KERNEL DENSITY ESTIMATION, BAYESIAN INFERENCE AND RANDOM EFFECTS MODEL

by

KAREN PUI-SHAN CHAN

Thesis is submitted for the degree of Doctor of Philosophy

in the University of Edinburgh

1990

DECLARATION

The following record of research work is submitted as a thesis for the degree of Doctor of Philosophy in the University of Edinburgh, having being submitted for no other degree. The research work was carried out under the supervision of Dr. C.G.G. Aitken. Except where acknowledgement is made, the work is original.

KAREN PUI-SHAN CHAN

Dedicated to my parents

We work not only to produce but to give value to time.

—— Eugène Delacroix (1798-1863)

TABLE OF CONTENTS

ABSTRACT

This thesis contains the results of a study in kernel density estimation. Bayesian inference and the random effects model, with application to a forensic problem.

Estimation of the Bayes' factor in a forensic science problem involved the derivation of predictive distributions in non-standard situations. The distribution of the values of a characteristic of interest among different items in forensic science problems is often non-Normal. Background, or training, data were available to assist in the estimation of the distribution for measurements on cat and dog hairs. An informative prior, based on the kernel method of density estimation, was used to derive the appropriate predictive distributions. The training data may be considered to be derived from a random effects model. This was taken into consideration in modelling the Bayes' factor. The usual assumption of the random factor being Normally distributed is unrealistic, so a kernel density estimate was used as the distribution of the unknown random factor. Two kernel methods were employed: the ordinary and adaptive kernel methods. The adaptive kernel method allowed for the longer tail, where little information was available.

Formulae for the Bayes' factor in a forensic science context were derived assuming the training data were grouped or not grouped (for example, hairs from one cat would be thought of as belonging to the same group), and that the within-group variance was or was not known. The Bayes' factor, assuming known within-group variance, for the training data, grouped or not grouped, was extended to the multivariate case. The method was applied to a practical example in a bivariate situation. Similar modelling of the Bayes' factor was derived to cope with a particular form of mixture data. Boundary effects were also taken into consideration.

Application of kernel density estimation to make inferences about the variance components under the random effects model was studied. Employing the maximum likelihood estimation method, it was shown that the between-group variance and the smoothing parameter in the kernel density estimate were related. They were not identifiable separately. With the smoothing parameter fixed at some predetermined value, the within- and between-group variance estimates from the proposed model were equivalent to the usual ANOVA estimates. Within the Bayesian framework, posterior distributions for the variance components, using various prior distributions for the parameters, were derived incorporating kernel density functions. The modes of these posterior distributions were used as estimates for the variance components.

A Student-t kernel within a Bayesian framework was derived after introduction of a prior for the smoothing parameter. Two methods of obtaining hyper-parameters for the prior were suggested, both involved empirical Bayes methods. They were a modified leave-one-out maximum likelihood method and a method of moments based on the optimum smoothing parameter determined from Normality assumption.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Dr. C.G.G. Aitken and thank him for his encourgement, patience and endless help, without which I certainly would not have completed this thesis. I also wish to thank the staff in the Department of Statistics, especially Mr. P.R. Fisk and Dr. C. Theobald for their help and comments.

I am very grateful to Carnegie Trust for the Universities of Scotland for providing me financial support and an Overseas Research Student Award for subsidising part of the fee.

Also I would like to take this opportunity to thank my family and friends in Hong Kong as well as in Scotland for their constant encouragement and moral support. My thanks to the PGs in the Department who make Ph.D. life so enjoyable and Mr. David McGuinness for his help and encouragement.

Last but not least, I thank the heavenly Father and the faith of the Roman Catholic Church for seeing me through.

SOME USEFUL NOTATIONS

The following notations are used in this thesis:

Symbol	Usual representation (unless stated separately)
$X/\underset{\sim}{X}$	An univariate/multivariate random variable which represents control data and has a known source
$Y/\underset{\sim}{Y}$	An univariate/multivariate random variable which represents recovered data and has an unknown source
$Z/\underset{\sim}{Z}$	An univariate/multivariate random variable represents training data (TS for short)
T, U, V, W	Univariate random variables for general purposes
z_{ij}	j^{th} observation from the i^{th} group of the training data
n	Number of groups in the training data Z
J	Common number of observations within each group in the training data
N (= n×J)	Total number of observations in the training data
m	Number of observations arise from X
r	Number of observations arise from Y
μ_i	Population mean from the i^{th} group
τ	(= σ^{-2}) Precision of population mean
α, β	Hyperparameters of the prior for τ
σ^2 or σ_e^2	Within-group variance
σ_a^2	Between-group variance
λ	Smoothing parameter
δ	Sensitivity parameter

CHAPTER 1

INTRODUCTION

1.1 Aims

Bayesian approaches to statistical analysis are often based on parametric models. For instance, the underlying density function is assumed to be a member of a specified family, $\{f(t;\theta): t \in \mathbb{R}, \theta \subseteq \Theta\}$ where Θ is a set of possible values of the parameter θ and $\mathbb{R}$ is the sample space. The standard Bayesian model adds two assumptions: (1) that the parameter θ can be regarded as a random variable, and (2) the prior distribution π of this random variable is known, either to make inferences about the parameter θ or to obtain marginal distributions of the data. Lindley (1965), Box and Tiao (1973), Press (1982) and among others employ improper types of prior distribution for the parameter. The resultant distributions may be undefinable and mathematically unacceptable. Nevertheless one could adopt the approach suggested by Raiffa and Schlaifer (1961), who introduced conjugate prior distribution. As a result if the underlying density function of the data belongs to a parametric family, then the conjugate prior also belongs to a parametric family. However, this is often not the case especially in a random effects model situation, while the within group observations are Normally distributed but the between group random factors are usually not. Assuming the between group random factor is distributed Normally may lead to misleading results and invalid model (Scheffé (1959), Tiao and Ali (1971)). It is possible to construct non-parametric estimates of the density function for the random variable in a

context of empirical Bayes approach. This provides the possibility of using such estimates to provide a means of non-parametric modelling and analysis. The aim of the major part of this thesis is to develop the ideas of implementing the kernel density estimation method in Bayesian modelling and making use of additional information such as the training data. Development of these ideas were motivated by a forensic problem.

In order to allow the concept described above parallel to the Bayesian framework, it would be more acceptable if we use a Bayesain approach to estimate the density of the random factors. Bayesian models are applicable when the distribution F underlying the data is unknown and can itself be thought of as being generated by some random mechanism. Much of the Bayesian approach to non-parametric density estimation is concentrated on inducing prior information about the unknown underlying density function f. In the context of non-parametric models, to make F random we must define a Probability P on a non-parametric family $\mathfrak{p}$ of distributions. At the same time, P should be a plausible probability distribution for F. A breakthrough in the problem of finding such a P was made by Ferguson (1973). He defined the Dirichlet prior P, the finite dimensional distributions which are the family of Dirichlet distributions. Wahba (1976) developed approximate estimates for the density f based upon Fourier expansions and Bayesian argument involving a covariance Kernel and a uniform prior estimates for f. The mean of the posterior distribution of f is used as the estimate for f. This approach also provides the posterior confidence intervals for f (see Wahba (1983)). Leonard (1978) proposed a parallel approach to Wahba's for

non-parametric estimation of the probability density based upon a finite number of observations and prior information about the smoothness of the density. Leonard's approach depends substantially on the particular choice of prior parameters. The Bayesian approach to the density estimation problem proposed here is different from those stated above, since in a kernel density estimation problem, the smoothing parameter λ may be regarded as a random variable depending on the data (see Loftsgaarden and Quesenbery (1965), Cover (1972)). So the method proposed here is based on an introduction of a prior density to the smoothing parameter after specifying the underlying density function f takes the kernel density form.

1.2 Background of the Forensic problem

In September last year one of the article heading in the "Scotsman" newspaper was 'Hair of the dog traps a Bogeyman', see Fig. 1.1. Robin Smith, the 'Bogeyman', tried to extract a million pounds by contaminating products from some of the top supermarket suppliers. However, he was finally trapped by his own dog because Forensic scientists identified one hair found in one of the contaminated products as belonging to a dog. This eventually leaded to his arrest and conviction. The hair in question matched with those hairs obtained from his dog.

Hair of the dog traps a Bogeyman

By JAMES ROUGVIE

ROBIN SMITH, a self-styled "Bogeyman" who thought he could get more than half a million pounds from some of the country's top supermarket suppliers, was finally trapped by the hair of the dog — ironically his devoted companion, a pedigree boxer called Abra.

Smith, 33, an uncommunicative loner lives on his own in his flat in the Stobswell area of Dundee after a divorce from his wife several years ago. He thought he had hatched a meticulous plot to force chain stores and manufacturers to part with more than £625,000 by threatening to spike their products.

His threats to Safeway, Littlewoods, and other manufacturers came amid a wave of copycat attacks on stores in which babyfood, tinned goods and nappies were dangerously contaminated.

But what Smith failed to realise was that his plan was neither as meticulous as he thought nor were the supermarkets standing by helplessly.

He made two blunders — his carefully prepared packages of spiked shampoo, which contained hair remover and glass particles, was also contaminated, by a single hair from his dog.

Further, he did not realise when he made his first contact that he was in direct touch with a senior policeman, one of a special squad from London assembled to combat the attacks being made on supermarkets throughout the spring and summer.

Smith started his campaign by sending tins of food to the managers of local Safeway and Littlewoods shops, threatening to contaminate their shelf products unless they both paid him £200,000.

Two cosmetics manufacturers, Proctor and Gamble and Elida Gibbs, were his next targets. He sent them bottles of shampoo and his demands were made in black envelopes signed "the Bogeyman".

He told the executives of the companies to place adverts in the "lost and found" columns of a Scottish daily paper in order to give him direct contact with senior management.

He then sent Rowntree McIntosh a tube of Smarties. One of the children's sweets had been hollowed out and an anti-depressive pill inserted.

But when Smith responded to the first advert, he was unaware Proctor and Gamble had already called in the special crime squad and on the end of the line was a police inspector.

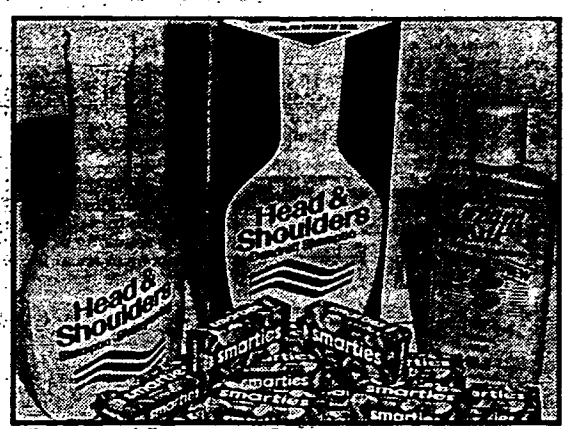
Astonishingly, Smith guessed correctly and asked "Is that a police inspector?" but appeared satisfied when told this was not the case. He made further calls to the special number always giving the code word "this is the Bogeyman calling."

Throughout May, forensic scientists, assisted by Proctor and Gamble's own laboratories, had minutely analysed the contents of the packages and discovered the one clue which would ultimately trap Smith. It was a single dog hair.

Police managed to locate the general area the calls were coming from — the Stobswell area of Dundee where Smith had his flat.

Tayside police were told to stake out every one of the 45-50 public phone boxes and watch for a man with a dog.

On 5 June, Smith strolled the 100 yards from his flat to the nearest phone box and tethered Abra there. He was on the phone when police opened the box.



"Bogeyman's" targets: Shampoo spiked with hair remover and Smarties with an anti-depression pill

Fig. 1.1 News article from the 'Scotsman' newspaper on 14th, September 1989.

Locard's principle in forensic science states that every contact leaves a trace. A criminal, in the course of committing a crime, may leave something behind, and he may take something away with him. From the examples given in Chan and Aitken (1989) for instance, suppose a crime is committed in which the criminal entered the house through a broken window and assaulted the residents, in the course of which assault blood was spilt. He left behind stains of his own blood and fibres from the jacket on the window and he may have taken away stains of the residents' blood on his clothes and fragments of glass from the broken window on his shoes. A suspect is later

apprehended whose blood group matches that left behind at the crime scene by the criminal. A jacket is found in his possession with fibres of a similar kind to that found on the crime window and with a blood stain whose blood group matches that of the victims, which is assumed to be different from that of the criminal. Fragments of glass are found in a pair of shoes belonging to the suspect with refractive index similar to that of the crime window. The problem of the assessment of the weight of such evidence, known as transfer evidence, is important in the administration of justice.

Good (1985) reviewed the problems associated with the weighing of evidence and proved that the only probabilities of interest in such circumstances are the probability of the evidence if the suspect is guilty and the probability of the evidence if the suspect is innocent. These probabilities are combined to form a measure of the weight of the evidence by constructing the ratio of the former probability to the latter probability in a likelihood ratio. This likelihood ratio may also be considered as a Bayes' factor, adjusting the prior odds in favour of guilt before the presentation of the evidence under consideration to provide posterior odds in favour of guilt after the presentation of the evidence. In certain situations, such as fibre transfer, the assumptions of the guilt or innocence of the suspect are too strong to make. It is more correct to replace these with assumptions that the suspect was present or not present at the scene of the crime.

The estimation of the Bayes' factor in the evaluation of evidence has been discussed in several papers (Lindley (1977), Evett

(1982), Evett (1984), Evett, Cage and Aitken (1987) (Evett et al hereafter), Chan and Aitken (1989)). Lindley (1977) suggested that evidence of contact should depend not only on the measurements but also on the distribution of the material of interest in the population which would form an additional objective source of information. Such information is relevant because when the control and recovered data are close enough to suggest a same source of material, there is greater evidence for the suspect not being present the scene of the crime when that same source is uncommon in the population than it is not. Seheult (1978) gave a hypothesis testing version to Lindley's arguement. Two years later Grove discussed the likelihood ratio approach to interpret forensic evidence. Evett (1982) showed that Bayesian inference can assist the forensic scientist to evaluate the evidence in the case where transfer has occurred from criminal to crime scene, illustrated by an example involving blood transfer. Evett (1984) distinguished between transfers from the criminal to the crime scene and from the crime scene to the criminal and derived general expressions for evaluating the evidential strength for either direction of transfer. In the former case, evidence found at the crime scene is assumed to have come from the criminal. In the latter, evidence found on the criminal may have come from some other source; for example in a case concerning broken glass, a suspect with glass in the soles of his shoes may be a glazier by trade and the glass may have been acquired perfectly innocently. Evett et al (1987) considered a particular bivariate problem relating to fibre evidence and used kernel density estimation to estimate the distribution of the recovered measurement assuming it comes from a random source of the population. Makov

(1987) considered a Bayesian method to assess the degree of evidence against each of the suspects, taking into account a 'missing suspect', that is an individual not included in the group of suspects who are under investigation.

