnosis on the basis of the class most heavily represented by that node.

In practice of course such methods are only useful when classification itself is the aim rather than estimation of posterior probabilities or odds of class membership.

1.5 ODDS RATIO ESTIMATION

The diagnostic method involves direct modelling of one or more odds ratios of the form $p(\pi_1|x)/p(\pi_2|x)$. While in theory any suitable function defined on $[0, \infty)$ may be used, in practice a parametric approach is the most common, using a logistic regression model.

1.5.1 Logistic regression

As discussed in Section 1.4.2, if $\{f_i(x)\}$ are assumed to have the Multivariate Normal form with equal covariance structure then the log-odds ratio is seen to be linear in the observations $\underline{x}$.

$$\log\left\{\frac{f_1(\underline{x})}{f_2(\underline{x})}\right\} = \alpha_0 + \underline{\beta}^T\underline{x}, \text{ where } \alpha_0 \text{ and } \underline{\beta} \text{ are unknown parameters.} \qquad (1.9)$$

Day and Kerridge (1967) noted that this also holds for a much wider class of models of the form

$$f_i(\underline{x}) = \alpha_i \exp\{-\tfrac{1}{2}(\underline{x} - \underline{\mu}_i)^T\Sigma^{-1}(\underline{x} - \underline{\mu}_i)\}\, \phi(\underline{x}),\ i = 1, \ldots, k, \qquad (1.10)$$

where $\phi(\underline{x})$ is an arbitrary non-negative scalar function, $\underline{\mu}_i$ is the mean vector in population π_i, Σ is the common covariance matrix, and $\{\alpha_i\}$ are normalising constants. Model (1.10) is satisfied by a wide range of distributions, including

1) the Multivariate Normal with equal covariance matrices
2) multivariate independent binary variables

3) multivariate binary variables satisfying the log-linear model with 2nd and higher order effects equal in each population

and

4) a combination of 1) and 3).

The linear logistic model (Day and Kerridge, 1967; Cox, 1970, p. 104),

$$\left.\begin{aligned} p(\pi_i|\underline{x}) &= \exp\{\beta_{i0} + \underline{\beta}_i^T \underline{x}\}.\ p(\pi_k|\underline{x}) \\ p(\pi_k|\underline{x}) &= 1/\{1 + \exp \sum_{s=1}^{k-1} (\beta_{s0} + \underline{\beta}_s^T \underline{x})\} \end{aligned}\right\},\ i = 1,\ldots,k-1 \quad (1.11)$$

where $\underline{\beta}_s^T = (\beta_{s1}, \beta_{s2}, \ldots, \beta_{sd})$ in d dimensions and for $s = 1,..,k-1$, with $(k-1)(d+1)$ parameters, is seen to be equivalent to (1.9), putting $\beta_0 = \alpha_0 + \ln(\theta_1/\theta_2)$ where $\underline{x}$ arises from π_i with prior probability θ_i. The model can be generalised by relaxing the assumption of linearity and incorporating quadratic or interaction terms in $\underline{\beta}_s^T$, as would be required with different dispersion matrices in 1) or for a binary vector $\underline{x}$ following a log-linear model having unequal 1st and 2nd order interactions but equal higher order interactions. Although a full quadratic model involves many more parameters than can be reliably estimated in more than a very few dimensions, especially as iterative maximum likelihood estimation is required (Day and Kerridge, 1967; Anderson, 1972), Anderson (1975) suggested an approximation to $\beta_0 + \underline{\beta}T\underline{x} + \underline{x}^T \Omega \underline{x}$, where $\Omega = \Sigma_i^{-1} - \Sigma_k^{-1}$, based on the spectral decomposition of Ω, which greatly reduces the number of parameters (from $(k-1)(d+1) + (k-1)d(d+1)/2$ to $(k-1)(d+1) + (k-1)d$). Transformations can also easily be incorporated by including functions of $(x_1, x_2, \ldots, x_d)$ in the exponent, making it a flexible and widely applicable model. Various authors have considered logistic type models for ordered dependent and explanatory variables, and these are discussed in Chapter 3.

Rather than assume a particular parametric model for $\{f_i(\underline{x})\}$, estimate its parameters, and derive $\hat{\underline{\beta}}_0$ and $\{\hat{\underline{\beta}}_s\}$ indirectly, logistic regression provides a unified approach for all underlying distributions, finding the MLEs of the parameters in (1.11) directly (so that fewer parameter estimates may be required). Halperin, Blackwelder and Verter (1971) compared the 2 approaches assuming model 1) to hold and found that the results were similar

though direct maximum likelihood tended to produce better fits to the data. On the other hand, surprisingly, Efron (1975) showed that in terms of asymptotic relative efficiency the logistic approach is typically (for root Mahalanobis distance > 2.5, and equal prior probabilities) 1/2 to 2/3 as effective as the LDF approach when multivariate normality with equal covariance matrices holds. O'Neill (1980) extended these results to non-normal continuous distributions, concluding that wherever possible the maximum likelihood approach should be used for the specific distributions at hand.

Complete separation of the samples causes non-uniqueness of the MLEs although such parameter estimates should still yield reasonably good discriminators (Day and Kerridge, 1967), while problems can also arise with small samples if zeroes occur in the one-way margins with qualitative data, causing the procedure to be singular (Anderson, 1974). Anderson and Richardson (1979) reduce the bias in the MLEs and hence in the resulting LDF, as do Moran and Murphy (1979) (see Section 1.4.2).

Day and Kerridge (1967) assumed sampling was from the mixture distribution $\Sigma_i \ \theta_i f_i(\underline{x})$, in unknown proportions, while Anderson (1972) and Anderson and Blair (1982) considered various sampling schemes, showing the model to hold subject to suitable re-interpretation of the parameters. Anderson and Blair also introduce penalised likelihood estimators (see Section 1.4.3) for continuous variables in separate and mixture sampling, from which the parameters $\underline{\beta}$ are as before, β_0 is approximately the same for large samples and the estimate of the mixture density $f(\underline{x})$ turns out to be a spline function. Splines are introduced in the next section.

1.5.2 Nonparametric approaches

A spline ratio estimator

Silverman (1978a) took a completely nonparametric approach to direct estimation of the odds ratio. Given 2 independent samples of size n_1 and n_2 respectively from densities f(x) and g(x), let $N = n_1 + n_2$, $Z_1, \ldots, Z_N$ be the combined order statistic of the data and ϵ_i be an indicator function, $\epsilon_i = 1$ if Z_i arose from f(x) and 0 otherwise. Assuming n_1 and n_2 reasonably large, Silverman showed that a smooth MLE of f/g is found by maximisation of the penalised

conditional log-likelihood

$$\log L(\alpha) = \sum_{i=1}^{N} \left[\epsilon_i \, \alpha(Z_i) - \log \left[1 + \exp\{\alpha(Z_i)\} \right] \right] - \beta \int_I \{\alpha''(x)\}^2 \, dx \qquad (1.12)$$

where $\alpha(x) = \log\left\{ \dfrac{n_2 . f(x)}{n_1 . g(x)} \right\}$, I is an arbitrary interval containing the data and the roughness penalty $\int_I \{\alpha''(x)\}^2 \, dx$ imposes smoothness on α. The degree of smoothness of $\hat{\alpha}$ is determined by β, a smoothness parameter to be chosen. Expression (1.12) is maximised over the class of functions on I which are continuous, have continuous 1st derivative and piecewise continuous 2nd derivative. Silverman states the unique maximiser of (1.12), for a specified β to be the cubic spline s (Section 1.4.3) with knots at the data points $Z_1, \ldots, Z_N$ satisfying the conditions that

1) $s''(Z_1^-) = s'''(Z_1^-) = s''(Z_N^+) = s'''(Z_N^+) = 0$, and

2) for $i = 1, .., N$, $s'''(Z_i^+) - s'''(Z_i^-) = \Phi\{s(Z_i), \epsilon_i\}$ where

$$\Phi\{t, \epsilon\} = \tfrac{1}{2}\beta^{-1}\{\epsilon - e^t/(1 + e^t)\}.$$

Numerical methods are required to solve 2 non-linear equations in 2 unknowns, in order to identify s for any given value of β. As it is given, this method is only applicable in 1 dimension and therefore its usefulness in practice, where multiple feature variables are more common, is limited.

An alternative nonparametric approach was provided by Lauder (1983) who used kernel smoothing to find a direct estimate of $p(\pi|x)$ rather than use kernel density estimation to estimate it by the sampling method. On several real data sets the method was compared to logistic regression and found to be comparable but more conservative in its diagnostic probabilities than the estimative parametric approach, though less so than a predictive logistic method. However, when a large number of variables relative to the

size of the training set were used it was consistently overconfident and in some cases reversed predictions from those of the logistic methods.

1.6 ASSESSING PERFORMANCE

1.6.1 Error rates

Most often, error rate (or equivalently non-error rate) has been used to assess a classification rule and variants of error rate are the most widely used assessment criteria in the literature. These are of two types. Traditionally the emphasis has been on parametric estimation of error rates, assuming multivariate normality with equal covariance matrices (2 populations) and use of the LDF, while more recently this has given way to empirical methods which do not assume a specific distributional form for the data.

Parametric estimators

The optimal error rate is that associated with the optimal (Bayes') classifier when all parameters are known, and will always be unknown. In practice parameters will be estimated, and the error rate associated with use of a sample-based rule, the true/actual/conditional error rate will also depend on unknown parameters. The unconditional or expected error rate is the expectation of the latter over the distribution of all possible parameter estimates. In practice these may be estimated by plug-in estimators, substituting estimates for the unknown parameters in the appropriate expression. For instance, Hills (1966) showed that average actual probability of error and the mean plug-in version of an unbiased estimator will always be greater and less than optimal respectively. Using the sample LDF, Dunn (1971) estimated the overall unconditional probability of correct classification by sample means of both the conditional probability and the plug-in estimator, and found the former to considerably underestimate the true probability for moderately sized samples, while the latter tended to be higher than optimal, confirming the results of Hills (1966). In each case the difference increased with dimension.

Various less biased estimators have been suggested. Lachenbruch and Mickey (1968) proposed a less biased version ("DS") of the usual plug-in estimator ("D") of conditional error, and

2 estimators ("O" and "OS") based on the expansion of expected error of Okamoto (1963), while Lachenbruch (1968) ("L") and Sorum (1971) suggested others. Numerous studies have compared the available methods, mostly as estimators of the conditional error, but also, e.g. Sorum (1972, 1973), Page (1985), of expected and optimal error rates. No one estimator is uniformly superior though Lachenbruch and Mickey (1968), in simulations, found D consistently poor while overall OS was superior to both O and DS, particularly as dimensionality increased, in terms of absolute deviation from the true error. McLachlan (1974a) proposed an asymptotically unbiased technique ("M"), with lower actual bias than OS, and requiring less computation than the leaving-one-out method (see below) (though see Fukunaga and Kessell (1971), and Habbema, Hermans and van der Burgt (1974), for the multivariate normal case and also the kernel method) and found it to be comparable, in terms of asymptotic mean square error, and average absolute deviation from the actual error over simulations, to Okamoto's estimator. McLachlan's method tended to underestimate actual error. McLachlan (1974b) using the asymptotic unconditional mean squared error criterion broadly confirmed the results of Lachenbruch and Mickey (1968) and also those of McLachlan (1974a), finding M to be comparable to or better than OS especially as dimension increased, and M and OS always and generally better than L respectively. Page (1985) extended the previous studies by evaluating a range of techniques as estimators of both conditional and optimal error, in terms of absolute deviation from the appropriate error rate, and for conditional error obtained similar conclusions, though the best estimators of each error rate differed. Snapinn and Knoke (1984) considered non-normal populations as well, and for small samples compared D and DS with the nonparametric apparent and leaving-one-out estimators described below, using unconditional mean square error.

Empirical estimators

In practice, empirical methods are probably more useful and can be superior even when normality does hold (Snapinn and Knoke, 1984). Of these, the apparent or resubstitution error rate or proportion of training cases misclassified, is well recognised as being an over-optimistic estimate of future performance (Hills, 1966; Lachenbruch and Mickey, 1968). A less biased estimate may be

obtained either by the holdout method, using a test set (a further set obtained from past records, ongoing data collection, or, where the sample is sufficiently large, by randomly splitting the available data into a training and test set before constructing the discriminant rule) or by the leaving-one-out method of Lachenbruch and Mickey (1968) where the rule is obtained on each set of n-1 observations in turn to allocate the remaining case. This is a cross-validatory assessment in the sense of Stone (1974a) using the (0,1) loss function. The former, although unbiased, can have large variance unless the sample is large (Hand, 1986a), while the latter is nearly unbiased but can also have high variability, especially for small data sets (Efron, 1983). In fact Toussaint (1974) quotes Glick as showing that for discrete distributions the leaving-one-out method has much greater variablility than the apparent error rate. Between these two extremes is the V-fold cross-validation method (see, for instance Breiman et. al., 1984, p. 12) which divides the available data into V subsets of as nearly equal size as possible, and classifies the cases in each subset on the basis of the rule constructed on the remaining n-V cases. The average of the V resulting error rates is then taken as the estimate. As all n observations are used to construct the final classifier, the procedure might be criticised in that the estimated error rate may not be that of the final classification rule. Efron (1983) gives some improvements on cross-validation, based on the bootstrap method of correcting the apparent error rate for bias (Efron, 1979), as do Chernick, Murthy and Nealy (1985). The smoothed error rate estimators of Glick (1978) and Tutz (1985) reduce variance as well as bias, and one such estimator was found by Snapinn and Knoke (1985) to be considerably superior to the resubstitution error and frequently also to leaving-one-out and an ideal bootstrap, in terms of unconditional mean squared error, for samples from both normal and non-normal populations.

Toussaint (1974) gives a bibliography of misclassification rate estimation, while Hand (1986a) and McLachlan (1986) review subsequent developments, thereby concentrating on empirical techniques.

1.6.2 Reliability measures

There are, however, 2 aspects to performance of a discriminant rule :

1) group separation, concerned with how accurately the rule classifies, and
2) reliability, or calibration, concerned with good estimation of the posterior probabilities.

Plainly, no matter which error rate is used, it only measures separation, not the actual value of the estimated probabilities, with the result that a near miss and an extremely poor prediction are treated equally, as are a correct confident, and a correct but diffident prediction. Often we will require not only that a rule classifies well, but also that the predicted probabilities are reliable i.e. realistic. To some extent, the degree of confidence or "sharpness" of a method may be assessed from error rates by introducing a category of doubt to which cases are assigned whose highest predicted probability lies below a specified threshhold value. Alternatively, the average probability assigned to the correct population of the test cases may be used. Knoke (1986) notes that the latter may be considered as a smoothed error rate estimator. Habbema, Hilden and Bjerregaard (1978, 1981), Habbema and Hilden (1981) and Hilden et al. (1978a, 1978b) consider various aspects of performance and present a number of functions of the assigned probabilities which are sensitive to calibration as well as separation. Particular continuous scores discussed with respect to reliability are the Brier or quadratic score of Brier (1950), the average logarithmic score and ϵ-modified logarithmic score, all of which make use of the actual values of the predicted probabilities. For example, for k = 2, if $\hat{p}(\pi_1|x)$ and $\hat{p}(\pi_2|x)$ are 0.8 and 0.2 respectively, for a case originating from π_1, then its contribution to the Brier score is $(1-.8)^2 + (.2)^2$. Averaging contributions over the sample, the formal definition of the Brier score is

$$S_B = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left\{ [1-p(\pi_j|x_{ji})]^2 + \sum_{\substack{l=1 \\ l\neq j}}^{k} [p(\pi_l|x_{ji})]^2 \right\}, \qquad (1.13)$$

where x_{ji} is the ith observation from population j, $i = 1,\ldots, n_j$, $j = 1,\ldots, k$, $\Sigma\, n_j = n$. Each contribution takes a minimum value of 0 and maximum of 2, when probabilities of 1 and 0 respectively are assigned to the correct group. Hence the score itself can vary over this range, with low values indicating good performance.

The average log score is an alternative to the Brier score and is defined as

$$S_L = \frac{1}{n} \sum_j \sum_i \log [p(\pi_j|x_{ji})] \qquad (1.14)$$

$\in (-\infty, 0]$, with a best possible value of 0. A disadvantage of S_L is that a few very poor predictions will result in an extremely low score due to the properties of the log function.

The ϵ-modified log score S_{EL} essentially sets a lower bound to the contribution which any one observation can make to S_L and is defined as

$$\frac{1}{n} \sum_j \sum_i \{ \log w\,(p(\pi_j|x_{ji})) + \epsilon \sum_{l \neq j} \log [w(p(\pi_l|x_{ji}))/\epsilon] \} \qquad (1.15)$$

$\in [\,(1-\epsilon) \log \epsilon, 0]$, where $w(z) = (1-\epsilon)z + \epsilon$ for $0 \leqslant z \leqslant 1$, so that $\epsilon \leqslant w(z) \leqslant 1$ and, say, $\epsilon = 0.01$. It can be shown that

$$S_{EL} \approx \frac{1}{n} \sum_j \sum_i \log [p(\pi_j|x_{ji}) + \epsilon].$$

Unlike S_L, S_{EL} penalises very small and zero values roughly equally.

Again the leaving-one-out method may be employed with any of these scores if a test set is unavailable.

Blattenberger and Lad (1985) show formally that in certain contexts the Brier score can be separated into 2 distinct parts reflecting the separation and calibration aspects of performance. In theory, reliability could also be assessed by ongoing data collection, by comparing the predicted posterior probabilities for a specified feature vector $\underline{x}$ with the proportion of cases with those observed indicants $\underline{x}$ and confirmed as belonging to each population. Where some of the feature variables are continuous, the probability of a particular observation recurring is of course zero and grouping of some sort would be necessary. A "pure" reliability score of this type which is practicable and can be derived from the test set is given by Titterington et al. (1981), p. 154.

Finally, although these measures do make use of the actual estimated probability values, they still do not take account of the relative seriousness of different types of error. When this is important and a loss matrix can be specified a priori, an average loss measure will also be a useful means of assessment.

1.7 VARIABLE SELECTION

Given a large set of potential predictors, correlations may render some variables redundant and reliable parameter estimation, reduction of data collection costs and computational difficulties may necessitate some further reduction of dimensionality while hopefully retaining most of the information in the original variable set.

Ideally one would compare the classification rules corresponding to all possible subsets of predictors in terms of a given criterion. In practice this will only be feasible for a moderate number of variables, say 15 at most. Otherwise (suboptimal) stepwise procedures (forward selection, backward elimination or a combination of the two), or intermediate accelerated search procedures (Hand, 1981b) will be necessary .

Traditionally measures of distance between populations have been used to assess how well a given set of variables separates them, for instance the F-statistic or related measures such as Wilks' Λ or Mahalanobis distance, appropriate for testing equality of population means assuming multivariate normality with equal Σ_i, (e.g. Cochran, 1964; Rencher and Larson, 1980) and allow tests of sufficiency of a specified set of variables (Rao, 1965, pp. 467-470 and p. 482). Due to the analogy of discriminant analysis with regression (Lachenbruch, 1975a, pp. 17-19), R^2, the squared multiple correlation of predictor and response variables, or equivalently residual sum of squares (Draper and Smith, 1966, Chapter 6), have also been used (e.g. Weiner and Dunn, 1966). More recently Krzanowski (1983b) used a stepwise procedure based on a distance measure for mixed variables (Krzanowski, 1983a). For discrete variables, based on work of Glick (1973) which suggested that large values of the distance

$D = \min_{\underline{x}} | \sqrt{g_1(\underline{x})} - \sqrt{g_2(\underline{x})} |$, where $g_j(\underline{x})$ is the jth discriminant score, were associated with low misclassification rate, Goldstein and Rabinowitz (1975) proposed selecting variables in 2 population discrete classification by maximising sample-based versions of D over all possible subsets:

$$\max_j \max_{1 \leq i \leq \binom{d}{j}} \min_{\underline{x}_i} | n_1(\underline{x}_i)^{1/2} - n_2(\underline{x}_i)^{1/2} | \qquad (1.16)$$

where $\underline{x}_i$ denotes the ith subset of j variables. Variants of (1.16) to correct for dimension were also discussed. Also for discrimination, Hills' (1967) based a stepwise procedure on the discrete Kullback-Leibler divergence

$$\sum_{\underline{x}} \{p_1(\underline{x}) - p_2(\underline{x})\} \ln \frac{p_1(\underline{x})}{p_2(\underline{x})}.$$

Goldstein and Dillon (1978, Chapter 4) discuss several other approaches for multinomial data.

Kittler (1975) discusses numerous distance measures closely related to the Bayes' error rate including the Kullback-Leibler divergence and also the Kolmogorov variational distance, but notes that most rely on multivariate integration and are therefore generally not computationally feasible as means of feature selection. Different evaluation criteria may lead to different variable subsets being chosen. Habbema and Hermans (1977) therefore argued in favour of basing variable selection in discriminant analysis on error rate or expected loss, preferably using a leaving-one-out method, as, unlike measures of separation such as the F-statistic, these are directly related to the eventual aim of the procedure, and also provide a natural stopping rule. They note in passing that criteria taking account of the actual estimated posterior probabilities would be preferable to error rate. Habbema et al. (1981) refer to such a variable selection program using the continuous scores of Section 1.6.

McLachlan (1976) who considered multivariate normality with equal covariance matrices and use of the LDF, proposed the use of an asymptotically unbiased estimator of the difference in conditional error rate by deletion of a subset of variables, and derived its asymptotic distribution, thus associating a confidence level with the increase in error rate.

While theoretical error rate declines with the inclusion of each extra variable, in practice with finite sample sizes inclusion of additional variables may not add discriminatory power. Murray (1977a) cautioned against the indiscriminate use of stepwise procedures based on apparent error rate, noting that if the best subset of each size is selected, although error rates do intitially decrease with increasing dimension they will eventually rise again so that a smaller subset can give superior discrimination results.

This was explained in terms of bias due to selection of the subset which performs best on the given data set, the bias being greatest for moderately sized variable sets. Apparent error rate was found to be a hopelessly misleading estimate of the true error. Including "dummy" variables (i.e. variables known to be uninformative) in the original variable set may be helpful in indicating at which point the subset chosen becomes "too large" (Hill, 1979, pp. 8, 25 and 37).

Habbema, Hermans and van den Broek (1974) used a stepwise procedure with product kernel density estimation (see Section 2.1.3), with an expected loss criterion, estimated by leaving-one-out. Pfeiffer (1985) also used a stepwise kernel-based procedure but his criterion considered for a given point $\underline{x}$ an average measure based on the ratio of the estimated correct class conditional p.d.f. to the maximum of those of the remaining classes, claiming that while distance measures may be appropriate when multivariate normality with equal covariance matrices is assumed, for nonparametric discrimination functions of the p.d.f.s are more useful. Pfeiffer considered that error rate or loss measures are too insensitive and may not be able to distinguish between several subsets of a given order.

Finally, the choice of variables can influence performance more than the choice of method or model employed to construct the classification rule (Titterington et al., 1981). Given this, it is natural to require that the criteria by which variable selection is made be related to both the context and aims of discrimination (Pfeiffer, 1985; Habbema and Hermans, 1977). Although variable selection is clearly an important aspect of discriminant analysis, it will not be considered any further here.

CHAPTER 2 CONTINUOUS KERNEL DENSITY ESTIMATION IN DISCRIMINANT ANALYSIS

In this chapter we consider use of the kernel method to estimate the class conditional densities as required by the sampling or indirect approach.

2.1 FIXED KERNELS

2.1.1 Introduction

There is a vast literature on kernel density estimation. Historically, the idea originated with Fix and Hodges' (1951) "running histogram", also considered by Rosenblatt (1956).

Viewed as a density estimate, the conventional histogram has a number of disadvantages (see, for instance, Tarter and Kronmal, 1976):

1) Its appearance depends crucially on the choice of classes or bins, both on choice of origin and, more particularly, on bin width.

2) The fixed nature of the cell structure means that the estimated density at two points a given distance apart may be very different depending on whether or not they lie in the same cell or in adjacent cells.

3) Although the underlying density will often be assumed smooth, the histogram is a step function - a particular problem if derivatives are required.

4) The estimate is zero outside of a given range.

The "running histogram" is defined, for a univariate sample $X_1, \ldots, X_n$, as

$$\hat{f}(x) = \frac{1}{2h} [\, F_n(x + h) - F_n(x - h) \,]$$

where $F_n(x)$ is the empirical distribution function, and derives naturally from the fact that $f(x) = F'(x)$ where F is the cumulative distribution function (c.d.f.). Silverman (1986, pp. 11-12) calls this the "naive estimator".

Equivalently, $\hat{f}(x) = \frac{1}{2hn}$ {no. of X_i in $[x - h, x + h]$}

$$= \frac{1}{nh} \sum_{i=1}^{n} W\left[\frac{x-X_i}{h}\right] \quad \text{where } W(u) = \begin{cases} \tfrac{1}{2} & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases} \qquad (2.1)$$

and hence is seen to be an average of "boxes" of width 2h and height 1/2h centred at each data point X_i, so that the points within distance h of x contribute 1/2h, the remainder contributing zero. Compared to the usual histogram where the estimate at x is given by the height (at the midpoint) of the cell containing x, x itself becomes the centre of a cell of width 2h and the estimate the height of the histogram at x. Difficulty 2) has therefore been overcome, as has the choice of origin. The estimate is still discontinuous and h remains to be chosen.

2.1.2 Univariate kernels

The kernel estimator of Parzen (1962) and Rosenblatt (1956) generalises the naive estimator by replacing W by a function K, where K is symmetric and

$\int K(u)\, du = 1$. If K(u) is defined on the whole real line (and decreases monotonically with increasing |u|), all data points now contribute to the estimate, and the problem of the lack of tails is removed. Since

$$\hat{f}(x) = \frac{1}{nh} \sum_{i} K\left[\frac{x-X_i}{h}\right], \qquad (2.2)$$

it inherits the properties of K and so K is usually chosen to be a smooth p.d.f. with the result that $\hat{f}(x)$ is now smooth and also a p.d.f.. However non-negativity is not essential, and allowing K to take negative values may allow bias reduction (Bartlett, 1963; Schucany and Sommers, 1977), as may relaxing the requirement that the kernel integrates to 1 (Terrell and Scott, 1980).

If X is a univariate, continuous random variable and if we have a sample $X_1, \ldots, X_n$, a p.d.f. is centred on each observation and the resulting estimate at any point x is given by averaging the contributions of these p.d.f.s over the sample, so that points close to x contribute more than those further away. See Figures

2.1(a) and (b).

K is called the "kernel" function and h, the "smoothing parameter", "window width" or "band width", determines the width of K and hence the degree of smoothness of the resulting estimate. Both the form of K and size of h remain to be chosen.

Epanechnikov (1969) derived the kernel which asymptotically minimises mean integrated square error (MISE)

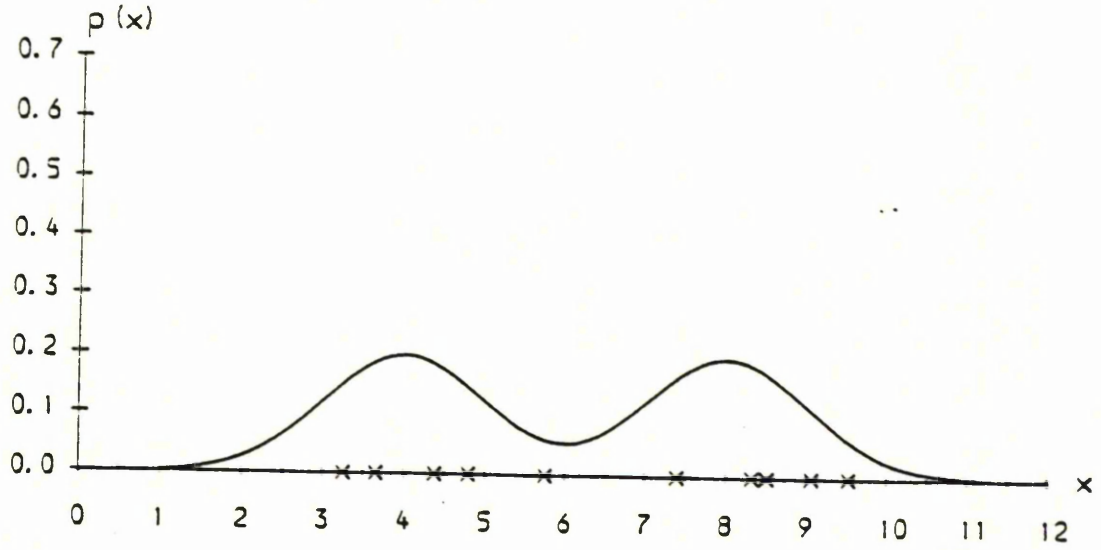
$$K(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left[\frac{1-t^2}{5}\right], & |t| \leq \sqrt{5} \\ 0, & \text{otherwise} \end{cases} \qquad (2.3)$$

However, on grounds of relative asymptotic MISE efficiency there is little to choose between kernels (even the naive estimator's rectangular kernel), and Silverman (1986, p. 43) recommends choice of kernel on considerations of ease and speed of computation (especially important with multivariate data) or for desirable differentiability properties. For speed, kernels of finite support, such as the Epanechnikov kernel itself, are appropriate but at the expense of the resulting estimate having no tails. In general the conclusion in the literature is that the form of K is relatively unimportant. For instance, van Ness and Simpson (1976) in a study described in more detail in Section 2.3.3, compared use of normal and Cauchy kernels and found their performance to be very similar. Consequently K is usually chosen for mathematical ease and is most often taken to be Normal with h as its standard deviation.

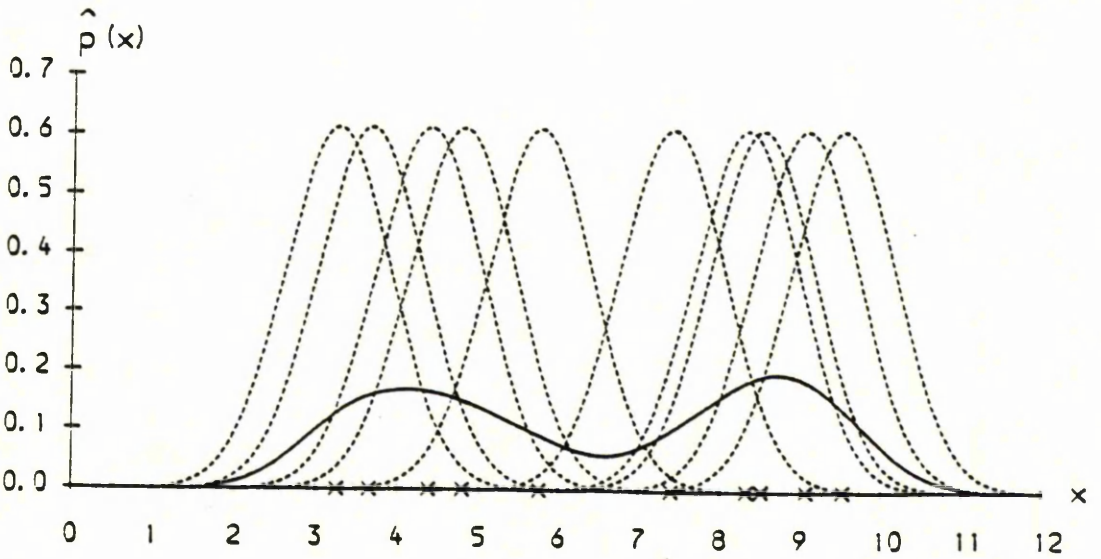
By contrast, the choice of h is crucial, as can be seen from Figures 2.1 (c)-(e). As $h \to 0$ the estimate tends to a series of infinite spikes at the data points. At the other extreme, as $h \to \infty$ the resulting estimate becomes increasingly flat and the effect of the data is lost altogether, although in general oversmoothing will be less serious than the reverse (Fryer, 1976). We consider in Section 2.3.2 how to choose an appropriate value of h between these two extremes.

The asymptotic behaviour of the kernel estimator is well documented. For instance while results of Rosenblatt (1956) and Yamato (1972) show that for finite samples and non-negative kernels the kernel estimator is biased for any choice of density f(x),

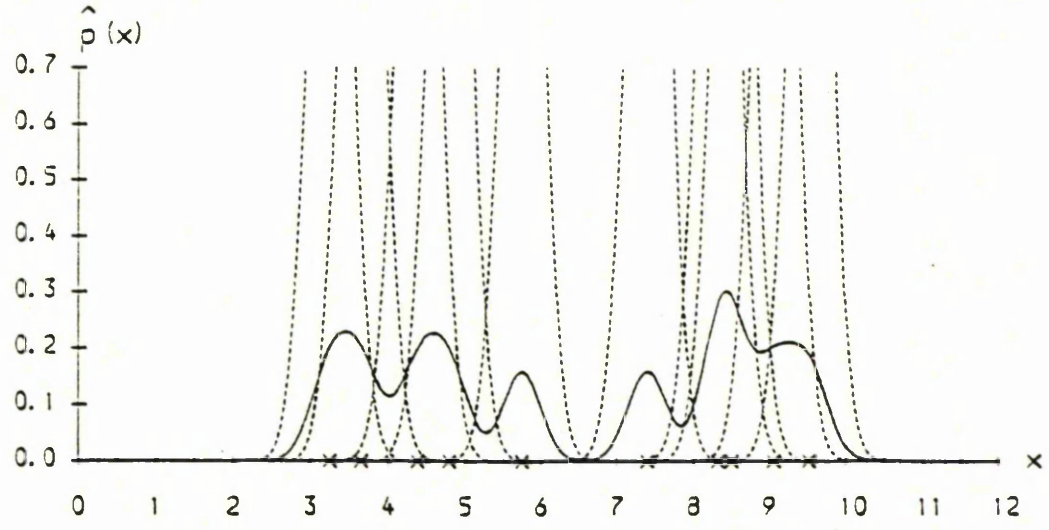
Figure 2.1 Fitting a kernel density estimate to a random sample of 10 observations from the bimodal Normal mixture ½ N(4, 1) + ½ N(8, 1) using a Normal kernel function.



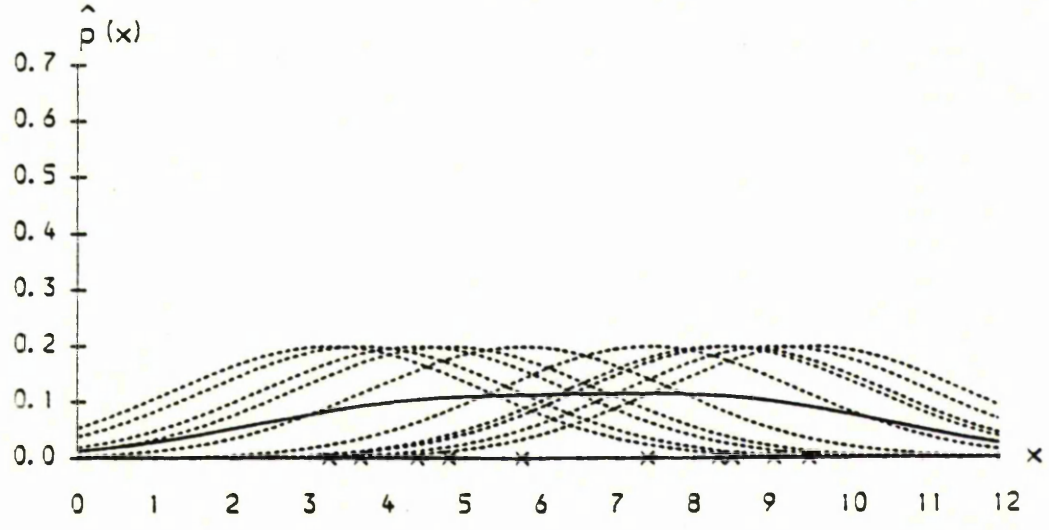
(a) the true mixture density.



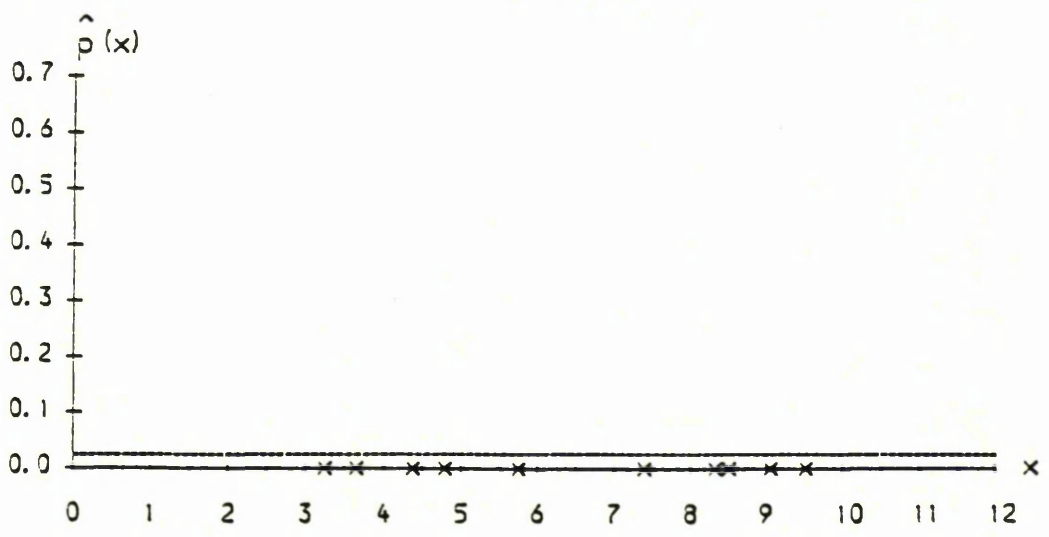
(b) h = .65 - an appropriate amount of smoothing which retains the bimodality.



(c) h = .25 - too little smoothing.



(d) h = 2.0 - oversmoothing loses the bimodality.



(e) h = 15.0 - effectively infinite smoothing.

Parzen (1962) showed that subject to certain conditions (1.6) on K and f, if lim h = 0 as n → ∞ then $\hat{f}_n(x)$ is asymptotically unbiased at all points of continuity of f(x), consistent and asymptotically normal. Van Ryzin (1969) states conditions under which strong consistency obtains. Fryer (1977) reviews various methods of density estimation including the kernel method, as does Wegman (1972), with Wertz and Schneider (1979) providing a more complete bibliography. More recently, Hand (1982) summarises some of the more important mathematical results while Silverman (1986) provides a less mathematical overview of the subject.

2.1.3 Multivariate kernels

The kernel method extends straightforwardly for multivariate data. At its most general the estimate in d dimensions is given by

$$\hat{f}(\underline{x}) \propto \frac{1}{n} \sum_{i=1}^{n} K(\underline{x} - \underline{X}_i ; S)$$

where K is a d-variate kernel function and S is a d×d array of smoothing parameters. Usually S is taken to be diagonal giving, for suitable choice of K,

$$\hat{f}(\underline{x}) = \frac{1}{n\, h_1 \ldots\ldots h_d} \sum_{i=1}^{n} K\left[\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_d - X_{id}}{h_d}\right], \qquad (2.4)$$

or for common smoothing parameters, $h_1 = \ldots\ldots = h_d = h$ say,

$$\hat{f}(\underline{x}) = \frac{1}{n\, h^d} \sum_{i=1}^{n} K\left[\frac{\underline{x} - \underline{X}_i}{h}\right]. \qquad (2.5)$$

A further simplification of (2.4) and (2.5), immediate for certain forms of K, gives

$$\hat{f}(\underline{x}) = \frac{1}{n \prod_j h_j} \sum_{i=1}^{n} \left\{ \prod_j K\left[\frac{x_j - X_{ij}}{h_j}\right] \right\}, \qquad (2.6)$$

and

$$\hat{f}(\underline{x}) = \frac{1}{n\, h^d} \sum_{i=1}^{n} \left\{ \prod_j K\left[\frac{x_j - X_{ij}}{h}\right] \right\} \qquad (2.7)$$

respectively. Kernels of this type are known as "product kernels". They can be generalised by allowing the form of the marginal kernel to vary, as is appropriate for discrete or mixed data, giving

$$\hat{f}(\underline{x}) = \frac{1}{n \prod_j h_j} \sum_{i=1}^{n} \left\{ \prod_j K_j \left[\frac{x_j - X_{ij}}{h_j} \right] \right\}. \qquad (2.8)$$

Continuous data may be standardised before applying a symmetric kernel and then transformed back (Habbema, Hermans and van den Broek, 1974), allowing use of a single smoothing parameter and hence forms (2.5) or (2.7). Similarly, Fukunaga, (1972, p. 175), suggests linear transformation of multivariate data to a unit covariance matrix prior to use of a radially symmetric kernel.

Cacoullos (1966) extended the results of Parzen (1962) for multivariate kernels and product kernels of the form (2.5) and (2.6), also commenting that the product kernel has stronger invariance properties, as (2.6) is invariant under different scale transformations of each variable while (2.5) requires a common scale transformation for invariance. Apart from its simplicity, this may partly explain why in practice the product kernel is by far the most commonly used for continuous data, as does ease of parameter estimation since marginal or univariate methods may be used. Usually the K_j are of the same parametric form, most often normal. Epanechnikov's kernel (2.3) minimises the asymptotic MISE when product kernels of the form (2.7) are used. Sacks and Ylvisaker (1981) found the kernel of type (2.5) minimising the asymptotic mean squared error (MSE) at a point $\underline{x}$ (which they take to be $\underline{0}$). Studying the asymptotic relative efficiency of product kernels, they found Rosenblatt's uniform kernel to be inferior to that of Epanechnikov but concluded that standard kernels were not far from optimal provided that the point of estimation is an interior point of the support of f. Otherwise standard kernels introduced a large bias term. This would suggest that in high-dimensional spaces, where the tails exert more influence, the product kernel may be inappropriate. There is also some evidence from a study by Remme, Habbema and Hermans (1980) that product kernels may not be optimal when used on correlated data. They found, in a study described in more detail in Section 2.3.3, using normal product kernels on multivariate normal data with correlated variables, that a slight decrease in performance was observed with

correlations from 0.4 to 0.5 but that in 6 dimensions and with correlations over 0.6 a more serious decrease in performance occurred as compared with results for uncorrelated variables.

Van Ness and Simpson (1976) compared normal product kernels with a multivariate Cauchy kernel and found little difference between them but the data were from Multivariate Normal distributions with uncorrelated variables of unit variance.

Murphy and Moran (1986) studied sensitivity of the product kernel method to correlation between variables, comparing it to the usual estimative LDF, an unbiased LDF (Moran and Murphy, 1979) and the predictive LDF in terms of classification and estimation of the log-odds. Equal sample sizes and equi-correlated variables were used with similar configurations to those of van Ness and Simpson (1976). (See Section 2.3.3). A single smoothing parameter was chosen to minimise MISE exactly. They confirmed the superiority of the kernel method, found by van Ness and Simpson, for high dimensions (> 8) and uncorrelated variables, though this was not as marked, and unlike the previous study classification with the kernel method was inferior to the parametric method for all cases for dimensions ≤ 8, nor did error rates follow those of the optimal LDF using the known covariance matrix. For moderately positively correlated variables the product kernel was still superior to the best parametric method in terms of both classification and estimation of log-odds, but with other correlation structures could behave erratically.

2.2 ADAPTIVE METHODS

In our discussion so far the width of the kernel centred on $\underline{X}_i$, $i = 1,.., n$ has been determined by a constant h, independent of both the point of estimation, $\underline{x}$, and the location of the sample points. Such kernels are called "fixed". As a consequence of h being fixed there is a tendency to oversmooth in areas of high density and undersmooth where the data are sparse. This is a particular problem for long-tailed distributions, especially in more than 1 dimension.

2.2.1 Nearest-neighbour estimators.

Instead of fixing the window width h, Loftsgaarden and Quesenberry (1965) fixed the number of points, k, spanned by the kernel. If $r_k(\underline{x})$ is the kth largest Euclidean distance of $\underline{x}$ from

the n data points and $V_k(\underline{x})$ is the volume of the d-dimensional hypersphere of radius $r_k(\underline{x})$ centred on $\underline{x}$, then the "kth-nearest-neighbour" estimator is derived by equating the expected number of observations in the hypersphere, which as $r_k(\underline{x}) \to 0$ is given by $n\hat{f}(\underline{x})\, V_k(\underline{x})$, with the observed number k. We then have

$$\hat{f}_n(\underline{x}) = k / nV_k(\underline{x}).$$

Equivalently, $$\hat{f}(\underline{x}) = \frac{k}{n\, c_d\, r_k^d(\underline{x})} \qquad (2.9)$$

where c_d is the volume of the unit sphere in d dimensions. Loftsgaarden and Quesenberry prove the estimator to be consistent. The role of the smoothing parameter is now taken by k, though unlike h its value is not critical. It is chosen to be rather less than n itself and the authors recommended $k \approx n^{1/2}$. The estimator can be seen to adapt the degree of smoothing to an initial estimate of the local density of the data, since for a point $\underline{x}$ in the tails of the distribution the distance $r_k(\underline{x})$ to its kth-nearest neighbour will be larger than for a point in an area of high density, but for a given point $\underline{x}$ the estimator is essentially of the fixed kernel type.

Hand (1982, p. 235) comments that by using the distance from $\underline{x}$ to a single point as the radius this estimator may be expected to have a larger standard deviation than the kernel estimator which uses an averaging process. Mack and Rosenblatt (1979) showed that nearest-neighbour estimators reduce noise in the tails at the expense of introducing considerable bias, which will then dominate global goodness-of-fit measures such as MISE (Rosenblatt, 1979). Bias is caused by the inverse variation of $\hat{f}(\underline{x})$ with $r_k^d(\underline{x})$ which increases as $|\underline{x}|$ increases, no matter what the tail behaviour of the data. This slow rate of decay, dependent on $\underline{x}$ itself, also causes the estimate to integrate to ∞. Though a continuous positive function, it is not therefore a p.d.f., and its derivative is discontinuous at every point of discontinuity of the derivative of $r_k(\underline{x})$.

The generalised kth-nearest-neighbour estimator (Moore and Yackel, 1977) is defined as

$$\hat{f}(\underline{x}) = \frac{1}{nr_k^d(\underline{x})} \sum_i K\left[\frac{\underline{x} - \underline{X}_i}{r_k(\underline{x})}\right], \qquad (2.10)$$

which is the kernel estimator at $\underline{x}$ with window width $r_k(\underline{x})$, i.e. kernels of common width $r_k(\underline{x})$ are used to obtain the estimate at $\underline{x}$, the width depending on $\underline{x}$ itself and the density of data near $\underline{x}$. The overall degree of smoothing is controlled by k. Using a uniform kernel, (2.10) reduces to (2.9), the relationship being analagous to that between the kernel and naive estimator. It can therefore be seen that the estimate at a single point for any order of nearest-neighbour, k, will be identical to that for a certain value of h given by the fixed kernel method. For more than one point the values would not coincide. The tail behaviour of the generalised estimator depends on the kernel used, and Moore and Yackel (1977) prove consistency properties, but again $\hat{f}(\underline{x})$ will not generally integrate to 1. In view of these difficulties the nearest-neighbour estimator is not ideal for an overall density estimate.

2.2.2 Variable kernels

The "variable" kernel estimator of Breiman, Meisel and Purcell (1977) combines the smoothness properties of the fixed kernel method with the adaptive properties of the nearest-neighbour estimator. It is defined as

$$\hat{f}(\underline{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{hd_{ik}} K\left[\frac{\underline{x}-\underline{X}_i}{hd_{ik}}\right]$$

where d_{ik} is the distance from $\underline{X}_i$ to its kth-nearest neighbour. Thus, a greater amount of smoothing is done in regions of sparse data, where d_{ik} is large, while more peaked kernels are associated with high density areas as d_{ik} decreases. For given k, the overall level of smoothing depends on h, while k determines the degree of responsiveness to local density. Unlike the nearest-neighbour method, the window width is independent of the point of estimation, $\underline{x}$, depending only on the distance between the data points themselves, and so, provided K is a p.d.f., $\hat{f}(\underline{x})$ also integrates to 1, and it inherits the smoothness properties of K. The variable kernel is therefore to be preferred to the nearest-neighbour

method. There are now 2 parameters to determine, h and k. Breiman et al. (1977) used a grid search method to optimise a goodness-of-fit statistic in terms of h and k, as did Raatgever and Duin (1978), both finding that the value of k was not critical. Habbema, Hermans and Remme (1978) used a two-stage procedure, reaching the same conclusion.

Silverman (1986, pp. 100-106) unifies the fixed and variable kernel methods and the method of Abramson (1982) (see below) into the "adaptive" method. A pilot estimate of $f(\underline{x})$ is found such that $\tilde{f}(\underline{X}_i) > 0$, $\forall i$. "Local bandwidth factors" $\{\lambda_i\}$,

$$\lambda_i = \left[\frac{\tilde{f}(\underline{X}_i)}{g}\right]^{-\alpha},$$

are defined where α, the "sensitivity parameter", lies in the range $0 \leqslant \alpha \leqslant 1$ and g is the geometric mean of the $\tilde{f}(\underline{X}_i)$, i.e.

$$g = \left[\prod_i \tilde{f}(\underline{X}_i)\right]^{1/n}.$$

The adaptive kernel estimator in d dimensions is then given by

$$\hat{f}(\underline{x}) = \frac{1}{n}\sum_i \frac{1}{(h\lambda_i)^d} K\left[\frac{\underline{x}-\underline{X}_i}{h\lambda_i}\right].$$

As usual h is the bandwidth and K a symmetric kernel integrating to 1. Setting α to 0 and d^{-1} give the fixed and variable kernel methods respectively. The overall degree of smoothing is determined by h, while the larger α is the more sensitive the λ_i will be to variations in the pilot density. Including g^{α} explicitly frees the λ_i from the scale of the data and forces their geometric mean to equal 1. As with the fixed and variable kernel methods, provided K is non-negative the estimator is a p.d.f., it inherits the smoothness properties of K and does not suffer from heavy tails. It is relatively insensitive to the pilot estimate (Breiman, Meisel and Purcell, 1977; Abramson, 1982) so that a quick, simple method such as the fixed kernel method with an Epanechnikov type kernel and deterministic choice of h is appropriate for the pilot stage. Silverman (1986, pp. 106-110) found the adaptive method to improve on the fixed method both in the tails and main part of the density. His examples suggest that

while α should be rather less than 1, the variable kernel ($\alpha = 1/d$) adapts too much to the pilot estimate. The method of Abramson (1982) uses $\alpha = \frac{1}{2}$ based on a result that if the λ_i were exactly equal to $f(X_i)^{-\frac{1}{2}}$ the asymptotic bias reduces to order h^4 rather than h^2 for the fixed kernel. No other value of α will achieve this. Abramson also found $\alpha = \frac{1}{2}$ to give good results in practice and Silverman's results confirmed this. As usual, h remains to be chosen.

Although adaptive methods can therefore be expected to give a better estimate of a continuous density than the fixed kernel method, for reasons explained in Section 2.7.1 we do not consider their use any further.

2.3 CHOICE OF SMOOTHING PARAMETER IN DENSITY ESTIMATION

Much of the literature concentrates on establishing mathematical conditions on the sequence $\{h_n\}$ as $n \to \infty$ to produce certain asymptotic behaviour. For a given finite sample size n, however, these are unhelpful in suggesting a suitable value for the smoothing parameter. The optimal choice of h depends on the smoothness of the density itself, which is of course unknown.

2.3.1 Subjective criteria

Where a density estimate is required for purposes of exploratory data analysis a subjective choice of h may be adequate, one method being to plot the density estimate for various values of h, choosing a value that is judged, in accordance with prior ideas, neither to over- nor undersmooth the data (see e.g. Scott, Tapia and Thompson, 1977). Silverman (1978b) suggested the "test graph" method, based on a theoretical result concerning the rate of uniform convergence. This involves plotting the graph of the estimated second derivative $\hat{f}''(x)$ for various h and choosing a value producing "rapid fluctuations which are quite marked but do not obscure the systematic variation completely". Both of these methods would be difficult to apply successfully in more than 1 dimension.

2.3.2 Automatic choice of h

Automatic procedures are based on optimisation of criteria of goodness-of-fit, whether asymptotic or exact. Some require numerical optimisation or iterative procedures, others yield

deterministic formulae, while others require cross-validation to yield a non-trivial solution.

An asymptotic method

While pointwise goodness-of-fit properties are of interest, global measures are more intuitively appealing. The most tractable and commonly used of these is Mean Integrated Square Error (MISE),

$$\int E\{[f(x) - \hat{f}(x)]^2\}\, dx,$$

where the expectation is with respect to the true density f(x). Minimisation of the Taylor expansion of MISE, from Parzen (1962) gives the asymptotically optimal smoothing parameter

$$h = \left[\int t^2 K(t)\, dt\right]^{-2/5}\left[\int K^2(t)\, dt\right]^{1/5}\left[\int [f''(x)]^2 dx\right]^{-1/5} n^{-1/5} \qquad (2.11)$$

where $\int [f''(x)]^2\, dx$ is a measure of smoothness of the true density.

An obvious way to proceed is to use a pilot estimate of h, perhaps chosen subjectively, to construct an estimate of the measure of smoothness which is then substituted into (2.11). (Nadaraya, 1974; Woodroofe, 1970). Scott, Tapia and Thompson (1977) extended this to an iterative procedure where h_{i+1} =

$$\left[\int t^2 K(t)\, dt\right]^{-2/5}\left[\int K^2(t)\, dt\right]^{1/5}\left[\int [f_n''(x, h_i)]^2 dx\right]^{-1/5} n^{-1/5} \qquad (2.12)$$

Using normal kernels an analytic expression may be found for (2.12). While h = 0 is always a solution, choosing h_0 as the sample range guarantees convergence to the largest solution such that $h \geq 0$.

Bowman (1981) compared the resulting solution for h with values obtained by direct minimisation of MISE for mixtures of Normal distributions and for two sample sizes, finding that while the asymptotically optimal choice was always smaller, "the difference was not serious".

Deterministic methods

Using normal data and normal kernels, Fryer (1976) minimised

the exact MISE, which is a function of h, n and σ, the population standard deviation, by plotting MISE/σ as a function of log(n) and log(h/σ), for sample sizes $0 \leq n \leq 22{,}000$. Approximating the line of best fit through the contours gave the simple formula

$$h \approx 1.31\, n^{-0.205}\, \sigma \qquad (2.13)$$

He also studied mixtures of Normals, finding that the optimal degree of smoothing was close to that given by (2.13) for both bimodal and skew distributions, as well as being insensitive to kurtosis.

Bowman (1981) modified Fryer's formula for smaller sample sizes, $20 \leq n \leq 100$, giving

$$h \approx 1.261\, n^{-0.226}\, \sigma \qquad (2.14)$$

and used Hogg's median absolute deviation estimator

$$\hat{\sigma} = \text{median}\,[\, |X_i - \text{median}\{X_i\}| \,/0.6745] \qquad (2.15)$$

as a robust estimator for σ to guard against outliers and heavy-tailed distributions. (See, for example, Hogg, 1979). As discussed in more detail below, Bowman found the method to work well for a variety of unimodal distributions.

A similar formula was derived by Silverman (1986, pp. 45-48). As an alternative to successive re-estimation of h in (2.11) $\int [f''(x)]^2\, dx$ may be calculated by reference to a standard parametric distribution. Silverman also used a Normal distribution and normal kernels, which gives

$$h = 1.06\, n^{-1/5}\, \sigma. \qquad (2.16)$$

He remarks that for multimodal data this will oversmooth somewhat since $\left[\int [f''(x)]^2\, dx\right]^{1/5}$ will be large relative to the standard deviation. Unlike Fryer (1976) (who did not estimate σ), he found, replacing σ by the sample standard deviation (s.d.), that as a

mixture of 2 unit Normal distributions became more strongly bimodal that (2.16) oversmoothed more and more. It also oversmoothed heavily skewed data but was insensitive to kurtosis. Using inter-quartile (IQ) range as a robust estimate of a multiple of σ improved performance on the long-tailed and skew distributions but oversmoothed still more for the bimodal case. Silverman recommended a compromise choice of

$$0.9\ A\ n^{-1/5} \tag{2.17}$$

where A = min {s.d., (IQ range)/1.34 }

as giving near-optimal MISE for all except skewness > 1.8 and separations > 3σ where both skewness and bimodality should still be clear for samples of size > 100.

Methods based on cross-validation

Habbema, Hermans and van den Broek (1974) and Duin (1976) took a maximum likelihood approach instead.

Maximising $\prod_{i=1}^{n} \hat{f}_n(X_i)$ is readily seen to yield the solution h = 0 and the estimate then consists of a set of spikes, one at each data point. Habbema et al. resolved this by using a leaving-one-out modification, maximising instead

$$\prod_{i=1}^{n} \hat{f}_{n-1}^{(i)}(X_i) \tag{2.18}$$

where $\hat{f}_{n-1}^{(i)}(X_i)$ is the estimate of the density at the point X_i, obtained by using the remaining (n-1) sample points.

Maximising (2.18) may be shown (Bowman, 1984) to be equivalent to minimising

$$\frac{1}{n} \sum_{i=1}^{n} [\ I\{\delta(x-X_i),\ \hat{f}_{n-1}^{(i)}(x)\} - I(\delta(x-X_i,\ f(x)\}\ ] \tag{2.19}$$

where $I(g,\ \hat{g}) = \int g(x) \log \frac{g(x)}{\hat{g}(x)}\, dx$ is the Kullback-Leibler loss function (Kullback, 1959, pp. 6-7) and $\delta(\cdot)$ is the Dirac delta function. This is a cross-validatory choice in the sense of Stone (1974a) (Titterington, 1980). It can be shown (Silverman, 1986,

pp. 52-53) that, up to a constant, (2.19) is an unbiased estimator of the expected Kullback-Leibler error for an estimate using the same smoothing parameter on a sample of size n-1, subject to certain restrictions on f and K. Scott and Factor (1981) observed that this method is very sensitive to outliers, tending to oversmooth to avoid very small values of $\log \hat{f}_{n-1}^{(i)}(x)$ at outlying observations, although in simulations it performed well on average. Bowman (1981) suggests that oversmoothing will not be serious unless the outlier is a considerable distance from the rest of the data, but that apparent gross outliers should be removed for the purpose of applying (2.19). Schuster and Gregory (1981) showed that for distributions where either tail decays at or more slowly than an exponential rate, this method leads to inconsistent density estimates while Chow, Geman and Wu (1983) prove consistency when both kernel and density have finite support. Recently Hall (1987) has shown that the asymptotic properties of Kullback-Leibler cross-validation depend heavily on interaction between the tail properties of the kernel and the unknown density. There are also potential difficulties with discretised data. Silverman (1986, pp. 54-55) shows that if there are no isolated data points then (2.19) tends to infinity as $h \to 0$ (Scott and Factor (1981) provide an example) and that if there are some isolated points but a lot of ties in the data the behaviour of (2.19) as $h \to 0$ will depend on the tail properties of the kernel.

Another loss function, used by Bowman (1981, 1984) and Rudemo (1982), yielding a mathematically tractable criterion from (2.19) is Integrated Square Error, ISE,

$$I(g, \hat{g}) = \int [g(x)-\hat{g}(x)]^2 \, dx \qquad (2.20)$$

In this case we minimise

$$\frac{1}{n} \sum_{i=1}^{n} \int [\hat{f}_{n-1}^{(i)}(x)]^2 \, dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{n-1}^{(i)}(X_i) \qquad (2.21)$$

Using normal kernels and results on convolutions of Normal densities the integration can be done analytically though as with the Kullback-Leibler criterion the optimising value of h is found numerically. Minimising (2.21) may be seen as maximising

$$\frac{1}{n} \sum_{i} \hat{f}_{n-1}^{(i)}(X_i)$$

subject to a roughness penalty

$$-\frac{1}{2n} \sum_i \int [\, \hat{f}_{n-1}^{(i)}(x) \,]^2 \, dx.$$

Again, it can be shown that up to a constant factor (2.21) is an unbiased estimator of MISE between f(x) and $\hat{f}_{n-1}^{(i)}(x)$ (Silverman, 1986, pp. 48-49), and Stone (1984) has shown that subject to mild conditions on f(x), and if K has finite support, the ratio of minimised ISE using (2.21) to the true minimised ISE tends to 1 as $n \to \infty$, with probability 1. Hall (1983a) also proves that minimising (2.21) is equivalent asymptotically to minimising ISE and also MISE, but does not restrict the domain of K. Unfortunately there may be severe problems with discretised data. Silverman (1986, pp. 51-52) shows that if the number of ties in the data relative to sample size is greater than a threshhold value dependent on the form of the kernel (.55 for normal kernels) then (2.21) tends to $-\infty$ as $h \to 0$ and hence an answer of no smoothing will be returned. In the light of this, and the difficulties described above with the modified maximum likelihood method, he urges caution in applying these methods without first restricting the range of h, suggesting $(.25h^*, 1.5h^*)$, where h^* is given by (2.17), as a sensible range to use.

Bowman (1981, 1985) in simulation studies using normal kernels on samples of size 25, 50 and 100 from mixtures of Normal distributions, Student's t, Cauchy, x^2 and Beta distributions compared use of his modification of Fryer's method (2.14), the asymptotically optimal MISE choice (2.12), cross-validation with both the Kullback-Leibler and ISE measures, a method based on the Cramer-von-Mises statistic

$D(F, F_n) = n \int [F(x) - F_n(x)]^2 \, dF(x)$ and several variable kernel techniques. Performance was measured in terms of 3 goodness-of-fit statistics of the form $\int [f(x) - \hat{f}_n(x)]^2 \, w(x) \, dx$. Fryer's Normal Optimal method was found to be superior to all others for unimodal data except for the long-tailed distribution with the larger sample size, where, not surprisingly, a variable kernel method was the best. Caution was advised in applying it indiscriminately, especially in higher dimensions where multimodality is more likely.

Cross-validation when used with the Kullback-Leibler criterion performed well in all cases (except with variable kernels) and was usually superior to the ISE criterion, but not for long-tailed distributions. Bowman (1984) found that in simulations from a standard Cauchy distribution that ISE was greatly superior to Kullback-Leibler cross-validation, even for samples as small as 25, and gives an example where the latter vastly oversmooths. Bowman (1985) recommended the ISE criterion as being the most generally dependable, at least for reasonably large samples. The asymptotically optimal method tended to undersmooth and did not perform very well except for the bimodal case. Scott and Factor (1981) also compared the modified maximum likelihood and asymptotic optimal methods on 25 samples of size 50 and 100 for normal data and a mixture of Normals, and found that in general the former method outperformed the latter in terms of MISE although the difference was not great.

2.3.3 Multiple smoothing parameters

The discussion above applies to univariate data, but any of the univariate methods may also be applied when product kernels of the type (2.6) are used, a separate window width being chosen for each variable marginally before multiplying individual kernels to give the estimate $\hat{f}(\underline{x})$ in d dimensions. However continuous data are usually pre-standardised so that observations on each variable are pooled to estimate a single parameter h, as in (2.7) (use of (2.7) without prior standardisation would be expected to result in a degraded performance). Habbema, Hermans, and van den Broek (1974) applied this method with normal kernels giving

$$\hat{f}(\underline{x}) = \frac{1}{(h\sqrt{2\pi})^d \, s_1 \ldots s_d} \; \frac{1}{n} \sum_{i=1}^{n} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{d} \left[\frac{x_j - X_{ij}}{h s_j} \right]^2 \right\}$$

where s_j is the sample standard deviation of the jth variable. The modified maximum likelihood method applied to the standardised data,

$$\left\{ Y_{ij} = \frac{X_{ij}}{s_j} \right\}, \; i = 1, .., n \; ; \; j = 1, \ldots, d, \text{ was used to choose h.}$$

The method was applied to a 2 population discrimination problem using 75 training and 23 test cases from what appeared to be near Bivariate Normal distributions with equal covariance matrices. Comparing estimated posterior probabilities, the authors found that the results were very similar for most of the test cases to those obtained by fitting Multivariate Normal distributions (i.e. using the LDF). Remme, Habbema and Hermans (1980) used the same method, in the same context, on samples of 15 or 35 training cases from each of populations of Multivariate Normal, logNormal and mixture distributions in 2 to 10 dimensions, evaluating the performance of the kernel, LDF and QDF methods on samples of 50 test cases from each population. Their conclusions were that the kernel method was the best or near-best method in all cases considered except for the multinormal case with equal covariance matrices where as expected the LDF was optimal. The kernel method was disappointing for the lognormal case, where a variable kernel was found by Habbema, Hermans and Remme' (1978) to be more appropriate. Except in this case, where improvement was slow, the kernel method performed increasingly well as sample sizes were increased (to 100 or 200 cases) though did surprisingly well even for relatively small samples. The modified maximum likelihood method was compared to a choice of h over the range 0.0 to 1.0, setting h equal in both populations, and found to produce satisfactory though slightly high values of the smoothing parameter (except for the lognormal case) leading to rather conservative estimates of the posterior probabilities. This accounts for the method making fewer very poor predictions than either the LDF or QDF. The results of Remme et al. (1980) also suggested caution in applying product kernels to highly correlated data, as mentioned in Section 2.1.3.

Van Ness and Simpson (1976), who compared the effects of dimension on kernel and parametric methods, took a different approach by choosing h to give the best classification rate for each separation of 2 Multivariate Normal populations (which had equal unit covariance structure) in each dimension (between 1 and 30) studied. Unfortunately they used completely independent sets of training and test cases to select h before applying it to the final simulations, giving the kernel method a headstart over the LDF and QDF. As might be expected, h was seen to increase with both separation and dimensionality, though sensitivity to separation was slight for high dimensionality. Error rate was

found to be extremely insensitive to h once h was sufficiently large, though van Ness (1979) found that this did not maintain for populations with unequal covariances. For equal sample sizes of 10 and 20 the kernel method performed similarly to the optimal LDF even in high dimensions, and performed well compared to parametric methods involving estimation of Σ even for very low dimensions. Van Ness (1979) used a cross-validation method to select h (on the basis of error rate). The performance of the kernel method with normal kernels and equal smoothing parameters, h_1 and h_2, in the 2 populations, was compared with the same method but where h_2 was set proportional to h_1, the constant of proportionality being determined by the sample standard deviations. Within populations the same value of h was used for each variable since the data were generated from $MVN_d(\underline{0}, I)$ and $MVN_d(\underline{u}, \frac{1}{2}I)$ distributions respectively. As would be expected, the latter method was found to perform much better than the former for moderately large dimensions ($d \geqslant 4$ or 5). In general using a single h was comparable to the LDF, while QDF was poorest, attributed to difficulties in parameter estimation for high dimensions relative to sample size. The stability of the kernel method as dimension increases was confirmed, provided separate smoothing parameters are used.

2.4 CHOICE OF SMOOTHING PARAMETERS IN A DISCRIMINANT ANALYSIS CONTEXT

While there are many criteria by which to choose the smoothing parameter for good estimation of a single density, simultaneous estimation of more than 1 density (and hence choice of more than 1 smoothing parameter) to produce a good discrimination procedure is largely ignored in the literature.

In discriminant analysis we explicitly or implicitly estimate one or more ratios of densities. Hand (1982, pp. 73, 85) recognises that good estimation of each separate p.d.f. will give good estimation of their ratio and accordingly recommends simultaneous estimation of the smoothing parameters using a criterion of the form

$$\int\left\{ |D(f_1(x), \hat{f}_1(x))| + |D(f_2(x), \hat{f}_2(x))| \right\} . w(f_1(x), f_2(x)) \, dx$$

where D(a, b) is a difference function, such as MSE, and w(a, b) is a weight function which decreases monotonically with increasing

distance between a and b. In practice of course $\{f_i(x)\}$ are unknown and, again, a cross-validatory approach would be necessary.

However, whether or not they are chosen simultaneously, it is easily seen that the optimal choice of smoothing parameter for each separate density will not necessarily be optimal with respect to estimation of the density ratio(s). Although a rather contrived example, this is illustrated for the 2 population case (considered henceforth) where both populations are identical and the priors are known. Writing $p_T(\cdot|\cdot)$ for the true function, we therefore have

$\frac{p_T(\pi_1|x)}{p_T(\pi_2|x)} = \frac{\theta_1}{\theta_2}$, $\forall x$, the prior odds ratio, and see that if infinite smoothing is applied to both samples (c.f. Figure 2.1(e)) we also have $\frac{\hat{p}(\pi_1|x)}{\hat{p}(\pi_2|x)} = \frac{\theta_1}{\theta_2}$, $\forall x$, so that the effect of the data is lost. Infinite smoothing, clearly far from optimal from the "marginal" point of view, produces exactly the right answer in this case. (In theory there may be other choices of $\underline{h}$ which would also achieve this but, due to sampling variability, the estimated densities will not be identical). We note in passing that in general θ_1/θ_2 is not known, so that the related but separate issue of reliable estimation of the prior odds ratio also arises, considered for instance by Hand (1986b).

Rather than estimate densities separately we therefore consider the posterior probability function $p(\pi_1|x)$ to be of more direct interest and in particular would choose (h_1, h_2) to optimise a goodness-of-fit criterion between the estimated and true predicted probability functions, such as MSE. In the simulation study reported in Section 2.5 below we illustrate the poor performance of several of the marginal methods discussed above, with respect to MSE.

In practice of course $p_T(\pi_1|x)$ is not known and wholly data-based procedures are required. As noted by Hand (1982, pp. 66-67, 83-85), the optimality of a particular (h_1, h_2) depends on the criterion/criteria by which our final discriminant rule is to be assessed. It is therefore natural to choose the smoothing parameters directly to optimise the assessment score of most interest, if such a score can be provided at the outset.

Tutz (1986) considered choosing the smoothing parameters to

minimise the leaving-one-out error rate (Lachenbruch and Mickey, 1968) and showed the resulting rule to be Bayes' risk consistent (van Ryzin, 1966; Glick, 1972) when used with discrete Aitchison and Aitken (1976) kernels. He applied the method to a 2 group allocation problem using 6 discrete feature variables, comparing it to variants of the modified maximum likelihood method used by Aitchison and Aitken (1976) in a discrete kernel context, but originally advocated by Habbema, Hermans and van den Broek (1974) with continuous data. In practice it is awkward to optimise a discrete function such as error rate. Tutz avoided the problem by using a smoothed approximation to the error rate, proposed by Glick (1978). Nevetheless, in terms of non-error rates using a test sample of 90 cases, allowing separate smoothing parameters for each variable and in each population, Tutz's method allocated only one more test case correctly than the best of the maximum likelihood methods considered. This is not altogether surprising as error rate is of course measuring only group separation and is known to be very insensitive (Shapiro, 1977; Hilden, 1984; Remme et al., 1980; Titterington et al., 1981) to small changes in the estimated predicted probabilities. In some contexts, separation may be all that is of interest, but we would argue that often the value of the probabilities themselves is at least as important. For instance, in medical applications the clinician may be provided with the predicted probabilities, which he can treat like the result of a clinical test, using them as just one of several aids in arriving at his final diagnosis. He can then assess the "cut-off" point at which the odds ratio $p(\pi_1|x) \; / \; p(\pi_2|x)$ becomes indicative of π_1 rather than π_2 in a more natural manner than having to specify an appropriate cost/loss structure a priori, or assuming equal costs as is often done. (Kerridge, 1966; Aitchison and Aitken (1976) also comment on lack of agreed loss structure necessitating provision of realistic estimates of type probabilities rather than an allocation rule). In this case, and where we know from past experience that it is unlikely that any new model will effect a great improvement in error rate, fine-tuning of probabilities will be important, and assessing reliability more important than separation. Reliable probabilities should of course yield a near-optimal error rate automatically.

For this reason, we consider optimisation of the continuous assessment scores (Brier (1.13), log (1.14) and modified log (1.15)

scores) discussed in Section 1.6.2. As a squared error loss function, the Brier score is of particular interest, and reflects both reliability and separation. In practice cross-validation is required when choosing $\underline{h}$ to optimise these scores on the training data, otherwise no smoothing will be indicated. This is because each score involves

$$\hat{p}(\pi_j|x) = \frac{\hat{p}(x|\pi_j)\hat{\theta}_j}{\hat{p}(x|\pi_1)\hat{\theta}_1 + \hat{p}(x|\pi_2)\hat{\theta}_2}, \quad j=1, 2 \text{ and for } x \in \pi_1, \ \hat{p}(\pi_1|x)$$

is maximised when $\hat{p}(x|\pi_2) = 0$ and similarly for $x \in \pi_2$. We note that as maximising average log score (1.14) means maximising

$\Pi^2_{i=1} \Pi^{n_i}_{j=1} \hat{p}(\pi_i|X_{ij})$, where X_{ij} is the jth observation in sample i,

this is seen to be analagous to cross-validation with the Kullback-Leibler loss function. As the modified log score (1.15) essentially sets a lower bound to the average log score we would expect the 2 methods to give similar answers. Also, in the 2 group case optimisation of the Brier score (1.13) requires minimisation of

$$\frac{1}{n} \sum^2_{i=1} \sum^{n_i}_{j=1} [\hat{p}(\pi_i|X_{ij})]^2 - \frac{2}{n} \sum^2_{i=1} \sum^{n_i}_{j=1} \hat{p}(\pi_i|X_{ij})$$

and is therefore analagous to cross-validation using the ISE criterion (2.20). Again numerical optimisation is required.

Henceforth we refer to smoothing parameter estimation methods based on (1.13)-(1.15) as "direct" or "assessment" methods, and those which estimate parameters separately or without reference to the predicted probabilities as "marginal" or "indirect" methods.

2.5 SIMULATION STUDY

2.5.1 Description and methods

In the simplest case where we have a single continuous feature variable x, 2 populations, π_1 and π_2, and where the true densities $\{p_T(x|\pi_i)\}$ and hence (given the mixing weights $\{\theta_i\}$) $\{p_T(\pi_i|x)\}$ are known, we can compare the performance of various methods with respect to mean square error (MSE)

$$\int_x [p_T(\pi_1|x) - \hat{p}(\pi_1|x)]^2 \, f(x) \, dx,$$

where $f(x) = \theta_1 \, p(x|\pi_1) + \theta_2 \, p(x|\pi_2)$,

by means of contour plots of MSE as a function of the 2 smoothing parameters h_1 and h_2.

A simulation study was carried out, using normal random variables and fixed normal kernels. The prior probabilities, θ_1 and θ_2, were estimated using the sample proportions, and the density estimates were calculated by means of the fast Fourier transform (Silverman, 1982b) with modifications as in Jones and Lotwick (1984). Numerical integration was used to find MSE using the trapezoidal rule over a fine grid.

Apart from considerations of sample size (as $n \to \infty$, $h_n \to 0$) we would expect the optimal degree of smoothing for a ratio to depend on the degree of separation of π_1 and π_2 (and hence on expected error rate), less being required for well separated populations. It was also of interest to vary the ratio $\sigma_1^2 : \sigma_2^2$. For given equal sample sizes n_1 and n_2 and no separation, we would expect that if $\sigma_2^2 > \sigma_1^2$ then optimal $h_2 >$ optimal h_1. How different would σ_1^2 and σ_2^2 need to be before the difference between h_1 and h_2 becomes important ? We might also hope that the behaviour of $\underline{h}$ as separation and $\sigma_1^2 : \sigma_2^2$ alter, would be fairly stable across sample sizes.

Accordingly, π_1 was chosen to be standard Normal and $\pi_2 \sim N(\mu, \sigma^2)$ where $\sigma^2 = 1^2, 2^2$, and 3^2. Both balanced and unbalanced cases were considered, for small and medium-sized samples, $n_1 : n_2$ taken as 10 : 10, 25 : 25, 10 : 25 and 25 : 10. The separation μ was standardised by choosing μ to give an expected Bayes' optimal error rate (Lachenbruch 1975a, pp. 10-11, 29-30) of 5%, 20%, or 50%, assuming the 0/1 cost structure, for each (n_1, n_2, σ^2) combination. Further details are given in Appendix 1, and of the means of simulation in Appendix 2. Table 2.1 shows the configurations used.

The marginal methods considered were :

1) the Normal Optimal method (2.14) (NOPT), using (2.15) for $\hat{\sigma}$
2) the Asymptotically Optimal MISE method (2.12) (ASOPT)
3) Cross-validation with the Kullback-Leibler criterion (2.19) (XVKL) and

Table 2.1 Configurations used in the simulation study

$(n_1 : n_2)$	σ^2	Error rate	5%	20%	50%	
		μ				
(10 : 10)	1^2		3.29	1.68	0.00	
and	2^2		4.83	2.31	0.00	(33.87%)
(25 : 25)	3^2		6.26	2.40	0.00	(25.78%)
	1^2		3.16	1.40	0.00	(28.57%)
(10 : 25)	2^2		4.92	2.47	1.16(2)	(28.57%)
	3^2		6.56	2.92	0.00	(26.99%)
	1^2		3.16	1.40	0.00	(28.57%)
(25 : 10)	2^2		4.35	1.14	0.00	(22.73%)
	3^2		5.42	0.00	-	(17.29%)

Note : 1) Error rate interacts with σ^2 and $n_1 : n_2$. The figures in brackets are the closest possible to the required error rate for the configurations in the last column.

2) 1.16 is the smallest μ for which the roots of the quadratic equation are real. (see Appendix 1).

4) Cross-validation with the Integrated Square Error (2.21) (XVISE) criterion.

Method 5 is notional and corresponds to optimisation of the true MSE.

The direct methods with the cross-validated Brier, log and ϵ-log criteria were also used. These are methods 6-8, denoted XV BRIER, XV LOG and XV ϵ-LOG respectively. Numerical optimisation for methods 6-8 was carried out using NAG (1984) subroutine E04JBF, a quasi-Newton type algorithm.

Preliminary work suggested considerable variability of contours between simulations for a given configuration. In the light of this, rather than display average contours, we simply illustrate each case with a small number of individual plots. Figures 2.2-2.36 display contour plots of MSE for 3 simulations from each configuration in Table 2.1, from which the relative performance of the methods may be judged.

2.5.2 Contour plots and discussion

Figures 2.2-2.36 (overleaf)

Contour plots of MSE of $\hat{p}(\pi_1|x)$ as a function of smoothing parameters (h_1, h_2) used in the normal kernel density estimates of each class conditional distribution, for the configurations of sample size and population moments given in Table 2.1. The configuration of sample sizes n_1, and n_2, normal mean μ and standard deviation σ in population π_2 is denoted by (n_1, n_2, μ, σ). The smoothing parameters given by methods 1-4 and 6-8, as described in the text, are superimposed on each plot, or, where these would be off the scale or are unclear, the values are given.

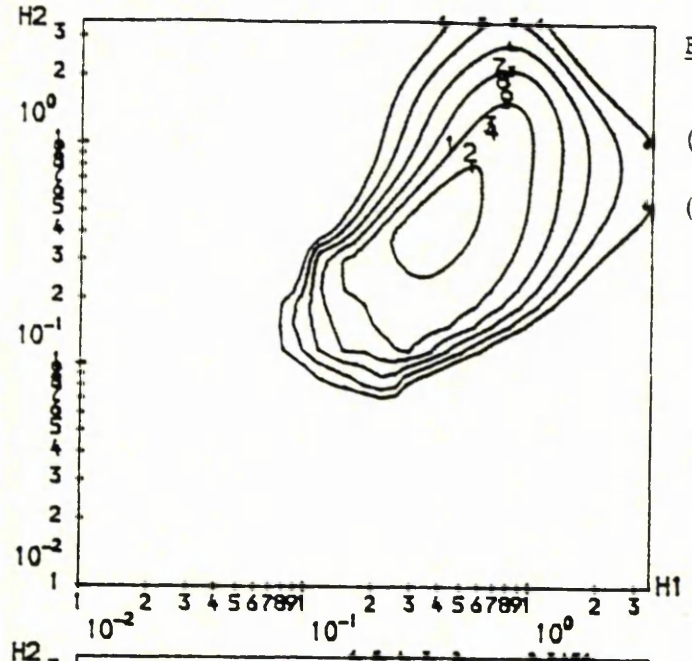
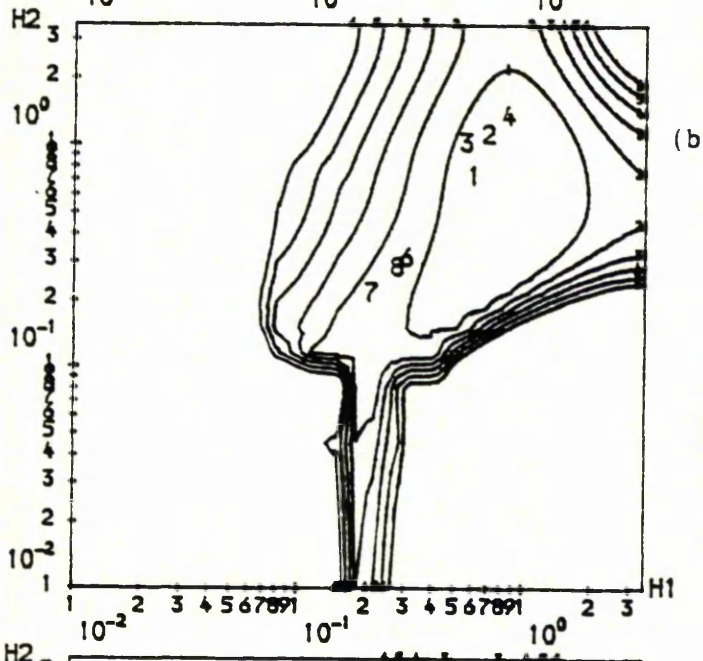


Figure 2.2

$(n_1, n_2: \mu, \sigma) = (10, 10; 3.29, 1)$

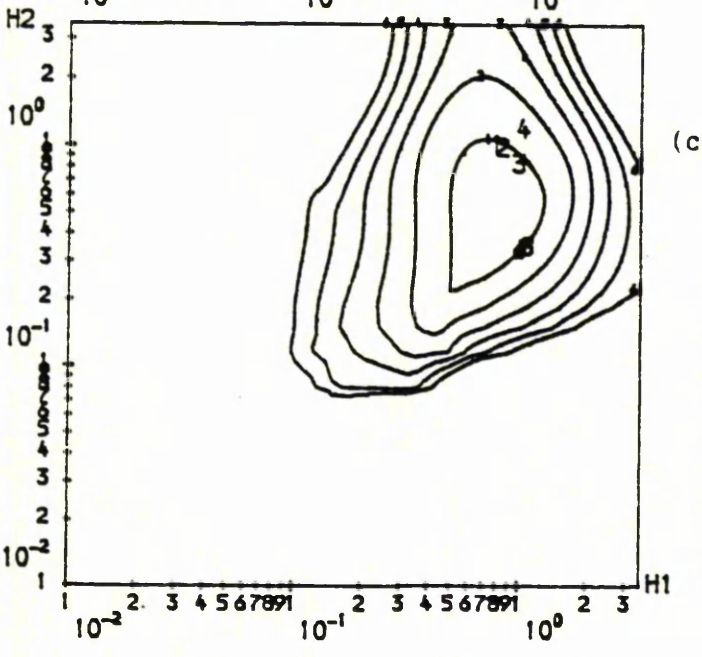
(a)

CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600



(b)

CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600



(c)

CONTOUR KEY	
1	0.0200
2	0.0300
3	0.0400
4	0.0500
5	0.0600
6	0.0700

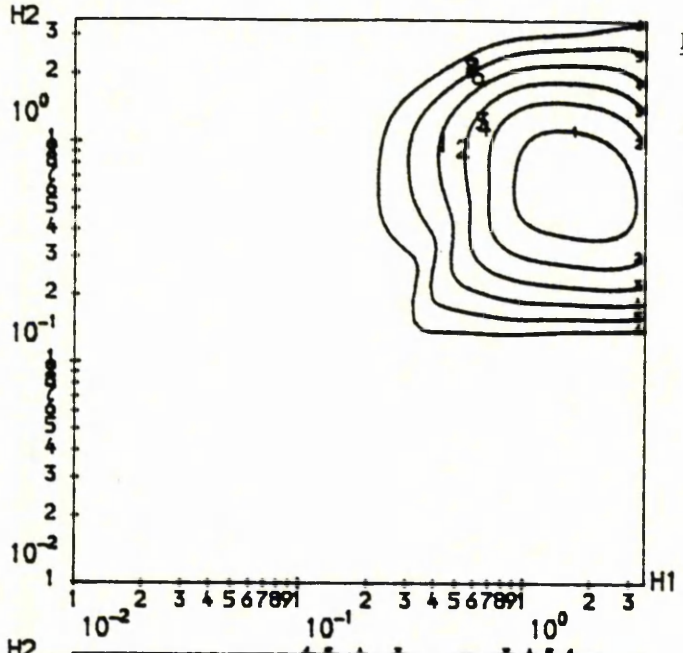
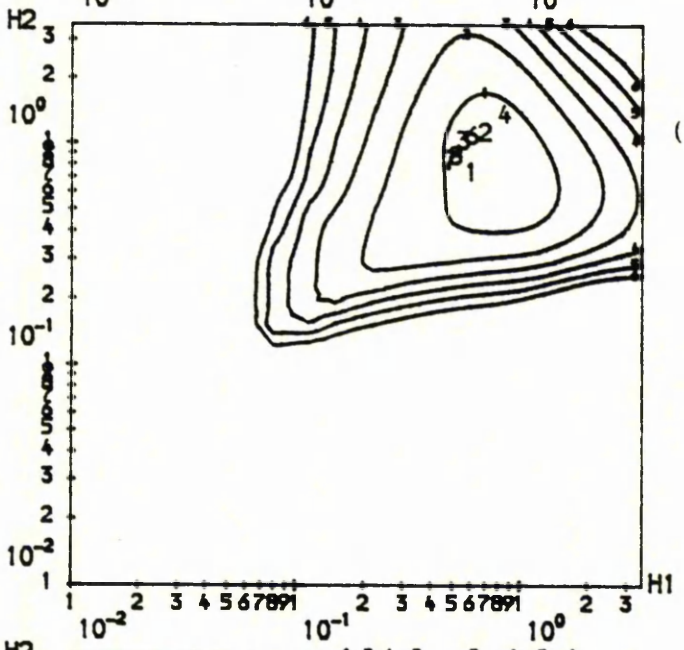


Figure 2.3

$(n_1, n_2; \mu, \sigma) = (10, 10; 1.68, 1)$

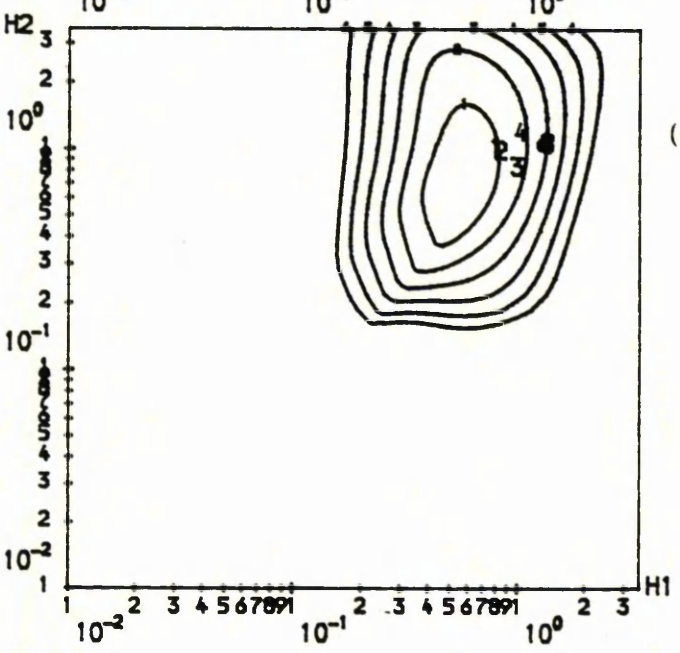
(a)

CONTOUR KEY	
1	0.0400
2	0.0500
3	0.0600
4	0.0700
5	0.0800
6	0.0900



(b)

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550



(c)

CONTOUR KEY	
1	0.0300
2	0.0400
3	0.0500
4	0.0600
5	0.0700
6	0.0800

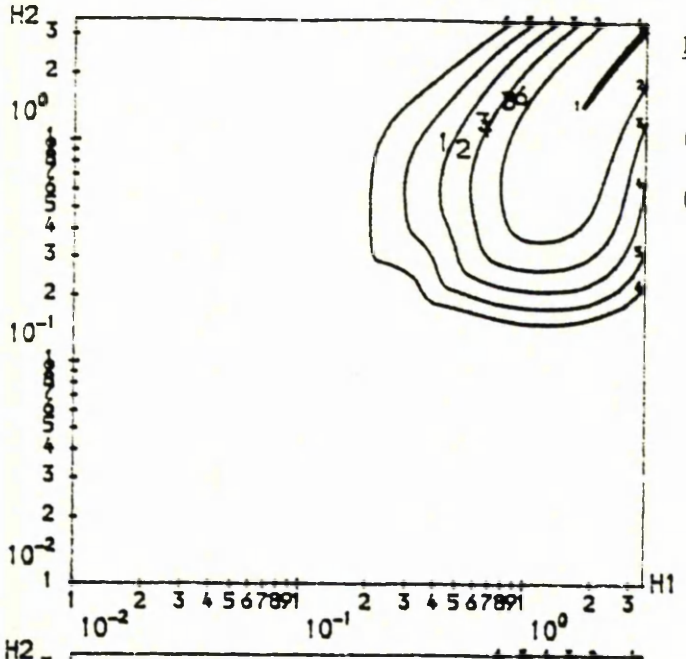
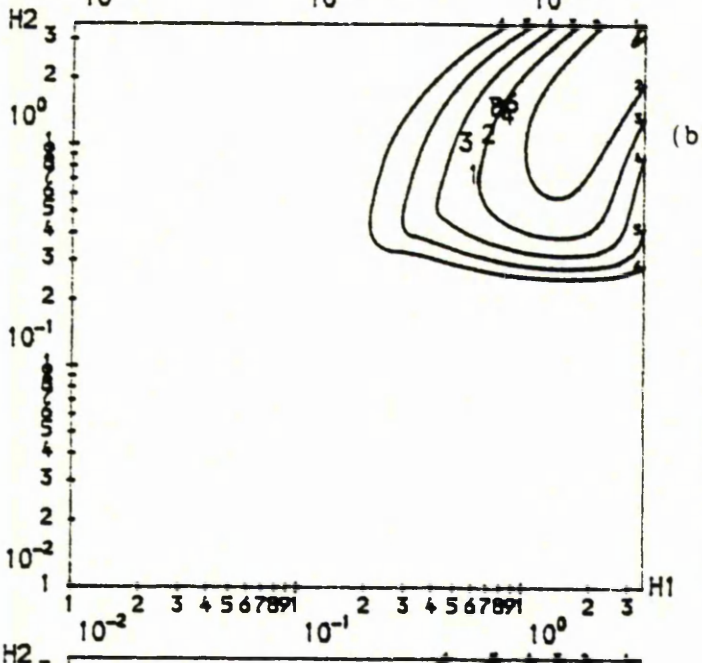


Figure 2.4

$(n_1, n_2; \mu, \sigma) = (10, 10; 0.00, 1)$

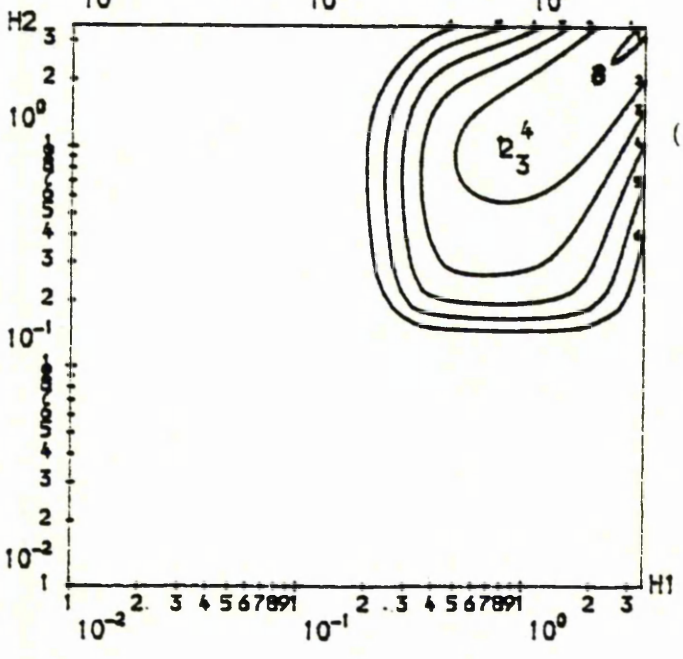
(a)

CONTOUR KEY	
1	0.0001
2	0.0101
3	0.0201
4	0.0301
5	0.0401
6	0.0501



(b)

CONTOUR KEY	
1	0.0005
2	0.0105
3	0.0205
4	0.0305
5	0.0405
6	0.0505



(c)

CONTOUR KEY	
1	0.0005
2	0.0105
3	0.0205
4	0.0305
5	0.0405
6	0.0505

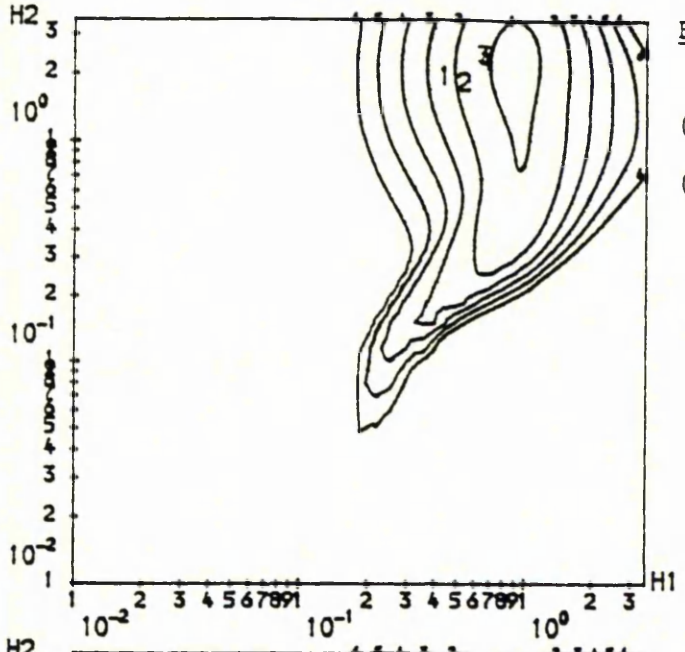


Figure 2.5

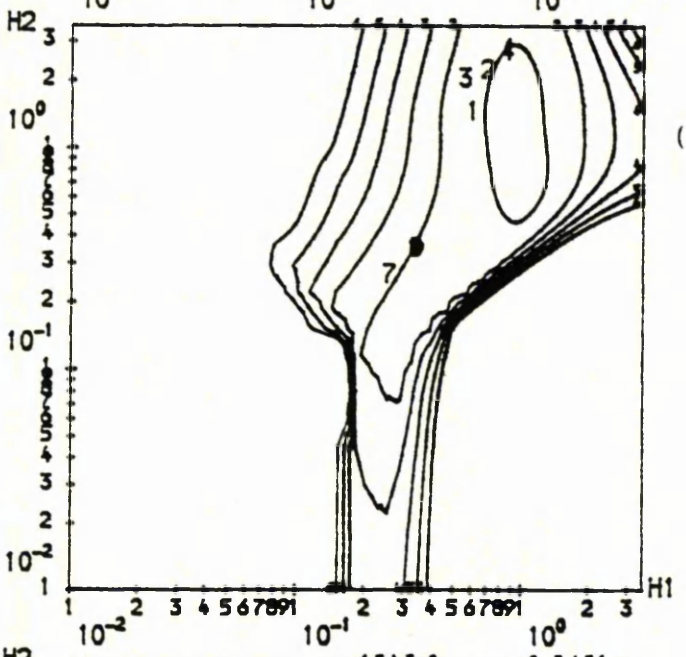
$(n_1, n_2; \mu, \sigma) = (10, 10; 4.83, 2)$

(a) 6- .581, 4.108

7- .538, 4.351

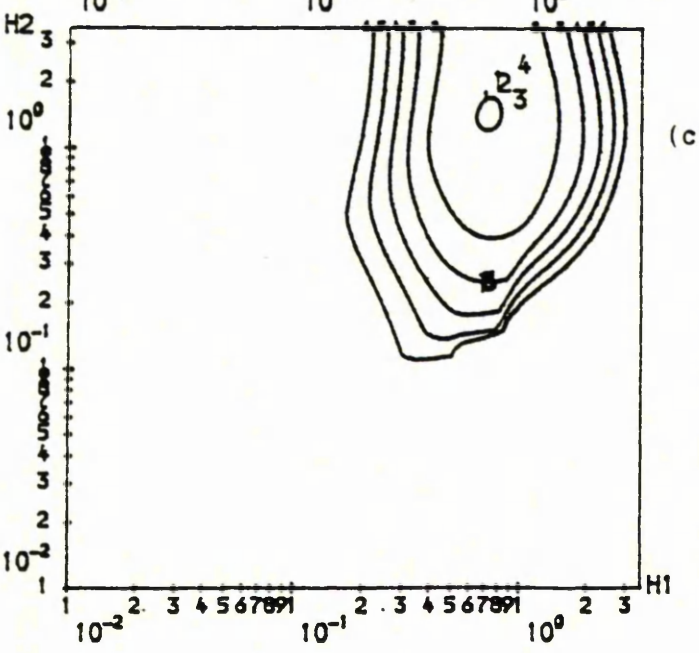
8- .541, 4.584

CONTOUR KEY	
1	0.0225
2	0.0325
3	0.0425
4	0.0525
5	0.0625
6	0.0725



(b)

CONTOUR KEY	
1	0.0025
2	0.0125
3	0.0225
4	0.0325
5	0.0425
6	0.0525



(c)

CONTOUR KEY	
1	0.0155
2	0.0255
3	0.0355
4	0.0455
5	0.0555
6	0.0655

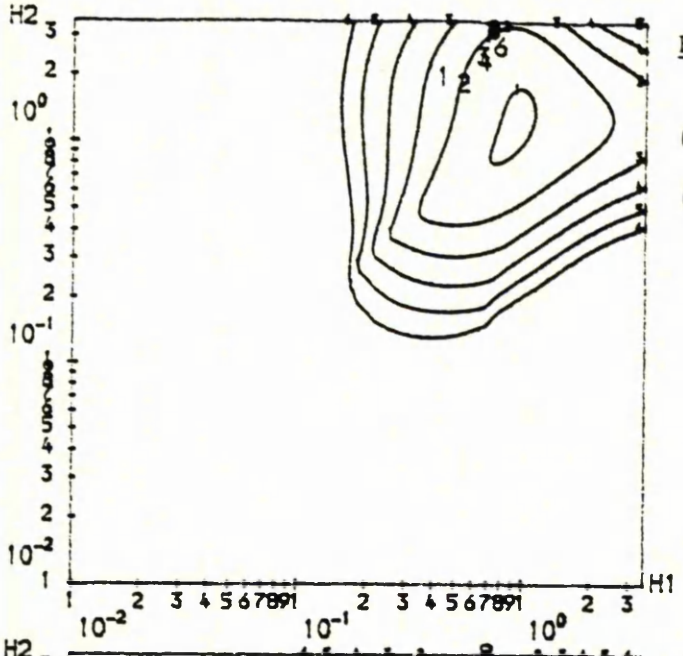
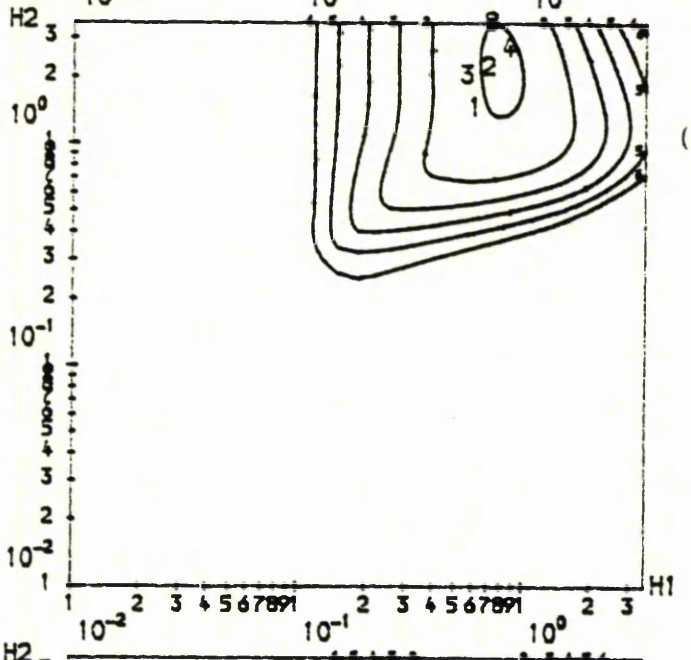


Figure 2.6

$(n_1, n_2; \mu, \sigma) = (10, 10; 2.31, 2)$

(a)

CONTOUR KEY	
1	0.0300
2	0.0400
3	0.0500
4	0.0600
5	0.0700
6	0.0800

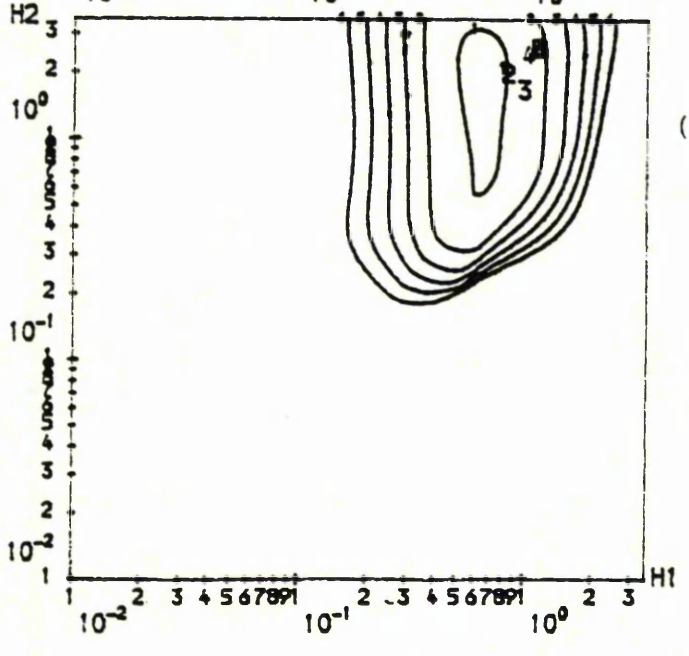


(b) 6- .780, 3.782

7- .667, 3.682

8- .711, 3.467

CONTOUR KEY	
1	0.0020
2	0.0120
3	0.0220
4	0.0320
5	0.0420
6	0.0520



(c)

CONTOUR KEY	
1	0.0150
2	0.0250
3	0.0350
4	0.0450
5	0.0550
6	0.0650

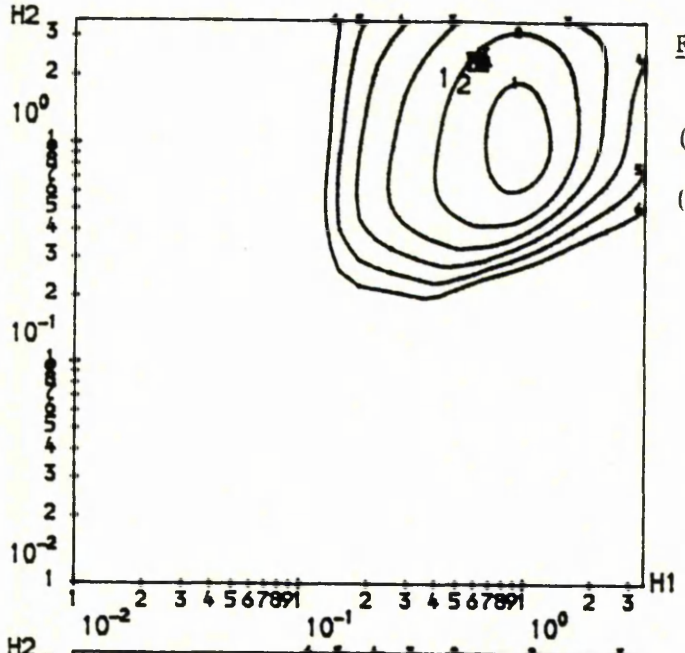
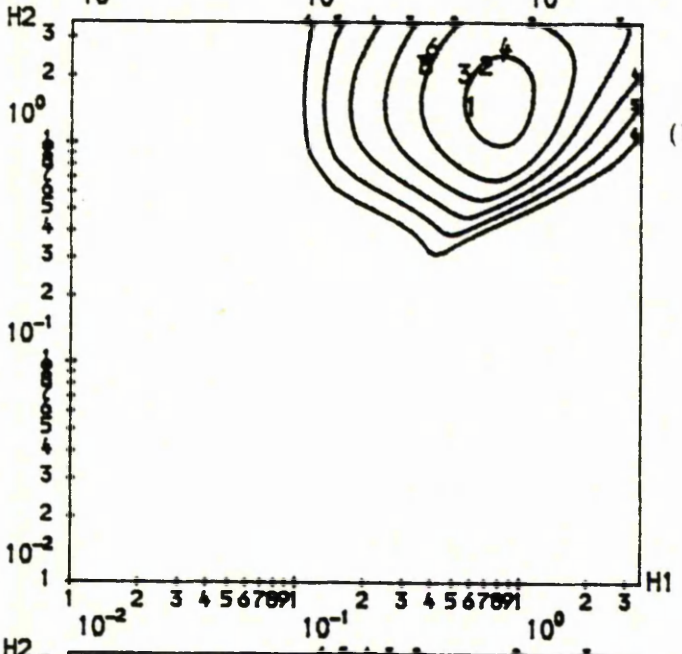


Figure 2.7

$(n_1, n_2; \mu, \sigma) = (10, 10; 0.00, 2)$

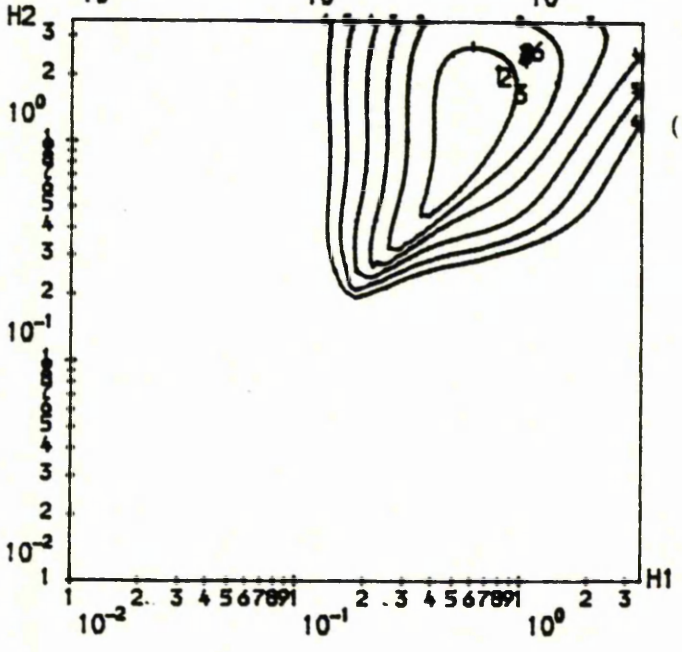
(a)

CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600



(b)

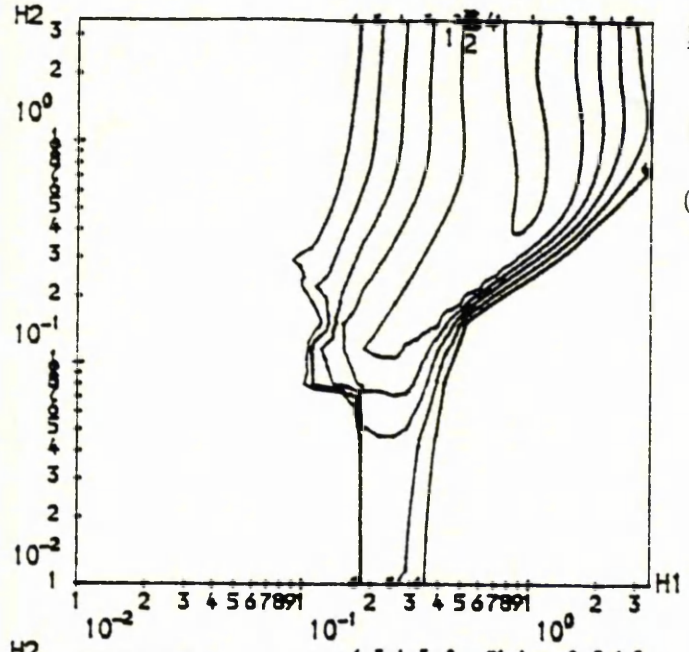
CONTOUR KEY	
1	0.0175
2	0.0275
3	0.0375
4	0.0475
5	0.0575
6	0.0675



(c)

CONTOUR KEY	
1	0.0150
2	0.0250
3	0.0350
4	0.0450
5	0.0550
6	0.0650

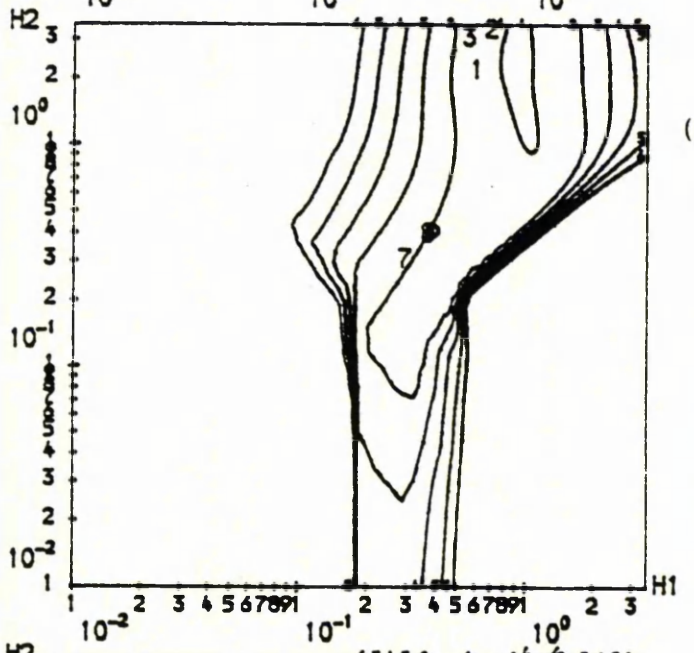
Figure 2.8



$(n_1, n_2; \mu, \sigma) = (10, 10; 6.26, 3)$

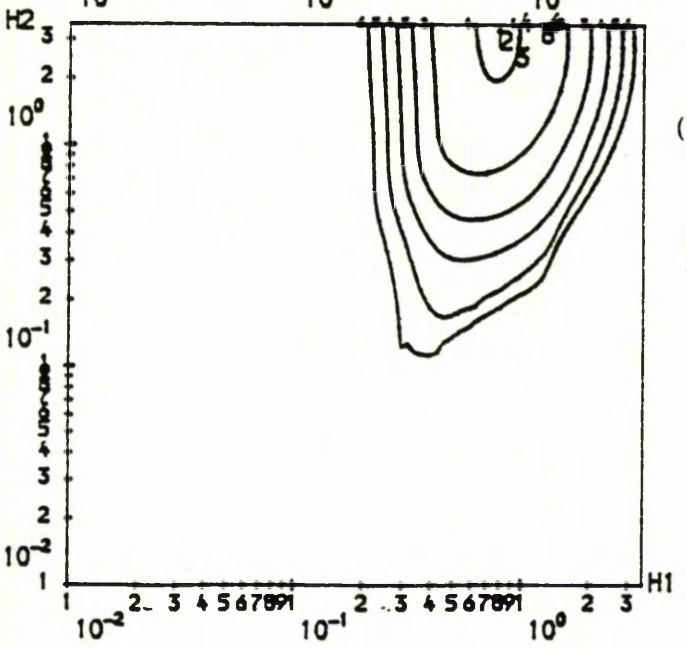
(a) 3- .650, 3.532 6- .561, 3.363
4- .663, 3.349 7- .531, 3.442
8- .537, 3.357

CONTOUR KEY	
1	0.0165
2	0.0265
3	0.0365
4	0.0465
5	0.0565
6	0.0665



(b) 4- .847, 3.951

CONTOUR KEY	
1	0.0010
2	0.0110
3	0.0210
4	0.0310
5	0.0410
6	0.0510



(c) 4- 1.031, 3.482 6- 1.383, 3.482
7- 1.261, 3.140
8- 1.286, 3.078

CONTOUR KEY	
1	0.0130
2	0.0230
3	0.0330
4	0.0430
5	0.0530
6	0.0630

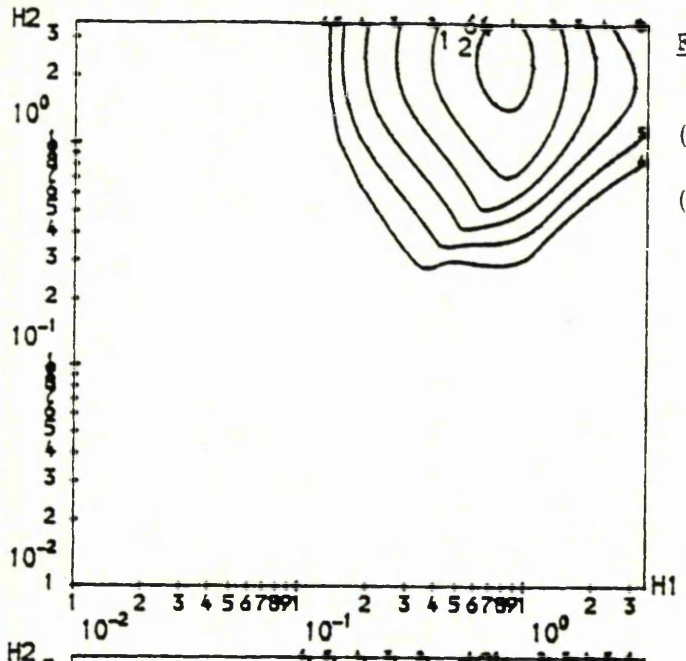
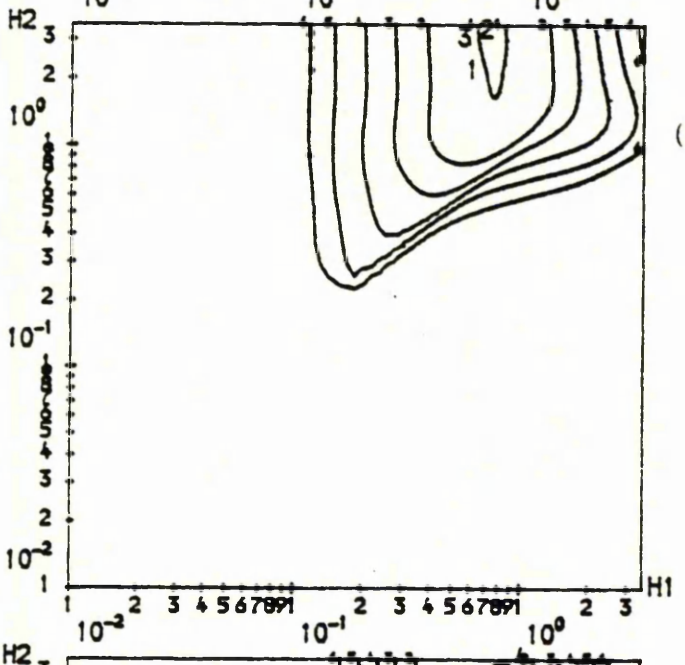


Figure 2.9

$(n_1, n_2; \mu, \sigma) = (10, 10; 2.40, 3)$

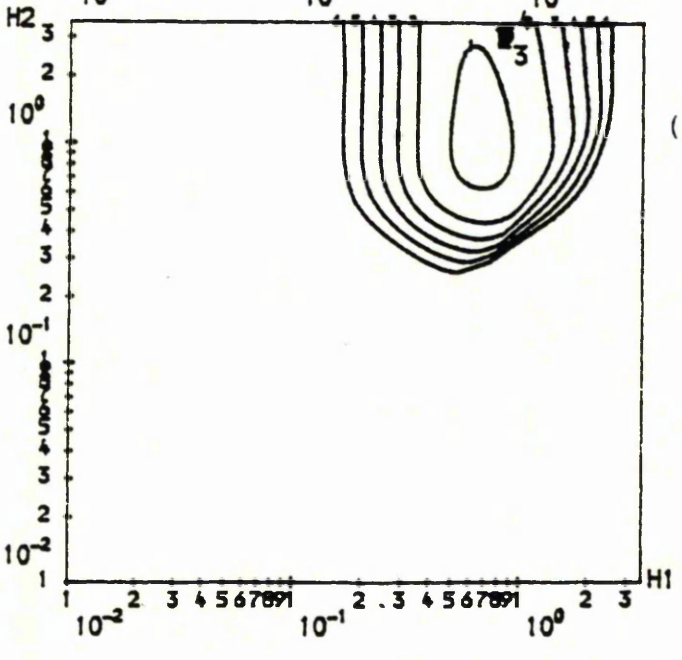
(a) 3- .650, 3.532 7- .554, 4.363
4- .663, 3.349 8- .565, 4.022

CONTOUR KEY	
1	0.0250
2	0.0350
3	0.0450
4	0.0550
5	0.0650
6	0.0750



(b) 4- .847, 3.951 6- .815, 4.999
7- .696, 6.341
8- .727, 5.892

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550



(c) 4- 1.031, 3.482

CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575

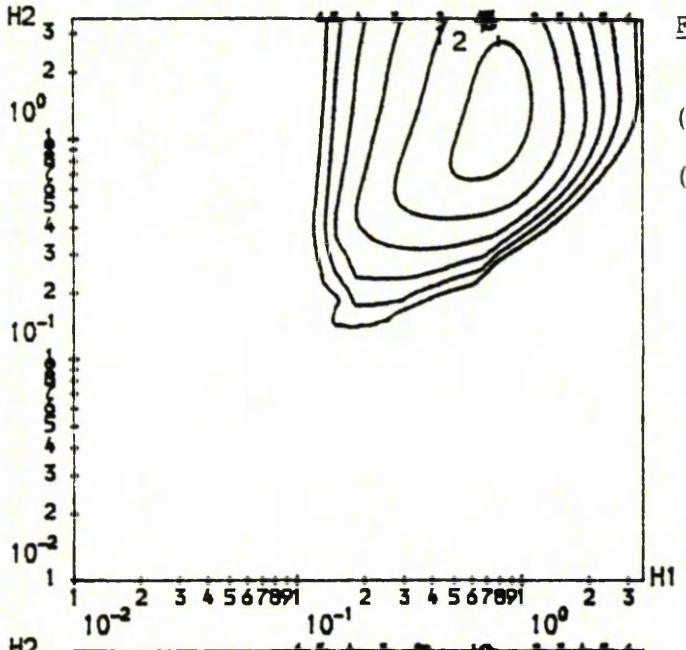
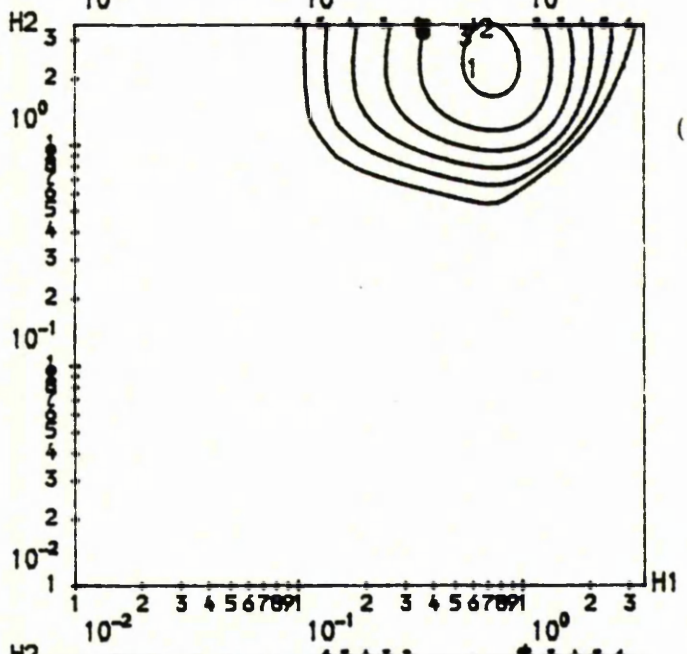


Figure 2.10

$(n_1, n_2; \mu, \sigma) = (10, 10; 0.00, 3)$

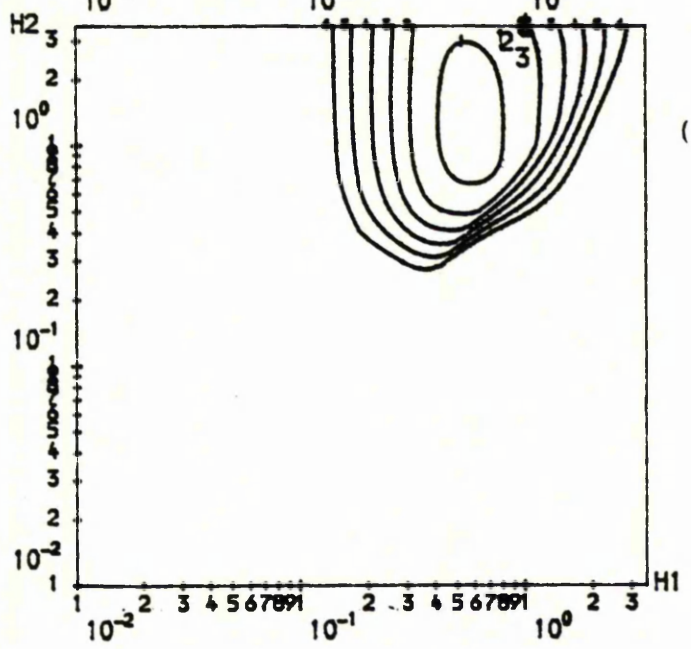
(a) 3- .650, 3.532 6- .735, 3.333
4- .663, 3.349 7- .692, 3.220
8- .713, 3.323

CONTOUR KEY	
1	0.0120
2	0.0220
3	0.0320
4	0.0420
5	0.0520
6	0.0620



(b) 4- .847, 3.951 6- .359, 3.204
7- .368, 3.202
8- .367, 3.212

CONTOUR KEY	
1	0.0150
2	0.0250
3	0.0350
4	0.0450
5	0.0550
6	0.0650



(c) 4- 1.031, 3.482 6- 1.030, 3.324
7- 1.013, 3.328
8- 1.006, 3.349

CONTOUR KEY	
1	0.0125
2	0.0225
3	0.0325
4	0.0425
5	0.0525
6	0.0625

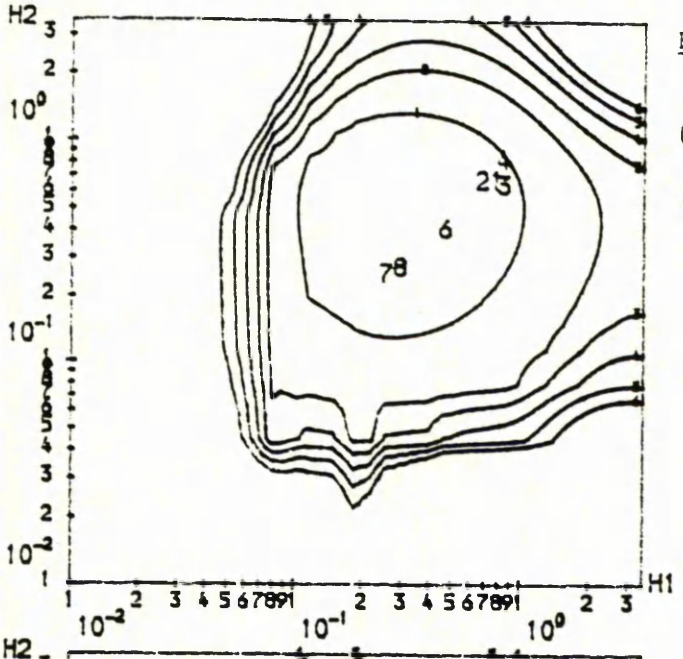
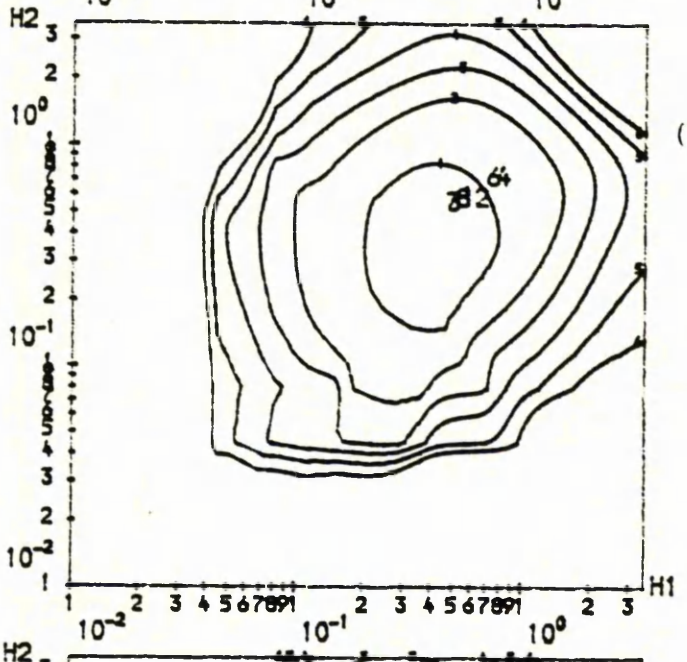


Figure 2.11

$(n_1, n_2; \mu, \sigma) = (25, 25; 3.29, 1)$

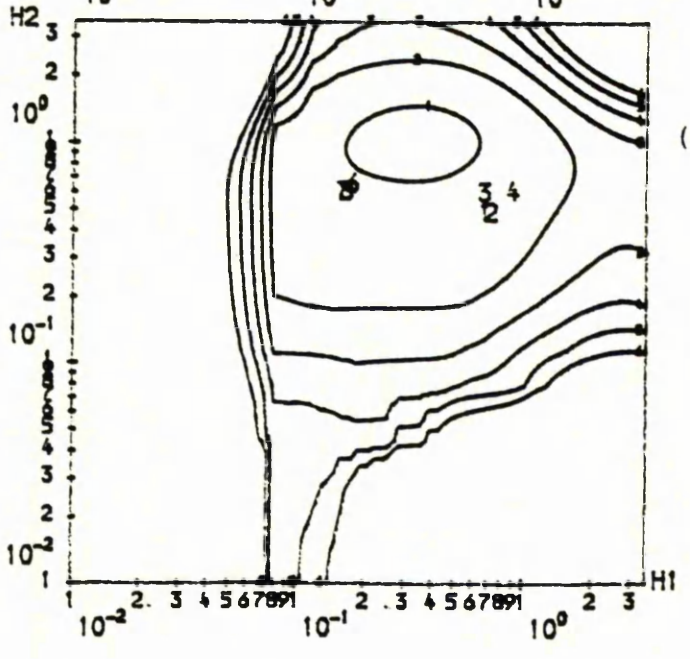
(a)

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550



(b)

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550



(c)

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550

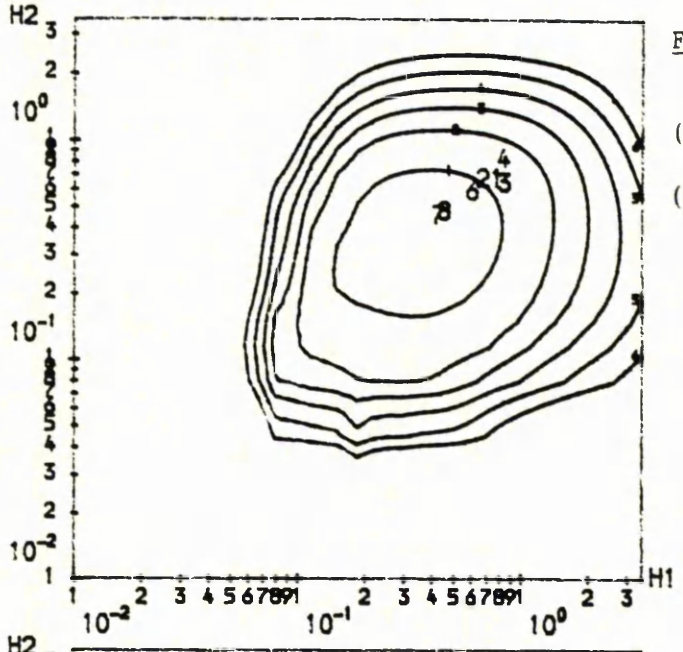
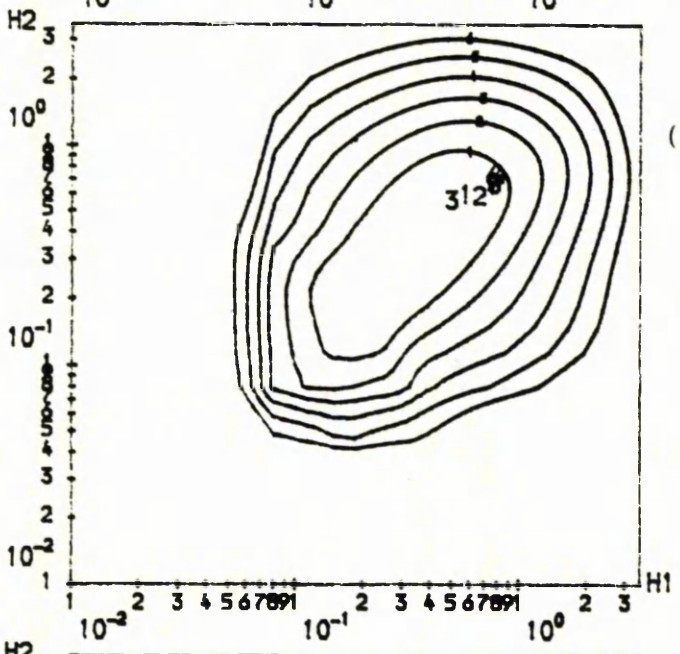


Figure 2.12

$(n_1, n_2; \mu, \sigma) = (25, 25; 1.68, 1)$

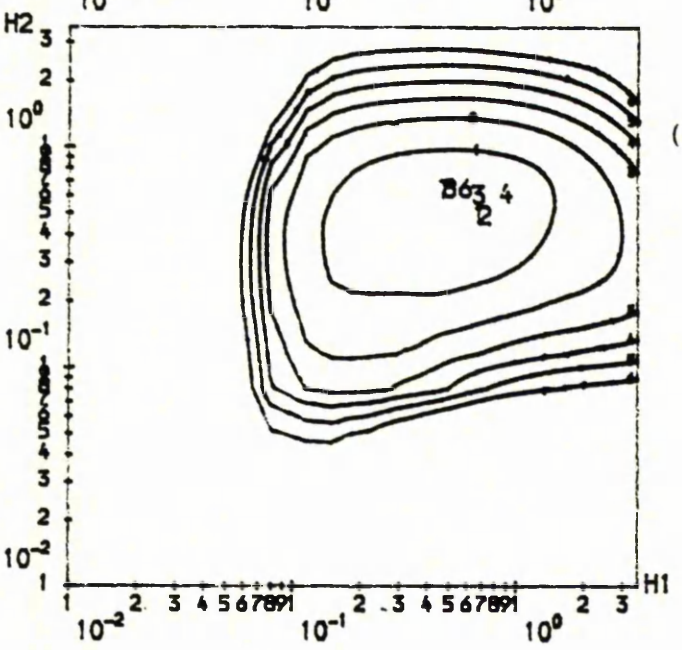
(a)

CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600



(b)

CONTOUR KEY	
1	0.0200
2	0.0300
3	0.0400
4	0.0500
5	0.0600
6	0.0700



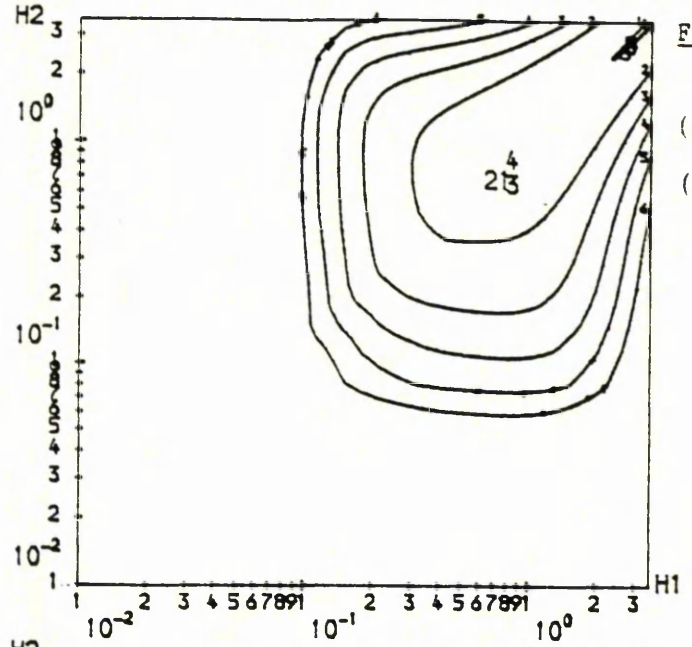
(c)

CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600

Figure 2.13

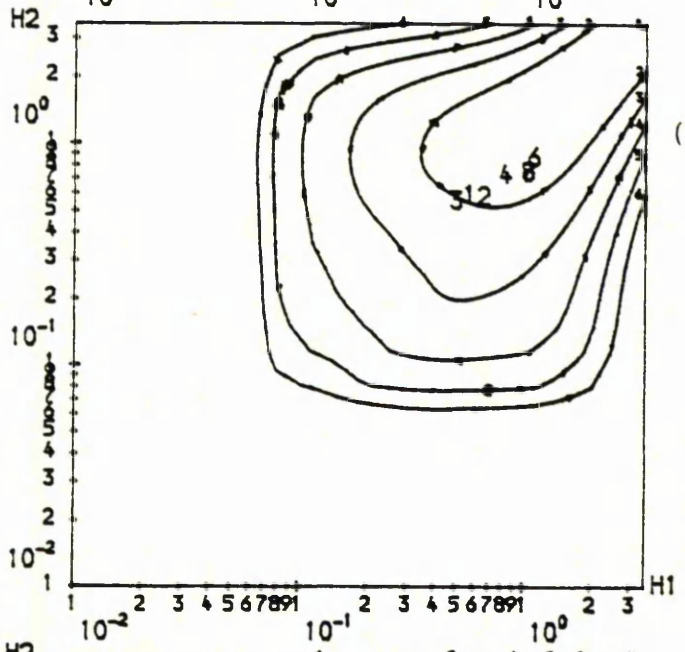
$(n_1, n_2; \mu, \sigma) = (25, 25; 0.00, 1)$

(a)



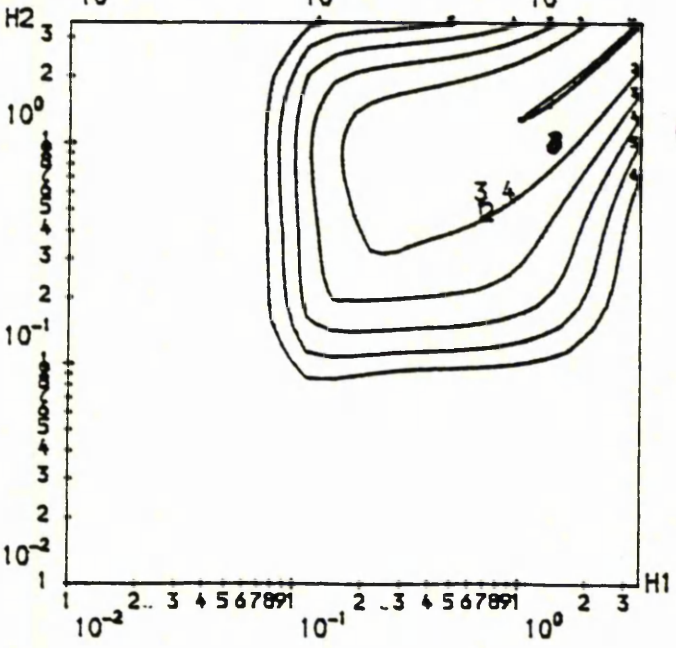
CONTOUR KEY	
1	0.0001
2	0.0101
3	0.0201
4	0.0301
5	0.0401
6	0.0501

(b)



CONTOUR KEY	
1	0.0001
2	0.0101
3	0.0201
4	0.0301
5	0.0401
6	0.0501

(c)



CONTOUR KEY	
1	0.0001
2	0.0100
3	0.0201
4	0.0300
5	0.0400
6	0.0501

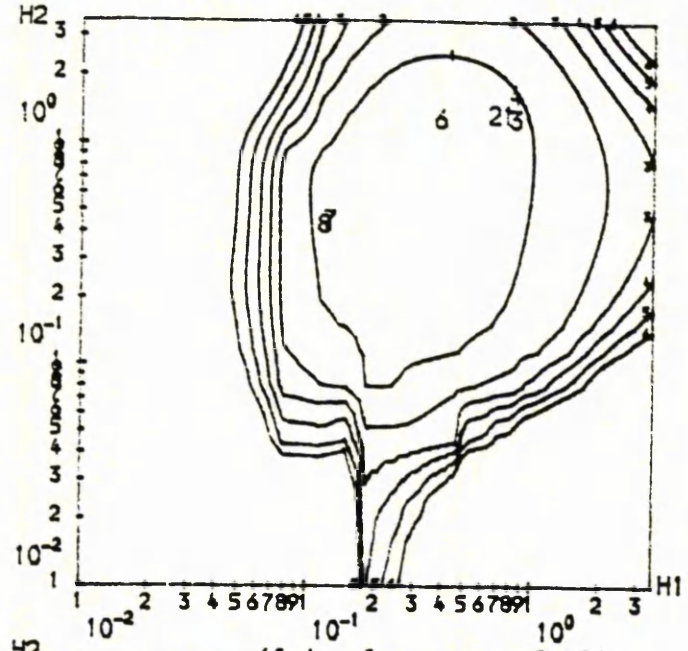
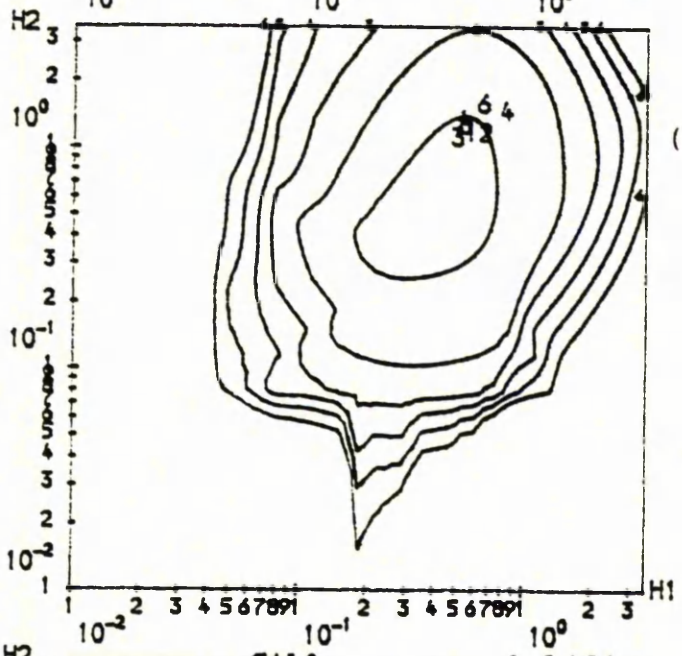


Figure 2.14

$(n_1, n_2; \mu, \sigma) = (25, 25; 4.83, 2)$

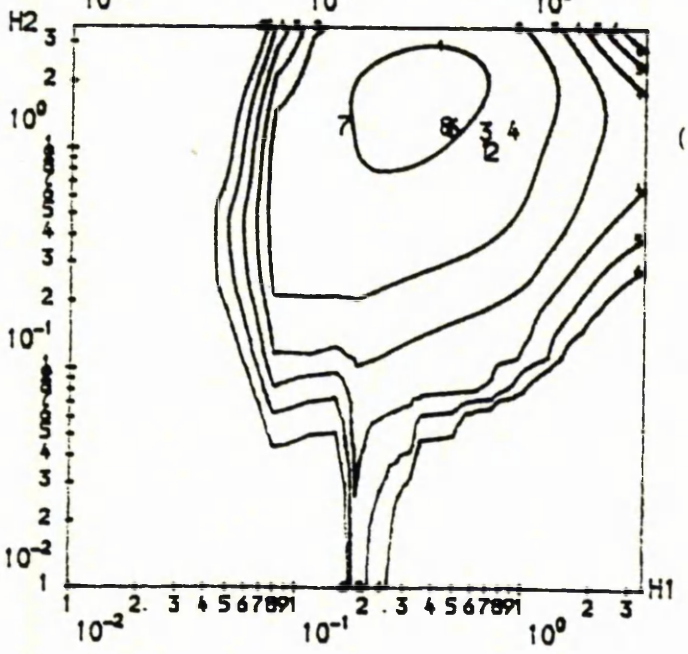
(a)

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550



(b)

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550



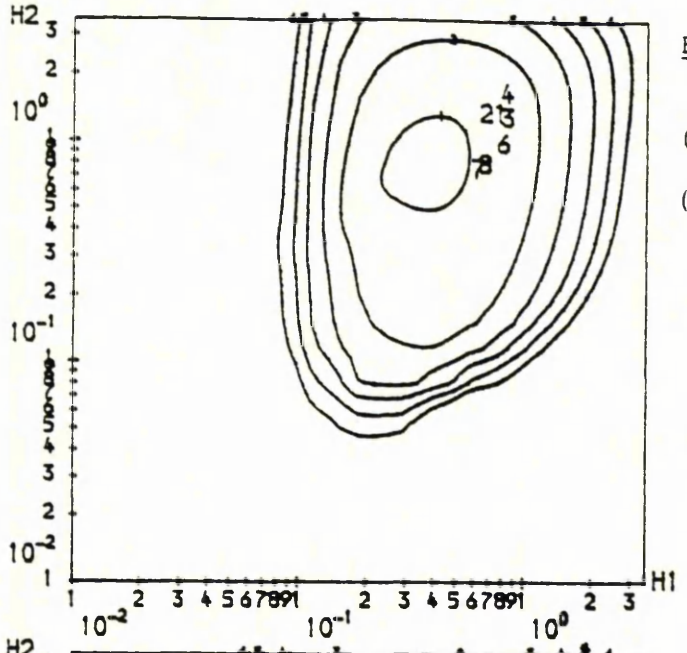
(c)

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550

Figure 2.15

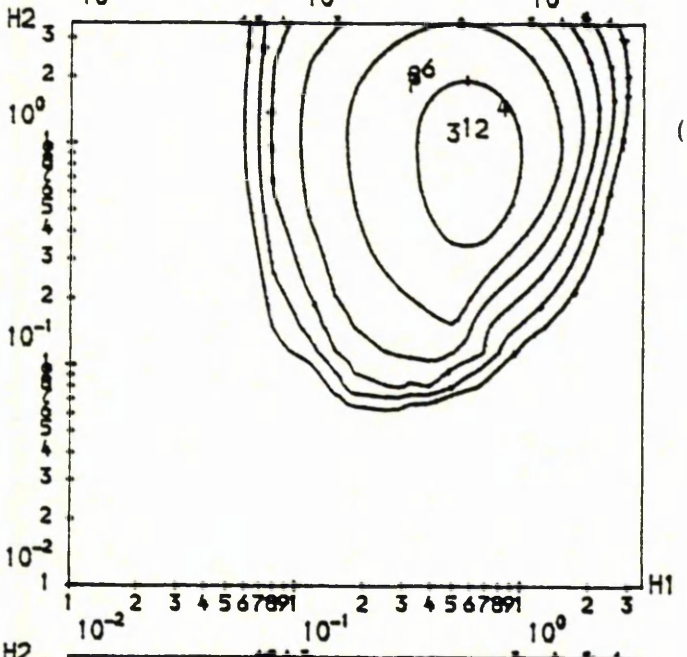
$(n_1, n_2; \mu, \sigma) = (25, 25; 2.31, 2)$

(a)



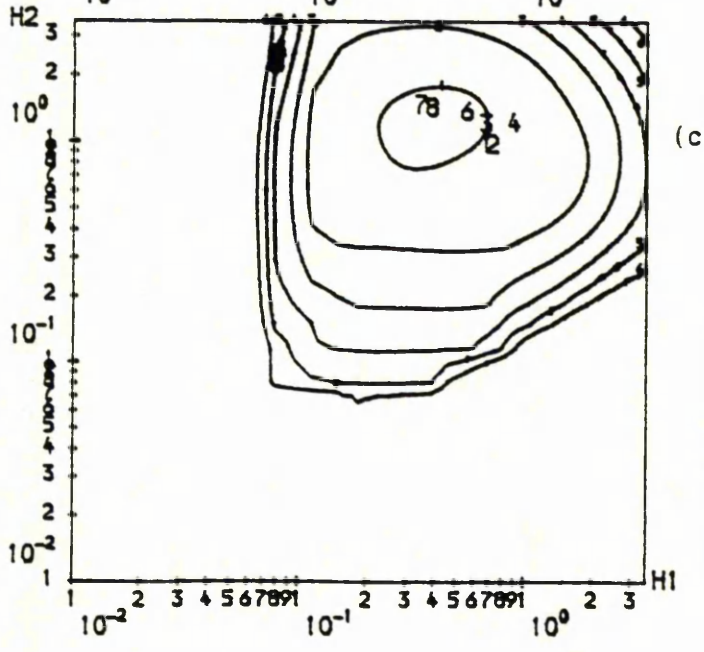
CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600

(b)



CONTOUR KEY	
1	0.0150
2	0.0250
3	0.0350
4	0.0450
5	0.0550
6	0.0650

(c)



CONTOUR KEY	
1	0.0080
2	0.0180
3	0.0280
4	0.0380
5	0.0480
6	0.0580

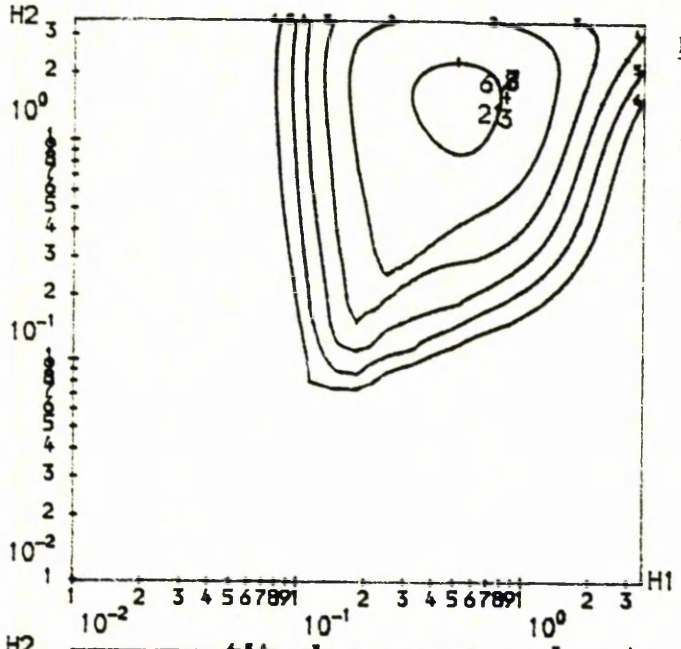
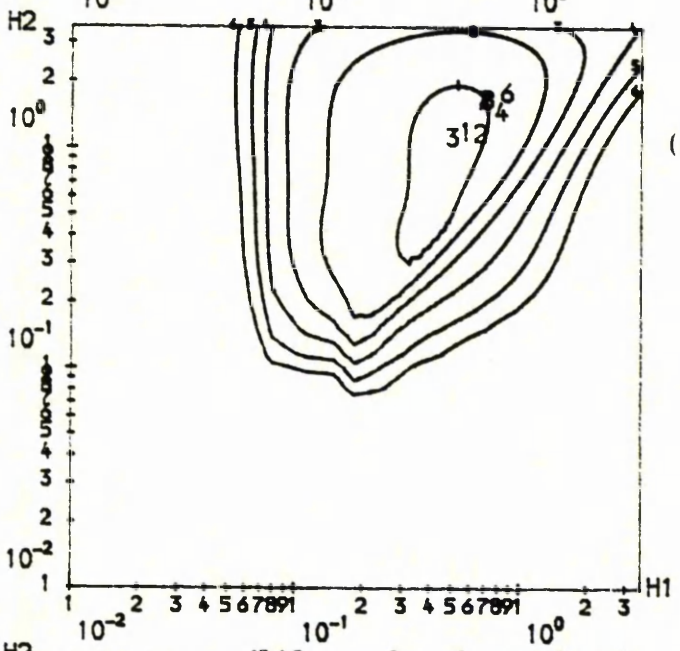


Figure 2.16

$(n_1, n_2; \mu, \sigma) = (25, 25; 0.00, 2)$

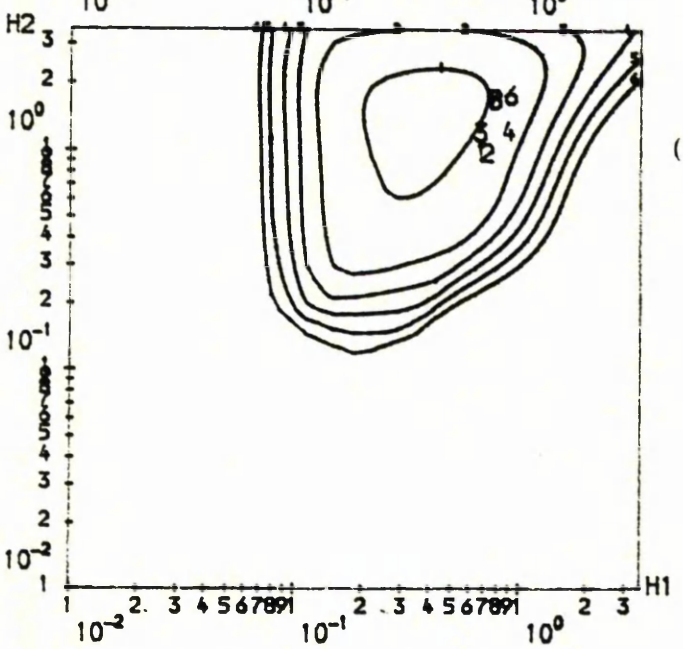
(a)

CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600



(b)

CONTOUR KEY	
1	0.0080
2	0.0180
3	0.0280
4	0.0380
5	0.0480
6	0.0580



(c)

CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575

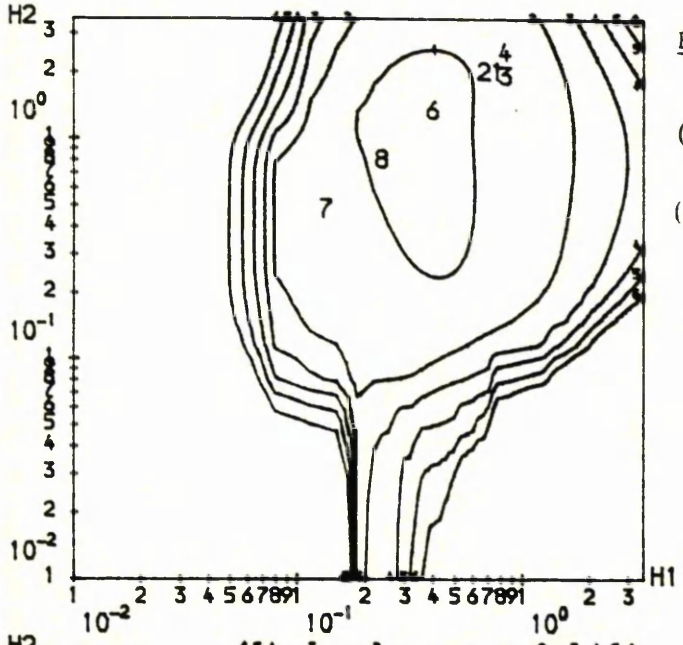
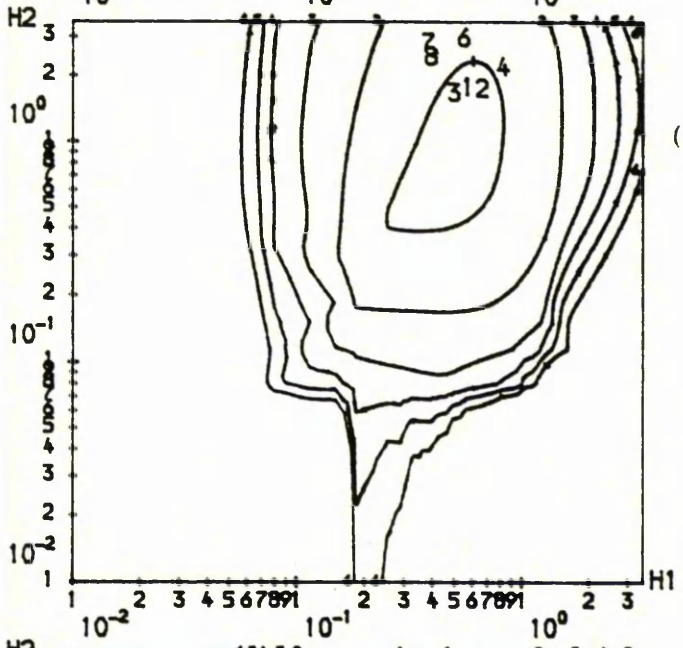


Figure 2.17

$(n_1, n_2; \mu, \sigma) = (25, 25; 6.26, 3)$

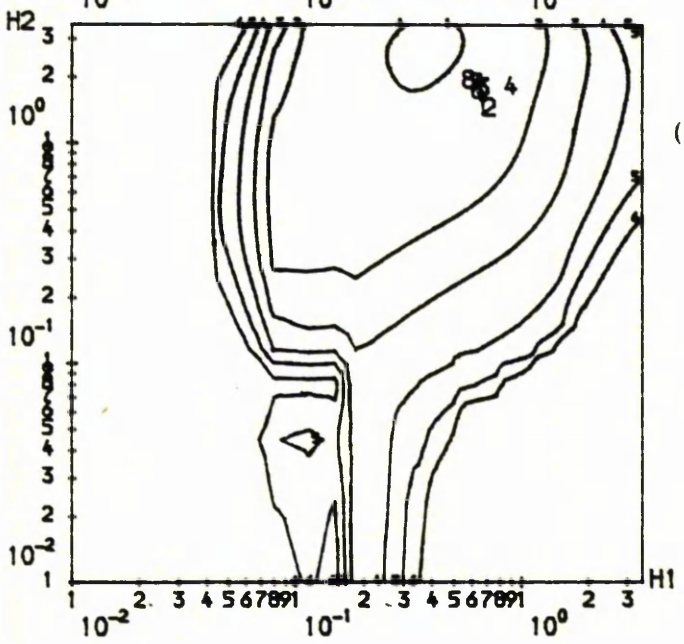
(a)

CONTOUR KEY	
1	0.0020
2	0.0120
3	0.0220
4	0.0320
5	0.0420
6	0.0520



(b)

CONTOUR KEY	
1	0.0060
2	0.0160
3	0.0260
4	0.0360
5	0.0460
6	0.0560



(c)

CONTOUR KEY	
1	0.0030
2	0.0130
3	0.0230
4	0.0330
5	0.0430
6	0.0530

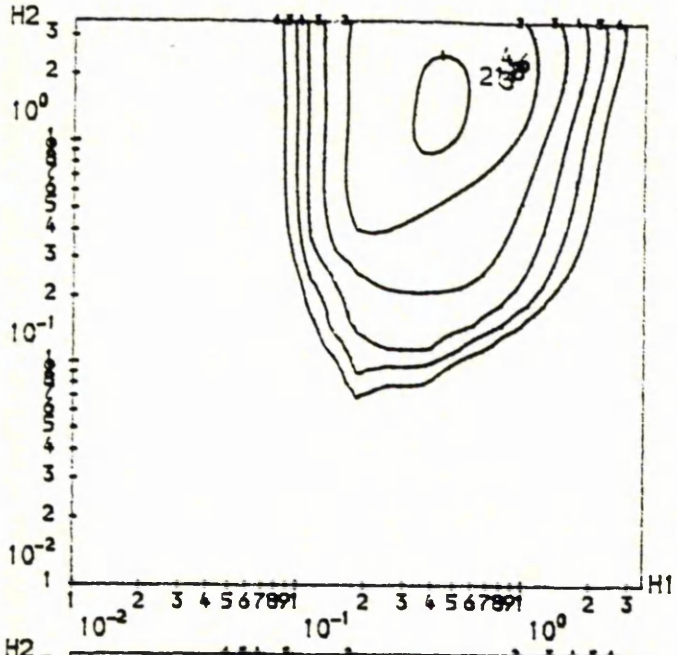
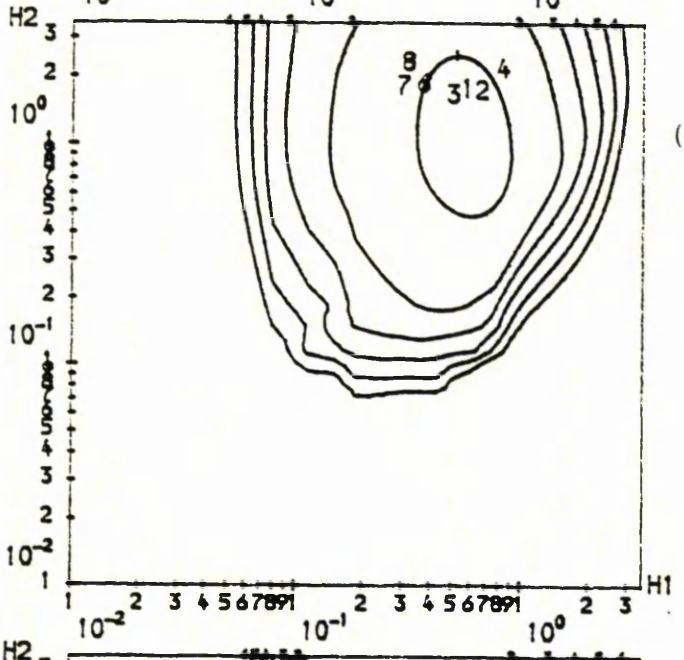


Figure 2.18

$(n_1, n_2; \mu, \sigma) = (25, 25; 2.40, 3)$

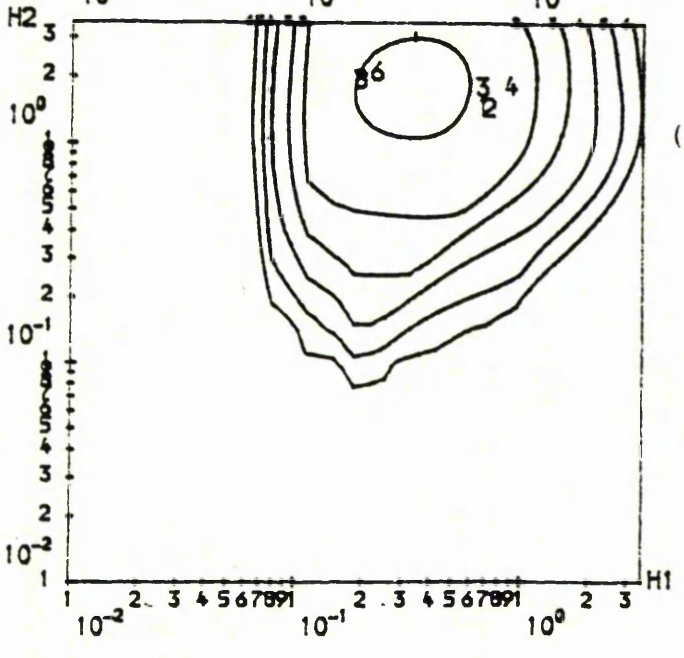
(a)

CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600



(b)

CONTOUR KEY	
1	0.0090
2	0.0190
3	0.0290
4	0.0390
5	0.0490
6	0.0590



(c)

CONTOUR KEY	
1	0.0080
2	0.0180
3	0.0280
4	0.0380
5	0.0480
6	0.0580

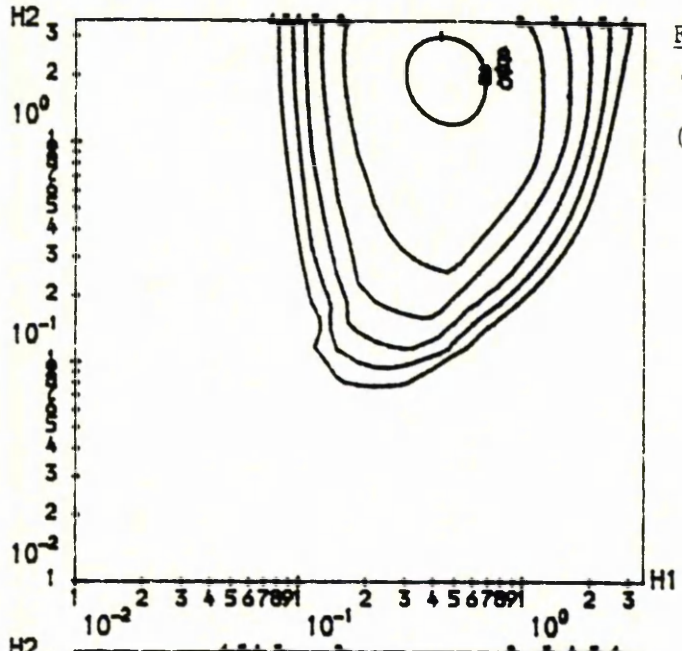
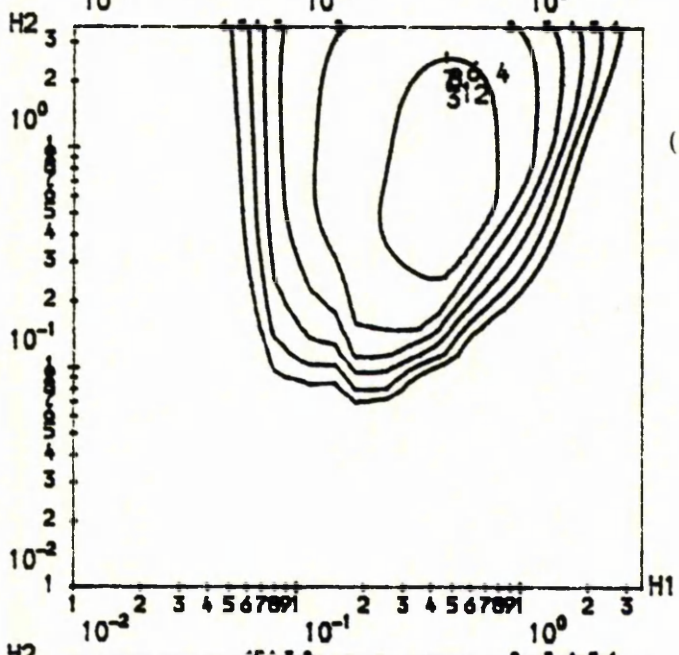


Figure 2.19

$(n_1, n_2; \mu, \sigma) = (25, 25; 0.00, 3)$

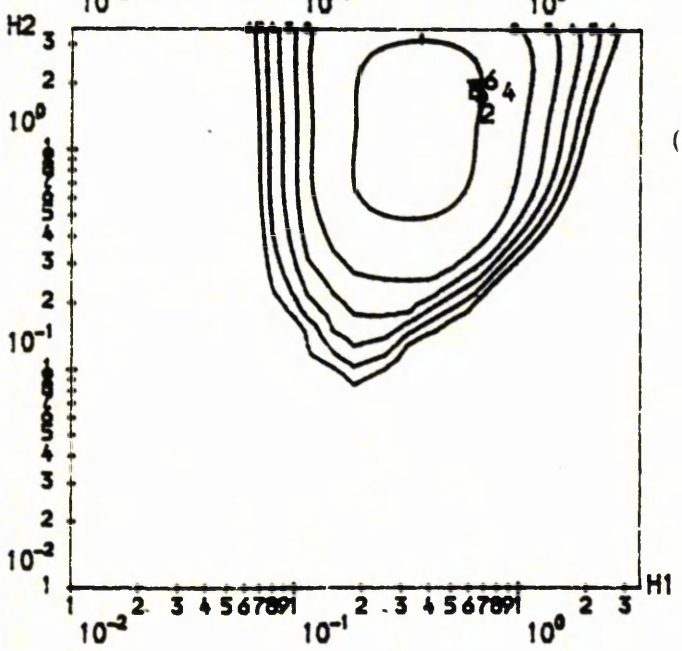
(a) 1- .772, 1.944 6- .675, 1.944
2- .672, 1.942 7- .815, 1.959
3- .837, 1.857 8- .814, 2.347
4- .830, 2.390

CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600



(b)

CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575



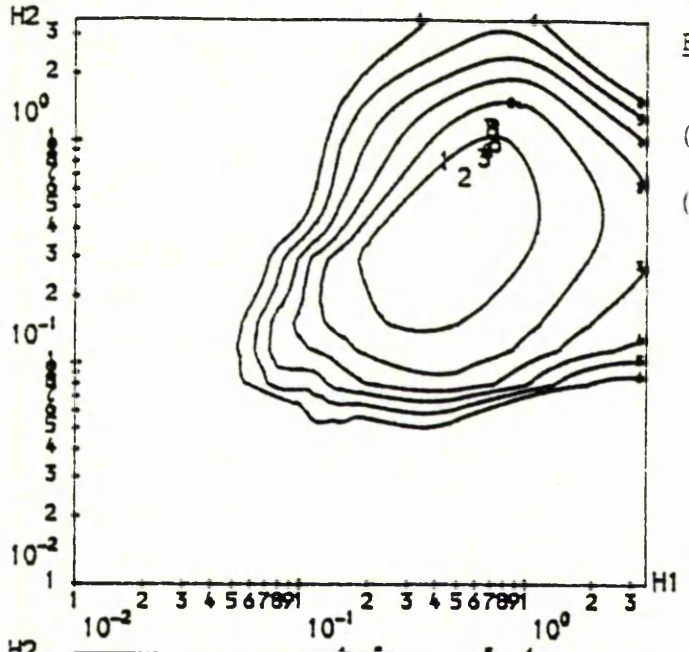
(c)

CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575

Figure 2.20

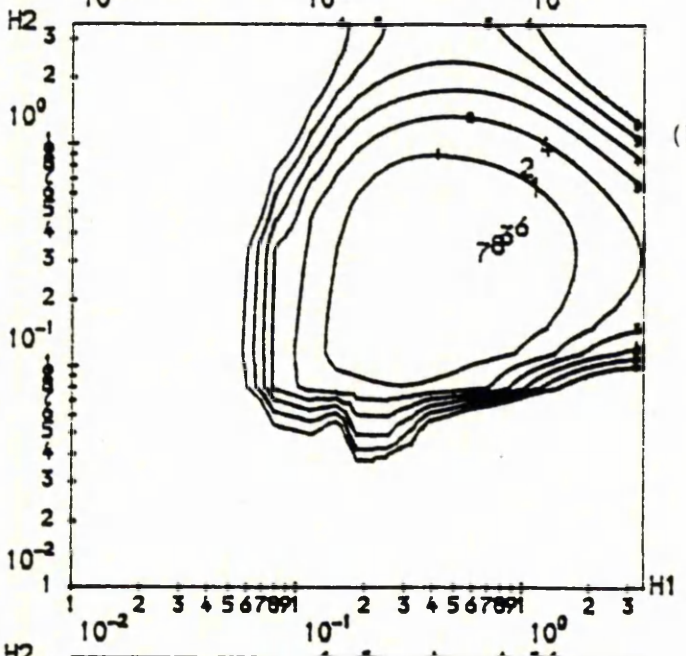
$(n_1, n_2; a, \sigma) = (10, 25; 3.16, 1)$

(a)



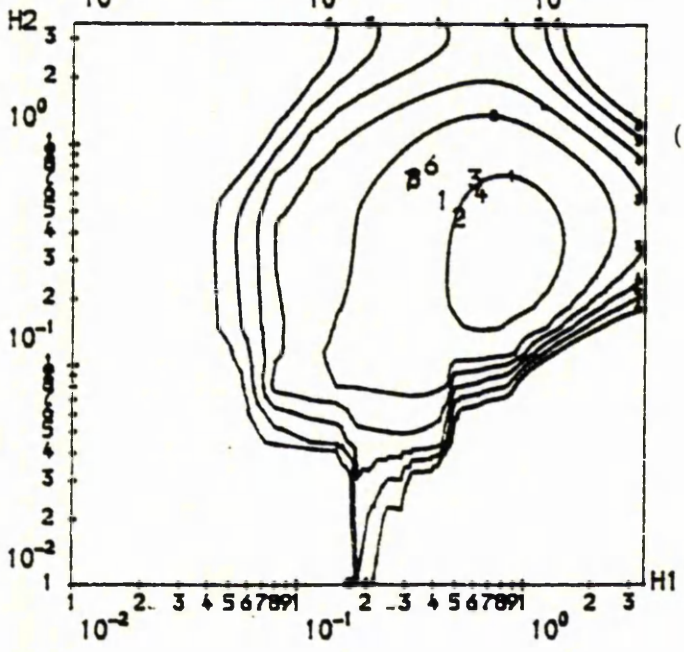
CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600

(b)



CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600

(c)

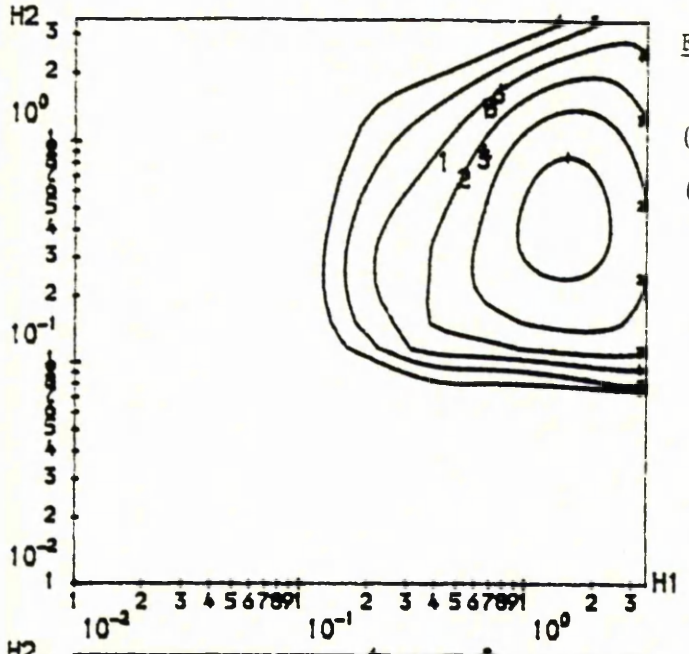


CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550

Figure 2.21

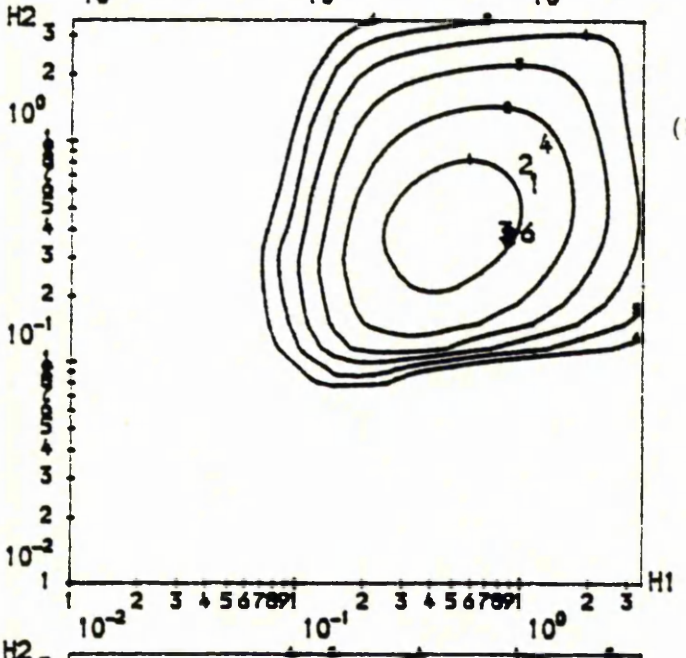
$(n_1, n_2; \mu, \sigma) = (10, 25; 1.40, 1)$

(a)



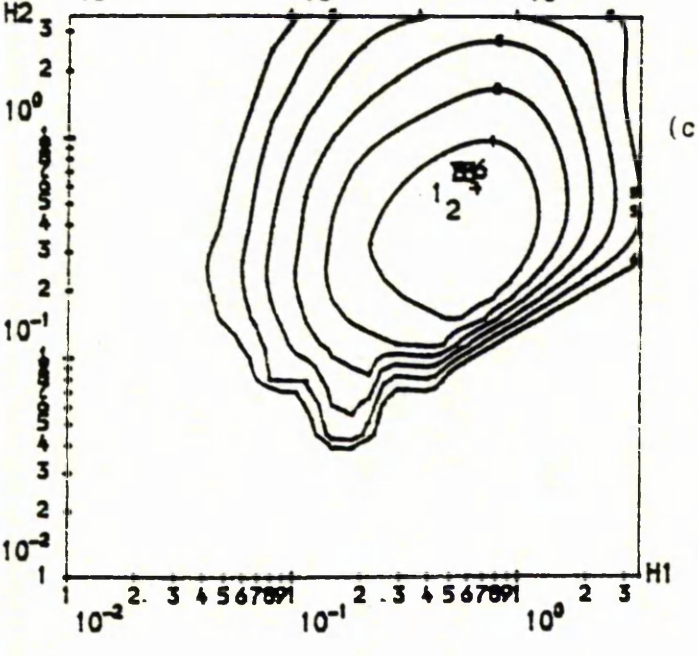
CONTOUR KEY	
1	0.0200
2	0.0300
3	0.0400
4	0.0500
5	0.0600
6	0.0700

(b)



CONTOUR KEY	
1	0.0200
2	0.0300
3	0.0400
4	0.0500
5	0.0600
6	0.0700

(c)

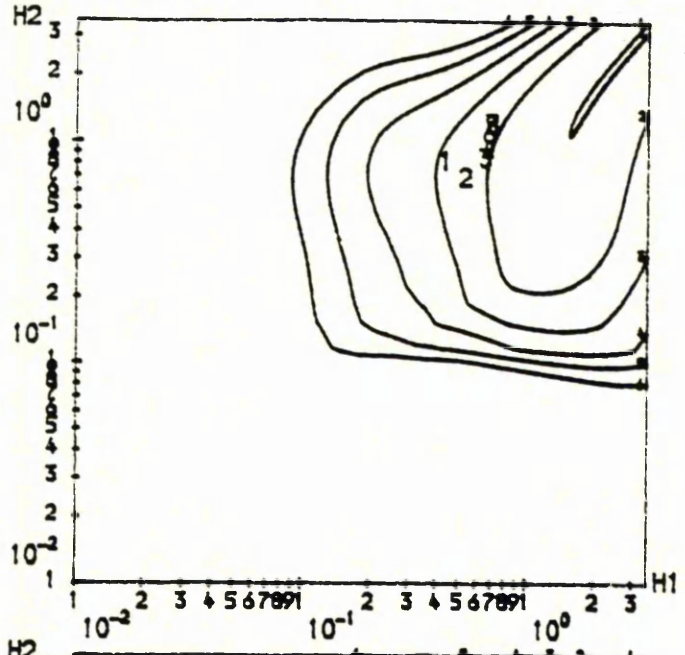


CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600

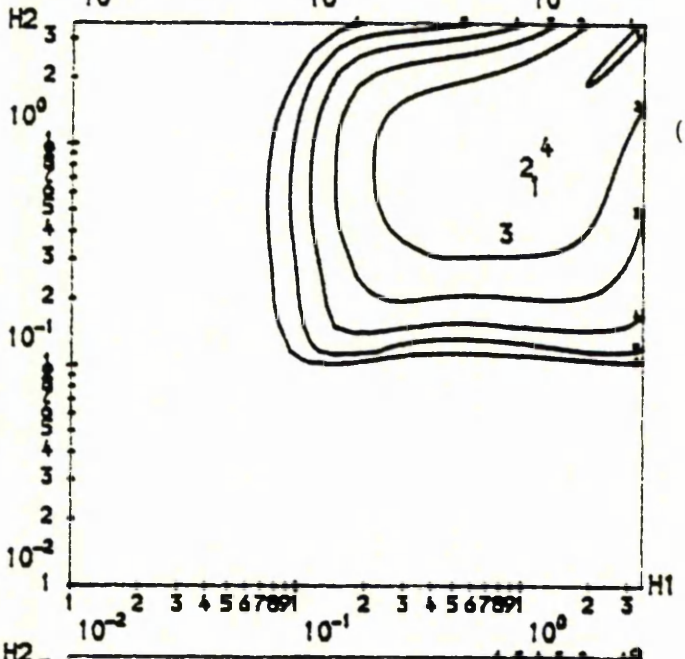
Figure 2.22

$(n_1, n_2; \mu, \sigma) = (10, 25; 0.00, 1)$

(a)