1.3 Observational data

Apart from simulated data which were used to validate and demonstrate the model and methods proposed in this thesis, practical data were also used. For instance, the methods discussed in Chapter 3 and 5 are applied to data on cat and dog hairs, respectively, available from the Home Office Forensic Science Service Central Research and Support Establishment at Aldermaston. These data were previously discussed by Peabody *et al* (1983) and Aitken (1986) and are known as the cats and dogs data throughout this thesis. The data sets consist of 22 cats and 20 dogs (10 hairs from each cat/dog) bivariate hair measurements, namely the hair width and medullary fraction (i.e. medullary width/hair width). The following diagram shows a cross section of a filament of hair, which indicates the hair and medullary width.

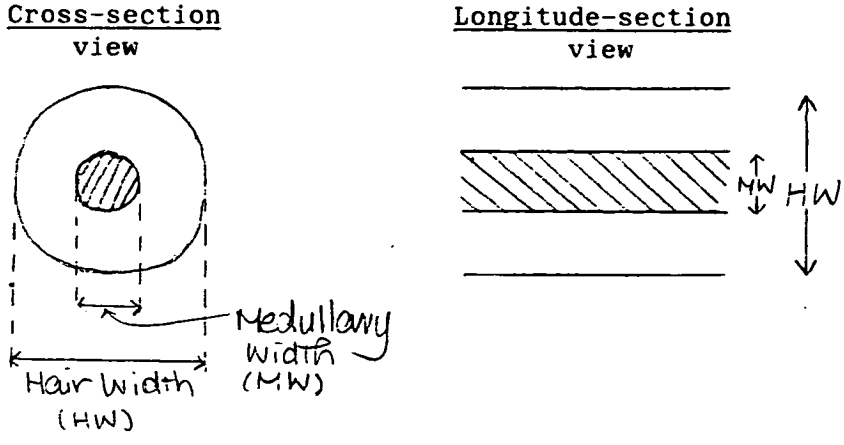


Diagram 1.1

1.4 Outline of the thesis

Kernel density estimation is a useful tool when the data are known to be not Normal. Some aspects of kernel density estimation are reviewed in Chapter 2. Also in chapter 2 Bayesian methods and introduction of the variance components problem are briefly summarised. In Chapter 3, the estimation of Bayes' factor in a forensic context is modelled under different assumption about the training data and parameters. Random effects model arises when we have training data consisting of between and within individual measurements.

In a random effects model, interest often lies in the estimation of the variance components and making relevant inference about them under the Normality assumption. However, this assumption is often not valid especially for the between groups variation. Kernel density is utilized to model the sample group means to investigate the effect of Non-normality about the between group variations. This was discussed in Chapter 4.

Chapter 5 is an adaptation of Chapter 3 to model a particular mixture data. It is motivated by a set of data which has a positive probability at a particular point, in this case it is at zero. Extension of Chapter 3 in a multivariate version, for the case where the within group variance is assumed known, is presented in Chapter 6.

Chapter 7 consists of a study on a new kernel density estimate, known as a Student-t. The Student-t kernel is a hybrid of the Gaussian kernel and the Bayesian method.

CHAPTER 2

DENSITY ESTIMATION, BAYESIAN METHODS AND RANDOM EFFECTS MODEL

2.1 Density estimation

Density estimation is a construction of an estimate of the density function from the observed data. A very natural use of density estimation is the informal investigation of the properties of a given set of data. Suppose that there is a set of observed data assumed to be a sample from an unknown *probability density function* (p.d.f.), f say. The probability density function can be estimated using either a *parametric* or a *non-parametric* approach. The former approach assumes the data are drawn from one of a known parametric family of distributions. An obvious example for this, is the Normal distribution with mean μ and variance σ^2. The density f underlying the observed data can be estimated by substituting estimates for μ and σ^2 from the observed data into the formula for the Normal density. The non-parametric approach, however, is more flexible and it lets the data speak for themselves.

Non-parametric density estimation can give valuable indications of features such as skewness and multimodality in the data. The oldest and most widely used density estimator is the *histogram*. The histogram has long been used as a means of displaying the distributional shape of a set of univariate data $D_n = \{t_1, t_2, \ldots, t_n\}$, assumed to be realisations of independent, identically distributed random variables. Usually, the histogram is required for pictorial

representation only, although it may be regarded as a formal estimate of the underlying density function (Tarter and Kronmal (1976)). Viewed as a density estimate, the histogram may be criticised in a number of ways; mainly that the underlying density is assumed to be smoothed, but the histogram is not and information has been thrown away in replacing $\{t_1,\ldots,t_n\}$ by $\{v_1,\ldots,v_{n^*}\}$, the mid-points of the bin containing the observations $\{t_i\}$ and the bin frequencies $\{fr_1,\ldots,fr_{n^*}\}$, where n^* is the number of bins.

2.1.1 The Kernel method

Rosenblatt (1956), Whittle (1958) and Parzen (1962) developed an approach to the problem which removed the difficulties the histogram method created. With the same notation as in the previous section, the estimator is of the form

$$f_n(t;D_n,h) = n^{-1} \sum_{i=1}^{n} K(t;t_i,h)$$

where K is itself a symmetric probability density, called a *kernel function* centred on t_i, whose variance is controlled by the parameter h. For example, it is often convenient to use for K a Normal density mean t_i and standard deviation h. Because of its role in determining how the probability associated with each observation is spread over the surrounding space, h is called the *smoothing parameter*. Since the properties of K are inherited by f_n, choosing K to be smooth will produce a density estimate which is also smooth. Furthermore each observation now has a kernel function centred directly over it and so the criticism associated with the histogram method discussed in Section 2.1 have been overcome. However, as the value of the smoothing parameter, h, can greatly affect the estimator (Bowman

(1985)) it must be chosen with care.

Other types of density estimator exist, such as those based on orthogonal series (see Fryer (1977)) and the penalised likelihood approach with Bayesian interpretation (Good and Gaskins (1971)). But no particular method can be regarded as superior over all others and the choice of a smoothing parameter analogous to the kernel smoothing parameter, h, is always required.

Here, the kernel approach is used because it is computationally straightforward, is conceptually simple being derived naturally from the histogram, and is closely related to other density estimation techniques. It is especially useful since it may be combined with other density functions in a obvious manner. However, the kernel method has a slight drawback when applied to data from long-tail distributions. It is because the smoothing parameter is fixed across the entire sample, so some spurious noise will appear in the tails of the estimates; if the estimates are smoothed effectively to deal with this, then we have problem of losing essential detail in the main part of the distribution.

A good general introduction to the subject of density estimation is given by Fryer (1977), with Wertz and Schneider (1979) providing an extensive additional list of references. The common structure of smoothing techniques is summed up by Titterington (1985).

In Chapters 3, 6 and 7 an adaptive kernel method, which is a modification of the ordinary kernel method, is also used and a brief

description is given below. The adaptive kernel provides better fit, especially, for smoothing long-tail distributions. Further discussion of the method is given by Silverman (1986).

2.1.2 The adaptive kernel method

The *adaptive* kernel method is an extension of the variable kernel method based on a nearest neighbours approach (see Silverman (1986)). The adaptive kernel method consists of a two-stage procedure in which an initial estimate is used to get a rough estimate of the density, then this estimate yields a pattern of smoothing factors corresponding to the various observations and these smoothing factors are used to construct the adaptive estimator itself.

Define the adaptive kernel estimate $\hat{f}(t)$ (assuming that the data points lie in p-dimensional space) by

$$\hat{f}(t) = n^{-1} \sum_{i=1}^{n} h^{-p} \lambda_i^{-p} K\{h^{-1} \lambda_i^{-1} (\underset{\sim}{t} - \underset{\sim}{t}_i)\}$$

where K is the kernel function, h is the smoothing parameter, λ_i is a smoothing factor given by

$$\lambda_i = \{\tilde{f}(\underset{\sim}{t}_i)/g\}^{-\delta},$$

$\tilde{f}(\underset{\sim}{t})$ is a pilot estimate of f(t) which satisfies $\tilde{f}(\underset{\sim}{t}_i) > 0$ for all i, g is the geometric mean of the $\tilde{f}(\underset{\sim}{t}_i)$ given by

$$\log_e g = n^{-1} \Sigma \log_e \tilde{f}(\underset{\sim}{t}_i),$$

and δ is a sensitivity parameter, a number satisfying $0 \leqslant \delta \leqslant 1$. As in the ordinary kernel method, K is a symmetric function integrating

to unity. Throughout the thesis $\delta = 1/2$ is used since Abramson (1982) gave an interesting and convincing argument that $\delta = 1/2$ is a reasonable choice for both the univariate and the multivariate case. However Breiman et al (1977) chose the reciprocal of p for δ.

The adaptive kernel method is used rather than a nearest neighbour method since the adaptive kernel is easy to compute. Choice of the sensitivity parameter δ is required and Abramson (1982) suggested that δ equal to one half is a reasonable choice. Silverman (1986) commented on the practical advantages of the adaptive over both the kernel and the nearest neighbour methods for smoothing long-tail distribution. He suggested that if undersmoothing in the tails is likely to cause difficulties, then the adaptive kernel approach is well worth considering. Abramson (1982) remarked, with reference to Breiman's findings, that the performance of the adaptive kernel method in a univariate study was considered disappointing, whereas in a bivariate study, excellent. Experience here is that the adaptive kernel behaved well.

2.1.3 Method of choosing smoothing parameter

For a density estimate to be fully defined, a value must be chosen for h. The value of h can be chosen subjectively. For example Scott, Tapia and Thompson (1977) recommended the subjective choice of a suitable smoothing parameter by decreasing it from values which gave estimates which were judged to be oversmoothed to values which seemed to give a density which was too rough and then marking the transition point. The "test graph" procedure of Silverman (1978a) requires the examination of graphs of the second derivative

of the density estimate. These techniques are difficult to evaluate because of their subjective nature.

There are many objective criteria in existence for the choice of the value of the smoothing parameter. They include asymptotic criteria (see Fryer (1977) and Silverman (1978b)), goodness-of-fit criteria based on empirical distributions (Good and Gaskins (1980)), methods based on Normality (Fryer (1976)), goodness-of-fit criteria in terms of Mean Integrated Square Error (MISE) (see Rosenblatt (1956), Parzen (1962), Woodroofe (1970), Nadaraja (1974) and Scott, Tapia and Thompson (1977)).

Two methods which will be used in this thesis are the method based on Normality (Fryer (1976)) and on cross validation due to Habbema et al (1974). The cross validation or, pseudo-maximum likelihood method is briefly described below.

A likelihood approach to the estimation of smoothing parameter problem was proposed by Habbema et al (1974) and by Duin (1976). If h is chosen to maximise

$$\prod_{i=1}^{n} \hat{f}_n(t_i;D_n,h)$$

then it is easily seen that the nuisance value of zero is obtained. Habbema et al and Duin therefore chose h to maximise

$$\prod_{i=1}^{n} \hat{f}_{n-1}(t_i;\ D_n\backslash\{t_i\},h) \qquad (2.1)$$

which leads to a reasonable degree of smoothing.

2.1.4 Measures of discrepancy

Some sort of measure or criterion is required to assess the performance of the kernel density estimation method. Various measures have been employed to study the discrepancy of the density estimator $\hat{f}$ from the true density f. When one considers estimation at a single point, a natural measure is the **mean square error** (MSE), defined by

$$MSE_t(\hat{f}) = E\{\hat{f}(t)-f(t)\}^2.$$

However, if one wants to have an overall picture of how the kernel performed then a measure over a wide range of t values would be more suitable. The following three measures provide such requirement, they are

(a) **Mean Integrated Square Error** (MISE) defined as

$$MISE(\hat{f}) = E\int\{\hat{f}(t)-f(t)\}^2\ dt,$$

(b) **Expected Mean Integrated Square Error** (EMSE)

$$EMSE(\hat{f}) = \int [E\{\hat{f}(t)-f(t)\}^2]f(t)\ dt, \text{ and}$$

(c) **Integrated Square Error** (ISE), defined by

$$ISE(\hat{f}) = \int\{\hat{f}(t)-f(t)\}^2\ dt.$$

Evaluation of the measure ISE is quite straight forward. Whereas, computation of MISE and EMSE is slightly more complicated. It involves, first of all, breaking down the term $E\{\hat{f}(t)-f(t)\}^2$ in (a) and (b) which is exactly the MSE above. MSE is the sum of the squared bias and the variance of $\hat{f}$ at the value t (see Silverman 1986

for details), namely,

$$\{E\hat{f}(t)-f(t)\}^2 + \text{Var } \hat{f}(t).$$

When these measures are evaluated all the integrations in (a), (b) and (c) are done numerically.

2.1.5 Examples of application of kernel density estimation

One of the most successful applications of density estimation techniques has been to discrimination problems. In the simplest situation, data arise from one of two classes, C1 and C2, each of which has an associated distribution defined by the density functions $f_1(t)$ and $f_2(t)$ respectively. Given data from each of these classes, the problem is to assign further observations, of unknown origin, to C1 or C2. Allocation to a particular class is usually based on the "$\log_e$ odd ratio"

$$\text{Log}_e \left\{ \frac{f_1(t)}{f_2(t)} \right\}.$$

The assumption of Normality leads to the familiar linear and quadratic discriminant functions, which were compared, by Remme, Habbema and Hermans (1980), with the use of nonparametric estimation of f_1 and f_2 from the data. The conclusion of the study was that the nonparametric approach is a very attractive one, performing well under a variety of situations, whereas the parametric procedures can give poor results when the underlying distribution is non-Normal.

Other areas to which kernel density estimation can be applied are: cluster analysis, bump hunting and testing for multimodality

(Cox (1966), Good and Gaskins (1980), Silverman (1981)) , simulation (Ripley (1983), Devroye and Györfi (1985)) and bootstrap applications (Efron (1981,1982)) etc.. Further details and references for these applications can be obtained from Silverman (1986)

A recent paper by Huang (1987) developed a two-sample nonparametric likelihood ratio test which involved the use of the kernel density method to estimate the likelihood ratio in order to test, with or without the assumption of common variance, whether two samples had a common mean.