CONTOUR KEY	
1	0. 0001
2	0. 0101
3	0. 0201
4	0. 0301
5	0. 0401
6	0. 0501

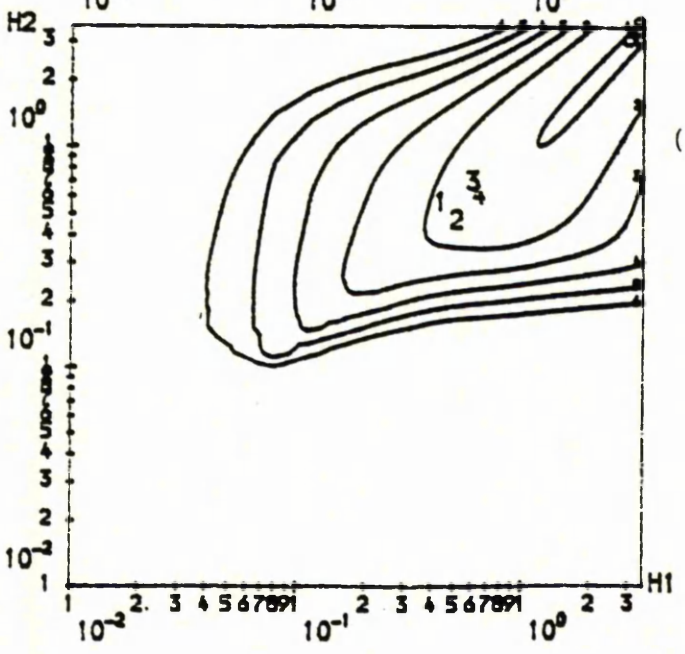


(b) 6- 3.994, 3.964

7- 5.513, 5.516

8- 6.937, 6.936

CONTOUR KEY	
1	0. 0002
2	0. 0102
3	0. 0202
4	0. 0302
5	0. 0402
6	0. 0502



(c) 6- 2.989, 3.057

7- 5.636, 5.666

8- 3.333, 3.415

CONTOUR KEY	
1	0. 0005
2	0. 0105
3	0. 0205
4	0. 0305
5	0. 0405
6	0. 0505

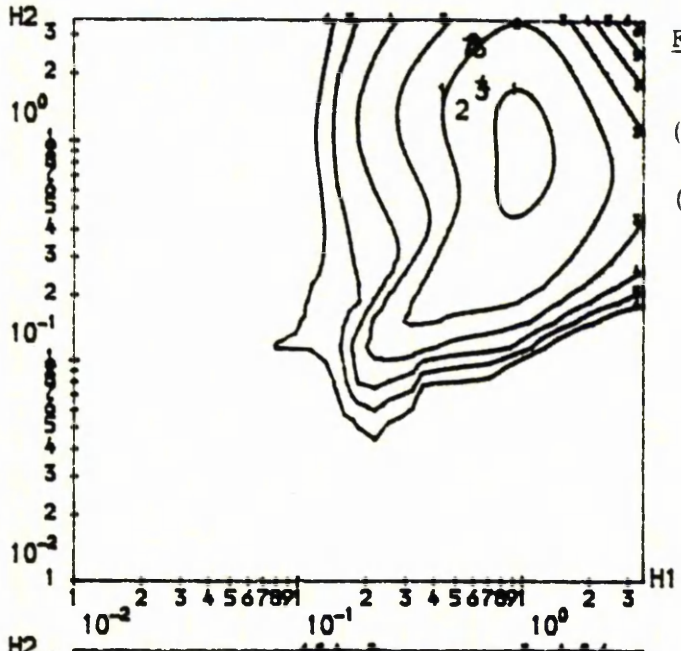
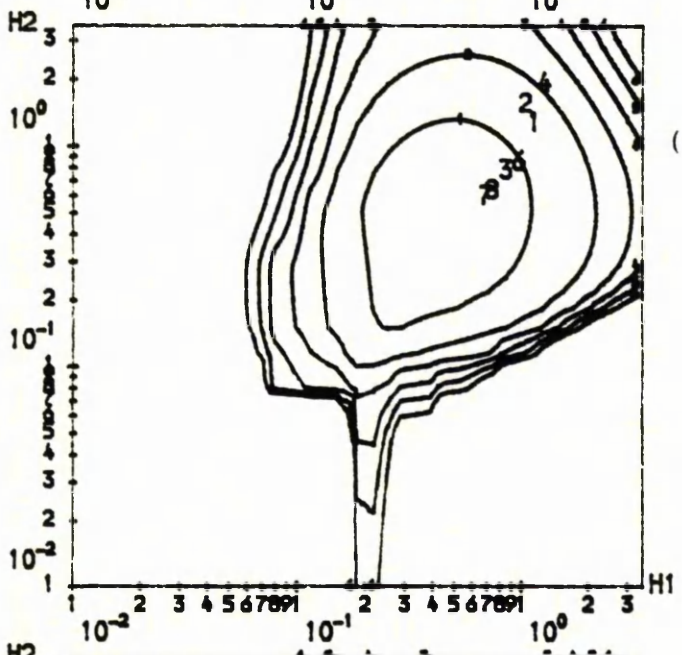


Figure 2.23

$(n_1, n_2; \mu, \sigma) = (10, 25; 4.92, 2)$

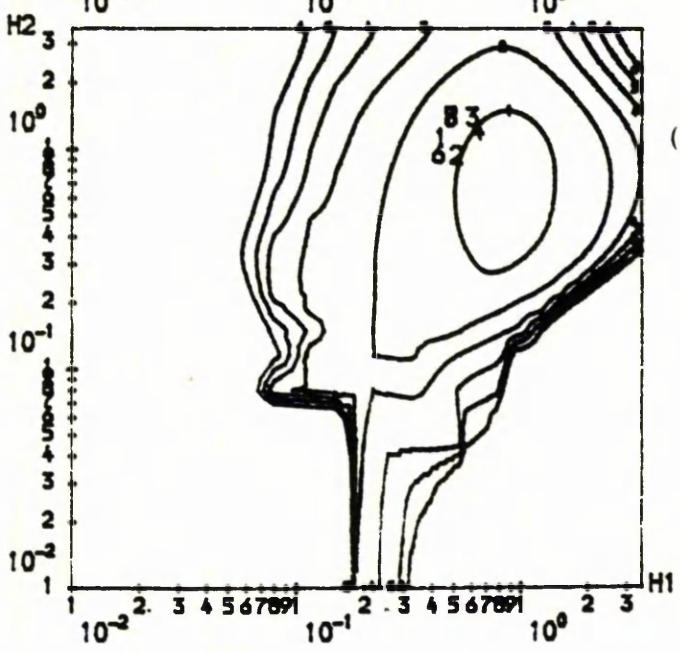
(a)

CONTOUR KEY	
1	0.0125
2	0.0225
3	0.0325
4	0.0425
5	0.0525
6	0.0625



(b)

CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575



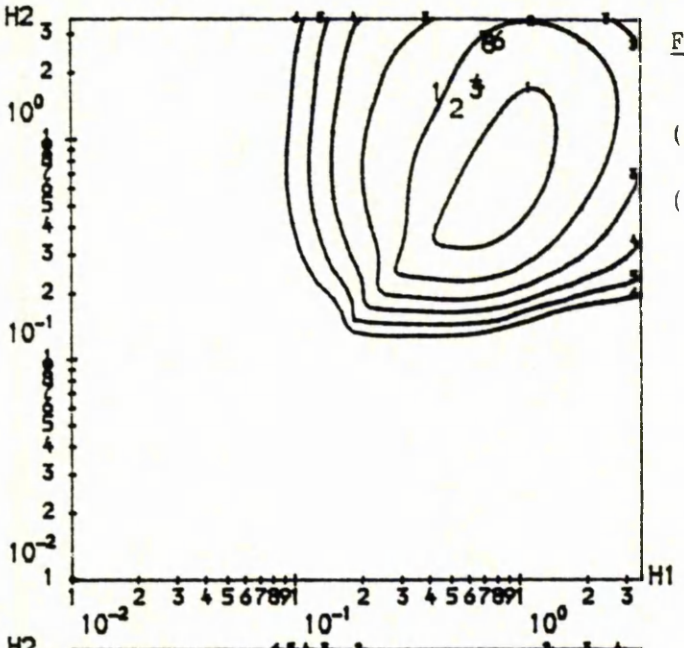
(c)

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550

Figure 2.24

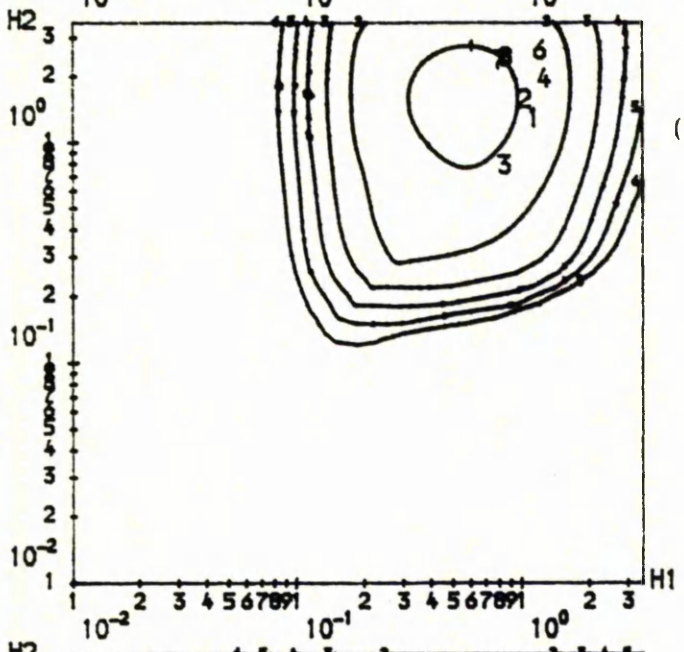
$(n_1, n_2; \mu, \sigma) = (10, 25; 2.47, 2)$

(a)



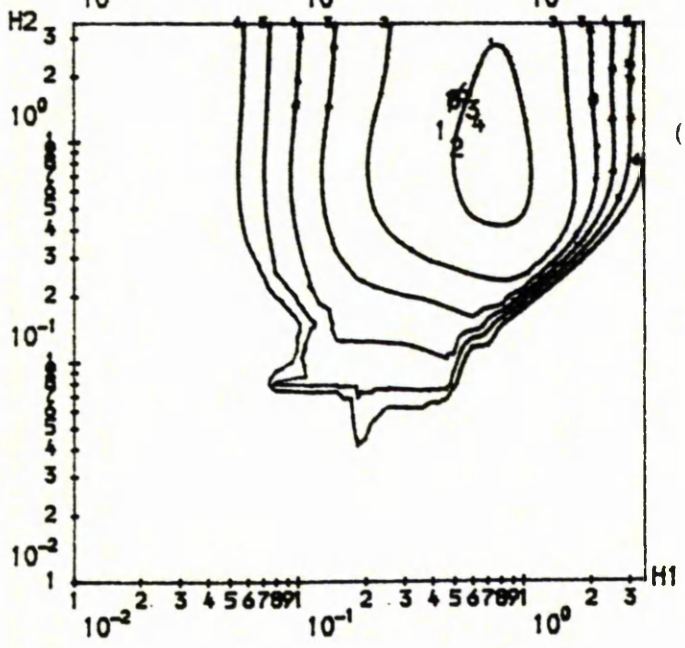
CONTOUR KEY	
1	0.0200
2	0.0300
3	0.0400
4	0.0500
5	0.0600
6	0.0700

(b)



CONTOUR KEY	
1	0.0150
2	0.0250
3	0.0350
4	0.0450
5	0.0550
6	0.0650

(c)



CONTOUR KEY	
1	0.0040
2	0.0140
3	0.0240
4	0.0340
5	0.0440
6	0.0540

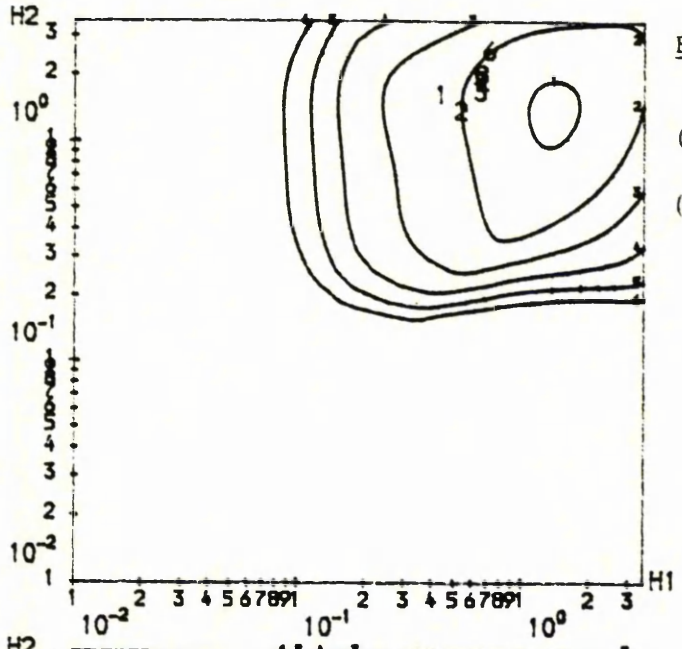
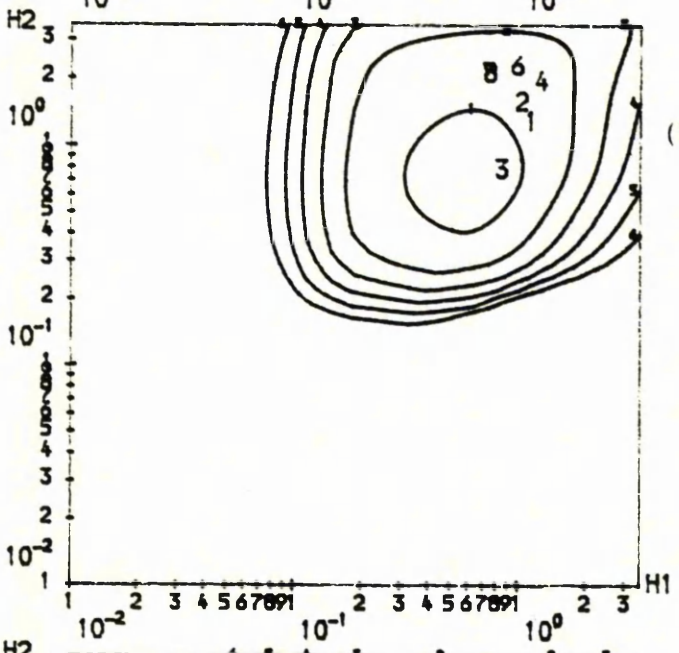


Figure 2.25

$(n_1, n_2; \mu, \sigma) = (10, 25; 1.16, 2)$

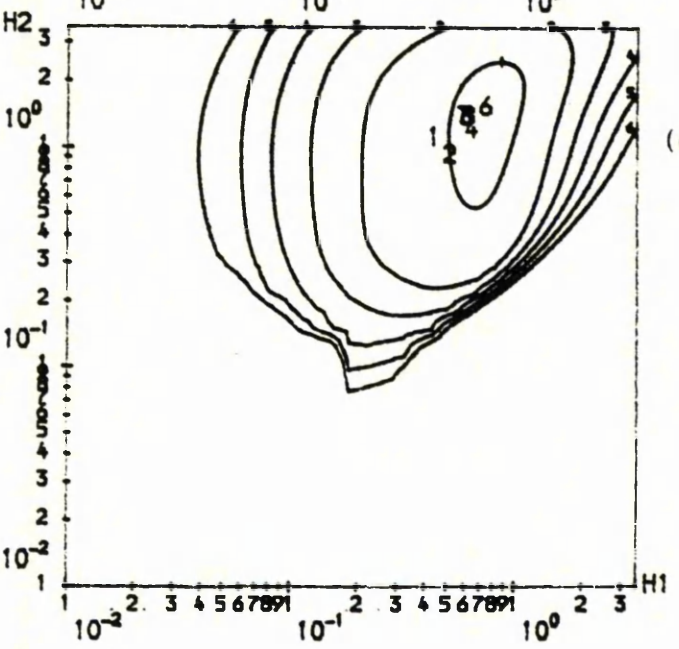
(a)

CONTOUR KEY	
1	0.0150
2	0.0250
3	0.0350
4	0.0450
5	0.0550
6	0.0650



(b)

CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575



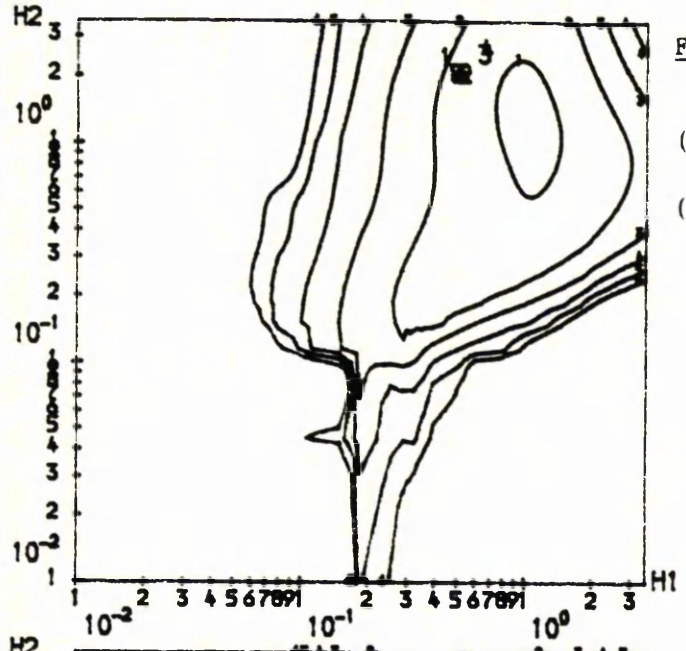
(c)

CONTOUR KEY	
1	0.0025
2	0.0125
3	0.0225
4	0.0325
5	0.0425
6	0.0525

Figure 2.26

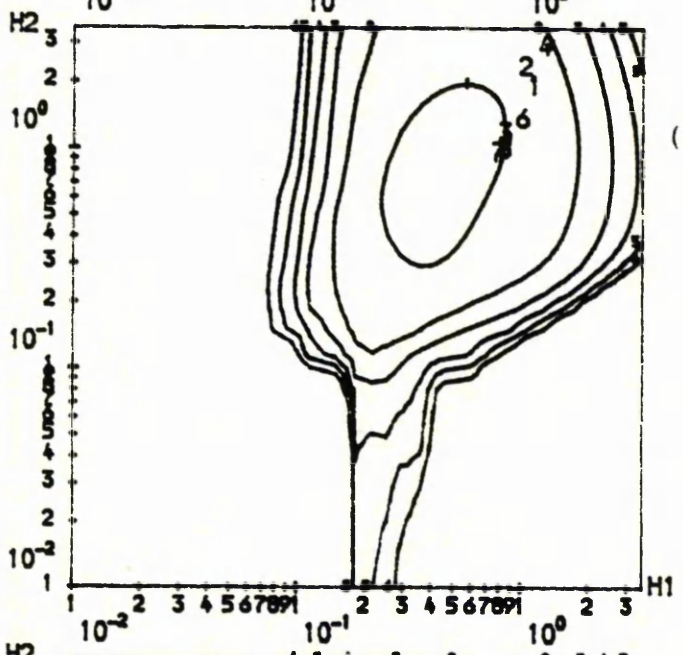
$(n_1, n_2; \mu, \sigma) = (10, 25; 6.56, 3)$

(a)



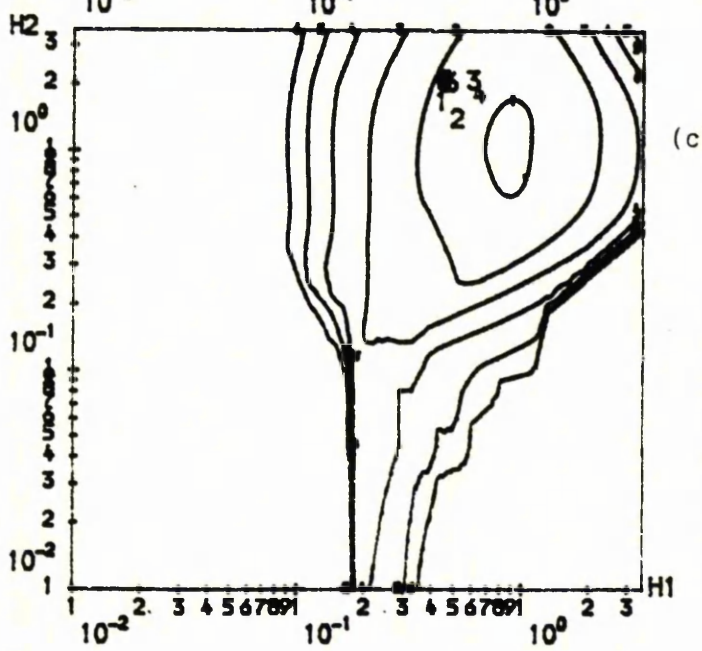
CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600

(b)



CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575

(c)



CONTOUR KEY	
1	0.0025
2	0.0125
3	0.0225
4	0.0325
5	0.0425
6	0.0525

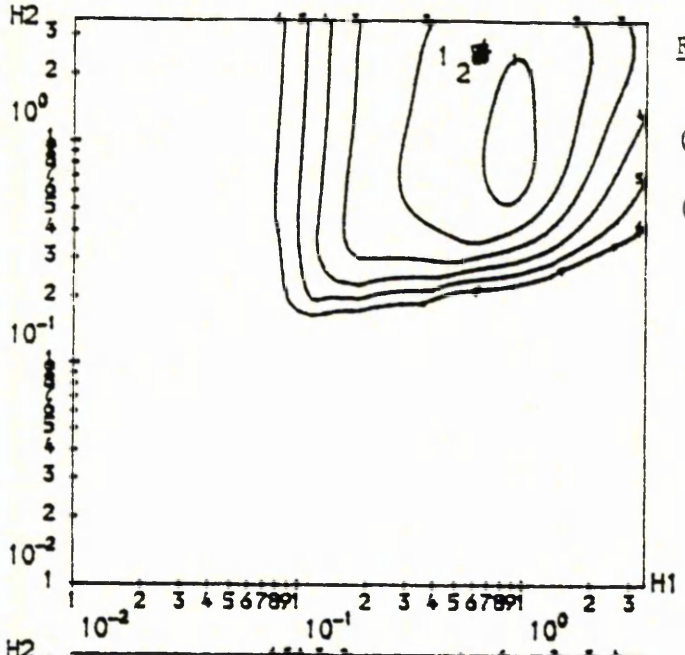
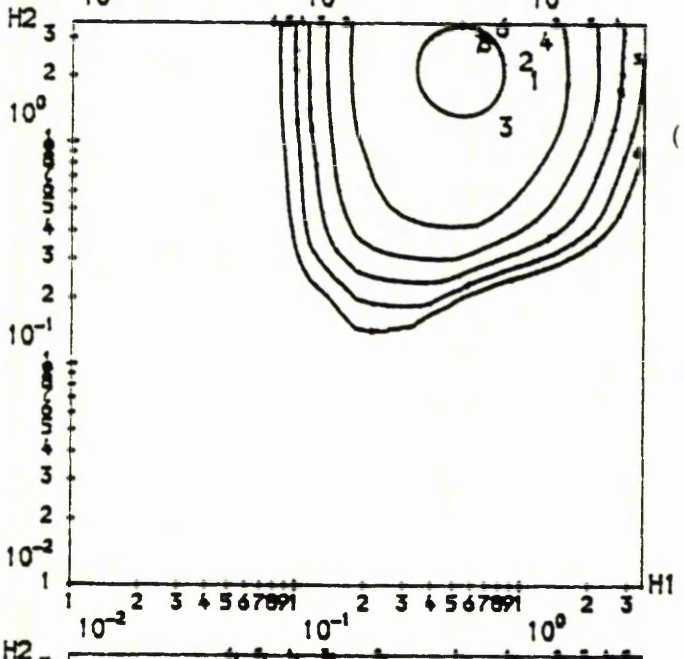


Figure 2.27

$(n_1, n_2; \mu, \sigma) = (10, 25; 2.92, 3)$

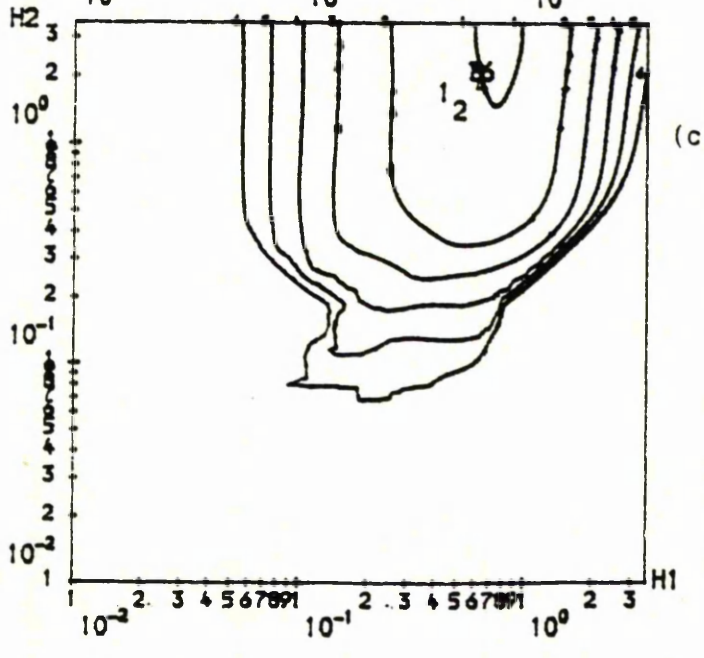
(a)

CONTOUR KEY	
1	0.0175
2	0.0275
3	0.0375
4	0.0475
5	0.0575
6	0.0675



(b)

CONTOUR KEY	
1	0.0090
2	0.0190
3	0.0290
4	0.0390
5	0.0490
6	0.0590



(c)

CONTOUR KEY	
1	0.0015
2	0.0115
3	0.0215
4	0.0315
5	0.0415
6	0.0515

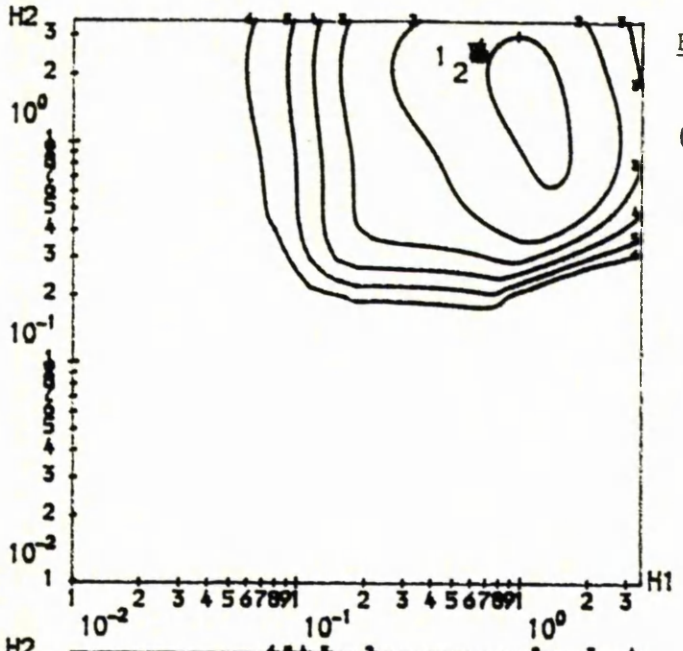
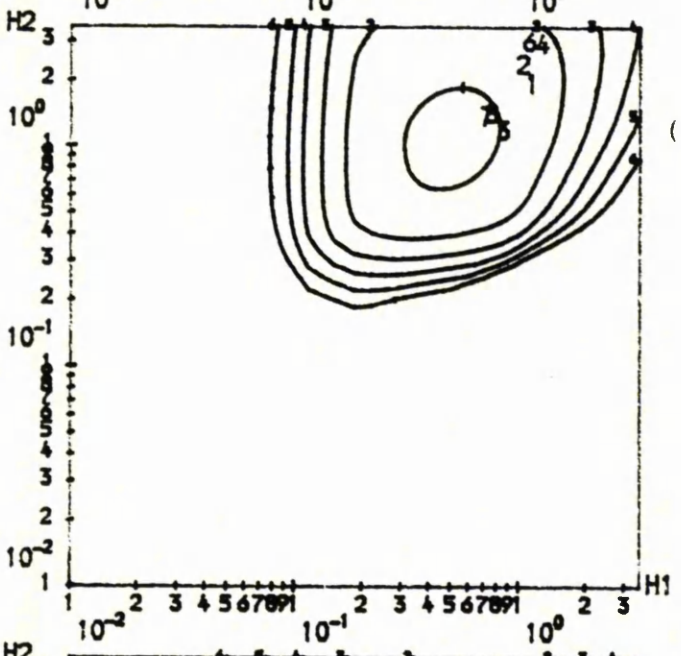


Figure 2.28

$(n_1, n_2; \mu, \sigma) = (10, 25; 0.00, 3)$

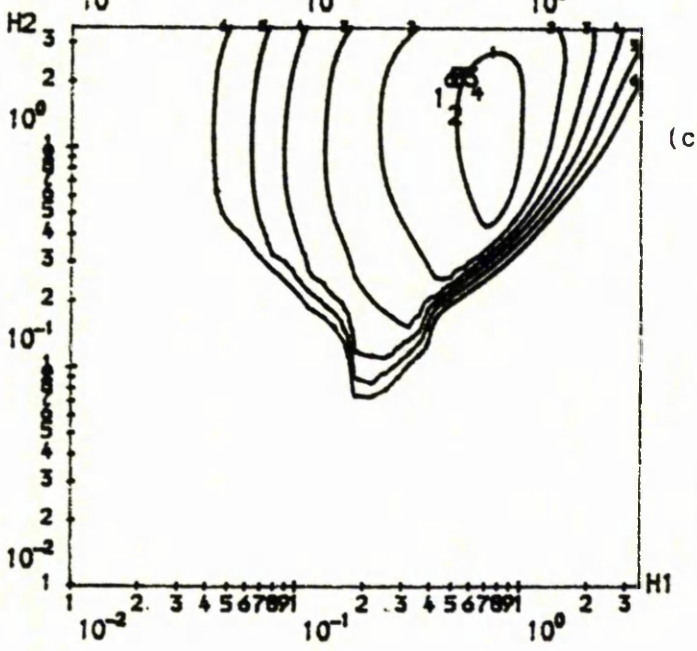
(a)

CONTOUR KEY	
1	0.0125
2	0.0225
3	0.0325
4	0.0425
5	0.0525
6	0.0625



(b)

CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575



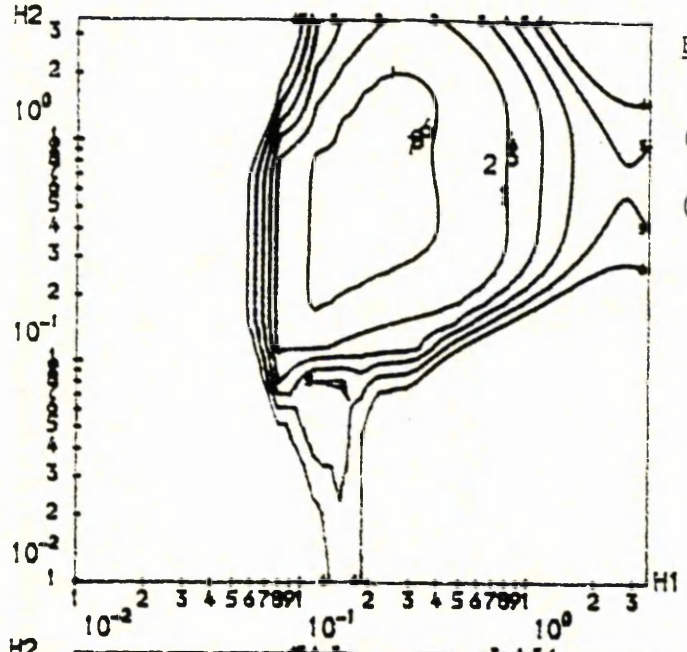
(c)

CONTOUR KEY	
1	0.0025
2	0.0125
3	0.0225
4	0.0325
5	0.0425
6	0.0525

Figure 2.29

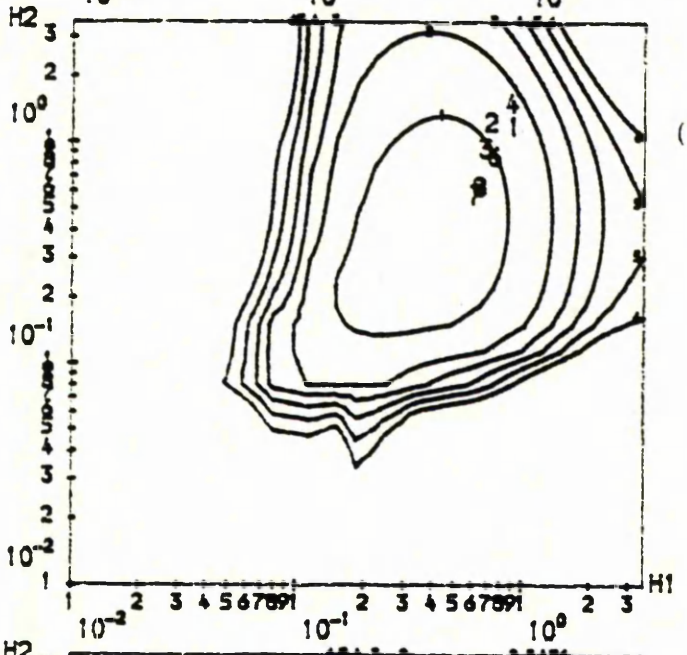
$(n_1, n_2: \mu, \sigma) = (25, 10: 3.16, 1)$

(a)



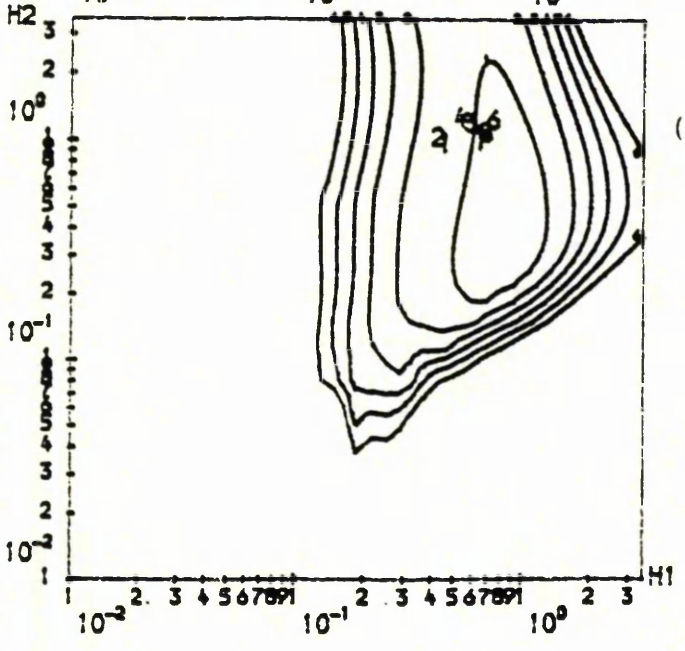
CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575

(b)



CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575

(c)

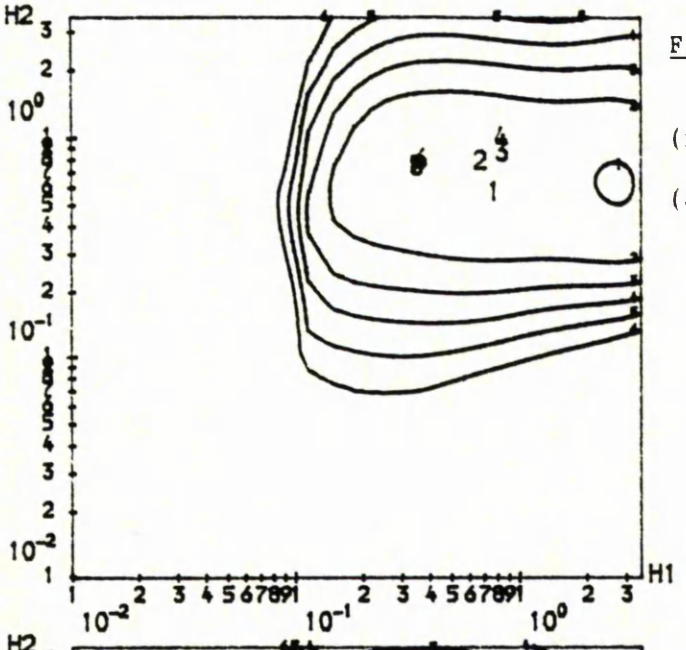


CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575

Figure 2.30

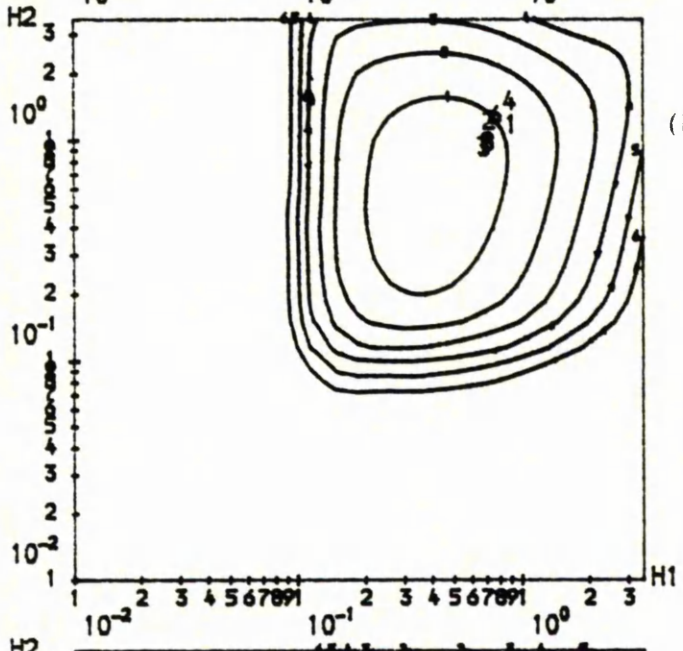
$(n_1, n_2; \mu, \sigma) = (25, 10; 1.40, 1)$

(a)



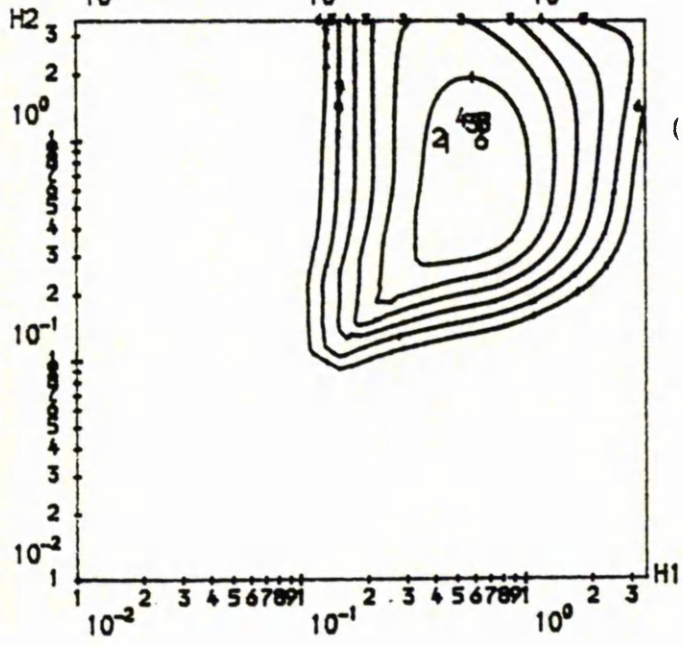
CONTOUR KEY	
1	0.0150
2	0.0250
3	0.0350
4	0.0450
5	0.0550
6	0.0650

(b)



CONTOUR KEY	
1	0.0150
2	0.0250
3	0.0350
4	0.0450
5	0.0550
6	0.0650

(c)



CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600

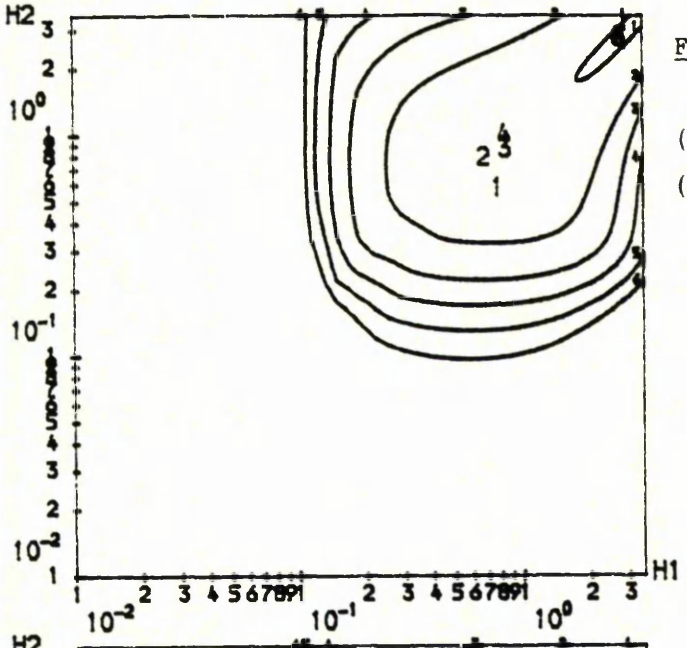
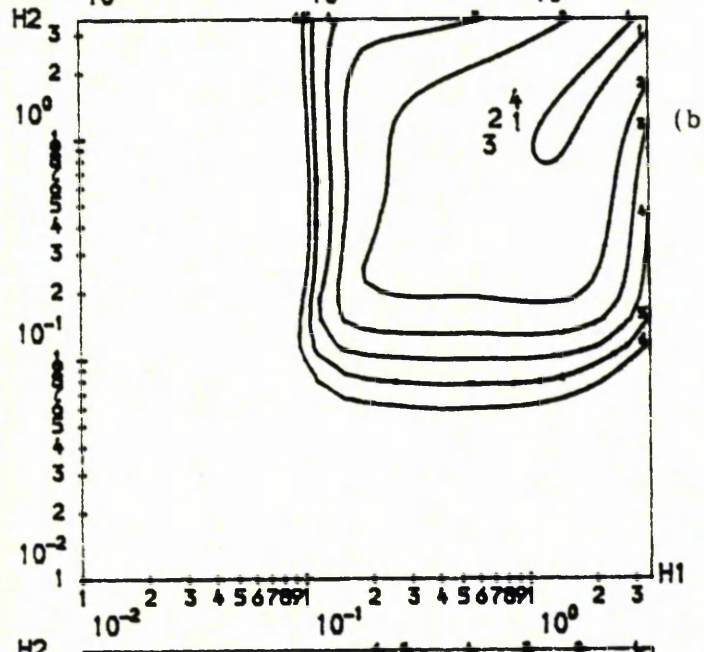


Figure 2.31

$(n_1, n_2; \mu, \sigma) = (25, 10; 0.00, 1)$

(a)

CONTOUR KEY	
1	0.0005
2	0.0105
3	0.0205
4	0.0305
5	0.0405
6	0.0505

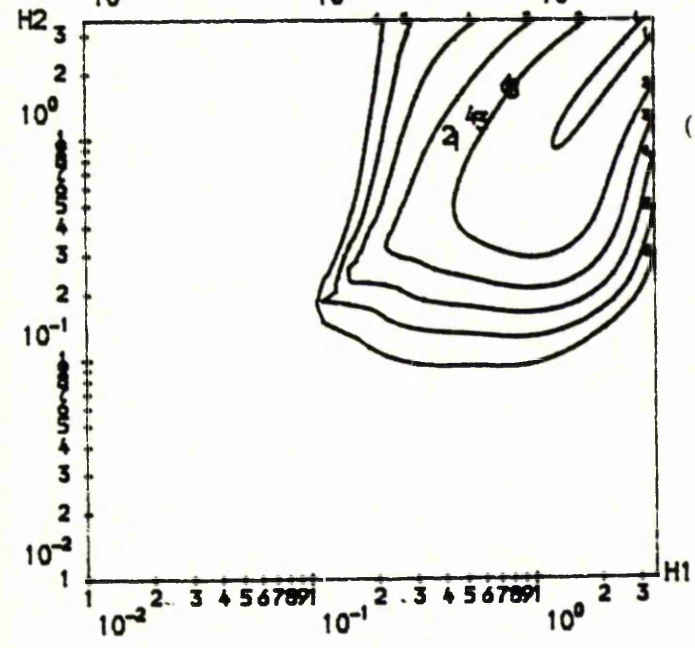


(b) 6- 5.381, 5.405

7- 5.990, 5.996

8- 6.586, 6.622

CONTOUR KEY	
1	0.0005
2	0.0105
3	0.0205
4	0.0305
5	0.0405
6	0.0505



(c)

CONTOUR KEY	
1	0.0005
2	0.0105
3	0.0205
4	0.0305
5	0.0405
6	0.0505

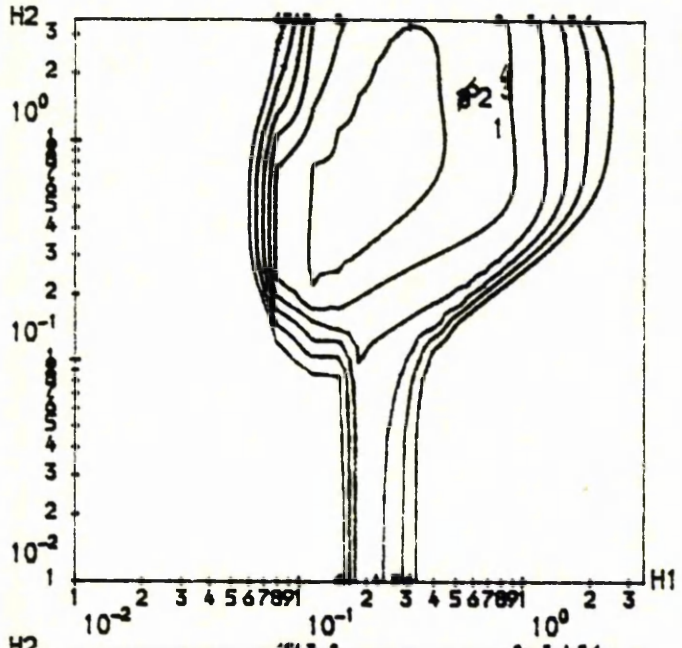
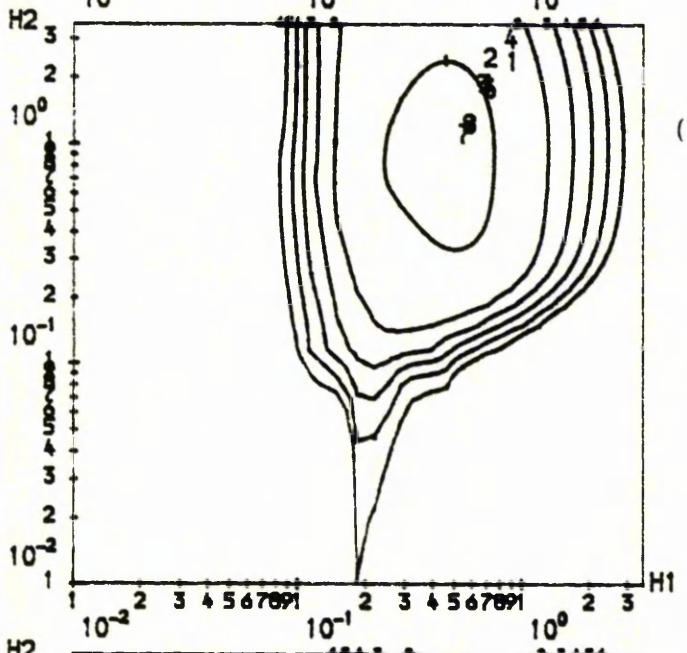


Figure 2.32

$(n_1, n_2; \mu, \sigma) = (25, 10; 4.35, 2)$

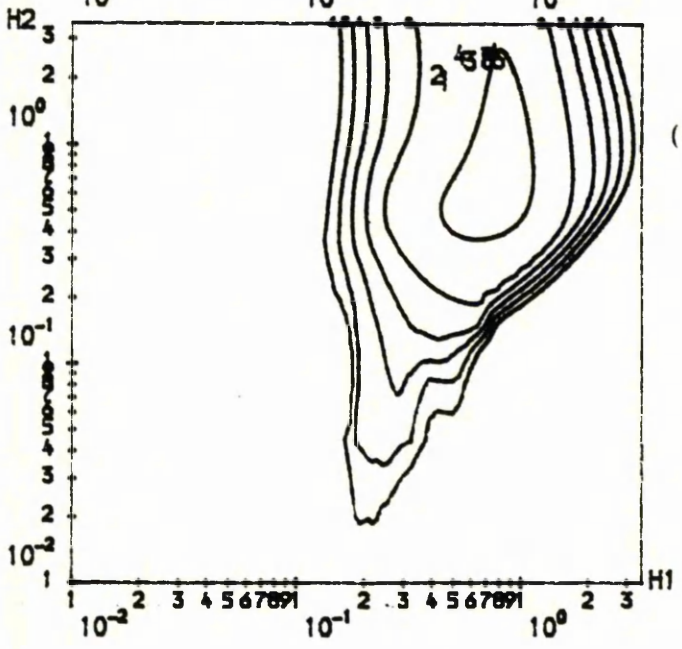
(a)

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550



(b)

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550



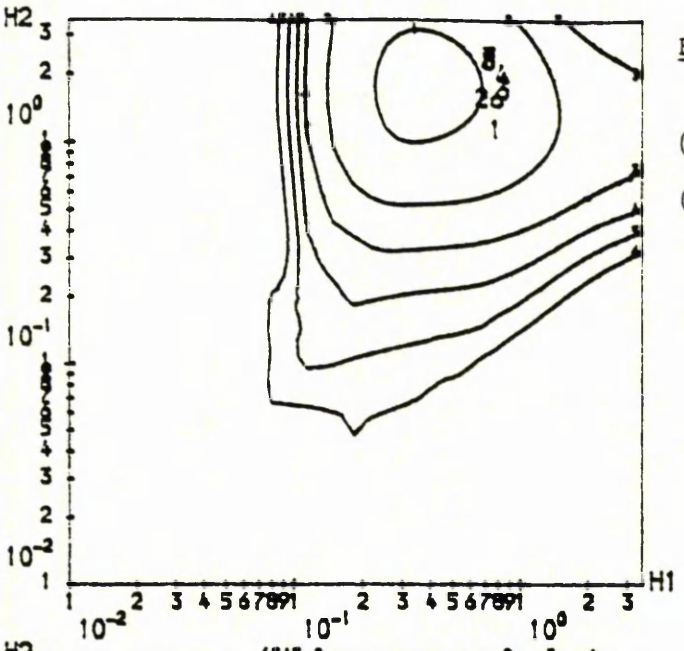
(c)

CONTOUR KEY	
1	0.0035
2	0.0135
3	0.0235
4	0.0335
5	0.0435
6	0.0535

Figure 2.33

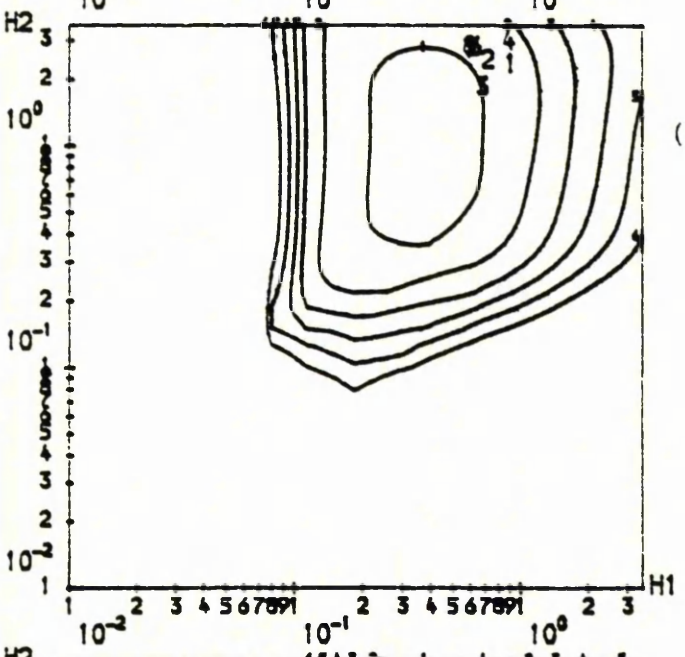
$(n_1, n_2; \mu, \sigma) = (25, 10; 1.14, 2)$

(a)



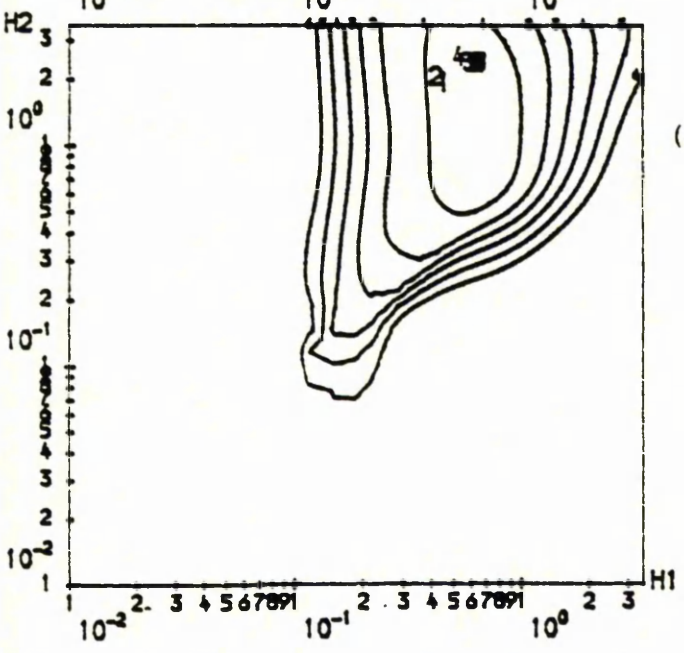
CONTOUR KEY	
1	0.0150
2	0.0250
3	0.0350
4	0.0450
5	0.0550
6	0.0650

(b)



CONTOUR KEY	
1	0.0095
2	0.0195
3	0.0295
4	0.0395
5	0.0495
6	0.0595

(c)

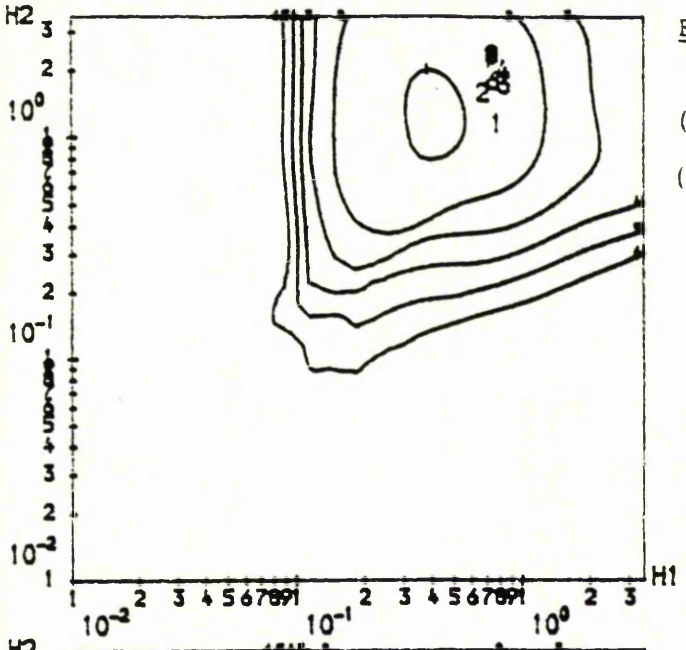


CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550

Figure 2.34

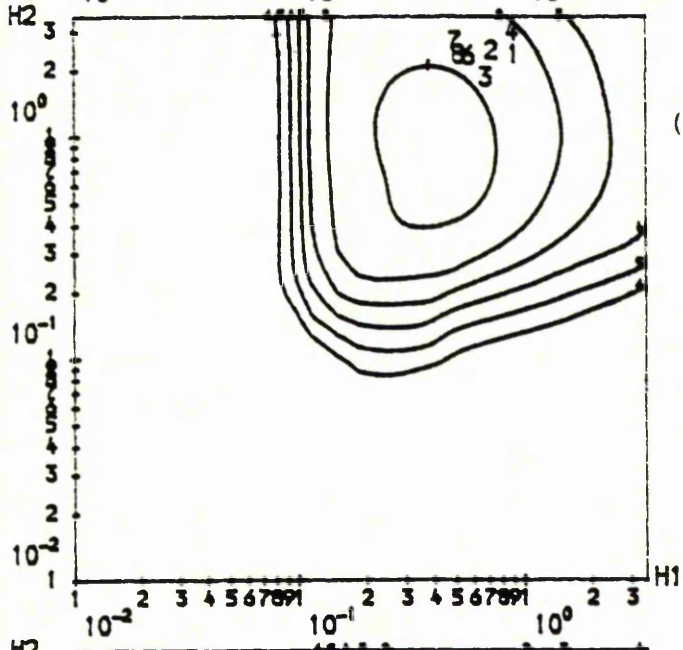
$(n_1, n_2; \mu, \sigma) = (25, 10; 0.00, 2)$

(a)



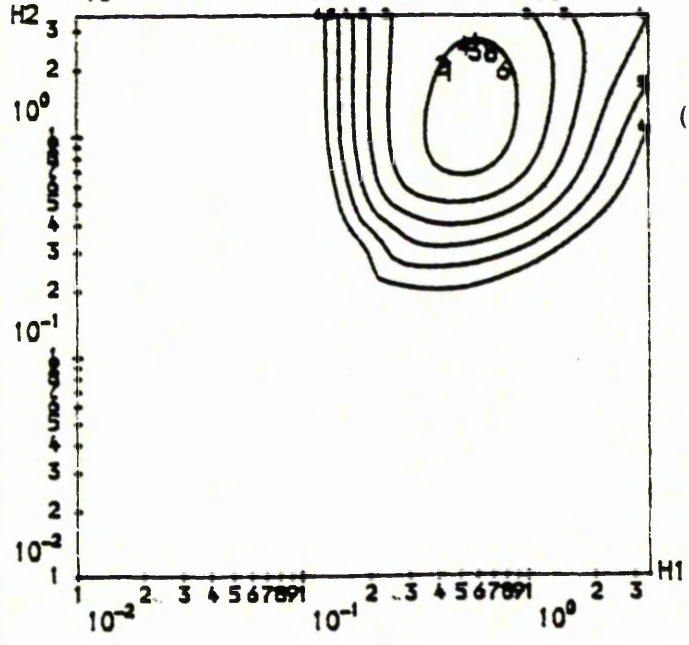
CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600

(b)



CONTOUR KEY	
1	0.0075
2	0.0175
3	0.0275
4	0.0375
5	0.0475
6	0.0575

(c)



CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550

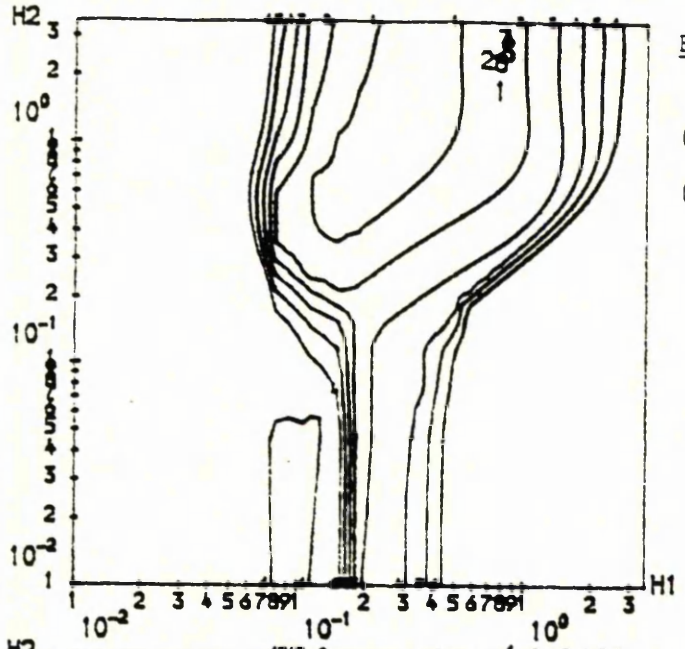
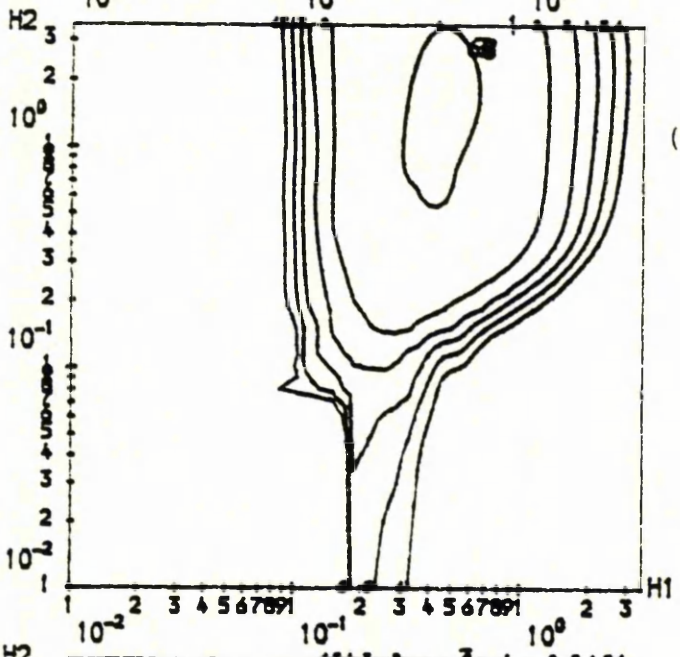


Figure 2.35

$(n_1, n_2; \mu, \sigma) = (25, 10; 5.42, 3)$

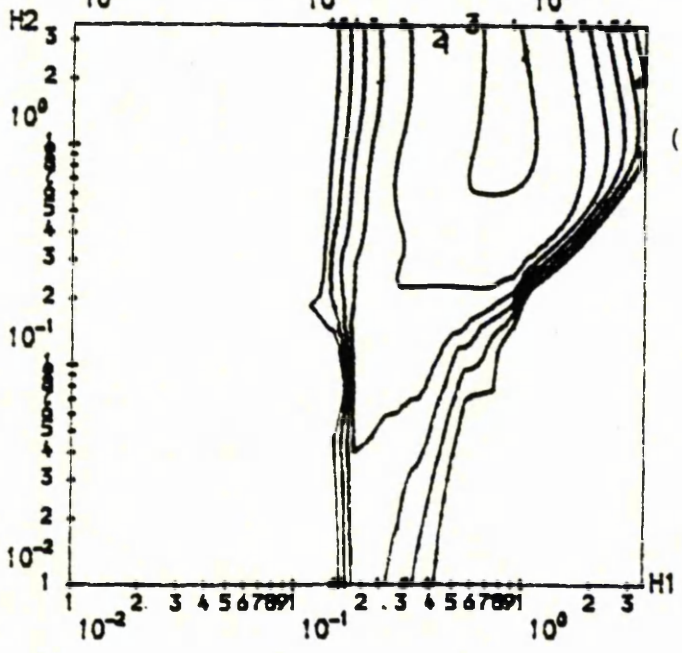
(a)

CONTOUR KEY	
1	0.0050
2	0.0150
3	0.0250
4	0.0350
5	0.0450
6	0.0550



(b) 1- .902, 3.433
2- .717, 3.578
4- .885, 4.372

CONTOUR KEY	
1	0.0035
2	0.0135
3	0.0235
4	0.0335
5	0.0435
6	0.0535



(c) 3- .591, 3.465 6- .764, 3.609
4- .531, 3.676 7- .677, 3.914
8- .692, 4.139

CONTOUR KEY	
1	0.0025
2	0.0125
3	0.0225
4	0.0325
5	0.0425
6	0.0525

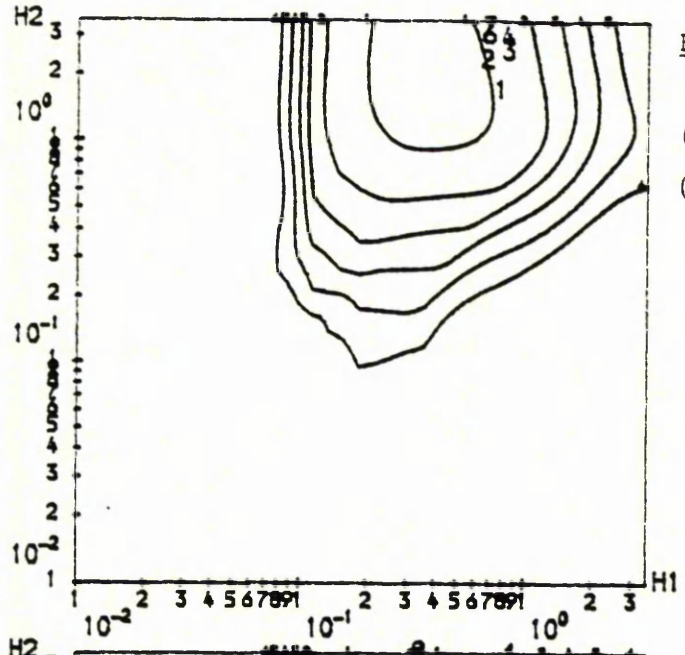
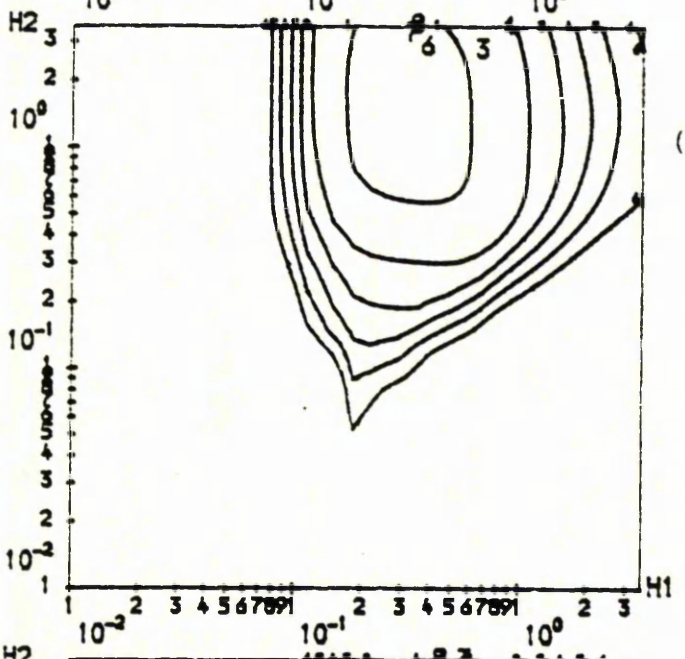


Figure 2.36

$(n_1, n_2; \mu, \sigma) = (25, 10; 0.00, 3)$

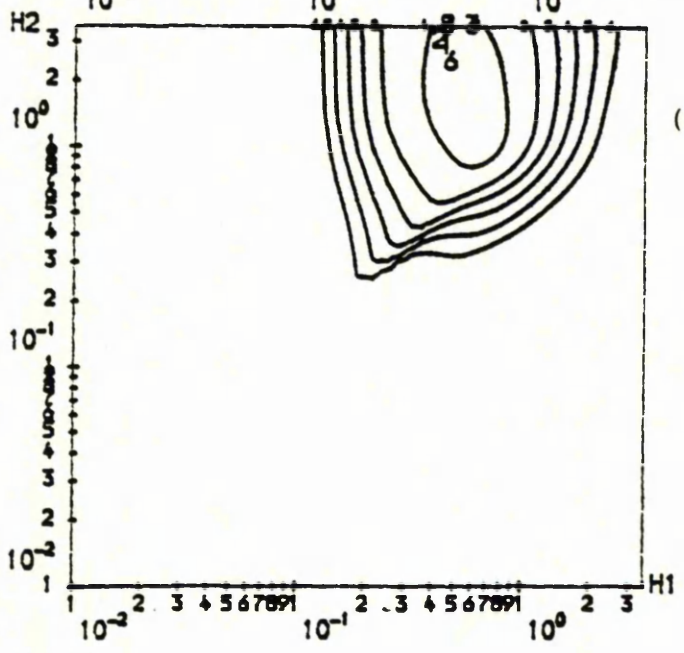
(a) 7- .688, 3.240
8- .685, 3.920

CONTOUR KEY	
1	0.0150
2	0.0250
3	0.0350
4	0.0450
5	0.0550
6	0.0650



(b) 1- .902, 3.433 7- .330, 3.211
2- .717, 3.579 8- .348, 3.497
4- .885, 4.372

CONTOUR KEY	
1	0.0100
2	0.0200
3	0.0300
4	0.0400
5	0.0500
6	0.0600



(c) 3- .591, 3.465 7- .458, 3.455
4- .531, 3.676 8- .457, 3.454

CONTOUR KEY	
1	0.0040
2	0.0140
3	0.0240
4	0.0340
5	0.0440
6	0.0540

Discussion of Figures 2.2-2.36

Several trends may be detected from the contour plots. Comparing equal-sized samples with n_i=10 and 25, the contours of the former are always closer, at least near the centre, than those of the latter, especially noticeable as the ratio of the standard deviations, σ, becomes more extreme (c.f. especially Figures 2.2 and 2.11 (5%, σ=1), and Figures 2.6 and 2.15 (20%, σ=2)), and tend to be located further to the north east of the plot. Also for unbalanced σ, h_2 may be larger relative to h_1 for the smaller sample sizes, see e.g. Figures 2.10 and 2.19 (50%, σ=3). The 10:10 examples also display less of a central plateau (e.g. Figures 2.5 and 2.14 (5%, σ=2), and Figures 2.10 and 2.19 again (50%, σ=3)). This is to be expected as it indicates that more smoothing of smaller data sets is appropriate, and that the degree of smoothing and ratio of smoothing parameters is more crucial than with the larger sample size. Not surprisingly, contours also tend to be less stable across simulations than for bigger samples (c.f. for example Figures 2.3 and 2.12 (20%, σ=1), Figures 2.5 and 2.14 again (5%, σ=2), and Figures 2.8 and 2.17 (5%, σ=3)) and also methods more variable across simulations (e.g. Figures 2.8 and 2.17 (5%, σ=3)). Allowing for these differences, the two sample sizes produce qualitatively similar contour plots and the conclusions from these are much the same.

In particular, for a given σ and fixed sample sizes, as error rate increases contours become more regular, they become steeper and/or there is less of a central plateau - compare for instance the very irregular contours of Figure 2.11 (5% error, σ=1) where there is considerable scope for variation of both the magnitude of smoothing parameters and of their ratio, with Figures 2.12 (20%, σ=1) and 2.13 (50%, σ=1) where this is much less evident. This is to be expected. Also, relative to each other methods 1-4 tend to be fairly stable across error rates (in fact for our simulations methods 1-4 are fixed for given σ_2 and sample sizes (see Appendix 2)) and similar, while 6-8 vary more, e.g. for n_i=25, Figures 2.11-2.13 (σ=1) and Figures 2.17-2.19 (σ=3). The increase in regularity of contours is more marked between 5% and 20% than between 20% and 50% (though note that "50%" is in fact always less than 35% (see Table 2.1)), and also for the smaller sample size (c.f. e.g. Figures 2.8 and 2.9 relative to Figures 2.9 and 2.10 (n_i=10, σ=3) with Figures 2.17 and 2.18 relative to Figures 2.18

and 2.19 (n_i=25, σ=3)).

For given sample sizes and fixed error rate, as σ increases not surprisingly the contour plots are qualitatively comparable but are centred higher up, indicating a larger h_2, e.g. Figures 2.13 and 2.19 (n_i=25, 50%, σ=1 and σ=3). The exact multiple h_2 of h_1 may not be crucial, at least for lower error rates, and varies.

The features discussed above are shared to some extent by the unbalanced sample size cases though these especially can produce very variable and irregular plots, e.g. Figure 2.24 (10:25, 20%, σ=2), Figure 2.30 (25:10, 20%, σ=1) and Figure 2.35 (25:10, 5%, σ=3), (a) especially. The effect of increasing σ in the 10:25 case is also not surprisingly somewhat less clear, as the imbalance in the sample sizes balances out that in the standard deviations, and can produce plots which on a linear scale would be fairly symmetric with approximately equal h_i being optimal, e.g. Figure 2.25 (10:25, 50%, σ=2).

The practical effect of differences in position of the various methods relative to the contours of MSE may be judged from plots of the predicted probability function $\hat{p}(\pi_1|x)$, some of which are given in Figures 2.37-2.44.

From both the contour plots and the latter, most notably methods 6-8 improve at least slightly on the marginal methods for the equal population case, (σ=1, 50%) where a high degree of smoothing is indicated, for all sample sizes, especially for 2 out of 3 simulations for $n_1:n_2$ = 10:25 (Figure 2.37) where 6-8 are much better than the marginal methods for (2) and (3), and 25:10 where 6-8 are uniformly better, markedly so for the first 2 simulations, especially (2). (Of 1-4 themselves, for equal-sized samples 4 is best, but all are very poor, while for 10:25 and 25:10 4 or 3 and 4 are best). This is no longer true as σ increases for equal sample sizes, where for zero separation and both σ=2 and σ=3, 1-4 could generally be at least slightly better for (1) and (3) though the best of 1-4 are near the optimal kernel (method 5) for (2) in each case. (For the smaller sample size 1 and 2 are best on (2) and (3), and 2 on (1) where method 1 is poor. For n_i=25, 4 is worst in each case, and 3 best at least on (2) and (3).) However, though similar to the best of 1-4 for (1), 6-8 are worse or compare to the worst for (2) (and are bimodal) and (3) (very poor for σ=2) for the smaller sample sizes n_i=10, while for n_i=25, 7 and 8 at least are comparable to or slightly better than 1-4 in each case. For the

Figures 2.37-2.44 (overleaf)

Examples of predicted probability functions $\hat{p}(\pi_1|x)$, corresponding to the contour plots in Figures 2.22, 2.36, 2.3, 2.5, 2.11, 2.18, 2.21 and 2.33 respectively. In each case kernel methods 1-4 and 6-8 are represented by broken lines, and the spline estimate by the dashed and dotted line, as shown in the key below. The true function is given by the heavy solid line and the MSE-optimal kernel (method 5) by the thinner solid line.