Evett et al (1987) applied kernel density estimation to evaluate a bivariate probability density function of a vector arising randomly from a population in a forensic problem.

2.2 Some aspects of the Bayesian approach to statistical modelling

The early work in Bayesian statistics of this century was done by de Finetti (1930), Jeffreys (1939, reprinted with corrections, 1983) and Ramsey (1931/1964). Jeffreys gave the foundation for Bayesian inference, which was continued by Lindley (1965), Box and Tiao (1973) and Press (1982). Most of these writers employed improper type prior distributions for the parameters, but Raiffa and Schlaifer (1961) introduce conjugate prior distributions, which are proper probability distributions.

In this section, some basic principles and concepts of Bayesian analysis are summarised. The Bayesian approach to statistical modelling and other problems is relatively simple but important.

2.2.1 The Bayes' Theorem

An essential element of the Bayesian approach is Bayes' theorem. Here the theorem for continuous random variables is stated. Let $f(\underset{\sim}{t},\underset{\sim}{\theta})$ denote the joint probability density function (p.d.f.) for a random observation vector $\underset{\sim}{t}$ and a parameter vector $\underset{\sim}{\theta}$, also considered random. The parameter vector $\underset{\sim}{\theta}$ may consist of some elements which are unknown and which it is derived to estimate in order to specify a model. Then, according to usual operations with the p.d.f. and with obvious notation, assuming both $\underset{\sim}{\theta}$ and $\underset{\sim}{t}$ have underlying probability distributions, we have

$$\begin{aligned} f(\underset{\sim}{t},\underset{\sim}{\theta}) &= f(\underset{\sim}{t}|\underset{\sim}{\theta})\ f(\underset{\sim}{\theta}) \\ &= f(\underset{\sim}{\theta}|\underset{\sim}{t})\ f(\underset{\sim}{t}) \end{aligned}$$

and thus

$$f(\underset{\sim}{\theta}|\underset{\sim}{t}) = \frac{f(\underset{\sim}{t}|\underset{\sim}{\theta})\ f(\underset{\sim}{\theta})}{f(\underset{\sim}{t})}$$

with $f(\underset{\sim}{t}) \neq 0$. The above expression can be written as follows:

$$\begin{aligned} f(\underset{\sim}{\theta}|\underset{\sim}{t}) &\propto f(\underset{\sim}{\theta})\ f(\underset{\sim}{t}|\underset{\sim}{\theta}) \\ &\propto \text{prior p.d.f} \times \text{likelihood function} \end{aligned} \qquad (2.2)$$

where $\propto$ denotes proportionality, $f(\underset{\sim}{\theta}|\underset{\sim}{t})$ is the *posterior p.d.f.* for the parameter vector $\underset{\sim}{\theta}$, given the sample information $\underset{\sim}{t}$, $f(\underset{\sim}{\theta})$ is the *prior p.d.f.* for the parameter vector θ, and $f(\underset{\sim}{t}|\underset{\sim}{\theta})$, viewed as a function of $\underset{\sim}{\theta}$, is the well-known *likelihood function*. Equation (2.2) is a statement of Bayes' theorem, a simple mathematical result in the theory of probability. Note that the joint posterior p.d.f. of the unknown parameter $\underset{\sim}{\theta}$, $f(\underset{\sim}{\theta}|\underset{\sim}{t})$, has all the prior and sample information

incorporated in it. The prior information enters the posterior p.d.f. via the prior p.d.f., whereas all the sample information enters via the likelihood function. In this latter connection the 'likelihood principle' states that $f(\underset{\sim}{t}|\underset{\sim}{\theta})$, considered as a function of $\underset{\sim}{\theta}$, quoted from Savage (1962)

> "... constitutes the entire evidence of the experiment, that is, it tells all that the experiment has to tell ".

In the usual Bayesian analysis, the posterior p.d.f. is employed to make inferences about parameters.

Example 2.1 Assume that we have n independent observations, $\underset{\sim}{t}' = \{t_1, t_2, \ldots, t_n\}$, drawn from a Normal population with unknown mean μ and known variance $\sigma^2 = \sigma_0^2$. We wish to obtain the posterior p.d.f. for μ. Applying (2.2) to this particular problem, we have

$$f(\mu|\underset{\sim}{t},\sigma_0^2) \propto f(\mu|\sigma_0^2)\ f(\underset{\sim}{t}|\mu,\sigma_0^2)$$

where $f(\underset{\sim}{\mu}|\underset{\sim}{t},\sigma_0^2)$ is the posterior p.d.f for the parameter μ, given the sample information $\underset{\sim}{t}$ and the assumed known value σ_0^2, $f(\mu)$ is the prior p.d.f for μ, and $f(\underset{\sim}{t}|\mu,\sigma_0^2)$, viewed as a function of the unknown parameter μ is the likelihood function. The likelihood function is given by $\prod_{i=1}^{n} f(t_i|\mu,\sigma_0^2)$, or

$$\begin{aligned} f(\underset{\sim}{t}|\mu,\sigma_0^2) &= (2\pi\sigma_0^2)^{-n/2} \exp\{-[\sum_{i=1}^{n}(t_i-\mu)^2]/2\sigma_0^2\} \\ &= (2\pi\sigma_0^2)^{-n/2} \exp\{-[\nu s^2 + n(\mu-\hat{\mu})^2]/2\sigma_0^2\} \end{aligned} \qquad (2.3)$$

where $\nu = n-1$, $\hat{\mu} = (1/n)\sum_{i=1}^{n} t_i$, the sample mean, and $s^2 = (1/\nu)\sum_{i=1}^{n}(t_i-\bar{t})^2$, the sample variance.

As regards a prior p.d.f. for μ, we assume that our prior information regarding this parameter can be represented by the following univariate Normal p.d.f., independent of σ_0^2

$$f(\mu|\sigma_0^2) = (2\pi\sigma_a^2)^{-1/2} \exp\{-(\mu-\mu_a)^2/2\sigma_a^2\} \tag{2.4}$$

where μ_a is the prior mean and σ_a^2 is the prior variance, parameters whose values are assigned by the investigator on the basis of his initial information. Then, on using Bayes' theorem to combine the likelihood function in (2.3) and dropping the notational dependency of the prior distribution of μ on σ_0^2, we obtain the following posterior p.d.f

$$\begin{aligned} f(\mu|\underset{\sim}{t},\sigma_0^2) &\propto f(\mu)\, f(\underset{\sim}{t}|\mu,\sigma_0^2) \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{(\mu-\mu_a)^2}{\sigma_a^2} + \frac{n(\mu-\hat{\mu})^2}{\sigma_0^2}\right]\right\} \\ &\propto \exp\left\{-\left[\frac{\sigma_a^2+\sigma_0^2/n}{2\sigma_a^2\sigma_0^2/n}\right]\left[\mu - \frac{\hat{\mu}\sigma_a^2 + \mu_a\sigma_0^2/n}{\sigma_a^2 + \sigma_0^2/n}\right]\right\} \end{aligned}$$

from which it is seen that μ is Normally distributed, a *posteriori*, with mean

$$E(\mu) = \frac{\hat{\mu}\sigma_a^2 + \mu_a\sigma_0^2/n}{\sigma_a^2 + \sigma_0^2/n} = \frac{\hat{\mu}(\sigma_0^2/n)^{-1} + \mu_a(\sigma_a^2)^{-1}}{(\sigma_0^2/n)^{-1} + (\sigma_a^2)^{-1}} \tag{2.5}$$

and variance given by

$$\mathrm{Var}(\mu) = \frac{\sigma_a^2\sigma_0^2/n}{\sigma_a^2 + \sigma_0^2/n} = \frac{1}{(\sigma_0^2/n)^{-1} + (\sigma_a^2)^{-1}}\,. \tag{2.6}$$

Note that the posterior mean in (2.5) is a weighted average of the sample mean $\hat{\mu}$ and the prior mean $\hat{\mu}_a$, with the weights being the

reciprocals of σ_o^2/n and σ_a^2. If we let $\tau_o = (\sigma_o^2/n)^{-1}$ and $\tau_a = (\sigma_a^2)^{-1}$, then $E(\mu) = (\hat{\mu}\tau_o + \mu_a\tau_a)/(\tau_o + \tau_a)$, where the τ's are often referred to as "precision" parameters. Also we have $\text{Var}(\mu) = 1/(\tau_o + \tau_a)$ from (2.6), and thus the precision parameter associated with the posterior mean is just $[\text{Var}(\mu)]^{-1} = \tau_o + \tau_a$, the sum of the sample and prior precision parameters.

2.2.2 Prior probability density functions

The prior p.d.f., denoted $f(\underset{\sim}{\theta})$ in (2.2), represents our prior information about the parameters of a model; that is, in the Bayesian approach, the prior information about parameters of a model is usually represented by an appropriately chosen p.d.f. In the example 2.1, for instance, prior information about a mean μ is represented in (2.4) by a Normal p.d.f. with prior mean μ_a and variance σ_a^2. The prior mean and variance μ_a and σ_a^2 are assigned values by the investigator in accord with his prior information about the parameter μ. If this Normal prior p.d.f. is judged an adequate representation of the available prior information, it can be used , as demonstrated, to obtain the posterior p.d.f. for μ. On the other hand, if the prior information is not adequately represented by a Normal prior p.d.f, another prior p.d.f. that does so may be used. For example, if from the past data it is known that μ is not Normally distributed, or if we have a scalar parameter θ, say, a proportion, that by its very nature is limited to an interval [0,1], it would not be appropriate to employ a Normal p.d.f. for θ, since a Normal p.d.f. does not limit the range of θ to the interval [0,1]. The p.d.f. for θ should be one, say a beta p.d.f., that can incorporate the available information on the range of θ.

As regards the nature of prior information, it is to be recognised that it may include information contained in samples of past data or samples randomly gathered to represent the distribution of the parameter. When a prior p.d.f. represents information of this kind, the prior p.d.f. is called an '*informative*' or '*reference*' prior. Otherwise, it is called '*diffuse*' or '*vague*'. A situation involving vague or diffuse priors may be one in which an investigator has little idea about the parameters under study.

Example 2.2. Consider the Example 2.1 again and assume that our prior information regarding the value of μ is vague or diffuse. To represent such information about the value of μ, we follow Jeffreys (1939/83) by taking

$$f(\mu) \propto \text{constant} \qquad -\infty < \mu < \infty$$

as our prior p.d.f.. This prior p.d.f. is improper, i.e., $\int f(\mu)\, d\mu$ is not finite. Then the posterior p.d.f. for μ, $f(\mu|\underset{\sim}{t},\sigma=\sigma_0)$ is given by

$$\begin{aligned} f(\mu|\underset{\sim}{t},\sigma=\sigma_0) &\propto f(\mu)\ \ell(\mu|\underset{\sim}{t},\sigma=\sigma_0) \\ &\propto \exp\{-[(\mu-\hat{\mu})^2/2\sigma_0^2]\}, \end{aligned}$$

where $\ell(\mu|\underset{\sim}{t},\sigma=\sigma_0) \propto f(\underset{\sim}{t}|\mu,\sigma=\sigma_0)$, is the likelihood function of μ conditional on $\underset{\sim}{t}$ and $\hat{\mu} = \sum_{i=1}^{n} t_i/n$, the sample mean. It is seen that posterior p.d.f. is Normal with mean $\hat{\mu}$ and variance σ_0^2/n. The same result would be obtained in Example 2.1 if there we spread out the Normal prior p.d.f. for μ (i.e. allowed $\sigma_a \to \infty$).

2.2.3 Marginal distribution of the observations

In certain situations it is of interest to obtain the marginal p.d.f. for the observations, denoted by $f(\underset{\sim}{t})$. The p.d.f. can be obtained as follows:

$$\begin{aligned} f(\underset{\sim}{t}) &= \int_{\Omega} f(\underset{\sim}{\theta},\underset{\sim}{t})\, d\underset{\sim}{\theta} \\ &= \int_{\Omega} f(\underset{\sim}{t}|\underset{\sim}{\theta})\, f(\underset{\sim}{\theta})\, d\underset{\sim}{\theta}, \end{aligned} \tag{2.7}$$

where Ω is the parameter space of $\underset{\sim}{\theta}$. Equation (2.7) indicates that the marginal p.d.f. of the observations is an average of the conditional p.d.f. $f(\underset{\sim}{t}|\underset{\sim}{\theta})$ with prior p.d.f. $f(\underset{\sim}{\theta})$ serving as the weight function.

Example 2.3 Let t_1 be an observation from a Normal distribution with unknown mean μ and known standard deviation $\sigma=\sigma_0$. Then

$$f(t_1|\mu,\sigma=\sigma_0) = (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp\{-(t_1-\mu)^2/2\sigma_0^2\}.$$

If the prior p.d.f. for μ is as in (2.4), then the marginal p.d.f. for t_1 is

$$\begin{aligned} f(t_1) &= \int_{-\infty}^{\infty} f(t_1|\mu,\sigma=\sigma_0)\, f(\mu)\, d\mu \\ &= (2\pi\sigma_0\sigma_a)^{-1} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\frac{(t_1-\mu)^2}{\sigma_0^2} + \frac{(\mu-\mu_a)^2}{\sigma_a^2}\right]\right\} d\mu \end{aligned}$$

On the completing the square for μ in the exponent and performing the integration, the result is

$$f(t_1) = \frac{1}{[2\pi(\sigma_a^2+\sigma_0^2)]^{\frac{1}{2}}} \exp\left\{-\frac{(t_1-\mu_a)^2}{2(\sigma_a^2+\sigma_0^2)}\right\}.$$

Thus the marginal p.d.f.for t_1 is normal with mean μ_a, the prior mean for μ, and variance $\sigma_a^2 + \sigma_o^2$. Since μ_a, σ_a^2 and σ_o^2 are assumed known, it is possible to use $f(t_1)$ to make probability statements about t_1, a fact that is often useful before t_1 is observed.

Example 2.3 is an example of the well-known result that a convolution of two Normal distribution gives a Normal distribution. Two basic results are proved in Appendix 2, which will be used in the later chapters.