Key:-

METHOD	LINE
1	----------
2	----------
3	----------
4	----------
OPTIMAL KERNEL	----------
6	----------
7	----------
8	----------
SPLINE	----------
TRUE $P(\Pi_1/X)$	----------

Figure 2.37 $(n_1, n_2; \mu, \sigma) = (10, 25; 0.00, 1)$

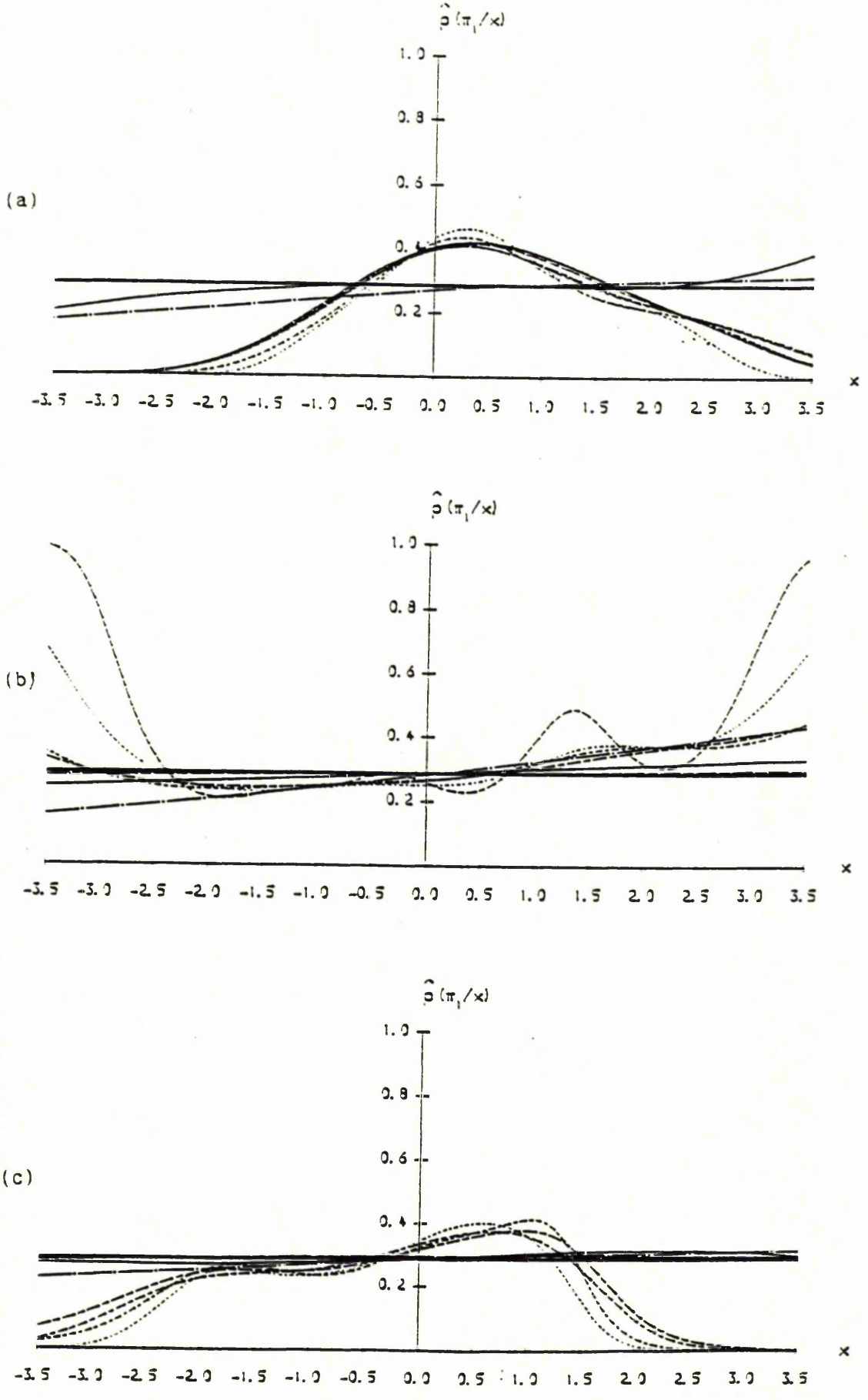


Figure 2.38 $(n_1, n_2; \mu, \sigma) = (25, 10; 0.00, 3)$

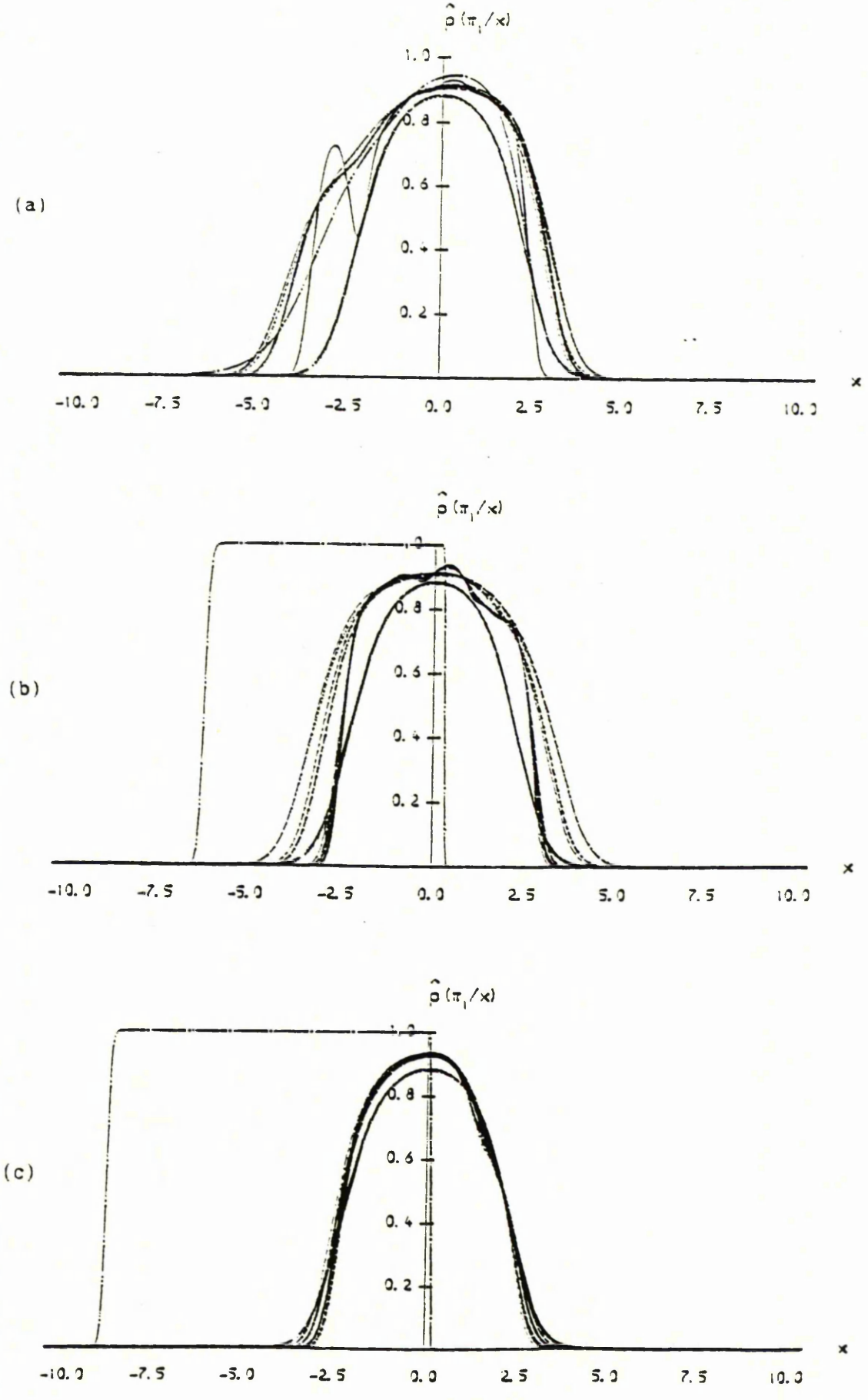


Figure 2.39 $(n_1, n_2; \mu, \sigma) = (10, 10; 1.68, 1)$

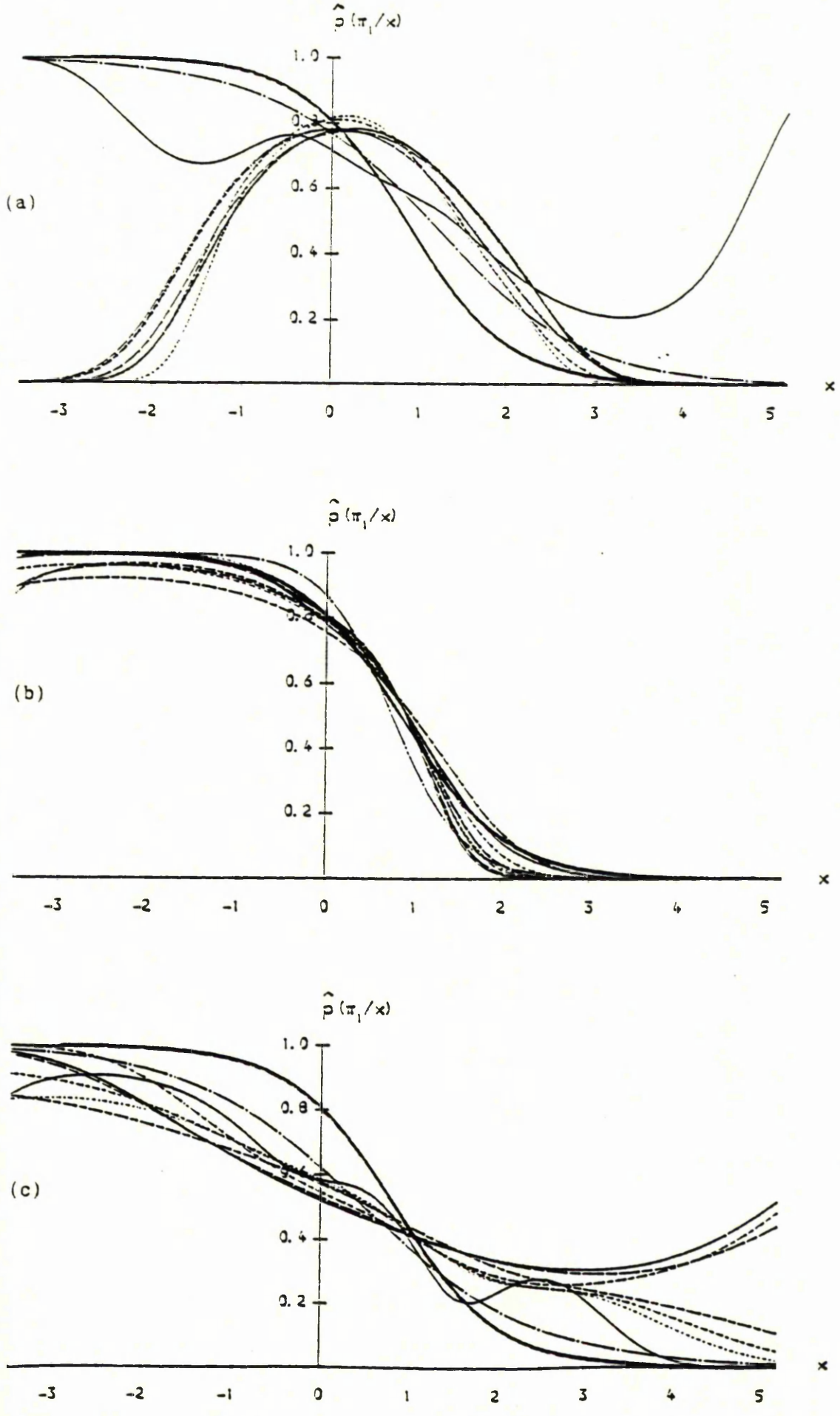


Figure 2.40 $(n_1, n_2; \mu, \sigma) = (10, 10; 4.83, 2)$

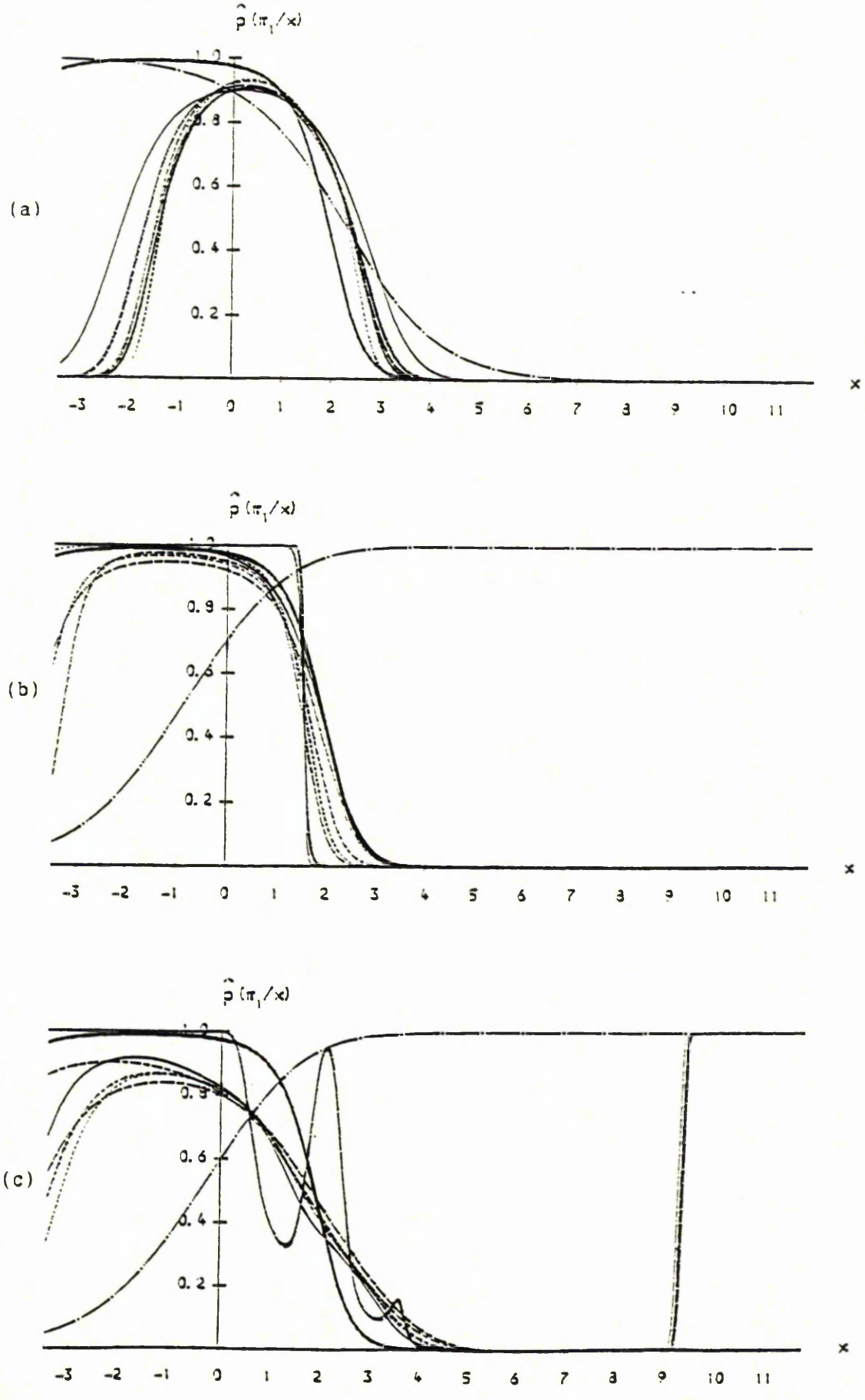


Figure 2.41 $(n_1, n_2; \mu, \sigma) = (25, 25; 3.29, 1)$

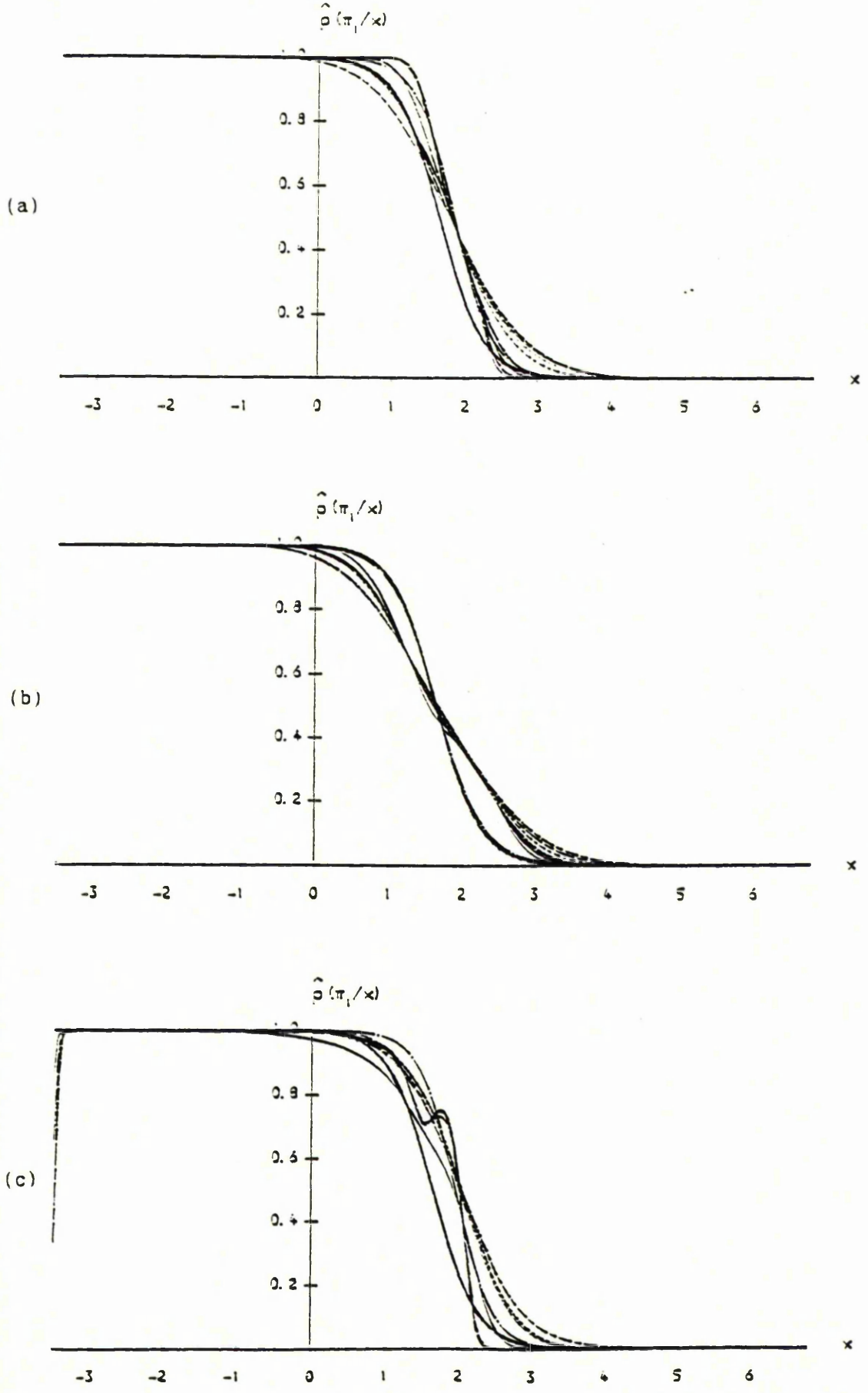


Figure 2.42 $(n_1, n_2; \mu, \sigma) = (25, 25; 2.40, 3)$

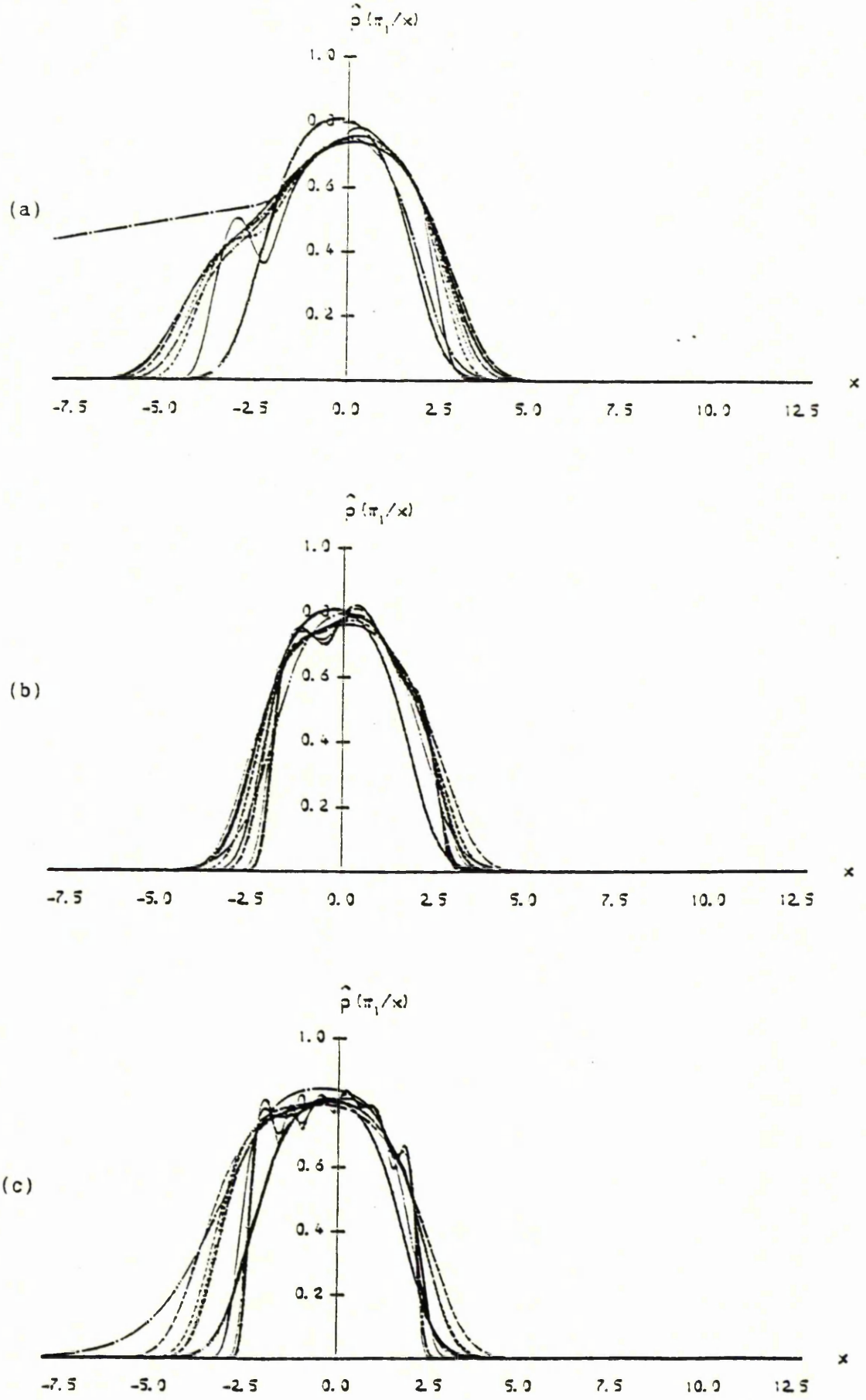


Figure 2.43 $(n_1, n_2: \mu, \sigma) = (10, 25: 1.40, 1)$

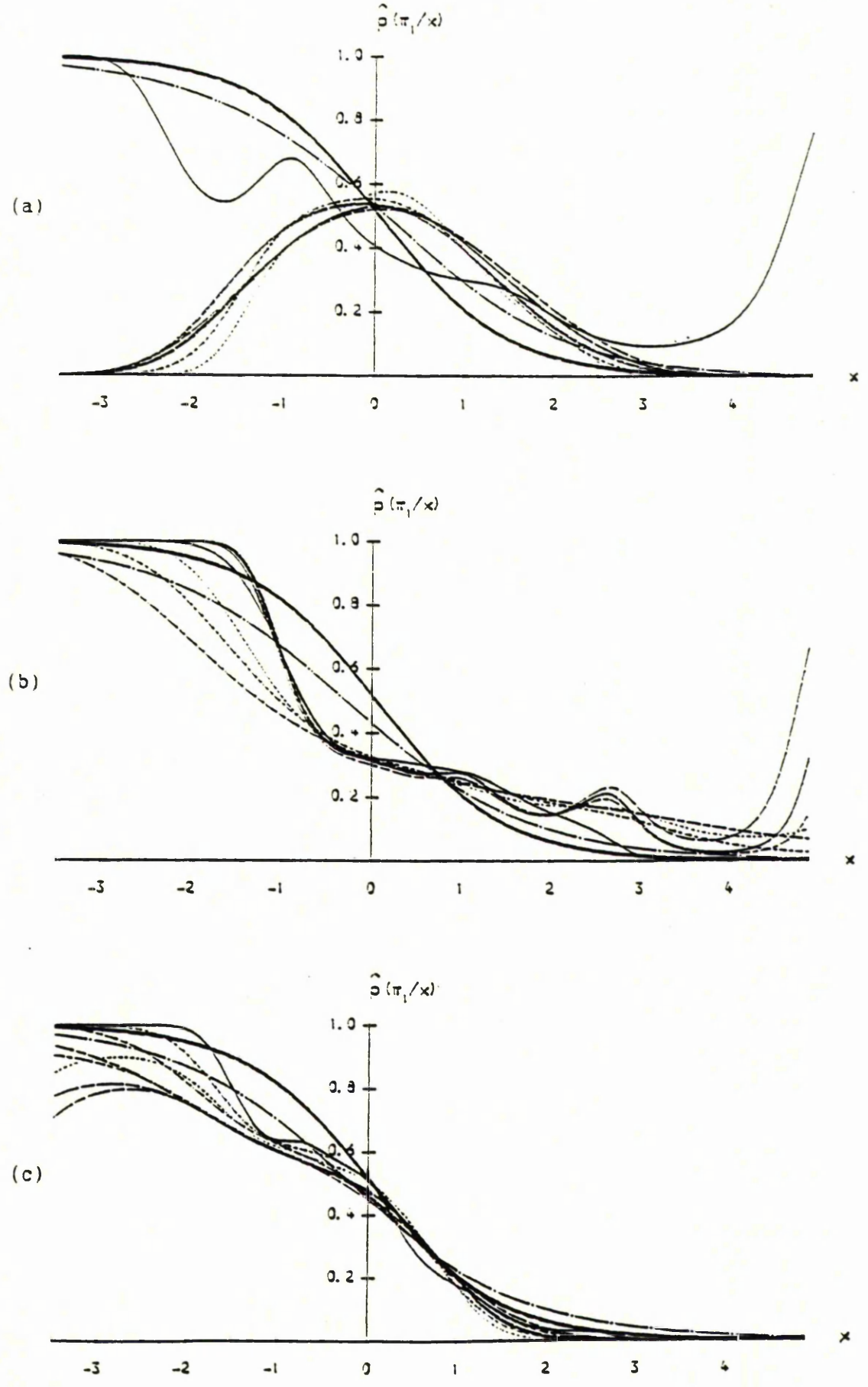
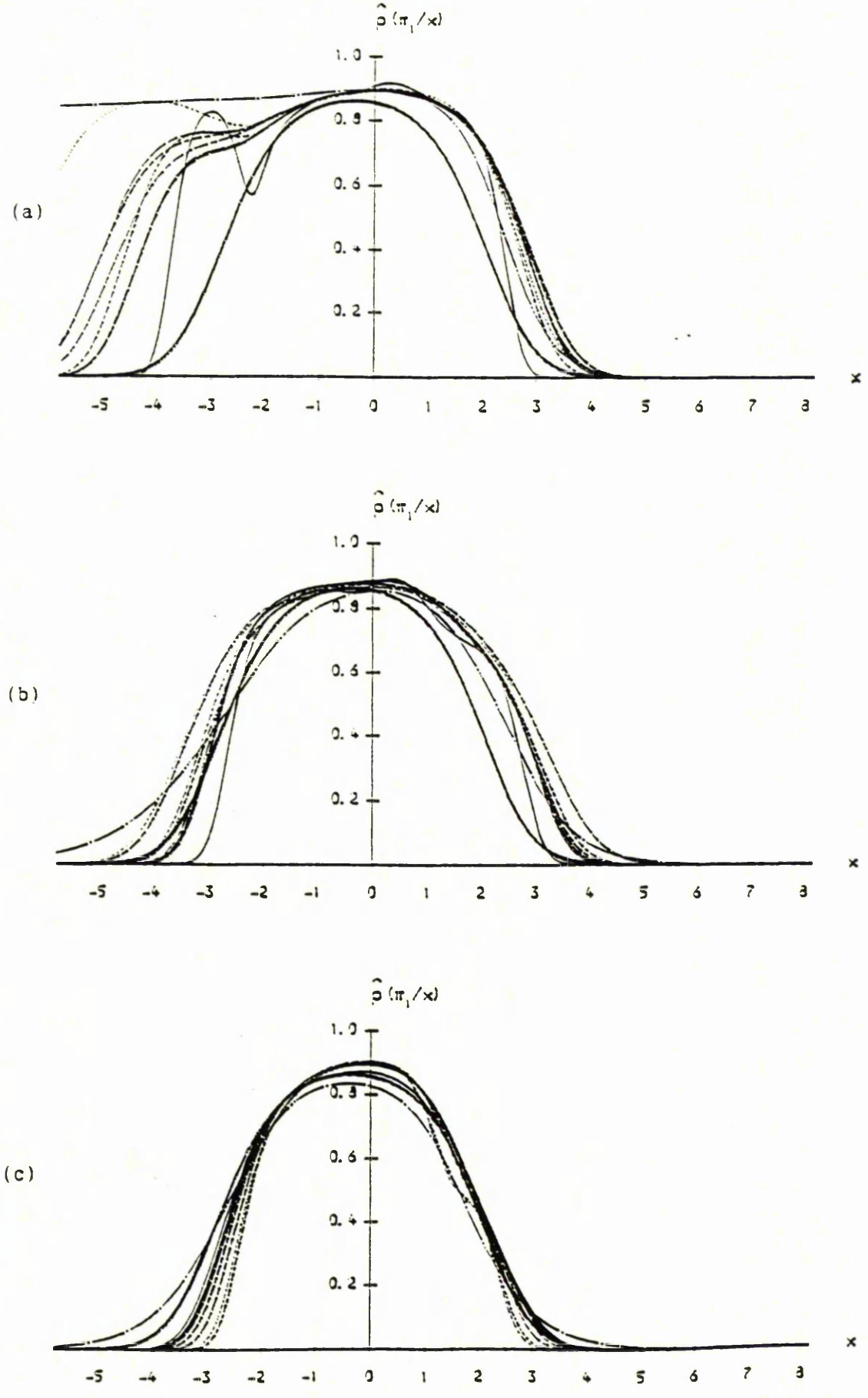


Figure 2.44 $(n_1, n_2; \mu, \sigma) = (25, 10; 1.14, 2)$



unbalanced cases, for 25:10 and σ=2, none is optimal for (1) or (2) where for (1) 6-8 compare to (the best of) 1-4, which are poor, and are among the best for (2) and not unlike the optimal kernel though a little undersmoothed. For (3) 6-8 are slightly better but all are within the 1st contour, if slightly suboptimal. (Of 1-4, 4 is worst on (1) and (2), 3 best on (2) and 3 and 4 best on (3).) For σ=3, all are similar for (1) and (3), again none quite optimal for (1) and are best for (2), though 6-8 again slightly undersmoothed (see Figure 2.38), being inside the 1st contour. (On (1) 2 is best and 4 worst, 3 best on (2) and all similar on (3).) For 10:25, for σ=2 (u_2 = 1.16), 6-8 are at least as good as 1-4 for (1), next best to 3, which is near-optimal, for (2) but could be better in each case and are near-optimal for (3) but comparable to 3 (and also 4). For σ=3, 6-8 compare to the best of 1-4 (3 and 4) for (1), are similar for (3) and for (2) 7 and 8 compare to 3, the best but not quite optimal, while 6 is poorer and compares to 1, 2 and 4. In each case 6-8 still leave room for improvement. (Of 1-4, for both σ=2 and 3, 3 and 4 are best for (1) and (3), and 3 clearly best for (2).)

For n_i=10, of the marginal methods 4 is slightly poorer in all cases on simulation (3), 1 poorer on (1) and 2 and 4 better on (2), so that the best one varies. For σ=1, 5% and more especially for 20% (see Figure 2.39) 1-4 are very poor on both (1) and (3), but not as bad for (2). For σ=2 at 5% on (1) and (3) 1-4, though not unlike optimal kernel method 5, are poor but for (2) the best of 1-4 is reasonable (see Figure 2.40). For 20% for (3) again all are relatively poor. Here unusually 2 and 4 are the best for (2), where all are acceptable, and (1), where there is room for improvement. For σ=3, for 5%, on (3) though all are poor, 1 and 2 especially are near to optimal method 5, as are 1-4 on (1), again poor. Again on (2) each estimate is acceptable. For 20%, in each case the best of 1-4 are close to method 5 though none quite optimal for (1) nor (3) but are good for (2), especially 2 and 4.

Methods 6-8 are typically very similar, though 6 may be better than 7 and 8, seen for 5% and simulations (1) and (2) (Figures 2.2, 2.5 and 2.8), and to some extent at 20% for σ=1 and 2 (Figures 2.6(a) and 2.3) and 7 slightly worse than 6 and 8 on (2) for 5% error (again Figures 2.2, 2.5 and 2.8). For σ=2, 20%, 8 is better than both 6 and 7 on (2), Figure 2.6(b). Generally for n_i=10, 6-8 offer no improvement on 1-4 where for σ=1 for both 5% and 20% (see

Figure 2.39), they are similar or worse. For 5%, 6-8 are far too sharp on (2) but, though noticeably undersmoothed, on (3) 6-8 are very similar to optimal kernel method 5. For σ=2, at 5% 6-8 compare to 1-4 for (1) but especially for (3) they are a lot worse as they are very badly undersmoothed, having h_2 too small, and also too sharp for (2) (see Figure 2.40). For 20%, 6-8 are comparable to 1-4 for (3), compare to the pooorer of 1-4 for (2) but to the best for (1). For σ=3, for 5%, 6-8 are again worse and far too sharp for (2) and also poorer for (3) while for (1) they compare to the better marginal ones though none are optimal. For 20%, again 6-8 compare in each case to method 5. For (1) 6-8 are comparable to 1-4, and for (3) compare to the best. For (2) 7 appears poor, 8 relatively so and 6 compares to the better marginal methods, which are good, but in terms of $\hat{p}(\pi_1|x)$ all are similar.

For n_i=25, 4 now appears consistently poorer than 1-3, and now 3 is generally best or near-best though method 2 (and 1) appears better in simulation (1) throughout. For σ=1 and 5% error, there is a little improvement potential at least on (3), Figure 2.11(c). For 20%, there is more room for improvement, though all are near to optimal method 5 on (3). On (2) 2 and 4 are slightly poorer than 1 and 3. For σ=2, at 5%, 1-4 are reasonably good, with 3 best on (2) but none are quite right. For 20%, for (1) all are very poor, also poor on (3), and on (2) though 1 especially is near the optimal kernel, again all are poor. For σ=3, at 5% on (2) 1-3 are near-optimal, but there is scope to improve on (1) and (3). At 20%, the pattern is similar, with 1-4, especially method 4, relatively poor on (1) and (3) but again good on (2).

Apart from the zero-separation case where they are very similar, 6-8 differ more for n_i=25 than for 10:10. Not surprisingly 7 and 8 are very similar and slightly distinct from 6 except for σ=2, 5% on (3), Figure 2.14(c), where 7 differs and is worse, as it is also for σ=3, 5% on (1), Figure 2.17(a), where there are larger differences. 7 and 8 are now superior at 5% for σ=1 and 2 on (2), Figures 2.11 and 2.14(b) and slightly better for σ=3 on (3), Figure 2.17(c) and also at 20% for σ=1 and 2, at least for (1), Figures 2.12 and 2.15. However 6 is obviously superior on (1) at 5% for both σ=2 and 3, Figures 2.14 and 2.17(a), and also for σ=3, 20%, Figures 2.18(b) and (c).

Compared to n_i=10, 6-8 are now more successful. For σ=1, for 5% (Figure 2.41), 6-8 are good especially for (1) and 7-8 very good

for (2), though no method is far from optimal, but despite appearing comparable (in terms of MSE) to 1-4 for (3) in contour plot Figure 2.11(c), 6-8 are quite distinct, undersmooth π_1 and $\hat{p}(\pi_1|x)$ is a little jagged. For 20%, 6-8 are at least as good on both (1) and (3), especially 7 and 8 which are near-optimal for (1), but compare to the poorer marginal methods (2 and 4) on (2). For σ=2, 6-8 again appear better for 5%, especially 6 for (1) and (3), although in fact on (1) while 6 is near-optimal 7 and 8 greatly undersmooth and produce a jagged curve, and while 6 and 8 slightly improve on 1-4 in (3), 7 is again undersmoothed. 7 and 8 compare to 3, the best of 1-4, for (2) but are still not quite right. For 20%, 6-8 appear slightly better for (1) though still poor, better for (3) where 7 and 8 especially are near-optimal, but worse and bimodal for (2). For σ=3, at 5%, 6 and 8 are slightly better than 1-4 for (1) though 7 undersmooths and $\hat{p}(\pi_1|x)$ is jagged, each of 6-8 is slightly better for (3) where 6 compares to 3, but worse again for (2), where 1-3 are near-optimal, as they are also for 20% where 6-8 undersmooth (2) and the only improvement is for simulation (3), but despite having the correct general shape for $\hat{p}(\pi_1|x)$ 6-8 are actually extremely jagged as h_1 is very small (see Figure 2.42).

For 10:25, 3 is generally the best of 1-4 on (2) where 4 is poorest, while 3 and 4 are superior on (1) and (3) with 1 poorest. For σ=1, at 5%, of 1-4 1 and 4 are poorest for (1), 3 best on (2) but none quite right for (3). Again for 20%, all appear good but similar for (3), though in terms of $\hat{p}(\pi_1|x)$ none are quite right, 3 is best on (2), and for (1) all are very poor. For σ=2, 5%, 3 is best on (2) and 3 and 4 best on (1) and (3) but not quite optimal. For 20%, all are similar but poor for (1), very poor for (2) and 1 the poorest for (3). For σ=3, at 5%, 1-4 are all poor for (1) and (3) while for (2) 3 is obviously better than 1, 2 and 4. For 20%, none are optimal.

Again 6 can be distinguished from 7 and 8. In each case 7 and 8 are at least slightly superior on (2), with 6 slightly better or similar on (1) and (3). Comparing 6-8 with 1-4, for σ=1, 5%, 6-8 are good only for (2) where they are near-optimal and compare to 3, the best, but are slightly worse for (3) and compare to the poorest for (1). For 20%, on (1) and (3) 6-8 compare to the poorest, but again are at least as good and compare to 3 for (2), though $\hat{p}(\pi_1|x)$ is strikingly undersmoothed (Figure 2.43). For σ=2, 5%, again they

are slightly worse for (1), but similar to 1-4 for (3), good and comparable to 3 for (2). For 20%, 6-8 are similar and do not improve on 1-4 for (1), 7 and 8 are slightly better for (2) but also all very poor, and although 6-8 are slightly worse for (3) they compare to method 1. For σ=3, at 5%, 6-8 are similar to method 1 on (1) while for (2) 6-8 compare to 3, the best. For (3) 6-8 are slightly worse than the better marginal ones. For 20%, 6-8 are among the best for (1) and also for (3), where they are good, but 7 and 8 at least are better for (2) and near-optimal, though still relatively poor.

For 25:10, 4 is consistently poorest on (2) and 3 best, with 3 and 4 best on (3). For (1), 2 appears best except for σ=1, 20% (Figure 2.30(a)) where again 3 and 4 are better. More specifically, for σ=1, 5%, 1-4 could be better especially for (1) and (3). 3 is better on (2). For 20% none are good on (1) and (3) though for the latter 2 is as the optimal kernel, while on (2) 3 is again superior to 1, 2 and 4. For σ=2, 5%, all have the correct shape on (1) but are poor on (3) and again 3 is best for (2). For 20%, all methods are within the 1st contour for (3) but none quite right for (1) and (2) where for the former 1-4 are seen to vary in terms of $\hat{p}(\pi_1|x)$ (Figure 2.44) with 1 especially poor. For σ=3, 5%, for (2) 1-4 seem relatively poor with 1 and 4 slightly poorer in terms of $\hat{p}(\pi_1|x)$, while for (1) all are similar with 1 slightly worse, though none optimal. None are quite right for (3). For 20%, Figure 2.38, for (1) all are similar but none optimal, nor are they for (2) while for (3) all are similar and inside the 1st contour.

The pattern is not quite as consistent now, but as usual 6 is often slightly different from 7-8, which are as good as or often superior to 6 on (2), at least at 5%, but all appear similar in terms of MSE on (1) and (3). For σ=1, for 5%, 6-8 are best and near to optimal method 5, though could still be better, for (1) and (3), as they are also for (2) where they compare to 3, the best of 1-4. For 20%, 6-8 are better for (3), though none good, while for (2) 7 and 8 compare to 3, the best, and are better than 6 which compares to the poorer of 1-4. 6-8 are no better than 1-4 for (1). Now for σ=2, 5%, 6-8 are slightly better for (1) and (3) and near to the optimal kernel on (3), but still not good though all have the correct shape on (1). For (2) 7 and 8 are also better and near-optimal with 6 similar to 3. For 20%, they are slightly

better for (3) while for (2) they compare to the best of 1-4 and for (1) 7 and 8 are better (6 compares to 3) but none is optimal. For σ=3, 5%, for (2) 6-8 are best, being on the edge of contour 1, but are similar in terms of $\hat{p}(\pi_1|x)$ to 1-4, and compare to the best of 1-4 for (1). For (3), 6-8 are superior (and optimal) though still not quite right. For 20%, Figure 2.38, 6-8 compare on (1) and (3) to 1-4, but for (2) though 6-8 are better and near-optimal they are multimodal.

In short, the scope to improve on marginal methods 1-4, which seem similar (on the contour plots, though are seen to vary more in Figures 2.37-2.44, with 3 often best or near-best and 1 sometimes very poor), appears limited except for equal populations, where 6-8 are considerably better and often near-optimal. In general the improvement potential is greater for balanced small samples (n_i=10) where 1-4 are often poor, though 6-8 in fact tend to be similar or worse. For n_i=25, 6-8 are usually comparable or slightly better (but generally for low error rate where 1-4 themselves are not far from optimal), as is also true for unbalanced samples, although for the latter the performance of 1-4 and potential to improve on them varies. In particular 6-8 can be very undersmoothed, even when in terms of MSE they seem to improve on 1-4.

Further examination of the plots comparing $p(\pi_1|x)$ and $\hat{p}(\pi_1|x)$ however, shows that the MSE-optimal kernel estimator (method 5) itself leaves something to be desired. In general method 5 is near the truth for identical populations, e.g. Figure 2.37, though not quite as good for equal samples when n_i=10. At 5% for σ=1 though 5 may leave a little room to improve especially for n_i=10 where it can be poor, (on (3), though (2) is excellent) mostly it is acceptable (e.g. Figure 2.41). For σ=2 it is not just as good. For n_i=25 its performance varies but again for n_i=10, for all error rates (e.g. Figure 2.40), (3) is poor and undersmoothed though (1) and (2) vary. For sample sizes 25:10 there could be at least a little improvement and more still for 10:25. It is poorer still at 20%, where for σ=1, though a little undersmoothed for 2 out of 3 simulations with n_i=25, (1) and (3) are extremely poor for n_i=10, especially (1) which is vastly undersmoothed, though (2) is very good (see Figure 2.39). (3) is slightly undersmoothed for unbalanced samples, for which (2) is poor, as is (1) also for 10:25 which is again very undersmoothed (Figure 2.43). For σ=2, 20%, for n_i=25 method 5 is poor on all simulations, grossly undersmoothed

on (1), and also poor at 50% where (1) and (2) are bimodal. For both error rates 5 is poor for n_i=10, at least on (3). Again at 20%, while (3) is reasonable, (1) and especially (2), are poor for 10:25 and (1) poor for 25:10 (Figure 2.44). The same is true for 25:10 at 50% while for 10:25 2 out of 3 estimates could be better. Finally, for σ=3 generally method 5 leaves at least some room for improvement, and can be very poor, e.g. at 5% for 10:25 on (1) and at both 5% and more especially 20% for 25:10 for (1) again, Figure 2.38, where it is very undersmoothed. (1) is also poor for n_i=25 at each error rate (e.g. Figure 2.42) though for n_i=10 is not as bad. Given that an estimate may be undersmoothed yet its MSE be relatively small, MSE may not be the best criterion by which to judge reliability.

2.5.3 Comparison of kernel and spline methods

In the light of the frequent inability of these kernel methods to produce a near-optimal predicted probability curve, in each case we compare the MSE-optimal kernel estimates (method 5) with the estimates provided by the spline ratio estimator of Silverman (1978a) (Section 1.5.2), which, as it estimates directly the density ratio or a one-to-one function of the ratio and is subject to a roughness penalty, might be expected to be a more flexible method. The smoothing parameter of the spline, β, was chosen to approximately minimise MSE using a grid search method. For a given β, a Simplex method (NAG (1984) algorithm E04CCF) was used to solve numerically the 2 non-linear equations whose solution determines the spline.

For n_i=10, the spline was generally unsuccessful. It avoids the conspicuous undersmoothing of the kernel method which can produce a multimodal curve (notably σ=1, 20%, simulations (1) and (3), see Figure 2.39, to a lesser extent σ=2, 50%, (3), and also 5% (3), seen in Figure 2.40), but tends instead to oversmooth, producing too gradual an "S-shape", too wide and flat (and often skew) a "density shape", or even more extreme, a sigmoid rather than a concave curve. It did not improve on the variability across simulations of the kernel method, and sometimes reversed the direction of the "S". It's generally poor performance may be due to a suboptimal β and the possibly ill-conditioned nature of the spline-fitting problem (see Section 2.6, where the smoothing parameter is selected to give a curve of comparable smoothness to

those of the other methods). The spline only rarely improved on method 5. For σ=1, 5%, the spline was superior for (1) and (3) but inappropriately smoothed on (2) where method 5 was excellent. Similarly, for σ=1, 20%, Figure 2.39, it was again considerably better both for (1) and (3) but much poorer for (2). The only other success was for σ=2, 5%, simulation (1), where its rather gradual but correct S-shape improved to some extent on method 5's concave estimate, though (2) and (3) were both very poor (see Figure 2.40).

For equal populations, σ=1, 50%, where method 5 does very well, and for all sample sizes, the spline is generally poorer (Figure 2.37 shows the 10:25 case).

For n_i=25, the spline's performance is less variable - it does not, for instance, reverse the direction of the curve. For σ=1, 5%, Figure 2.41, the spline improves for (2), where it is very good, but is at least slightly poorer for (1), where method 5 was not far from optimal, and also for (3), while for σ=1, 20%, it is comparable or better, especially for (1), and has the correct shape whereas method 5 undersmoothed both (1) and especially (2). For σ=2, 50%, where method 5 generally was slightly undersmoothed, the spline oversmoothed, and was worse - again wide, and also skew for (1) and (2). For 5%, the spline improves on the shape of method 5, especially for (2) which was undersmoothed, and is slightly better for (1) but worse for (3). For 20%, while especially for (1) method 5 was greatly undersmoothed and poor, despite having a concave rather than an S-shape, the spline, which is sigmoid, is equally poor. For σ=3, 5%, the spline is consistently sigmoid rather than flatly concave and worse for 2 out of 3 simulations, even for (1) where method 5 was also poor. For 20% too, Figure 2.42, it oversmooths and is wide, skew and poor for (3) and particularly skew for (1), though slightly better than method 5 for (2). For 50%, it was smoother and slightly better or comparable to method 5.

For the unbalanced 10:25 situation, again the spline was the most successful for σ=1, 5% and 20%, especially for (1) and (2) of 5% and (1) of 20% where method 5 is again extremely poor. Here, despite slightly oversmoothing, the spline is uniformly better (see Figure 2.43). For σ=2, the spline is generally better for (1), where for both 5% and 20% method 5 is poor, but worse for (2) and (3). For 5%, the S is a little too gradual, and it is also a

gradual S rather than concave for 20% where it was especially poor for (3) unlike the kernel method. For 50%, it is again very oversmoothed, and also skew for (2) and (3). This is also true of σ=3 for 20% and especially 50% where it is uniformly very poor. For 5%, the spline is not concave, again a gradual S, and is slightly better only for (1).

For $n_1:n_2$ = 25:10, most notably there are rare examples of undersmoothing of the spline on simulations (2) and (3) of σ=3, 20%, where it is extremely poor (Figure 2.38). On (1) however, where method 5 was clearly bimodal, it is better although rather skew. For σ=3, 5%, the spline is not concave but improves on method 5 for (1) which was undersmoothed and is comparable for (2) but worse for (3). For σ=2, 5%, where method 5 was reasonable, if a little undersmoothed, at least for (1) and (2), the spline is now poor, especially for (2) and (3), with too gradual a curve or a completely wrong direction. For 20%, the spline was only very slightly worse for (3), which was good, better for (2) despite being oversmoothed (a little wide) and comparable for (1) in terms of MSE despite its lack of concavity, but here method 5 was obviously undersmoothed (Figure 2.44). For 50%, method 5 was extremely poor for (1) and the spline is even worse, being very skew, slightly better for (2), but slightly worse for (3). For σ=1, again for 20%, where method 5 is very undersmoothed for (2), the spline is good, while (1) and (3) are also comparable or better and now the correct S-shape, if a little too gradual a curve. For 5%, again the spline is slightly better for (1) and smoother, better and very good for (2) but worse for (3).

In general, therefore, in contrast to the kernel method which tends to undersmooth, the spline does the opposite. It occasionally improves on the optimal kernel but is more usually disappointing and can be very poor, especially for smaller sample sizes. Given the spline estimator's asymptotic foundation (Silverman, 1978a) this is possibly not surprising. We compare its performance to the empirical kernel methods for a much larger data set in Section 2.6.

2.5.4 T-statistics, discussion and conclusions

To enable more objective assessment of the differences between the kernel methods, 200 simulations of each configuration were carried out and paired t-statistics derived for all possible

comparisons between methods in terms of error rate, Brier score, log and modified log scores, and MSE.

Figures 2.45-2.48 provide summary plots of the significant differences between methods for all 5 scores, ranking the methods from best to worst by their mean score and connecting with a straight line those methods with nonsignificant t-statistics, in the usual way. To correct to some extent for making multiple comparisons a critical value of t(199|.995) = 2.576 was used for a 2-sided test

Tables 2.2-2.5 present the numerical results for the Brier, modified log and MSE scores, where the signs of the t-statistics have been adjusted so that a positive t-statistic for methods A and B (denoted A, B) indicates superiority of method A. Of the initial 5 scores, MSE was retained to allow comparison with the contour plots, error rate dropped as it is the least sensitive measure, and modified log score taken as representative of the two log scores, being more sensitive in general than the ordinary log score and having smaller standard deviation (the latter judged from tables of means and standard deviations for each score and method, which are not presented here).

Discussion of t-test Tables 2.2-2.5 and summary Figures 2.45-2.48

Although no one method seemed consistently superior in Section 2.5.2, and method 1 particularly could be poor (as, for instance, in Figures 2.40 and 2.44), Figures 2.45-2.48 in general give the impression that on average marginal methods 1 and 3 are overall best, 4 and 2 poorer, and direct methods 6-8 poorest for easier discrimination problems (5% error). For σ=1 usually 1 is significantly better than the rest, or at least better than all except 3 (and sometimes 4), and at least 7 and 8 poorer than 1-4 (not for n_i=25), and for n_i=10 3 is also better than 6-8. As σ increases the pattern is similar but typically there are fewer differences, although 3 still betters 7 if not 8 also, except for σ=3 and equal n_i where at least 6 and 8 move up past 2 and for n_i=10 1, 3 and 4 are not distinguishable from 8.

As classification becomes more difficult and error increases 6-8 appear to improve relative to the marginal methods and are best or near-best for identical populations and those differing only in variance, though rarely significantly better than the very simple marginal method 1. Except for σ=3 and 25:10 where there is little

Figures 2.45-2.48 and Tables 2.2-2.5 :-

Methods (as described above):

Marginal methods:

1) Normal Optimal (NOPT)
2) Asymptotically Optimal MISE (ASOPT)
3) Cross-validation with Kullback-Leibler (XVKL)
4) " " " Integrated Square Error (XVISE)

Direct methods:

6) Cross-validation with the Brier score (XV BRIER)
7) " " " the log score (XV LOG)
8) " " " the modified log score (XV ϵ-LOG).

Figure 2.45(a) $(n_1, n_2) = (10, 10)$: Summary of t-tests

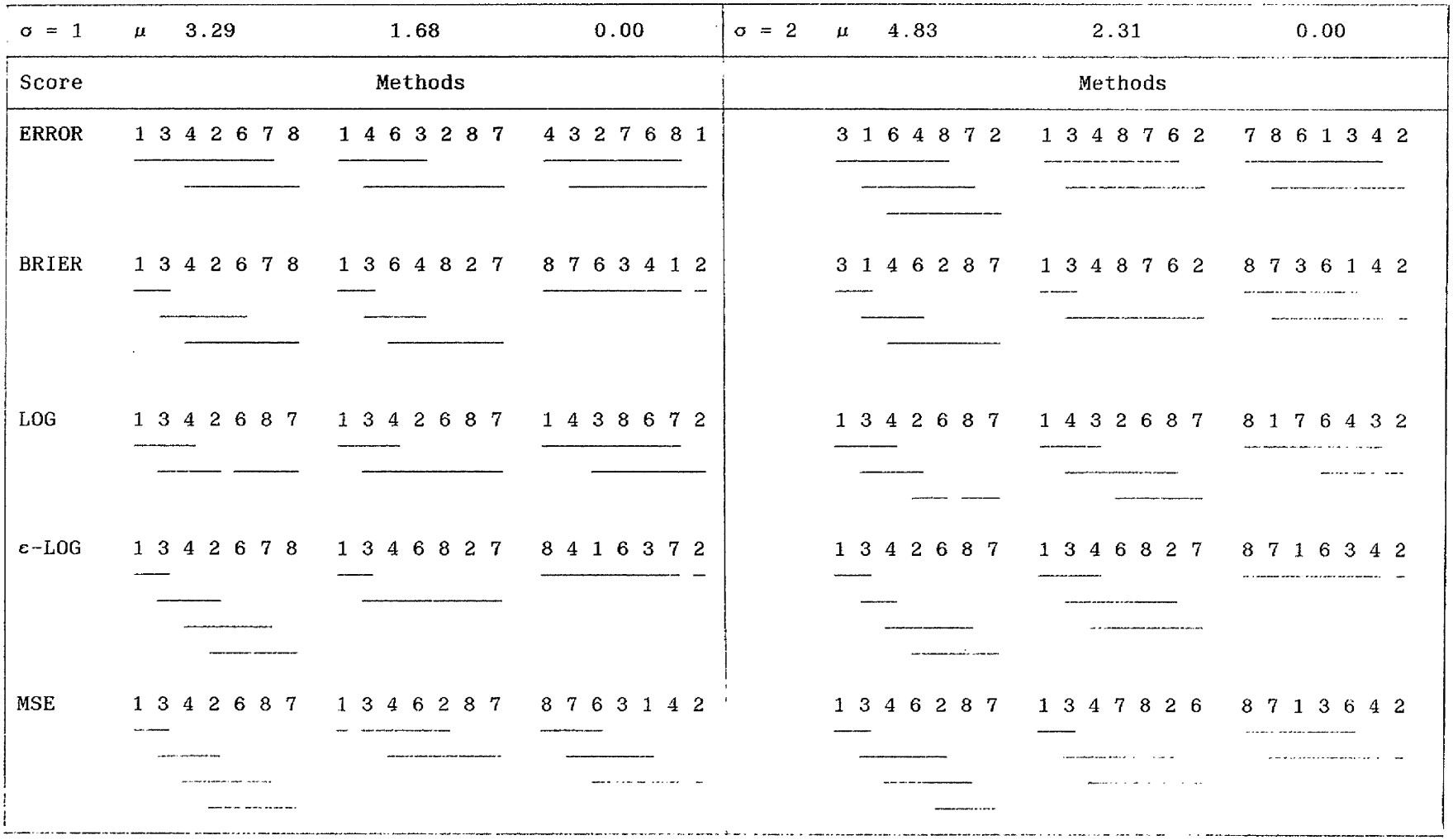


Figure 2.45(b) $(n_1 , n_2) = (10, 10)$: Summary of t-tests

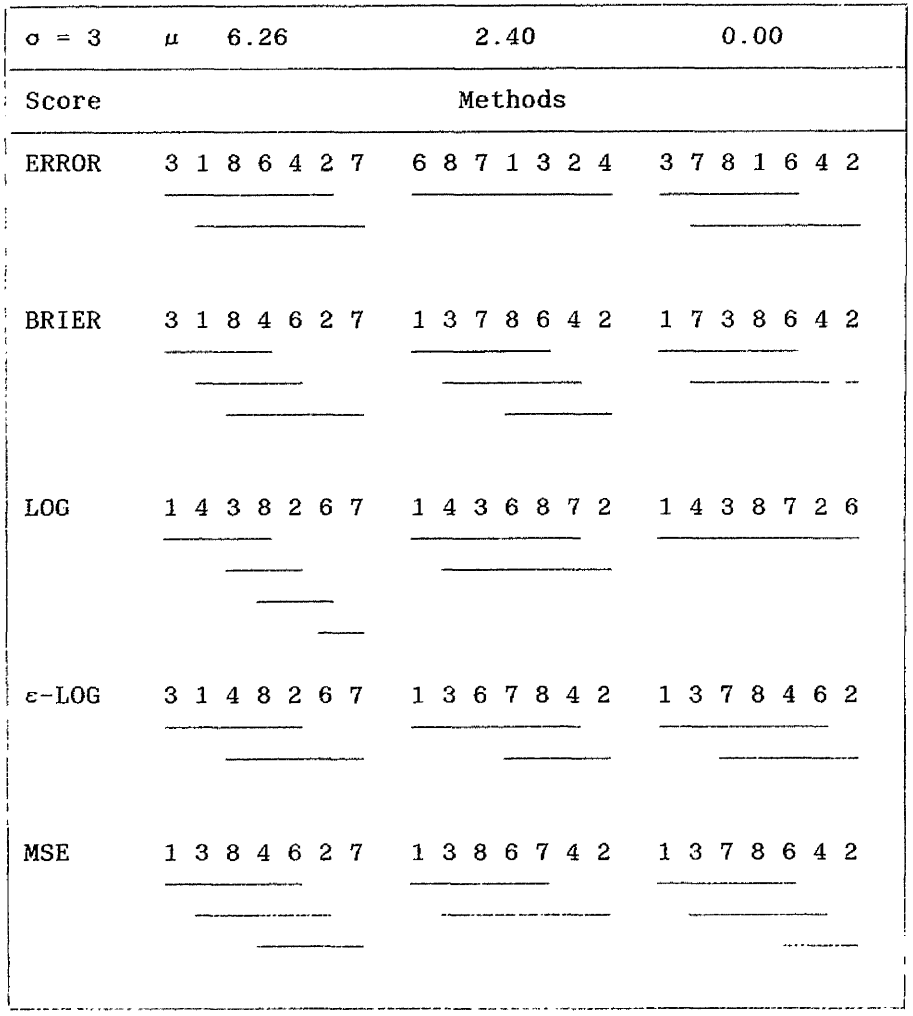


Figure 2.46(a) $(n_1, n_2) = (25, 25)$: Summary of t-tests

$\sigma = 1$	μ 3.29	1.68	0.00	$\sigma = 2$	μ 4.83	2.31	0.00
Score	Methods				Methods		
ERROR	3 1 4 6 8 7 2	1 4 6 3 7 8 2	2 4 1 7 8 3 6		1 3 6 4 8 7 2	1 3 4 6 8 2 7	7 1 8 3 6 4 2
BRIER	1 3 4 2 6 8 7	1 3 6 8 4 7 2	7 8 1 6 4 3 2		1 3 4 2 6 8 7	1 3 4 6 8 7 2	1 7 8 3 6 4 2
LOG	1 4 3 2 6 8 7	1 4 3 2 6 8 7	1 8 4 7 3 6 2		1 4 3 2 6 8 7	1 4 3 6 2 8 7	1 8 7 4 3 6 2
ϵ-LOG	1 3 4 2 6 8 7	1 3 4 6 8 2 7	1 8 7 4 6 3 2		1 3 4 2 6 8 7	1 3 4 6 8 7 2	1 8 7 3 6 4 2
MSE	1 3 4 2 8 6 7	1 3 6 4 8 7 2	7 8 1 6 3 4 2		1 3 4 6 2 8 7	1 3 4 6 8 7 2	1 8 7 3 6 4 2

Figure 2.46(b) $(n_1, n_2) = (25, 25)$: Summary of t-tests

$\sigma = 3$	μ 6.26	2.40	0.00
Score	Methods		
ERROR	1 3 8 6 7 4 2	1 3 4 7 8 6 2	1 3 7 8 4 6 2
BRIER	1 3 4 6 8 7 2	1 8 3 4 7 6 2	1 7 8 3 4 6 2
LOG	1 4 3 2 6 8 7	1 4 8 3 6 2 7	1 4 3 8 7 2 6
ϵ-LOG	1 4 3 6 2 8 7	1 8 3 4 7 6 2	1 8 3 7 4 6 2
MSE	1 3 4 6 8 2 7	1 3 8 4 7 6 2	1 3 7 8 6 4 2

Figure 2.47(a) $(n_1, n_2) = (10, 25)$: Summary of t-tests

$\sigma = 1$	3.16	1.40	0.00	$\sigma = 2$ μ 4.92	2.47	1.16
Score	Methods			Methods		
ERROR	1 3 4 2 6 8 7	3 1 6 8 4 7 2	7 6 8 1 3 4 2	3 1 4 6 2 7 8	1 3 8 4 7 6 2	7 8 6 1 3 4 2
BRIER	1 3 4 2 6 8 7	1 3 4 8 6 7 2	1 7 6 8 4 3 2	1 3 4 2 6 7 8	1 4 3 7 6 8 2	1 7 3 8 4 6 2
LOG	1 4 3 2 6 8 7	1 4 3 8 2 7 6	1 4 7 3 8 6 2	1 4 3 2 6 7 8	1 4 3 2 6 7 8	1 4 3 7 2 6 8
ε-LOG	1 3 4 2 6 8 7	1 4 3 8 2 6 7	1 7 4 6 8 3 2	1 3 4 2 6 7 8	1 4 3 6 7 8 2	1 4 3 7 8 6 2
MSE	1 3 4 2 6 8 7	1 3 4 8 7 6 2	7 6 8 1 3 4 2	1 3 4 2 6 8 7	1 3 4 7 8 6 2	1 7 3 8 4 6 2

Figure 2.47(b) $(n_1, n_2) = (10, 25)$: Summary of t-tests

$\sigma = 3$	μ 6.56	2.92	0.00
Score	Methods		
ERROR	3 1 6 4 7 8 2	6 7 8 1 3 4 2	7 4 8 6 3 1 2
BRIER	1 3 4 2 6 8 7	1 3 6 7 4 8 2	1 3 7 8 6 4 2
LOG	1 4 3 2 6 8 7	1 4 3 6 2 8 7	1 4 3 8 6 7 2
ε-LOG	1 4 3 2 6 8 7	1 3 4 6 7 8 2	1 3 4 8 7 6 2
MSE	1 3 4 2 6 8 7	1 3 4 7 8 6 2	1 3 8 7 6 4 2

Figure 2.48(a) $(n_1, n_2) = (25, 10)$: Summary of t-tests

$\sigma = 1$	μ 3.16	1.40	0.00	$\sigma = 2$	μ 4.35	1.14	0.00
Score	Methods				Methods		
ERROR	1 3 4 6 7 8 2	1 3 2 4 6 7 8	8 6 7 1 3 4 2		1 3 2 8 4 7 6	1 3 2 4 7 6 8	3 1 8 7 4 2 6
BRIER	1 3 4 2 7 6 8	1 3 4 6 2 7 8	8 7 6 1 4 3 2		1 3 4 2 7 8 6	1 3 4 7 6 8 2	1 3 8 7 4 6 2
LOG	1 3 4 2 6 7 8	1 3 4 2 6 8 7	1 8 7 6 4 3 2		1 3 4 2 6 7 8	1 3 7 6 4 8 2	1 3 8 7 4 6 2
ϵ-LOG	1 3 2 4 6 7 8	1 3 4 2 6 8 7	8 7 1 6 4 3 2		1 4 3 2 6 7 8	1 3 4 6 8 7 2	1 3 8 7 4 6 2
MSE	1 3 4 2 6 8 7	1 3 4 6 2 8 7	8 6 7 1 4 3 2		1 3 4 2 8 7 6	1 3 4 7 6 8 2	1 3 4 8 7 6 2

Figure 2.48(b) $(n_1, n_2) = (25, 10)$: Summary of t-tests

$\sigma = 3$	μ 5.42	0.00
Score		Methods
ERROR	1 2 4 3 7 8 6	1 3 7 8 2 6 4
BRIER	1 4 3 2 6 7 8	1 3 4 2 7 8 6
LOG	1 4 3 2 6 7 8	1 3 4 2 7 8 6
ϵ-LOG	1 4 3 2 6 7 8	1 3 4 2 7 8 6
MSE	1 3 4 2 6 8 7	1 3 4 2 6 7 8

Table 2.2(a) Paired t-statistics of differences between methods

$(n_1, n_2) = (10, 10)$									
Methods	Scores								
	Brier	ε-Log	MSE	Brier	ε-Log	MSE	Brier	ε-Log	MSE
μ / σ=1		3.29			1.68			0.00	
1,2	3.11	3.22	3.85	4.84	4.34	5.27	3.43	3.86	5.17
1,3	0.58	0.72	0.63	1.90	1.59	2.76	-1.10	0.09	-0.02
1,4	3.00	3.18	3.83	4.42	3.35	4.68	-0.71	-0.04	0.42
1,6	2.68	3.59	3.20	3.05	2.81	3.45	-1.55	0.05	-1.92
1,7	3.17	4.47	4.79	3.51	3.15	5.12	-1.60	0.13	-2.20
1,8	4.01	4.88	4.69	3.82	3.37	5.22	-2.00	-0.60	-2.72
2,3	-2.10	-2.17	-2.39	-2.63	-2.64	-2.31	-3.89	-3.36	-4.10
2,4	-0.38	-0.90	-1.67	-1.74	-2.74	-2.82	-5.94	-6.54	-9.26
2,6	0.51	1.27	0.23	-1.08	-0.67	-0.29	-3.94	-2.84	-5.68
2,7	0.61	1.56	2.32	0.06	0.24	0.91	-4.02	-2.95	-5.80
2,8	1.35	2.10	1.99	-0.25	-0.09	0.54	-4.31	-3.48	-6.17
3,4	2.14	2.02	2.12	2.02	1.47	1.26	0.54	-0.13	0.33
3,6	2.09	2.79	2.63	1.51	1.69	1.92	-0.91	-0.05	-2.47
3,7	2.68	3.64	4.96	2.64	2.38	4.11	-1.07	0.09	-4.00
3,8	3.74	4.50	5.08	2.62	2.57	3.98	-1.63	-1.06	-4.57
4,6	0.66	1.63	0.75	-0.19	0.56	0.64	-1.12	0.07	-2.04
4,7	0.81	2.06	2.84	0.80	1.23	1.91	-1.17	0.16	-2.35
4,8	1.58	2.60	2.55	0.61	1.07	1.61	-1.58	-0.57	-2.81
6,7	0.04	0.15	2.53	1.17	0.94	1.56	-0.06	0.15	-0.73
6,8	0.93	0.77	2.02	1.13	0.76	1.23	-1.31	-1.88	-1.83
7,8	1.37	1.02	-1.18	-0.46	-0.48	-0.78	-1.22	-1.68	-1.64
μ / σ=2		4.83			2.31			0.00	
1,2	3.26	3.20	3.79	3.59	3.44	4.39	4.26	4.08	5.35
1,3	-0.32	0.05	0.51	1.03	0.83	2.22	-0.54	0.45	0.20
1,4	2.91	2.81	3.40	2.82	2.04	3.18	2.22	1.64	2.95
1,6	2.42	3.78	2.81	2.89	2.82	4.12	-0.42	0.20	0.42
1,7	3.90	4.89	5.02	2.92	2.90	3.56	-1.03	-0.65	-0.19
1,8	3.43	4.50	3.76	2.86	2.87	3.66	-1.33	-1.05	-0.90
2,3	-2.90	-2.64	-2.57	-2.01	-2.34	-1.58	-4.04	-3.22	-3.82
2,4	-1.38	-1.77	-2.08	-1.95	-3.21	-3.31	-4.01	-4.85	-5.39
2,6	-0.69	0.53	-0.34	-0.03	-0.24	0.30	-3.28	-2.95	-3.72
2,7	1.29	2.33	2.49	-0.11	0.03	-0.14	-3.85	-3.75	-4.10
2,8	0.50	1.52	0.93	-0.21	-0.14	-0.06	-4.16	-4.09	-4.66
3,4	2.69	2.28	1.99	1.23	0.90	0.41	2.22	0.81	1.81
3,6	3.39	4.53	3.38	3.08	3.19	3.19	0.09	-0.22	0.16
3,7	4.56	5.29	5.40	2.26	2.71	1.91	-0.54	-1.01	-0.45
3,8	3.91	4.73	4.11	2.20	2.60	2.12	-0.86	-1.37	-1.20
4,6	0.01	1.51	0.40	0.76	1.15	1.42	-1.74	-0.97	-1.78
4,7	1.97	3.18	3.13	0.68	1.35	0.94	-2.37	-1.91	-2.27
4,8	1.15	2.36	1.61	0.59	1.24	1.02	-2.71	-2.35	-2.91
6,7	3.14	3.01	3.94	-0.15	0.46	-0.76	-1.06	-1.16	-0.91
6,8	1.83	1.66	2.05	-0.34	0.18	-0.66	-1.53	-1.64	-1.94
7,8	-1.49	-1.62	-3.10	-0.53	-0.81	0.40	-1.75	-2.71	-4.10

Table 2.2(b) Paired t-statistics

$(n_1, n_2) = (10, 10)$									
Methods	Scores								
	Brier	ε-Log	MSE	Brier	ε-Log	MSE	Brier	ε-Log	MSE
μ / σ=3	6.26			2.40			0.00		
1,2	2.93	2.88	3.48	4.16	4.10	4.54	4.38	4.15	5.34
1,3	-0.56	-0.29	0.33	0.34	0.58	1.62	0.21	0.98	0.94
1,4	2.31	1.87	2.50	2.67	2.14	3.39	3.29	2.58	4.38
1,6	2.09	3.00	2.20	0.87	0.94	2.24	1.48	1.96	2.33
1,7	2.73	3.75	3.67	0.56	1.09	2.21	0.15	1.04	1.03
1,8	1.19	1.29	0.66	0.83	1.29	2.03	0.27	1.08	1.05
2,3	-2.84	-2.57	-2.44	-3.20	-3.26	-2.10	-3.58	-2.78	-3.38
2,4	-1.67	-2.33	-2.82	-2.09	-3.59	-3.10	-2.85	-3.82	-3.48
2,6	-0.22	0.56	-0.55	-2.45	-2.67	-1.80	-1.32	-0.61	-1.72
2,7	0.81	1.81	1.69	-2.68	-2.50	-1.71	-3.02	-2.32	-3.24
2,8	-1.05	-0.88	-2.08	-2.54	-2.52	-1.94	-2.94	-2.30	-3.25
3,4	2.35	1.73	1.41	2.24	1.43	0.95	2.32	0.94	2.11
3,6	3.08	3.37	2.27	0.62	0.42	0.66	1.39	1.29	1.45
3,7	3.49	4.03	3.83	0.30	0.69	0.73	-0.05	0.02	0.00
3,8	1.65	1.44	0.41	0.63	0.89	0.45	0.09	0.05	0.03
4,6	0.64	1.81	0.50	-1.38	-0.80	-0.54	-0.15	0.74	-0.50
4,7	1.60	2.90	2.51	-1.62	-0.58	-0.50	-1.87	-0.74	-2.08
4,8	-0.31	0.15	-1.08	-1.42	-0.43	-0.72	-1.76	-0.71	-2.08
6,7	1.45	1.86	2.55	-0.46	0.22	0.06	-1.81	-1.57	-1.94
6,8	-1.00	-1.58	-2.05	-0.04	0.42	-0.44	-1.70	-1.54	-1.93
7,8	-1.91	-2.64	-3.74	1.10	0.51	-0.76	1.71	0.65	0.35

Table 2.3(a) Paired t-statistics

$(n_1, n_2) = (25, 25)$									
Methods	Scores								
	Brier	ε-Log	MSE	Brier	ε-Log	MSE	Brier	ε-Log	MSE
μ / σ=1		3.29			1.68			0.00	
1,2	4.06	3.79	4.53	5.09	4.72	6.23	7.41	7.00	8.47
1,3	2.23	2.98	2.10	3.54	3.46	4.55	4.00	4.12	3.95
1,4	4.06	4.30	4.51	5.70	4.60	6.94	3.90	3.84	4.50
1,6	5.35	5.68	7.35	4.10	4.11	5.60	1.50	2.86	1.60
1,7	5.34	6.06	6.63	5.03	5.18	6.10	-0.66	0.92	-0.67
1,8	5.08	5.69	6.06	4.68	4.81	5.72	-0.59	0.91	-0.66
2,3	-2.52	-1.81	-2.90	-2.55	-2.04	-3.22	-3.65	-3.38	-5.36
2,4	-2.34	-2.01	-1.84	-2.08	-2.85	-2.53	-9.67	-8.35	-10.15
2,6	0.84	2.23	1.33	-1.83	-0.91	-1.69	-4.77	-3.31	-5.78
2,7	1.62	3.26	1.93	-0.90	0.43	-0.11	-7.02	-6.32	-8.85
2,8	0.88	2.62	1.17	-1.34	-0.07	-0.70	-7.06	-6.39	-8.86
3,4	1.77	0.92	2.91	1.99	0.84	2.73	-0.63	-1.08	0.51
3,6	3.56	4.18	4.76	0.88	1.16	1.48	-2.52	-0.56	-1.86
3,7	4.13	4.95	4.93	2.94	3.57	3.70	-5.06	-3.68	-5.43
3,8	3.44	4.27	4.19	1.83	2.45	2.54	-5.02	-3.71	-5.47
4,6	2.15	3.22	2.35	-0.89	0.46	-0.52	-1.28	0.41	-1.78
4,7	2.75	4.10	2.86	0.15	1.92	1.18	-3.56	-2.28	-4.30
4,8	2.08	3.48	2.08	-0.31	1.33	0.49	-3.52	-2.31	-4.29
6,7	1.44	2.24	1.27	1.21	1.89	2.54	-2.74	-2.60	-2.76
6,8	0.19	1.10	-0.09	0.72	1.24	1.69	-2.68	-2.62	-2.79
7,8	-1.77	-2.02	-2.50	-0.94	-1.25	-1.59	0.74	-0.21	0.03
μ / σ=2		4.83			2.31			0.00	
1,2	3.99	3.77	4.90	5.67	5.75	6.75	5.51	5.26	6.20
1,3	1.97	2.62	2.43	3.39	3.50	4.17	2.39	2.60	3.13
1,4	3.28	3.01	4.59	4.85	4.22	5.78	4.56	4.12	4.87
1,6	4.09	4.73	4.96	4.78	4.21	5.97	2.09	2.33	3.80
1,7	4.80	6.01	5.96	5.20	5.18	6.15	0.74	1.66	2.92
1,8	4.84	5.72	5.73	4.56	4.61	6.33	0.97	1.62	3.01
2,3	-2.46	-1.66	-3.07	-3.05	-2.66	-4.33	-3.61	-3.11	-3.90
2,4	-2.38	-2.70	-2.67	-3.92	-4.75	-5.43	-3.71	-4.06	-4.36
2,6	0.37	2.13	-0.25	-1.92	-2.05	-1.95	-2.98	-2.11	-3.14
2,7	1.44	3.46	1.39	-1.05	-0.40	-0.86	-5.12	-4.20	-4.71
2,8	1.33	3.18	0.58	-1.34	-1.05	-1.17	-5.04	-4.30	-5.05
3,4	1.56	0.17	2.42	1.37	0.16	2.17	2.20	1.46	1.98
3,6	3.07	3.56	2.74	1.58	0.71	2.94	0.01	0.30	0.20
3,7	4.49	5.55	4.99	3.06	2.85	4.05	-2.26	-1.59	-1.38
3,8	4.52	5.20	4.05	2.35	1.98	3.88	-1.99	-1.65	-1.53
4,6	1.72	3.33	1.01	0.01	0.51	0.72	-1.58	-0.52	-1.23
4,7	2.58	4.62	2.64	1.22	2.20	1.84	-3.97	-2.66	-2.87
4,8	2.57	4.35	1.87	0.71	1.44	1.61	-3.86	-2.80	-3.21
6,7	1.48	2.17	2.00	1.62	2.42	1.60	-1.84	-1.47	-1.84
6,8	1.42	1.89	1.14	0.96	1.33	1.23	-1.66	-1.57	-2.13
7,8	-0.41	-0.84	-1.78	-0.88	-1.71	-0.95	1.07	-0.47	-0.41

Table 2.3(b) Paired t-statistics

$(n_1, n_2) = (25, 25)$									
Methods	Scores								
	Brier	ε-Log	MSE	Brier	ε-Log	MSE	Brier	ε-Log	MSE
σ=3 \ μ	6.26			2.40			0.00		
1,2	4.55	4.43	5.30	5.38	5.27	6.27	4.61	4.72	5.35
1,3	1.84	2.71	2.94	2.68	2.78	3.99	1.91	2.22	2.19
1,4	3.36	2.83	4.49	4.14	3.56	5.35	4.52	4.19	5.09
1,6	3.00	4.00	4.93	4.24	4.50	6.04	3.62	3.66	4.46
1,7	3.80	5.26	5.26	3.09	3.86	5.29	1.83	2.47	3.16
1,8	3.49	4.87	4.98	2.35	3.01	4.05	2.01	2.42	3.17
2,3	-3.03	-2.20	-3.24	-3.00	-2.63	-3.80	-2.95	-2.89	-4.10
2,4	-3.31	-4.21	-4.08	-3.83	-4.12	-4.54	-1.83	-3.07	-2.89
2,6	-1.49	-0.23	-0.81	-1.12	-0.83	-1.69	-0.63	-0.33	-1.13
2,7	-0.44	1.52	0.05	-2.03	-1.11	-1.64	-3.06	-2.86	-3.67
2,8	-0.85	1.09	-0.28	-3.25	-2.73	-3.06	-3.06	-2.99	-3.66
3,4	1.51	-0.41	1.59	0.86	0.22	1.63	2.71	1.96	3.87
3,6	1.71	2.22	2.48	1.78	1.66	2.28	2.01	2.08	2.65
3,7	3.27	4.46	3.58	0.95	1.61	2.44	-0.20	0.08	0.51
3,8	2.93	3.94	3.21	-0.53	-0.17	0.39	-0.11	-0.05	0.49
4,6	0.28	2.22	1.11	1.11	1.48	0.73	0.18	1.03	-0.00
4,7	1.29	3.77	1.95	0.04	1.12	0.66	-2.43	-1.55	-2.79
4,8	0.88	3.36	1.61	-1.21	-0.33	-0.95	-2.41	-1.69	-2.77
6,7	1.52	2.96	1.49	-1.44	-0.50	-0.03	-2.51	-2.37	-2.78
6,8	0.97	2.26	0.96	-3.55	-2.73	-3.00	-2.49	-2.48	-2.75
7,8	-1.48	-2.49	-1.65	-1.88	-2.16	-3.18	0.22	-0.47	0.01

Table 2.4(a) Paired t-statistics

$(n_1, n_2) = (10, 25)$									
Methods	Scores								
	Brier	ε-Log	MSE	Brier	ε-Log	MSE	Brier	ε-Log	MSE
μ / σ=1		3.16			1.40			0.00	
1,2	3.66	3.65	4.63	4.03	4.22	5.03	4.95	5.06	5.99
1,3	2.69	2.73	2.68	2.11	2.50	2.78	2.08	2.50	1.74
1,4	4.33	4.42	5.54	3.37	2.98	4.24	1.96	1.80	2.21
1,6	4.46	5.27	4.01	4.13	4.04	5.73	0.71	1.20	-0.05
1,7	5.88	6.38	5.72	3.86	3.87	5.25	0.22	0.79	-0.46
1,8	5.78	6.41	5.27	3.60	3.52	4.72	0.79	1.38	-0.05
2,3	-1.68	-1.67	-3.24	-2.45	-2.33	-3.33	-3.28	-3.06	-5.22
2,4	-1.92	-2.17	-2.37	-3.12	-4.45	-4.69	-7.72	-7.65	-10.60
2,6	1.30	2.45	1.88	-0.28	0.04	-0.37	-2.93	-2.39	-4.68
2,7	3.37	4.17	3.92	-0.14	0.07	-0.36	-3.50	-2.93	-5.36
2,8	2.72	3.78	3.63	-0.50	-0.40	-0.78	-3.14	-2.52	-5.07
3,4	0.75	0.57	2.82	0.86	-0.09	1.11	-0.26	-1.04	0.45
3,6	2.88	3.89	3.31	2.57	2.44	3.51	-0.66	-0.35	-1.27
3,7	4.91	5.65	5.27	2.99	2.67	4.22	-1.48	-1.08	-2.21
3,8	4.29	5.28	4.76	2.44	2.06	3.40	-0.72	-0.31	-1.52
4,6	2.22	3.53	2.38	1.48	2.12	2.15	-0.43	0.25	-1.37
4,7	4.14	5.02	4.37	1.58	2.14	2.03	-0.93	-0.20	-1.88
4,8	3.62	4.78	4.04	1.18	1.69	1.50	-0.40	0.39	-1.46
6,7	2.98	3.00	4.21	0.23	0.06	-0.11	-0.91	-0.76	-0.72
6,8	2.30	2.62	3.62	-0.42	-0.71	-0.72	0.12	0.20	0.01
7,8	-1.96	-1.71	-0.16	-1.38	-1.78	-3.13	1.65	1.79	1.30
μ / σ=2		4.92			2.47			1.16	
1,2	3.52	3.36	3.71	4.72	4.55	5.95	5.24	4.84	6.19
1,3	1.44	1.83	1.26	2.34	2.55	2.40	2.23	2.18	2.76
1,4	3.70	3.30	4.05	2.74	2.55	3.46	3.61	2.73	4.20
1,6	3.92	5.03	5.45	3.27	3.21	4.09	3.24	3.51	4.40
1,7	4.06	5.39	5.46	3.23	3.29	4.59	0.66	1.42	1.51
1,8	4.35	5.63	5.84	3.69	3.49	4.68	2.39	2.58	3.00
2,3	-1.84	-1.67	-2.70	-2.17	-2.00	-3.90	-3.56	-3.00	-4.83
2,4	-1.93	-2.50	-2.38	-4.93	-5.78	-7.69	-5.64	-5.90	-8.06
2,6	0.59	2.14	1.58	-0.81	-0.30	-0.71	-1.90	-0.71	-2.38
2,7	1.29	2.86	3.25	-0.90	-0.22	-1.13	-3.65	-2.03	-4.22
2,8	1.58	3.15	3.17	-0.59	-0.02	-0.94	-2.16	-0.79	-2.97
3,4	1.35	0.74	2.15	-0.06	-0.43	0.49	1.02	-0.01	1.49
3,6	2.57	3.75	4.53	1.23	1.49	2.45	1.65	2.16	2.37
3,7	3.64	5.00	5.12	1.36	1.82	3.09	-1.28	0.03	-0.92
3,8	3.72	5.06	5.32	1.67	1.99	3.24	0.92	1.60	1.02
4,6	1.35	3.30	2.72	1.25	1.75	2.18	0.53	1.88	0.94
4,7	1.96	3.84	3.93	1.03	1.70	2.12	-1.73	0.03	-1.76
4,8	2.25	4.13	3.97	1.41	1.90	2.31	-0.03	1.29	-0.16
6,7	1.04	1.50	3.04	-0.17	0.07	-0.43	-2.70	-1.94	-3.21
6,8	1.50	2.10	2.89	0.25	0.35	-0.20	-1.15	-0.40	-1.66
7,8	0.62	0.71	-0.53	1.04	0.79	0.88	2.43	2.22	2.93

Table 2.4(b) Paired t-statistics

$(n_1, n_2) = (10, 25)$									
Methods	Scores								
	Brier	e-Log	MSE	Brier	e-Log	MSE	Brier	e-Log	MSE
μ / σ=3	6.56			2.92			0.00		
1,2	3.37	3.38	3.37	4.96	4.46	6.38	3.75	3.64	5.43
1,3	1.88	2.11	1.01	1.66	1.88	2.71	0.97	1.17	1.44
1,4	3.11	2.56	2.89	3.12	2.50	4.20	3.06	2.17	4.58
1,6	2.99	4.08	4.30	2.14	2.26	4.25	1.75	1.84	3.29
1,7	3.81	5.55	4.92	2.13	2.77	3.62	0.92	1.42	2.34
1,8	3.79	5.04	4.96	2.64	2.97	3.74	1.00	1.46	2.24
2,3	-1.05	-1.18	-2.44	-3.04	-2.40	-4.61	-2.82	-2.19	-4.73
2,4	-2.52	-3.38	-3.30	-5.62	-5.55	-7.62	-3.34	-3.91	-5.53
2,6	0.37	1.64	0.98	-2.41	-1.32	-2.39	-2.38	-1.41	-2.58
2,7	1.27	3.10	2.73	-1.89	-0.40	-2.56	-2.59	-1.73	-3.78
2,8	1.20	2.61	2.36	-1.61	-0.24	-2.53	-2.62	-1.78	-3.92
3,4	0.28	-0.22	1.33	0.97	0.04	1.30	1.48	0.32	2.58
3,6	1.95	3.46	4.64	0.77	0.99	2.29	0.85	0.92	2.14
3,7	3.40	5.28	4.88	1.08	1.82	2.09	0.14	0.66	1.37
3,8	2.84	4.39	4.86	1.55	2.02	2.18	0.17	0.60	1.14
4,6	1.19	2.98	2.39	-0.28	0.83	0.81	-0.79	0.62	-0.05
4,7	2.13	4.48	3.71	-0.02	1.48	0.51	-1.25	0.13	-1.28
4,8	1.99	3.89	3.48	0.36	1.67	0.57	-1.26	0.10	-1.49
6,7	1.37	2.42	2.79	0.32	1.23	-0.40	-0.86	-0.57	-1.31
6,8	1.65	2.17	2.52	0.95	1.53	-0.31	-0.91	-0.63	-1.54
7,8	-0.06	-0.57	-1.06	1.27	0.86	0.22	0.06	-0.20	-0.85

Table 2.5(a) Paired t-statistics

$(n_1, n_2) = (25, 10)$									
Methods	Scores								
	Brier	e-Log	MSE	Brier	e-Log	MSE	Brier	e-Log	MSE
σ=1 \ μ		3.16			1.40			0.00	
1,2	3.25	3.17	4.01	3.84	3.94	5.22	4.77	4.39	6.93
1,3	2.25	2.80	2.19	0.84	1.63	1.74	2.05	1.98	2.22
1,4	3.03	3.00	3.68	3.18	2.73	4.05	1.63	1.45	1.73
1,6	4.13	5.01	4.49	3.21	3.58	4.76	-0.13	0.39	-2.46
1,7	3.65	4.54	5.18	3.78	4.48	5.51	-0.28	-0.01	-2.16
1,8	3.48	4.30	4.69	3.80	4.39	5.37	-0.88	-0.14	-2.84
2,3	-1.45	-0.96	-2.53	-3.12	-2.80	-4.29	-2.08	-2.41	-4.33
2,4	-0.63	0.03	-0.07	-2.88	-3.16	-3.40	-7.17	-6.15	-9.20
2,6	1.69	3.18	2.81	-0.12	0.32	-0.11	-3.70	-3.25	-7.38
2,7	1.55	3.10	3.86	0.47	1.17	0.92	-3.98	-3.78	-7.85
2,8	2.02	3.56	3.79	0.52	1.08	0.76	-4.09	-3.52	-7.57
3,4	0.80	0.72	2.14	1.82	0.76	2.42	-1.12	-1.06	-0.93
3,6	2.69	3.71	3.93	3.50	3.12	3.81	-2.25	-1.63	-4.62
3,7	2.62	3.57	4.83	5.52	5.02	6.96	-3.07	-2.74	-5.60
3,8	2.90	3.78	4.37	5.26	4.65	5.97	-3.13	-2.43	-5.11
4,6	1.87	2.85	2.69	1.21	1.76	1.27	-1.13	-0.57	-3.33
4,7	1.69	2.70	3.70	1.81	2.55	2.41	-1.28	-0.97	-3.24
4,8	2.14	3.15	3.72	1.86	2.46	2.23	-1.74	-1.05	-3.50
6,7	-0.12	0.04	1.36	1.00	1.38	1.53	-0.30	-0.78	0.20
6,8	0.43	0.69	0.95	1.10	1.31	1.54	-1.54	-1.04	-0.60
7,8	0.61	0.74	-0.37	0.16	-0.41	-0.39	-1.04	-0.28	-0.73
σ=2 \ μ		4.35			1.14			0.00	
1,2	3.11	3.01	4.11	4.33	4.15	4.95	4.11	4.49	5.04
1,3	2.29	2.68	2.31	1.94	2.08	1.90	0.42	1.31	0.47
1,4	3.26	2.95	3.68	3.20	2.62	3.42	3.09	2.99	2.97
1,6	4.29	5.25	6.00	2.78	2.71	3.56	2.03	2.51	3.05
1,7	3.98	5.02	5.78	2.90	2.95	4.99	1.80	2.46	3.22
1,8	4.15	5.05	5.80	3.00	2.97	4.65	1.72	2.37	2.98
2,3	-1.02	-0.73	-2.42	-2.27	-2.42	-3.70	-3.14	-3.39	-4.57
2,4	-1.07	-0.94	-0.78	-2.38	-2.42	-3.23	-2.20	-2.78	-4.60
2,6	1.71	3.04	3.00	-1.02	-1.31	-1.50	-0.99	-1.26	-1.43
2,7	1.51	2.97	2.79	-1.17	-1.39	-1.49	-1.35	-1.58	-1.57
2,8	1.67	2.99	2.73	-0.59	-0.58	-0.77	-1.46	-1.73	-1.80
3,4	0.32	-0.08	1.63	0.73	0.44	1.76	1.70	1.28	2.04
3,6	3.05	3.93	5.24	1.30	0.97	2.46	1.99	1.58	2.98
3,7	2.76	3.90	5.04	1.23	0.92	2.60	1.60	1.36	2.73
3,8	2.93	3.94	5.11	1.90	1.78	3.47	1.48	1.18	2.45
4,6	2.22	3.35	3.45	0.29	0.33	0.33	0.02	0.13	0.60
4,7	1.98	3.21	3.30	0.15	0.21	0.22	-0.35	-0.14	0.43
4,8	2.15	3.24	3.27	0.63	0.75	0.89	-0.44	-0.27	0.19
6,7	-0.42	0.11	-0.64	-0.31	-0.28	-0.16	-0.66	-0.51	-0.27
6,8	-0.05	0.20	-0.88	0.59	0.73	1.04	-0.86	-0.76	-0.70
7,8	0.88	0.29	-0.56	1.26	1.29	1.58	-0.81	-1.32	-1.38

Table 2.5(b) Paired t-statistics

$(n_1, n_2) = (25, 10)$						
Methods	Scores					
	Brier	ε-Log	MSE	Brier	ε-Log	MSE
μ / σ=3	5.42			0.00		
1,2	2.83	2.86	4.04	4.07	4.10	4.44
1,3	2.65	2.79	2.56	1.20	2.04	1.23
1,4	2.76	2.40	3.41	3.03	2.88	3.86
1,6	3.55	3.56	5.66	3.43	4.05	4.57
1,7	3.84	4.43	6.47	3.77	4.22	4.83
1,8	3.78	4.43	6.36	3.80	4.21	4.93
2,3	-0.16	-0.30	-1.85	-3.27	-3.04	-3.57
2,4	-1.73	-2.35	-2.34	-0.41	-0.89	-0.78
2,6	0.99	1.51	2.00	0.56	1.07	0.08
2,7	1.67	2.67	2.92	0.58	0.97	0.31
2,8	1.69	2.80	2.83	0.61	0.98	0.37
3,4	-0.99	-1.52	0.24	1.98	1.33	2.68
3,6	1.28	1.92	4.07	3.29	3.65	3.91
3,7	2.30	3.32	5.39	3.90	4.01	4.49
3,8	2.31	3.43	5.10	3.94	4.01	4.61
4,6	2.19	2.85	3.62	0.82	1.65	0.54
4,7	2.53	3.62	4.28	0.81	1.48	0.75
4,8	2.48	3.62	4.16	0.84	1.52	0.81
6,7	1.19	1.85	1.88	-0.07	-0.38	0.38
6,8	1.22	2.04	1.77	-0.03	-0.37	0.48
7,8	0.05	0.50	-0.32	0.33	0.12	0.94

change from 5%, at 20% error at least one of 6-8 moves up to 4th place or better (displacing 2) for each σ, though less noticeably for σ=1 and 10:25. For σ=1 however 1 is still generally superior, though for equal samples 3 is no longer better than 6 and 8, but for σ=2 and 3, 2 is generally poor if not last and for n_i=25 at σ=3 8 is significantly better than 2, and 1 better than 2 and 4 alone for n_i=10 and possibly no better than 8 for n_i=25. For 10:25 and σ=3, 1 may be no better than 6 and 7 except on MSE. At 50% error, 2 is last and generally poorest for all σ, and now at least 2 of 6-8 (usually 7 and 8) are in the 1st four, most notably for σ=1, n_i=10 and 25:10, where uniquely 1 is poorer than a direct method as 8 improves on each of 1-4, at least on MSE. For n_i=25 1 is no better than (3 and) 6-8 except on MSE, nor are 1 and 3 different from 6-8 for σ=3 and n_i=10, and 1 clearly betters 6 (and 2) only for n_i=25. For unbalanced samples, 25:10 has 1, 8 and 7 all better than 2, if not 3 also. For σ=2 at least one of 7 and 8 are in the 1st four, and for 10:25 both better 6, which is generally last along with 2. 2 is worse than at least 1 for σ=3 also, where there are no other differences.

More specifically, from Tables 2.2-2.5, for 10:10, there is little difference between 3 and 4. While some differences are nearly significant, notably for 5%, the only significant result is that 3 betters 4 for 5% with σ=2 (on Brier score only). For 5%, 3 is better than 6-8 for σ=1 and 2, and better than 6 and 7 for σ=3. For 20%, 3 is better only than 7 and 8 for σ=1, better than each of 6-8 on at least one score for σ=2, but no longer better for σ=3 where there are no differences for 50% either. For 50%, the only differences between 3 and 6-8 are on MSE for which 3 is worse than 7-8 for σ=1. Comparing 4 to 6-8, for 5% 4 is better than 7 on MSE for σ=1 and better than 8 on at least one score. 4 is also better than 7 for σ=2 and on at least one score for σ=3. For 20% there are no differences, while for 50%, for σ=1 and 2, 4 is worse than 8 on MSE at least. 6-8 are the same at 20% for all σ, 6 better than 7 at 5%, σ=2 only where 7 is also worse than 8 on MSE, as is also true for σ=3 (on both ε-log and MSE). For σ=2 alone, at 50% 7 is poorer than 8. 2 is consistently poorer than 3 and 4 at 50%, is poorer on at least 2 scores for 20% (not quite significantly poorer than 3 for σ=2). For 5%, it is worse than 4 only on MSE (at least for σ=3), but declines as error rate increases, being poorer than 6-8 for all scores at 50% for σ=1 and 2, and poorer than 7 and 8

only, on at least 2 scores. for σ=3 where it is approaching significance or worse than 6-8 for 20% as well. 1 is better than 2 for each case, and for 5% 1 is better than all except 3 for σ=1 and 2. For σ=3 1 is better than 6 and 7 on at least one score and is not far from improving on 4 also. For 20%, again 1 is better than all except 3 for σ=1 and 2. and better than 3 also. on MSE only, for σ=1 (and almost σ=2), but for σ=3 1 is better than 2 and 4 alone (and not quite significantly better than 6-8 on MSE). For 50%, for σ=1 1 is now worse than 8 on MSE but not significantly worse than 6 and 7. For σ=2, 1 is better than 4 on MSE only, and for σ=3 does not decline as much relative to 6-8, though the differences between 1 and 6-8 are nonsignificant, and 1 is still better than 4.

For 25:25 there are more differences, which are now more consistent. 3 is better than 4 on MSE only for σ=1, 5% and 20%, and on 2 out of 3 scores for 50%, σ=3. For 5%, σ=1 and 2, 3 is better than 6-8 while for σ=3 it betters 7 and 8 only. For 20%, 3 is now only better than 6 on MSE for σ=2 and better than 8, on MSE at least, for σ=2 and possibly σ=1, but still better than 7 for both σ=1 and 2. As error rate increases 3 declines relative to 6-8 (though for σ=3, at 50% 3 is still better than 6 on at least MSE) so that at 50% there are no differences for σ=2, while for σ=1 3 is now clearly worse than both 7 and 8. The same is also evident for 4 relative to 6-8, as at 5%, for σ=1 and 2, 4 betters 7, and 6 and 8 on at least one score, though 4 is better than 7 and 8 only and on MSE alone for σ=3, but for 20% there are no significant differences, while for 50% 4 is worse than both 7 and 8 for σ=1 and 2, and also for σ=3 (on MSE at least). Between 6-8 there are few differences. 6 is worse than both 7 and 8 for σ=1, 50% and on at least MSE for σ=3, where for 20% error 7 is also worse than 8, again on MSE. For 5%, σ=3, 6 is better than 7 but only on ϵ-log. 2 is consistently poorer than 3 and 4 especially for higher error rates. It is better than 7 and 8 only for σ=1 and 2, 5%, (on ϵ-log) and declines, being poorer than 8 for 20% and σ=3, poorer than 6-8 for both σ=1 and 2 at 50% and also for σ=3 poorer than 7 and 8. Once again 1 is better than 2 in all cases. 1 is better than all others for 20% and all σ, and for 5% clearly better than all except 3 which it betters on at least one score. For 50%, 1 is better than 4 for all σ, better than 3 for σ=1 and 2, better than 6 on at least one score especially as σ increases, and better than 7

and 8 for at least one score for σ=2 and 3 only. As it does for 6, 1 improves relative to 7 and 8 as σ increases.

For unbalanced sample sizes, for $n_1:n_2$ = 10:25 there are few significant differences between 3-4 and 6-8 for any σ and 20-50% error, except that 3 is better than 6-8, at least for σ=1. For 5%, 3 and 4 are better than 6-8. 6 improves on 7 and 8, especially for 5% and 20%. In each case 2 is generally worse than 3 and 4, at least for higher error rates, and also declines relative to 6-8 as error increases, from being superior at 5% to inferior at 50%. For 1 versus the rest, as for balanced sample sizes, 1 is always better than 2. For 5%, σ=1, 1 is best overall and also better than all except 3 for σ=2 and 3, as is true for 20% for all σ (at least on MSE). For 20%, 1 is also better than 3 for σ=1, borderline for σ=2, and better for σ=3 on MSE only, while for 50%, for σ=1 it is only better than 2 with 3 borderline (in particular 1 is no better than 6-8) but for σ=2, 1 is better than all except 7, and for σ=3, it is clearly better only than 2 and 4, bettering 6 on MSE alone.

For the 25:10 configurations, there is generally little difference between 3 and 4, as 3 improves on 4 only for MSE and σ=3, 20%. For 5%, 3 betters 6-8 for σ=1 and 2, and for σ=3 on at least MSE. It is also better for σ=1 and 3 at 20%, and better on MSE for σ=2, 20% or higher, but worse at least on MSE for σ=1, 50%. For 5% and each σ, 4 is better than 6-8 on at least 2 scores including MSE, but there is no difference for σ=2 and 3 for higher error rates. For σ=1, 6-8 are better on MSE for 50% but 4 nearly better than 7 and 8 for 20%. There are no differences between 6-8. 2 is worse than 3 and 4 for 20% or higher (worse than 3 only for σ=3) and for σ=3 may be worse than 4 at 5%. For σ=1, 2 again declines relative to 6-8 from better to markedly worse as error rate increases, but for σ=2 and 3 from better (or same) to nonsignificant. Comparing 1 to the rest, as usual 1 is better than 2 in all cases, for 5% is better than all others, including 3 on at least one score, and for 20% is better than all except 3. For 50%, σ=1, 1 is borderline with 3, but worse than or borderline with 6-8 on MSE, and for σ=2, 1 betters 4, and is superior to 6-8 at least on MSE.

Conclusions

Of the marginal methods, in all cases method 1 is consistently better than 2. 2 is consistently poorer than 3 and 4 which are

usually similar, or 3 slightly superior.

Comparing 1 to the rest, at 5% and 20%, 1 was generally best of all or better than all except method 3, though for n_i=10 at 20% and σ=3 1 could not be distinguished from 6-8, nor from 8 at 5%.

Direct methods 6-8 are similar, especially for 25:10, but 6 may be better than 7 and 8 for 10:25, especially for 5% and 20% error. For equal smaller samples, 6-8 are similar at 20%, but 8 superior for σ=3 and n_i=25, while at 5% 7 may be poorer, and at 50% 7 and 6 respectively worse than 8 for n_i=10 and 25.

Relative to 6-8 the marginal methods generally deteriorated as error rate increased, most notably relative to method 2 (the poorest marginal method) against which 7 and 8, at least, improved from poorer or comparable at 5% to comparable or better at 50%. By 50%, 3, which was generally superior to 6-8 at 5%, may be worse than 7 and 8 for equal samples for σ=1 but comparable to 6-8 for larger σ, and 4 also worse than 7 and 8 for larger samples (n_i=25), being comparable to 6, but worse than 8 only for n_i=10. For 25:10 at 50%, and σ=1 only, 6-8 were better than 3 and superior to 4 on MSE. For 10:25 6-8 were comparable to 3 at 50% for all σ. 6-8 were comparable to 4 by 20% for 10:25, and for equal samples (for which 7 and 8 were superior at 50%), and also for $\sigma > 1$ for 25:10. At 50% for σ=1, 1 was no different than 7 and 8 for n_i=25, worse than 8 and comparable to 6 and 7 for n_i=10, no better than each of 6-8 for 10:25 and worse or nearly worse than 6-8, on MSE, for 25:10. For equal n_i, n_i=10 only, for σ=2 and 3 1 was no better than each of 6-8, but still superior for n_i=25. For σ=2, 1 was better than all except 7 for 10:25 and still superior on MSE for 25:10. For σ=3 and 10:25 1 was still better on MSE than 6 but comparable to 7 and 8.

2.6 APPLICATION OF 2.5 TO A REAL DATA SET

The data used consist of observations on a set of 1000 patients who suffered a severe head injury. These cases were randomly split into a training and test set, each of 500 cases.

The aim was to predict a patient's recovery status as one of 2 outcomes at 6 months after injury, on the basis of age (in years). Population π_1 consists of those who die or will remain in a permanent vegetative state and π_2 those who are severely disabled, moderately disabled or who make a good recovery. Age is one of 2 particularly important feature variables in this context, and was

chosen here to allow comparison with results in Chapter 4, where age has been grouped into 5-year categories and treated as a discrete variable. A fuller description of the data may also be found in Chapter 4. For reasons explained there, the data were reduced to 472 training and 476 test cases, comprising 248 and 239 cases originating from π_1 and 224 and 237 cases of type π_2 respectively. The data are given in Table 2.6 and displayed in Figures 2.49 and 2.50, where the category 70-75 includes all cases of $\geqslant$ 70 years.

Table 2.7 shows the Brier score, log scores and error rate achieved by marginal methods 1-4 and direct methods 6-8, as used in Section 2.5. Of the former, methods 2 and 3 are almost identical while method 4 smooths slightly more in each population, and achieves slightly better scores as a result. Method 1 is strikingly different, smoothing considerably more, and, unlike any other method, gives more smoothing to π_1 than π_2. It achieves better scores all round, and while the differences are not great, the estimated posterior probability function, seen in Figure 2.51(a), is rather more realistic than any other method achieved, including the direct methods 6-8. Of the latter, method 6 is best, indicating the most smoothing and achieving slightly better scores, followed by method 8 with method 7 in 3rd place. Methods 7 and 8 are comparable to method 4 while 6 is slightly better. However in all cases the differences are small and the direct methods do not even approach the very simple Normal Optimal method. Direct optimisation of the test Brier score (see again Table 2.7) shows that method 1 is not quite optimal and that more smoothing of each population increases the Brier score and produces an even smoother predicted probability function (Figure 2.51(b)), although the other scores deteriorate very slightly. It will be noted that while the difference between the scores corresponding to method 1 and the rest may not be thought important (and the differences in error rate are negligible) Figure 2.51 clearly illustrates the potential benefit in attempting to optimise the Brier score, although in this instance cross-validation is not the best means of achieving this. Figure 2.52 which shows a contour plot of the Brier score, calculated on the test data, and the performance of the various methods, also demonstrates this. It will be noted that in relation to the simulations considered in Section 2.5 the contour plot is rather flat, presumably due to the much larger sample sizes.

Table 2.6 Age data

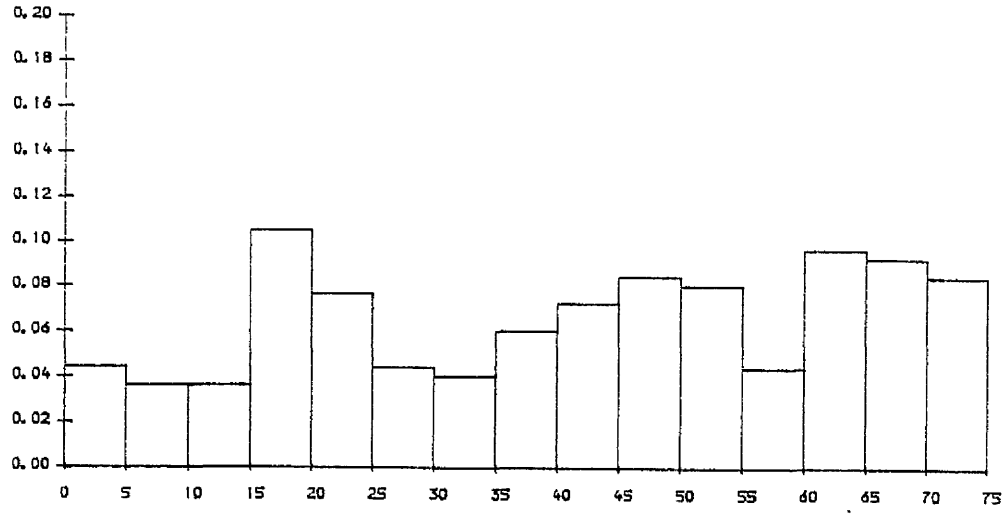
	Training frequencies		Test frequencies			Training frequencies		Test frequencies	
Age	π_1	π_2	π_1	π_2	Age	π_1	π_2	π_1	π_2
0	0	1	0	0	22	3	2	3	5
1	1	1	2	0	23	2	8	8	3
2	7	1	1	5	24	5	6	1	1
3	3	3	1	3	25	3	5	7	8
4	0	4	3	2	26	2	3	6	3
5	2	4	3	9	27	2	7	3	3
6	4	5	1	6	28	0	4	2	1
7	1	3	1	3	29	4	1	2	6
8	2	5	1	9	30	3	5	8	4
9	0	2	3	2	31	4	3	5	3
10	2	3	3	3	32	3	1	1	4
11	1	9	0	2	33	0	2	3	5
12	3	4	3	5	34	0	2	4	2
13	2	4	2	5	35	4	6	4	0
14	1	8	2	2	36	3	2	4	2
15	5	6	1	6	37	2	1	2	1
16	6	10	1	9	38	4	2	2	2
17	3	10	8	2	39	2	3	1	7
18	8	4	5	15	40	4	3	2	6
19	4	13	6	12	41	3	4	4	4
20	3	6	5	6	42	2	2	5	1
21	6	2	2	6	43	4	3	3	3

Continued.

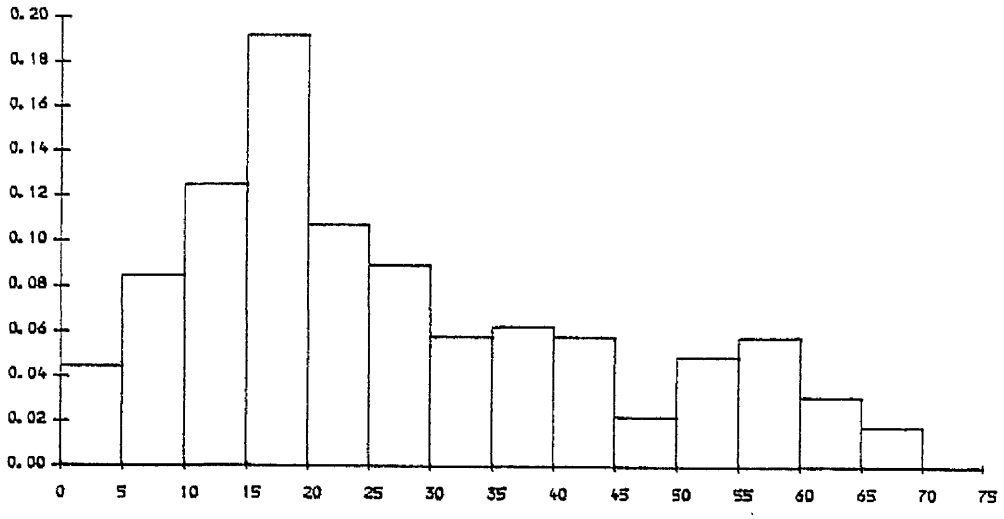
Table 2.6 cont'd. Age data

	Training frequencies		Test frequencies			Training frequencies		Test frequencies	
Age	π_1	π_2	π_1	π_2	Age	π_1	π_2	π_1	π_2
44	5	1	1	5	66	6	1	3	0
45	6	2	5	2	67	4	1	2	1
46	6	2	1	2	68	4	0	3	0
47	3	0	2	2	69	5	0	2	0
48	4	1	3	5	70	1	0	5	1
49	2	0	4	5	71	4	0	2	0
50	4	2	4	3	72	3	0	2	0
51	6	1	5	2	73	0	0	0	0
52	3	4	0	3	74	1	0	2	0
53	3	1	1	4	75	0	0	0	0
54	4	3	4	2	76	3	0	2	0
55	2	2	8	2	77	2	0	3	0
56	0	1	2	0	78	1	0	2	0
57	3	3	1	1	79	3	0	0	0
58	1	2	6	2	80	1	0	2	0
59	5	5	3	3	81	0	0	0	0
60	7	1	7	0	82	1	0	0	0
61	4	2	2	0	83	1	0	1	0
62	4	1	3	1					
63	4	1	3	2					
64	5	2	4	0					
65	4	2	3	3					

Figure 2.49 Histograms of Age distribution for the training data.

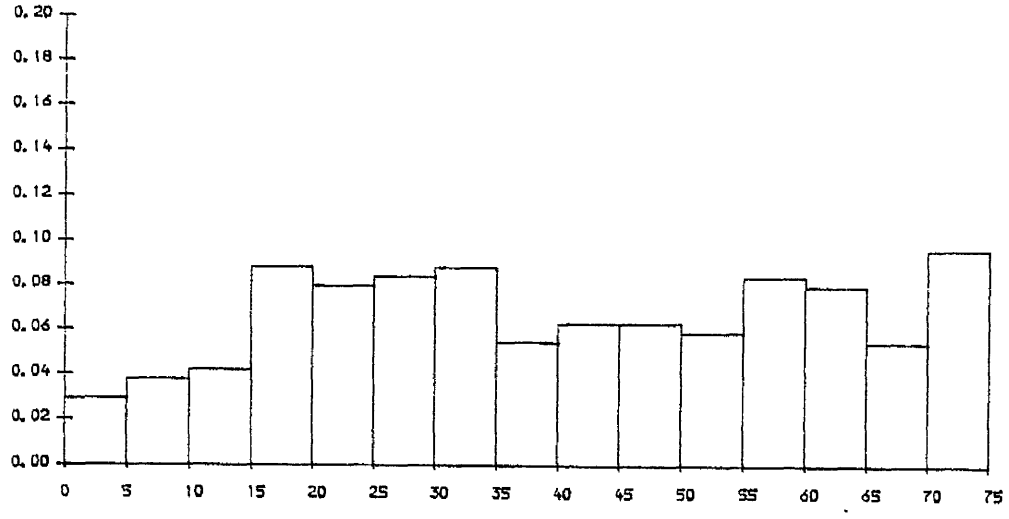


(a) $p(x \mid \pi_1)$

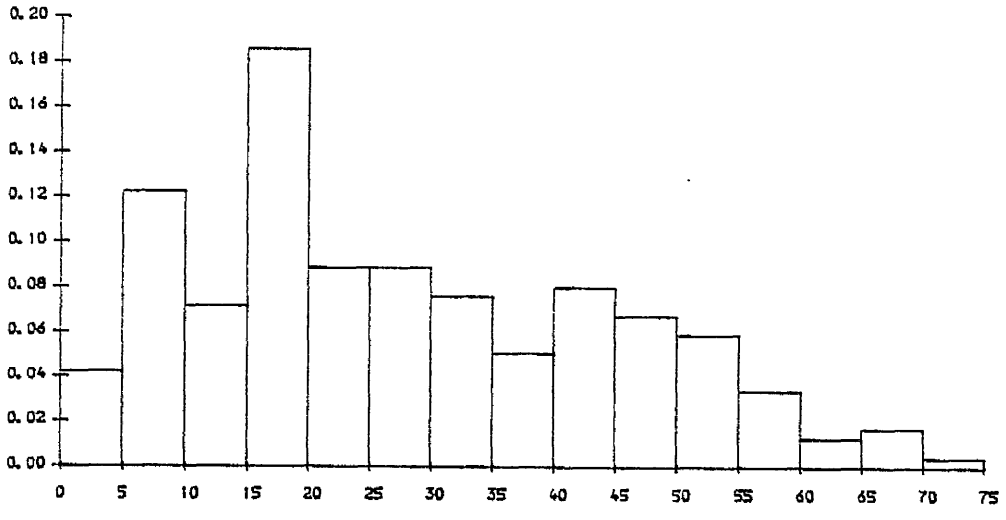


(b) $p(x \mid \pi_2)$

Figure 2.50 Histograms of Age distribution for the test data.



(a) $p(x \mid \pi_1)$



(b) $p(x \mid \pi_2)$

Table 2.7 RESULTS FOR REAL CONTINUOUS DATA

Method	Smoothing Parameters		Scores			
	h_1	h_2	Brier Score	Log Score	ϵ-Log Score	Error Rate (%)
Marginal methods						
1 NOPT (1)	11.917	7.045	.7730	-.6413	-.5941	38.2
2 ASOPT	2.459	2.812	.7671	-.6524	-.6053	40.3
3 XVKL	2.288	2.735	.7667	-.6533	-.6062	40.3
4 XVISE	3.390	3.848	.7692	-.6477	-.6007	39.9
Direct methods						
6 XV BRIER	3.429	4.286	.7694	-.6474	-.6004	39.9
7 XV LOG	3.416	3.552	.7692	-.6478	-.6008	39.9
8 XV ϵ-LOG	3.428	3.703	.7693	-.6476	-.6006	39.9
Spline	β = 220,500		.7309	-.8264	-.7562	38.9
TEST BRIER optimisation	16.589	12.778	.7739	-.6420	-.5945	39.9

NOTE : (1) Fryer's method (2.13) was used here since each sample size is larger than those to which modification (2.14) applies, though the degree of smoothing indicated by the latter was only slightly less in each population.

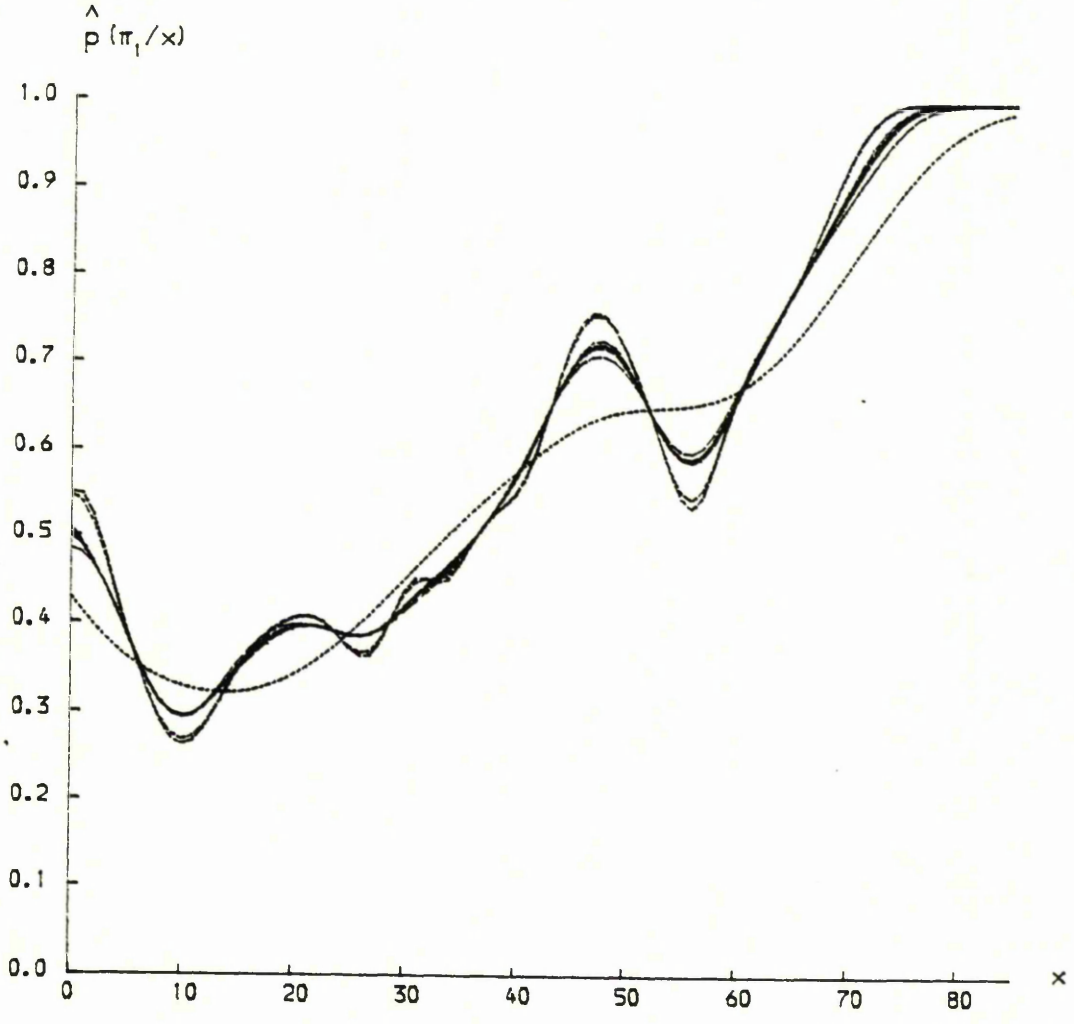


Figure 2.51(a) Predicted probability of π_1 as a function of Age for marginal methods 1-4 and direct methods 6-8. Method 1 clearly stands out as the smoothest curve, methods 2 and 3 smooth least and 4 and 6-8 are intermediate.

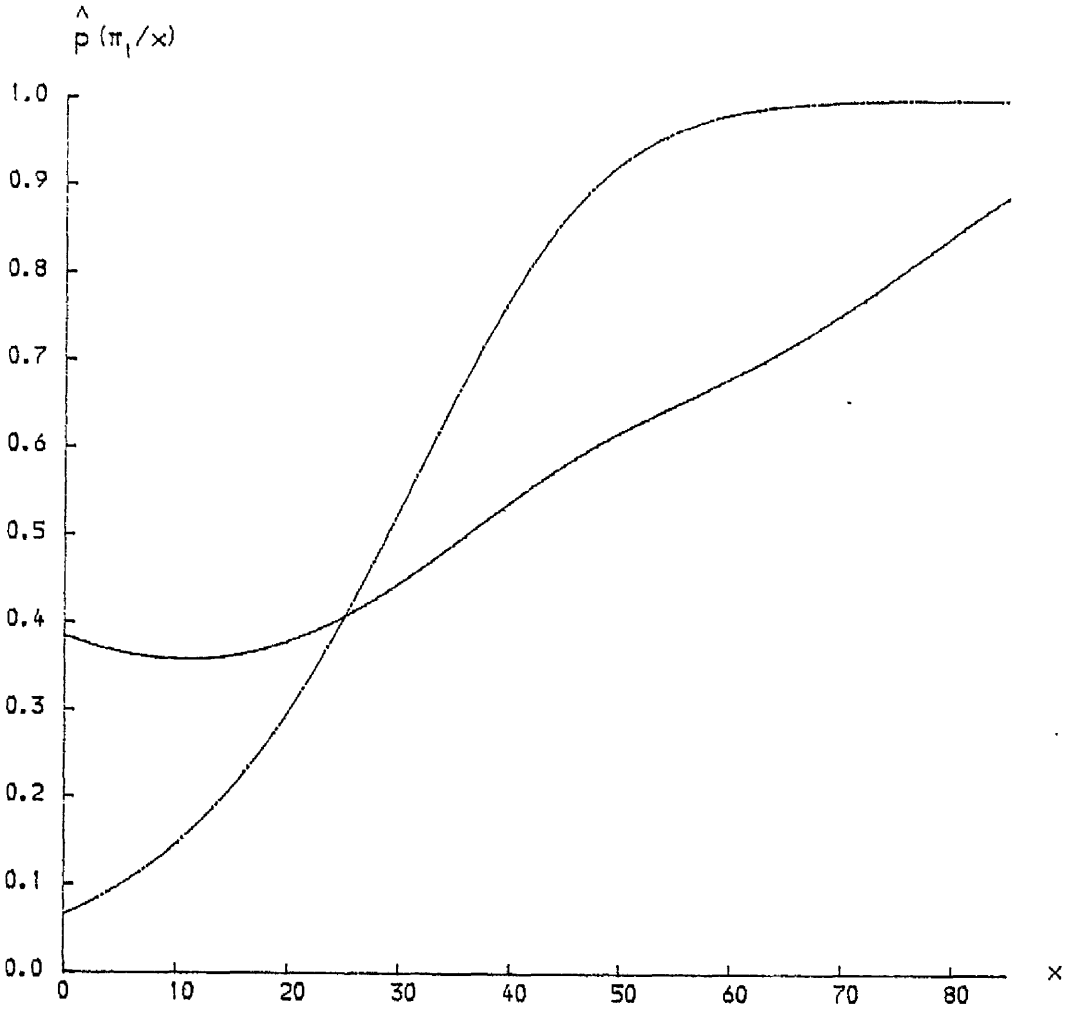


Figure 2.51(b) Predicted probability of π_1 as a function of Age for the Brier-optimal kernel and the spline ratio estimate (sigmoid).

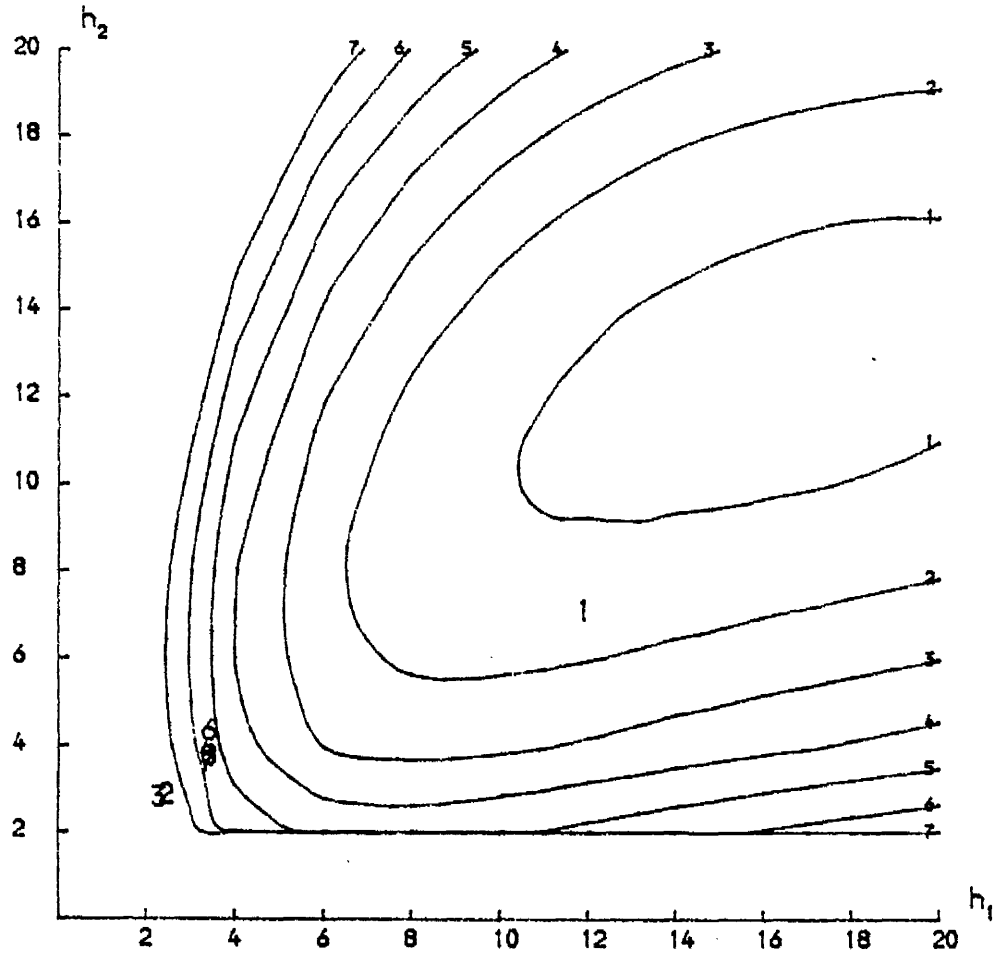


Figure 2.52 Contour plot of the test Brier score as a function of the smoothing parameters (h_1, h_2), where h_i is associated with Age in the ith population, and parameter estimates for methods 1-4 and 6-8. Contour heights are :- .7735, .7715, .7695, .7675, .7655, .7635 and .7615.

Nevertheless Figure 2.51 still shows clear differences between methods.

Methods 2 and 6 are taken as representative of each distinct group of methods, being the best of each group, and estimated class conditional distributions of age for these and method 1 are found in Figures 2.53 and 2.54, together with the Brier-optimal ones. Comparing these to the raw relative frequency histograms in Figures 2.49 and 2.50, method 1, although a marginal method, is seen to grossly oversmooth the raw data in terms of providing an acceptable density estimate, although the effect of the data is still visible using methods 2 and 6. Especially in population 1, method 1, although marginal, effectively applies infinite smoothing. In each case, especially in π_2, the Brier-optimal curves are flatter still. This exemplifies our argument of Section 2.4 that more smoothing than is marginally optimal is likely to be required for optimal estimation of the posterior probability function, or, equivalently, the density ratio, although here it is in fact a marginal method which "oversmooths".

In order to apply the spline ratio estimator of Silverman (1978a), the algorithm for which rests heavily on successive differences in the combined order statistic of the 2 training samples, the many ties in the data were broken by adding to each observation in the training samples an observation generated from a Un(-½, ½) random variable. Here, the smoothing parameter was selected by eye to give a suitable degree of smoothing and comparable scores to those of the kernel methods. The approximately optimal parameter was taken to be β = 220,500, but the problem seems ill-conditioned in that the solution of the non-linear equations (see Sections 1.5.2 and 2.5 above) is heavily dependent on the initial estimate provided, much heavier smoothing than is approximately optimal produces a rougher curve of completely different appearance, comparable to that of a smaller β, and the degree of smoothness/direction of the curve is not a unimodal function of β. The resulting spline estimate is shown superimposed on Figure 2.51(b). Even on this much larger data set, although again smoother than most of the kernel estimates the spline has poorer performance.

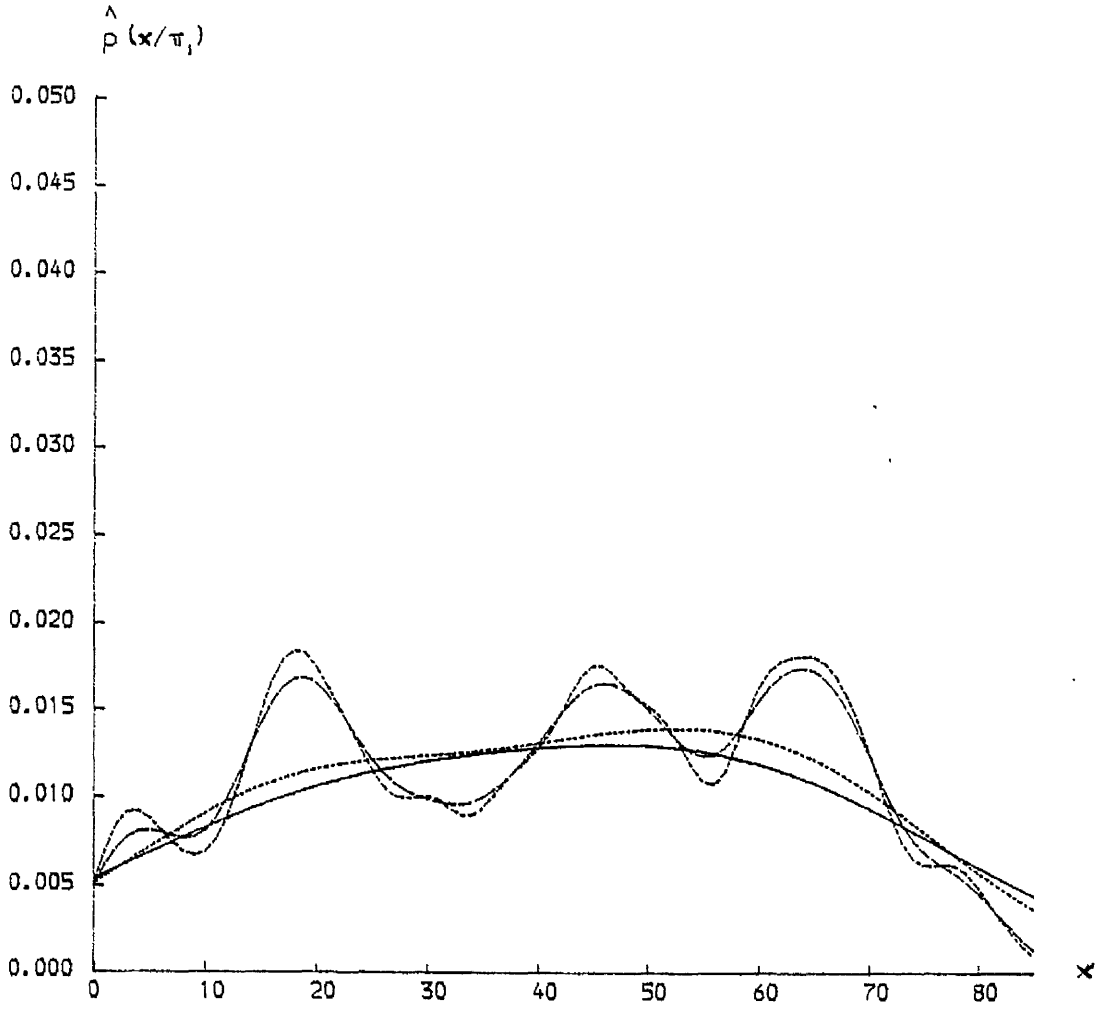


Figure 2.53 Estimated class conditional density of Age in π_1 for marginal method 2, slightly smoother direct method 6 and best method 1. The solid line denotes the Brier-optimal kernel which is slightly flatter again.

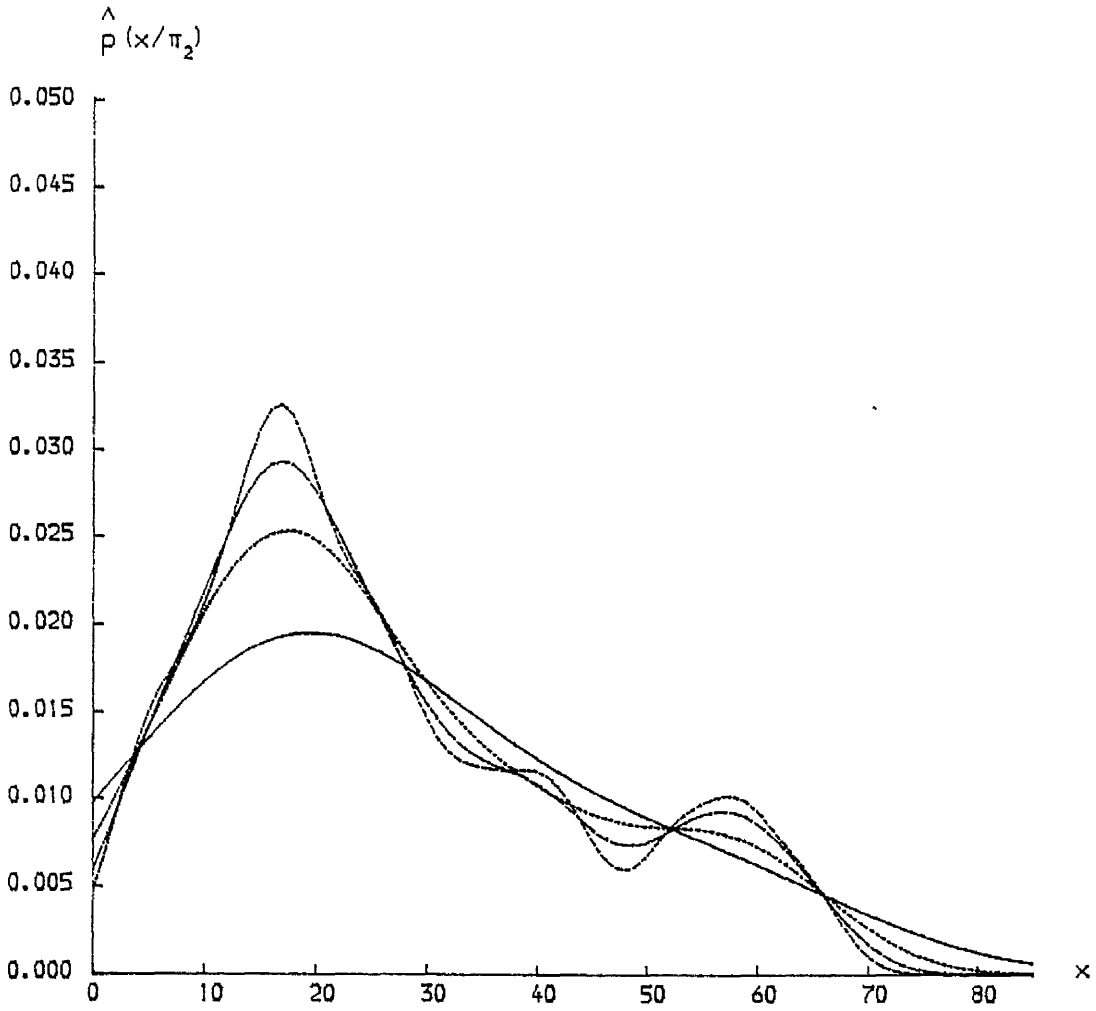


Figure 2.54 Estimated class conditional density of Age in π_2 for marginal method 2, direct method 6 (smoother) and best method 1. The solid line denotes the Brier-optimal kernel which is much smoother still.

2.7 EXTENSIONS

2.7.1 Variable kernels

Variable kernels were introduced in Section 2.2.2. Breiman, Meisel and Purcell (1977), using normal product kernels and samples of size 400 from both a Bivariate Normal and a mixture of Bivariate Normal distributions and estimating smoothing parameters to optimise directly 3 sample-based measures of error, found in each case that the best variable kernel was markedly superior to the best fixed kernel. Habbema, Hermans and Remme (1978) used Multivariate Normal, mixtures of Multivariate Normal, and logNormal distributions in 2 dimensions (and 1 case in 6 dimensions), with sample-based performance measures of a discriminant analysis nature using posterior probabilities, weighting small disagreements between $\hat{p}(\pi_i|\underline{x})$ and $p(\pi_i|\underline{x})$ less than larger ones. They concluded that for skew distributions, as exemplified by the logNormal, variable kernel methods led to substantial improvement over fixed kernels but that for symmetric distributions and mixtures little improvement was made. Raatgever and Duin (1978) studied Normal and logNormal distributions in 1, 2 and 5 dimensions, using smaller samples. Assessing results in terms of the Kolmogorov variational distance,

$$1 - \int_{\underline{x}} \min \{f(\underline{x}), \hat{f}(\underline{x})\} \, d\underline{x},$$

they found that for normal data, variable kernels seemed to improve on fixed ones only for sample sizes large relative to the dimension, but again there was considerable improvement for lognormal distributions. Copas and Fryer (1980) and Bowman (1981) considered transforming univariate skew data to approximate symmetry, applying fixed kernels, then transforming back. Bowman's transformation was in fact an application of variable kernel techniques. He found using data from mixtures of Normals, and measures of goodness-of-fit of the form

$$\int [f(x) - \hat{f}(x)]^2 \, w(x) \, dx \quad ,$$

that one such technique did particularly well for a one-tailed distribution but not so well for a skew distribution. Hand (1982, p. 121) summarised this by saying that "it seems that the variable

kernel should be used in preference to fixed kernels, as it never performs substantially worse than the latter and, especially for skew distributions, will often do markedly better". Abramson (1982) also comments favourably on their use in practice. While it seems therefore that using variable kernels will give an improved fit to a single density and therefore, probably, to the density ratio, we would expect that the conclusions of experiments similar to those in Sections 2.5 and 2.6, but using variable kernels, would be qualitatively the same i.e. that marginal approaches to smoothing parameter estimation are often suboptimal and can be very poor, so that more direct methods appear necessary, choosing smoothing parameters simultaneously. However it is possible that any potential improvement may not be as great with variable kernels, as Breiman et al. (1977) noted that the estimated optimal smoothing parameter for fixed kernel estimators was much more sensitive to the error criterion chosen to assess it than for variable methods.

2.7.2 Multivariate kernels

Rarely in practice would we have only 1 feature variable. We would also expect a more dramatic improvement for the direct assessment methods over "marginal" methods in the multivariate case, although illustrating this by means of contour plots as above will only be possible if a single smoothing parameter is used in each population. (For example, by standardising each variable as in Habbema, Hermans and van den Broek, 1974).

Calculation of MSE extends directly although numerical integration in more than 2 dimensions will now be necessary and may be time consuming. The wholly data-based assessment methods also extend readily but unless standardisation is used optimisation may be difficult as technical problems were encountered with numerical optimisation even in 2 dimensions.

Of the "marginal" methods those based on cross-validation extend directly, but may be impracticably time consuming. In principle the asymptotic optimal method (2.12) extends to multiple dimensions. For a common smoothing parameter h and identical product kernels, Epanechnikov (1969) gives the asymptotically minimising MISE choice for h, while a similar expression may be obtained from results of Cacoullos (1966) for a single h but with a general multivariate kernel. In either case h is $O(n^{-1/(d+4)})$ and

for multivariate normal kernels is defined up to a factor involving multiple integration of a function of second partial derivatives of the true density.

In general, a direct extension of Fryer's Normal Optimal method (2.13) is not possible. Using $MVN_d(\underline{\mu}, \Sigma)$ data, and multivariate normal kernels with an array S of smoothing parameters,

$$MISE(S, n, \Sigma, d) =$$

$$(4\pi)^{-d/2} \{ |\Sigma|^{-\frac{1}{2}} - 2^{d/2+1} |2\Sigma + S|^{-\frac{1}{2}} + n^{-1} |S|^{-\frac{1}{2}} + (n-1)/n \; |\Sigma + S|^{-\frac{1}{2}} \}$$

using results on convolutions of Normal densities (Miller, 1964, pp. 25-26).

Even in the simplest case, using product kernels and common smoothing parameter h, a similar approach to that of Fryer (1976) is not possible. In any given instance, after (possibly robust) estimation of Σ, numerical optimisation would be required to identify the optimal smoothing parameter(s). Murphy and Moran (1986) quote the corresponding formula using the kernel $MVN_d(\underline{0}, hM)$, where M is a positive definite matrix, but minimised it directly in the context of a simulation study.

2.7.3 Multiple populations

With $k > 2$, the MSE generalises to

$$\int_{\underline{x}} \left\{ \sum_{i=1}^{k} [p_T(\pi_i|\underline{x}) - \hat{p}(\pi_i|\underline{x})]^2 \right\} f(\underline{x}) \, d\underline{x},$$

though again it is not possible to identify the minimum MSE choice of $\underline{h}$ diagrammatically and while the data-based assessment methods generalise straightforwardly, again numerical optimisation may be required in more than 2 dimensions.

2.7.4 Discrete variables

Feature variables are often discrete or binary rather than continuous. Product kernels are usually used for multivariate discrete or mixed data, and since computation is easier kernels are more easily applied than in the multivariate continuous case. Direct methods will be appropriate no matter what the data type. Discrete kernels are considered in more detail in Chapters 3 and 4.

CHAPTER 3 ORDERING IN DISCRIMINANT ANALYSIS

3.1 INTRODUCTION

We noted in Section 1.4.2 the nature of ordinal variables. These arise commonly in many fields, particularly in medicine and the social sciences. Some examples are :

1) severity of pain, assessed as "nil/mild/moderate/severe"
2) degree of pain relief after treatment, with categories "worse/same/slight improvement/marked improvement/complete relief"
3) degree classification : 1 / 2(1) / 2(2) / 3 / Unclassified /Fail
4) number of years of education
5) age group : 0-9 / 10-19 / 20-29 /.....

These are seen to range from the very soft "assessed" variables such as 1), through more quantitative orderings to categorisations of a continuous scale of measurement such as 5) at the other extreme. Commonly, methods of modelling ordinal variables assume that the underlying distribution is smooth. For variables such as 5) with an interval scale arising from equal groupings of a continuous variable, a smooth histogram is to be expected, provided that the true distribution is itself smooth. However had age been grouped into categories of unequal width, the smoothness would be lost. Similarly there is no reason to assume that data collected on degree of pain relief would yield a smooth histogram since the steps on scale 2) are not equal, and there is no reason why the proportion of cases experiencing "slight improvement" should be close to that of "marked improvement", say. This has important implications for modelling as the degree to which an ordered variable resembles an interval scale should be reflected in the approach chosen. Anderson (1984) also notes this.

Most standard categorical data methods treat all variables as nominal in that the results are invariant under permutations of the categories. Agresti (1984, p. 3) notes that the distinction between continuous and discrete variables is often rather less important than that between nominal or qualitative variables and ordinal or quantitative ones. Since the latter are inherently quantitative they should be treated more like interval than nominal variables. Some advantages of using ordinal methods are that they have greater power to detect some important alternatives to

independence, and within the standard classes such as log-linear or logistic models, provide a richer hierarchy of models than is possible with the standard nominal models, yet are mostly more parsimonious, easier to interpret, and can be applied where standard models are either trivial or else have too many parameters to be tested for goodness-of-fit. (Agresti, 1984, p. 3).

3.2 TYPES OF ORDERING

Ordering can affect a data set in various ways :

1) through one or more explanatory variables being ordinal in nature, examples of which were given above.

2) through an ordinal response variable, such as recovery status after injury, which might be recorded as "dead / poor / moderate / good recovery". Various authors have recently dealt with ordered outcome models by fitting generalised logistic regression type models, notably McCullagh (1980) and Anderson (1984). (See Section 3.3).

3) A further possibility is that, viewed as a function of $\underline{x}$, response may be ordered with respect to the value or ordering of one or more independent variables, in the sense that the posterior probability of at least one outcome $p(\pi_i|\underline{x})$ is an ordered function of $\underline{x}$, either decreasing or increasing as one or more elements in $\underline{x}$ increase. More generally, rather than be monotonic $p(\pi_i|\underline{x})$ might be convex/concave. The condition

$\sum_{i=1}^{k} p(\pi_i|\underline{x}) = 1$ requires that if $p(\pi_1|\underline{x})$ is ordered in the sense described, then $p(\pi_2|\underline{x})$ obeys the reverse form of ordering if $k = 2$, whereas $p(\bar{\pi}_1|\underline{x})$ does so if $k > 2$.

Conditioning on $\underline{x}$, the training data can be regarded as having arisen from a series of related Binomial experiments. If $\underline{x}$ is univariate then monotonicity of $p(\pi_1|\underline{x})$ requires the same ordering of the Binomial parameters i.e. they are "simply" ordered (see Section 3.5.1), while in the multivariate case $p(\pi_1|\underline{x})$ is monotonic in each element of $\underline{x}$ if the remaining variables are fixed, so that the parameters are non-decreasing or non-increasing in each row/column/layer etc., having a "partial" order (see Section 3.5.1). In this case the degree of smoothness of the marginal p.d.f. $p(\underline{x}|\pi_i)$ is irrelevant. Rather, we require a particular type

of ordering to be exhibited in $\hat{p}(\pi_i|\underline{x})$. Therefore to exploit ordering of this kind, the diagnostic approach, modelling $p(\pi_i|\underline{x})$ directly, appears to be necessary. The logistic regression class of models (Section 1.5.1) is well developed but either typically brings in variables in a linear fashion, which is too strong an assumption with non-interval variables, or more generally relaxes linearity by using a separate parameter to contrast all but one category back to a reference category. The latter, however, ignores the ordered structure completely. More generally still we might express $p(\pi_i|\underline{x})$ as $\phi(\underline{x})$ but bring in the appropriate form of ordering by suitable constraints on the parameters $\{\phi(\underline{x})\}$. This brings us to isotonic regression techniques.

In d dimensions, if $(n_{ij\ldots l})$, the number of observations in π_1 with $\underline{x} = (i, j, \ldots, l)$, are realisations of Binomial $(N_{ij..l}, p_{ij..l})$ random variables, it can be shown (Barlow et al., 1972, p. 102) that the MLEs of the $(p_{ij..l})$ subject to non-increasing rows/columns etc. are the solution of the problem,

$$\underset{G}{\text{minimise}} \sum_i \sum_j \ldots \sum_l \left[\frac{n_{ij..l}}{N_{ij..l}} - g_{ij..l}\right]^2 N_{ij..l}$$

where G is the class of functions $\{g_{ij\ldots l} : i \leqslant p,\ j \leqslant q, \ldots$ and $l \leqslant s \Rightarrow g_{ij\ldots l} \leqslant g_{pq\ldots s}\}$, re-ordering categories if necessary. This is an example of the isotonic regression problem described in Section 3.5.1 below.

MODELS AND METHODS

3.3 MODELS FOR ORDERED RESPONSE

McCullagh (1980) developed a general class of models for a single ordinal response with one or more independent variables, either continuous or unordered categorical variables, based on the assumption of an underlying continuous latent variable, either observable (as with grouped continuous response) or not (for assessed qualitative response). Categories can therefore be thought of as contiguous intervals on a continuous scale.

For k ordered categories the models have the form
link $\{p_j\} = \theta_j - \underline{\beta}^T\underline{x}$ where $\underline{x}$ is the vector of explanatory variables, $p_j = \Pr\{Y \leqslant y_j\}$ $j = 1, \ldots, k$ where y_j is the jth outcome category, and (θ_j), $j = 1, \ldots, k$, are the "cut-points" on the link

function scale. The response $Y = y_j$ can be interpreted as indicating that the underlying continuous variable Z is such that $\theta_{j-1} \leq z \leq \theta_j$. Equivalently $Pr(Y \leq y_j|\underline{x}) = F(\theta_j - \underline{\beta}^T\underline{x})$ where F is any suitable c.d.f.. The logistic link function has the advantage that the resulting parameter estimates are invariant to reversal of the order of the categories, as is desirable for ordinal variables, rather than to arbitrary permutations thereof, as occurs with standard models for nominal variables. This gives the model

$$\left.\begin{array}{l} \log\left[\dfrac{p_j}{1-p_j}\right] = \theta_j - \underline{\beta}^T\underline{x}\ ,\ j = 1,\ldots,k-1 \\ \\ \theta_1 \leq \theta_2 \ldots \leq \theta_{k-1} \end{array}\right\},$$

fitted by means of iterative maximum likelihood techniques. McCullagh (1980) used a version of iteratively reweighted least squares.

Anderson and Phillips (1981) apply this model with more than 1 explanatory variable, and also consider y-conditional sampling. They noted that with respect to parameter estimation, amalgamation of categories gave similar parameter estimates but the loss of information was reflected in higher standard errors. Ashby, Pocock and Shaper (1986) applied the method to a very large data set with 5 explanatory variables.

Anderson (1984) argued against the general applicability of McCullagh type models on the grounds that the "cut-points" $\{\theta_s\}$ are difficult to interpret unless response is directly related to a latent variable, as in the case of a grouped continuous outcome. Secondly, they involve only 1 function $\underline{\beta}^T\underline{x}$ to distinguish between all categories. The "stereotype regression" models on the other hand provide a hierarchy of models within which to test dimensionality (number of independent linear functions needed to distinguish between different categories) and, once this is established, distinguishability. Two categories are "indistinguishable" with respect to $\underline{x}$ if $\underline{x}$ cannot separate them. Only if the model is 1-dimensional should an ordered model be considered. The McCullagh model is inherently a 1-dimensional ordered regression model suitable for a grouped continuous response. For assessed variables, the stereotype model generalises the standard logistic model (1.9) or equivalently (1.11), and therefore has the advantage of belonging to the exponential family.

Structure is imposed on the $(\underline{\beta}_s)$ to generate m-dimensional models. m = 1,.., q where q is the rank of d by k array $(\underline{\beta}_1, \ldots, \underline{\beta}_k)$ i.e. q $\leq$ min(d, k-1). A 1-dimensional model is obtained by constraining the $(\underline{\beta}_s)$ to be parallel i.e. $\underline{\beta}_s = -\psi_s \underline{\beta}$, s = 1,.., k and $\psi_k = 0$, $\psi_1 = 1$ for identifiability. The stereotype ordered regression model requires in addition that $1 = \psi_1 > \psi_2 > \ldots > \psi_k = 0$ (or the reverse, by changing the sign of $\underline{\beta}$). Similarly, a 2-dimensional model has $\underline{\beta}_s = -\psi_s\underline{\beta} - \phi_s\underline{\gamma}$, s = 1,.., k with $\psi_k = \phi_k = 0$, and constraints such as $\psi_1 = 1$, $\phi_1 = 0$ and $\psi_2 = 0$, $\phi_2 = 1$ for identifiability.

Parameter estimation again requires numerical maximum likelihood estimation but is simpler for the y-conditional sampling case than if McCullagh models are used.

The simplest model is chosen on the basis of goodness-of-fit. If a 1-dimensional model is adequate one could then test for orderedness. (The model is ordered automatically if the parameter estimates are.) Indistinguishability is tested for by means of hypotheses of the forms $H_0 : \underline{\beta}_s = \underline{\beta}_t$.

The stereotype ordered model is also ordered in the sense that the class conditional distributions $\{f_s(\underline{x})\}$ are ordered with respect to each other. If $z = \underline{\beta}^T\underline{x}$ then Y_z, where the subscript makes explicit the dependence of outcome Y on z, is stochastically increasing with respect to index z in the sense that $Pr(Y_p > z) \geq Pr(Y_s > z)$ $\forall z$ <=> p > s, $\forall s$, p. Anderson (1984) comments that the models can be modified for use with ordinal regressors.

3.4 MODELS FOR ORDERED EXPLANATORY VARIABLES

3.4.1 Log-linear and logistic models

For ordinal explanatory variables Simon (1974), Haberman (1974), Goodman (1979, 1983) and Agresti (1983, 1984) amongst others have used log-linear and logit models, assuming that meaningful scores can be assigned to the categories. In the bivariate r by c case, for instance, the standard log-linear model has the saturated form

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad \sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_i \lambda_{ij}^{XY} = \sum_J \lambda_{ij}^{XY} = 0$$

with rc - 1 parameters to be estimated, where (m_{ij}) are the expected cell counts. The independence model is the only standard

alternative. If both variables are ordered and suitable scores $u_1 < u_2 < \ldots < u_r$ and $v_1 < v_2 < \ldots < v_c$ can be assigned to the categories of X and Y respectively (usually the integers, $(u_i = i)$, $(v_j = j)$, or at least evenly spaced scores for ease of interpretation), the model

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \beta(u_i - \bar{u})(v_j - \bar{v}) \qquad (3.1)$$

describes the association between X and Y but with only 1 more parameter than the independence model, no matter how many categories there are, and is a special case of the saturated model while $\beta = 0$ gives the independence model. (3.1) is called the "linear-by-linear association" model since for a fixed X the deviation of $\log m_{ij}$ from independence is linear in Y through the scores (v_j) with slope $\beta(u_i - \bar{u})$ and similarly for fixed Y. Non-linear generalisations are also possible by inclusion of interaction terms. Equally spaced scores give the "uniform association" model of Goodman (1979). As with the usual log-linear models iterative maximum likelihood estimation techniques are required for fitting. With 1 ordinal variable Y, say, or where the ordering of the other variable is not relevant, the "linear row effects" model is

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \tau_i(v_j - \bar{v}), \quad \sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_i \tau_i = 0$$

with $(r-1)$ independent association parameters. Similarly, a "linear column effects" model can be defined. Either has the independence model as a special case, putting $\beta = 0$, or $\tau_i = 0$, $\forall i$. The models generalise in the obvious way to higher dimensions. Log-multiplicative models (Agresti, 1984, pp. 138-147) dispense with the need to assign scores a priori by replacing them with parameters to be estimated, but at the expense of increased difficulty in fitting.

Alternatively, logit models of a similar nature are available. For instance, if m_{ijk} is the expected frequency of a binary response k given X = i and Y = j, then the simple additive model for the logit,

$$\log (m_{ij2} / m_{ij1}) = \alpha + \tau_i^X + \tau_j^Y, \quad \sum_i \tau_i = \sum_j \tau_j = 0,$$

for nominal variables X and Y, with r + c - 1 parameters, gives way

to models with a linear effect on the logit through the scores $\{v_j\}$ or $\{u_i\}$,

$$\log(m_{ij2}/m_{ij1}) = \alpha + \tau_i^X + \beta^Y(v_j-\bar{v})$$

or

$$\log(m_{ij2}/m_{ij1}) = \alpha + \beta^X(u_i-\bar{u}) + \beta^Y(v_j-\bar{v})$$

for ordinal Y or both X and Y ordinal, having r + 1 and 3 parameters respectively. $\{\tau_i\}$ and $\{\tau_j\}$ pertain to the partial association between X or Y and the response. α represents the mean of the r x c logits and the $\{\tau\}$ are deviations from the mean due to the location of X or Y. Again β^Y has a slope interpretation, representing the change in the logit, and $\exp\{\beta\}$ the multiplicative change in the odds, for a unit change in the ordinal variable. If response is ordered also, the "cumulative" logits (as used by McCullagh (1980) and Anderson (1984), see Section 3.3) or other suitable logits (Agresti, 1984, pp. 113-115) may be expressed in a similar manner.

We noted that for non-interval ordinal variables, methods such as logistic regression make assumptions which may be inappropriate. The nonparametric kernel method of density estimation can also recognise ordered categories. We describe below its extension to discrete and ordered variables.

3.4.2 Discrete kernels

Smoothed relative frequency estimators

We noted in Section 1.4.2 the difficulties of using the MLE of a probability function given sparse data. In attempts to overcome the problem of empty cells a number of authors (e.g. Good, 1965, pp. 23-25; Fienberg and Holland, 1973; Stone, 1974b) have considered smoothed relative frequency estimators of the form

$$\hat{p}(x) = (1-\alpha)\frac{n(x)}{n} + \frac{\alpha}{k}, \qquad (3.2)$$

k = no. of cells, $0 \leq \alpha \leq 1$, or more generally

$$\hat{p}(x) = (1-\alpha)\frac{n(x)}{n} + \alpha\phi(x) \qquad (3.3)$$

where $\underline{\phi}$ is a vector of probabilities and n(x) is the frequency of outcome x. α is a smoothing parameter to be chosen (see below), so that we are smoothing to a greater or lesser extent away from the MLE and towards uniformity (3.2) or towards a parametric or other suitable model (3.3). The models can be derived from geometrical (Fienberg and Holland, 1973) or Bayesian arguments. From the latter, Good (1965), p. 25, and Fienberg and Holland (1973) put a Dirichlet prior on $\{p(x)\}$ with mean $\underline{\phi}(x)$, deriving $\hat{p}(x)$, though with a slightly different parameterisation, as the mean of the posterior distribution. Therefore (3.2) reflects prior ignorance. For multiple variables, Fienberg and Holland (1973) chose $\underline{\phi}(\underline{x})$ to reflect the multivariate structure of the data. Leonard (1977) provides an approximate alternative to exact Bayesian estimates. For the univariate case with ordered categories, an alternative Bayesian approach is that of Leonard (1973), extending work of Whittle (1958), who incorporates ideas of smoothness into the covariance matrix of a Multivariate Normal distribution on the logit cell probabilities, though this leads to non-linear estimators. Leonard (1975) generalised this to 2-dimensional tables. We noted in Section 1.4.3 the Bayesian spirit of maximum penalised likelihood methods. Titterington and Bowman (1985) consider related discrete minimum penalised distance methods, some of which also yield linear estimators.

Fienberg and Holland (1973) noted that since the MLE is also the minimum variance unbiased estimator, no unbiased estimator can have a smaller risk function, defining risk as $n\, E\, \Sigma(\hat{p}(\underline{x}) - p(\underline{x}))^2$, nor can a biased estimator have uniformly lower risk as Johnson (1971) has shown that the MLE is admissible with respect to risk. However this is due to the MLE performing well where $\underline{p}(\underline{x})$ is extreme (one element of $\underline{x}$ being close to 1). In some small sample comparisons (3.3) was found to be superior in terms of risk to the MLE for more moderate $\underline{p}(\underline{x})$ over a large region of the sample space especially as the number of categories increases. As $n \to \infty$ for fixed k, the risks are approximately equal so that (3.3) is also consistent, while for $n \to \infty$ and $k \to \infty$ such that n/k is constant (corresponding to sparse data) the smoothed estimator has almost uniformly smaller risk, $\forall \{\underline{p}(\underline{x})\}$.

The kernel estimator

The kernel method is equally applicable to discrete as to

continuous data. Instead of centering a continuous p.d.f. at each observation, a histogram is used so that the resulting estimate is built up by each observation contributing a fixed proportion λ to the cell in which it lies, rather than the usual weight of 1. λ plays the role of the smoothing parameter and, as in the continuous case, essentially determines the shape of the resulting estimate. The exact manner in which the remaining probability mass is allocated to the other cells depends on the type of data. For nominal discrete variables it is divided equally between the remaining cells whereas for ordinal ones adjacent cells recieve more weight than more distant ones. This may be seen to be more appropriate for grouped continuous data where the underlying distribution is in fact smooth, so that the probabilities of neighbouring cells are very similar, than for more qualitative non-interval scaled ordered variables such as severity of pain. Therefore one might expect kernels for ordinal data to perform rather better on the former type of data than the latter.

Aitchison and Aitken kernels

The simple form of kernel just described is due to Aitchison and Aitken (1976) and is the most widely used discrete kernel. Formally, for a single binary variable and data $X_1, \ldots, X_n$, the kernel estimator

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x|X_i, \lambda),$$

becomes $$\sum_{j=1}^{k} r_j K(x|j, \lambda) \qquad (3.4)$$

where r_j is the relative frequency of the jth cell and $K(x|X_i, \lambda) = K(x|j, \lambda)$ if $X_i = j$. (3.4) is therefore seen to be a weighted average of relative frequencies.

The Aitchison and Aitken kernel is given by

$$K(x|X_i, \lambda) = \lambda^{1-|x-X_i|} (1-\lambda)^{|x-X_i|} \qquad (3.5)$$

$$\text{so that } K(x|X_i, \lambda) = \left.\begin{array}{ll} \lambda, & X_i = x \\ 1-\lambda, & X_i \neq x \end{array}\right\},$$

and where the usual monotonicity requirement is satisfied by the condition $\frac{1}{2} \leq \lambda \leq 1$. No smoothing corresponds to $\lambda = 1$, (c.f. $h = 0$ in the continuous case), yielding the usual MLE, while $\lambda = \frac{1}{2}$

gives the uniform distribution and corresponds to infinite smoothing. $K(x|X, \lambda)$ is seen to define a discrete probability function. The d-dimensional equivalent of (3.5), employing a common smoothing parameter λ, is given by Aitchison and Aitken (1976) as

$$K(\underline{x}|\underline{X}_i, \lambda) = \lambda^{d-d(\underline{X}_i,\underline{x})} (1-\lambda)^{d(\underline{X}_i,\underline{x})}, \qquad \tfrac{1}{k} \leqslant \lambda \leqslant 1 \tag{3.6}$$

$$(= \lambda^{d} \alpha^{d(\underline{X}_i,\underline{x})}, \quad \text{where } \alpha = (1-\lambda)/\lambda, \ 0 \leqslant \alpha \leqslant 1)$$

where $d(\underline{X},\underline{x})$ is the Euclidean distance $(\underline{X}-\underline{x})^T(\underline{X}-\underline{x})$ which corresponds to the number of variables on which $\underline{X}$ and $\underline{x}$ differ. $K(\underline{x}|\underline{X}, \lambda)$ in the form (3.6) provides a discrete counterpart of the commonly used spherical normal kernel,

$$K\left[\frac{\underline{x}-\underline{X}}{h}\right] = (2\pi h)^{-d/2} \exp\left\{-\frac{1}{2h}(\underline{x}-\underline{X})^T(\underline{x}-\underline{X})\right\}$$

and is seen to be equivalent to the product kernel in d-dimensions,

$$\prod_{j=1}^{d} \lambda^{1-|x_j-(\underline{X})_j|} (1-\lambda)^{|x_j-(\underline{X})_j|},$$

the "cubical binomial" density or product density function of d independent Binomial trials. As standardisation is not possible in the same way as for continuous data (though see the single parameter kernel estimator of Titterington et al. (1981) described in Section 4.2), a different kernel K_j may be required in each dimension $j = 1,\ldots, d$, corresponding to distinct λ_j, whence

$$K(\underline{x}|\underline{X}, \underline{\lambda}) = \prod_{j=1}^{d} \lambda_j^{1-|x_j-(\underline{X})_j|} (1-\lambda)^{|x_j-(\underline{X})_j|} \qquad \text{(Aitken, 1978)} \tag{3.7}$$

Mixed kernels of the form (1.7) are also possible.

For the general nominal variable with k_j categories Aitchison and Aitken (1976) suggested

$$K(x|X, \lambda_j) = \left.\begin{array}{ll} \lambda_j, & X = x \\ \dfrac{1-\lambda_j}{k_j-1}, & X \neq x \end{array}\right\}, \quad 1/k_j \leqslant \lambda_j \leqslant 1, \tag{3.8}$$

for monotonicity, which, for $k_j = 2$, reduces to the binary kernel.

Titterington (1980) demonstrated that for a single k-category variable Aitchison and Aitken's kernel can be written in the form of the convex relative frequency estimator (3.3) with $\alpha = k(1-\lambda)/(k-1)$. Here $\phi(x) = 1/k$, $\forall$ x and hence we are smoothing towards uniformity.

For an ordered kernel, the weighting given to observation X_i should reflect the distance $|x-X_i|$. Aitchison and Aitken (1976) gave one such kernel for the trinomial case, defining $K(x|X, \lambda)$ as follows :

X \ x	1	2	3
1	λ^2	$2\lambda(1-\lambda)$	$(1-\lambda)^2$
2	$\frac{1}{2}(1-\lambda^2)$	λ^2	$\frac{1}{2}(1-\lambda^2)$
3	$(1-\lambda)^2$	$2\lambda(1-\lambda)$	λ^2

so that $\forall$ X, $\sum_{x=1}^{3} K(x|X, \lambda) = 1$. This is only sensible for $\lambda > 2/3$.

Alternatively, Titterington (1980) suggested one based on a linear decline of $K(x|X, \lambda)$ with increasing distance $|X-x|$. For example, for a 4-cell Multinomial, $K(x|X, \lambda)$ takes the form:

X \ x	1	2	3	4
1	λ	$3/6(1-\lambda)$	$2/6(1-\lambda)$	$1/6(1-\lambda)$
2	$3/9(1-\lambda)$	λ	$4/9(1-\lambda)$	$2/9(1-\lambda)$
3	$2/9(1-\lambda)$	$4/9(1-\lambda)$	λ	$3/9(1-\lambda)$
4	$1/6(1-\lambda)$	$2/6(1-\lambda)$	$3/6(1-\lambda)$	λ

For a k-cell variable, write

$\hat{p}(x) = \sum_{j=1}^{k} r_j K(x|j, \lambda)$ as $\hat{\underline{p}}(\underline{x}) = C^T \underline{r}$ where $\underline{r} = (r_1, r_2, \ldots, r_k)$ is the vector of relative frequencies and C is a $k \times k$ array such that $C_{ij} \geq 0 \ \forall i,j$ and each row sums to 1. Letting $C = I+(1-\lambda)G$ where

$$\left.\begin{array}{l} G_{ii} = -1, \ i = 1, \ldots, k \text{ and} \\ G_{ij} \geq 0, \ i \neq j, \text{ such that } G\underline{1} = 0 \text{ where } \underline{1}^T = (1, \ldots, 1) \end{array}\right\} \qquad (3.9)$$

provides the Aitchison and Aitken kernel (3.8) if $G_{ij} = (k-1)^{-1}$, $i \neq j$, and the general form of Titterington's (1980) ordered kernel is given by

$$G_{ij} = \left.\begin{array}{ll} \dfrac{2j}{(k-1)i}, & j < i \\ \dfrac{2(k+1-j)}{(k-1)(k+1-i)}, & j > i \end{array}\right\}. \qquad (3.10)$$

For $k = 2$, this again reduces to the binary Aitchison and Aitken kernel. For multiple variables, ordered or not, the product kernel is given in the bivariate case by

$\hat{P}(\underline{x}) = C_1^T R C_2$ where $C_i \in M^{k_i x k_i}$ corresponds to the ith variable with k_i categories and $\hat{P}$ and $R \in M^{k_1 x k_2}$ are now arrays of estimated probabilities and relative frequencies. In general we have

$$\hat{p}_{i\ldots l} = \sum_{j=1}^{k_1} \sum_{m=1}^{k_d} (C_1^T)_{ij} \ldots (C_d^T)_{lm} r_{j\ldots m}.$$

A more general prescription still is that of Habbema, Hermans and Remme (1978) who proposed

$K(x|X_i, \lambda_j) \propto \lambda_j^{D^2(x, X_i)}$ where $D^2(\cdot, \cdot)$ is a suitable distance measure.

Each of these kernels can be adapted to deal with missing data. There are 2 possibilities; either "missing" may be treated as an extra category, destroying any ordinal nature (unless $k = 2$ in which case "missing" is considered to lie between "absent" and "present"), as in Titterington (1977), or a new kernel form developed. Murray and Titterington (1978) provide such a kernel which may be used with either ordered or nominal variables and the appropriate form of basic kernel such as those described above.

Titterington (1980) provides another similar one.

Aitken (1978) compared performance of Aitchison and Aitken's (1976) kernels with both a single and separate λ_j (chosen marginally), with that of the generalised Hills' estimator (see below) and 2 parametric models, namely the independent binary model and a (predictive) logistic model, using Anderson et al.'s (1972) Keratoconjunctivitis Sicca data with 10 binary indicants and 2 outcomes and a second data set with 7 and 5 binary symptoms and 2 outcomes also. Comparing methods in terms of leaving-one-out error rate, log-likelihood and the number of classifications when a grey area was introduced, the conclusion was that overall no one method was better or worse than any other. In particular the performance of the kernel (with common λ) and nearest-neighbour methods were identical, though the latter was much more computationally time consuming, especially as the number of variables increased, since various values of t (see below) were used. The kernel method did best on grounds of doubt, and well on log-likelihood where the multivariate kernel improved slightly, and has only one parameter to estimate. The multivariate kernel could perform very poorly with respect to error rate.

Hills' estimator

Although Aitchison and Aitken type kernels are the most commonly used, probably on grounds of simplicity and consistency properties, the earliest discrete kernel estimator was provided by Hills (1967) who used the kernel

$$K(\underline{x} \mid \underline{X}_i) = \left.\begin{array}{ll} 1 \ , & d(\underline{x} \ , \ \underline{X}_i \ ) \leqslant t \\ 0 \ , & \text{otherwise} \end{array}\right\} , \qquad (3.11)$$

where $d(\cdot \ , \ \cdot)$ is a measure of distance and t is a threshhold distance beyond which observations $\underline{X}_i$ cease to contribute to the estimate at $\underline{x}$. Hills termed this a near-neighbour estimator but in fact it is a discrete analogy of Rosenblatt's rectangular kernel for continuous data (2.1) and suffers from the same lack of smoothness properties. If t is chosen to be less than the smallest inter-cell distance the MLE results so that zero cell estimates are still possible, with consequent inability to assign test cases which fall into such cells. Also consistency is only guaranteed for $t = 0$ i.e. the MLE, noted by Aitchison and Aitken (1976) who

provided a generalised version of Hills' near-neighbour estimator based on their own kernel, namely

$$K(\underline{x}|\ \underline{X},\ \lambda) = \left.\begin{array}{ll} \dfrac{\lambda^{d-d(\underline{X},\ \underline{x})}\ (1-\lambda)^{d(\underline{X},\ \underline{x})}}{B(d,\ t,\ \lambda)}, & d(\underline{X},\ \underline{x}) \leqslant t \\ 0\ , \text{ otherwise} & \end{array}\right\}$$

with $\frac{1}{2} \leqslant \lambda \leqslant 1$ and $B(d,\ t,\ \lambda)$, a normalising factor, =

$$\sum_{j=0}^{t} \begin{bmatrix} d \\ j \end{bmatrix} \lambda^{d-j}\ (1-\lambda)^{j}\ ,$$

which reduces to Hills' estimator for $\lambda = \frac{1}{2}$, and for $t = d$ becomes the usual Aitchison and Aitken kernel (Aitken, 1978). For $\lambda > \frac{1}{2}$ this assigns a smoother sequence of weights than Hills' method, more distant neighbours receiving lower weights. To choose λ, the same criterion would be used as for the usual kernel. Again, t, the order of nearest-neighbour, must be chosen to avoid occurrence of points for which there is no near-neighbour. Aitchison and Aitken (1976) found it necessary to go to $t = 3$ in an application of the method to an example involving 10 binary variables and 77 training cases of 2 types. With a suitable choice of λ, the kernel method, however, will always assign non-zero probability to every point in the sample space, and the extent to which neighbouring points contribute is determined by the data through the choice of λ.

Wang and van Ryzin kernels

While the multivariate kernels of Aitchison and Aitken are of product kernel form, more genuinely multivariate kernel estimators are provided by the weighted relative frequency estimators of Wang and van Ryzin (1981) and Hall (1981b). For univariate data Wang and van Ryzin provide the discrete analogue of the continuous Parzen type estimators and hence generalise the method of Hills (1967). As above, but with $\gamma = 1-\lambda$, the estimator is

$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x|X_i,\ \gamma)$. $K(x|X,\ \gamma)$ is a probability function

defined on the integers {0, ±1, ±2,,..} and subject to the additional constraints

$$\left.\begin{array}{l} K(x|X,\ 0) = 0\ ,\ X \neq x \\ \qquad\qquad 1\ ,\ X = x \end{array}\right\} \quad \text{corresponding to delta functions for}$$

continuous variables as $h \to 0$, so that with no smoothing the multinomial estimator results. γ belongs to an interval R' of the real line containing the origin. If $pr(\hat{\gamma} \in R') = 1$ and $\hat{\gamma}_n \to 0$ in probability (with probability 1) the estimator is (strongly) consistent and in addition asymptotically normal as $n \to \infty$ if $\hat{\gamma}_n \sqrt{n} \to 0$ in probability, as is also true if the MLE (as a consistent estimator) is substituted in any expression obtained for $\hat{\gamma}_n$ (Wang and van Ryzin, 1981). They considered the Uniform weight function,

$$\left.\begin{array}{lll} K(x|X,\ \gamma) = \gamma\ /\ 2t\ , & |X-x| = 1,..,\ t \\ \qquad 1-\gamma\ , & X = x \\ \qquad\quad 0\ , & |X-x| > t \end{array}\right\}\ , \qquad (3.12)$$

t a fixed constant, $t \geq 1$, and $\gamma \in (0,\ 1)$, similar to the Aitchison and Aitken kernel, and a geometric function

$$\left.\begin{array}{l} K(x|X,\ \gamma) = \tfrac{1}{2}\ (1-\gamma)\ \gamma^{|X-x|}\ ,\ |X-x| \geq 1 \\ \qquad\qquad\qquad\qquad 1-\gamma\ ,\ X = x \end{array}\right\}\ ,\ \gamma \in (0,\ 1), \qquad (3.13)$$

which has a more gradual decline. The former was recommended with t = 1 or 2 on the grounds of its simplicity and the fact that it leads to an exact optimal $\hat{\gamma}$ (see below). Where the probability function to be estimated is supported on the non-negative integers, Wang and van Ryzin suggest adding the weights which would be assigned to X, X < 0, to that given to cell X = x.

Hall's relative frequency estimator

While the conclusion of Wang and van Ryzin (1981) was that the form of the kernel is not vitally important, Hall (1981b) provided an alternative relative frequency estimator for multivariate binary data, choosing the linear combination of frequencies to minimise directly the discrete analogue of MISE (c.f. the (asymptotically)

optimal kernel of Epanechnikov, 1969)

$E\{\sum_{\underline{x}} (\hat{p}(\underline{x}) - p(\underline{x}))^2\}$. For the nearest-neighbour type estimator,

$\hat{p}(\underline{x}) = \frac{1}{n} \sum_{j=0}^{t} w_j N_j(\underline{x})$ where t is a threshhold distance, $t \leq d-1$,

and $N_j(\underline{x})$ is the no. of observations at distance j from $\underline{x}$, i.e.

$$N_j(\underline{x}) = \sum_{\{\underline{X}_i : (\underline{X}_i-\underline{x})^T(\underline{X}_i-\underline{x}) = j\}} N_0(\underline{X}_i),$$

Hall shows that the optimal choice of $\underline{w}$ is given by

$\underline{w} = \{ P + n^{-1}(D-P) \}^{-1} \underline{p}$ where $\underline{w} = (w_0, w_1, \ldots, w_t)^T$,

$$D = \text{diag}\left\{ \begin{bmatrix} d \\ 0 \end{bmatrix}, \ldots\ldots, \begin{bmatrix} d \\ t \end{bmatrix} \right\}, \quad \underline{p} = \sum_{\underline{x}} p_0(\underline{x})\, \underline{p}(\underline{x}), \quad \text{where}$$

$$\underline{p}(\underline{x}) = (p_0(\underline{x}), p_1(\underline{x}), \ldots, p_t(\underline{x}))^T, \quad P = \sum_{\underline{x}} \underline{p}(\underline{x})\, \underline{p}(\underline{x})^T,$$

$$p_j(\underline{x}) = \sum_{\{\underline{X}: (\underline{x}-\underline{X})^T(\underline{x}-\underline{X}) = j\}} p_0(\underline{X}), \quad j = 1, \ldots, t,$$

and $p_0(\underline{X})$ is the probability of observing feature vector $\underline{X}$. Rewriting $\underline{w}$ and substituting relative frequencies for the $\{N_j(\underline{x})\}$ to give MLE $\hat{P}$, gives

$$\hat{\underline{w}} = \{(1-n^{-1})\, I_{t+1} + n^{-1}\, \hat{P}^{-1}\, D\}^{-1}\, \underline{i}$$

where $\underline{i} = (1, 0, \ldots, 0)^T \in R^{t+1}$ and P is non-singular. A modified version $\underline{w}_1$ is given which will ensure that the resulting estimates will sum to 1, namely,

$$\underline{w}_1 = A_n^{-2} \{ \underline{p} + (1-\underline{h}^T A_n^{-2}\, \underline{p})\, \underline{h} \,/\, (\underline{h}^T A_n^{-2}\, \underline{h}) \}$$

or

$$\underline{w}_1 = \underline{w} + B_n^{-1}\, \underline{1}_{t+1}\, (1-\underline{h}^T B_n^{-1}\, \underline{i}) \,/\, (\underline{h}^T B_n^{-1}\, \underline{1}_{t+1}), \quad \text{where}$$

$$\underline{h} = \left\{ \begin{bmatrix} d \\ 0 \end{bmatrix}, \begin{bmatrix} d \\ 1 \end{bmatrix}, \ldots\ldots, \begin{bmatrix} d \\ t \end{bmatrix} \right\}^T, \quad A_n^{-2} \equiv P + n^{-1}\,(D-P), \quad \text{and}$$

$$B_n \equiv (1-n^{-1})\, I_{t+1} + n^{-1}\, P^{-1}\, D.$$

Probability estimates using either $\hat{\underline{w}}$ or $\hat{\underline{w}}_1$ may be negative and are slightly biased. Improved estimates of the weights may be obtained by iteration using $\hat{\underline{w}}$ or $\hat{\underline{w}}_1$ as a startpoint and re-estimating $\underline{p}(\underline{x})$. For t = 1, 2, and 3, Hall compared his

estimators with those of Hills (1967) (3.11) and Aitchison and Aitken (1976) (3.6), which are also linear combinations of the $\{N_j(\underline{x})\}$, on simulated data (Hall, 1981a). With 2 disease classes and 4 binary variables, $\hat{\underline{w}}$ and $\hat{\underline{w}}_1$ gave very similar predicted probabilities, which for each t were similar to those using Aitchison and Aitken's kernel with the smoothing parameter of Hall (1981a), especially for t = 1. Hills' method, for each t, produced quite different probabilities from either method. Hall's method, for given t, is simple to calculate, and, being based on a measure of distance, extends immediately to categorical and multiple variables.

In conclusion, while the above are all weighted combinations of relative frequencies, the method of Aitchison and Aitken (1976) is the most general since it extends to mixed and ordered variables as well as multivariate categorical ones, via the product kernel form. Those of Hills and Hall would require more generalised distance functions to cope with mixed data. The latter method, though quick to compute, can lead to negative estimates while the former can be time consuming, especially in multiple dimensions, its consistency is not guaranteed and, unlike the kernel method (for suitable choice of λ) can lead to zero cell estimates. Like both Hills' and Hall's methods, in principle Wang and van Ryzin's estimator is potentially more genuinely multivariate than that of Aitchison and Aitken.

Choice of smoothing parameter λ

Most of the literature makes use of product rather than genuinely multivariate kernels and assumes a common form of kernel for each variable. We may either assume a common smoothing parameter a priori, or allow separate degrees of smoothing, possibly combining these (for instance by averaging) if a single parameter is required. Secondly, if the $\{\lambda_j\}$ are not constrained to be equal these may either be estimated on a marginal basis or simultaneously estimated to optimise a global multivariate criterion. One would expect a multivariate density estimate to require more smoothing than marginal ones, and the study of Titterington (1980) found that the latter approach does produce smaller $\{\lambda_j\}$.

As with continuous kernels there are numerous methods by which to choose $\underline{\lambda}$.

Maximum likelihood

Aitchison and Aitken (1976) showed that maximisation of the likelihood

$\prod_{j=1}^{n} \hat{p}(\underline{X}_j)$ yields $\lambda_j = 1$, $\forall j$, indicating no smoothing, as in the continuous case. Substitution of $\underline{\lambda} = \underline{1}$ into the discrete kernel estimator (3.7) gives the usual MLE $\hat{\underline{p}}(\underline{x}) = n(\underline{x})/n$ and the usual problems associated with the multinomial estimator arise unless $n/\prod k_i$ is sufficiently large, and no empty cells occur - often not the case.

Methods based on cross-validation

Given this difficulty Aitchison and Aitken (1976) adopted the methodology which Habbema, Hermans and van den Broek (1974) used for continuous variables, maximising instead the modified likelihood

$\prod_{i=1}^{n} \hat{p}_{(i)}(\underline{X}_i)$ where $\hat{p}_{(i)}(\underline{x})$ is the kernel estimate obtained by omitting $\underline{X}_i$ from the design set, that is

$$\hat{p}_{(i)}(\underline{x}) = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} K(\underline{x} | \underline{X}_j, \underline{\lambda}).$$

Analagous to the procedure for continuous variables, the cross-validatory choice of $\underline{\lambda}$ for a given loss function $L(\delta(\underline{x}), p(\underline{x}))$ which measures the distance between the vectors $\delta(\underline{x})$ and $p(\underline{x})$, with the multinomial indicator

$$\left.\begin{array}{l} \delta_i(\underline{x}) = 1, \ \underline{x} = \underline{X}_i \\ \qquad\quad 0, \text{ otherwise} \end{array}\right\},$$

minimises $\frac{1}{n} \sum_{i=1}^{n} L(\delta_i(\underline{x}), p_i(\underline{x}))$. (Bowman, 1980).

The discrete Kullback-Leibler distance

$$L(\underline{\delta}_i, \underline{p}_i) = \sum_{\underline{x}} \delta_i(\underline{x}) \log \frac{\delta_i(\underline{x})}{p_i(\underline{x})} = c - \log p_i(\underline{X}_i), \ c \text{ a constant, can}$$

again be shown to yield the modified maximum likelihood estimate (MMLE) of λ_j. Similarly, the discrete analogue of ISE, the Summed Square Error, $L(\underline{\delta}_i, \underline{p}_i) = \sum_{\underline{x}} (\delta_i(\underline{x}) - p_i(\underline{x}))^2$ (3.14) again yields a tractable criterion, $\underline{\lambda}$ being chosen to minimise

$$\sum_{\underline{x}} \sum_{i=1}^{n} p_i(\underline{X}_i)^2 - 2 \sum_{i=1}^{n} p_i(\underline{X}_i),$$

as does the weighted version $L(\underline{\delta}_i, \underline{p}_i) = \sum_{\underline{x}} [\delta_i(\underline{x}) - p_i(\underline{x})]^2 / p_i(\underline{x})$ which emphasises good estimation of the cell probabilities of less common outcomes. In the context of smoothed relative frequency estimators, Stone (1974b) chose $\hat{\alpha}$ by cross-validation, to minimise both quadratic loss (3.14) and modulus loss functions $\Sigma_j |\delta_j - p_j|$, noting that $\hat{\alpha}$ is not robust to the choice of loss function. He compared the use of these, in terms of cross-validated quadratic loss (Stone, 1974a, 1974b) with that of Fienberg and Holland (1973), (see below), one of Good (1965), and the MLE on 6 examples, finding that none was uniformly superior. Use of modulus loss results in an estimator that can still have some zero estimates (Hand, 1982, p. 145).

For a single binary variable the modified maximum likelihood estimator of λ can be found explicitly, giving $\hat{\lambda} = \max\{\frac{1}{2}, -(r_1^2 a_1 + r_2^2 a_2)n/a_1 a_2\}$ (Titterington, 1980), where r_1 and r_2 are the relative frequencies of each outcome and $a_1 = n(r_1 - r_2) - 1$ and $a_2 = n(r_2 - r_1) - 1$. However, full multivariate optimisation requires numerical methods, even if a single smoothing parameter is assumed.

Using common λ and binary variables Aitchison and Aitken (1976) showed that $\hat{\lambda} \xrightarrow{p} 1$ as $n \to \infty$ and hence $\hat{p}(\underline{x}) \xrightarrow{p} p(\underline{x})$. Bowman (1980) established pointwise consistency for the same estimator.

Hall (1981a) noted that if several cells are empty and all other cells contain more than 1 observation that the modified llikelihood can have a local maximum at $\lambda = 1$.

Hall's method

In view of this potential difficulty Hall (1981a) suggested in such situations minimising the expected sum of squares (c.f. MISE

for continuous variables),

$E\{\sum_{\underline{x}} (p(\underline{x}) - \hat{p}(\underline{x}))^2\}$ or weighted mean sum of squares,

$\sum_{\underline{x}} p(\underline{x})\, E\{(p(\underline{x}) - \hat{p}(\underline{x}))^2\}$. Based on a Taylor expansion of $\hat{p}(\underline{x})$ in terms of powers of $(1-\lambda)$, and using the Aitchison and Aitken kernel with common λ,

$$\hat{\lambda}(\underline{x}) = 1 - p(\underline{x})\{d\{1 - p(\underline{x}) + p_1(\underline{x})\} / n\{p_1(\underline{x}) - dp(\underline{x})\}^2$$

was found to minimise the asymptotic mean square error, where $p_i(\underline{x})$ now denotes the probability of falling into a cell at distance i away from $\underline{x}$. This is location dependent and would yield a discrete analogue of the variable estimator, but suffers from the problem that if sample estimates are substituted for $p(\underline{x})$ and $p_1(\underline{x})$ and $\underline{x}$ is empty then $\hat{\lambda}(\underline{x}) = 1$.

Asymptotic minimisation of the 2 suggested global criteria yields overall estimators

$$\hat{\lambda} = 1 - [d + \sum_{\underline{x}} p(\underline{x})\{p_1(\underline{x}) - dp(\underline{x})\}] / n\sum_{\underline{x}} (p_1(\underline{x}) - dp(\underline{x}))^2$$

and

$$\hat{\lambda} = 1 - [\sum_{\underline{x}} p(\underline{x})^2\{d(1 - p(\underline{x})) + p_1(\underline{x})\}] / n\sum_{\underline{x}} p(\underline{x})\, (p_1(\underline{x}) - dp(\underline{x}))^2$$

respectively.

Again $p(\underline{x})$ and $p_1(\underline{x})$ are replaced by their MLEs, or, alternatively an iterative procedure may be used, starting with $\hat{\lambda}_0 = 1$, say, evaluating the criterion based on the kernel estimate of $\{p(\underline{x})\}$, re-evaluating $\hat{\lambda}$ etc. until convergence is achieved.

Hall's method has the advantage over those of the previous section of not requiring numerical optimisation.

The above estimates rely on 1st order expansions of $\hat{p}(\underline{x})$. Hall (1981a) also gives the corresponding $\hat{\lambda}$ for the mean summed squared error criterion based on the 2nd order expansion. Hand (1983) compared this with the modified maximum likelihood method on 5 moderately large real data sets involving 10, 5, 12, 8 and 20 binary variables respectively. The 3rd data set was an extension of the 2nd. Each involved 3 populations except the 1st, which had 2 outcomes. Allowing separate λs for each class, but common for

each variable, it was found that MMLE consistently produced slightly smaller λ, and hence more smoothing, than Hall's method but in terms of error rate on a separate test set there was negligible difference, suggesting that the choice of method may not be critical and that simpler methods may be adequate. For the 5 variable case and assuming common λ in each class, sensitivity to the value of λ was investigated in terms of the same error rate and it was found that the range of error rates was very small, any $\lambda \in (.7, 1.0)$ giving very similar results. Both MMLE and Hall's method using separate λs produced near-optimal error rates. However using more general loss functions, appropriate where different costs of misclassification apply, the choice of loss criterion was seen to have a critical effect, not only on the size of the range of λ producing acceptable results but also on the optimal value of λ.

For data sets 1 and 4, involving more variables, and in each case using 2 classes only, a range of (λ_1, λ_2) were chosen, λ_i referring to the smoothing parameter used in class i. In contrast to the continuous variable situation, where different degrees of smoothing are to be preferred in different classes, the smallest error rates consistently occurred for $\lambda_1 = \lambda_2$, for both data sets for both apparent and test set error rates. The leaving-one-out error rate was minimised for the overall best of these. Both MMLE and Hall's 2nd order method undersmoothed slightly but with only a slight increase in error rate.

Wang and van Ryzin (1981) also minimised mean summed squared error. For the uniform kernel (3.12) an exact expression was obtained for the parameter γ while for the geometric kernel (3.13) an approximate solution was found by truncating the weight function. In either case γ is dependent on $\{p(\underline{x})\}$ and again relative frequencies are substituted, yielding consistent estimators. The asymptotic mean summed square error was shown to be strictly less than that of the MLE (except in the degenerate case).

For samples of size 10, 20, 50, 100 and 200, and 500 simulations of data from Negative Binomial and Poisson distributions, the uniform kernel with t=1 and 2 and the geometric kernel were compared with the MLE and the parametric models with estimated parameters, on grounds of mean summed squared error, and found to perform uniformly better than the MLEs though worse than the appropriate parametric model. There was little to choose

between these kernels and more complex ones. As Aitchison and Aitken (1976) also commented of their estimator, the advantage of such smoothed estimators is that their large sample properties equal those of the MLE but are better for small samples when some degree of smoothness is present, while choice of the wrong parametric model will not give a consistent estimator.