2.2.4 Predictive probability density functions

On many occasions, given sample information $\underset{\sim}{t}$, we are interested in making a probability statement about a 'new' observation t_1, given $\underset{\sim}{t}$. In the Bayesian approach the p.d.f. for the 'new' observation t_1, given sample information, can be obtained and is known as the predictive p.d.f.; for example, let $\tilde{\underset{\sim}{t}}$ represent a vector of the 'new' observations. We write

$$f(\tilde{\underset{\sim}{t}},\underset{\sim}{\theta}|\underset{\sim}{t}) = f(\tilde{\underset{\sim}{t}}|\underset{\sim}{\theta},\underset{\sim}{t})\ f(\underset{\sim}{\theta}|\underset{\sim}{t}) \qquad (2.8)$$

as the joint p.d.f. for $\tilde{\underset{\sim}{t}}$ and a parameter vector $\underset{\sim}{\theta}$, given the sample information $\underset{\sim}{t}$. On the right of (2.8) $f(\tilde{\underset{\sim}{t}}|\underset{\sim}{\theta},\underset{\sim}{t})$ is the conditional p.d.f. for $\tilde{\underset{\sim}{t}}$, given $\underset{\sim}{\theta}$ and $\underset{\sim}{t}$, whereas $f(\underset{\sim}{\theta}|\underset{\sim}{t})$ is the conditional p.d.f. for $\underset{\sim}{\theta}$ given $\underset{\sim}{t}$, that is, the posterior p.d.f. for θ. Note that $f(\tilde{\underset{\sim}{t}}|\theta,\underset{\sim}{t}) = f(\tilde{\underset{\sim}{t}}|\theta)$ since the 'new' and the past observations are assumed to be independent. To obtain the predictive p.d.f., $f(\tilde{\underset{\sim}{t}}|\underset{\sim}{t})$, we merely integrate (2.8) with respect to $\underset{\sim}{\theta}$ that is

$$f(\tilde{\underset{\sim}{t}}|\underset{\sim}{t}) = \int_\Omega f(\tilde{\underset{\sim}{t}},\underset{\sim}{\theta}|\underset{\sim}{t})\ d\underset{\sim}{\theta}$$

$$\int_{\Omega} f(\tilde{\underset{\sim}{t}}|\underset{\sim}{\theta})\, f(\underset{\sim}{\theta}|\underset{\sim}{t})\, d\underset{\sim}{\theta}. \qquad (2.9)$$

Again, equation (2.9) indicates that the predictive p.d.f. can be viewed as an average of conditional predictive p.d.f.'s, $f(\tilde{\underset{\sim}{t}}|\underset{\sim}{\theta})$, with the posterior p.d.f. for $\underset{\sim}{\theta}$, $f(\underset{\sim}{\theta}|\underset{\sim}{t})$ serving as the weighting function. Examples for a wide range distributions can be found in Aitchison and Dunsmore (1975).

2.2.5 Empirical Bayes method

The *empirical Bayes'* method is employed in most chapters of this thesis to estimate a density function of an unknown parameter. Prominent amongst workers in this area are Robbins (1955), who pioneered the idea and adopted the terminology 'empirical Bayes approach', and Maritz (1970). Maritz (1970) described the approach as a 'hybrid' one.

As an illustration of a Bayes' procedure suppose that data t, arise as an observation of a random variable, T. The distribution of T, specified by the probability model, is assumed to belong to some family $\mathfrak{p}$, indexed by a parameter θ. It is assumed that the probability density function of the random variable T has a known form, $f(t|\theta)$, depending on θ; but that θ is unknown, except that it lies in a parameter space, Ω. For the purpose of the point estimation of the parameter θ, one could use the mode or the *mean* of the posterior distribution of θ, given the sample, t. Assume that the mean of the posterior distribution of θ is the estimator of θ and $\pi(\theta)$ denotes the prior for θ then, from (2.2) we could estimate θ by

$$\tilde{\theta}_{\pi}(t) = \frac{\int_{\Omega} \theta\, f(t|\theta)\, \pi(\theta)\, d\theta}{\int_{\Omega} f(t|\theta)\, \pi(\theta)\, d\theta} . \qquad (2.10)$$

So if the prior distribution were known, we would have in (2.10) a reasonable estimator of θ. In a typical empirical Bayes situation, it is assumed that in addition to the current observation t when the parameter value is θ, a set of 'past' observations $t_1,\ldots,t_n$ obtained when the parameter values were $\theta_1,\ldots,\theta_n$, say (these θ values being known) is given. It is assumed that θ_i (i=1,...,n) arise as a random sample from the prior distribution, $\pi(\theta)$, and that the t_i (i=1,...n) are independent sample observations arising under the values of θ. The previous observations 'reflect' the prior distribution, $\pi(\theta)$ and, in the general empirical Bayes' approach, are used to estimate $\pi(\theta)$ for use in Bayesian inference. In some cases direct estimation of $\pi(\theta)$ is unnecessary and may be by-passed (see Maritz (1970)).

If the estimation of the prior p.d.f is an objective of the problem and suppose $\pi(\theta)$ depends on hyperparameters $\alpha \in H$, then from (2.9) the predictive density of t is given by

$$f(t|\alpha) = \int_{\Omega} f(t|\theta)\, \pi(\theta|\alpha)\, d\theta, \qquad t \in S, \quad \alpha \in H$$

and one may use this integral equation to find values of α which support the fit of the predicted observation t (sampled from the population with density $f(t|\alpha)$) or to fit past observations which were sampled from a population with density $f(t|\alpha)$. If one observes values $t_1, t_2, \ldots, t_n$ from this distribution, one may find values of α compatible with these observations by the method of moments or maximum likelihood or some other principle of estimation (see Maritz (1970) for examples). Or one might put a known prior density on α

and estimate α from the conditional distribution of α given the predicted observations.

One interest to many observers will be the extent to which Bayesians are able in practice to depart from the standard Bayesian model with a subjective guess of π, and can instead imbed the problem at hand as the $(n+1)^{th}$ one in the empirical Bayes model, to yield and use a more formally described guess $\hat{\pi}_n$ based on the past n observations. To non-Bayesians, Robbins' model will seem much more acceptable in many practical settings than the original Bayesian formulation. For example, t_i might be an observation of some biological charateristic of a worker i in a hospital or a large plant, and θ_i an index of the underlying condition having an unknown distribution charateristic of this population of workers.

2.2.6 Bayesian approach to hypothesis testing

Suppose that there are two hypotheses H_i and H_j, with prior probabilities $P(H_i)$ and $P(H_j)$. Let $\underset{\sim}{\theta}_i$ denote the parameter vector associated with hypothesis H_i under which the p.d.f. for the observation vector $\underset{\sim}{t}$ is $f(\underset{\sim}{t}|\underset{\sim}{\theta}_i)$ and $\underset{\sim}{\theta}_j$, the parameter vector associated with hypothesis H_j under which the p.d.f. for $\underset{\sim}{t}$ is $f(\underset{\sim}{t}|\underset{\sim}{\theta}_j)$. Then the posterior probability associated with H_i is given by

$$P(H_i|\underset{\sim}{t}) = \frac{P(H_i)\ f(\underset{\sim}{t}|H_i)}{f(\underset{\sim}{t})} \qquad (2.10)$$

and similarly

$$P(H_j|\underset{\sim}{t}) = \frac{P(H_j)\ f(\underset{\sim}{t}|H_j)}{f(\underset{\sim}{t})}. \qquad (2.11)$$

By Bayes' rule, the relative posterior probabilities of two hypotheses can be written as

$$\frac{P(H_i|\underset{\sim}{t})}{P(H_j|\underset{\sim}{t})} = \left[\frac{f(\underset{\sim}{t}|H_i)}{f(\underset{\sim}{t}|H_j)}\right] \times \left[\frac{P(H_i)}{P(H_j)}\right].$$

The second factor in brackets is the prior odds ratio in favour of H_i. The data-dependent term in the first set of brackets is the "Bayes' factor".

The data are said to favour H_i relative to H_j if the Bayes' factor exceeds one, that is, if the observed data $\underset{\sim}{t}$ are more likely under hypothesis H_i than under hypothesis H_j. The densities of $\underset{\sim}{t}$, implied by the hypotheses in (2.10) and (2.11) are conditional on the parameters, u_i and σ_i^2 say, but may be straightforwardly "mixed" into a marginal density as

$$f(\underset{\sim}{t}|H_i) = \int_{u_i}\int_{\sigma_i^2} f(\underset{\sim}{t}|H_i,u_i,\sigma_i^2)\ f(u_i,\sigma_i^2)\ d\sigma_i^2\ du_i \qquad (2.12)$$

where $f(u_i,\sigma_i^2)$ is the prior density. The conditional p.d.f. $f(\underset{\sim}{t}|H_i,u_i,\sigma_i^2)$ is a likelihood function of (u_i,σ_i^2), and (2.12) defines $f(\underset{\sim}{t}|H_i)$ as a weighted or marginal likelihood.

The Bayes' factor may be contrasted with the likelihood ratio, which is used classically to summarize the data evidence. The likelihood ratio is

$$L(H_i,H_j) = \frac{\max_{u_i,\sigma_i} f(\underset{\sim}{t}|u_i,\sigma_i^2,H_i)}{\max_{u_j,\sigma_j} f(\underset{\sim}{t}|u_j,\sigma_j^2,H_j)} .$$

The Bayes' factor considers the ratio of the averages of the likelihood function over all value of (u_i,σ_i^2) and (u_j,σ_j^2). The likelihood ratio involves taking the ratio of maximised likelihood functions under H_i and H_j, a procedure that amounts to using maximum likelihood estimates as if they are true values of the unknown parameters. The Bayesian approach, however, presupposes prior distributions that can be used to weight the evidence at different values of the parameters.

The Bayes' factor has been employed to make comparison of alternative models denoted by different hypotheses. Smith and Spiegelhalter (1980) used a Bayesian approach to comparing alternative nested linear models and provided a unified development of a number of model choice criteria by examining some prior specifications. A measure of the weight of evidence provided by the data for H_i against H_j Jeffrey (1939/83, appendix B), suggested the following grouping of the Bayes' factor into 'order of magnitude' based on the logarithmic scale

$BF > 1$	Evidence support H_i.
$1 > BF > 10^{-½}$	Evidence against H_i, but not worth more than a bare mention.
$10^{-½} > BF > 10^{-1}$	Evidence against H_i substantial.
$10^{-1} > BF > 10^{-3/2}$	Evidence against H_i strong.
$10^{-3/2} > BF > 10^{-2}$	Evidence against H_i very strong.
$10^{-2} > BF$	Evidence against H_i decisive.

These groupings are to be used in Chapter 6 to interpret the behaviour of the Bayes' factor.

2.2.7 Hypothesis testing: A Judicial analogy

This section establishes some judicial concepts and notations in conjunction with Chapter 3, though the words 'guilty' and 'innocent' are in fact too strong for our model in Chapter 3, as explained earlier in Chapter 1.

The subject of hypothesis testing in this context may be usefully introduced by an analogy. Based on the evidence presented, a judge and/or jury in a legal proceeding decide whether a defendant should be innocent or guilty. The assumption of innocence until proven guilty beyond a reasonable doubt explicitly favours the hypothesis of innocence. The hypothesis of innocence is taken as the null hypothesis; the hypothesis of guilt is taken as the alternative hypothesis.

The more critical error — finding an innocent man guilty — is called an *error of the first kind* or a *type I error*. Acceptance of the null hypothesis when it is in fact false — finding a guilty man

innocent — is called *an error of the second kind* or a *type II error.* Schematically we have

		Actions	
Hypotheses (States)		Find innocent (accept H_0)	Find guilty (reject H_0)
H_0:	Innocent		Type I error
H_1:	Guilty	Type II error	

If a man is innocent, we want to have a low probability of finding him guilty. Let this probability be

$$\alpha = \Pr(\text{guilty}|\text{innocent}).$$

Analogously, let

$$\beta = \Pr(\text{innocent}|\text{guilty}).$$

If both α and β are defined before the judicial process commences, then the quality of the evidence may be predicted effectively. For example, a zero value of α amounts to the prediction that if the defendant is innocent, the evidence will be so unambiguous and the process by which a verdict is rendered will be so perfect that with probability one he will be justly found innocent.

2.3 Random effects model

The random effects models for the analysis of variance are also called variance-components models. The origin of the random effects models lie in astronomical problems, statisticians re-invented random effects models long after they were introduced by astronomers and

then developed more complicated ones.

2.3.1 One-way classification model

It is easiest to introduce the random effects model by an example. Young et al (1965) designed an experiment to study the maternal ability of mice. Weights of ten-day-old litters as a measure of maternal ability were used. Six litters from each of four dams, all of one breed, constitute the data. A suitable model for analysing the data is the one-way classification model

$$t_{ij} = A_i + e_{ij} \qquad (i=1,\ldots,n;j=1,\ldots,J) \qquad (2.13)$$

where t_{ij} is the weight of the j^{th} litter from the i^{th} dam, A_i being the 'true' mean weight for the i^{th} dam and e_{ij} the usual error term. These two random factors are assumed independent.

Consider the A_i's and the dam they represent. The data relate to maternal ability, a variable that is assuredly subject to biological variation from animal to animal. The aim of the experiment is therefore unlikely to centre on specifically the 4 female mice used in the experiment. After all, they are only a sample form a large population of mice, the females of the breed, each of which has some ability in a maternal capacity. The animals that are in the experiment are therefore envisaged as a random sample of 4 from a population of females.

In the usual one-way random effects model, interest lies not in the difference between any one of the 4 mice and any other of them, in the experiment described above, but interest does lie in the

extent to which maternal ability varies throughout the population of mice, and to this end the model (2.13) is directed.

The sampling process involved in obtaining such data is taken as being such that any one of many possible sets of data could be derived from repetitions of the data-gathering process. But now, in concentrating attention on repetitions, we do not confine ourselves to always having the same 4 mice - we imagine getting a random sample of 4 on each occasion from the population of mice. Thus the A_i's, of the mice data described, are a random sample from a population of A's. Hence, so far as the data are concerned, the A_i's therein are random variables and the model associated with this type of data is called random effects model or, sometimes, the random model. Eisenhart (1947) called it Model II, a name that continues to receive widespread use.

Let the A_i's and e_{ij}'s of model (2.13) have variances σ_a^2 and σ_e^2 respectively. Then the variance of an observation is from (2.13), assuming independence between A_i and e_{ij}, $\sigma_t^2 = \sigma_a^2 + \sigma_e^2$. The variances, σ_a^2 and σ_e^2, are accordingly called variance components; each is a variance in its own right and is a component of σ_t^2. The model is sometimes referred to as a variance components model. Estimation of the variance components and inferences about them are the objectives of using such a model.

2.3.2 Analysis of variance

The earliest methodology for the estimation of the variance components was to equate the analysis of variance sums of squares to

their expectations and solve the resulting system of linear equations for estimates of the variance components. The methodology was developed by Daniels (1939) and Winsor and Clarke (1940), and the sampling properties of the estimators studied by Graybill (1954), Graybill and Wortham (1956), and Graybill and Hultquist (1961). For example, in the one-way random effects model, the between and within mean squares are equated to their expectations, giving analysis of variance estimates of the between and within components (See Searle (1971)). It is customary to summarise the results in a Analysis of Variance (ANOVA) table. The form for the one-way classification is given in Table 2.1.

Table 2.1 Analysis of Variance for the one way classification

Source of variation	Sum of Square	d.f.	Mean Square	Expected MS	*F*-ratio
Between	BSS	n-1	BSS/(n-1)=BMS	$\sigma_e^2 + J\sigma_a^2$	BMS/WMS
Within	WSS	n(J-1)	WSS/n(J-1)=WMS	σ_e^2	
Total	TSS	N-1			

A Bayesian approach to the estimation of variance components problem was taken by Hill (1965,1967), who studied the one-way model and Tiao & Tan (1965, 1966). Stone and Springer (1965) criticized Tiao and Tan's choice of prior distribution. Box and Tiao (1973) is the first book on Bayesian analysis to deal with the variance components of random effects model. They give a very thorough treatment of the subject and the methodology is based on numerical

determination of the one- and two- dimensional marginal posterior distribution of the variance components. Broemeling (1985) gives a more general and basic theory of linear models from a Bayesian viewpoint.

Another development of the analysis of variance problem is a maximum likelihood estimation technique given by Hartley and Rao in 1967 and since then there have appeared many new methodologies including restricted maximum likelihood, minimum norm quadratic unbiased estimation or MINQUE, iterative MINIQUE, MIVGUE, or minimum variance quadratic unbiased estimation. Searle (1977) gave a summary of the recently developed methods.

One difficulty which has concerned many of these writers is the so-called 'negative estimated variance' problem. For instance, under the one-way random effects model, with the assumption that the A_i's and e_{ij}'s are independent among themselves, the following unbiased estimator $\hat{\sigma}_a^2$ for σ_a^2, the between group variance

$$\hat{\sigma}_a^2 = (\text{BMS}-\text{WMS})/J$$

where BMS and WMS are the Between and Within Mean Squares (see Table 2.1), respectively, may, with positive probability, take a negative value. This problem does not occur if the Bayesian approach is employed. A second difficulty within the traditional framework is the sensitivity of inferences to the departures from the underlying assumptions. Scheffé (1959) showed that non-Normality in the A_i's will have serious effects on the distributions of the criteria which one uses to make inference about the parameters in the one way model.

Tiao and Ali (1971) investigated the effect of non-Normality on inference about the variance components by assuming the distribution of A_i was in a form of a mixture of two Normals. Their investigation concluded that inferences regarding the between group variance σ_a^2 are very sensitive to the Normality assumption. However, inference concerning the within group variance σ_e^2 is not so sensitive to failure of the distributional assumptions.

CHAPTER 3

ESTIMATION OF BAYES' FACTOR IN A FORENSIC CONTEXT

3.1 Introduction

In Chapter 2 we have already seen the general Bayesian treatment of hypothesis testing and statistical modelling. This forms a foundation for modelling the forensic problem in this chapter. The Bayes' factor, or likelihood ratio, plays an important role in the assessment of forensic evidence. The general background of a forensic problem was briefly reviewed in Chapter 1. Here we consider a particular problem. The method developed here is applied to the cat hairs data, see Section 1.3. The most likely scenario is that a criminal would pick up cat hairs at the crime scene and the transfer of evidence would be from the crime scene to the criminal. A full consideration of the strength of the evidence would require knowledge of the probability that any suspect present at the crime scene may have picked up cat hairs from some innocent source. This possibility is not considered here and the assumption is made that hairs found on a suspect and assessed under the assumption of presence at the crime scene could only have come from the crime scene. This is done so that progress may be made in the evaluation of the evidence in the situation of relaxed assumptions from Lindley (1977) and Evett et al (1987).

Background data collected by forensic scientists often have a random effects structure where the random effects do not have a Normal distribution. The methods of assessing these data compare

results obtained where a group structure in the background data is and is not assumed, and where the within group variance is and is not assumed known. The distribution of the random effects is modelled using kernel density estimation. Much of this chapter is published in Chan and Aitken (1989).

3.2 Notation

A crime is committed. A suspect is apprehended and transfer evidence is found which associates him with the crime scene. For example, cat hairs on his clothes may have come from a cat resident at the crime scene. The evidence is assessed by considering the probability of the evidence if the suspect was present at the crime scene and if he was not.

In a view of the problem set out above, the following notation is used:

Let C be the hypothesis that there is a contact between the suspect and the crime scene and let $\bar{C}$ be the hypothesis that there is no contact. The transfer evidence will be denoted by E and consists of two sets of data, X and Y, so that E=(X,Y). Data $\underset{\sim}{X} = (x_1,\ldots,x_m)'$ are control data consisting of m measurements whose source is known, for example, measurements on representative hairs from the cat at the crime scene. Data $\underset{\sim}{Y} = (y_1,\ldots,y_r)'$ are recovered data consisting of r measurements and consist of measurements of material similar to that which provides the control data, taken from what is known as a receptor body. In the example described earlier they would be measurements of cat hairs taken from the suspect's clothing. If the

suspect was present at the scene of the crime then X and Y could have the same source. If the suspect was not present at the scene then X and Y have different sources. To assist in the evaluation of evidence there is a set, Z, of measurements, known as the training data, which is taken to be a representative sample of the whole population of measurements of the material of interest. The training data collected often has a random effects structure, that is, there are variations between and within individuals or items. Then it is neccessary to take this into account in modelling the Bayes' factor.

The definition of 'grouped' or 'grouping' is that the data are generated from a random effects model. If the training data Z are said to be grouped, it means that Z are available in grouped form, namely, $\{Z_{ij};\ i=1,2,\ldots,n,\ j=1,2,\ldots,J\}$ where n is the number of groups, J is the number of observations (assumed constant) in each group and Z_{ij} is the j^{th} observation in the i^{th} group. For example, the training data could consist of measurements from each of J hairs taken from each of n cats. Particular measurements $\{z_{ij}\}$ are assumed to be generated by a random effects model

$$z_{ij} = \mu_i + \epsilon_{ij}, \qquad i=1,2,\ldots,n;\ j=1,2\ldots J \tag{3.1}$$

where μ_i is a realisation of a random variable U_i in the ith group, denoting group membership. The U_i's are independent, identically distributed and the distribution of U_i is not assumed to be Normal. The residual terms ϵ_{ij} are assumed to be realisations of a random variable which is Normally distributed, independently of U_i, with mean 0 and a variance σ^2 constant over all groups.

3.3 Assumptions and general formulation of Bayes' factor

In Chapter 2, we have already established a Bayesian approach to hypothesis testing. Here under the notation in Section 3.2 we have, by Bayes' rule, the relative posterior probabilities of two hypotheses given the evidence E, which can be written as

$$\frac{Pr(C|E)}{Pr(\bar{C}|E)} = \left[\frac{Pr(E|C)}{Pr(E|\bar{C})}\right] \times \left[\frac{Pr(C)}{Pr(\bar{C})}\right].$$

The second factor in brackets is the prior odds ratio in favour of C. The data-dependent term in the first set of brackets is the "Bayes' factor". The evidence is said to be in favour of C relative to $\bar{C}$ if the Bayes' factor exceeds one, that is , if the evidence E is more likely under hypothesis C than it is under hypothesis $\bar{C}$. Here only the evidence E is used in measuring the weight of evidence, but in some occasions there are other factors may be take into consideration. A fuller exposition is given in Evett (1984).

The interest here is the estimation of the Bayes' factor (BF)

$$Pr(E|C)/Pr(E|\bar{C}) \qquad (3.2)$$

in the particular case where the training data Z are generated by a random effects model with a non-Normal distribution for the random effects. Strictly speaking, the probabilities in (3.2) are conditional on Z and this is assumed implicitly in what follows.

In the example described earlier the observations are continuous measurements from hairs from individual cats and the data are taken to be continuous with E=(X,Y). Thus, the probability operators Pr in

(3.2) may be replaced by probability density functions f and the numerator of (3.2) may be written as

$$Pr(E|C) = f(x,y|C)\,dx\,dy = f(y|x,C)\,dy \times f(x|C)\,dx,$$

where x and y are particular realisations of X and Y, respectively. The denominator of (3.2) is a product of two marginal probability density functions. If $\bar{C}$ is true then X and Y are independent and

$$Pr(E|\bar{C}) = f(x|\bar{C})\,dx \times f(y|\bar{C})\,dy.$$

Also the probability density function of X is independent of C and $\bar{C}$ and so

$$f(x|\bar{C}) = f(x|C).$$

Thus, the Bayes factor (3.2) reduces to

$$\frac{f(y|x,C)}{f(y|\bar{C})}, \qquad (3.3)$$

which is a ratio of predictive and marginal p.d.fs. of Y.

In order to formulate the Bayes' factor (3.3), a further assumption is made: under the hypothesis C it is assumed that measurements of X and Y are Normally distributed about the true mean μ_C of their common source and have constant variance σ^2.

If C holds, the parameters of interest are μ_C and σ^2, denoted by $\theta = (\mu_C, \sigma^2)$. The numerator of the Bayes factor may now be written as

$$f(y|x,C) = \int f(y|\theta,C) \times f(\theta|x,C)\,d\theta.$$

The second term of the integrand is the posterior distribution of θ

given x and C. This is equal to

$$\frac{f(x|\theta,C) \times f(\theta|C)}{f(x|C)} .$$

Here $f(\theta|C)$ is a prior distribution of θ. The term $f(x|C)$ equals

$$\int f(x|\theta,C) \times f(\theta|C)\, d\theta.$$

The conditional probability density function, $f(x|\theta,C)$, for a particular value of X is a likelihood function of θ, and $f(x|C)$ is defined as a weighted or marginal likelihood.

If $\bar{C}$ holds the denominator $f(y|\bar{C})$ is the density of Y implied by the hypothesis $\bar{C}$ and may be written as

$$f(y|\bar{C}) = \int f(y|\theta^*,\bar{C}) \times f(\theta^*|\bar{C}) d\theta^*$$

where $\theta^* = (\mu^*,\sigma^2)$ and μ^* is the true, unknown, mean of the source of the measurements of Y. The values of the density functions are independent of C and $\bar{C}$. The conditioning on C or $\bar{C}$ may now be dropped to give the Bayes' factor, as was shown by Lindley (1977),

$$\frac{\int f(y|\theta)\, f(x|\theta)\, f(\theta)\, d\theta}{\int f(x|\theta)\, f(\theta)\, d\theta \int f(y|\theta^*)\, f(\theta^*)\, d\theta^*} . \qquad (3.4)$$

The density functions $f(\theta)$ and $f(\theta^*)$ may be assumed to be equivalent since the assumption is made earlier that the between group random factor U_i is independent and identically distributed. Notice also that the structure of (3.4) is such that the difference between μ_C and μ^* is not important and hence the difference is ignored and both are denoted by μ.

The factor (3.4) will be evaluated under four different sets of assumptions for the training data and the within-group variance σ^2 as follows:

1. Training data grouped, within-group variance known.
2. Training data ungrouped, within-group variance known.
3. Training data grouped, within-group variance unknown.
4. Training data ungrouped, within-group variance unknown.

3.4 Sampling distribution of the control and recovered data

If X consists of m measurements then a sufficient statistic for the true mean, if the variance σ^2 is assumed known, is the sample mean $\bar{X}$, which is Normally distributed about the unknown true value with variance σ^2/m. Similarly, $\bar{Y}$ denotes the mean of r measurements of Y. Under C, $\bar{Y}$ is also Normally distributed with mean μ_C and variance σ^2/r. The density functions $f(y|\theta)$ and $f(x|\theta)$ are replaced by the density function of $\bar{Y}$ and $\bar{X}$ for cases 1 and 2 in Section 3.3, namely

$$f(\bar{y}|\mu,\sigma^2) = \frac{\sqrt{r}}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{r(\bar{y}-\mu)^2}{2\sigma^2}\right\} \qquad (3.5)$$

and

$$f(\bar{x}|\mu,\sigma^2) = \frac{\sqrt{m}}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{m(\bar{x}-\mu)^2}{2\sigma^2}\right\}. \qquad (3.6)$$

However for cases 3 and 4 in the previous section, if σ^2 is unknown, the sample mean and variance of X are jointly sufficient for μ and σ^2. We can write $f(x|\theta)$ as $f(\bar{x},S_X^2|\mu,\sigma^2)$ which can be factorised to

give

$$f(\bar{x},S_X^2|\mu,\sigma^2) = f(\bar{x}|\mu,\sigma^2) \times f(S_X^2|\sigma^2), \qquad (3.7)$$

using the fact that $\bar{Y}$ and S_X^2 are statistically independent. Further the first term on the right of (3.7) is a Normal density function and the second term is a x^2 density function with (n-1) degrees of freedom. Similarly $f(y|\mu,\sigma^2)$ can be expressed as $f(\bar{y},S_y^2|\mu,\sigma^2)$, that is,

$$f(\bar{y},S_y^2|\mu,\sigma^2) = f(\bar{y}|\mu,\sigma^2) \times f(S_y^2|\sigma^2). \qquad (3.8)$$

The joint density functions (3.7) and (3.8) are usually written in the form of a Normal-gamma function.

3.5 Estimation of the Bayes' factor

3.5.1 Distribution of the group population mean μ

The formula of the Bayes' factor in (3.4) required the knowledge of the probability density function $f(\theta)$ where $\theta = (\mu,\sigma^2)$. The joint probability density function of the unknown parameters (μ,σ^2) can be factorised as follows, assuming independence between the unknown parameters,

$$f(\theta) = f(\mu,\sigma^2) = f(\mu|\nu) \times f(\sigma^2). \qquad (3.9)$$

where ν denotes one or more nuisance parameters, either assummed known or unknown. If σ^2 is known, (3.9) reduces to $f(\mu)$. Whereas if σ^2 is unknown, one has to specify the probability density function $f(\sigma^2)$ in modelling the Bayes' factor. This will be discussed where appropriate in the later sections.