Minimum mean squared error (MMSE)

Fienberg and Holland (1973) chose the degree of smoothing to minimise risk, following the standard procedure of substituting relative frequencies in the resulting expression for α. Hand (1982, p. 149) notes that here too an iterative procedure is possible. As for MMLE full multivariate optimisation requires numerical methods, even if a single smoothing parameter is assumed, which can be time consuming. For multiple dimensions increasing sparsity of the data may mean multivariate MMSE is not possible. For a single k-cell variable an exact MMSE solution can be obtained, using the formulation $\hat{p}(\underline{x}) = C^T\underline{r}$, either substituting relative frequencies or using an iterative approach. In principle an optimal form for G may be found to minimise MSE, subject to the constraints (3.9), though not explicitly. If G is specified marginal estimation will be quick. Brown and Rundell (1985) minimised instead an unbiased estimator of mean summed square error, to avoid substitution of probability estimates in the resulting parameter estimator.

For the purposes of smoothing ordered discrete data, Titterington and Bowman (1985) compared unordered kernels with an ordered kernel of Habbema, Hermans and Remme (1978) and Wang and van Ryzin's uniform-1 kernel, with smoothing parameter chosen by MMSE and cross-validation with both quadratic and Kullback-Leibler loss, with several logistic-normal Bayesian estimates. Unusually, variation amongst means of choosing the degree of smoothing was less than that amongst the methods used. For the purposes of smoothing sparse data the uniform kernel was found to undersmooth, removing rather few zero cells. A logistic-normal approximation also undersmoothed, and the ordered kernel was superior to both these and unordered kernels. The same was true for simulated data from ordered Multinomial distributions, assessing methods in terms of average squared error relative to that of the MLE, where the Habbema kernel excelled. For unordered distributions unordered

kernels were best, but the former were close behind. In all cases the logistic-normal methods were outperformed by a kernel method.

Pseudo-Bayes methods

Titterington (1980) estimated λ by giving it a Beta prior. The posterior for λ is then a mixture of Betas to which in principle exact Bayesian methodology may be applied. To avoid heavy computation, a fractional updating method was proposed, approximating the mixture by a single Beta distribution whose parameters are updated iteratively. The posterior mean is then readily estimated from the final parameter estimates, setting it to ½ if the estimate is smaller than this. Over 20 simulations with sample sizes of 10, 30, and 50 of univariate binary observations with p = .1, .3 and .5 and 2 sets of initial parameters (β_0, γ_0) = (1, 1) and (6, 1), in terms of bias and root mean square error the approximate and exact Bayesian methods were found to give similar results but the former achieved considerable saving in time. Both were competitive with the use of MMLE. Fienberg and Holland's MMSE approach was at least as good as MMLE and a deterministic method similar. The unsmoothed MLE had the smallest bias though the other estimators were often close behind and good in terms of root mean square error.

For multivariate binary observations, if independent Beta priors are assumed for the $\{\lambda_i\}$ the posterior is also a product of Betas and in principle both exact and pseudo-Bayes methods extend, though the former quickly becomes prohibitive as n and the number of variables increase. For polychotomous variables the same applies.

Titterington (1980) compared 12 methods on the 10-dimensional data of Anderson et al. (1972), in terms of the estimates of $\{\lambda_i\}$ and ordering of test cases on the basis of estimated likelihood ratio of one population rather than the other. He found that marginal methods, while very quick, produced rather sharp estimates compared to those of the corresponding multivariate ones and that MMSE produces less smoothing than MMLE. The fractional updating methods compared well to full multivariate MMLE with separate λ_i but was much faster. The different methods produced roughly similar orderings on the basis of likelihood ratio and therefore for practical use he recommended use of a couple of marginal approaches plus a fractional updating method for speed. Treating the data as a 1024-cell Multinomial and applying smoothed relative

frequency estimators led to misclassification of 4 rather than 0 test cases, so that it appears worthwhile to make use of the multivariate structure of the data.

Summarising, we have described a number of means of automatic choice of smoothing parameter for a density estimator. As in the continuous case, within the context of discrimination these methods may be expected to perform suboptimally in terms of good likelihood ratio estimation, a more direct approach being appropriate.

3.5 ISOTONIC REGRESSION

3.5.1 The isotonic regression problem

Let X be a finite set $(x_1, x_2, \ldots, x_n)$. A binary relation $\prec$ on X establishes a "simple order" on X if $\prec$ is

1) reflexive : $x \prec x$, $x \in X$
2) transitive : $x \prec y$, $y \prec z \Rightarrow x \prec z$, $x, y, z \in X$
3) antisymmetric : $x \prec y$, $y \prec x \Rightarrow x = y$, $x, y \in X$ and
4) every two elements are comparable : $x, y \in X \Rightarrow x \prec y$ or $y \prec x$.

A "partial order" is reflexive, transitive and antisymmetric but there may be non-comparable elements. Thus every simple order is a partial order.

A real-valued function g on X is defined to be "isotonic" ("antitonic"), or order preserving, with respect to the order $\prec$ if $x \prec y$, $x, y \in X \Rightarrow g(x) \leqslant g(y)$ $(g(x) \geqslant g(y))$.

Given a positive weight function $w(\cdot)$, the (least squares) "isotonic regression" g^* with weight w of a fixed function $f(\cdot)$, f and w both defined on X, minimises

$$\sum_{i=1}^{n} (f(x_i)-g(x_i))^2 \, w(x_i)$$

over the class of isotonic functions g on X. Barlow et al. (1972), pp. 24-26, show that g^* is unique.

Solving for g^* is a constrained quadratic programming problem but due to its wide applicability (see Barlow et al., 1972) a number of specialised algorithms have been proposed.

3.5.2 Algorithms

For a univariate simple ordering or amalgamation of simple

orderings, g^* is easily found, commonly by using the Pool-Adjacent-Violators (PAV) algorithm of Ayer et al. (1955) (or see Barlow et al., 1972, pp. 13-15). This progressively pools adjacent points x_i, x_{i+1}, where the ordering is violated i.e. $f(x_i) > f(x_{i+1})$, defining a block $\{x_i, x_{i+1}\}$ with function value $f(x_i, x_{i+1}) = \{f(x_i)w(x_i) + f(x_{i+1})w(x_{i+1})\} / \{w(x_i) + w(x_{i+1})\}$ and associated weight $\{w(x_i) + w(x_{i+1})\}$. The ordering is such that $\{x_i, x_{i+1}\} \prec x_j$ if $x_i \prec x_j$ and $x_{i+1} \prec x_j$. Once all such pairs of points have been pooled, blocks are amalgamated in the same manner, if necessary, until no reversals of the ordering occur. Another version of this is the Up-and-Down-Blocks algorithm of J.B. Kruskal (Barlow et al., 1972, pp. 72-73). Cran (1980) gives a Fortran algorithm based on the latter. Both algorithms apply only to simple orderings.

For partial orders and/or more than 1 dimension, more complex algorithms are required. An example of a d-dimensional partial ordering was given in Section 3.2, with $X = \{(i,j,..,l) : (i,j..,l) \prec (p,q,..,s)$ if $i \leqslant p$, $j \leqslant q$,.... and $l \leqslant s\}$, with a finite range for each of i, j,...,l. Again such algorithms involve averaging of function values over subsets of X. Four are discussed in Section 2.3 of Barlow et al. (1972). Murray (1983) cautions that the Minimax Order algorithm of Alexander (Barlow et al., 1972, pp. 81-88), can lead to cycling. Others are given by Gebhardt (1970), Lee (1983), who notes a problem with a sub-algorithm of the former, and Dykstra (1981). The latter is efficient when f is nearly isotonic. Lee (1983) compared his Min-Max algorithm with those of Gebhardt and 2 others on 2 real data sets and 3 simulated ones, showing his to be much the most efficient. All of these are exact methods but most require considerable computation and identification of a great many subsets of X, tending to be slow. Alternatively, approximate iterative methods are available which converge to the correct solution. In principle, the following methods should be very quick.

Dykstra and Robertson (1982) give an iterative algorithm based on successive 1-dimensional smoothings for the least squares isotonic regression which is increasing in each of several variables, while Dykstra (1983) gives a generalisation of this to solve general restricted least squares regression problems. Both successively project the current solution (less the incremental change due to enforcing the same constraint previously) onto the

solution space corresponding to a given constraint. In the latter, each constraint is treated separately, whereas the former treats blocks of constraints simultaneously, using the PAV algorithm (see Section 4.4 of Dykstra (1983)). Since isotonic regression is a special case of quadratric programming subject to linear constraints, the general approach can be used and there can be advantages in doing so (see below). Section 4.2 of Dykstra (1983) gives the means of projection.

A Fortran implementation for the 2-dimensional case of Dykstra and Robertson (1982) is given by Bril et al. (1984), using the PAV algorithm to perform the univariate smoothings, though see Murray and Wilson (1987) for some corrections to this. The solution is not uniquely determined for zero-weighted cells as these may be replaced by any values satisfying the order constraints.

For the purposes of smoothing, Dykstra and Robertson (1982) replaced near zero-weights with a small positive weight, δ, suggesting the value 10^{-5}. However we have found in practice with several real data sets that this can lead to very slow convergence (Murray and Wilson, 1986). When a near-zero weighted cell is involved in a row or column smoothing its smoothed value is almost entirely determined by the values of adjacent cells, meaning that the smoothed value can be very different from the original one. In turn this implies that certain components of the matrices of row and column adjustments can become very large. This means that in practice G, the isotonic regression of the array F, is approached so slowly that it is not possible to know when the solution has been obtained to within a desired accuracy. It can be shown for the simple

$$2 \text{ by } 2 \text{ array } F = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \text{ with weights } W = \begin{bmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{bmatrix}, \quad (\delta < \tfrac{1}{2}),$$

that after 2n iterations of the algorithm (n row and n column smoothings) the estimate of G, $\hat{G}_{2n} = (1-\delta)(1-2\delta)^{2n-1} \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$.

Using $\delta = 10^{-5}$ as suggested, after 50,000 and 100,000 iterations respectively, we have

$$\hat{G}_{50,000} = \begin{bmatrix} \cdot 37 & -\cdot 37 \\ \cdot 37 & -\cdot 37 \end{bmatrix} \quad \text{and} \quad \hat{G}_{100,000} = \begin{bmatrix} \cdot 14 & -\cdot 14 \\ \cdot 14 & -\cdot 14 \end{bmatrix}$$

whereas the solution is given by

$\hat{G} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ (obtained, for instance, by using the algorithm of Dykstra, 1981). Thus, after 100,000 iterations the results are not even correct to 1 significant figure.

In practice we have overcome the problem of slow convergence by setting the F value of zero-weighted cells to the overall weighted mean of the data array and by taking δ to be a substantial fraction (e.g. 0.5) of the smallest non-zero cell weight. Murray and Wilson (1986) provide the necessary Fortran code to implement this suggestion within the algorithm of Bril et al. (1984). While this moves the solution away from the isotonic regression, it introduces extra smoothing in areas of the array where the reponse is poorly determined. Section 4.4.2 reports an example in which this was found to be beneficial.

In general, starting with row rather than column smoothing, or vice-versa, can also substantially affect the speed of convergence, as also noted by Dykstra and Robertson (1982), although this was not in itself sufficient to overcome the zero-weighted cell problem.

Wollan and Dykstra (1987) provide a Fortran implementation of Dykstra (1983), Section 4.2, to solve the problem

$$\text{minimise} \quad (\underline{f}-\underline{g})^T S^{-1} (\underline{f}-\underline{g}) \qquad (3.15)$$
$$\underline{g} : A\underline{g} \leq \underline{b}$$

where $\underline{f}$ and $\underline{b}$ are specified vectors of length n and k respectively, $S \in M^{n \times n}$ is a given positive definite matrix and

$A \in M^{k \times n} = \begin{bmatrix} \underline{a}(1)^T \\ \vdots \\ \underline{a}(k)^T \end{bmatrix}$ corresponds to the k linear inequality constraints. The isotonic regression problem is seen to be a special case of (3.15), with $S^{-1} = \text{diag}\ (w_1^{-1}, w_2^{-1}, \ldots\ldots, w_n^{-1})$, where $(w_1, w_2, \ldots\ldots, w_n)$ is the vector of weights, and $\underline{b} = \underline{0}$, arranging the arrays F and W into vector form for the multivariate case. Immediately it is seen that any zero weights must again be

replaced by a positive weight δ.

Not surprisingly, applying the algorithm to the 2 by 2 example above, again with $\delta = 10^{-5}$, caused exactly the same problems, the same formula for the current solution applying after n applications of each constraint. However, Wollan and Dykstra (1987) warn that unequal weights may cause slow convergence, and state that the addition of redundant constraints can speed up convergence. With the only possible extra constraint, $g_{11} \leqslant g_{22}$, in the above example, convergence to the correct solution was in fact achieved in a very few iterations. This is encouraging since the more general algorithm also has the advantage of being able to find the closest convex/concave function (in terms of least squares) to $\underline{f}$ i.e. the closest function which is isotonic on one side of an unspecified turning point and antitonic on the other. Section 4.1 of Dykstra (1983) relates to concave restrictions. Other than in 1 dimension, this is non-trivial using the isotonic regression algorithm.

Finally, we note that the isotonic regression is defined only at the points $x_i \in X$. For intermediate values it can only be given within limits imposed by the ordering constraints. It is characterised, by means of the pooling procedures used to derive it, as a rough step function or surface, taking a uniform value within a "solution block" or "level set". We return to this point in Section 4.4.

CHAPTER 4 ORDERING-A COMPARATIVE STUDY

4.1 THE DATA AND BACKGROUND

The data arise from an ongoing prospective study of patients suffering severe head injury ("severe" being defined as being in post-injury coma for at least 6 hours), initiated at the Southern General Hospital, Glasgow, but currently involving several centres internationally. The "Head Injury Study" was set up to investigate the feasibility of predicting eventual recovery status on the basis of data collected shortly after injury, both in order to enable concentration of limited and expensive resources on those patients most likely to "recover", since almost half of such patients will die even with intensive care, and provide an objective means by which to carry out clinical trials of new management regimes. In practice the response variable used is recovery status at 6 months after injury, by which time the condition has usually stabilised (Jennett et al., 1976). A fuller description of the data, prediction problem and references may be found in Titterington et al. (1981) who used the data in an extensive comparative study of discrimination techniques. This is described in Section 4.2 below.

The "Glasgow Outcome Scale" (Jennett and Bond, 1975) defines recovery status as one of 5 outcome categories, namely "death", "persistent vegetative state", "severe disability", "moderate disability", or "good recovery", so that outcome is seen to be an ordered response variable. The important predictor variables are age and various measures of brain damage (as evidenced by degree of brain dysfunction), such as eye, motor and verbal response to stimulation. These 3 scores collectively constitute the "Glasgow Coma Scale" (Teasdale and Jennett, 1974). Their sum, ranging from 3 (no response) to 15 (normal), is referred to as the "Coma Sum" or "Coma Score" and is a particularly powerful single predictor (Teasdale et al., 1979). All such predictor variables are either binary or ordered discrete variables, apart from age although in the 1981 study age was grouped into decades and also treated as an ordered categorical variable. The data used here consist of observations on the same 1000 patients (collected between 1968 and 1976) as were used in the existing study. These were split randomly into a training and test set, each of 500 cases and the sample relative frequencies of each outcome used as plug-in estimates of the prior probabilities.

Like the kernel estimator (see Section 3.4.2) the isotonic method may be used with incomplete data, by treating "missing" as an extra category but constraining the isotonic estimate in those cells to lie between those corresponding to the lowest and highest values of the variable which is missing. In practice, given existing problems with isotonic regression algorithms (see Section 3.5.2), this may prove difficult and to facilitate model comparisons only complete cases were used.

4.2 AN EXISTING STUDY

The paper of Titterington et al. (1981) describes an extensive study to compare numerous discrimination techniques applied to the large and complex data set described above. The complexity is due to multidimensionality, substantial missing data, and a mixture of variable types, namely ordered categorical variables (which may be regarded as either nominal categorical or continuous), binary variables and a continuous variable.

Three ordered categories of outcome were used, "dead/vegetative", "severely disabled", and "moderately disabled/good recovery", although none of the models used took account of this ordering. Four subsets of variables (comprising 4, 4, 6, and 10 variables repectively) displaying differing degrees of dependence, and with different proportions of missing data, were chosen (on grounds of prior knowledge) on which to compare the methods. Set 1 comprised Age, EMV Score (i.e. Coma Sum) and 2 others - weakly dependent and with appreciable missing data. Set 2 consisted of the highly dependent variables Age and the raw Eye, Motor, and Verbal Scores, on which there was little missing data. Set 3 was an extension of Set 1 and Set 4 resulted from Set 3 by breaking down the "created indicants" such as EMV Score.

Three discrete parametric models were considered, based on assumptions of independent variables but allowing for a single overall association factor (equal to 1 for complete independence), Lancaster models allowing for 1st order interactions, and latent class (i.e. mixture) models assuming the underlying models to have the independence form (see Section 1.4.2). All of these coped straightforwardly with missing data but none took account of ordinality. Discrete kernel methods were used, both treating "missing" as an extra category and hence losing the ordered nature of the variables, and using the missing data kernel suggested by

Murray and Titterington (1978) which adapts a basic kernel (either ordered or nominal). Product kernels were used with Aitchison and Aitken's (1976) kernel (3.8) for general categorical variables and Titterington's (1980) kernel (3.10) for ordered categorisations. Two means of choosing the smoothing parameters were employed, namely marginal minimisation of mean squared error, and the multivariate pseudo-Bayesian fractional updating technique of Titterington (1980). Further, for the case where "missing" was treated as an extra category, a single smoothing parameter, u, $0 < u < 1$, was chosen in order to optimise assessment criteria and the degree of smoothing for a k-category variable controlled by λ_k where λ_k was set to $1-(k-1)u/k$. This compares to the method for continuous variables whereby each variable is pre-standardised and a single parameter h used, the effective smoothing in the ith dimension then being proportional to the ith standard deviation. Finally, the variables were treated as continuous, and the linear and quadratic discriminant rules applied, with the usual assumptions of normality. Missing data were treated both by proper maximum likelihood estimation and the cruder method of substituting sample means. The latter was also used in the linear logistic model, again as if the data were continuous. The methods were compared in terms of error rate, average logarithmic and Brier scores, and an average loss measure as well as 2 measures of reliability, one based on comparison of predicted and actual degree of recovery at 6 months after injury and also, for discrete parametric models, one given by Hilden et al. (1978a).

Overall, the independence model was the best, proving surprisingly robust to variable interdependence, with a moderate association factor being optimal. It was only bettered, in data set 2, by a Lancaster model and has the advantage of ease in dealing with incomplete data. Latent class models were generally poor, only doing well for data set 4. Of the continuous models, the results of the linear discriminant function (with maximum likelihood estimation of missing values) were comparable to those of the discrete parametric models, but the former provides a single method which did well for each variable set, whereas choice of an appropriate discrete model could be critical. It also has the advantage of recognising ordered categories, as does the linear logistic model, which despite its cruder treatment of missing data, was comparable to the LDF for all except data set 4. The QDF was

consistently poor.

The kernel methods had consistently disappointing log and Brier scores and only even approached other models for variable set 2 (which had little missing data). This was thought to be due to the high dimensionality of the data and a discrete kernel being unable to cope well with such sparse data. Neither the marginal MMSE nor the multivariate pseudo-Bayes methods provided sufficient smoothing in this multivariate problem, resulting in very low density estimates and unreliable predicted probabilities. The single parameter method was much the best of the kernel methods despite taking no account of ordering. Of the methods which used Murray and Titterington's (1978) missing data kernel, those using an ordered kernel were better for all variable sets than those which did not, and overall they outperformed the methods using an extra category for "missing" with the marginal, and pseudo-Bayes choices of smoothing parameter. This suggests that the single parameter method might more closely approach the better parametric models if used with an ordered kernel, although Titterington et al. (1981) noted that if we can assume normality for the LDF then continuous or mixed kernel models rather than discrete ones might also prove useful in this context.

A general conclusion of the study was that variation in performance among the methods tended to be smaller than that amongst variable sets and therefore, given sufficient informed prior knowledge to construct sensible "created indicants", such as the Coma Sum, a simple model such as the independence model will do just as well if not better than either a more complex parametric model or the robust LDF used in an uninformed manner.

4.3 A COMPARATIVE STUDY

As in Section 2.6 we treat the 2 population case, collapsing outcome to "dead/vegetative" (π_1), and "severely disabled/moderately disabled/good recovery" (π_2). Age was grouped into 15 5-year categories allowing treatment as an ordinal variable.

4.3.1 A univariate example

Age was chosen as a single predictor, both to allow comparison with the results of Section 2.6, which used it as a continuous variable, and since it is the single most important indicator of

recovery at or after 24 hours post-injury (prior to that Coma Sum is known to be more informative). Table 4.1 and the histograms in Figures 4.1 and 4.2 display the data for both the training and test samples.

While the likelihood of death after severe head injury increases with age (Figures 4.1(c) and 4.2(c)), it will be noted that the conditional distribution of π_1 given age is convex rather than strictly isotonic, due to the different response to injury of the brain of a very young child. Modification of the isotonic regression algorithm to allow for this is straightforward in 1 dimension. However, as we noted in Section 3.5.2, with multiple predictors this is awkward and it is then simpler to collapse the categories to induce isotonicity.

The ordered discrete kernel (3.10) of Titterington (1980), and the isotonic method (denoted ISO) of Section 3.5.2 are used to estimate $p(\pi_1|x)$, using both marginal and direct means to choose the smoothing parameters, namely

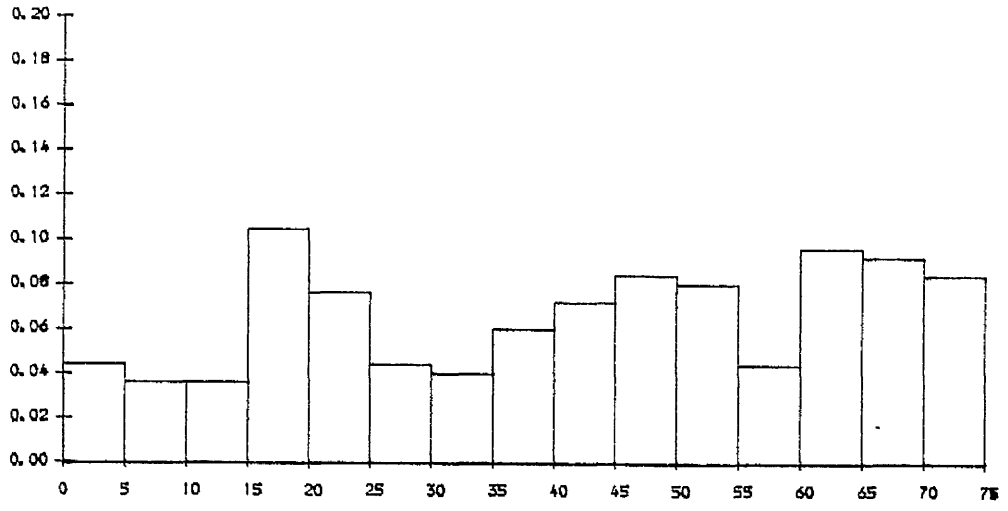
1) Marginal Minimum Mean Squared Error (MMSE) with 2 parameters.
2) Marginal Kullback–Leibler Cross–validation (XVAL) with 2 parameters.
3) Marginal Minimum Mean Squared Error (MMSE) with common parameter.
4) Marginal Kullback-Leibler Cross-validation (XVAL) with common parameter.
5) Multivariate Minimum Mean Squared Error (MV MMSE) with common parameter.
6) Multivariate Kullback-Leibler Cross-validation (MV XVAL) with common parameter.
7) Bivariate Brier Score optimisation (MV BRIER) with cross-validation.
8) Brier Score optimisation (BRIER) with cross-validation and common parameter.

In principle one might collapse the 2-dimensional contingency table across "population", analagous to the usual procedure for marginal parameter estimation where the table is collapsed across one or more variables. However the former would base estimation of $\underline{\lambda}$ upon a larger than appropriate sample size, and so marginal methods 3 and 4 simply average the parameter estimates of methods 1 and 2 respectively, whereas the corresponding "multivariate"

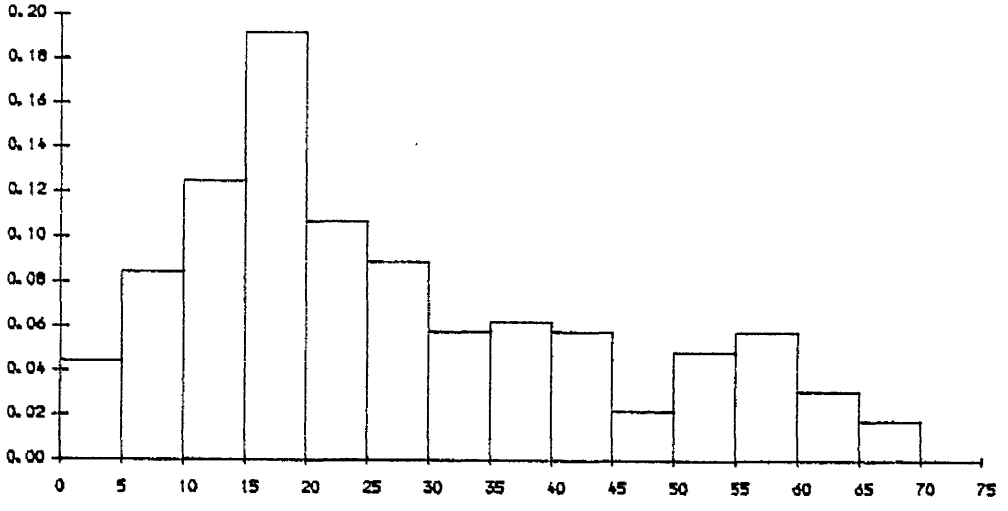
Table 4.1 Univariate data

	Training sample			Test sample		
Age	Cell count	Frequency of π_1	Relative Frequency of π_1	Cell count	Frequency of π_1	Relative Frequency of π_1
0-4	21	11	.524	17	7	.412
5-9	28	9	.321	38	9	.237
10-14	37	9	.243	27	10	.370
15-19	69	26	.377	65	21	.323
20-24	43	19	.442	40	19	.475
25-29	31	11	.355	41	20	.488
30-34	23	10	.435	39	21	.538
35-39	29	15	.517	25	13	.520
40-44	31	18	.581	34	15	.441
45-49	26	21	.808	31	15	.484
50-54	31	20	.645	28	14	.500
55-59	24	11	.458	28	20	.714
60-64	31	24	.774	22	19	.864
65-69	27	23	.852	17	13	.765
≥70	21	21	1.000	24	23	.958
Total	472	248	.525	476	239	.502

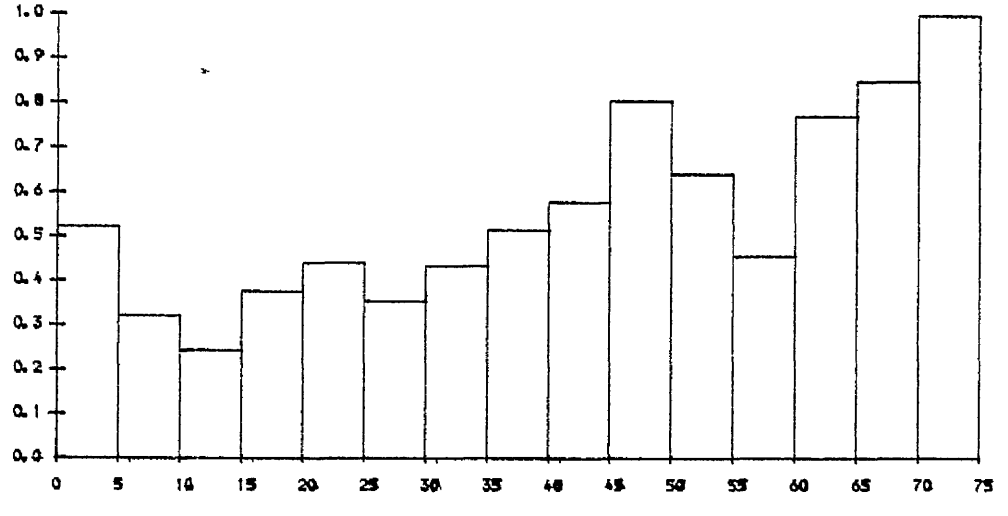
Figure 4.1 Histograms of the raw training relative frequencies.



(a) $p(x \mid \pi_1)$

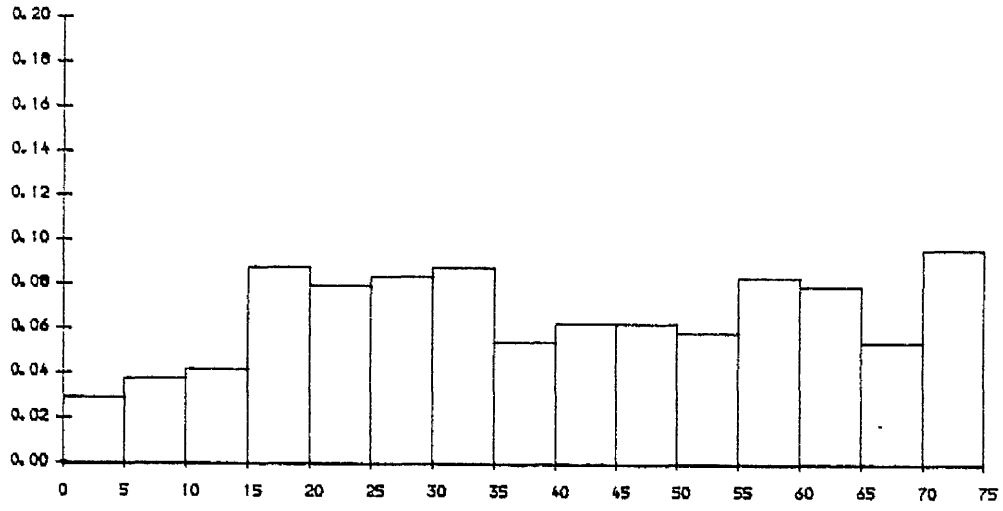


(b) $p(x \mid \pi_2)$

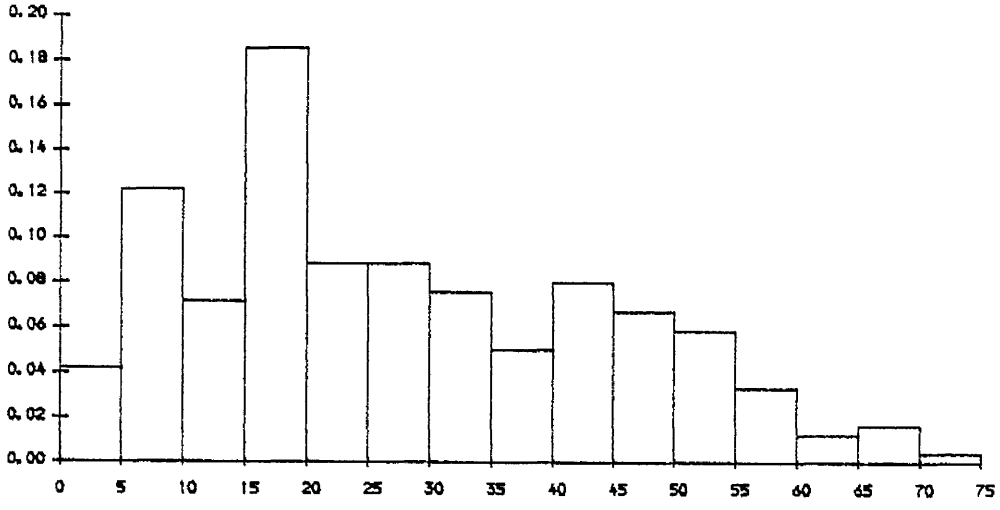


(c) $p(\pi_1 \mid x)$

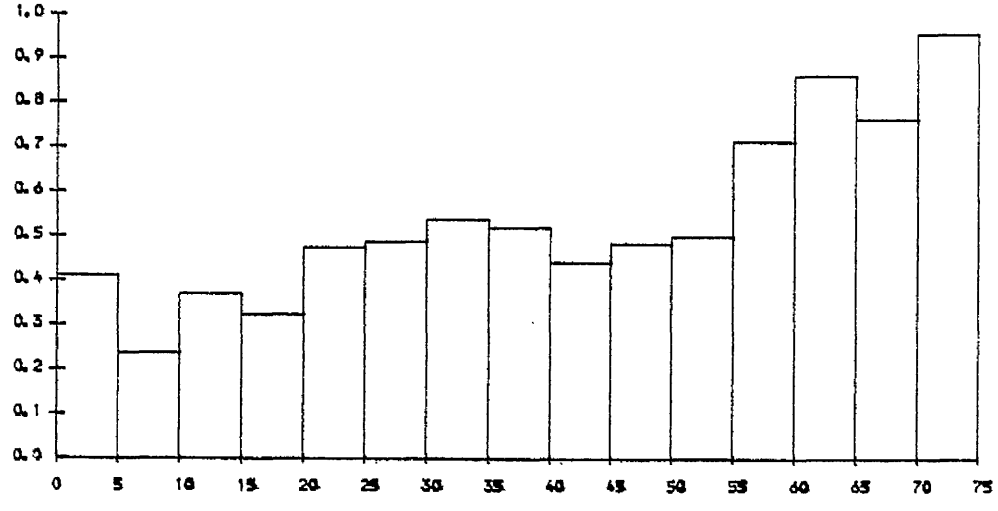
Figure 4.2 Histograms of the raw test relative frequencies.



(a) $p(x \mid \pi_1)$



(b) $p(x \mid \pi_2)$



(c) $p(\pi_1 \mid x)$

methods 5 and 6 set λ_1 equal to λ_2 in the original expression to be optimised (as does method 8). The estimated smoothing parameters and resulting Brier scores achieved on the 500 test cases for the various methods are shown in Table 4.2(a), where λ_i denotes the smoothing parameter used in population π_i. Error rates are also provided for interest and are seen to be very similar. Lower and upper limits on the Brier score for these methods, achieved by no smoothing at all of the raw MLEs (denoted RFTR), and use of the cell relative frequencies in the test set (RFTS), respectively, are provided for comparison in Table 4.2(b). Table 4.3 contains the fitted proportions for each method.

XVAL is seen to perform slightly better on both criteria than MMSE in each case, both marginal and multivariate, smoothing slightly more, although differences in the Brier score are small. Although differences in computation time are minimal in this case, little is gained by using 2 parameters rather than a common one, nor by a multivariate approach, as methods 3 and 4 smooth slightly more than 5 and 6 respectively and are slightly superior. Differences in the fitted distributions are small and the corresponding histograms are visually very similar. Method 2 is presented, as the best, in Figure 4.3(a). Table 4.2(b) provides the highest Brier scores achievable with these kernel methods, by direct optimisation of the test Brier score (using either 2 parameters (method 7T), or a common parameter (8T)), and we find that the isotonic method (Table 4.2(a)) achieves a Brier score very close to the optimum. The potential for improvement over the indirect methods (1-6) is considerable but, disappointingly, the direct methods (7 and 8) using cross-validation, which are virtually identical (see Table 4.3), are poorer on Brier score than any other method, smoothing rather less than would be optimal, especially in π_1, as is also seen from the fitted class conditional distributions in Figures 4.5 and 4.6. While the value of smoothing is seen in that each method outperforms the unsmoothed relative frequencies, in terms of Brier score, isotonisation alone comes close to being optimal (see Figures 4.3 and 4.4). Figure 4.7, a contour plot of the Brier score as a function of the smoothing parameters (λ_1, λ_2), shows the relative performance of some of the kernel methods.

It is of interest to compare the results here to those obtained in Section 2.6 using age as a continuous variable. Particularly

Table 4.2(a) RESULTS FOR THE UNIVARIATE EXAMPLE

Method	Smoothing Parameters		Scores	
	λ_1	λ_2	Brier Score	Error Rate (%)
Marginal methods				
1 MMSE - 2 parameters	.811	.862	.76933	41.8
2 XVAL - 2 parameters	.761	.827	.77027	39.3
3 MMSE - 1 parameter	.836	.836	.76926	41.8
4 XVAL - 1 parameter	.794	.794	.77018	41.8
Multivariate methods				
5 MV MMSE - 1 parameter	.842	.842	.76915	41.8
6 MV XVAL - 1 parameter	.796	.796	.77013	41.8
Direct methods				
7 MV BRIER SCORE - 2 parameters	.946	.935	.76649	41.8
8 BRIER SCORE - 1 parameter	.939	.939	.76655	41.8
ISOTONIC REGRESSION (with quadratic modification)			.77397	39.3

Table 4.2(b) RESULTS FOR THE UNIVARIATE EXAMPLE

Method	Smoothing Parameters		Scores	
	λ_1	λ_2	Brier Score	Error Rate (%)
Reference methods				
Direct optimisation on the test set :				
7T MV BRIER SCORE - 2 parameters	.322	.565	.77399	38.0
8T BRIER SCORE - common parameter	.475	.475	.77336	38.6
TEST PROPORTIONS (RFTS) (resubstitution)			.78341	37.0
TRAINING PROPORTIONS (RFTR)			.76455	41.8

Table 4.3 Fitted distributions: $\hat{p}(x|\pi_1)$

Age	RFTR	1	2	3	4	5	6	7	7T	8	8T	RFTS
0-4	.044	.040	.038	.040	.039	.040	.039	.043	.028	.043	.032	.029
5-9	.036	.037	.037	.037	.037	.037	.037	.036	.039	.036	.038	.038
10-14	.036	.040	.041	.040	.041	.040	.040	.037	.050	.038	.047	.042
15-19	.105	.097	.094	.098	.096	.098	.096	.102	.076	.102	.082	.088
20-24	.077	.076	.076	.076	.076	.076	.076	.076	.076	.076	.076	.080
25-29	.044	.052	.054	.051	.053	.051	.053	.047	.073	.047	.066	.084
30-34	.040	.050	.052	.049	.051	.048	.051	.043	.075	.043	.067	.088
35-39	.061	.066	.068	.066	.067	.065	.067	.062	.082	.062	.077	.054
40-44	.073	.076	.077	.076	.076	.075	.076	.074	.085	.074	.082	.063
45-49	.085	.085	.085	.085	.085	.085	.085	.085	.087	.085	.086	.063
50-54	.081	.081	.081	.081	.081	.081	.081	.081	.082	.081	.082	.058
55-59	.044	.051	.053	.050	.052	.050	.052	.046	.069	.046	.063	.084
60-64	.097	.090	.088	.091	.090	.091	.090	.095	.073	.095	.079	.080
65-69	.093	.084	.082	.085	.083	.085	.083	.090	.062	.090	.069	.054
≥70	.084	.073	.070	.075	.072	.075	.072	.081	.044	.081	.053	.096

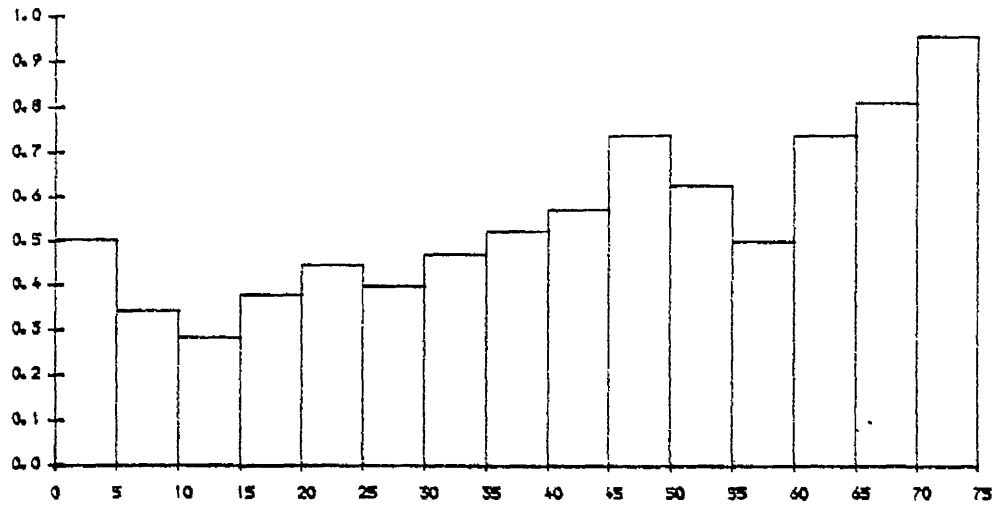
$\hat{p}(x|\pi_2)$

	RFTR	1	2	3	4	5	6	7	7T	8	8T	RFTS
0-4	.045	.042	.042	.042	.042	.042	.042	.044	.038	.044	.037	.042
5-9	.085	.080	.079	.080	.078	.080	.078	.083	.071	.083	.068	.122
10-14	.125	.117	.115	.116	.113	.116	.114	.121	.101	.122	.096	.072
15-19	.192	.176	.172	.173	.168	.173	.168	.184	.141	.185	.130	.186
20-24	.107	.105	.104	.105	.104	.105	.104	.106	.101	.106	.099	.089
25-29	.089	.090	.090	.090	.091	.090	.091	.090	.092	.090	.093	.088
30-34	.058	.064	.065	.064	.066	.064	.066	.061	.075	.060	.079	.076
35-39	.063	.067	.068	.068	.069	.067	.069	.064	.076	.064	.078	.051
40-44	.058	.062	.063	.063	.064	.063	.064	.060	.071	.060	.073	.080
45-49	.022	.031	.033	.032	.035	.032	.035	.026	.049	.026	.054	.067
50-54	.049	.052	.053	.052	.053	.052	.053	.050	.058	.050	.060	.059
55-59	.058	.058	.058	.058	.058	.058	.058	.058	.058	.058	.058	.034
60-64	.031	.033	.034	.034	.034	.034	.034	.032	.038	.032	.039	.013
65-69	.018	.020	.020	.020	.021	.020	.021	.019	.024	.019	.026	.017
≥70	.000	.002	.003	.003	.004	.003	.004	.001	.008	.001	.009	.004

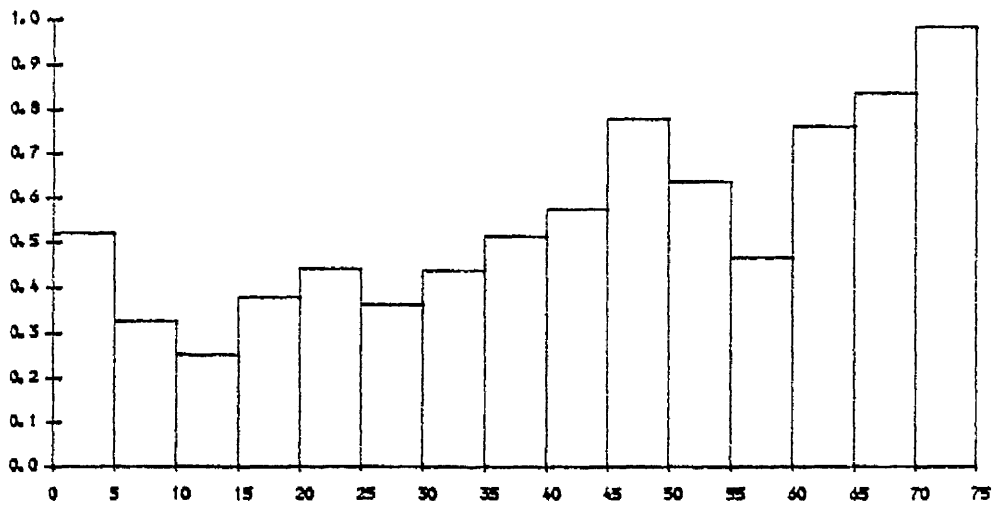
$\hat{p}(\pi_1|x)$

	RFTR	1	2	3	4	5	6	7	7T	8	8T	RFTS	ISO
0-4	.524	.508	.504	.514	.512	.515	.512	.522	.448	.520	.488	.412	.524
5-9	.321	.338	.343	.340	.345	.339	.345	.328	.380	.328	.386	.237	.321
10-14	.243	.275	.284	.275	.284	.274	.283	.254	.357	.255	.354	.370	.243
15-19	.377	.378	.379	.385	.388	.385	.388	.381	.373	.380	.412	.323	.377
20-24	.442	.446	.447	.447	.448	.447	.448	.444	.455	.444	.459	.475	.405
25-29	.355	.391	.400	.386	.393	.385	.392	.365	.466	.367	.442	.488	.405
30-34	.435	.466	.472	.455	.459	.454	.459	.440	.524	.443	.485	.538	.435
35-39	.517	.524	.526	.518	.518	.518	.518	.516	.544	.518	.520	.520	.517
40-44	.581	.576	.575	.571	.569	.572	.569	.576	.571	.577	.554	.441	.581
45-49	.808	.754	.742	.745	.730	.746	.731	.781	.662	.783	.637	.484	.642
50-54	.645	.633	.630	.631	.627	.631	.627	.639	.610	.640	.601	.500	.642
55-59	.458	.494	.503	.490	.497	.489	.497	.469	.568	.470	.548	.714	.642
60-64	.774	.749	.743	.749	.742	.750	.743	.765	.681	.765	.688	.864	.774
65-69	.852	.824	.816	.823	.815	.824	.816	.842	.737	.841	.748	.765	.852
≥70	1.000	.971	.962	.966	.957	.968	.957	.987	.864	.988	.864	.958	1.000

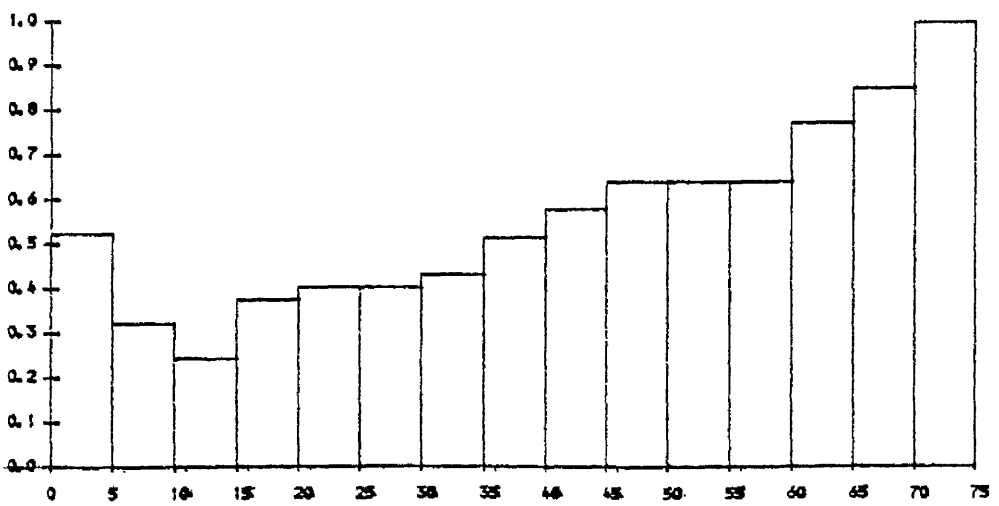
Figure 4.3 $\hat{p}(\pi_1 \mid x)$



(a) XVAL with 2 parameters (2).

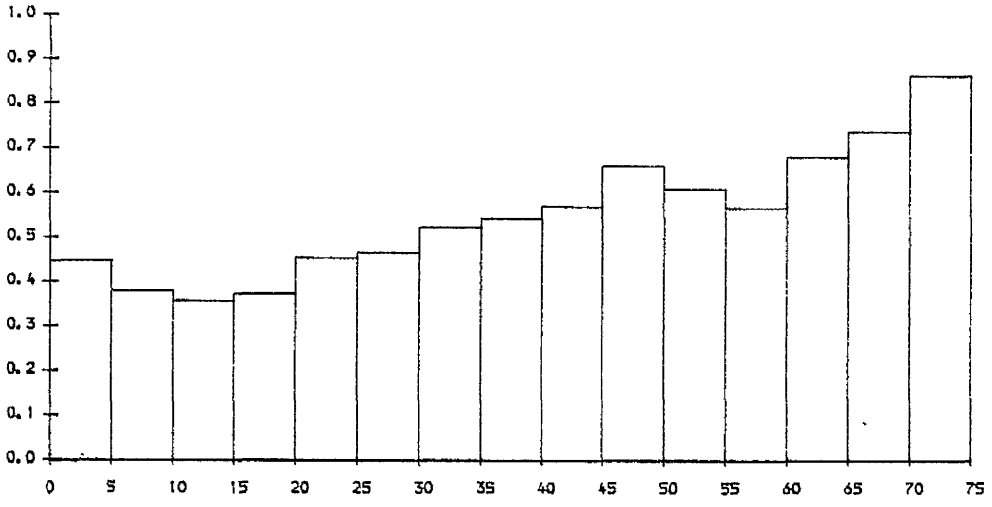


(b) MV BRIER with 2 parameters using cross-validation (7).



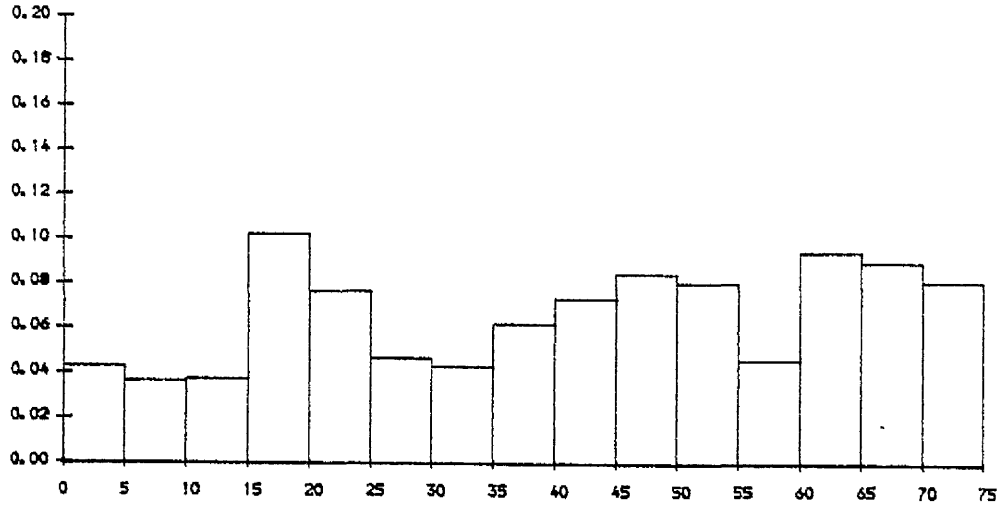
(c) Isotonic regression (ISO).

Figure 4.4 $\hat{p}(\pi_1 \mid x)$

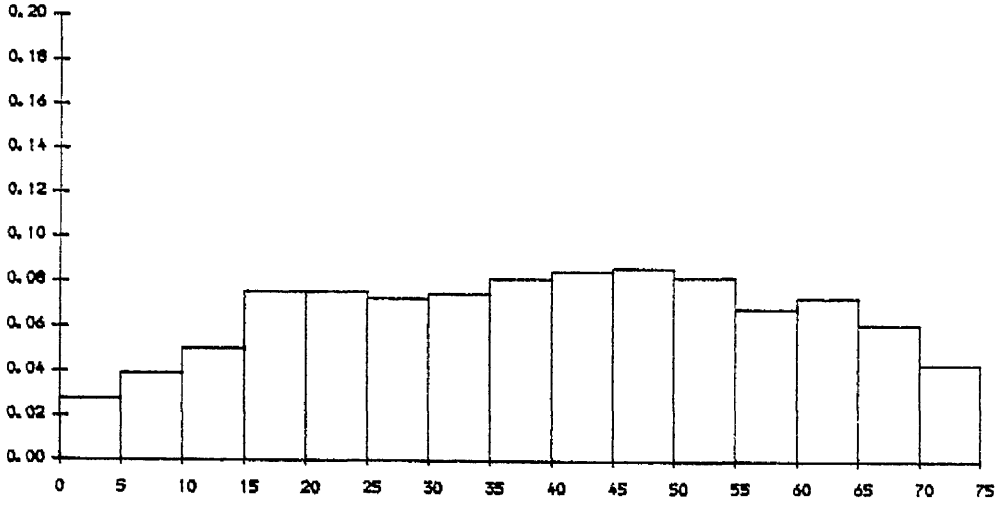


Brier score optimisation on the test data (7T).

Figure 4.5 $\hat{p}(x \mid \pi_1)$

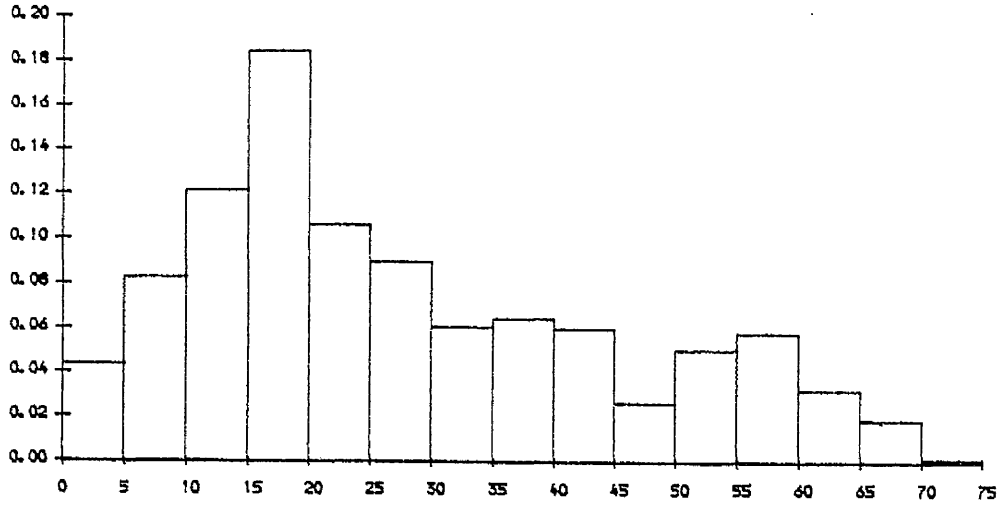


(a) MV BRIER using cross-validation (7).

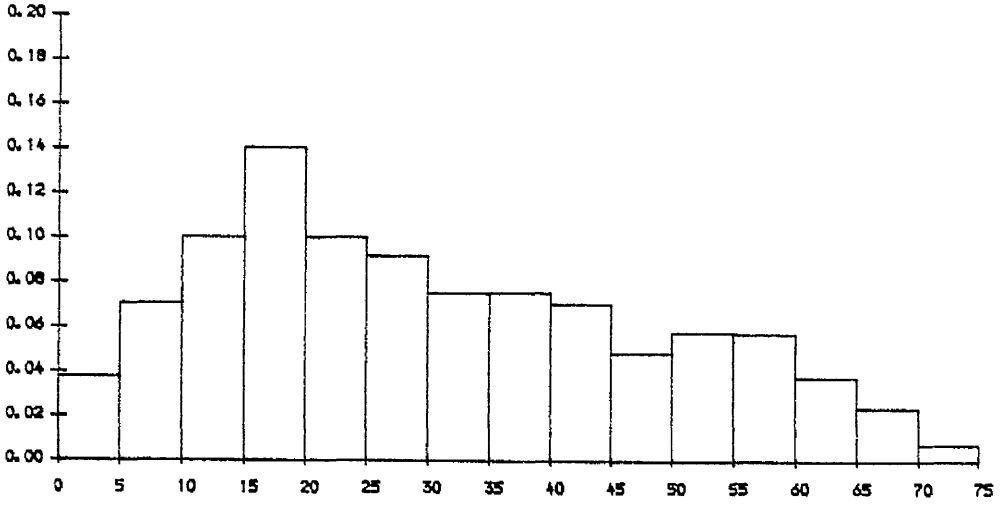


(b) Brier score optimisation on the test set (7T).

Figure 4.6 $\hat{p}(x \mid \pi_2)$



(a) MV BRIER using cross-validation (7).



(b) Brier score optimisation on the test set (7T).

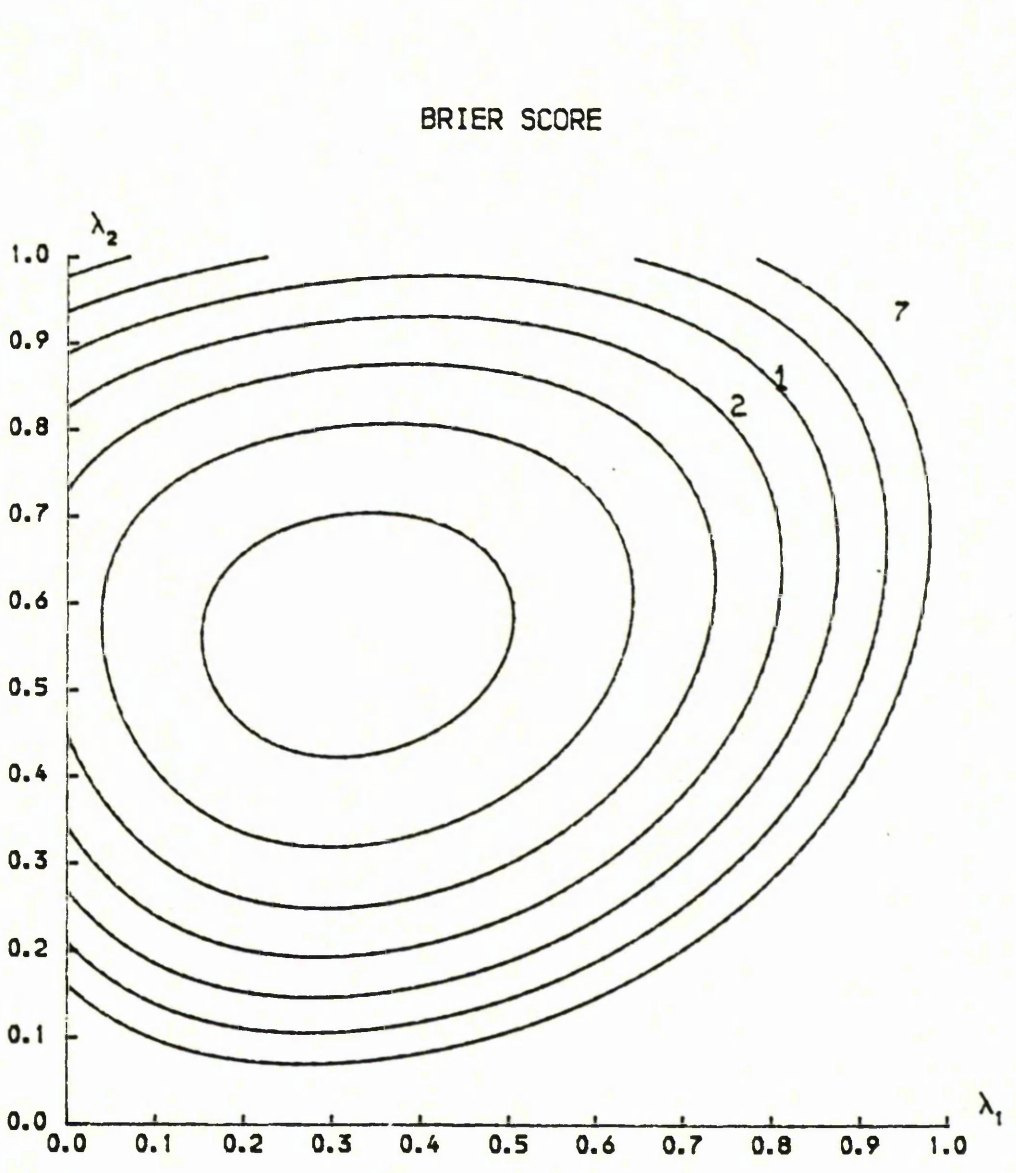


Figure 4.7 Contour plot of the test Brier score as a function of the smoothing parameters (λ_1, λ_2) used in populations π_1 and π_2 respectively. Methods 1, 2 and 7 are as in Table 4.2(a). Contour heights are :- .7735, .7725, .7715, .7705, .7695, .7685 and .7675.

striking is the fact that the best continuous method, namely the Normal Optimal, achieves a predicted probability function, Figure 2.51(a), which strongly resembles a smoothed version of the isotonic estimate, Figure 4.3(c), while the other continuous functions, the result of not smoothing nearly as much, follow the irregularities of the raw relative frequency histogram, Figure 4.1(c), and are comparable to the discrete kernel methods. Surprisingly, perhaps, the Brier score achieved by NOPT, although similar, is not quite as high as that of the rougher discrete isotonic estimate.

4.3.2 Bivariate examples

A 15 x 10 table

The example above is unrealistic in that in practice multiple predictors are much more common. Here we use both age, categorised as above, and Coma Sum, with 10 categories (namely 3, 4,....., 11, and 12-15), as indicants, with product kernels. The data are shown in Tables 4.4(a) and 4.4(b) and as 3-dimensional isometric projections in Figures 4.8 and 4.9, where X and Y denote age and Coma Sum respectively. It will be seen that while $p(\pi_1|x)$ increases with age it is antitonic with respect to Coma Sum, i.e. it decreases as Coma Sum increases. (See also Figures 4.1(c), 4.2(c) and 4.10).

With the kernel method there are now 4 smoothing parameters to choose and the following criteria were used :

1) Marginal Minimum Mean Squared Error (MMSE) with 4 parameters.
2) Marginal Kullback-Leibler Cross-validation (XVAL) with 4 parameters.
3) Marginal Minimum Mean Square Error (MMSE) with 2 parameters.
4) Marginal Kullback-Leibler Cross-validation (XVAL) with 2 parameters.
6) Multivariate Kullback-Leibler Cross-validation (MV XVAL) with 4 parameters.
8) Multivariate Kullback-Leibler Cross-validation (MV XVAL) with 2 parameters.
10) Standardisation Kullback-Leibler Cross-validation (STD XVAL) with 2 parameters.

Table 4.4(a) Relative frequencies of π_1 for the 15 x 10 table

Training data

Age	Coma Sum 3	4	5	6	7	8	9	10	11	≥12
0-4	1.000	1.000	.500	.667	.500	.250	.525	.000	.525	.500
5-9	.525	.500	.714	.167	.286	.000	.000	.000	.525	.525
10-14	.525	.500	.400	.000	.428	.167	.333	.000	.000	.000
15-19	1.000	.714	.500	.381	.167	.125	.000	.000	.525	.000
20-24	1.000	.800	.500	.462	.222	.444	.500	.525	.000	.000
25-29	1.000	.667	.600	.428	.167	.333	.000	.000	.000	.525
30-34	1.000	1.000	.667	.454	.000	.000	.000	.525	.525	.000
35-39	1.000	1.000	.500	.600	.250	.333	.000	.525	.000	.000
40-44	1.000	1.000	1.000	.750	.556	.200	.000	.525	.333	.000
45-49	1.000	1.000	.333	.875	.500	.525	.525	.525	.525	.000
50-54	1.000	1.000	1.000	.667	.600	.500	.525	.000	.525	.250
55-59	.525	1.000	.500	.400	.750	.143	1.000	.525	.525	.000
60-64	.525	.750	.500	1.000	1.000	1.000	1.000	.000	.000	.000
65-69	.525	1.000	1.000	1.000	1.000	.500	.667	.000	.000	.525
≥70	1.000	1.000	1.000	1.000	1.000	1.000	.525	1.000	.525	.525

Test data

Age	Coma Sum 3	4	5	6	7	8	9	10	11	≥12
0-4	.502	.667	.333	.500	.667	.000	.000	.502	.000	.502
5-9	.502	.600	1.000	.071	.375	.000	.000	.000	.502	.502
10-14	1.000	.667	.400	.286	.333	.500	.250	.502	.000	.000
15-19	1.000	.875	.700	.200	.071	.167	.502	.000	.000	.000
20-24	1.000	.778	.667	.583	.111	.000	.000	.000	1.000	.502
25-29	1.000	.500	.500	.533	.400	.000	.500	.502	.502	.333
30-34	1.000	.857	1.000	.636	.333	.600	.000	.000	.000	.000
35-39	1.000	.667	1.000	.667	.167	.667	.000	.502	.502	.000
40-44	.000	1.000	1.000	.625	.200	.167	.000	.000	.000	.000
45-49	1.000	1.000	.333	.500	.333	.167	1.000	.502	.000	.000
50-54	.500	.714	.750	.250	.200	1.000	.000	.667	.000	.502
55-59	1.000	1.000	1.000	.714	.667	.500	.502	.000	.502	.333
60-64	1.000	1.000	1.000	1.000	1.000	.000	.502	.000	.000	1.000
65-69	.502	1.000	1.000	.500	.833	.502	.000	1.000	.502	.000
≥70	1.000	1.000	1.000	1.000	.833	1.000	1.000	.502	.502	.502

Note: The data presented are slightly smoothed in that the relative frequency of π_1 in empty cells has been set to the overall relative frequency, .525 or .502 for the training and test data respectively.

Table 4.4(b) No. of cases in each cell

Training data

Age	Coma Sum									
	3	4	5	6	7	8	9	10	11	≥12
0–4	1	2	2	3	6	4	0	1	0	2
5–9	0	2	7	6	7	4	1	1	0	0
10–14	0	4	5	7	7	6	3	3	1	1
15–19	7	7	6	21	12	8	4	3	0	1
20–24	1	5	2	13	9	9	2	0	1	1
25–29	1	3	5	7	6	3	4	1	1	0
30–34	2	1	3	11	1	2	2	0	0	1
35–39	2	7	2	5	4	3	1	0	2	3
40–44	1	4	3	4	9	5	1	0	3	1
45–49	2	10	3	8	2	0	0	0	0	1
50–54	1	4	4	3	10	4	0	1	0	4
55–59	0	3	2	5	4	7	1	0	0	2
60–64	0	4	4	7	6	5	1	1	2	1
65–69	0	4	2	8	6	2	3	1	1	0
≥70	1	4	2	3	7	3	0	1	0	0

Test data

Age	Coma Sum									
	3	4	5	6	7	8	9	10	11	≥12
0–4	0	3	3	4	3	2	1	0	1	0
5–9	0	5	2	14	8	5	2	2	0	0
10–14	1	3	5	7	3	2	4	0	1	1
15–19	1	8	10	15	14	12	0	3	1	1
20–24	1	9	3	12	9	1	3	1	1	0
25–29	3	4	2	15	10	2	2	0	0	3
30–34	1	7	2	11	6	5	2	3	1	1
35–39	1	3	1	9	6	3	1	0	0	1
40–44	1	5	3	8	5	6	1	2	1	2
45–49	4	6	3	2	3	6	1	0	2	4
50–54	2	7	4	4	5	1	1	3	1	0
55–59	2	6	2	7	3	4	0	1	0	3
60–64	2	1	2	8	5	1	0	1	1	1
65–69	0	1	5	2	6	0	1	1	0	1
≥70	1	5	2	7	6	1	2	0	0	0

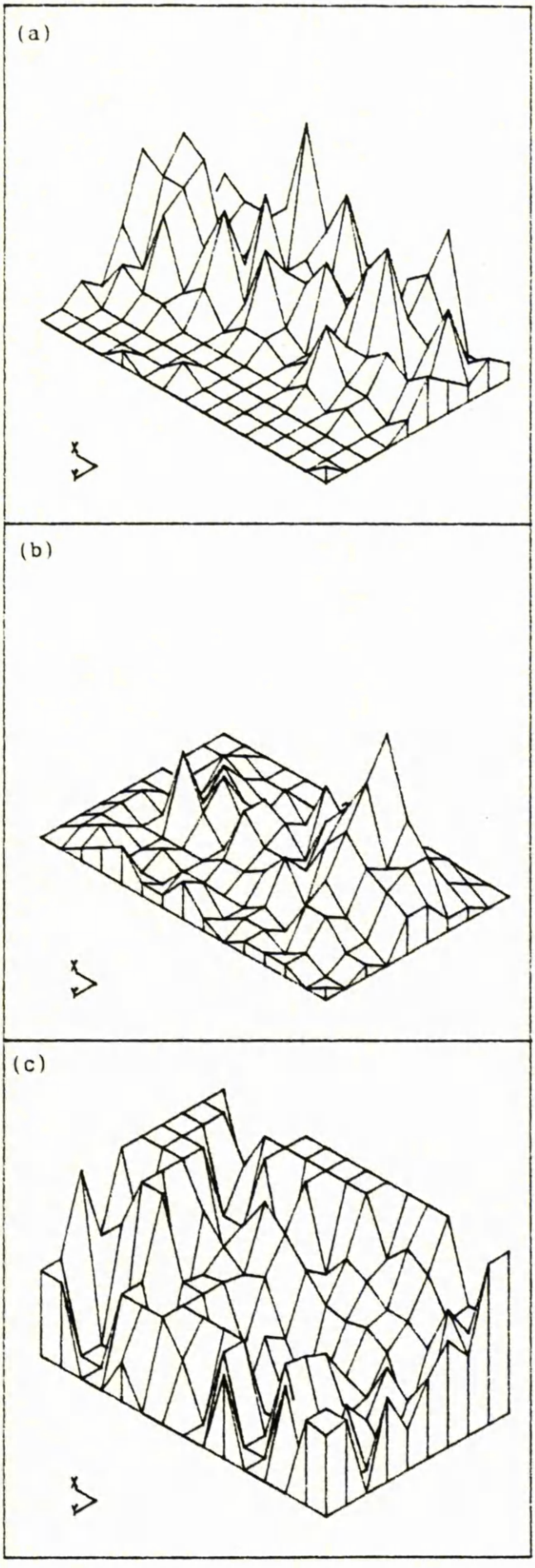


Figure 4.8 Relative frequencies displayed as 3-dimensional isometric projections for the training data.

(a) $p(\underline{x} \mid \pi_1)$

(b) $p(\underline{x} \mid \pi_2)$

(c) $p(\pi_1 \mid \underline{x})$

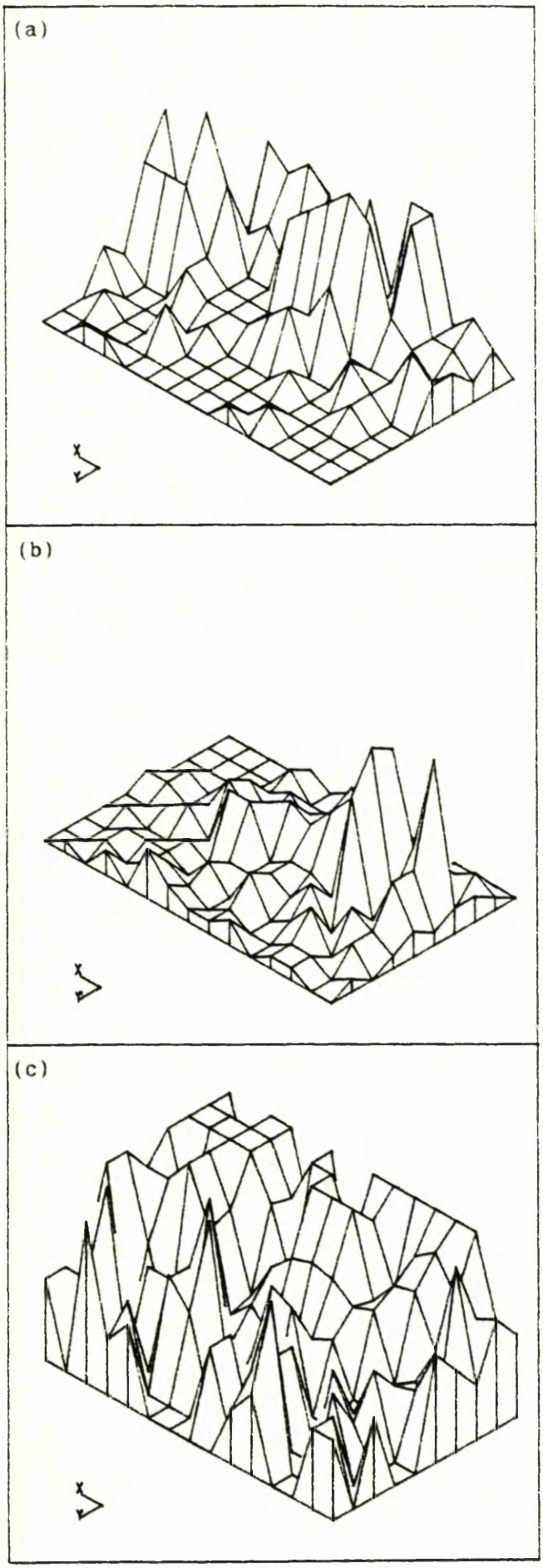


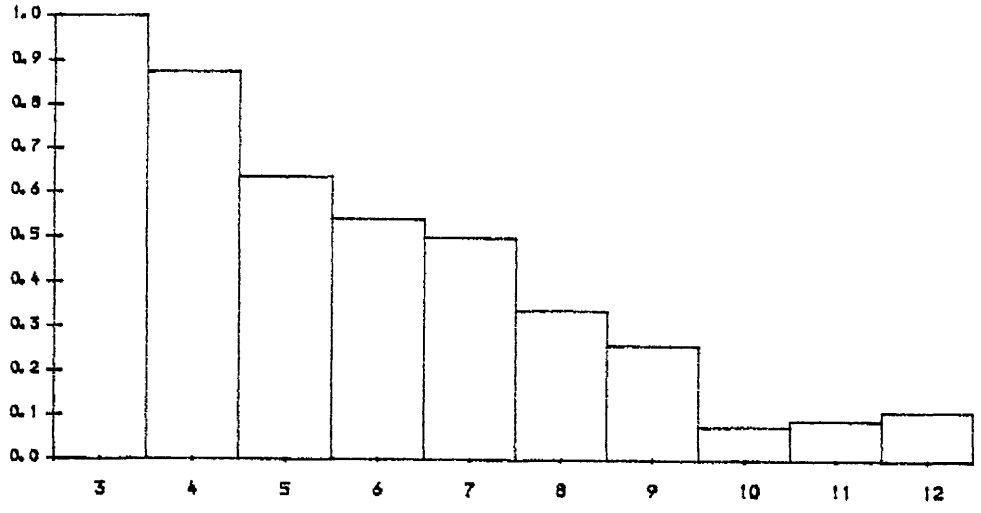
Figure 4.9 Relative frequencies displayed as 3-dimensional isometric projections for the test data.

(a) $p(\underline{x} \mid \pi_1)$

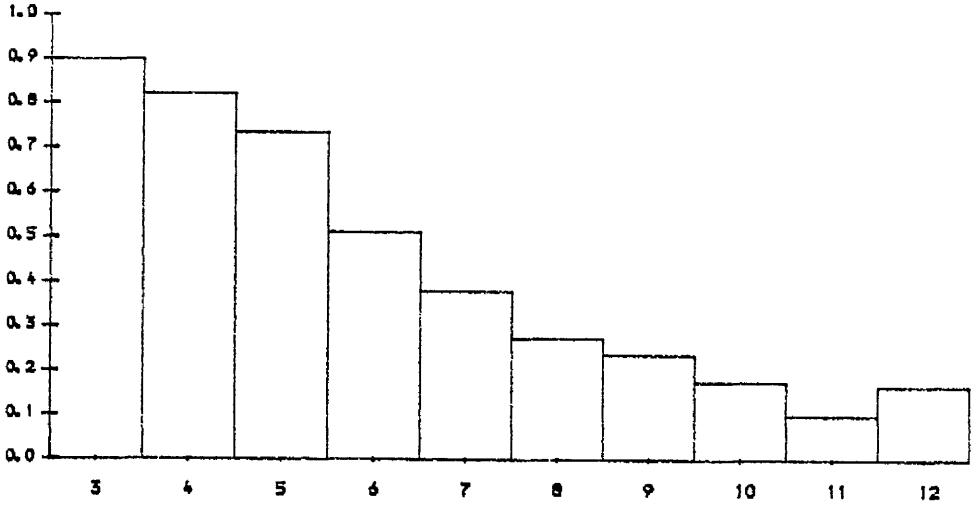
(b) $p(\underline{x} \mid \pi_2)$

(c) $p(\pi_1 \mid \underline{x})$

Figure 4.10 Relative frequency of π_1 given Coma Sum.



(a) Training data.



(b) Test data.

12) Standardisation Kullback-Leibler Cross-validation (STD XVAL) with 1 parameter.
13) Multivariate Brier Score method (MV BRIER) with 4 parameters.
14) Multivariate Brier Score method (MV BRIER) with 2 parameters.
15) Standardisation Brier Score method (STD BRIER) with 2 parameters.
16) Standardisation Brier Score method (STD BRIER) with 1 parameter.

For methods 1-4 parameters are chosen to optimise functions of the marginal distributions whereas in 6-16 simultaneous optimisation of multivariate criteria is used. The multivariate methods were not used with the MSE criterion as the resulting computation proved too time consuming, hence the absence of methods 5, 7, 9 and 11 (which are used below). The 4-parameter methods allow a different parameter for each variable within each of the 2 populations, whilst the 2-parameter ones assume a common smoothing parameter for a given variable across populations. (As in Section 4.3.1, our 2-dimensional marginal methods simply average across populations the appropriate 4-dimensional marginal parameter estimates.) Although this is generally thought to lead to a suboptimal rule (though see Hand, 1982, pp. 162-164, for some counter-examples) it was considered that age is sufficiently different in type from Coma Sum to warrant a different degree of smoothing. Since age is of interval type we would expect $\hat{p}(\pi_1|\text{age})$ to be smoother than the corresponding function of Coma Sum. An alternative, which still allows a degree of difference in smoothing of the 2 variables for a given population, is the method of Titterington et al. (1981) (see Section 4.2), which adjusts the smoothing parameter according to the number of categories in a roughly analagous manner to pre-standardisation of continuous variables followed by use of a single smoothing parameter. The 2- and 1-dimensional "standardisation" methods 10 and 12 may be compared to the 4- and 2-dimensional methods 6 and 8, and 15 and 16 with 13 and 14.

Again the Brier scores are tabulated, in Tables 4.5(a) and (b), together with the corresponding smoothing parameters and the results for the isotonic method. Results for the reference methods are given in Table 4.5(c).

Compared to the results of Section 4.3.1, in each case the

Table 4.5(a) RESULTS FOR THE 15 X 10 TABLE

Method	Smoothing Parameters		Scores	
	π_1 Age Coma Sum	π_2	Brier Score	Error Rate (%)
Marginal methods				
1 MMSE - 4 parameters	.811 .941	.862 .902	.82150	27.1
2 XVAL - 4 parameters	.761 .942	.827 .972	.82089	26.3
3 MMSE - 2 parameters	.836 .922	.836 .922	.82117	26.3
4 XVAL - 2 parameters	.794 .957	.794 .957	.82258	26.3
Multivariate methods				
6 MV XVAL - 4 parameters	.447 .893	.519 .716	.82379	25.0
8 MV XVAL - 2 parameters	.442 .861	.442 .861	.82409	25.8
10 STD XVAL - 2 parameters	.704 .715	.614 .628	.82090	26.3
12 STD XVAL - 1 parameter	.664 .676	.664 .676	.82010	27.3

Table 4.5(b) RESULTS FOR THE 15 X 10 TABLE

Method	Smoothing Parameters		Scores	
	π_1 Age Coma Sum	π_2	Brier Score	Error Rate (%)
Direct methods				
13 MV BRIER SCORE - 4 parameters	.838 .966	.806 .923	.82236	26.0
14 MV BRIER SCORE - 2 parameters	.808 .971	.808 .971	.82222	26.3
15 STD BRIER SCORE - 2 parameters	.906 .909	.846 .852	.82054	27.3
16 STD BRIER SCORE - 1 parameter	.876 .880	.876 .880	.81912	27.7
ISOTONIC REGRESSION			.82599	26.0

Table 4.5(c) REFERENCE METHODS FOR THE 15 X 10 TABLE

Method	Smoothing Parameters		Scores	
	Age Coma Sum	π_1 π_2	Brier Score	Error Rate (%)
Direct optimisation on test set :				
13T MV BRIER SCORE - 4 parameters	.566 .935	.575 .894	.82518	25.8
14T MV BRIER SCORE - 2 parameters	.558 .932	.558 .932	.82512	25.8
15T STD BRIER SCORE - 2 parameters	.819 .826	.712 .722	.82224	27.5
16T STD BRIER SCORE - 1 parameter	.758 .766	.758 .766	.82098	27.9
TEST PROPORTIONS (RFTS) (1)			.86951	19.8
TRAINING PROPORTIONS (RFTR) (1)			.80221	30.7

NOTE (1) : Relative frequencies in zero weighted cells have been set to the overall relative frequency or estimated prior in the population of interest.

value of the extra variable is demonstrated in higher Brier scores and lower error rates. It is also evident that less smoothing is desirable for Coma Sum than age (as would be expected), as each method indicates this.

Of the marginal methods 1-4, the differences in fitted class conditional distributions (not shown) are small as all smooth similarly, although there is a noticeable difference in both the 2- and 4-parameter cases between XVAL and MMSE in the predicted distributions $\{\hat{p}(\pi_1|\underline{x})\}$ (see Figure 4.11).

The multivariate approach (methods 6 and 8) is now superior for XVAL in each case (here, unexpectedly, 2 parameters are better than 4, as is also true of the marginal XVAL methods) and as expected indicates rather more smoothing, of age especially and most noticeably in π_1, than the marginal methods (see Figure 4.12 for methods 4 and 8). Method 6 smooths π_1 a little less but π_2 rather more and overall appears slightly smoother than method 8 (Figure 4.11), though its Brier score is lower.

The estimated class conditional distributions of the direct Brier methods are very similar although there are some differences in the predicted distributions (Figure 4.13). In each case the data-based method smooths less (age especially, for the non-standardisation methods) than direct optimisation suggests and differences can be seen in the marginals also (Figure 4.14 shows methods 13 and 13T). They now perform better than all marginal methods except 2-D XVAL (4) on Brier score, though not quite as well as the corresponding MV XVAL methods which specify near optimal degrees of smoothing. Again the isotonic method does well in that its Brier score in fact outperforms that achieved by direct optimisation of the test Brier score using the kernel approach. (Table 4.5(c)). Figure 4.15 shows the unsmoothed data and fitted predicted probabilities for the best Brier method (13), best non-Brier kernel method (8) and the isotonic method.

Of the standardisation methods, Table 4.5(c) shows that they do not allow sufficient variability of smoothing parameters between variables, as the optimum achievable Brier score is rather lower in each case than the usual 4- and 2-parameter methods with less and more than optimal smoothing of age and Coma Sum respectively in each population (see Figures 4.13 and 4.14). The data-based standardisation methods are also poorer than the corresponding methods which use more parameters, both the multivariate and direct

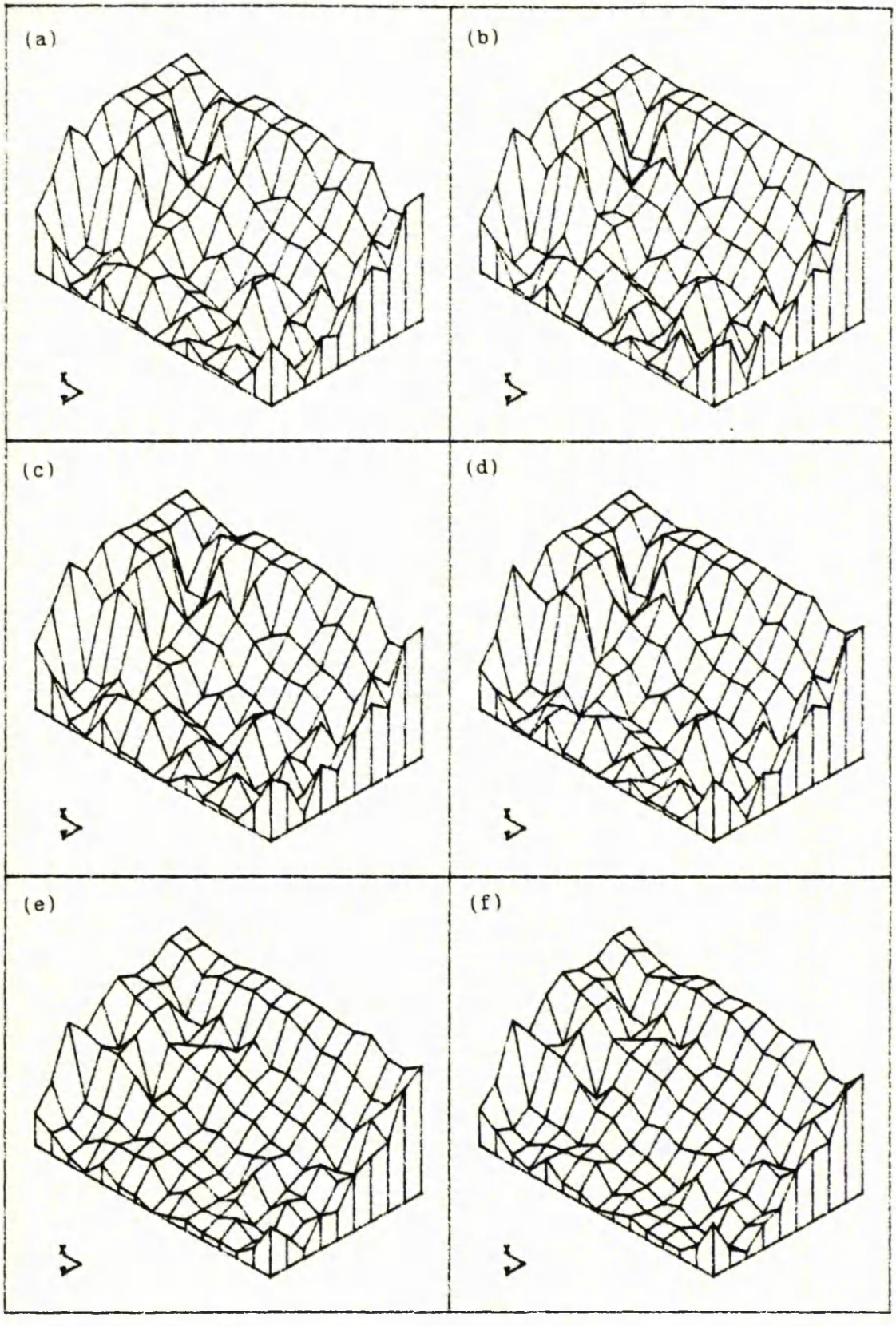


Figure 4.11 $\hat{p}(\pi_1 \mid \underline{x})$

(a) MMSE with 4 parameters (1).

(b) XVAL with 4 parameters (2).

(c) MMSE with 2 parameters (3).

(d) XVAL with 2 parameters (4).

(e) MV XVAL with 2 parameters (8).

(f) MV XVAL with 4 parameters (6).

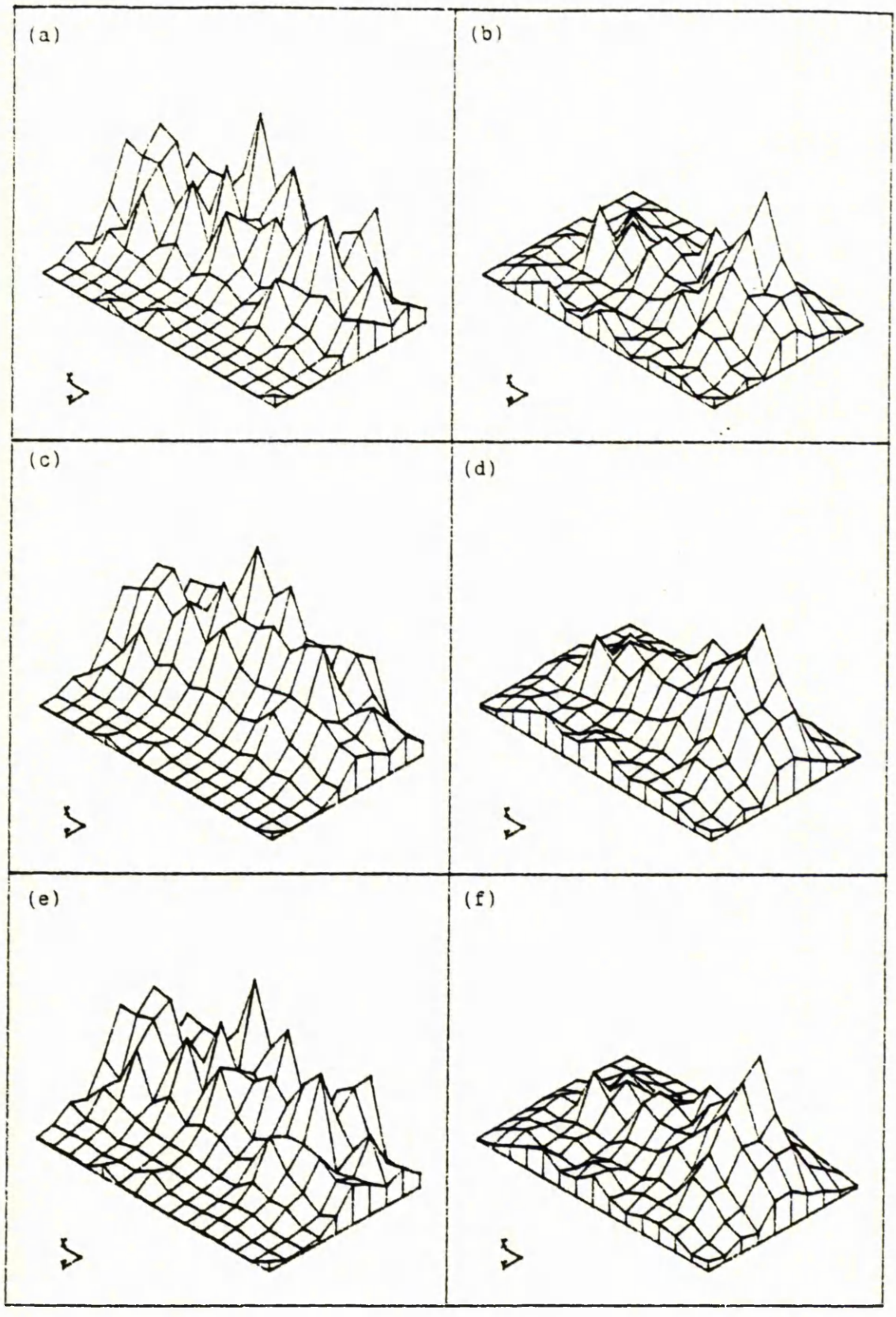


Figure 4.12

(a) $\hat{p}(\underline{x} \mid \pi_1)$ for XVAL with 2 parameters (4).

(b) $\hat{p}(\underline{x} \mid \pi_2)$ for (4).

(c) $\hat{p}(\underline{x} \mid \pi_1)$ for MV XVAL with 2 parameters (8).

(d) $\hat{p}(\underline{x} \mid \pi_2)$ for (8).

(e) $\hat{p}(\underline{x} \mid \pi_1)$ for STD XVAL with 2 parameters (10).

(f) $\hat{p}(\underline{x} \mid \pi_2)$ for (10).

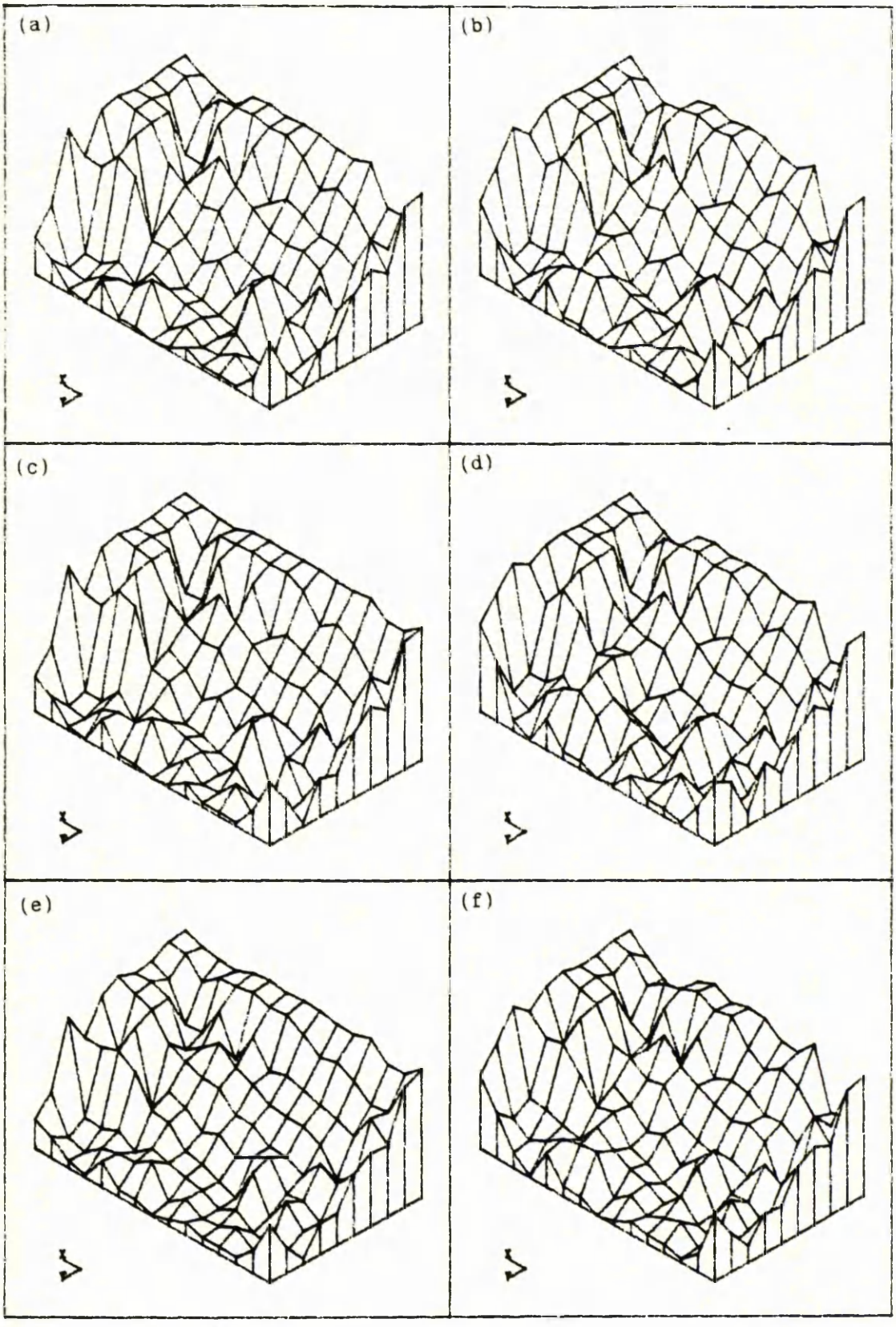


Figure 4.13 $\hat{p}(\pi_1 \mid \underline{x})$

(a) MV BRIER with 4 parameters (13).

(b) STD BRIER with 2 parameters (15).

(c) MV BRIER with 2 parameters (14).

(d) STD BRIER with 1 parameter (16).

(e) 13 optimised on test data (13T).

(f) 15 optimised on test data (15T).

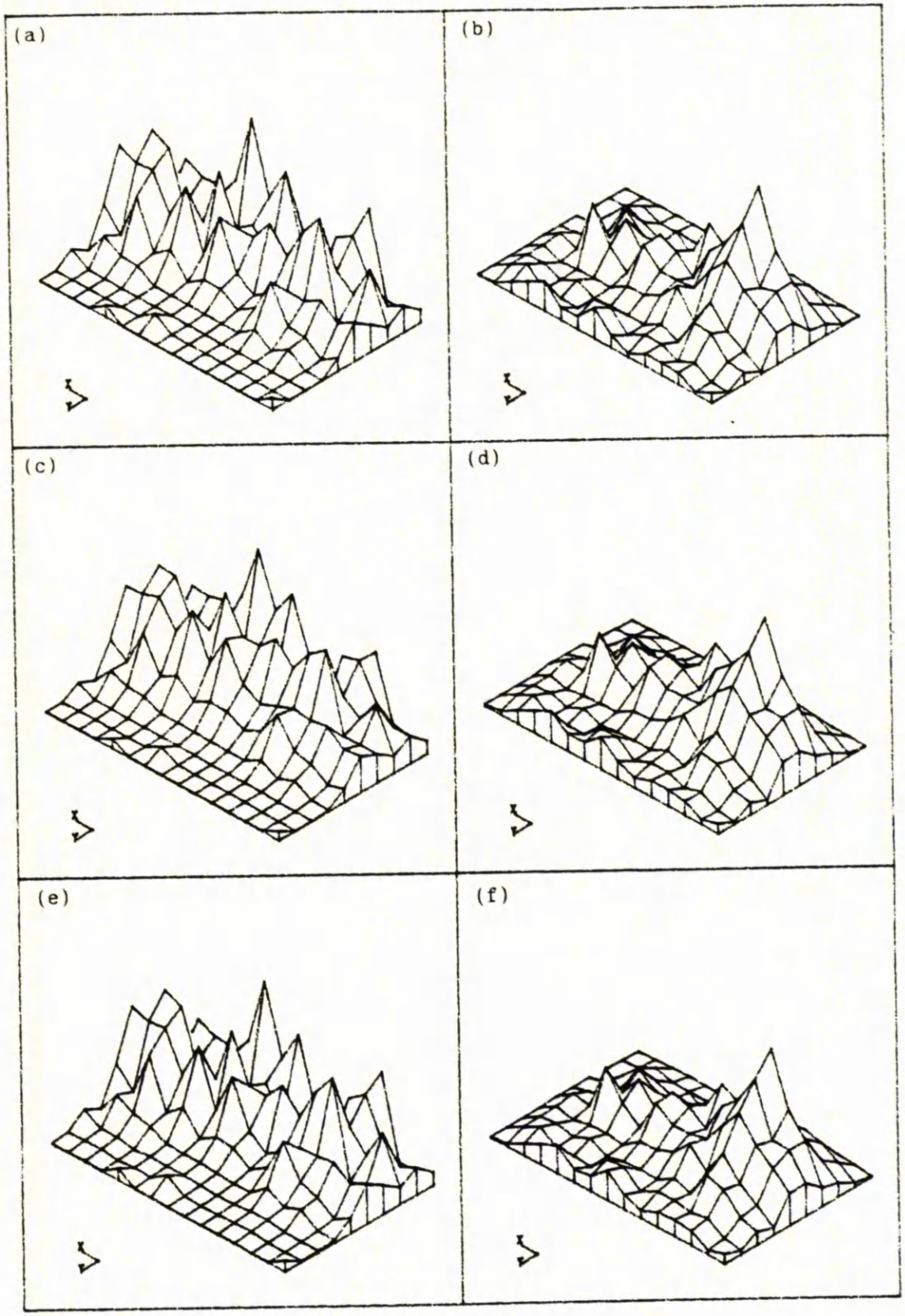


Figure 4.14

(a) $\hat{p}(\underline{x} \mid \pi_1)$ for MV BRIER with 4 parameters (13).

(b) $\hat{p}(\underline{x} \mid \pi_2)$ for (13).

(c) $\hat{p}(\underline{x} \mid \pi_1)$ for (13) optimised on the test data (13T).

(d) $\hat{p}(\underline{x} \mid \pi_2)$ for (13T).

(e) $\hat{p}(\underline{x} \mid \pi_1)$ for STD BRIER with 2 parameters optimised on the test data (15T).

(f) $\hat{p}(\underline{x} \mid \pi_2)$ for (15T).

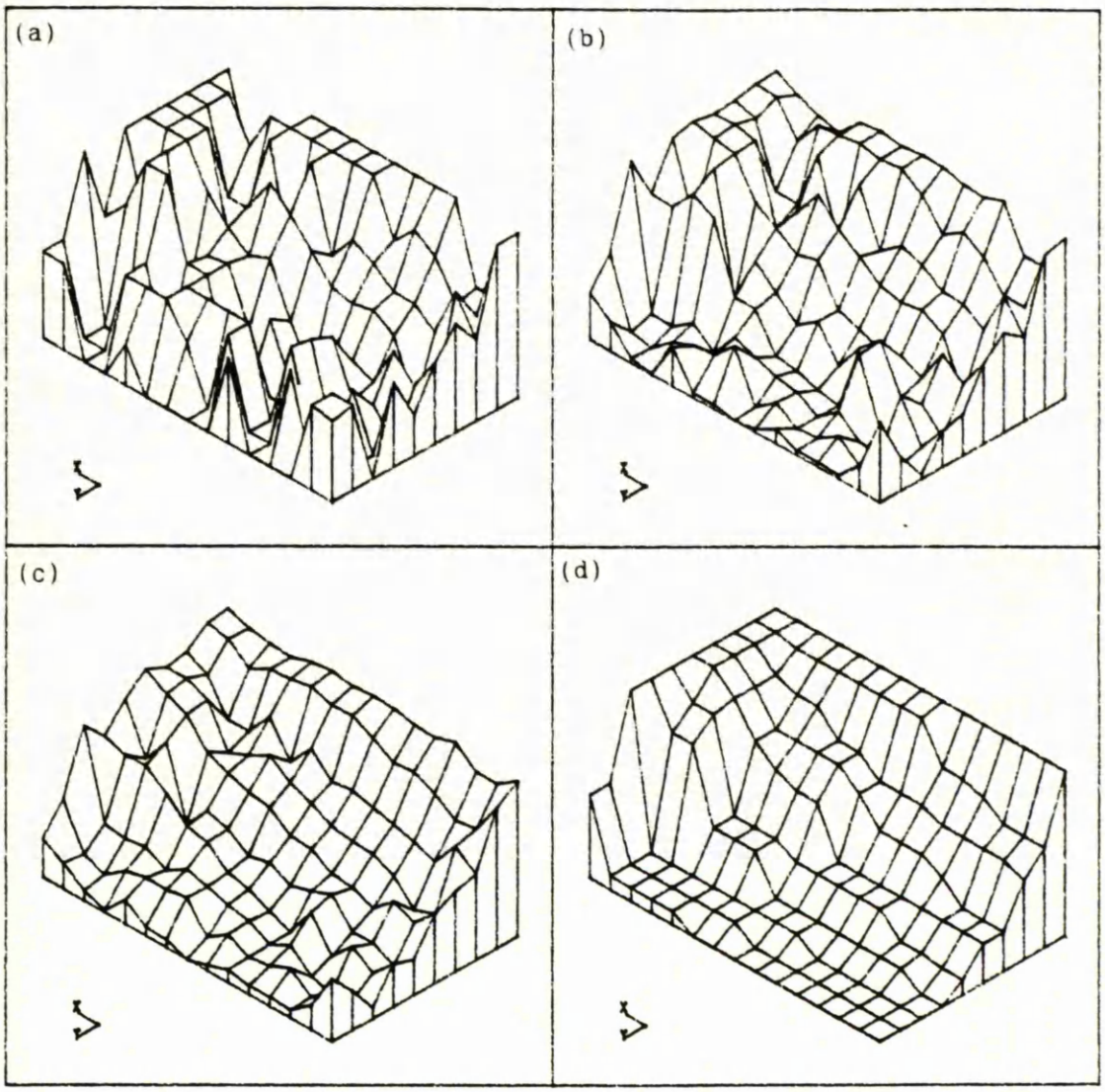


Figure 4.15 $\hat{p}(\pi_1 \mid \underline{x})$

(a) Unsmoothed training relative frequencies.

(b) MV BRIER with 4 parameters (13).

(c) MV XVAL with 2 parameters (8).

(d) Isotonic regression.

Brier ones, on both Brier score and error rate, and compare poorly to the marginal methods. The fitted class conditional distributions are intermediate between those of the marginal and multivariate methods, but generally speaking are more like the former (Figure 4.12). Again there are differences in the predicted distributions (see Figure 4.16). Once more MV XVAL is superior to the direct Brier methods, smoothing more, although not as noticeably.

The contour plots, Figures 4.17 and 4.18, show the test Brier score as a function of smoothing parameters λ_1 and λ_2. In the former λ_1 is associated with age and λ_2 with Coma Sum, taken to be common to both populations, while the latter associates λ_i with age in the ith population and uses the standardisation approach. The former shows the value of allowing a separate smoothing parameter for Coma Sum as this clearly requires rather less smoothing than age, while in the latter less smoothing is indicated for π_1 than π_2 as may be expected from Figures 4.2(a)-(b). In both cases data-based methods fall short of optimal, though less so in the latter instance.

A 5 x 5 table

Strictly speaking, in the 15 x 10 example the isotonic method is slightly handicapped in that no modification has been made to the isotonic algorithm to allow for the quadratic effect of age upon outcome. Collapsing the data from a 15 × 10 table to, say, a 5 × 5 one, removes any observed convexity in the raw relative frequencies, while retaining age groups of equal width gives the ordered kernel its best possible chance to perform well. Use of 15-year age groups will reduce the effect of the few cases of very young age, making isotonicity a reasonable assumption, and although collapsing Coma Sum to 5 categories reduces informativeness it also removes empty cells. Table 4.6 and Figures 4.19 and 4.20 show the data for the 5 x 5 table, and the corresponding results for the smaller table are found in Tables 4.7(a)-(c). Here the multivariate MMSE methods are now feasible and corresponding MMSE results are also shown for the standardisation approach, so that in addition to the methods above we have:-

5) Multivariate Minimum Mean Square Error (MV MMSE) with 4 parameters.

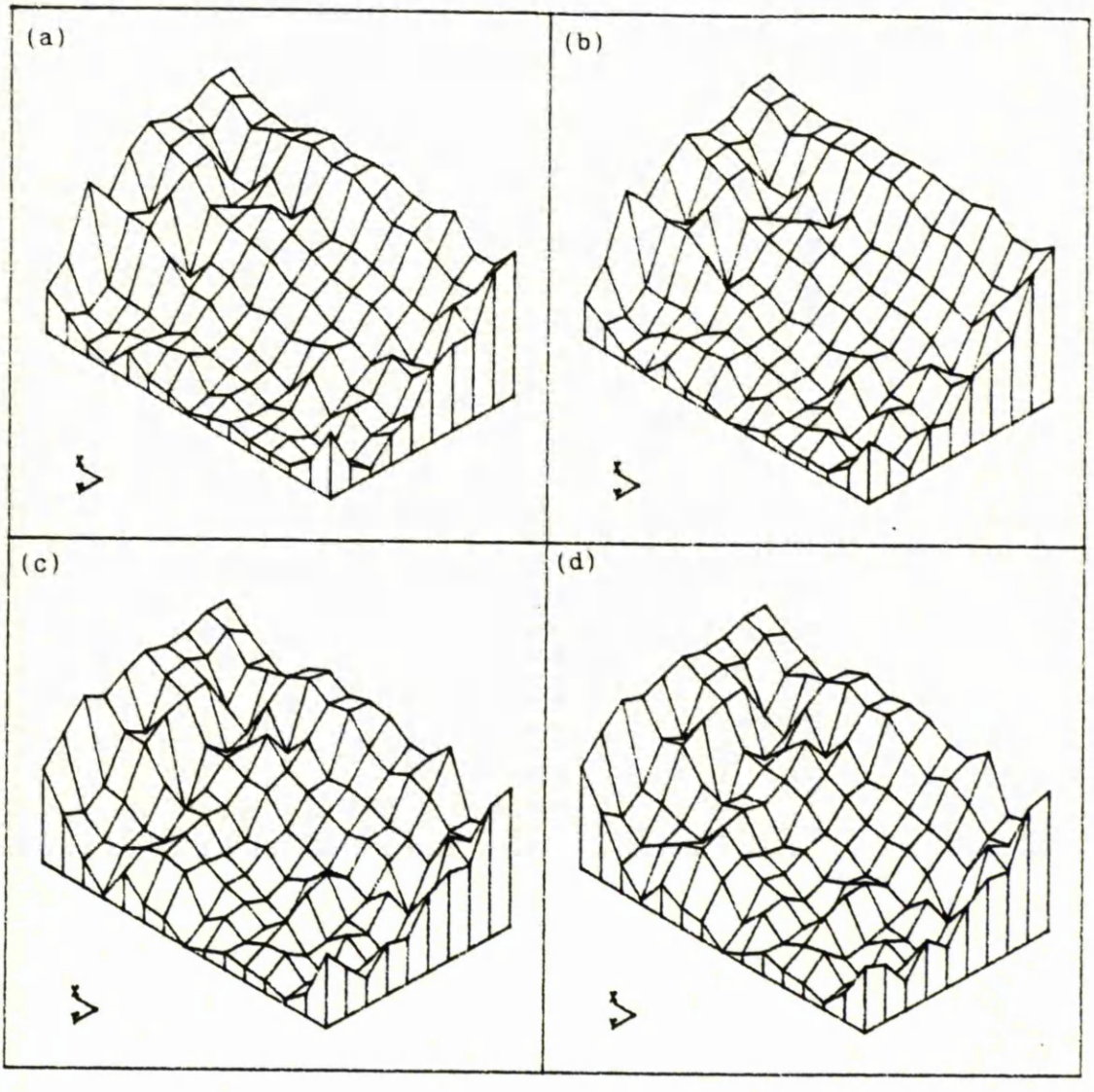


Figure 4.16 $\hat{p}(\pi_1 \mid \underline{x})$

(a) MV XVAL with 4 parameters (6).

(b) MV XVAL with 2 parameters (8).

(c) STD XVAL with 2 parameters (10).

(d) STD XVAL with 1 parameter (12).

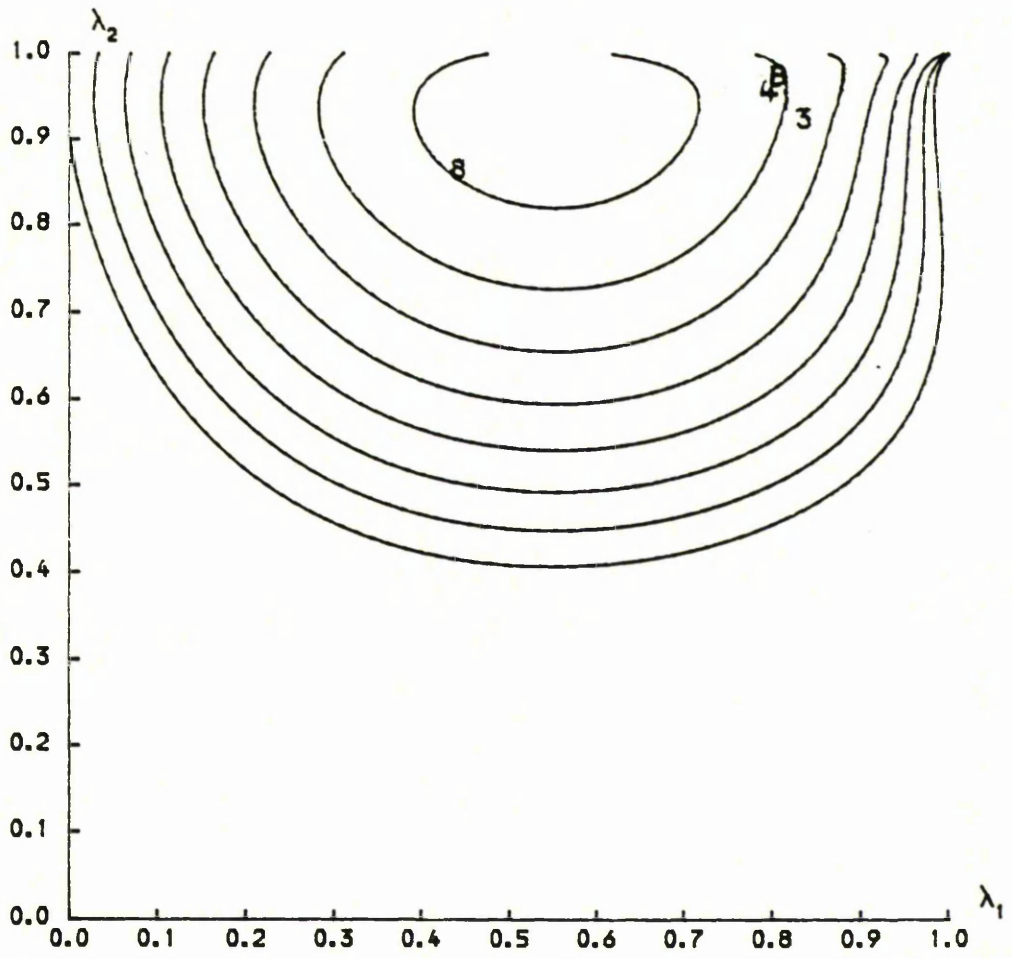


Figure 4.17 Contour plot of the test Brier score as a function of the smoothing parameters (λ_1, λ_2), where λ_1 and λ_2 are associated with Age and Coma Sum respectively in each population. Symbol B denotes method 14. Methods are as in Tables 4.5(a) and (b). Contour heights are :- .824, .822, .820, .818, .816, .814, .812 and .810.

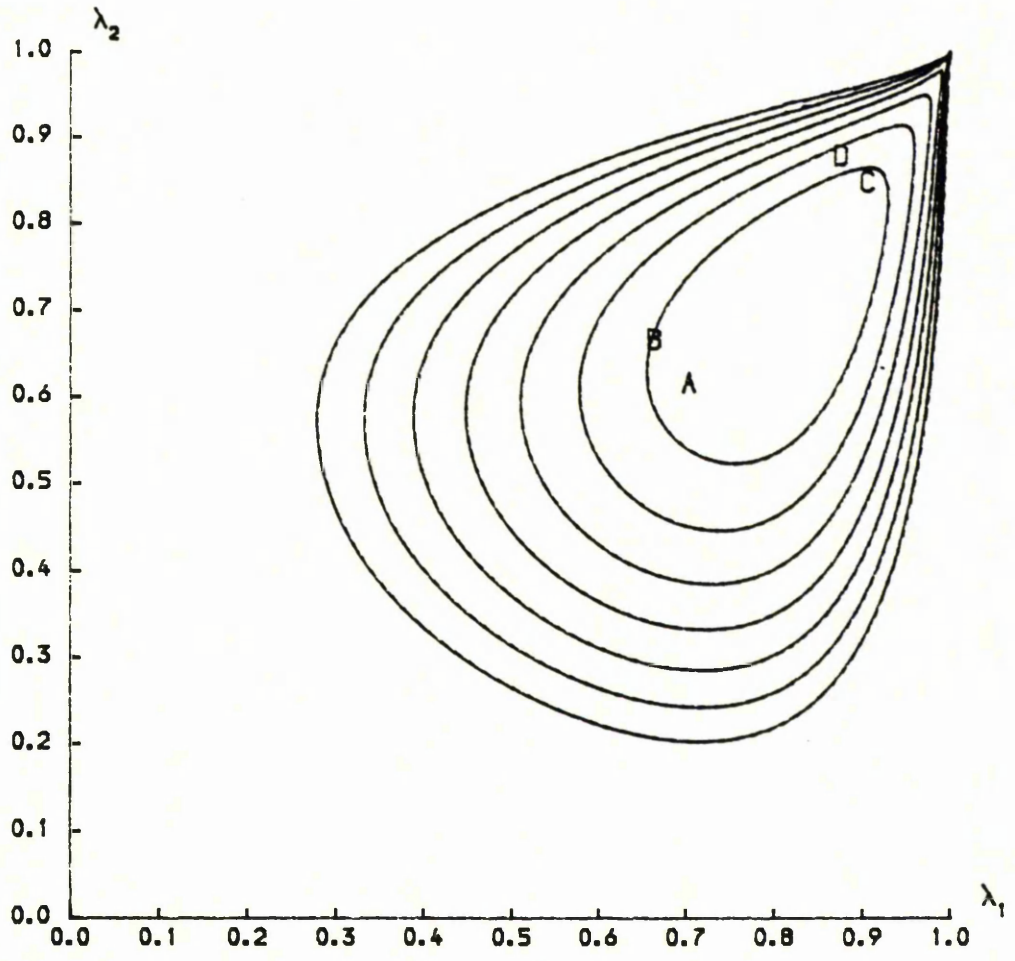


Figure 4.18 Contour plot of the test Brier score as a function of the smoothing parameters (λ_1, λ_2), where λ_i is associated with Age in the ith population. Symbols A, B, C and D denote methods 10, 12, 15 and 16 respectively (see Tables 4.5(a) and (b)). Contour heights are :- .820, .818, .816, .814, .812, .810 and .808.

Table 4.6 Raw data for the 5 x 5 table

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Raw training proportions	0-14	.667	.367	.294	.111	.250
	15-29	.833	.444	.234	.071	.000
	30-44	1.000	.607	.333	.000	.100
	45-59	1.000	.680	.481	.500	.143
	≥60	.923	.923	.966	.571	.000
No. of cases in each cell		9	30	34	9	4
		24	54	47	14	4
	"	17	28	24	4	10
		20	25	27	2	7
		13	26	29	7	4
Raw test proportions		.667	.286	.304	.111	.000
		.808	.491	.167	.111	.333
	"	.833	.706	.322	.000	.000
		.889	.591	.364	.500	.100
		1.000	.962	.842	.600	.333
No. of cases in each cell		12	35	23	9	3
		26	57	48	9	6
	"	18	34	31	9	6
		27	22	22	6	10
		10	26	19	5	3

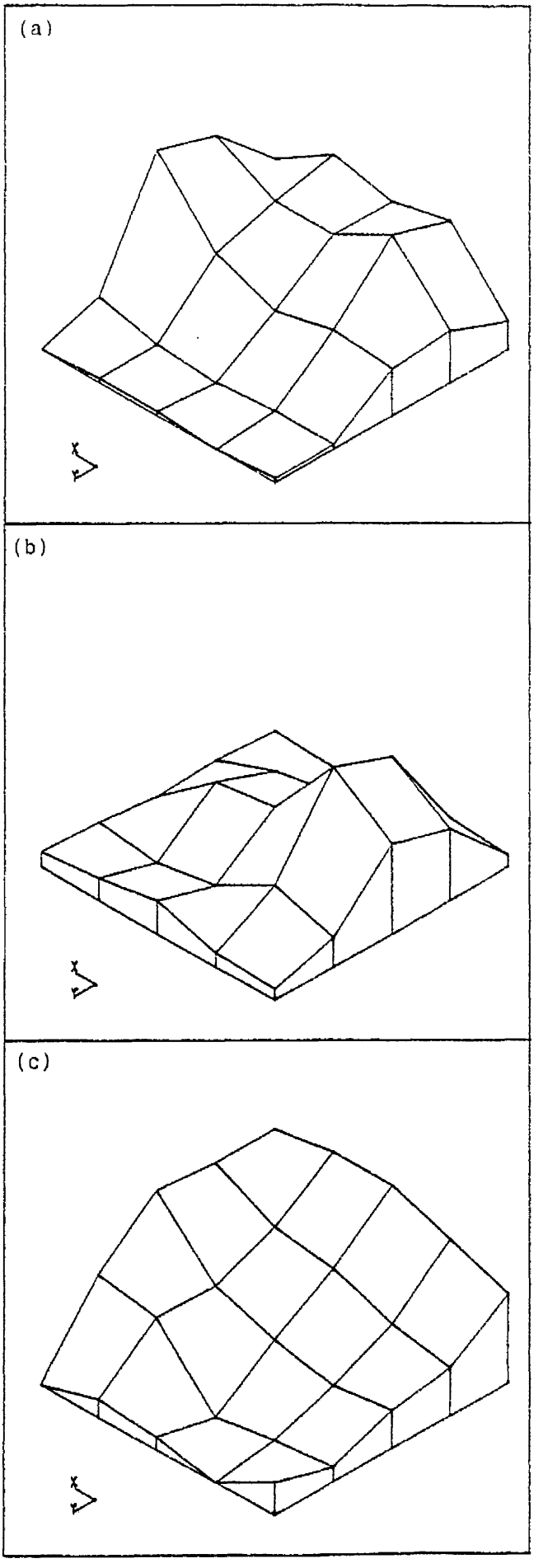


Figure 4.19 Relative frequencies displayed as 3-dimensional isometric projections for the training data and the 5x5 table.

(a) $p(\underline{x} \mid \pi_1)$

(b) $p(\underline{x} \mid \pi_2)$

(c) $p(\pi_1 \mid \underline{x})$

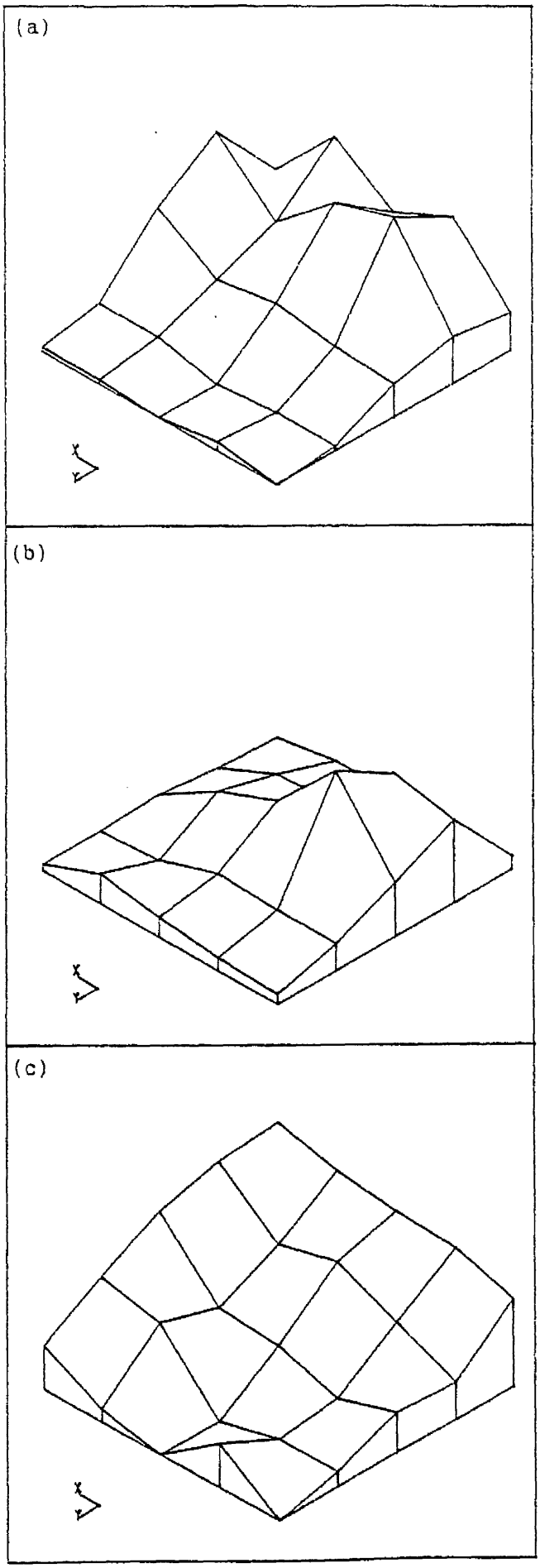


Figure 4.20 Relative frequencies displayed as 3-dimensional isometric projections for the test data and the 5x5 table.

(a) $p(\underline{x} \mid \pi_1)$

(b) $p(\underline{x} \mid \pi_2)$

(c) $p(\pi_1 \mid \underline{x})$

Table 4.7(a) RESULTS FOR THE 5 X 5 TABLE

Method	Smoothing Parameters		Scores	
	π_1 Age Coma	π_2 Sum	Brier Score	Error Rate (%)
Marginal methods				
1 MMSE - 4 parameters	.906 .970	.947 .956	.82453	25.2
2 XVAL - 4 parameters	.871 .971	.931 .950	.82443	25.2
3 MMSE - 2 parameters	.926 .963	.926 .963	.82464	25.2
4 XVAL - 2 parameters	.901 .960	.901 .960	.82458	25.2
Multivariate methods				
5 MV MMSE - 4 parameters	.850 .941	.910 .916	.82439	26.5
6 MV XVAL - 4 parameters	.702 .942	.862 .901	.82340	26.5
7 MV MMSE - 2 parameters	.885 .929	.885 .929	.82451	25.2
8 MV XVAL - 2 parameters	.805 .928	.805 .928	.82393	25.2
9 STD MMSE - 2 parameters	.903 .903	.913 .913	.82433	25.2
10 STD XVAL - 2 parameters	.903 .903	.883 .883	.82441	25.2
11 STD MMSE - 1 parameter	.908 .908	.908 .908	.82446	25.2
12 STD XVAL - 1 parameter	.895 .895	.895 .895	.82425	25.2

Table 4.7(b) RESULTS FOR THE 5 X 5 TABLE

Method	Smoothing Parameters		Scores	
	π_1 Age Coma Sum	π_2	Brier Score	Error Rate (%)
Direct methods				
13 MV BRIER SCORE - 4 parameters	.963 .956	.941 .961	.82474	25.2
14 MV BRIER SCORE - 2 parameters	.945 .958	.945 .958	.82469	25.2
15 STD BRIER SCORE - 2 parameters	.957 .957	.951 .951	.82474	25.2
16 STD BRIER SCORE - 1 parameter	.956 .956	.956 .956	.82470	25.2
ISOTONIC REGRESSION			.82548	25.2

Table 4.7(c) REFERENCE METHODS FOR THE 5 X 5 TABLE

Method	Smoothing Parameters		Scores	
	π_1 Age	π_2 Coma Sum	Brier Score	Error Rate (%)
Reference methods				
Direct optimisation on test set :				
13T MV BRIER SCORE - 4 parameters	.990 .940	.918 .902	.82497	25.2
14T MV BRIER SCORE - 2 parameters	.959 .943	.959 .943	.82473	25.2
15T STD BRIER SCORE - 2 parameters	.952 .952	.913 .913	.82486	25.2
16T STD BRIER SCORE - 1 parameter	.946 .946	.946 .946	.82472	25.2
TEST PROPORTIONS (RFTS)			.83213	25.2
TRAINING PROPORTIONS (RFTR)			.82392	25.2

7) Multivariate Minimum Mean Square Error (MV MMSE) with 2 parameters.

9) Standardisation Minimum Mean Square Error (STD MMSE) with 2 parameters.

11) Standardisation Minimum Mean Square Error (STD MMSE) with 1 parameter.

Tables 4.8(a)-(c) give the predicted probabilities, some of which are also shown in isometric plots, Figures 4.21-4.25. Tables 4.9(a)-(c) and 4.10(a)-(c) contain the corresponding class conditional density estimates for the kernel methods.

Error rates are not in fact seriously affected, and, surprisingly perhaps, fractionally improved. The smaller table retains the advantage over use of age alone. Each marginal method (all fairly comparable both in terms of Brier score and fitted distributions) improves slightly over use of the 15 x 10 table, whereas the multivariate XVAL methods are slightly poorer on Brier score. Collapsing age indicates rather less smoothing, as would be expected, whereas the Coma Sum requires only a little less smoothing than previously.

In each case MMSE outperforms XVAL on Brier score, though the former smooths less. Both multivariate methods 5 and 7 are noticeably less smooth on age, best seen in π_1, than 6 and 8, with slight differences in predicted probabilities. (See Figure 4.21 for the best method of each type, and 4.22 and 4.25(b) for the predicted probabilities.) For both MMSE and XVAL, 2-parameter methods are superior to 4-parameter ones and marginal methods are superior to multivariate ones, though this is less marked for MMSE than XVAL as methods 5 and 7 produce similar fitted distributions. Compared to the marginal methods, 7 especially is a little smoother in π_1 but the differences are small, whereas again 2 and 4 are noticeably less smooth than 6 and 8, of which 8 is smoother (see Figure 4.22).

Now both Brier score methods outperform the other data-based methods slightly, although marginal 2-D MMSE (3) is close to multivariate 2-D BRIER (14), and are near to optimal, (c.f., for the better Brier method, Figures 4.23(a)-(c) and 4.23(d)-(f)), especially in the 2-dimensional case. Figure 4.24 compares the optimal 2- and 4-dimensional methods.

Since we are now using a square table the standardisation

Table 4.8(a) 5 x 5 table : $\hat{p}(\pi_1|x)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Marginal methods						
1) MMSE 4-D	0-14	.669	.378	.299	.128	.238
	15-29	.815	.448	.251	.099	.063
	30-44	.974	.608	.364	.108	.113
	45-59	.983	.677	.493	.466	.160
	≥60	.921	.904	.937	.574	.078
2) XVAL 4-D		.674	.382	.300	.127	.232
		.813	.449	.256	.100	.067
	"	.970	.609	.372	.117	.113
		.980	.677	.496	.449	.158
		.920	.898	.928	.565	.079
3) MMSE 2-D		.671	.378	.300	.134	.246
		.820	.450	.252	.105	.068
	"	.974	.600	.354	.111	.117
		.983	.671	.488	.464	.165
		.921	.899	.930	.583	.090
4) XVAL 2-D		.678	.382	.302	.136	.243
		.820	.452	.256	.109	.075
	"	.969	.598	.359	.122	.118
		.979	.668	.488	.449	.166
		.921	.891	.919	.578	.096

Continued.

Table 4.8(a) cont'd. 5 x 5 table : $\hat{p}(\pi_1|x)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Multivariate methods						
5) MV MMSE 4-D	0-14	.664	.385	.303	.143	.235
	15-29	.799	.450	.264	.122	.101
	30-44	.953	.607	.382	.167	.128
	45-59	.968	.675	.501	.457	.179
	≥60	.918	.890	.916	.581	.138
6) MV XVAL 4-D		.687	.399	.305	.142	.214
		.793	.452	.281	.127	.117
	"	.942	.615	.415	.203	.129
		.958	.678	.512	.434	.173
		.915	.867	.884	.541	.141
7) MV MMSE 2-D		.665	.384	.304	.154	.248
		.806	.453	.264	.133	.111
	"	.953	.596	.367	.173	.133
		.969	.666	.493	.464	.188
		.919	.885	.909	.596	.157
8) MV XVAL 2-D		.690	.397	.310	.158	.239
		.811	.461	.278	.141	.120
	"	.942	.590	.379	.190	.137
		.960	.657	.493	.428	.185
		.917	.859	.872	.575	.159
9) STD MMSE 2-D		.650	.378	.302	.166	.258
		.796	.448	.263	.148	.135
	"	.951	.597	.370	.205	.144
		.967	.669	.496	.505	.206
		.918	.892	.919	.618	.201
10) STD XVAL 2-D		.640	.384	.308	.165	.250
		.787	.456	.269	.147	.128
	"	.934	.594	.369	.193	.146
		.956	.665	.498	.461	.207
		.916	.884	.906	.616	.204
11) STD MMSE 1-D		.647	.380	.304	.162	.255
		.795	.450	.263	.143	.127
	"	.948	.597	.368	.193	.142
		.966	.668	.496	.489	.203
		.918	.892	.917	.616	.192
12) STD XVAL 1-D		.645	.382	.305	.170	.256
		.789	.451	.268	.153	.140
	"	.941	.595	.372	.211	.148
		.960	.667	.498	.489	.212
		.916	.887	.910	.620	.215

Table 4.8(b) 5 x 5 table : $\hat{p}(\pi_1|x)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Direct methods						
13) MV BRIER 4-D	0-14	.660	.374	.300	.138	.254
	15-29	.818	.449	.249	.109	.072
	30-44	.975	.598	.346	.110	.120
	45-59	.984	.670	.486	.482	.172
	⩾60	.921	.904	.937	.597	.103
14) MV BRIER 2-D	"	.662	.374	.299	.136	.250
		.816	.448	.250	.107	.072
		.974	.601	.352	.113	.118
		.983	.673	.488	.482	.170
		.921	.905	.938	.592	.099
15) STD BRIER 2-D	"	.655	.374	.299	.136	.250
		.813	.448	.249	.107	.071
		.973	.601	.350	.110	.119
		.982	.674	.489	.480	.172
		.921	.907	.940	.595	.102
16) STD BRIER 1-D	"	.657	.373	.298	.137	.251
		.815	.447	.248	.108	.073
		.975	.602	.351	.113	.119
		.984	.674	.489	.491	.172
		.921	.908	.942	.596	.103
ISOTONIC REGRESSION	"	.667	.367	.259	.089	.089
		.833	.444	.259	.089	.089
		.980	.607	.333	.089	.089
		.980	.680	.483	.483	.091
		.980	.945	.945	.571	.091

Table 4.8(c) 5 x 5 table : $\hat{p}(\pi_1|x)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Reference methods						
13T) MV BRIER 4-D	0-14	.629	.378	.308	.144	.246
	15-29	.798	.458	.256	.118	.079
	30-44	.948	.592	.342	.115	.129
	45-59	.966	.666	.490	.430	.186
	≥60	.918	.899	.926	.612	.137
14T) MV BRIER 2-D	"	.649	.372	.299	.143	.254
		.808	.447	.250	.116	.087
		.970	.602	.353	.133	.124
		.981	.675	.491	.502	.181
		.920	.908	.942	.606	.128
15T) STD BRIER 2-D	"	.643	.381	.307	.137	.239
		.802	.457	.257	.110	.071
		.952	.596	.351	.110	.124
		.968	.668	.492	.420	.176
		.919	.897	.923	.595	.112
16T) STD BRIER 1-D	"	.655	.374	.299	.142	.252
		.811	.448	.251	.115	.085
		.970	.601	.354	.132	.124
		.980	.673	.490	.490	.178
		.920	.905	.937	.600	.122

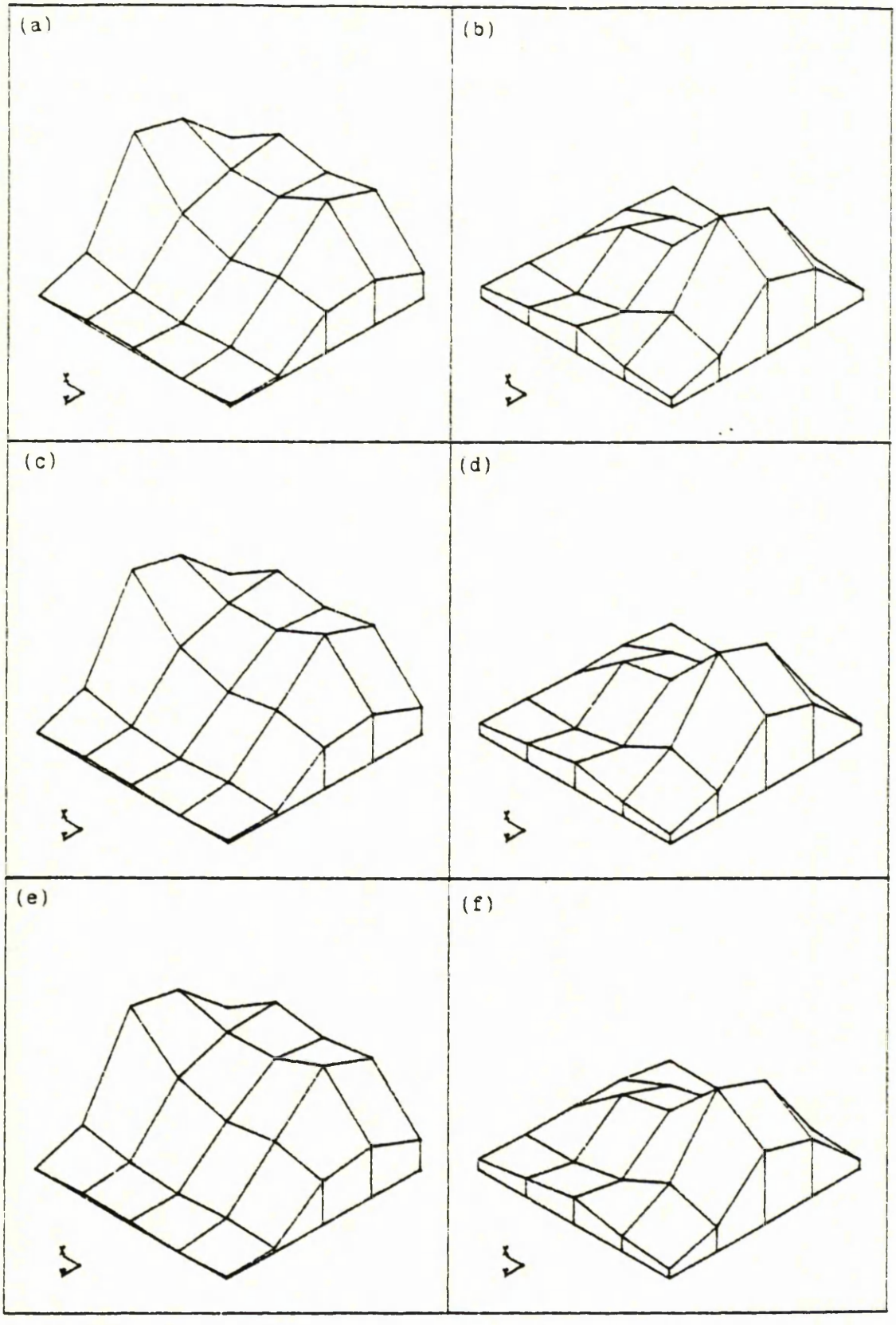


Figure 4.21 Fitted distributions.

(a) $\hat{p}(\underline{x} \mid \pi_1)$ for MMSE with 2 parameters (3).

(b) $\hat{p}(\underline{x} \mid \pi_2)$ for (3).

(c) $\hat{p}(\underline{x} \mid \pi_1)$ for MV MMSE with 2 parameters (7).

(d) $\hat{p}(\underline{x} \mid \pi_2)$ for (7).

(e) $\hat{p}(\underline{x} \mid \pi_1)$ for MV XVAL with 2 parameters (8).

(f) $\hat{p}(\underline{x} \mid \pi_2)$ for (8).

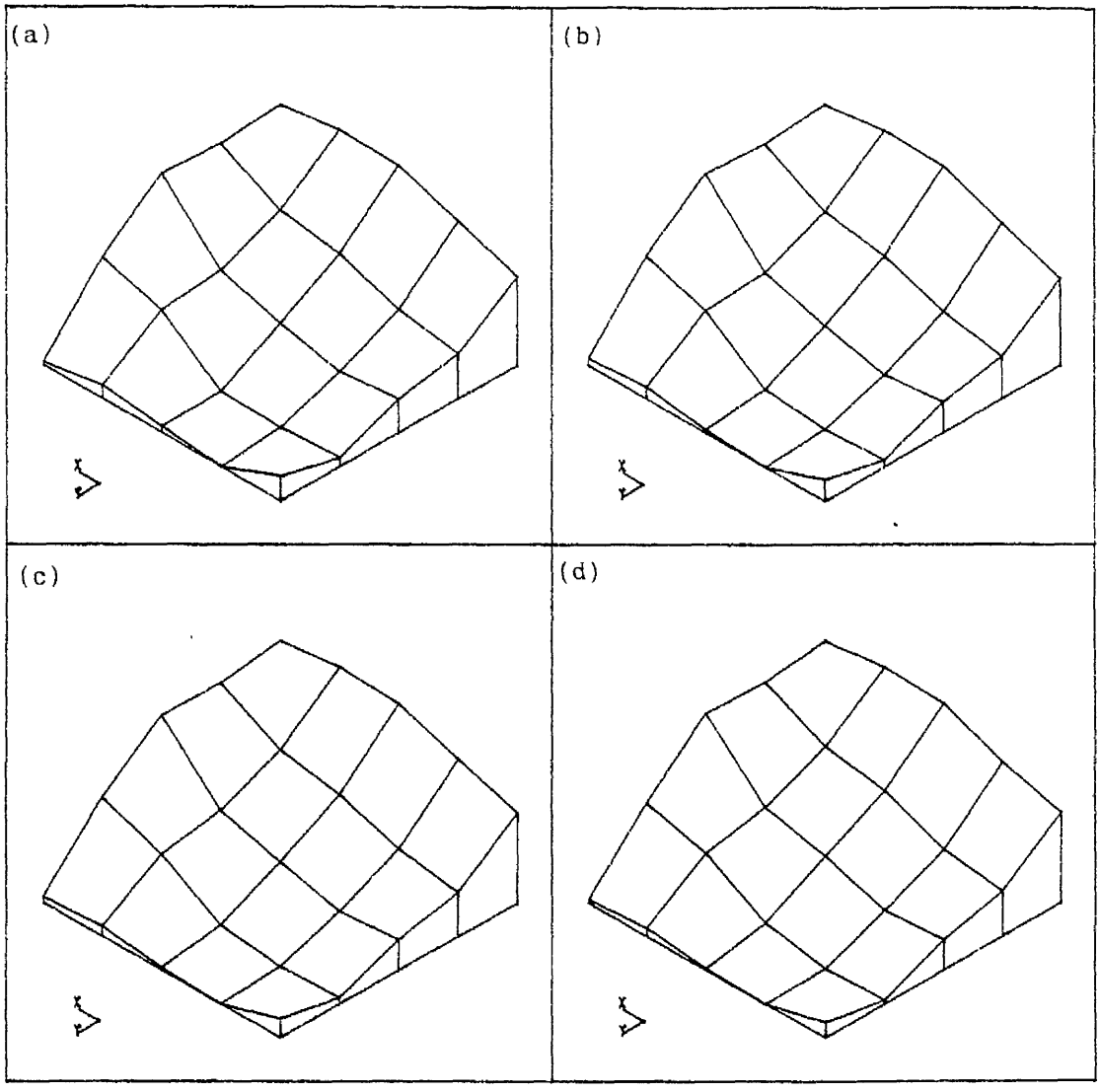


Figure 4.22 $\hat{p}(\pi_1 \mid \underline{x})$

(a) XVAL with 2 parameters (4).

(b) MV MMSE with 2 parameters (7).

(c) MV XVAL with 2 parameters (8).

(d) MV XVAL with 4 parameters (6).

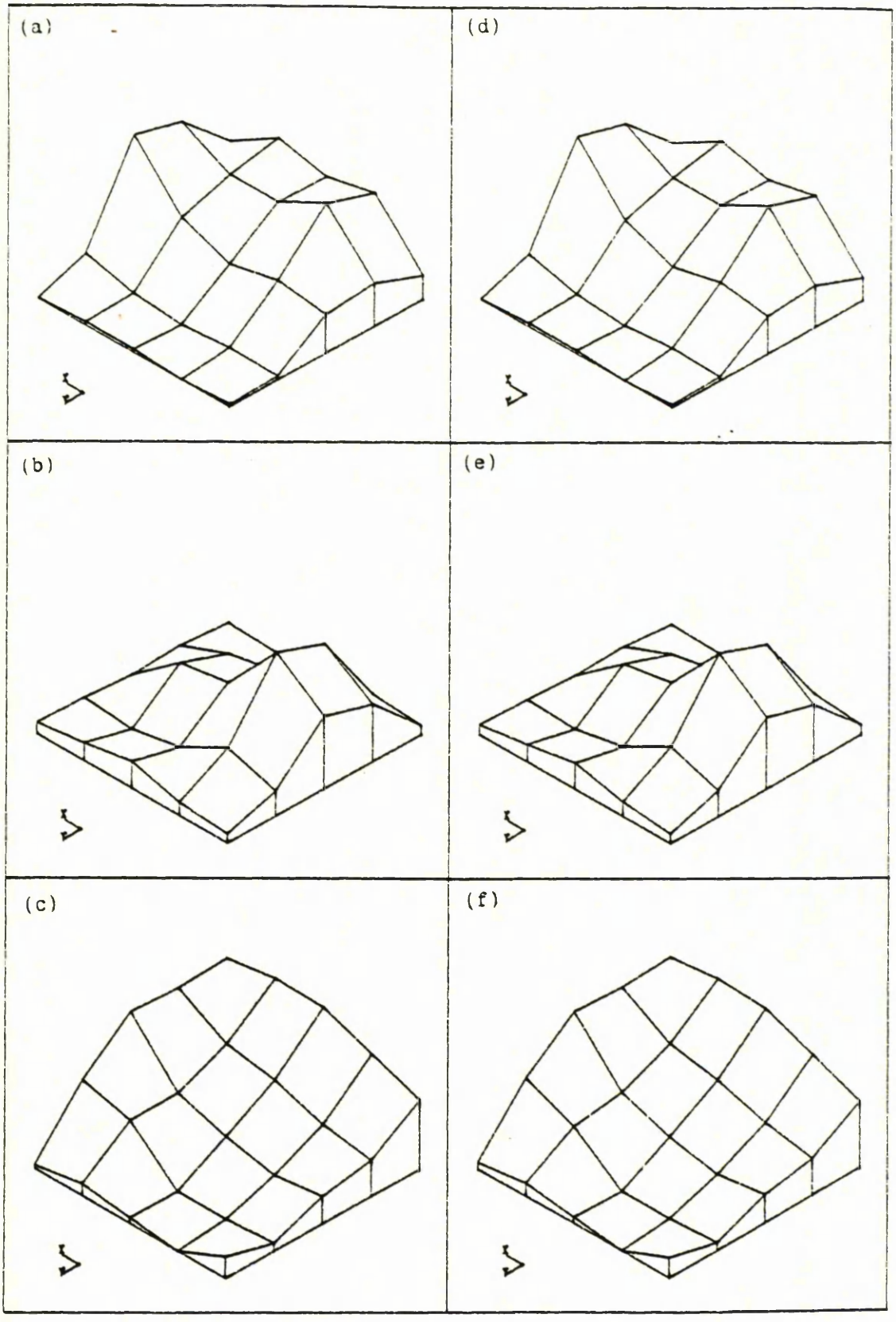


Figure 4.23 Fitted distributions.

(a) $\hat{p}(\underline{x} \mid \pi_1)$ for MV BRIER with 4 parameters (13).

(b) $\hat{p}(\underline{x} \mid \pi_2)$ for (13).

(c) $\hat{p}(\pi_1 \mid \underline{x})$ for (13).

(d) $\hat{p}(\underline{x} \mid \pi_1)$ optimising on the test data (13T).

(e) $\hat{p}(\underline{x} \mid \pi_2)$ for (13T).

(f) $\hat{p}(\pi_1 \mid \underline{x})$ for (13T).

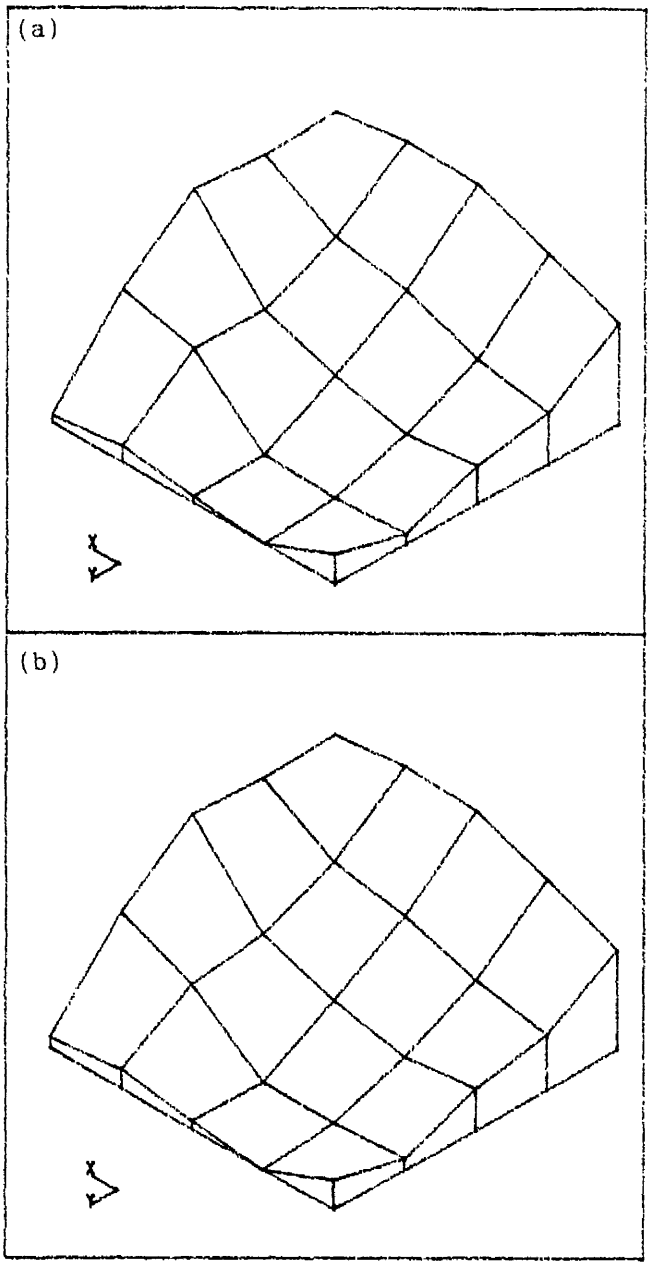


Figure 4.24 $\hat{p}(\pi_1 \mid \underline{x})$

(a) MV BRIER with 2 parameters optimised on the test data (14T)

(b) MV BRIER with 4 parameters optimised on the test data (13T).

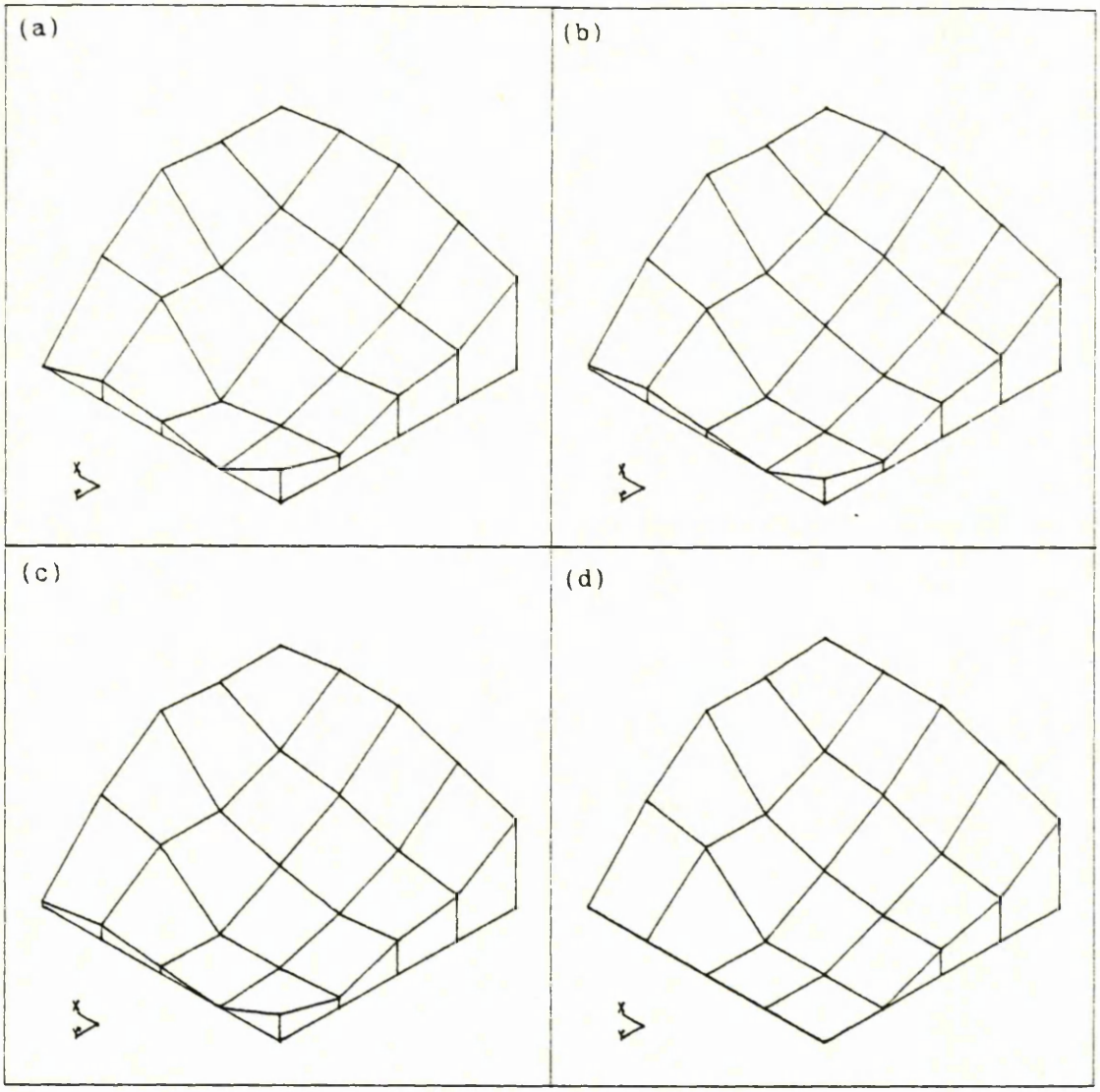


Figure 4.25 $\hat{p}(\pi_1 \mid \underline{x})$

(a) Unsmoothed training relative frequencies.

(b) MMSE with 2 parameters (3).

(c) MV BRIER with 4 parameters (13).

(d) Isotonic regression.

Table 4.9(a) 5 x 5 table : $\hat{p}(x|\pi_1)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Marginal methods						
1) MMSE 4-D	0-14	.0261	.0446	.0394	.0048	.0041
	15-29	.0775	.0935	.0465	.0058	.0012
	30-44	.0685	.0710	.0374	.0022	.0045
	45-59	.0778	.0704	.0545	.0059	.0046
	≥60	.0478	.0909	.1034	.0164	.0013
2) XVAL 4-D		.0268	.0450	.0392	.0047	.0040
		.0769	.0929	.0471	.0058	.0013
	"	.0690	.0722	.0390	.0025	.0044
		.0774	.0713	.0553	.0060	.0045
		.0476	.0891	.1005	.0158	.0013
3) MMSE 2-D		.0256	.0443	.0394	.0050	.0043
		.0775	.0936	.0463	.0060	.0013
	"	.0679	.0700	.0364	.0023	.0046
		.0777	.0698	.0540	.0062	.0047
		.0479	.0917	.1048	.0171	.0016
4) XVAL 2-D		.0261	.0445	.0392	.0050	.0042
		.0770	.0930	.0467	.0062	.0015
	"	.0681	.0709	.0377	.0027	.0046
		.0773	.0704	.0546	.0064	.0048
		.0477	.0903	.1025	.0168	.0017

Continued.

Table 4.9(a) cont'd. 5 x 5 table : $\hat{p}(x|\pi_1)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Multivariate methods						
5) MV MMSE 4-D	0-14	.0270	.0446	.0387	.0055	.0043
	15-29	.0752	.0911	.0479	.0073	.0022
	30-44	.0679	.0721	.0404	.0041	.0050
	45-59	.0756	.0712	.0557	.0076	.0051
	≥60	.0473	.0870	.0974	.0170	.0024
6) MV XVAL 4-D		.0302	.0460	.0380	.0054	.0039
		.0727	.0884	.0503	.0075	.0028
	"	.0700	.0773	.0475	.0055	.0049
		.0739	.0749	.0589	.0082	.0049
		.0465	.0796	.0854	.0146	.0025
7) MV MMSE 2-D		.0262	.0440	.0388	.0059	.0046
		.0752	.0912	.0475	.0079	.0024
	"	.0669	.0704	.0388	.0043	.0052
		.0754	.0701	.0549	.0080	.0055
		.0474	.0884	.0996	.0182	.0029
8) MV XVAL 2-D		.0278	.0448	.0384	.0059	.0044
		.0738	.0897	.0488	.0080	.0028
	"	.0680	.0732	.0427	.0051	.0052
		.0745	.0721	.0566	.0084	.0054
		.0470	.0844	.0932	.0170	.0029
9) STD MMSE 2-D		.0256	.0434	.0385	.0066	.0049
		.0743	.0903	.0476	.0091	.0031
	"	.0655	.0690	.0383	.0052	.0057
		.0743	.0692	.0545	.0092	.0061
		.0474	.0884	.0998	.0199	.0038
10) STD XVAL 2-D		.0256	.0434	.0386	.0066	.0049
		.0743	.0903	.0476	.0091	.0031
	"	.0655	.0690	.0383	.0052	.0057
		.0743	.0692	.0545	.0092	.0061
		.0474	.0885	.0998	.0199	.0038
11) STD MMSE 1-D		.0256	.0434	.0386	.0064	.0049
		.0747	.0907	.0474	.0088	.0029
	"	.0657	.0690	.0380	.0049	.0056
		.0747	.0692	.0544	.0089	.0060
		.0475	.0889	.1005	.0197	.0036
12) STD XVAL 1-D		.0257	.0433	.0384	.0068	.0050
		.0738	.0898	.0478	.0096	.0034
	"	.0652	.0691	.0388	.0057	.0058
		.0738	.0692	.0546	.0096	.0063
		.0473	.0878	.0987	.0202	.0041

Table 4.9(b) 5 x 5 table: $\hat{p}(x|\pi_1)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Direct methods						
13) MV BRIER 4-D	0-14	.0247	.0438	.0395	.0052	.0044
	15-29	.0778	.0939	.0458	.0064	.0014
	39-44	.0670	.0685	.0347	.0023	.0048
	45-59	.0778	.0687	.0532	.0063	.0050
	≥60	.0480	.0933	.1074	.0180	.0018
14) MV BRIER 2-D		.0251	.0440	.0394	.0051	.0044
		.0776	.0937	.0460	.0063	.0014
	"	.0674	.0692	.0356	.0023	.0047
		.0777	.0692	.0536	.0063	.0049
		.0480	.0925	.1060	.0176	.0017
15) STD BRIER 2-D		.0249	.0439	.0395	.0052	.0044
		.0778	.0938	.0458	.0063	.0014
	"	.0672	.0688	.0350	.0023	.0047
		.0778	.0689	.0534	.0063	.0049
		.0480	.0930	.1069	.0179	.0018
16) STD BRIER 1-D		.0249	.0439	.0395	.0052	.0044
		.0777	.0938	.0459	.0064	.0014
	"	.0672	.0688	.0351	.0023	.0047
		.0777	.0689	.0534	.0064	.0049
		.0480	.0930	.1068	.0179	.0018

Table 4.9(c) 5 x 5 table: $\hat{p}(x|\pi_1)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Reference methods						
13T) MV BRIER 4-D	0-14	.0241	.0433	.0394	.0056	.0047
	15-29	.0776	.0937	.0456	.0071	.0018
	30-44	.0659	.0671	.0336	.0027	.0050
	45-59	.0773	.0678	.0526	.0070	.0054
	⩾60	.0481	.0941	.1087	.0193	.0024
14T) MV BRIER 2-D		.0248	.0436	.0393	.0055	.0046
	"	.0772	.0932	.0460	.0070	.0018
		.0665	.0683	.0351	.0029	.0050
		.0771	.0686	.0533	.0070	.0052
		.0480	.0927	.1064	.0186	.0023
15T) STD BRIER 2-D		.0249	.0439	.0394	.0053	.0045
	"	.0775	.0936	.0460	.0065	.0015
		.0670	.0688	.0353	.0025	.0048
		.0775	.0689	.0535	.0065	.0050
		.0480	.0927	.1063	.0180	.0019
16T) STD BRIER 1-D		.0250	.0438	.0393	.0054	.0045
	"	.0771	.0932	.0462	.0069	.0017
		.0668	.0688	.0357	.0028	.0049
		.0771	.0690	.0536	.0069	.0052
		.0479	.0921	.1055	.0182	.0022
Raw training proportions (RFTR)		.0242	.0444	.0403	.0040	.0040
	"	.0806	.0968	.0444	.0040	.0000
		.0685	.0685	.0322	.0000	.0040
		.0806	.0685	.0524	.0040	.0040
		.0484	.0968	.1129	.0161	.0000
Raw test proportions (RFTS)		.0335	.0418	.0293	.0042	.0000
	"	.0879	.1172	.0335	.0042	.0084
		.0628	.1004	.0418	.0000	.0000
		.1004	.0544	.0335	.0126	.0042
		.0418	.1046	.0669	.0126	.0042

Table 4.10(a) 5 x 5 table : $\hat{p}(x|\pi_2)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Marginal methods						
1) MMSE 4-D	0-14	.0143	.0815	.1024	.0361	.0146
	15-29	.0194	.1278	.1533	.0580	.0202
	30-44	.0020	.0507	.0723	.0204	.0387
	45-59	.0015	.0371	.0621	.0075	.0264
	⩾60	.0045	.0107	.0077	.0134	.0173
2) XVAL 4-D		.0143	.0807	.1013	.0360	.0148
		.0195	.1263	.1516	.0576	.0206
	"	.0024	.0513	.0728	.0210	.0384
		.0018	.0376	.0622	.0082	.0264
		.0046	.0112	.0086	.0135	.0172
3) MMSE 2-D		.0139	.0809	.1018	.0356	.0145
		.0188	.1266	.1522	.0570	.0201
	"	.0020	.0517	.0735	.0206	.0386
		.0015	.0379	.0629	.0079	.0265
		.0045	.0114	.0087	.0135	.0174
4) XVAL 2-D		.0138	.0799	.1005	.0352	.0146
		.0186	.1247	.1501	.0562	.0204
	"	.0024	.0527	.0745	.0212	.0382
		.0018	.0388	.0634	.0087	.0265
		.0045	.0122	.0100	.0136	.0173

Continued.

Table 4.10(a) cont'd. 5 x 5 table : $\hat{p}(x|\pi_2)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Multivariate methods						
5) MV MMSE 4-D	0-14	.0152	.0788	.0986	.0366	.0156
	15-29	.0209	.1230	.1475	.0582	.0220
	30-44	.0037	.0517	.0725	.0226	.0377
	45-59	.0027	.0380	.0614	.0100	.0261
	≥60	.0047	.0119	.0099	.0136	.0169
6) MV XVAL 4-D	"	.0152	.0766	.0956	.0362	.0160
		.0210	.1189	.1428	.0572	.0230
		.0048	.0535	.0740	.0241	.0369
		.0036	.0395	.0620	.0118	.0260
		.0048	.0135	.0124	.0138	.0167
7) MV MMSE 2-D	"	.0146	.0783	.0981	.0358	.0154
		.0200	.1219	.1465	.0569	.0218
		.0036	.0529	.0741	.0227	.0376
		.0027	.0390	.0625	.0102	.0262
		.0046	.0127	.0111	.0136	.0170
8) MV XVAL 2-D	"	.0139	.0754	.0945	.0345	.0155
		.0190	.1163	.1405	.0542	.0224
		.0046	.0562	.0775	.0243	.0366
		.0035	.0416	.0645	.0125	.0264
		.0047	.0153	.0151	.0139	.0170
9) STD MMSE 2-D	"	.0153	.0789	.0986	.0368	.0157
		.0211	.1231	.1475	.0585	.0222
		.0037	.0515	.0722	.0226	.0377
		.0028	.0379	.0612	.0100	.0260
		.0047	.0118	.0098	.0136	.0168
10) STD XVAL 2-D	"	.0159	.0768	.0957	.0371	.0164
		.0222	.1194	.1430	.0586	.0235
		.0051	.0523	.0725	.0243	.0369
		.0038	.0386	.0608	.0119	.0258
		.0048	.0128	.0114	.0137	.0165
11) STD MMSE 1-D	"	.0154	.0786	.0981	.0368	.0158
		.0213	.1225	.1468	.0585	.0224
		.0039	.0517	.0723	.0229	.0376
		.0029	.0380	.0612	.0103	.0260
		.0047	.0120	.0100	.0136	.0168
12) STD XVAL 1-D	"	.0157	.0776	.0968	.0370	.0161
		.0218	.1208	.1448	.0586	.0230
		.0046	.0520	.0724	.0236	.0372
		.0034	.0383	.0610	.0111	.0259
		.0048	.0124	.0108	.0136	.0166

Table 4.10(b) 5 x 5 table: $\hat{p}(x|\pi_2)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Direct methods						
13) MV BRIER 4-D	0-14	.0141	.0813	.1023	.0359	.0145
	15-29	.0191	.1275	.1531	.0576	.0201
	30-44	.0019	.0511	.0727	.0204	.0387
	45-59	.0014	.0374	.0624	.0076	.0264
	≥60	.0045	.0109	.0080	.0135	.0174
14) MV BRIER 2-D	"	.0142	.0814	.1024	.0360	.0146
		.0193	.1277	.1533	.0578	.0201
		.0020	.0509	.0724	.0204	.0387
		.0014	.0372	.0622	.0075	.0264
		.0045	.0108	.0078	.0135	.0173
15) STD BRIER 2-D	"	.0145	.0814	.1023	.0363	.0147
		.0197	.1278	.1532	.0583	.0203
		.0021	.0505	.0719	.0205	.0387
		.0016	.0370	.0618	.0076	.0263
		.0046	.0106	.0075	.0134	.0172
16) STD BRIER 1-D	"	.0144	.0818	.1028	.0363	.0146
		.0196	.1284	.1539	.0583	.0201
		.0019	.0504	.0719	.0203	.0388
		.0014	.0368	.0618	.0073	.0264
		.0045	.0104	.0072	.0134	.0173

Table 4.10(c) 5 x 5 table: $\hat{p}(x|\pi_2)$

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Reference methods						
13T) MV BRIER 4-D	0-14	.0157	.0787	.0982	.0372	.0159
	15-29	.0218	.1228	.1470	.0591	.0225
	30-44	.0040	.0512	.0716	.0229	.0376
	45-59	.0030	.0376	.0607	.0102	.0259
	≥60	.0047	.0117	.0096	.0136	.0167
14T) MV BRIER 2-D	"	.0148	.0815	.1022	.0367	.0149
		.0202	.1279	.1532	.0589	.0206
		.0022	.0501	.0713	.0206	.0387
		.0017	.0366	.0613	.0076	.0262
		.0046	.0103	.0072	.0134	.0172
15T) STD BRIER 2-D	"	.0153	.0789	.0986	.0368	.0157
		.0211	.1231	.1475	.0585	.0221
		.0037	.0515	.0722	.0226	.0377
		.0028	.0379	.0612	.0100	.0260
		.0047	.0118	.0098	.0136	.0168
16T) STD BRIER 1-D	"	.0146	.0811	.1018	.0364	.0148
		.0199	.1272	.1525	.0583	.0206
		.0023	.0506	.0720	.0208	.0386
		.0017	.0371	.0617	.0079	.0263
		.0046	.0107	.0078	.0134	.0172
Raw training proportions (RFTR)	"	.0134	.0848	.1071	.0357	.0134
		.0178	.1339	.1607	.0580	.0178
		.0000	.0491	.0714	.0178	.0402
		.0000	.0357	.0625	.0045	.0268
		.0045	.0089	.0045	.0134	.0178
Raw test proportions (RFTS)	"	.0169	.1055	.0675	.0338	.0126
		.0211	.1224	.1688	.0338	.0169
		.0126	.0422	.0886	.0380	.0253
		.0126	.0380	.0591	.0126	.0380
		.0000	.0042	.0126	.0084	.0084

methods set the smoothing parameters to be common across variables and allow comparison of this approach and setting them equal in each population. The standardised methods now have more potential as the optimum Brier scores (Table 4.7(c)) are not far from those of the usual methods and the direct Brier methods 15 and 16 are very similar to 13 and 14 with nearly identical Brier scores and fitted distributions. Both STD XVAL methods (10 and 12) are superior to MV XVAL (6 and 8) in terms of Brier score, but still slightly inferior to the marginal methods (2 and 4) though similar to the latter in terms of their fitted distributions. STD 1-D XVAL (12) is inferior to STD 2-D XVAL (10), but again the fitted distributions are similar. STD 2-D MMSE (9) is inferior to both MV MMSE methods (5 and 7). STD 1-D MMSE (11) which is superior to 2-D (9), is also better than MV 4-D MMSE (5) but worse than 2-D (7), and therefore also worse than the marginal MMSE ones. In terms of the fitted probabilities 9 and 11 are similar and similar to 7, while these only differ slightly from the MV 4-D method 5. Comparing MMSE to XVAL, the 1-parameter methods are virtually identical but now STD 2-D MMSE (9) is inferior (in π_2) to the corresponding XVAL method (10) which produces a slightly smoother predicted distribution.

Despite the removal of convexity, the Brier score achieved by the isotonic method is in fact slightly lower than that of the 15 x 10 case. However, once more, it is overall best, although it is not surprising that with rather less sparse data the isotonic estimate does not improve as much over the kernel methods as with the larger table. Figure 4.25 shows the unsmoothed data, and predicted probabilities for the best Brier method (13), best non-Brier kernel method (3) and the isotonic method.

Conclusions

Overall, we conclude that marginal methods tend to perform similarly and produce similar fitted distributions. Multivariate methods not surprisingly were superior to marginal ones in the sparse table. In either table, the difference between marginal and multivariate methods is more marked for XVAL than MMSE, while standardisation methods of either type tend to be more like marginal methods than the usual multivariate methods. In the smaller table, where less smoothing is required, MMSE was generally superior to XVAL, though the opposite is true for the 1-dimensional

and 15 x 10 examples. Surprisingly, in the smaller table for both the XVAL and MMSE approaches, and in the larger table for the non-standardisation XVAL methods alone, using a common parameter for a given variable in each population is in fact slightly superior to using 4 separate parameters.

Except for the less sparse table, where the degree of smoothing is less crucial and the potential to improve on standard methods is smaller, the data-based Brier methods using cross-validation tend to be disapppointing and can be even poorer than marginal methods. Here 4 parameters rather than 2 are better.

The overall best method is the isotonic approach which alone recovers the required ordering and outperforms or has a near-optimal Brier score compared to the kernel methods, although again for the smaller table no method was far from isotonic.

4.4 SMOOTHING THE ISOTONIC ESTIMATE

The results of Section 4.3 suggest that where response is known to be ordered with respect to the explanatory variables, directly modelling its posterior distribution, rather than indirectly via estimation of the class conditional distributions, is the more appropriate approach. In these examples the isotonic regression method was the only one which completely recovered this underlying ordering. However, a major disadvantage of isotonic regression is that it produces a step function estimate with disappointing continuity properties, unlike the smoother kernel-based nonparametric estimator or a direct parametric method such as logistic regression. While the isotonic method is applying smoothing to the unconstrained MLEs, extra smoothing of some sort would seem to be desirable.

4.4.1 Convex smoothing

In density estimation we have seen (see Section 3.4.2) that the 1-dimensional kernel estimator of Aitchison and Aitken (1976) can be viewed as a convex combination of the data, expressed as relative frequencies, and a vector of probabilities, such convex estimators smoothing the MLE. In a similar spirit we might apply additional smoothing to the isotonic estimates of posterior probabilities $\{p(\pi_1|\underline{x})\}$ by smoothing towards those of a suitable parametric model, such as a logistic or independence model, or a smoother nonparametric model such as the kernel estimator, in order

to achieve a compromise between the ordered but rough nature of the former estimator, and the desirable continuity properties but overly structured nature of the latter ones. It can be shown (see Appendix 3) that the Brier score of a convex model such as λISO + $(1-\lambda)$PARA, $0 \leqslant \lambda \leqslant 1$, where ISO and PARA denote the isotonic and a parametric model respectively, is a quadratic in λ, so that the optimising value of λ is easily found.

Smoothing towards a kernel estimate unfortunately was unable to improve on the Brier score of the isotonic estimate in any of the 3 discrete examples above.

4.4.2 Adding in pseudo-observations

One motivation of smoothed relative frequency estimators is to avoid zero estimates. In the context of log-linear and logistic models, Goodman (1970, 1971) suggested adding a pseudo-count of ½ to every cell when empty cells occur, while Gart and Zweifel (1967) used ½ or 1, adding these either to all cells or to each empty cell. Grizzle, Starmer, and Koch (1969) replaced zero counts by 1/k, k = total number of cells. Following the same procedure for maximum likelihood estimation of multinomial probabilities is equivalent to convex smoothing (Fienberg and Holland, 1973). Goodman's procedure specifies $\alpha = \frac{1}{2}k/(n+\frac{1}{2}k)$ with $\theta(y) = 1/k$, $\forall y$, as in (3.2). More generally, adding a pseudo-count of $\frac{1}{2}k\theta(y)$ to each cell recovers the estimator (3.3), again with $\alpha = \frac{1}{2}k/(n+\frac{1}{2}k)$.

One means of further smoothing the isotonic estimate is to adjust the cell weights $\{w_{ij}\}$. We noted in Section 3.5.2 the difficulties caused by zero-weighted (i.e. empty) cells, and stated that the Brier score could be improved and these difficulties overcome, by setting the cell relative frequencies to the overall relative frequency of π_1 and assigning a non-trivial weight δ to them, $0 < \delta \leqslant 1$. This has the effect of introducing further smoothing in the region of each empty cell, smoothing away from the isotonic regression. An alternative, applicable whether or not zero-weighted cells are present, is to add in a small number of extra observations, λ, to every cell weight and to each population in proportion to $p(\pi_i)$. This is analagous to the type of convex smoothing described above, but is followed here by isotonic regression. It has the effect of biasing all the $\{\hat{p}(\pi_i|x)\}$ towards uniformity, in a Bayesian spirit, the degree of smoothing being controlled by the value of λ. For $\lambda = \delta \approx 0$ both reduce to the

usual isotonic method of Dykstra (1981), assuming that zero-weighted cells are in any case set to $\hat{p}(\pi_i)$. The methods are referred to below as Methods 1 and 2 respectively, while Dykstra's method is referred to as the basic isotonic procedure.

Figure 4.26 shows that in the 15 by 10 example both methods improve the Brier score on the test data over that of Dykstra's standard isotonic method with δ set to 10^{-5}, and, where there are zero-weighted cells, that overall additional smoothing produces a greater improvement than extra smoothing around the empty cells alone. In practice of course the parameters cannot be chosen to optimise performance on the test set and optimising on the training data will lead to an answer of no extra smoothing. Ideally a leaving-one-out procedure is required, but full cross-validation would be too time consuming to compute. Randomly partitioning the design set several times into a "design subset" and somewhat smaller "test subset", and then averaging the resulting parameter estimates may provide an acceptable compromise.

Ten random samples of size 157 were taken from the basic set of 472 observations and used as a test subset, the remaining 315 cases then being used to reformulate the isotonic regression in each case. The choice of δ was made by a rough grid search to optimise the "test" Brier score; candidate values of δ were taken in steps of .1 over the range (0, 1), and (0, 10) for methods 1 and 2 respectively. The average of $\hat{\delta}$ over the 10 simulations was used with the full training set to find the corresponding estimate. The Brier score quoted corresponds to the original test set.

The estimated smoothing parameters and Brier scores for the subsampling methods are given with those for full optimisation on the test set in Tables 4.11(a) and 4.11(b), for the 15 by 10 and 5 by 5 cases respectively. For the latter, since there are no empty cells, method 1 is not an appropriate way to apply extra smoothing.

Even this rather crude version of method 2 produced in the 5 by 5 example a slight improvement in the Brier score over the basic isotonic method, bringing it close to what would be achieved by optimisation on the full test set. The estimated degree of smoothing is of the same order of magnitude as that of the latter method and the resulting probabilities are also closer than they are to those of the basic isotonic regression. (See Table 4.12). The basic isotonic probabilities and those produced using method 2 with subsampling are shown in Figures 4.27(a) and (b) respectively.

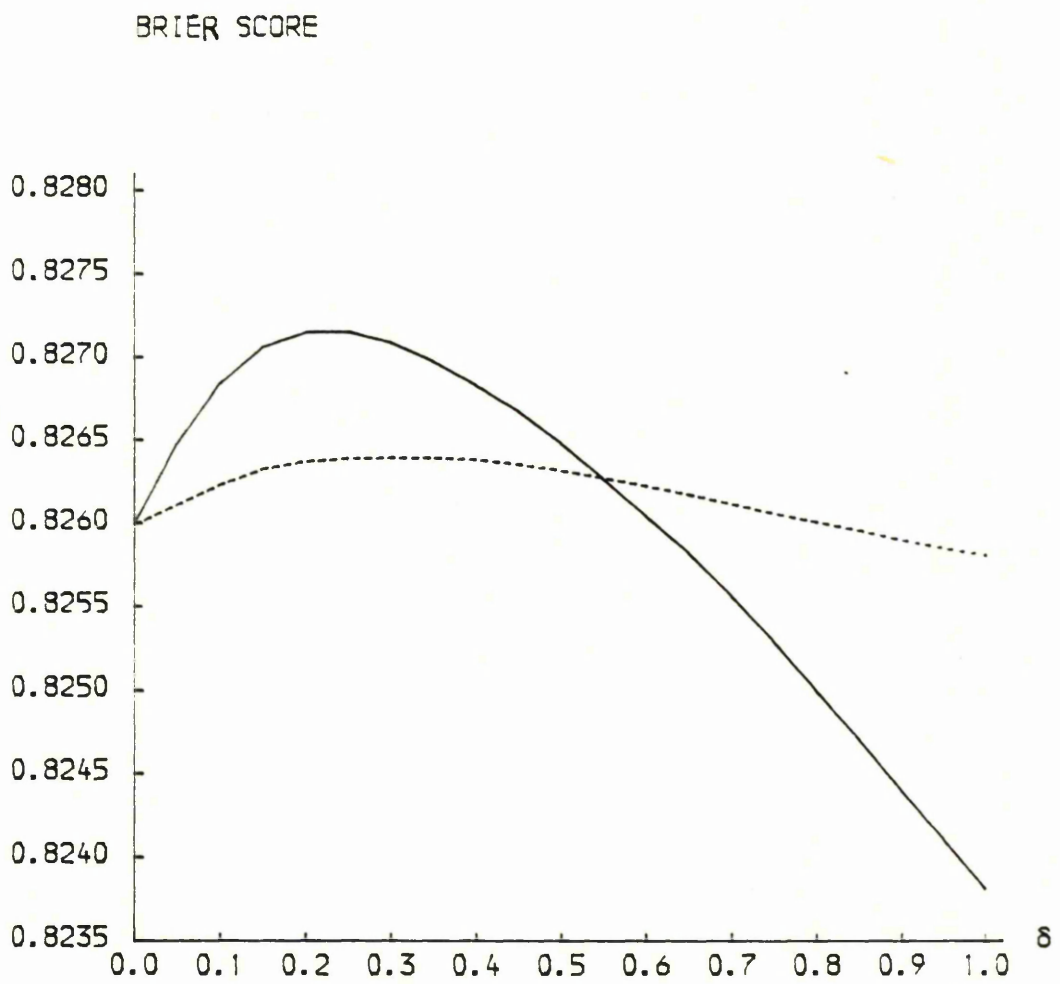


Figure 4.26 Brier score as a function of δ, for the 15 x 10 array, where δ controls the degree of extra smoothing imposed on the basic isotonic regression. The dotted line corresponds to method 1 and the solid line to method 2.