As mentioned earlier the distribution of the unknwon parameter μ is relevant in the evaluation of weight of evidence. So an informative prior for μ is used. In Section 3.2, it is assumed that the distribution of U has been taken to be non-Normal so a Normal prior distribution for U cannot be used. Instead a kernel density estimate is used to construct the distribution of μ based on the training data. This method of acquiring the prior distribution for an unknown parameter is so called the Empirical Bayes (EB) method. Brief details of the EB method is described in Section 2.2.5. The distribution of μ is to be estimated under the assumption that the training data is grouped or not grouped.

a. Assumed grouped training data

Since the training data is grouped, the existence of a random structure in the training data suggests that the group means $\bar{z}_{i.}$ may be used as the data points in constructing the kernel density estimate for $f(\mu)$ where

$$\bar{z}_{i.} = \frac{1}{J} \sum_{j=1}^{J} z_{ij} .$$

The sample variance of the group means is given by

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (\bar{z}_{i.} - \bar{z}_{..})^2$$

where $\bar{z}_{..} = \sum_{i=1}^{n} \bar{z}_{i.}/n$. Adopting the method due to Habbema, Hermans and van den Broek (1974) (see Section 2.1.3 for details) in a univariate case, the kernel prior density for μ, assuming the training data are grouped, is then given by

$$\hat{f}(\mu) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi s^2 \hat{\lambda}_1^2)^{1/2}} \exp\left\{-\frac{(\mu-\bar{z}_{i.})^2}{2s^2\hat{\lambda}_1^2}\right\}. \qquad (3.10)$$

The smoothing parameter for the kernel density estimate is denoted by λ_1, and its estimate, $\hat{\lambda}_1$, is determined by the pseudo-maximum likelihood method.

Note that $\bar{z}_{i.}$ is used as a substitute for μ_i. The discrepancy in the results that this will cause is small when J is large as in the case discussed here where J = 10, but it may be important if J were small; such as 2 or 3. Sensitivity analysis of small changes in $\bar{z}_i$'s to the Bayes' factor are examined later in this chapter to measure this importance.

b. Assumed ungrouped training data

If the grouping of the training data is ignored, then the training set may be represented as

$$Z = (Z_1, Z_2, \ldots, Z_N)^T$$

where N = n×J. The training data may be thought of as one observation from each of N items. Let the sample variance of the full data set ignoring grouping be given by

$$s'^2 = \frac{1}{N-1}\sum_{\ell=1}^{N}(z_\ell - \bar{z}_{.})^2.$$

The estimate $\hat{\lambda}_2$ of the smoothing parameter λ_2 is obtained using pseudo-maximum likelihood techniques as in part a of this section and the kernel prior density for μ is now given by

$$\hat{f}(\mu) = \frac{1}{N} \sum_{\ell=1}^{N} \frac{1}{(2\pi s'^2 \hat{\lambda}_2^2)^{1/2}} \exp\left\{ - \frac{(\mu - z_\ell)^2}{2 s'^2 \hat{\lambda}_2^2} \right\} . \qquad (3.11)$$

Again the discrepancy in the results of the substitution of μ_i by z_ℓ is small if N is sufficiently large.

3.5.2 Training data grouped, within-group varinace known

Under the assumption of known within-group variance σ^2, the parameter vector Θ in Section 3.3 is just a scalar μ. The numerator of (3.4) may be represented as

$$\int f(y|\mu)\ f(x|\mu)\ f(\mu)\ d\mu.$$

The "known" value of the within-group variance σ^2 is taken to be the sample estimate

$$\hat{\sigma}^2 = \frac{1}{n(J-1)} \sum_{i=1}^{n} \sum_{j=1}^{J} (z_{ij} - \bar{z}_{i.})^2 . \qquad (3.12)$$

The conditional density functions of $\bar{Y}$ and of $\bar{X}$ are given by (3.5) and (3.6) respectively. Upon combining (3.5), (3.6) and (3.10) and integrating over μ, the numerator of (3.4) may be written as

$$\frac{1}{(2\pi a b_k n \hat{\sigma})} \exp\left\{ - \frac{(\bar{x} - \bar{y})^2}{2a^2 \hat{\sigma}^2} \right\} \sum_{i=1}^{n} \exp\left\{ - \frac{(w - \bar{z}_{i.})^2}{2 b_k^2} \right\} \qquad (3.13)$$

where $a^2 = m^{-1} + r^{-1}$, $b_k^2 = s^2 \hat{\lambda}_1^2 + (\hat{\sigma}^2/k)$, $k = m + r$, $w = (m\bar{x} + r\bar{y})/k$.

In view of Section 2.2.4, the marginal density functions of $\bar{X}$ and $\bar{Y}$ may be obtained by combining (3.6) with (3.10) and (3.5) with (3.10), respectively. The denominator of (3.4) is then the product of these two marginal density functions and is given by

$$\frac{1}{2\pi b_m b_r n^2}\left[\sum_{i=1}^{n}\exp\left\{-\frac{(\bar{x}-\bar{z}_{i.})^2}{2b_m^2}\right\}\sum_{i=1}^{n}\exp\left\{-\frac{(\bar{y}-\bar{z}_{i.})^2}{2b_r^2}\right\}\right]. \qquad (3.14)$$

The ratio of expression (3.13) and (3.14) is the Bayes' factor for the assumptions of grouped training data and known within-group variance.

3.5.3 Training data ungrouped, within-group variance known.

It is instructive to investigate the effect of the grouping on the estimation of the Bayes' factor. Normally, an analysis of variance would be done to investigate the between- and within-group variances. This will be discussed in the next chapter. The implications of the results of such an investigation for the estimation of the Bayes's factor are of interest. The investigation of the effect of the grouping is done by evaluating the Bayes' factor under the assumption that the training data are not grouped.

The "known" value of σ^2 is taken to be the same as (3.12) for direct comparison with the results of Section 3.5.2. In a similar manner to Section 3.5.2, the marginal density functions for $f(y|\mu)$ and $f(x|\mu)$ are given by (3.5) and (3.6) and $f(\mu)$ is given by (3.11). After some simplifications and integration, the numerator of (3.4) is then given by

$$\frac{1}{(2\pi a b_k' \hat{\sigma} N)}\exp\left\{-\frac{(\bar{x}-\bar{y})^2}{2a^2\hat{\sigma}^2}\right\}\sum_{i=1}^{N}\exp\left\{-\frac{(w-z_\ell)^2}{2b_k'^2}\right\} \qquad (3.15)$$

where $b_k'^2 = s'^2\hat{\lambda}_z^2 + (\hat{\sigma}^2/k)$ and a, w, k are as before (see after (3.13)). The denominator of (3.4) is given by

$$\frac{1}{2\pi b_m' b_r' N^2}\left[\sum_{i=1}^{N}\exp\left\{-\frac{(\bar{x}-z_\ell)^2}{2b_m'^2}\right\}\sum_{i=1}^{N}\exp\left\{-\frac{(\bar{y}-z_\ell)^2}{2b_r'^2}\right\}\right]. \qquad (3.16)$$

The ratio of expressions (3.15) and (3.16) is the Bayes' factor for the assumptions of ungrouped training data and known within-group variance.

3.5.4 Training data grouped, within-group variance unknown.

Here we assume the training data have a grouped structure and the within-group variance σ^2 is unknown. For convenience alone, the unknown within-group variance σ^2 is to be replaced by a new parameter $\tau = \sigma^{-2}$, called the precision which is a measure of precision for the within-group Normal distributions. We express the prior in terms of τ and the conjugate prior density function for τ is given by

$$f(\tau) = \frac{\{\beta_0/2\}^{\alpha_0/2}\,\tau^{(\alpha_0-2)/2}}{\Gamma(\alpha_0/2)}\exp\left\{-\frac{\beta_0\tau}{2}\right\}. \qquad (3.17)$$

If the two components of $\theta = (\mu,\sigma^2)$ are assumed to have independent priors then the prior density for θ can be expressed as the product of the prior densities $f(\mu)$ and $f(\tau)$ where $f(\mu)$ has an estimate in the form of (3.10). An informative prior for τ is obtained from the training set Z with α_0 and β_0 estimated by

$$\hat{\alpha}_0 = n(J-1) \text{ and } \hat{\beta}_0 = \sum_{i=1}^{n}\sum_{j=1}^{J}(z_{ij}-\bar{z}_{i.})^2,$$

respectively. Under the assumption of unknown within-group variance, the density function $f(y|\mu,\sigma^2)$ in (3.4) is a joint probability density function of $\bar{y}$, the sample mean, and s_y^2, the sample variance.

Similarly we could obtain $f(x|\mu,\sigma^2)$ as $f(\bar{x},s_x^2|\mu,\sigma^2)$ (see Section 3.4 for details). Before proceeding, some notation is required. Let

$$s_{xx} = \sum_{p=1}^{m}(x_p - \bar{x})^2, \qquad s_{yy} = \sum_{q=1}^{r}(y_q - \bar{y})^2,$$

$$H_1 = \beta_0 + s_{xx}, \qquad H_2 = \beta_0 + s_{yy},$$

$$H = H_1 + s_{yy} + \left[\frac{(\bar{y} - \bar{x})^2}{a^2}\right],$$

$$S_1(\mu) = \sum_{i=1}^{n}\exp\left\{-\frac{(\mu - \bar{z}_{i.})^2}{2s^2\lambda_i^2}\right\},$$

$$I_g(X) = \int_{-\infty}^{\infty}\frac{S_1(\mu)}{\{1 + mH_1^{-1}(\mu - \bar{x})^2\}^{(m+\alpha_0)/2}}\,d\mu,$$

$$I_g(Y) = \int_{-\infty}^{\infty}\frac{S_1(\mu)}{\{1 + rH_2^{-1}(\mu - \bar{y})^2\}^{(r+\alpha_0)/2}}\,d\mu,$$

$$I_g(W) = \int_{-\infty}^{\infty}\frac{S_1(\mu)}{\{1 + (r+m)H^{-1}(\mu - w)^2\}^{(m+r+\alpha_0)/2}}\,d\mu,$$

$$D(t,u,v) = \Gamma(t)\Gamma(u)\Gamma(v) \,/\, \Gamma(t+u+v), \quad \Gamma(t) = \int_0^{\infty} s^{t-1}\,e^{-s}\,ds.$$

After tedious manipulations and simplifications, the numerator $f(y|x,C)$ in (3.3), which can be summarised as $f(\bar{y},s_{yy}|\bar{x},s_{xx},C)$, may be estimated by

$$\frac{s_{yy}^{(r-3)/2}\; r^{1/2}\; H_1^{(m+\alpha_0)/2}\; H^{-(m+r+\alpha_0)/2}\; \{I_g(W)\}}{[D\{(m+\alpha_0)/2,(r-1)/2,1/2\}] \times \{I_g(X)\}}. \qquad (3.18)$$

The denominator $f(y|\bar{C})$ in (3.3) is estimated by

$$\frac{s_{yy}^{(r-3)/2}\; r^{1/2}\; \beta_0^{\alpha_0/2}\; (2\pi n^2 s^2 \lambda_i^2)^{-1/2}}{[D\{(\alpha_0/2,(r-1)/2,1/2\}] \times H_2^{(r+\alpha_0)/2} \times I_g(Y)}. \qquad (3.19)$$

The ratio of (3.18) and (3.19) is the Bayes' factor for the

assumptions of grouped training data and unknown within-group variance. This is not the ratio of two vague priors and thus problems caused by undefined constants do not arise (Spiegelhalter and Smith (1982)).

3.5.5 Training data ungrouped, within-group variance unknown.

It is assumed that the training data have no grouping structure and the variance σ^2, defined in Section 3.4, is assumed to be not known. Then strictly speaking no information is available on σ^2, so a vague prior for τ, the precision where $\tau = \sigma^{-2}$, is used, namely

$$f(\tau) = \tau^{-1} , \quad \tau > 0.$$

The prior density of $f(\mu)$ is estimated with the estimate given by (3.11). Using a similar argument as in Section 3.5.4, we obtained the joint probability density function of $\bar{X}$ and S_X^2 and of $\bar{Y}$ and S_y^2 in place of $f(x|\mu,\sigma^2)$ and $f(y|\mu,\sigma^2)$ in (3.4) with $\theta = (\mu,\sigma^2)$. After some simplifications the Bayes' factor (BF) can be estimated as

$$BF = \frac{\{s_{XX}^{m/2} s_{yy}^{r/2} N\hat{\lambda}_2 s' (2\pi)^{\frac{1}{2}} I_g(NW)\}}{\{B(r/2,m/2) H^{(r+m)/2} I_g(NY) I_g(NX)\}} ,$$

where $B(u,v) = \Gamma(u)\Gamma(v)/\Gamma(u+v)$,

$$I_g(NW) = \int_{-\infty}^{\infty} \frac{S_2(\mu)}{\{1+(r+m)H^{-1}(\mu-w)^2\}^{(r+m)/2}} d\mu,$$

$$I_g(NY) = \int_{-\infty}^{\infty} \frac{S_2(\mu)}{\{1+rs_{yy}^{-1}(\mu-\bar{y})^2\}^{r/2}} d\mu,$$

$$I_g(NX) = \int_{-\infty}^{\infty} \frac{S_2(\mu)}{\{1+ms_{XX}^{-1}(\mu-\bar{x})^2\}^{m/2}} d\mu,$$

$$S_2(\mu) = \sum_{\ell=1}^{N} \exp\left\{ - \frac{(\mu - z_\ell)^2}{2\hat{\lambda}_2^2 s'^2} \right\},$$

$$H = S_{xx} + S_{yy} + \left[\frac{(\bar{y} - \bar{x})^2}{a^2} \right]$$

This Bayes' factor exists only when m and r are both greater than 1. If either m or r equals 1, it is zero.

3.6 Example

Data are available on 10 (J=10) hairs from each of 22 cats (n=22) to form the training set Z. The measurement taken is the value of the medullary fraction, the ratio of the width of the central core of the hair to the total width of the hair. The measurements are restricted to the interval (0,1). In practice, for cat hairs, they are sufficiently far removed from the ends of the interval being mainly in the interval (0.5,0.8) that this constraint should not be important for kernel density estimation and consequently the formulation of the Bayes' factor. The 22 group means from which the prior density of μ is obtained are shown in Table 3.1. The kernel density estimates for the prior density of μ described in Section 3.5.1 using the ordinary and adaptive kernel method under the assumption of the training data are grouped or not grouped are plotted in Fig. 3.1. The adaptive kernel estimate has a longer tail than the ordinary kernel.