Table 4.11(a) EXTRA SMOOTHING WITH THE 15 X 10 TABLE

Isotonic methods		Brier Score	Error Rate (%)
Method 1	$\delta = .29$	.826400	26.0
Method 2	$\delta = .23$	.827166	26.0
Method 1 with subsampling	$\delta = .55$	.826278	26.0
Method 2 with subsampling	$\delta = .47$	.826604	26.0

Table 4.11(b) EXTRA SMOOTHING WITH THE 5 X 5 TABLE

Isotonic methods		Brier Score	Error Rate (%)
Method 2	$\delta = .81$	.825789	25.2
Method 2 with subsampling	$\delta = .92$	.825783	25.2

Table 4.12 Fitted distributions $\hat{p}(\pi_1|x)$ for the 5 x 5 table

Method	Age	Coma Sum				
		3-4	5-6	7-8	9-10	11-15
Isotonic	0-14	.667	.367	.259	.089	.089
	15-29	.833	.444	.259	.089	.089
	30-44	.980	.607	.333	.089	.089
	45-59	.980	.680	.483	.483	.091
	≥60	.980	.945	.945	.571	.091
Method 2		.655	.371	.264	.131	.131
		.823	.446	.264	.131	.131
	"	.959	.605	.340	.131	.131
		.959	.675	.485	.485	.146
		.959	.933	.933	.567	.146
Method 2 with subsampling		.654	.371	.265	.136	.136
		.822	.446	.265	.136	.136
	"	.956	.604	.340	.136	.136
		.956	.674	.485	.485	.153
		.956	.932	.932	.566	.153

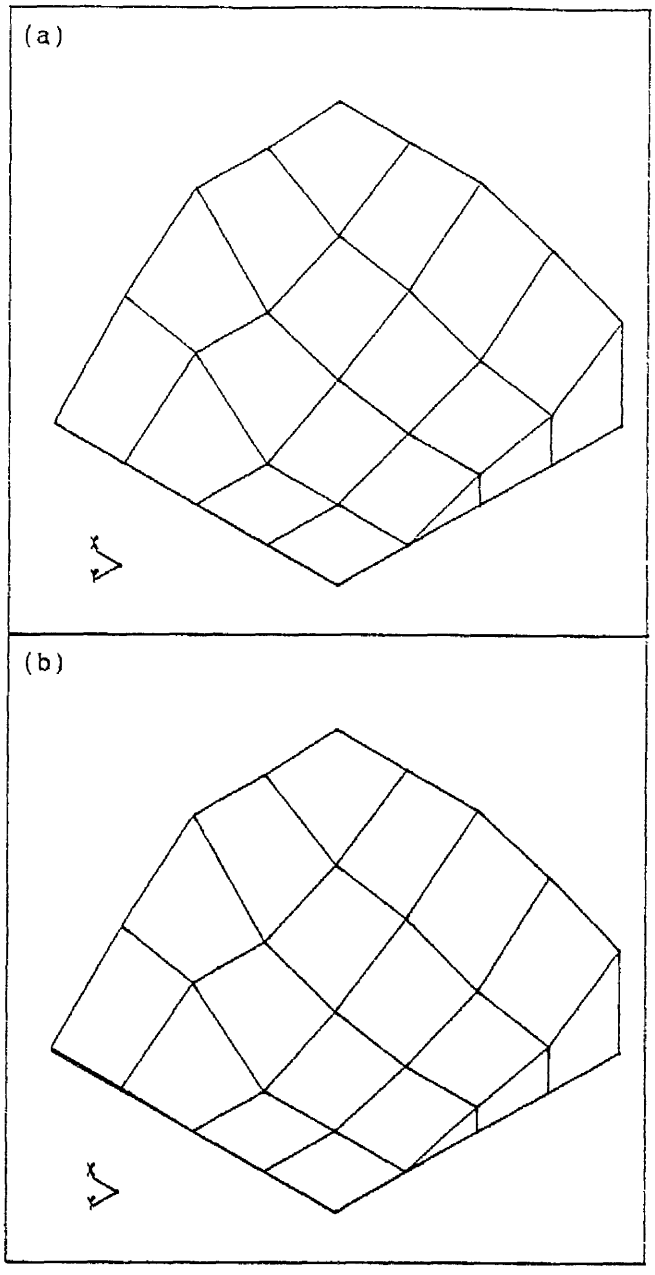


Figure 4.27 $\hat{p}(\pi_1 \mid \underline{x})$ displayed as 3-dimensional isotonic projections for the 5x5 table.

(a) Basic isotonic regression.

(b) Method 2 with subsampling.

Visually, there is little difference between them though the latter are fractionally smoother.

In the 15 by 10 case, both methods 1 and 2 with subsampling improved the Brier score of the basic isotonic estimate, although in both cases the suggested degree of smoothing is greater than optimal. In both cases the probabilities (not shown) are smoothed towards those of the optimal isotonic regression with extra smoothing and away from the basic one, although this is less noticeable for method 1. Sampling method 1 produced a slightly nearer optimal degree of smoothing than did sampling method 2, but nevertheless the latter still outperformed it and in fact achieved a better Brier score then method 1 would if optimisation were carried out directly on the full test set. Three-dimensional plots, Figures 4.28(a)-(c), show that while method 2 produces smoother probabilities than the basic isotonic regression, sampling method 2 is intermediate between the two but still very similar to optimal. Again a more marked difference is apparent between methods for the larger table, where the value of extra smoothing would be expected to be greater.

While $\hat{\delta}$ varies considerably between subsamples, and an occasional answer of $\hat{\delta} = 0$ is returned with the Brier score declining smoothly from that of the basic isotonic regression, where $\hat{\delta}$ is non-zero smooth curves of the type illustrated in Figure 4.26 are seen. These are encouraging results for the isotonic estimator. Hence it appears that taking the mean parameter estimate over a reasonable number of subsamples will suggest a sensible degree of extra smoothing and will effect an improvement, in terms of the Brier score, over the method of Dykstra. Particularly for sparse tables, we would recommend use of method 2.

4.4.3 Isotonisation of a parametric model

A further possibility (not implemented here) is to take a smooth parametric model as our starting point, constraining the parameter estimates suitably in order to produce an "isotonised" estimate of the $\{p(\pi_1|x_{ij})\}$. With the 2 population logistic model we would have

$p(\pi_1|x_{ij}) = \exp(\theta_{ij})/\{1 + \exp(\theta_{ij})\}$, $i = 1,\ldots, k_1$; $j = 1,\ldots, k_2$

and hence requiring that $p(\pi_1|x_{ij}) \leqslant p(\pi_1|x_{kl})$ iff $i \leqslant k$ and $j \leqslant l$ is equivalent to the conditions on the parameters $\{\theta_{ij}\}$ that $\theta_{ij} \leqslant \theta_{kl}$ iff $i \leqslant k$ and $j \leqslant l$. In principle the procedure would be

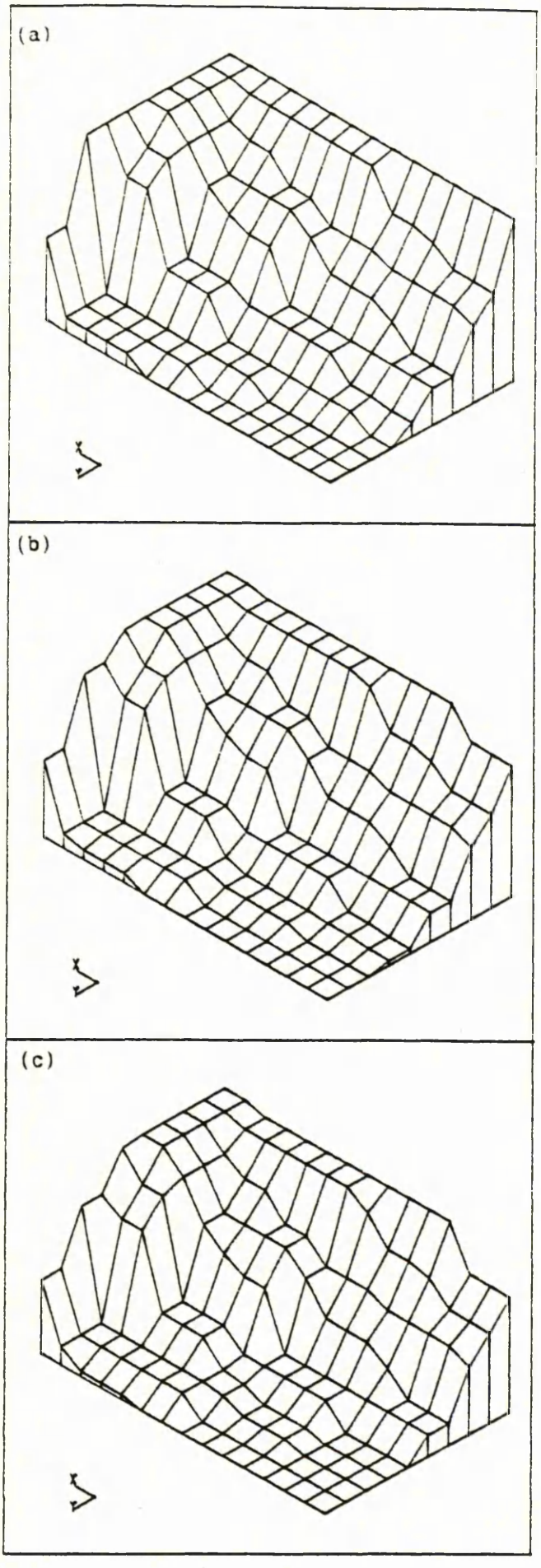


Figure 4.28 $\hat{p}(\pi_1 \mid \underline{x})$ for the 15 x 10 table.

(a) Basic isotonic regression.

(b) Method 2.

(c) Method 2 with subsampling.

to estimate $\{\theta_{ij}\}$ by constrained maximum likelihood, but this would be more readily accomplished in practice by fitting the logistic model in the usual way and then finding the isotonic regression of the $\{\theta_{ij}\}$, using the inverse estimated standard errors of the parameters as cell weights.

4.5 A 3-DIMENSIONAL EXAMPLE

4.5.1 Data

We consider now 3 variables, using the same data set, namely the raw (24 hour best) Eye, Motor and Verbal Scores (EMV Scores) from which the Coma Sum is derived, which have 4, 6, and 5 ordered categories respectively. In each case a low score is poor. Again the 6 month outcome is the response variable of interest.

Five sample sizes were used as training data, consisting of the 1st 100, 200, 300, 400 and the full set of 500 training cases, omitting incomplete cases as before. In each case the test set was as previously, namely the 2nd set of 500 cases. Use of the same test set allows comparisons to be made between training samples of different sizes. The full test and training samples were also reversed as a check against spurious results.

Initially the full 120-cell contingency table was used, (shown for the full sample of 500 cases in Tables 4.13(a)-(b)) and is obviously sparse, especially for the smaller sample sizes. Secondly, this was collapsed to a 24-cell table (Table 4.14) by recoding Eye and Verbal scores to 2 categories each (1, and > 1). Motor score is the most informative and its 6 categories were retained.

4.5.2 Models

At this point we dispense with the kernel method as it would seem that the diagnostic approach is the more appropriate for this data set. We therefore compare 2 logistic models with the isotonic method, but for comparison also include an independence model as a simple but robust example of the sampling approach, found by Titterington et al. (1981) to perform well on this data set. The models used are as follows :

1) The independence model (INDEP) is as used by Titterington et al. (1981), namely

Table 4.13(a) 120-CELL TABLE

Category		Training sample cell counts									
		π_1					π_2				
E	V / M	1	2	3	4	5	1	2	3	4	5
1	1	19	0	0	0	0	0	0	0	0	0
	2	55	2	0	0	0	8	1	0	0	0
	3	30	2	0	0	0	18	2	0	0	0
	4	57	7	1	0	0	49	7	1	0	0
	5	39	13	0	1	0	41	34	3	1	0
	6	2	0	0	0	0	0	1	0	1	0
2	1	1	0	0	0	0	0	0	0	0	0
	2	1	0	0	0	0	0	0	0	0	0
	3	1	0	0	0	0	0	0	0	0	0
	4	2	2	0	0	0	0	1	1	0	0
	5	4	4	0	0	0	7	9	1	0	0
	6	0	0	0	0	0	0	0	2	0	0
3	1	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0
	4	0	1	0	1	0	0	0	0	0	0
	5	0	0	0	0	0	3	6	1	2	1
	6	0	0	0	0	0	1	1	3	3	0
4	1	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0
	4	1	0	0	0	0	0	1	0	0	0
	5	0	0	0	0	0	2	5	0	1	0
	6	0	0	0	2	0	0	1	0	3	2

Table 4.13(b) 120-CELL TABLE

Category		Test sample cell counts									
		π_1					π_2				
E	V / M	1	2	3	4	5	1	2	3	4	5
1	1	18	0	0	0	0	2	0	0	0	0
	2	60	1	0	0	0	13	0	0	0	0
	3	33	1	0	0	0	12	3	0	0	0
	4	63	5	1	0	0	55	8	2	0	0
	5	27	8	0	0	0	41	21	2	1	0
	6	2	0	0	1	0	2	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0
	2	2	0	0	0	0	1	0	0	0	0
	3	0	0	0	0	0	2	0	0	0	0
	4	3	0	0	0	0	7	2	0	1	0
	5	1	3	1	0	0	10	7	1	0	0
	6	0	0	0	0	0	0	2	1	1	0
3	1	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	1	1	0	0	0
	3	0	0	0	0	0	0	0	0	0	0
	4	2	0	0	0	0	0	1	0	1	0
	5	2	2	0	1	0	5	4	1	2	1
	6	0	0	0	1	0	1	2	2	1	1
4	1	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	0	0	1	0	0	0	0
	5	0	0	0	0	0	4	4	0	2	0
	6	0	0	0	1	0	0	2	1	2	0

Table 4.14 24-CELL TABLE

Category		Training sample cell counts				Test sample cell counts			
		π_1		π_2		π_1		π_2	
E	V / M	1	2	1	2	1	2	1	2
1	1	19	0	0	0	18	0	2	0
	2	55	2	8	1	60	1	13	0
	3	30	2	18	2	33	1	12	3
	4	57	8	49	8	63	6	55	10
	5	39	14	41	38	27	8	41	24
	6	2	0	0	2	2	1	2	0
2	1	1	0	0	0	0	0	0	0
	2	1	0	0	0	2	0	2	1
	3	1	0	0	0	0	0	2	0
	4	3	4	0	3	5	0	8	5
	5	4	4	12	26	3	7	19	22
	6	0	2	1	15	0	2	1	15

$$p(\underline{x}|\pi_i) \propto \left\{ \prod_{r=1}^{d} \left[\frac{n_i(x_r) + 1/k_r}{N_i(r) + 1} \right] \right\}^B ,$$

where d is the number of variables, x_r denotes the observation on the rth variable, $n_i(x_r)$ is the number of cases in the training set with outcome π_i and score x_r on variable r, k_r is the number of categories of the rth variable, and $N_i(r)$ is the number of cases in the training set of outcome π_i for which variable r is not missing. B is an overall association factor.

For simplicity and more faithfulness to the basic model B was set to 1.

2) and 3) Logistic models

Both the linear logistic model (LINT), which treats each variable as continuous or interval-scaled, and the nominal categorical equivalent (LCAT) were used. In view of the ordered nature of the variables we would expect the former to be superior. These models were fitted using the BMDP program PLR (Dixon, 1985). Scores for the slightly smoothed raw test and training relative frequencies, replacing zero-weighted cells by the overall relative frequency of the appropriate outcome, are also provided as reference points and are denoted RFTS and RFTR respectively. These relative frequencies were used as input to the isotonic regression algorithm.

4) Isotonic Regression

In view of the difficulties with sparse data described previously, the algorithm of Dykstra (1981) was used with a weight of $\delta = .1$, as this was found to give convergence within a reasonable number of cycles, here taken as 500. Again zero-weighted cells were given a relative frequency of the overall mean. While this will impose a little extra smoothing (as in Method 1 of Section 4.4.2), practical considerations require it. Convergence was extremely slow when the training sample consisted of the first 100, 200, 300, 400, or 500 cases with $\delta = 10^{-5}$, as it was also with the alternative algorithm of Wollan and Dykstra (1987). The latter algorithm also encountered difficulties with the 2nd 500 cases in both the 24- and 120-cell multinomials and was disappointing here, perhaps not surprisingly as in the 120-cell

case it imposes a basic set of 286 constraints and an augmented one of 360. In the 24-cell example the corresponding numbers are 44 and 48. Being rather less sparse, all examples considered with the 24-cell problem converged quickly with the specialised isotonic algorithm as did the 120-cell case with the 2nd 500 cases used as training data, so that extra smoothing is not strictly speaking necessary here but was retained for better comparison between the 2 sets of results. The method is denoted ISO.

For the larger and smaller tables respectively, Tables 4.15(a)-(b) and 4.16(a)-(b) give continuous performance scores and error rates, for each sample size and model, together with benchmark models RFTR and RFTS as previously. Means of assessment of the confidence of each method, namely probability assignment and doubt matrices (see Section 1.6.2), are found in Tables 4.15(c)-(d) and 4.16(c)-(d). The threshhold value of .85 for a classification of doubt was chosen as being the level at which differences appear between the various methods.

4.5.3 Discussion

Especially for smaller sample sizes 100-200, LINT is noticeably better than LCAT, which is poorer than any other method on all scores, especially on Brier score, and even on error rate. For larger samples, however, error rates are closer, and LCAT is fractionally better on error rate for the 1st 500 cases then LINT, and fractionally better overall with the 2nd 500 cases.

For larger samples particularly, ISO is poorer on error rate relative to the parametric methods, and is characterised by more unbalanced error rates which may be a disadvantage, although, except for the 1st 100 cases where the reverse is true for all except INDEP, more type 2s than type 1s are misclassified by all methods. This is not surprising as π_2 represents more widely varying outcomes than π_1, but, given the potential use of such predictions for patient management, misclassification of type 2s may be considered to be the more serious mistake and ISO to be deficient in this respect. On Brier and modified log scores ISO tends to be slightly better than INDEP.

Overall LINT is best and better on Brier, log and modified log scores in all cases except that noted above, and for 24 cells and 100 training cases where INDEP was slightly better.

Hilden (1984) remarks that the independence model is typically

Table 4.15(a) RESULTS FOR THE 120-CELL MULTINOMIAL

Method	Brier Score	Log Score	ε-Log Score	Error Rate $\%\pi_1$	Error Rate $\%\pi_2$	Error Rate % Overall
Sample Size = 100						
LCAT	.7588	-.8941	-.6624	62.34	20.25	41.39
LINT	.7900	-.6100	-.5634	51.05	15.19	33.19
ISO	.7839	-.6885	-.5842	66.53	6.75	36.76
INDEP	.7743	-.6780	-.6123	34.73	45.15	39.92
RFTR	.7512	-.9511	-.6731	61.92	18.99	40.55
RFTS	.8128	-.5436	-.5017	23.85	35.86	29.83
Sample Size = 200						
LCAT	.7677	-.8231	-.6254	23.85	60.34	42.02
LINT	.7820	-.6265	-.5791	10.88	61.18	35.92
ISO	.7754	-.7375	-.6039	8.37	69.20	38.66
INDEP	.7630	-.7064	-.6391	10.88	62.87	36.76
RFTR	.7565	-1.1022	-.6811	13.39	74.26	43.70
RFTS	.8128	-.5436	-.5017	23.85	36.29	30.04
Sample Size = 300						
LCAT	.7816	-.6970	-.5849	12.97	61.60	37.18
LINT	.7930	-.6036	-.5568	14.23	54.85	34.45
ISO	.7827	-.7379	-.5997	8.37	70.04	39.08
INDEP	.7772	-.6692	-.6061	13.39	56.12	34.66
RFTR	.7716	-1.1683	-.6622	10.04	70.04	39.92
RFTS	.8128	-.5436	-.5017	23.43	36.71	30.04
Sample Size = 400						
LCAT	.7869	-.6843	-.5748	24.27	43.88	34.03
LINT	.7974	-.5944	-.5478	25.52	37.55	31.51
ISO	.7854	-.7347	-.5956	8.37	70.04	39.08
INDEP	.7851	-.6558	-.5936	13.39	56.12	34.66
RFTR	.7761	-1.1598	-.6539	9.62	67.93	38.66
RFTS	.8128	-.5436	-.5017	23.43	35.44	29.41

Table 4.15(b) RESULTS FOR THE 120-CELL MULTINOMIAL

Method	Brier Score	Log Score	ϵ-Log Score	Error Rate $\%\pi_1$	Error Rate $\%\pi_2$	Error Rate % Overall
Sample Size = 500						
LCAT	.7900	-.6720	-.5688	24.27	40.51	32.35
LINT	.7952	-.5976	-.5514	24.27	40.93	32.56
ISO	.7905	-.6746	-.5708	14.23	57.81	35.92
INDEP	.7862	-.6467	-.5913	24.69	38.82	31.72
RFTR	.7676	-1.4835	-.7168	21.34	48.52	34.87
RFTS	.8123	-.5436	-.5017	23.43	35.86	29.62
Sample = 2nd 500						
LCAT	.7960	-.5909	-.5455	32.66	35.27	33.90
LINT	.7976	-.5881	-.5423	32.26	34.82	33.47
ISO	.7961	-.5896	-.5441	33.47	33.93	33.69
INDEP	.7891	-.6185	-.5692	33.47	33.93	33.69
RFTR	.7826	-.9220	-.6144	33.06	37.95	35.38
RFTS	.8150	-.5291	-.4893	27.02	37.50	31.99

Table 4.15(c) RESULTS FOR THE 120-CELL MULTINOMIAL

Method	Average Probability assigned to correct outcome		Doubt Matrix (doubt ≤ 0.85) % classified to π_1 and π_2			
	π_1	π_2	π_1	π_2	π_1	π_2
Sample Size = 100						
LCAT	.513	.602	8.37	1.69	3.35	15.61
LINT	.539	.602	-	-	1.26	4.22
ISO	.541	.605	7.53	0.84	3.77	16.88
INDEP	.566	.577	7.53	0.84	5.02	20.68
RFTR	.508	.584	8.37	1.69	1.67	6.75
RFTS	.627	.624	9.20	0.84	0.42	20.25
Sample Size = 200						
LCAT	.674	.491	8.37	1.27	2.93	16.03
LINT	.693	.481	7.53	0.84	1.67	10.55
ISO	.686	.489	33.47	6.75	3.77	18.14
INDEP	.727	.472	32.64	6.33	3.77	19.83
RFTR	.677	.459	34.31	8.02	2.93	11.39
RFTS	.627	.624	9.20	0.84	0.42	20.25
Sample Size = 300						
LCAT	.642	.527	8.37	1.27	2.93	15.61
LINT	.647	.534	7.53	0.84	1.67	10.55
ISO	.638	.540	7.53	0.84	3.77	18.14
INDEP	.705	.513	32.64	6.33	4.18	20.68
RFTR	.633	.528	9.21	2.95	3.77	16.03
RFTS	.627	.624	9.21	0.84	0.42	20.25
Sample Size = 400						
LCAT	.630	.557	8.37	1.27	2.93	16.03
LINT	.626	.570	7.53	0.84	2.51	14.77
ISO	.627	.562	33.47	6.75	3.77	16.88
INDEP	.688	.555	33.47	6.75	5.44	23.63
RFTR	.625	.551	34.31	8.44	3.77	15.61
RFTS	.627	.624	9.20	0.84	0.42	20.25

Table 4.15(d) RESULTS FOR THE 120-CELL MULTINOMIAL

Method	Average Probability assigned to correct outcome		Doubt Matrix (doubt ⩽ 0.85) % classified to π_1 and π_2			
	π_1	π_2	π_1	π_2	π_1	π_2
Sample Size = 500						
LCAT	.642	.554	33.47	6.75	2.93	13.92
LINT	.638	.555	7.53	0.84	1.67	7.60
ISO	.642	.558	33.47	6.75	3.77	16.88
INDEP	.688	.562	33.47	6.75	6.28	22.36
RFTR	.645	.534	35.56	12.66	3.35	14.77
RFTS	.627	.624	9.20	0.84	0.42	20.25
Sample = 2nd 500						
LCAT	.583	.614	7.66	-	2.82	14.73
LINT	.581	.607	7.66	-	0.81	15.18
ISO	.588	.612	7.66	-	1.21	11.61
INDEP	.620	.628	29.84	3.57	6.05	26.34
RFTR	.582	.600	8.47	0.89	4.44	13.39
RFTS	.648	.610	33.87	3.57	-	18.75

Table 4.16(a) RESULTS FOR THE 24-CELL MULTINOMIAL

Method	Brier Score	Log Score	ε-Log Score	Error Rate $\%\pi_1$	Error Rate $\%\pi_2$	Error Rate % Overall
Sample Size = 100						
LCAT	.7648	-.8877	-.6395	64.02	13.92	39.08
LINT	.7837	-.6231	-.5761	49.79	18.14	34.03
ISO	.7830	-.7819	-.5927	66.53	7.17	36.97
INDEP	.7853	-.6339	-.5832	39.75	30.80	35.29
RFTR	.7570	-.9051	-.6564	62.34	18.99	40.76
RFTS	.8064	-.5672	-.5228	25.94	35.02	30.46
Sample Size = 200						
LCAT	.7740	-.7036	-.5995	23.85	55.70	39.71
LINT	.7839	-.6196	-.5729	9.62	62.87	36.13
ISO	.7777	-.7931	-.6035	8.79	65.82	37.18
INDEP	.7721	-.6604	-.6097	12.13	56.96	34.45
RFTR	.7696	-.9434	-.6379	16.74	64.14	40.34
RFTS	.8064	-.5672	-.5228	25.94	35.44	30.67
Sample Size = 300						
LCAT	.7899	-.6702	-.5671	14.64	53.16	33.82
LINT	.7947	-.5966	-.5506	14.23	55.27	34.66
ISO	.7871	-.7724	-.5835	8.79	65.82	37.18
INDEP	.7844	-.6375	-.5857	14.64	52.74	33.61
RFTR	.7815	-.9178	-.6132	9.20	62.87	35.92
RFTS	.8064	-.5672	-.5228	25.10	35.44	30.25
Sample Size = 400						
LCAT	.7948	-.6604	-.5576	25.94	35.86	30.88
LINT	.7978	-.5902	-.5445	25.52	37.98	31.72
ISO	.7886	-.7704	-.5815	8.79	65.82	37.18
INDEP	.7892	-.6348	-.5812	14.64	52.74	33.61
RFTR	.7853	-.9109	-.6064	9.20	62.87	35.92
RFTS	.8064	-.5672	-.5228	25.10	35.44	30.25

Table 4.16(b) RESULTS FOR THE 24-CELL MULTINOMIAL

Method	Brier Score	Log Score	ε-Log Score	Error Rate $\%\pi_1$	Error Rate $\%\pi_2$	Error Rate % Overall
Sample Size = 500						
LCAT	.7955	-.6592	-.5565	25.94	36.71	31.30
LINT	.7953	-.5961	-.5501	23.43	41.35	32.35
ISO	.7913	-.6696	-.5663	14.23	57.81	35.92
INDEP	.7901	-.6293	-.5770	25.94	35.44	30.67
RFTR	.7758	-1.2060	-.6689	20.50	45.57	32.98
RFTS	.8064	-.5672	-.5228	25.10	35.44	30.25
Sample = 2nd 500						
LCAT	.7973	-.5878	-.5424	32.66	34.82	33.69
LINT	.7968	-.5892	-.5436	32.66	34.82	33.69
ISO	.7949	-.5931	-.5474	33.87	33.93	33.90
INDEP	.7886	-.6185	-.5704	33.47	33.93	33.69
RFTR	.7864	-.8828	-.6037	33.06	34.82	33.90
RFTS	.8084	-.5553	-.5125	27.42	37.95	32.42

Table 4.16(c) RESULTS FOR THE 24-CELL MULTINOMIAL

Method	Average Probability assigned to correct outcome		Doubt Matrix (doubt ⩽ 0.85) % classified to π_1 and π_2			
	π_1	π_2	π_1	π_2	π_1	π_2
Sample Size = 100						
LCAT	.522	.580	7.53	0.84	2.09	7.60
LINT	.530	.599	-	-	-	-
ISO	.540	.599	7.53	0.84	1.26	6.75
INDEP	.558	.589	7.53	0.84	2.09	7.60
RFTR	.505	.601	8.37	1.69	1.26	6.75
RFTS	.614	.611	8.37	0.84	2.09	18.14
Sample Size = 200						
LCAT	.678	.470	32.64	6.33	0.84	6.33
LINT	.692	.469	7.53	0.84	-	-
ISO	.685	.476	33.47	7.17	1.26	6.75
INDEP	.721	.482	32.64	6.33	4.18	16.03
RFTR	.676	.474	34.31	8.02	1.26	6.75
RFTS	.614	.611	8.37	0.84	2.09	18.14
Sample Size = 300						
LCAT	.639	.533	7.53	0.84	1.26	6.75
LINT	.648	.530	7.53	0.84	0.84	6.33
ISO	.636	.537	7.53	0.84	1.26	6.75
INDEP	.695	.536	32.64	6.33	4.18	18.14
RFTR	.633	.535	9.20	2.53	1.26	6.75
RFTS	.614	.611	8.37	0.84	2.09	18.14
Sample Size = 400						
LCAT	.627	.564	32.64	6.33	1.26	6.75
LINT	.629	.563	7.53	0.84	0.84	6.33
ISO	.626	.561	33.47	7.17	4.18	16.03
INDEP	.681	.572	32.64	6.33	4.18	18.14
RFTR	.626	.559	34.31	8.02	4.18	16.03
RFTS	.614	.611	8.37	0.84	2.09	18.14

Table 4.16(d) RESULTS FOR THE 24-CELL MULTINOMIAL

Method	Average Probability assigned to correct outcome		Doubt Matrix (doubt ≤ 0.85) % classified to π_1 and π_2			
	π_1	π_2	π_1	π_2	π_1	π_2
Sample Size = 500						
LCAT	.643	.553	32.64	6.33	-	-
LINT	.641	.551	7.53	0.84	0.84	6.33
ISO	.642	.554	33.47	7.17	4.18	16.03
INDEP	.686	.569	32.64	6.33	4.18	18.14
RFTR	.650	.535	36.40	12.24	4.18	16.03
RFTS	.614	.611	8.37	0.84	2.09	18.14
Sample = 2nd 500						
LCAT	.583	.612	7.66	-	2.42	18.30
LINT	.582	.611	7.66	-	2.42	18.75
ISO	.585	.613	7.66	-	0.81	7.14
INDEP	.618	.629	29.84	3.57	5.64	26.34
RFTR	.581	.608	8.47	1.34	4.44	13.84
RFTS	.635	.596	33.06	3.57	2.42	19.64

overconfident, producing over-extreme odds, while not distorting much the ranking of the alternative diagnoses. He attributes this to violation of conditional independence. Titterington et al. (1981) also found this. We see that in terms of average probability assigned to the correct group for cases from π_1, while LCAT, LINT and ISO are reasonably similar, INDEP assigns higher probability. (Tables 4.15(c)-(d) and 4.16(c)-(d)). On the 1st and 2nd 500 cases it is also more confident at predicting π_2 cases though for 120 cells it gives lower probability to the latter when the training set is smaller (100 - 300 cases). There is little difference with 100 - 400 cases and the smaller table. However, since overall LINT performs the best of the parametric models, this was the one chosen to use in conjunction with ISO for convex smoothing.

Tables 4.17 and 4.18 overleaf give the results of smoothing using the model CON = λ ISO + $(1-\lambda)$ LINT for the 120- and 24-cell tables respectively. Only in the 24-cell case with the smallest sample size was there a marked improvement seen in scores (to the 3rd rather than 4th decimal place). In fact in this instance INDEP was the best parametric model on Brier score rather than LINT, and the convex method betters it. In all other cases if any improvement was possible it was very slight. Although we have not used it here, the isotonic Method 2 would presumably once more be a superior means of improving on the basic isotonic method, bringing it closer to LINT.

However, the superior performance of LINT may well be due to the nature of the Coma Score with its equal steps, and it may be expected to improve less on LCAT for an ordered scale with less even steps. In such a situation ISO might come into its own, especially if implemented with extra smoothing, and outperform both logistic models. In either situation it would be interesting to compare the various isotonic approaches with the isotonised logistic models suggested in Section 4.4.3.

Table 4.17 120-CELL TABLE :
CONVEX SMOOTHING USING MODEL λISO + (1-λ) LINT

Sample size	$\hat{\lambda}$	Method	Brier Score	Log Score	ϵ-Log Score	Error Rate (%)
100		LINT	.7900	-.6100	-.5634	33.19
		ISO	.7839	-.6885	-.5842	36.76
	.2157	CON	.7905	-.6085	-.5619	33.19
200		LINT	.7820	-.6265	-.5791	35.92
		ISO	.7754	-.7375	-.6039	38.66
	.0842	CON	.7821	-.6265	-.5790	35.92
300		LINT	.7930	-.6036	-.5568	34.45
		ISO	.7827	-.7379	-.5997	39.08
	0	CON	AS LINT			
400		LINT	.7974	-.5944	-.5478	31.51
		ISO	.7854	-.7347	-.5956	39.08
	0	CON	AS LINT			
1st 500		LINT	.7952	-.5976	-.5514	32.56
		ISO	.7905	-.6746	-.5708	35.92
	.1053	CON	.7952	-.5975	-.5512	32.56
2nd 500		LINT	.7976	-.5881	-.5423	33.47
		ISO	.7961	-.5896	-.5441	33.69
	.2275	CON	.7977	-.5872	-.5416	33.47

Table 4.18 24-CELL TABLE :
CONVEX SMOOTHING USING MODEL λISO + (1-λ) LINT

Sample size	$\hat{\lambda}$	Method	Brier Score	Log Score	ε-Log Score	Error Rate (%)
100		LINT	.7837	-.6231	-.5761	34.03
		ISO	.7830	-.7819	-.5927	36.97
	.4624	CON	.7858	-.6174	-.5706	33.61
200		LINT	.7839	-.6196	-.5729	36.13
		ISO	.7777	-.7931	-.6035	37.18
	.0552	CON	.7840	-.6194	-.5727	36.13
300		LINT	.7947	-.5966	-.5506	34.66
		ISO	.7871	-.7724	-.5835	37.18
	0	CON	AS LINT			
400		LINT	.7978	-.5902	-.5445	31.72
		ISO	.7886	-.7704	-.5815	37.18
	0	CON	AS LINT			
1st 500		LINT	.7953	-.5961	-.5501	32.35
		ISO	.7913	-.6696	-.5663	35.92
	.0884	CON	.7953	-.5962	-.5502	32.35
2nd 500		LINT	.7968	-.5892	-.5436	33.69
		ISO	.7949	-.5931	-.5474	33.90
	0	CON	AS LINT			

CHAPTER 5 CONCLUSIONS AND FURTHER WORK

Kernel smoothing and limitations of the present work

Hitherto we have addressed the question of how to achieve reliable estimation of posterior probabilities through the use of estimators involving data smoothing.

In Chapter 2, comparing marginal and direct methods for estimation of smoothing parameters in fixed (continuous) kernel density estimators in terms of MSE, using contour plots and plots of the predicted probability function $p(\pi_1|x)$, the scope to improve on the former appeared limited, except for equal populations where direct methods generally improved substantially. The potential to improve in general seemed greater for smaller samples but here direct methods were typically no better or worse than marginal ones and could be very undersmoothed. The MSE-optimal kernel itself was often poor, at its best again for equal populations, which are not of practical interest, and generally acceptable for well separated populations (expected error rate of 5%) but for intermediate problems frequently disappointing and undersmoothed. Trying to improve on the kernel method with a spline ratio estimator was also disappointing as the latter tended to over- rather than undersmooth and could be very poor, again especially for small samples.

From t-statistics, it was found that marginal method 1 (NOPT) was consistently better than 2 (ASOPT), and 2 consistently poorer than 3 (XVKL) and 4 (XVISE) which were similar, or 3 slightly superior. At 5% and 20%, 1 was in fact generally best of all or better than all except method 3. Direct methods 6-8 (XV BRIER, XV LOG and XV ϵ-LOG respectively) were similar also, especially for $n_1:n_2 = 25:10$, though for other sample sizes 6 might have the edge for well-separated populations but, especially for equal-sized samples, 8 the edge for higher error rates, when 6 is poorer.

Relative to 6-8 the marginal methods generally deteriorated as error rate increased, most notably relative to method 2 (the poorest marginal method), becoming comparable to or poorer than 6-8 for an expected error rate of 50%. Again this was especially marked for equal populations, where for equal smaller sample sizes even the overall best method 1 was inferior to method 8, rather than just comparable to the direct methods, and for the unbalanced 25:10 case worse or nearly worse than 6-8. These are isolated cases however, and the improvement of 6-8 relative to (the better

of) 1-4 was disappointingly inconsistent.

On the much larger real data set in Section 2.6 methods 2 and 3 were undersmoothed, as was 4 which was slightly smoother. Method 1 alone, despite grossly oversmoothing both π_1 and π_2 in terms of providing acceptable density estimates, gave a realistic predicted probability function and was near-optimal. Direct methods 7 and 8 were comparable to 4 while 6 smoothed only slightly more.

An obvious limitation of the work in Chapter 2 is the restriction to normal populations and a single feature variable. In particular, if a variety of distributional shapes were studied, we might see some deterioration of the simple Normal Optimal method in the same cases as in density estimation alone (i.e. long-tailed and multimodal distributions, especially in higher dimensions, Bowman (1981, 1985). See Section 2.3.2.) and better relative performance of the assessment methods, as there is no reason to expect a decline in performance of these for a change in density shape.

Less obviously, an unfortunate consequence of the manner of generating the simulated normal data used in Section 2.5 (see Appendix 2), chosen for better comparability of different (balanced) sample sizes, is that for a given simulation and constant sample sizes the sample from π_1 is the same, as is the sample from π_2 for specified sample sizes and fixed σ_2, so that the variability in the data is rather limited.

Smoothing parameter estimation method 1 depends by definition only on sample size and population variance, as for normal data do methods 2-4. The consequence of this and the lack of variability in the data is seen in Appendix 2 to be that for a given simulation, for fixed sample sizes and constant σ_2, methods 1-4 do not vary at all. Also, for a given simulation and given sample sizes, the samples from π_2 are functionally related. While methods 6-8 depend in a complex manner on both μ_2 and σ_2, as does the mean squared error of $\hat{p}(\pi_1|x)$, certain trends observed from the contour plots and t-statistics may depend to some considerable extent on the means of simulation rather than being invoked merely by increasing variance, separation etc.. Were such work to be extended it would be helpful for there to be more natural variability in the data.

While the work of Chapter 2 was largely based on simulated data, the opposite is true of Chapter 4 where the real data set

introduced in Section 2.6 was used more extensively to study the performance of class conditional and posterior probability function estimators on univariate and multivariate ordered categorical variables. Discrete kernel estimators are an example of smoothed relative frequency estimators and are a special case of convex estimators of the type (3.3). In Chapter 4, as in Chapter 2, data-based Brier methods using cross-validation tended to be disappointing and could be even poorer than marginal methods, except for the less demanding, less sparse problem where the degree of smoothing is less crucial and the potential to improve on standard methods is smaller.

Another approach to joint smoothing parameter estimation

An alternative approach to joint estimation of the smoothing parameters $\{\lambda_i\}$ (for ordered or unordered discrete kernels) or $\{h_i\}$ (for continuous kernels) for discrimination purposes is proposed in a recent paper by Hall and Wand (1988). Rather than base estimation on density ratios directly they use the equivalence of allocation to π_1 when

$\frac{f_1(\underline{x}).\theta_1}{f_2(\underline{x}).\theta_2} \geq 1$ with allocation on the basis of the difference in discriminant scores using the rule :

allocate to π_1 if $g(x) \equiv f_1(\underline{x})\theta_1 - f_2(\underline{x})\theta_2 \geq 0$,

and jointly estimate the smoothing parameters to minimise the difference between $g(\underline{x})$ and $\hat{g}(\underline{x})$. Minimising an unbiased estimator of (or equivalently using cross-validation with) mean summed square error or integrated square error they demonstrate consistency in that for the former the estimated smoothing parameters tend asymptotically to those minimising mean summed square error itself, and for the latter the minimised ISE tends to the true minimum ISE. As in Tutz (1986) the results of applying the method to 3 examples are quoted only in terms of error rate, with the usual small differences between methods. It would be interesting to compare Hall and Wand's estimation method with our direct assessment methods based on functions of the density ratio, and in terms of continuous performance scores as well as error rate. Also, Hall and Wand note, for categorical data, that optimal performance is

sometimes achieved using negative smoothing parameters. While our convex estimators λISO + (1-λ)LINT (Section 4.5) rarely yielded any improvement in Brier score over that of the better individual estimator for $0 \leqslant \lambda \leqslant 1$, some optimising values outside of this range were returned, although not used as the interpretation of the new estimator is then lost.

Smoothed isotonic estimators

Isotonic regression also provides smoothed relative frequency estimates, although it accomplishes the smoothing in a different manner and through a diagnostic rather than a sampling approach. For discrete data, we have considered the imposition of further smoothing on an isotonic estimate by imposing a convex structure upon the predicted probabilities, smoothing between a parametric model and the isotonic estimator. The former brings in prior ideas of smoothness while the latter imposes ordering. More successfully, a different approach smoothed the raw relative frequencies using a Bayesian type convex estimator by adding in extra pseudo-observations, and then allowed the standard isotonic algorithms to bring in the required ordering. For continuous feature variables, we introduced smoothing via spline estimators of densities and density ratios in Sections 1.4.3 and 1.5.2 respectively. Isotonic splines provide a further possibility for the addition of extra smoothing. Wright and Wegman (1980) note the analogy between isotonic regression estimators and penalised least squares smoothing splines. Both are implicitly defined as minimisers of a squared distance term (or penalised squared distance term) over a certain class of functions. The difference is the subspace of functions over which optimisation takes place. They remark that the marriage of the two approaches, by optimising over the intersection of the relevant subspaces, may give the desired ordering while overcoming the disappointing continuity properties of the isotonic estimators.

The predicted probability function of the best continuous (kernel) method in Section 2.6 strongly resembled a smoothed version of the isotonic estimate for the discretised problem (Section 4.3.1). Although the performance of a 1-dimensional spline estimator in Section 2.5 as well as 2.6 was often disappointing, it was generally smooth and it would be interesting to investigate the performance of isotonic splines in higher

dimensions, although implementation appears to be non-trivial (Wright and Wegman, 1980; Wegman and Wright, 1983).

The relevance of ordering

Implicitly we have assumed the ordered nature of our response variable with respect to the explanatory variables to be of fundamental importance in modelling predicted probabilities and argued that this is best accomplished by a diagnostic approach.

We noted however in Section 3.3 an advantage of Anderson's stereotype models for ordered response variables over McCullagh's models to be that while the latter inherently assume the relevance of ordering, the former allow this to be tested. Analagous to McCullagh models, isotonic models take our definition of ordering as their endpoint, which may be seen as a disadvantage. While ordering known to be present may not be strictly relevant for modelling purposes the nature of discriminant analysis is such that we are able to assess how well a model performs in practice with respect to a given criterion. Good performance is the aim and provided this is achieved the assumptions underlying the model may not be too important. In Chapter 4 however, while isotonic regression alone recovered the required ordering, in each case considered it also outperformed the kernel methods and achieved near-optimal Brier scores, especially when implemented with extra smoothing. In Section 4.5 a simple ordered logistic model, treating the feature variables as interval-scaled, improved on a basic isotonic estimator for a 3-dimensional problem while 2 unordered models were poorer than both.

It would be useful to verify the conclusions of Chapter 4 on a variety of distributions with inherent ordering, as for instance in Titterington and Bowman (1985) (see Section 3.4.2) who surveyed and compared, for ordered categorical variables, numerous smoothed relative frequency estimators of probability functions and means of estimating their smoothing parameters.

Multiple ordered populations

In both Chapters 2 and 4, we considered only the most commonly studied 2 population problem. Had we not dichotomised outcome in Chapter 4, 3 rather than 2 types of ordering would have been present as the outcome variable is also ordinal. For ordered populations some misclassification errors are inherently more

serious than others, allocation to an adjacent or nearby category being less serious than a more extreme misclassification. Ashby et al. (1986) use in this situation a modified error rate, comparing the number of classifications to the correct or an adjacent category, to assess performance of 3 allocation rules based on a McCullagh (1980) model. It is not immediately obvious how one would incorporate unequal misclassification costs into assessment criteria of the type emphasised in this thesis.

APPENDIX 1 STANDARDISATION OF CONFIGURATIONS USED IN SIMULATIONS IN SECTION 2.5

Using Lachenbruch's (1975a) notation, the Bayes' optimal error rate

$T(R,f) = \theta_1 \int_{R_2} f_1(x)\,dx + \theta_2 \int_{R_1} f_2(x)\,dx$, where $\{f_i(x)\}$ are the class conditional distributions and $\{R_i\}$ the decision regions for π_i, $i = 1, 2$, $R_1 = \left\{x : \frac{f_1(x)}{f_2(x)} \cdot \frac{\theta_1}{\theta_2} > 1\right\}$, $R_2 = \bar{R}_1$,

using the Bayes' optimal allocation rule with costs $C(j|i) = 0$ if $i = j$, 1 otherwise and $C(j|i)$ is the cost of allocating an observation from π_i to π_j. $\{\theta_i\}$ are known incidence rates.

For $\pi_1 \sim N(0, 1)$ and $\pi_2 \sim N(\mu, \sigma^2)$, $\sigma^2 \neq 1$, identifying $R_1 = \{ x: \psi_1 x^2 + \psi_2 x + \psi_3 > 0 \}$, for constants ψ_1, ψ_2, ψ_3, requires finding the roots of a quadratic equation in x, a and b, say.

$$T(R,f) = \theta_2 \left\{ \Phi\left[\frac{b-\mu}{\sigma}\right] - \Phi\left[\frac{a-\mu}{\sigma}\right]\right\} + \theta_1 \left\{\Phi(a) + 1 - \Phi(b)\right\} = D, \text{ say,}$$

if $R_1 = \{x : a < x < b\}$,

and $= 1 - D$ if $R_1 = \{x : x \in (-\infty, a) \cup (b, \infty)\}$, where Φ is the standard Normal c.d.f..

For $\sigma^2 = 1$, $T(R,f) = \theta_1 \Phi\left\{\frac{\ln(\theta_2/\theta_1) - \frac{1}{2}\delta^2}{\delta}\right\} + \theta_2 \Phi\left\{-\left\{\frac{\ln(\theta_2/\theta_1) + \frac{1}{2}\delta^2}{\delta}\right\}\right\}$

where $\delta^2 = \frac{\mu^2}{\sigma^2}$ is the Mahalanobis distance between π_1 and π_2.

If $\theta_1 = \theta_2$ and $\sigma^2 = 1$, $T(R,f) = \Phi\left\{\frac{-\delta}{2}\right\}$ and solving $T(R,f)-E = 0$, where E is the required error rate, is trivial. Otherwise, it was solved numerically, using NAG (1984) subroutines C05ADF, and S15ABF to evaluate the standard Normal c.d.f.. Maximum likelihood plug-in estimates were used for $\{\theta_i\}$.

APPENDIX 2 A NOTE ON THE MEANS OF SIMULATION OF THE DATA IN SECTION 2.5

The random number generator used to simulate the normal data in Section 2.5 from each configuration was initialised each time with the same seed, for better comparability of different (balanced) sample sizes, generating first π_1 then π_2 for each simulation in turn. For given sample sizes, NAG (1984) subroutine G05DDF generates an observation from a N(0, 1) random variable and derives a N(μ, σ) observation by direct transformation of the former. Consequently, initialising simulation 1 at the same place each time means that for a given simulation, and constant sample sizes the sample from π_1 is the same, and for specified sample sizes and fixed σ_2, so is the sample from π_2. (Equally for a given simulation and given sample sizes, the samples from π_2 are functionally related).

Smoothing parameter estimation method 1 depends trivially on sample size and population variance. Where each observation is drawn from a normal population, since N(μ, σ) $\equiv$ σ N(0, 1) + μ, methods 2-4 of smoothing parameter estimation, each of which optimises a function involving the observations only through data differences, for a given sample size also depend only on the population variance and not on its mean. The result of the means by which the data were simulated is therefore that for a given simulation, for constant sample sizes and fixed σ_2, the estimates from methods 1-4 do not vary at all.

APPENDIX 3 OPTIMISING THE BRIER SCORE OF A CONVEX COMBINATION OF 2 MODELS

If P_{ij} is the probability assigned to class j for the ith case, and $d_{(i)}$ is the outcome associated with that case,

the Brier score, $S_B = \frac{1}{n}\left\{\sum_i \left[(1-P_{id_{(i)}})^2 + \sum_{j\neq d_{(i)}} P_{ij}^2\right]\right\}$,

may be written as $1 - 2Q_1 + Q_2$ where $Q_1 = \frac{1}{n}\sum_i P_{id_{(i)}}$, and

$Q_2 = \frac{1}{n}\sum_i \sum_j P_{ij}^2$.

Let $\{P^*_{ij}\}$ and $\{P'_{ij}\}$ denote probabilities arising from two models with corresponding scores Q^*_i, Q'_i, $i = 1, 2$, and define Q_{12} as $\frac{1}{n}\sum_i \sum_j P^*_{ij} P'_{ij}$, then $S_B(\lambda P^* + (1-\lambda)P')$ where $0 \leq \lambda \leq 1$,

is easily shown to be of the form $a + b\lambda + c\lambda^2$,

with $a = 1 - 2Q'_1 + Q'_2$, $b = 2Q'_1 - 2Q^*_1 - 2Q'_2 + 2Q_{12}$,

and $c = Q^*_2 + Q'_2 - 2Q_{12}$.

Similarly, the rescaled Brier score, $1 - \frac{1}{2} S_B = a_1 + b_1\lambda + c_1\lambda^2$, where $a_1 = 1 - \frac{1}{2}a$, $b_1 = -\frac{1}{2}b$, and $c_1 = -\frac{1}{2}c$. Either, as quadratics in λ, are optimised by $\hat{\lambda} = -b/(2c)$, with a corresponding Brier score of $a - b^2/(4c)$, provided $c > 0$. Otherwise, we cannot improve on the Brier score by taking a simple linear combination of the 2 methods. Even with $c > 0$, $\hat{\lambda}$ does not necessarily lie in the range (0, 1), in which case we are again restricted to one or other endpoint.

REFERENCES

ABRAMSON, I.S. (1982). On bandwidth variation in kernel estimates - a square root law. Ann. Statist., 10, 1217-1223.

AGRESTI, A. (1983). A survey of strategies for modelling cross-classifications having ordinal variables. J. Amer. Statist. Assoc., 78, 184-198.

AGRESTI, A. (1984). Analysis of Ordinal Categorical Data. Wiley, New York.

AITCHISON, J. (1975). Goodness of prediction fit. Biometrika, 62, 547-554.

AITCHISON, J. and DUNSMORE, I.R. (1975). Statistical Prediction Analysis. Cambridge University Press.

AITCHISON, J. and AITKEN, C.G.G. (1976). Multivariate binary discrimination by the kernel method. Biometrika, 63, 413-420.

AITCHISON, J., HABBEMA, J.D.F. and KAY, J. (1977). A critical comparison of two methods of statistical discrimination. Appl. Statist., 26, 15-25.

AITKEN, C.G. (1978). Methods of discrimination in multivariate binary data. In Compstat 1978, ed. L.C.A. Corsten and J. Hermans, 155-161. Physica Verlag, Vienna.

ANDERSON, J.A. (1969). Constrained discrimination between k populations. J. R. Statist. Soc. B, 31, 123-139.

ANDERSON, J.A. (1972). Separate sample logistic discrimination. Biometrika, 59, 19-36.

ANDERSON, J.A. (1974). Diagnosis by logistic discriminant function : further practical problems and results. Appl. Statist., 23, 397-404.

ANDERSON, J.A. (1975). Quadratic logistic discrimination. Biometrika, 62, 149-154.

ANDERSON, J.A. and BLAIR, V. (1982). Penalised maximimum likelihood estimation in logistic regression and discrimination. Biometrika, 69, 123-136.

ANDERSON, J.A. and RICHARDSON, S.C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. Technometrics, 21, 71-78.

ANDERSON, J.A., WHALEY, K., WILLIAMSON, J. and BUCHANAN, W.W. (1972). A statistical aid to the diagnosis of keratoconjunctivitis sicca. Biometrika, 71, 353-360.

ANDERSON, J.A. (1984). Regression and ordered categorical variables (with Discussion). J. R. Statist. Soc. B, 46, 1-30.

ANDERSON, J.A. and PHILLIPS, P.R. (1981). Regression, discrimination, and measurement models for ordered variables. Appl. Statist., 30, 22-31.

ANDERSON, T.W. (1951). Classification by multivariate analysis. Psychometrika, 16, 631-650.

ANDERSON, T.W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley, New York.

ANDERSON, T.W. and BAHADUR, R.R. (1962). Classification into two multivariate normal distributions with different covariance matrices. Ann. Math. Statist., 33, 420-431.

ASHBY, D., POCOCK, S.J. and SHAPER, A.G. (1986). Ordered polychotomous regression: an example relating serum biochemistry and haematology to alcohol consumption. Appl. Statist., 35, 289-301.

AYER, M., BRUNK, H.D., EWING, G.M., REID, W.T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. Ann. Math. Statist., 26, 641-647.

BAHADUR, R.R. (1961). A representation of the joint distribution of responses to n dichotomous items. In Studies in Item Analysis and Prediction, ed. H. Solomon, 158-168. Stanford University Press, Palo Alto, CA.

BARLOW, R.E., BARTHOLOMEW, D.J., BREMNER, J.M. and BRUNK, H.D. (1972). Statistical Inference under Order Restrictions. Wiley, New York.

BARTLETT, M.S. (1963). Statistical estimation of density functions. Sankhya A, 25, 245-254.

BARTLETT, M.S. and PLEASE, N.W. (1963). Discrimination in the case of zero mean differences. Biometrika, 50, 17-21.

BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975). Discrete Multivariate Analysis. The MIT Press, Cambridge, MA.

BLATTENBERGER, G. and LAD, F. (1985). Separating the Brier score into calibration and refinement components : a graphical exposition. The American Statistician, 39, 26-32.

BONEVA, I.R., KENDALL, D.G. and STEVANOV, I. (1971). Spline transformations: three diagnostic aids for the statistical data-analyst (with Discussion). J. R. Statist. Soc. B, 33, 1-71.

BOWMAN, A.W. (1980). A note on consistency of the kernel method for the analysis of categorical data. Biometrika, 67, 682-684.

BOWMAN, A.W. (1981). Some Aspects of Density Estimation by the Kernel Method. Ph.D. dissertation, University of Glasgow.

BOWMAN, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71, 353-360.

BOWMAN, A.W. (1985). A comparative study of some kernel-based nonparametric density estimators. J. Statist. Comput. Simul., 21, 313-327.

BREIMAN, L., MEISEL, W. and PURCELL, E. (1977). Variable kernel estimates of multivariate densities. Technometrics, 19, 135-144.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. (1984). Classification and Regression Trees. Wadsworth Inc., Belmont, CA.

BRIER, G.W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78, 1-3.

BRIL, G., DYKSTRA, R., PILLERS, C. and ROBERTSON, T. (1984). Algorithm AS206. Isotonic regression in two independent variables. Appl. Statist., 33, 352-357.

BROWN, P.J. and RUNDELL, P.W.K. (1985). Kernel estimates for categorical data. Technometrics, 27, 293-299.

BRUNK, H.D. (1978). Univariate density estimation by orthogonal series. Biometrika, 65, 521-528.

BRUNK, H.D. and PIERCE, D.A. (1974). Estimation of discrete multivariate densities for computer-aided differential diagnosis of disease. Biometrika, 61, 493-499.

BUTLER, W.J. and KRONMAL, R.A. (1985). Discrimination with polychotomous predictor variables using orthogonal functions. J. Amer. Statist. Assoc., 80, 443-448.

CACOULLOS, T. (1966). Estimation of a multivariate density. Ann. Inst. Stat. Maths., 18, 179-189.

CENCOV, N.N. (1962). Evaluation of an unknown distribution density from observations. Soviet Math., 3, 1559-1562.

CHANG, P.C. and AFIFI, A.A. (1974). Classification based on dichotomous and continuous variables. J. Amer. Statist. Assoc., 69, 336-339.

CHERNICK, M.R., MURTHY, V.K. and NEALY, C.D. (1985). Application of bootstrap and other resampling techniques: evaluation of classifier performance. Pattern Recognition Letters, 3, 167-178.

CHOI, S.C. (ed.) (1986). Statistical Methods of Discrimination and Classification: Advances in Theory and Applications. Pergamon Press, Oxford. (Comp. and Maths. with Applics., 12A.)

CHOW, Y-S., GEMAN, S. and WU, L-D. (1983). Consistent cross-validated density estimation. Ann. Statist., 11, 25-38.

CLUNIES-ROSS, C.W. and RIFFENBURGH, R.H. (1960). Geometry and linear discrimination. Biometrika, 47, 185-189.

COCHRAN, W.G. (1964). On the performance of the linear discriminant function. Technometrics, 6, 179-190.

COCHRAN, W.G. and HOPKINS, C. (1961). Some classification methods with multivariate qualitative data. Biometrics, 17, 10-32.

COPAS, J.B., and FRYER, M.J. (1980). Density estimation and suicide risks in psychiatric treatment. J. R. Statist. Soc. A, 143, 167-176.

CORMACK, R.M. (1971). A review of classification (with Discusssion). J. R. Statist. Soc. A, 134, 321-367.

COVER, T.M. and HART, P.E. (1967). Nearest neighbour pattern classification. IEEE Trans. Inf. Theory, IT-13, 21-27.

COX, D.R. (1970). Analysis of Binary Data. Methuen, London.

CRAN, G.W. (1980). Algorithm AS149 : Amalgamation of means in the case of simple ordering. Appl. Statist., 29, 209-211.

CRITCHLEY, F. and FORD, I. (1985). Interval estimation in discrimination: the multivariate normal equal covariance case. Biometrika, 72, 109-116.

DAVIS, A.W. (1987). Moments of linear discriminant functions, and an asymptotic confidence interval for the log odds ratio. Biometrika, 74, 829-840.

DAWID, A.P. (1976). Properties of diagnostic data distributions. Biometrics, 32, 647-658.

DAY, N.E. and KERRIDGE, D.F. (1967). A general maximum likelihood discriminant. Biometrics, 23, 313-323.

DE DOMBAL, F.T., LEAPER, D.J., STANILAND, J.R., McCANN, A.P. and HORROCKS, J.C. (1972). Computer-aided diagnosis of acute abdominal pain. British Medical Journal, 2, 9-13.

DE MONTRICHER, G.F., TAPIA, R.A. and THOMPSON, J.R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. Ann. Statist., 3, 1329-1348.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). J. R. Statist. Soc. B, 39, 1-38.

DILLON, W.R. and GOLDSTEIN, M. (1978). On the performance of some multinomial classification rules. J. Amer. Statist. Assoc., 73, 305-313.

DIXON, W.J. (ed.) (1985). BMDP Statistical Software. University of California Press, Berkeley, CA.

DRAPER, N.R. and SMITH, H. (1966). Applied Regression Analysis. Wiley, New York.

DUIN, R.P.W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Trans. Computers, C-25, 1175-1179.

DUNN, O.J. (1971). Some expected values for probabilities of correct classification in discriminant analysis. Technometrics, 13, 345-353.

DYKSTRA, R.L. (1981). An isotonic regression algorithm. J. Statist. Planning and Inference, 5, 355-363.

DYKSTRA, R.L. (1983). An algorithm for restricted least squares regression. J. Amer. Statist. Assoc., 78, 837-842.

DYKSTRA, R.L. and ROBERTSON, T. (1982). An algorithm for isotonic regression for two or more independent variables. Ann. Statist., 10, 708-716.

EFRON, B. (1975). The efficiency of logistic regression compared to Normal discriminant analysis. J. Amer. Statist. Assoc., 70, 892-898.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. Ann. Statist., 7, 1-26.

EFRON, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. J. Amer. Statist. Assoc., 78, 316-331.

EPANECHNIKOV, V.A. (1969). Nonparametric estimation of a multidimensional probability density. Theory of Prob. Applics., 14, 153-158.

EVERITT, B.S. (1984). An Introduction to Latent Variable Models. Chapman and Hall, London.

FIENBERG, S.E. and HOLLAND, P.W. (1973). Simultaneous estimation of multinomial cell probabilities. J. Amer. Statist. Assoc., 68, 683-691.

FINNEY, D.J. (1971). Probit Analysis (3rd edition). Cambridge University Press.

FISHER, R.A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eugen., 7, 179-188.

FISHER, L. and VAN NESS, J.W. (1973). Admissible discriminant analysis. J. Amer. Statist. Assoc., 68, 603-607.

FIX, E. and HODGES, J. (1951). Discriminatory analysis, nonparametric discrimination: consistency properties. Report no. 4, project no. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

FRYER, M.J. (1976). Some errors associated with the nonparametric estimation of density functions. J. Inst. Maths. Applics., 18, 371-380.

FRYER, M.J. (1977). A review of some nonparametric methods of density estimation. J. Inst. Maths. Applics., 20, 335-354.

FUKUNAGA, K. (1972). Introduction to Statistical Pattern Recognition. Academic Press, New York.

FUKUNAGA, K. and KESSELL, D.L. (1971). Estimation of classification error. IEEE Trans. Computers, C-20, 1521-1527.

GART, J.J. and ZWEIFEL, J.R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. Biometrika, 54, 181-187.

GEBHARDT, F. (1970). An algorithm for monotone regression with one or more independent variables. Biometrika, 57, 263-271.

GEISSER, S. (1964). Posterior odds for multivariate normal classifications. J. R. Statist. Soc., 26, 69-76.

GILBERT, E.S. (1968). On discrimination using qualitative variables. J. Amer. Statist. Assoc., 63, 1399-1412.

GILBERT, E.S. (1969). The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. Biometrics, 25, 505-516.

GLICK, N. (1972). Sample-based classification procedures derived from density estimators. J. Amer. Statist. Assoc., 67, 116-121.

GLICK, N. (1973). Sample-based multinomial classification. Biometrics, 29, 241-256.

GLICK, N. (1978). Additive estimators for probabilities of correct classification. Pattern Recognition, 10, 211-222.

GOLDSTEIN, M. and RABINOWITZ, M. (1975). Selection of variates for the two-group multinomial classification problem. J. Amer. Statist. Assoc., 70, 776-781.

GOLDSTEIN, M. and DILLON, W.R. (1978). Discrete Discriminant Analysis. Wiley, New York.

GOOD, I.J. (1965). The Estimation of Probabilities. The MIT Press, Cambridge, MA.

GOOD, I.J. and GASKINS, R.A. (1971). Nonparametric roughness penalties for probability densities. Biometrika, 58, 255-277.

GOOD, I.J. and GASKINS, R.A. (1980). Density estimation and bump-hunting by the penalised likelihood method exemplified by scattering and meteorite data. J. Amer. Statist. Assoc., 75, 42-73.

GOODMAN, L.A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. J. Amer. Statist. Assoc., 65, 226-256.

GOODMAN, L.A. (1971). The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics, 13, 33-61.

GOODMAN, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. J. Amer. Statist. Assoc., 74, 537-552.

GOODMAN, L.A. (1983). The analysis of dependence in cross-classifications having ordered categories using log-linear models for frequencies and log-linear models for odds. Biometrics, 39, 149-160.

GRIZZLE, J.E., STARMER, C.F. and KOCH, G.G. (1969). Analysis of categorical data by linear models. Biometrics, 25, 489-504.

HABBEMA, J.D.F. and HERMANS, J. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. Technometrics, 19, 487-493.

HABBEMA, J.D.F., HERMANS, J., and VAN DEN BROEK, K. (1974). A step-wise discriminant analysis program using density estimation. In Compstat 1974, ed. G. Bruckman, 101-110. Physica Verlag, Vienna.

HABBEMA, J.D.F., HERMANS, J., and VAN DER BURGT, A.T. (1974). Cases of doubt in allocation problems. Biometrika, 61, 313-324.

HABBEMA, J.D.F., HERMANS, J., and REMME, J. (1978). Variable kernel density estimation in discriminant analysis. In Compstat 1978, ed. L.C.A. Corsten and J. Hermans, 178-185. Physica Verlag, Vienna.

HABBEMA, J.D.F., HILDEN, J. and BJERREGAARD, B. (1978). The measurement of performance in probabilistic diagnosis. I. The problem, descriptive tools, and measures based on classification matrices. Meth. Inform. Med., 17, 217-226.

HABBEMA, J.D.F. and HILDEN, J. (1981). The measurement of performance in probabilistic diagnosis. IV. Utility considerations in therapeutics and prognostics. Meth. Inform. Med., 20, 80-96.

HABBEMA, J.D.F., HILDEN, J. and BJERREGAARD, B. (1981). The measurement of performance in probabilistic diagnosis. V. General recommendations. Meth. Inform. Med., 20, 97-100.

HABERMAN, S.J. (1974). Log-linear models for frequency tables with ordered classifications. Biometrics, 30, 589-600.

HALL, P. (1981a). On nonparametric multivariate binary discrimination. Biometrika, 68, 287-294.

HALL, P. (1981b). Optimal near neighbour estimator for use in discriminant analysis. Biometrika, 68, 572-575.

HALL, P. (1983a). Large sample optimality of least squares cross-validation in density estimation. Ann. Statist., 11, 1156-1174.

HALL, P. (1983b). Orthogonal series methods for both qualitative and quantitative data. Ann. Statist., 11, 1004-1007.

HALL, P. (1987). On Kullback-Leibler loss and density estimation. Ann. Statist., 15, 1491-1519.

HALL, P. and WAND, M.P. (1988). On nonparametric discrimination using density differences. Biometrika, 75, 541-547.

HALPERIN, M., BLACKWELDER, W.C. and VERTER, J.L. (1971). Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. J. Chron. Dis., 24, 125-158.

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

HAND, D.J. (1981b). Branch and bound in statistical data analysis. The Statistician, 30, 1-13.

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

HAND, D.J. (1983). A comparison of two methods of discriminant analysis applied to binary data. Biometrics, 39, 683-694.

HAND, D.J. (1986a). Recent advances in error rate estimation. Pattern Recognition Letters, 4, 335-346.

HAND, D.J. (1986b). Estimating class sizes by adjusting fallible classsifier results. In CHOI, S.C. (ed.) (1986). Statistical Methods of Discrimination and Classification: Advances in Theory and Applications, 289-299. Pergamon Press, Oxford. (Comp. and Maths. with Applics., 12A.)

HELLMAN, M.E. (1970). The nearest neighbour classification rule with a reject option. IEEE Trans. Sys. Sci. Cyb., SSC-6, 179-185.

HILDEN, J. (1984). Statistical diagnosis based on conditional independence does not require it. Comput. Biol. Med., 14, 429-435.

HILDEN, J., HABBEMA, J.D.F., and BJERREGAARD, B. (1978a). The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. Meth. Inform. Med., 17, 227-237.

HILDEN, J., HABBEMA, J.D.F., and BJERREGAARD, B. (1978b). The measurement of performance in probabilistic diagnosis. III. Methods based on continuous functions of the diagnostic probabilities. Meth. Inform. Med., 17, 238-246.

HILL, M.A. (1979). Annotated computer output for regression analysis. BMDP-77 Technical Report no. 48. Health Services Computing Facility, UCLA.

HILLS, M. (1966). Allocation rules and their error rates (with Discussion). J. R. Statist. Soc. B, 28, 1-31.

HILLS, M. (1967). Discrimination and allocation with discrete data. Appl. Statist., 16, 237-250.

HOEL, P.G. and PETERSON, R.P. (1949). A solution to the problem of optimum classification. Ann. Math. Statist., 20, 433-438.

HOGG, R.V. (1979). Statistical robustness; one view of its use in applications today. The American Statistician, 79, 108-115.

HORROCKS, J.C., McCANN, A.P., STANILAND, J.R., LEAPER, D.J. and DE DOMBAL, F.T. (1972). Computer-aided diagnosis. Description of an adaptable system and operational experience with 2,034 cases. British Medical Journal, 2, 5-9.

JENNETT, B. and BOND, M. (1975). Assessment of outcome after severe head injury. Lancet i, 480-484.

JENNETT, B., TEASDALE, G., BRAAKMAN, R., MINDERHOUD, J. and KNILL-JONES, R. (1976). Predicting outcome in individual patients after severe head injury. Lancet i, 1031-1034.

JOHNSON, B.M. (1971). On the admissible estimators for certain fixed binomial problems. Ann. Math. Statist., 42, 1579-1587.

JONES, M.C. and LOTWICK, H.W. (1984). A remark on Algorithm AS 176. Kernel density estimation using the fast Fourier transform. Remark ASR50. Appl. Statist., 33, 120-122.

KERRIDGE, D.F. (1966). Contribution to the discussion of HILLS, M. (1966). Allocation rules and their error rates. J. R. Statist. Soc. B, 28, 1-31.

KIMELDORF, G.S. and WAHBA, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Statist., 41, 495-502.

KIMELDORF, G.S. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. J. Math. Anal. and Applics., 33, 82-95.

KITTLER, J. (1975). Mathematical methods of feature selection in pattern recognition. Int. J. Man-Machine Studies, 7, 609-637.

KLONIAS, V.K. (1984). On a class of nonparametric density and regression estimators. Ann. Statist., 12, 1263-1284.

KNOKE, J.D. (1986). The robust estimation of classification error rates. In CHOI, S.C. (ed.) (1986). Statistical Methods of Discrimination and Classification: Advances in Theory and Applications, 253-259. Pergamon Press, Oxford. (Comp. and Maths. with Applics., 12A.)

KRONMAL, R.A. and TARTER, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. J. Amer. Statist. Assoc., 63, 925-952.

KRZANOWSKI, W.J. (1975). Discrimination and classification using both binary and continuous variables. J. Amer. Statist. Assoc., 70, 782-790.

KRZANOWSKI, W.J. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. Technometrics, 19, 191-200.

KRZANOWSKI, W.J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. Biometrics, 36, 493-499.

KRZANOWSKI, W.J. (1983a). Distance between populations using mixed continuous and categorical variables. Biometrika, 70, 235-243.

KRZANOWSKI, W.J. (1983b). Stepwise location model choice in mixed variable discrimination. Appl. Statist., 32, 260-266.

KULLBACK, S. (1959). Information Theory and Statistics. Wiley, New York.

LACHENBRUCH, P.A. (1968). On expected probabilities of misclassification in discriminant analysis, necessary sample size and a relation with the multiple correlation coefficient. Biometrics, 24, 823-834.

LACHENBRUCH, P.A. (1975a). Discriminant Analysis. Hafner Press, New York.

LACHENBRUCH, P.A. (1975b). Zero-mean difference discrimination and the absolute linear discriminant function. Biometrika, 62, 397-401.

LACHENBRUCH, P.A. and MICKEY, M.R. (1968). Estimation of error rates in discriminant analysis. Technometrics, 10, 1-10.

LACHENBRUCH, P.A., SNEERINGER, C. and REVO, L.T. (1973). Robustness of the linear and quadratic functions to certain types of non-normality. Comm. Statist., 1, 39-56.

LANCASTER, H.O. (1969). The Chi-squared Distribution. Wiley, New York.

LAUDER, I.J. (1983). Direct kernel assessment of diagnostic probabilities. Biometrika, 70, 251-256.

LEE, C.C. (1983). The min-max algorithm and isotonic regression. Ann. Statist., 11, 467-477.

LEONARD, T. (1973). A Bayesian method for histograms. Biometrika, 60, 297-308.

LEONARD, T. (1975). Bayesian estimation models for two-way tables. J. R. Statist. Soc. B, 37, 23-37.

LEONARD, T. (1977). A Bayesian approach to some multinomial estimation and pretesting problems. J. Amer. Statist. Assoc., 72, 869-874.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information (with Discussion). J. R. Statist. Soc. B, 40, 113-116.

LINHART, H. (1959). Techniques for discriminant analysis with discrete variables. Metrika, 2, 138-149.

LOFTSGAARDEN, F. and QUESENBERRY, C. (1965). A nonparametric estimate of a multivariate density function. Ann. Statist., 36, 1049-1051.

MACK, Y.P. and ROSENBLATT, M. (1979). Multivariate k-nearest neighbour density estimates. J. Multivariate Anal., 9, 1-15.

MARDIA, K.V., KENT, J.T. and BIBBY, J.M. (1979). Multivariate Analysis. Academic Press, London.

MARKS, S. and DUNN, O.J. (1974). Discriminant functions when covariance matrices are unequal. J. Amer. Statist. Assoc., 69, 555-559.

MARTIN, D.C. and BRADLEY, R.A. (1972). Probability models, estimation, and classification for multivariate dichotomous populations. Biometrics, 28, 203-221.

MATUSITA, K. (1955). Decision rules, based on the distance, for problems of fit, two samples, and estimation. Ann. Math. Statist., 26, 631-640.

McCULLAGH, P. (1980). Regression models for ordinal data (with Discussion). J. R. Statist. Soc. B, 42, 109-142.

McLACHLAN, G.J. (1974a). An asymptotic unbiased technique for estimating the error rates in discriminant analysis. Biometrics, 30, 239-249.

McLACHLAN, G.J. (1974b). Estimation of the errors of misclassification on the criterion of asymptotic mean square error. Technometrics, 16, 255-260.

McLACHLAN, G.J. (1976). A criterion for selecting variables for the linear discriminant function. Biometrics, 32, 529-534.

McLACHLAN, G.J. (1986). Assessing the performance of an allocation rule. In CHOI, S.C. (ed.) (1986). Statistical Methods of Discrimination and Classification: Advances in Theory and Applications, 261-272. Pergamon Press, Oxford. (Comp. and Maths. with Applics., 12A.)

MILLER, K.S. (1964). Multivariate Gaussian Distributions. Wiley, New York.

MOORE, D.H. (1973). Evaluation of five discrimination procedures for binary variables. J. Amer. Statist. Assoc., 68, 399-404.

MOORE, D.S. and YACKEL, J.W. (1977). Consistency properties of nearest neighbour density function estimates. Ann. Statist., 5, 143-154.

MORAN, M.A. and MURPHY, B.J. (1979). A closer look at two alternative methods of statistical discrimination. Appl. Statist., 28, 223-232.

MURPHY, B.J. and MORAN, M.A. (1986). Parametric and kernel density methods in discriminant analysis: another comparison. In CHOI, S.C. (ed.) (1986). Statistical Methods of Discrimination and Classification: Advances in Theory and Applications, 197-207. Pergamon Press, Oxford. (Comp. and Maths. with Applics., 12A.)

MURRAY, G.D. (1977a). A cautionary note on selection of variables in discriminant analysis. Appl. Statist., 26, 246-250.

MURRAY, G.D. (1977b). A note on the estimation of probability functions. Biometrika, 64, 150-152.

MURRAY, G.D. and TITTERINGTON, D.M. (1978). Estimation problems with data from a mixture. Appl. Statist., 27, 325-334.

MURRAY, G.D. (1983). Nonconvergence of the minimax order algorithm. Biometrika, 70, 490-491.

MURRAY, G.D. and WILSON, A.J. (1986). Zero weighted cells. A remark on AS206 : Isotonic regression in two independent variates. Appl. Statist., 35, 312-314.

MURRAY, G.D. and WILSON, A.J. (1987). Correction to algorithm AS206 : Isotonic regression in two independent variates. Appl. Statist., 36, 120.

NADARAYA, E.A. (1974). On the integral mean square error of some estimates for the density function. Theory Prob. Applics., 19, 133-141.

NAG Fortran Library Manual, Mark 11. (1984). Numerical Algorithms Group, Oxford.

NELDER, J.A. (1974). Log-linear models for contingency tables: a generalisation of classical least squares. Appl. Statist., 23, 323-329.

OKAMOTO, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. Ann. Math. Statist., 34, 1286-1301. (Correction. (1968). Ann. Math. Statist. 39, 1358.)

O'NEILL, T.J. (1980). The general distribution of the error rate of a classification procedure with application to logistic regression discrimination. J. Amer. Statist. Assoc., 75, 154-160.

OTT, J. and KRONMAL, R.A. (1976). Some classification procedures for multivariate binary data using orthogonal functions. J. Amer. Statist. Assoc., 71, 391-399.

PAGE, J.T. (1985). Error rate estimation in discriminant analysis. Technometrics, 27, 189-198.

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

PFEIFFER, K.P. (1985). Stepwise variable selection and maximum likelihood estimation of smoothing factors of kernel functions for nonparametric discriminant functions evaluated by different criteria. Comp. Biomed. Res., 18, 46-61.

RAATGEVER, J.W. and DUIN, R.P.W. (1978). On the variable kernel model for nonparametric density estimation. In Compstat 1978, ed. L.C.A. Corsten and J. Hermans, 524-533. Physica Verlag, Vienna.

RAO, C.R. (1965). Linear Statistical Inference and Its Applications. Wiley, New York.

REMME, J., HABBEMA, J.D.F. and HERMANS, J. (1980). A simulative comparison of linear, quadratic and kernel discrimination. J. Statist. Comp. Simul., 11, 87-106.

RENCHER, A.C. and LARSON, S.F. (1980). Bias of Wilks' Λ in stepwise discriminant analysis. Technometrics, 22, 349-356.

REVO, L.T. (1970). On Classification with Certain Types of Ordered Qualitative Data: an Evaluation of Several Procedures. Unpublished Ph.D. dissertation, University of North Carolina, Chapel Hill.

RIGBY, R.A. (1982). A credibility interval for the probability that a new observation belongs to one of two multivariate normal populations. J. R. Statist. Soc. B, 44, 212-220.

RIPLEY, B.D. and TAYLOR, C.C. (1987). Pattern Recognition. Science Progress, Oxford., 71, 413-428.

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

ROSENBLATT, M. (1979). Global measures of deviation for kernel and nearest neighbour estimates. In Gasser, T. and Rosenblatt, M. (eds.), Smoothing Techniques for Curve Estimation, 181-190. Lecture Notes in Mathematics, 757. Springer Verlag, Berlin.

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

SACKS, J. and YLVISAKER, D. (1981). Asymptotically optimum kernels for density estimation at a point. Ann. Statist., 9, 334-346.

SCHUCANY, W.R. and SOMMERS, J.P. (1977). Improvement of kernel type density estimators. J. Amer. Statist. Assoc., 72, 420-423.

SCHUSTER, E.F. and GREGORY, G.G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. In Eddy, W.F. (ed.), Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, 295-298. Springer Verlag, New York.

SCOTT, D.W. and FACTOR, L.E. (1981). Monte Carlo study of three data-based nonparametric density estimators. J. Amer. Statist. Assoc., 76, 9-15.

SCOTT, D.W., TAPIA, R.A. and THOMPSON, J.R. (1977). Kernel density estimation revisited. Nonlinear Analysis, Theory, Methods and Applications, 1, 339-372.

SHAPIRO, A.R. (1977). The evaluation of clinical prediction. New England Journal of Medicine, 296, 1509-1514.

SILVERMAN, B.W. (1978a). Density ratios, empirical likelihood and cot death. Appl. Statist., 27, 26-33.

SILVERMAN, B.W. (1978b). Choosing the window width when estimating a density. Biometrika, 65, 1-11.

SILVERMAN, B.W. (1978c). Contribution to the discussion of LEONARD, T. (1978). Density estimation, stochastic processes and prior information. J. R. Statist. Soc. B, 40, 113-146.

SILVERMAN, B.W. (1982a). On the estimation of a probability density function by the maximum penalized likeihood method. Ann. Statist., 10, 795-810.

SILVERMAN, B.W. (1982b). Kernel density estimation using the fast Fourier transform. Algorithm AS 176. Appl. Statist., 31, 93-97.

SILVERMAN, B.W. (1984a). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. J. Amer. Statist. Assoc., 79, 584-589.

SILVERMAN, B.W. (1984b). Spline smoothing: the equivalent variable kernel method. Ann. Statist., 12, 898-916.

SILVERMAN, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with Discussion). J. R. Statist. Soc. B, 47, 1-52.

SILVERMAN, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

SIMON, G. (1974). Alternative analyses for the singly-ordered contingency table. J. Amer. Statist. Assoc., 69, 971-976.

SIMONOFF, J.S. (1983). A penalty function approach to smoothing large sparse contingency tables. Ann. Statist., 11, 208-218.

SKENE, A.M. (1978). Discrimination using latent structure models. In Compstat 1978, ed. L.C.A. Corsten and J. Hermans, 199-211. Physica Verlag, Vienna.

SMITH, C.A.B. (1947). Some examples of discrimination. Ann. Eugen., 13, 272-282.

SNAPINN, S.M. and KNOKE, J.D. (1984). Classification error rate estimators evaluated by unconditional mean squared error. Technometrics, 26, 371-378.

SNAPINN, S.M. and KNOKE, J.D. (1985). An evaluation of smoothed classification error-rate estimators. Technometrics, 27, 199-206.

SORUM, M.J. (1971). Estimating the conditional probability of misclassification. Technometrics, 13, 333-343.

SORUM, M.J. (1972). Estimating the expected and optimal probabilities of misclassification. Technometrics, 14, 935-943.

SORUM, M.J. (1973). Estimating the expected probability of misclassification for a rule based on the linear discriminant function: univariate normal case. Technometrics, 15, 329-339.

STONE, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. Ann. Statist., 12, 1285-1297.

STONE, M. (1974a). Cross-validatory choice and assessment of statistical predictions (with Discussion). J. R. Statist. Soc. B, 36, 111-147.

STONE, M. (1974b). Cross-validation and multinomial prediction. Biometrika, 61, 509-515.

TARTER, M.E. and KRONMAL, R.A. (1976). An introduction to the implementation and theory of nonparametric density estimation. The American Statistician, 30, 105-112.

TEASDALE, G. and JENNETT, B. (1974). Assessment of coma and impaired consciousness. A practical scale. Lancet ii, 81-84.

TEASDALE, G., MURRAY, G., PARKER, L. and JENNETT, B. (1979). Adding up the Glasgow Coma Score. Acta Neurochirurgica Suppl., 28, 13-16.

TERRELL, G.R. and SCOTT, D.W. (1980). On improving convergence rates for nonnegative kernel density estimators. Ann. Statist., 8, 1160-1163.

TITTERINGTON, D.M. (1977). Analysis of incomplete binary data by the kernel method. Biometrika, 64, 455-460.

TITTERINGTON, D.M. (1980). A comparative study of kernel-based density estimates for categorical data. Technometrics, 22, 259-268.

TITTERINGTON, D.M. and BOWMAN, A.W. (1985). A comparative study of smoothing procedures for ordered categorical data. J. Statist. Comput. Simul., 21, 291-312.

TITTERINGTON, D.M., MURRAY, G.D., MURRAY, L.S., SPIEGELHALTER, D.J., SKENE, A.M., HABBEMA, J.D.F. and GELPKE, G.J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients (with Discussion). J. R. Statist. Soc. A, 144, 145-175.

TITTERINGTON, D.M., SMITH, A.F.M. and MAKOV, U.E. (1985). Statistical Analysis of Finite Mixture Distributions. Wiley, Chichester.

TOUSSAINT, G.T. (1974). Bibliography on estimation of misclassification. IEEE Trans. Inf. Theory, IT-20, 472-479.

TRAMPISCH, H.J. (1978). Classical discriminant analysis and Lancaster models for qualitative data. In Compstat 1978, ed. L.C.A. Corsten and J. Hermans, 205-210. Physica Verlag, Vienna.

TUTZ, G.E. (1985). Smoothed additive estimators for non-error rates in multiple discriminant analysis. Pattern Recognition, 18, 151-159.

TUTZ, G. (1986). An alternative choice of smoothing parameters for kernel-based density estimates in discrete discriminant analysis. Biometrika, 73, 405-411.

VAN NESS, J.W. (1979). On the effects of dimension in discriminant analysis for unequal covariance populations. Technometrics, 21, 119-127.

VAN NESS, J.W. and SIMPSON, C. (1976). On the effects of dimension in discriminant analysis. Technometrics, 18, 175-187.

VAN RYZIN, J. (1966). Bayes risk consistency of classification procedures using density estimation. Sankhya A, 28, 261-270.

VAN RYZIN, J. (1969). On strong consistency of density estimates. Ann. Math. Statist., 40, 1765-1772.

VICTOR, N., TRAMPISCH, H.J., and ZENTGRAF, R. (1974). Diagnostic rules for qualitative variables with interactions. Meth. Inform. Med., 13, 184-186.

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

WAHBA, G. (1971). A polynomial algorithm for density estimation. Ann. Math. Statist., 42, 1870-1886.

WAHBA, G. (1975). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. Ann. Statist. 3, 15-29.

WAHBA, G. (1976). Histosplines with knots which are order statistics. J. R. Statist. Soc. B, 38, 140-151.

WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. J. R. Statist. Soc. B, 40, 364-372.

WAHBA, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. J. R. Statist. Soc. B, 45, 133-150.

WAHBA, G. and WOLD, S. (1975). A completely automatic French curve: fitting spline functions by cross-validation. Comm. Statist., 4, 1-18.

WALD, A. (1944). On a statistical problem arising in the classification of an individual into 2 groups. Ann. Math. Statist., 15, 145-163.

WANG, M.C. and VAN RYZIN, J. (1981). A class of smooth estimators for discrete distributions. Biometrika, 68, 301-309.

WARNER, H.R., TORONTO, A.F., VEASEY, L.G. and STEPHENSON, R. (1961). A mathematical approach to medical diagnosis. Application to congenital heart disease. J. Amer. Med. Assoc., 177, 177-183.

WECKER, W.E. and ANSLEY, C.F. (1983). The signal extraction approach to non-linear regression and spline smoothing. J. Amer. Statist. Assoc., 78, 81-89.

WEGMAN, E.J. (1972). Nonparametric probability density estimation: I. A summary of available methods. Technometrics, 14, 533-546.

WEGMAN, E.J. and WRIGHT, I.W. (1983). Splines in statistics. J. Amer. Statist. Assoc., 78, 351-365.

WEINER, J.M. and DUNN, O.J. (1966). Elimination of variables in linear discrimination problems. Biometrics, 22, 268-275.

WELCH, B.L. (1939). Note on discriminant functions. Biometrika, 31, 218-220.

WERTZ, W. and SCHNEIDER, B. (1979). Statistical density estimation : a bibliography. Int. Stat. Rev., 47, 155-175.

WHITTLE, P. (1958). On the smoothing of probability density functions. J. Statist. Soc. B, 20, 334-343.

WOLLAN, P.C. and DYKSTRA, R.L. (1987). Maximising linear inequality constrained Mahalanobis distances. Algorithm AS225, Appl. Statist., 36, 234-240.

WOODROOFE, M. (1970). On choosing a delta sequence. Ann. Math. Statist., 41, 1665-1675.

WRIGHT, I.W. and WEGMAN, E.J. (1980). Isotonic, convex and related splines. Ann. Statist., 8, 1023-1035.

YAMATO, H. (1972). Some statistical properties of estimators of density and distribution functions. Bull. Math. Statist., 15, 113-131 (and Correction, p. 133).

ZENTGRAF, R. (1975). A note on Lancaster's definition of higher-order interactions. Biometrika, 62, 375-378.