The Bayes' factor was evaluated for each of the four models [(3.5.2) to (3.5.5)] for different numbers of observations in the control and recovered measurements. In practice, we could always

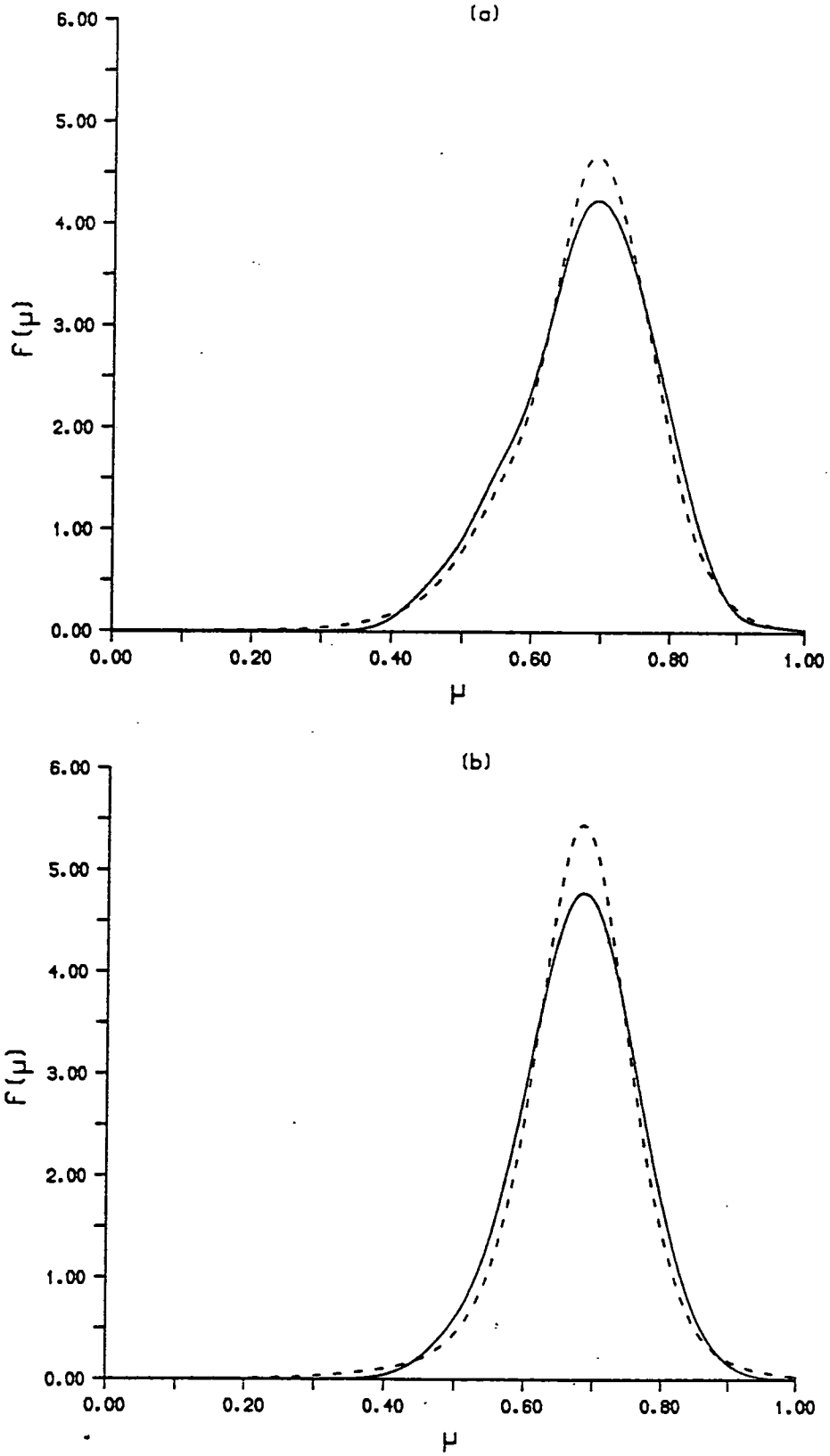


Fig. 3.1 Kernel density estimates for the prior density of μ using (a) 220 cat hairs (i.e. TS are ungrouped) and (b) 22 group means (i.e. TS are grouped); solid line represents the ordinary kernel method and dash line represents the adaptive kernel method.

ensure the number of control measurements is greater than the number of recovered measurements since we could take as many observations from the control as necessary, subject to constraints on laboratory resources. Different pairings of r and m will be chosen from, in this example, 1, 5 and 10. The value of m will always be at least r. The total number of combinations {r,m} for these chosen values is 6, that is {1,1}, {1,5}, {1,10}, {5,5}, {5,10} and {10,10}. The results from the assumed known within-group variance model are presented in Tables 3.2 - 3.7 for these respective combinations. Tables 3.2' - 3.7' show the results from the assumed unknown within-group variance model derived in Section 3.5.4 and 3.5.5. For example, Table 3.2 refers to the situation in which there is one control hair (m=1) and one recovered hair (r=1) with Table 3.3 referring to the situation with m=5, r=1 and so on. For these Tables the control hair measurement ($\bar{X}$) has three possible values 0.4, 0.6 and 0.8. The recovered hair measurement ($\bar{Y}$) takes values from 0.10 to 0.90 in steps of 0.05.

Table 3.1 The Ordered 22 Group means $\bar{z}_{i.}$

Group i	1	2	3	4	5	6	7	8	9
$\bar{z}_{i.}$	.5096	.5854	.5920	.6057	.6360	.6401	.6572	.6582	.6702
Group i	10	11	12	13	14	15	16	17	18
$\bar{z}_{i.}$	.6723	.6783	.6797	.6850	.6901	.6966	.7132	.7364	.7366
Group i	19	20	21	22					
$\bar{z}_{i.}$	.7451	.7530	.7671	.8187					

Table 3.2 Bayes' factor (with ordinary kernel) as a function of Y assuming σ^2 known given some values of $\bar{X}$ for r=1 & m=1.

TS	Grouped			Ungrouped		
Y $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	24.485	0.0090	0.0000	16.858	0.0118	0.0000
0.15	28.543	0.0214	0.0000	18.371	0.0231	0.0000
0.20	30.932	0.0484	0.0000	19.234	0.0452	0.0000
0.25	30.987	0.1045	0.0000	19.203	0.0884	0.0000
0.30	28.389	0.2146	0.0001	18.043	0.1731	0.0000
0.35	23.336	0.4147	0.0007	15.607	0.3342	0.0003
0.40	16.717	0.7342	0.0035	12.017	0.6157	0.0018
0.45	10.090	1.1454	0.0144	7.8943	1.0299	0.0085
0.50	5.0228	1.5197	0.0488	4.2407	1.4842	0.0348
0.55	2.0765	1.6934	0.1354	1.8106	1.7698	0.1144
0.60	0.7342	1.6009	0.3130	0.6157	1.7245	0.2958
0.65	0.2287	1.3057	0.6170	0.1728	1.3965	0.6156
0.70	0.0638	0.9284	1.0555	0.0419	0.9629	1.0687
0.75	0.0159	0.5771	1.5782	0.0090	0.5729	1.5853
0.80	0.0035	0.3130	2.0648	0.0018	0.2958	2.0299
0.85	0.0007	0.1478	2.3608	0.0003	0.1338	2.2600
0.90	0.0001	0.0609	2.3612	0.0000	0.0536	2.2003

Table 3.3 Bayes' factor (with ordinary kernel) as a function of Y assuming σ^2 known given some values of $\bar{X}$ for r=1 & m=5.

TS	Grouped			Ungrouped		
Y $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	63.087	0.0000	0.0000	35.105	0.0000	0.0000
0.15	89.878	0.0002	0.0000	44.233	0.0001	0.0000
0.20	108.35	0.0014	0.0000	50.175	0.0008	0.0000
0.25	109.62	0.0081	0.0000	50.606	0.0044	0.0000
0.30	91.720	0.0389	0.0000	44.485	0.0217	0.0000
0.35	61.907	0.1533	0.0000	33.020	0.0918	0.0000
0.40	32.476	0.4749	0.0000	19.755	0.3194	0.0000
0.45	12.665	1.1050	0.0002	8.9830	0.8622	0.0001
0.50	3.5478	1.8642	0.0018	2.9233	1.6988	0.0014
0.55	0.7094	2.2629	0.0128	0.6510	2.3326	0.0115
0.60	0.1032	2.0110	0.0653	0.0982	2.2043	0.0652
0.65	0.0112	1.3393	0.2497	0.0103	1.4683	0.2598
0.70	0.0009	0.6791	0.7266	0.0008	0.7144	0.7511
0.75	0.0001	0.2637	1.6190	0.0000	0.2597	1.6143
0.80	0.0000	0.0784	2.7612	0.0000	0.0714	2.6111
0.85	0.0000	0.0178	3.5981	0.0000	0.0150	3.2138
0.90	0.0000	0.0031	3.5873	0.0000	0.0024	3.0439

Table 3.4 Bayes' factor (with ordinary kernel) as a function of Y assuming σ^2 known given some values of $\bar{X}$ for r=1 & m=10.

TS	Grouped			Ungrouped		
Y $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	73.453	0.0000	0.0000	39.212	0.0000	0.0000
0.15	112.13	0.0001	0.0000	52.524	0.0000	0.0000
0.20	140.28	0.0006	0.0000	61.675	0.0003	0.0000
0.25	142.62	0.0042	0.0000	62.666	0.0020	0.0000
0.30	116.07	0.0251	0.0000	53.973	0.0126	0.0000
0.35	73.749	0.1185	0.0000	38.147	0.0654	0.0000
0.40	35.232	0.4225	0.0000	21.101	0.2678	0.0000
0.45	12.098	1.0186	0.0001	8.6045	0.8158	0.0000
0.50	2.8841	1.9455	0.0007	2.4327	1.7366	0.0006
0.55	0.4740	2.4095	0.0063	0.4553	2.4664	0.0059
0.60	0.0547	2.1008	0.0405	0.0558	2.3081	0.0420
0.65	0.0045	1.3206	0.1871	0.0046	1.4586	0.2001
0.70	0.0003	0.6082	0.6330	0.0003	0.6455	0.6654
0.75	0.0000	0.2065	1.5782	0.0000	0.2049	1.5832
0.80	0.0000	0.0517	2.8996	0.0000	0.0472	2.7294
0.85	0.0000	0.0095	3.9191	0.0000	0.0080	3.4491
0.90	0.0000	0.0013	3.9021	0.0000	0.0010	3.2323

Table 3.5 Bayes' factor (with ordinary kernel) as a function of Y assuming σ^2 known given some values of $\bar{X}$ for r=5 & m=5.

TS	Grouped			Ungrouped		
$\bar{Y}$ $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	0.0001	0.0000	0.0000	0.0014	0.0000	0.0000
0.15	0.0139	0.0000	0.0000	0.0492	0.0000	0.0000
0.20	0.6076	0.0000	0.0000	0.8462	0.0000	0.0000
0.25	9.9717	0.0000	0.0000	7.2660	0.0000	0.0000
0.30	61.468	0.0000	0.0000	30.587	0.0000	0.0000
0.35	141.32	0.0000	0.0000	60.833	0.0000	0.0000
0.40	118.19	0.0003	0.0000	53.087	0.0002	0.0000
0.45	33.418	0.0189	0.0000	17.197	0.0151	0.0000
0.50	27.353	0.4002	0.0000	1.5716	0.3588	0.0000
0.55	0.0556	2.2709	0.0000	0.0362	2.3084	0.0000
0.60	0.0003	3.3696	0.0001	0.0002	3.8239	0.0001
0.65	0.0000	1.4239	0.0076	0.0000	1.6015	0.0072
0.70	0.0000	0.1809	0.2222	0.0008	0.1943	0.2127
0.75	0.0000	0.0069	1.9655	0.0000	0.0069	1.8308
0.80	0.0000	0.0001	5.1782	0.0000	0.0001	4.5297
0.85	0.0000	0.0000	4.0271	0.0000	0.0000	3.3788
0.90	0.0000	0.0000	0.9487	0.0000	0.0000	0.8231

Table 3.6 Bayes' factor (with ordinary kernel) as a function of Y assuming σ^2 known given some values of $\bar{X}$ for r=5 & m=10.

TS	Grouped			Ungrouped		
$\bar{Y}$ $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.15	0.0005	0.0000	0.0000	0.0028	0.0000	0.0000
0.20	0.0989	0.0000	0.0000	0.1685	0.0000	0.0000
0.25	4.8677	0.0000	0.0000	3.7403	0.0000	0.0000
0.30	61.206	0.0000	0.0000	29.492	0.0000	0.0000
0.35	194.83	0.0000	0.0000	78.860	0.0000	0.0000
0.40	152.49	0.0000	0.0000	65.483	0.0000	0.0000
0.45	27.049	0.0037	0.0000	13.967	0.0026	0.0000
0.50	0.9175	0.2236	0.0000	0.5603	0.1838	0.0000
0.55	0.0050	2.2942	0.0000	0.0036	2.2649	0.0000
0.60	0.0000	3.8949	0.0000	0.0000	4.4520	0.0000
0.65	0.0000	1.2060	0.0008	0.0000	1.3694	0.0008
0.70	0.0000	0.0725	0.0838	0.0000	0.0776	0.0830
0.75	0.0000	0.0009	1.6281	0.0000	0.0008	1.5356
0.80	0.0000	0.0000	6.0745	0.0000	0.0000	5.2434
0.85	0.0000	0.0000	4.3130	0.0000	0.0000	3.4956
0.90	0.0000	0.0000	0.5992	0.0000	0.0000	0.5013

Table 3.7 Bayes' factor (with ordinary kernel) as a function of Y assuming σ^2 known given some values of $\bar{X}$ for r=10 & m=10.

TS	Grouped			Ungrouped		
Y $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.20	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
0.25	0.0529	0.0000	0.0000	0.0598	0.0000	0.0000
0.30	9.0059	0.0000	0.0000	5.0831	0.0000	0.0000
0.35	144.78	0.0000	0.0000	59.473	0.0000	0.0000
0.40	215.51	0.0000	0.0000	87.378	0.0000	0.0000
0.45	27.735	0.0000	0.0000	13.189	0.0000	0.0000
0.50	0.2581	0.0323	0.0000	0.1375	0.0289	0.0000
0.55	0.0001	1.5986	0.0000	0.0001	1.6040	0.0000
0.60	0.0000	4.7781	0.0000	0.0000	5.5142	0.0000
0.65	0.0000	0.9599	0.0000	0.0000	1.0924	0.0000
0.70	0.0000	0.0139	0.0173	0.0008	0.0151	0.0163
0.75	0.0000	0.0000	1.3620	0.0000	0.0000	1.2453
0.80	0.0000	0.0000	7.5926	0.0000	0.0000	6.4428
0.85	0.0000	0.0000	2.9701	0.0000	0.0000	2.4168
0.90	0.0000	0.0000	0.0848	0.0000	0.0000	0.0751

Table 3.2' Bayes' factor (with ordinary kernel) as a function of Y assuming σ^2 unknown given some values of $\bar{X}$ for r=1 & m=1. Grouped model

Y \ $\bar{X}$	0.4	0.6	0.8
0.10	24.984	0.0199	0.0000
0.15	28.161	0.0362	0.0000
0.20	30.103	0.0674	0.0000
0.25	30.106	0.1266	0.0000
0.30	27.690	0.2372	0.0003
0.35	22.879	0.4329	0.0013
0.40	16.459	0.7432	0.0050
0.45	9.9635	1.1446	0.0176
0.50	4.9746	1.5150	0.0539
0.55	2.0697	1.6910	0.1410
0.60	0.7432	1.6011	0.3167
0.65	0.2392	1.3054	0.6173
0.70	0.0708	0.9272	1.0537
0.75	0.0195	0.5774	1.5768
0.80	0.0050	0.3167	2.0627
0.85	0.0012	0.1539	2.3536
0.90	0.0003	0.0672	2.3482

Table 3.3' Bayes' factor (with ordinary kernel) as a function of Y assuming σ^2 unknown given some values of $\bar{X}$ for r=1 & m=5. Grouped model

Y \ $\bar{X}$	0.4	0.6	0.8
0.10	70.716	0.0002	0.0000
0.15	92.013	0.0009	0.0000
0.20	106.42	0.0035	0.0000
0.25	106.47	0.0139	0.0000
0.30	89.420	0.0521	0.0000
0.35	60.827	0.1750	0.0000
0.40	32.096	0.4969	0.0000
0.45	12.556	1.1128	0.0003
0.50	3.5344	1.8580	0.0025
0.55	0.7184	2.2575	0.0147
0.60	0.1092	2.0102	0.0688
0.65	0.0130	1.3381	0.2528
0.70	0.0013	0.6789	0.7265
0.75	0.0001	0.2668	1.6169
0.80	0.0000	0.0825	2.7573
0.85	0.0000	0.0204	3.5834
0.90	0.0000	0.0042	3.5646

Table 3.4' Bayes' factor (with ordinary kernel) as a function of Y assuming σ^2 unknown given some values of $\bar{X}$ for r=1 & m=10. Grouped model.

Y \ $\bar{X}$	0.4	0.6	0.8
0.10	84.555	0.0001	0.0000
0.15	115.93	0.0003	0.0000
0.20	138.03	0.0016	0.0000
0.25	138.42	0.0079	0.0000
0.30	113.10	0.0353	0.0000
0.35	72.487	0.1385	0.0000
0.40	34.848	0.4462	0.0000
0.45	12.004	1.0968	0.0001
0.50	2.8760	1.9395	0.0010
0.55	0.4816	2.4036	0.0075
0.60	0.0585	2.0999	0.0432
0.65	0.0054	1.3194	0.1901
0.70	0.0004	0.6084	0.6332
0.75	0.0000	0.2096	1.5762
0.80	0.0000	0.0550	2.8959
0.85	0.0000	0.0112	3.9032
0.90	0.0000	0.0018	3.8782

Table 3.5' Bayes' factor (with ordinary kernel) as a function of Y assuming σ^2 unknown given some values of $\bar{X}$ for r = 5 & m = 5.

TS	Grouped			Ungrouped		
$\bar{Y}$ \ $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	0.0021	0.0000	0.0000	0.0491	0.0019	0.0006
0.15	0.0577	0.0000	0.0000	0.1234	0.0032	0.0007
0.20	1.0879	0.0000	0.0000	0.3785	0.0054	0.0008
0.25	11.836	0.0000	0.0000	1.3892	0.0082	0.0008
0.30	62.701	0.0000	0.0000	5.0196	0.0100	0.0006
0.35	140.02	0.0000	0.0000	12.812	0.0107	0.0003
0.40	117.41	0.0005	0.0000	14.829	0.0163	0.0002
0.45	33.241	0.0224	0.0000	5.1156	0.0466	0.0003
0.50	2.8002	0.4105	0.0000	0.7783	0.2003	0.0006
0.55	0.0657	2.2657	0.0000	0.1037	0.8103	0.0017
0.60	0.0005	3.3663	0.0001	0.0163	1.3917	0.0055
0.65	0.0000	1.4215	0.0090	0.0032	0.5810	0.0227
0.70	0.0000	0.1854	0.2278	0.0009	0.1120	0.1230
0.75	0.0000	0.0081	1.9614	0.0004	0.0214	0.6636
0.80	0.0000	0.0001	5.1695	0.0002	0.0055	1.6496
0.85	0.0000	0.0000	4.0145	0.0002	0.0022	1.1455
0.90	0.0000	0.0000	0.9722	0.0003	0.0015	0.4065

Table 3.6' Bayes' factor (with ordinary kernel) as a_fucntion of Y assuming σ^2 unknown given some values of X for r = 5 & m = 10.

TS	Grouped			Ungrouped		
$\bar{Y}$ $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	0.0001	0.0000	0.0000	0.0027	0.0000	0.0000
0.15	0.0052	0.0000	0.0000	0.0155	0.0000	0.0000
0.20	0.2587	0.0000	0.0000	0.1106	0.0001	0.0000
0.25	6.5203	0.0000	0.0000	0.8968	0.0003	0.0000
0.30	63.841	0.0000	0.0000	6.0852	0.0005	0.0000
0.35	193.32	0.0000	0.0000	21.853	0.0010	0.0000
0.40	151.69	0.0000	0.0000	24.453	0.0029	0.0000
0.45	26.969	0.0049	0.0000	5.6768	0.0168	0.0000
0.50	0.9650	0.2351	0.0000	0.4308	0.1436	0.0000
0.55	0.0068	2.2911	0.0000	0.0243	0.9760	0.0001
0.60	0.0000	3.8920	0.0000	0.0016	1.9515	0.0007
0.65	0.0000	1.2053	0.0011	0.0001	0.6123	0.0058
0.70	0.0000	0.0763	0.0882	0.0008	0.0636	0.0682
0.75	0.0000	0.0011	1.6268	0.0000	0.0057	0.6870
0.80	0.0000	0.0000	6.0666	0.0000	0.0007	2.3037
0.85	0.0000	0.0000	4.3040	0.0000	0.0001	1.4534
0.90	0.0000	0.0000	0.6292	0.0000	0.0000	0.3370

Table 3.7' Bayes' factor (with ordinary kernel) as a_function of Y assuming σ^2 unknown given some values of X for r = 10 & m = 10.

TS	Grouped			Ungrouped		
$\bar{Y}$ $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	0.0000	0.0000	0.0000	0.0059	0.0000	0.0000
0.15	0.0000	0.0000	0.0000	0.0395	0.0000	0.0000
0.20	0.0003	0.0000	0.0000	0.2706	0.0000	0.0000
0.25	0.1060	0.0000	0.0000	1.3747	0.0000	0.0000
0.30	10.226	0.0000	0.0000	6.4308	0.0000	0.0000
0.35	144.70	0.0000	0.0000	31.078	0.0000	0.0000
0.40	214.75	0.0000	0.0000	47.386	0.0001	0.0000
0.45	27.771	0.0001	0.0000	7.5511	0.0017	0.0000
0.50	0.2922	0.0365	0.0000	0.2250	0.0482	0.0000
0.55	0.0003	1.6028	0.0000	0.0044	0.9458	0.0000
0.60	0.0000	4.7769	0.0000	0.0001	3.1469	0.0000
0.65	0.0000	0.9626	0.0000	0.0000	0.6480	0.0008
0.70	0.0000	0.0157	0.0196	0.0000	0.0253	0.0027
0.75	0.0000	0.0000	1.3657	0.0000	0.0008	0.7390
0.80	0.0000	0.0000	7.5877	0.0000	0.0000	3.6917
0.85	0.0000	0.0000	2.9767	0.0000	0.0000	1.4134
0.90	0.0000	0.0000	0.0961	0.0000	0.0000	0.1221

Various features of the Bayes' factor are apparent from the Tables 3.2 - 3.7. The results in Table 3.7 for any given model and value of $\bar{X}$ are much less variable than those in Table 3.2, reflecting the intuitive feeling that a more precise evaluation of the weight of evidence will be obtained due to the increasing number of control and recovered observations. The values of $\bar{X}$ were chosen to be of varying degrees of rarity for the value of the medullary fraction in cat hairs. The value of 0.6 is fairly common, that of 0.8 not so common and that of 0.4 quite rare. Suppose the control and recovered hairs have similar mean values for the medullary fraction measurements. It is desirable that a measure of the weight of evidence should give less weight to this similarity if the mean values are relatively common than if they are relatively rare. Medullary fraction values of about 0.6 are relatively common, those of about 0.4 are relatively rare. Thus, from Table 3.7, if $\bar{X}$ and $\bar{Y}$ are identical, far greater weight is given to the evidence if $\bar{X}$ and $\bar{Y}$ equal 0.4 than if they equal 0.6. A comparison of these results with those of Table 3.2 show that more weight is given to the match if more (r=m=10) hairs are involved than if few (r=m=1) hairs are involved. Notice that in Tables 3.2 - 3.7 and 3.2'- 3.7', there is little difference between the results obtained for grouped TS model from the assumed known and unknown within-group variance models. This is because the informative prior for τ used in the assumed known variance model is based on n(J-1) degrees of freedom, which in this case equals 198. However, in Tables 3.5 - 3.7 and 3.5' - 3.7' there is a substantial difference for ungrouped TS model between the assumed known and unknown within-group variance models. Such differences include having the maximum in different places (see Table 3.5 for example).

This is because a vague prior for τ is used in the assumed unknown within-group variance model, for the reason given in Section 3.5.5.

The effect of ignoring the grouping structure in the training data is that the maximum of BF is considerably reduced when $\bar{X} = 0.4$ in most of the combinations of r and m. In general the values of BF under the ungrouped model are slightly higher than the grouped model when $\bar{X}$ and $\bar{Y}$ are both common.

An adaptive kernel density estimate is also used to model the density function of μ in (3.9). This is to safeguard the possibility of trouble at the tails of the ordinary kernel estimate. The values of the Bayes' factor, using the adaptive kernel density estimate for μ, are shown in Tables 3.8 - 3.13. Consider the case where r=1 and m=1. Table 3.8 shows that there is not much difference between the ordinary and adaptive kernels when $\bar{X}$ equal 0.6 and 0.8. However, differences occur when $\bar{X}$ is rare, such as 0.4. The position of the maximum of the BF is being shifted to $\bar{Y} = 0.35$ when $\bar{X} = 0.4$, compared with $\bar{Y} = 0.25$ when the ordinary kernel was used (see Table 3.2). Also the values of BF when $\bar{X}$ and $\bar{Y}$ are both rare are reduced by approximately 50% in most cases but the values of BF are still much larger than the values when $\bar{X}$ and $\bar{Y}$ are common.

There is a slightly paradoxical feature of Tables 3.2 - 3.6. The maximum value of the Bayes' factor does not occur when $\bar{X} = \bar{Y}$. This is due to the effect of the training data Z. With only one control and one recovered hair the training data has an influence on the weight of the evidence. Suppose, for illustrative purposes, that

Table 3.8 Bayes' factor (with adaptive kernel) assuming σ^2 known given $\bar{X}$ = 0.4, 0.6 and 0.8; r=1 & m=1.

TS	Grouped			Ungrouped		
$Y^{\bar{X}}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	1.7036	0.0000	0.0000	1.3234	0.0001	0.0000
0.15	4.3569	0.0003	0.0000	3.4483	0.0006	0.0000
0.20	9.2049	0.0019	0.0000	7.1813	0.0037	0.0000
0.25	16.005	0.0109	0.0000	11.917	0.0186	0.0000
0.30	22.503	0.0525	0.0000	15.660	0.0745	0.0000
0.35	24.476	0.2012	0.0004	16.121	0.2316	0.0003
0.40	19.398	0.5617	0.0034	12.969	0.5518	0.0018
0.45	11.022	1.0913	0.0176	8.2541	1.0229	0.0098
0.50	4.7724	1.5370	0.0623	4.1927	1.5065	0.0411
0.55	1.7260	1.7001	0.1642	1.6937	1.7780	0.1311
0.60	0.5617	1.5829	0.3508	0.5518	1.7066	0.3221
0.65	0.1718	1.2945	0.6450	0.1523	1.3787	0.6382
0.70	0.0500	0.9460	1.0561	0.0376	0.9677	1.0724
0.75	0.0137	0.6159	1.5576	0.0086	0.5969	1.5712
0.80	0.0034	0.3508	2.0549	0.0018	0.3221	2.0196
0.85	0.0007	0.1694	2.3698	0.0003	0.1502	2.2635
0.90	0.0001	0.0662	2.2955	0.0001	0.0595	2.1824

Table 3.9 Bayes' factor (with adaptive kernel) assuming σ^2 known given $\bar{X}$ = 0.4, 0.6 and 0.8; r=1 & m=5.

TS	Grouped			Ungrouped		
$Y^{\bar{X}}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	0.3567	0.0000	0.0000	0.3663	0.0000	0.0000
0.15	1.9860	0.0000	0.0000	1.9829	0.0000	0.0000
0.20	7.7619	0.0000	0.0000	7.2205	0.0001	0.0000
0.25	21.106	0.0007	0.0000	17.596	0.0009	0.0000
0.30	38.874	0.0093	0.0000	28.405	0.0098	0.0000
0.35	45.779	0.0793	0.0000	29.919	0.0679	0.0000
0.40	31.966	0.4002	0.0000	20.452	0.3065	0.0000
0.45	12.863	1.1567	0.0002	9.2012	0.9108	0.0001
0.50	3.1603	2.0258	0.0024	2.7650	1.8095	0.0017
0.55	0.5232	2.3674	0.0153	0.5573	2.4136	0.0133
0.60	0.0637	2.0116	0.0708	0.0767	2.2007	0.0705
0.65	0.0061	1.3159	0.2531	0.0076	1.4373	0.2660
0.70	0.0005	0.6814	0.7168	0.0006	0.7054	0.7481
0.75	0.0000	0.2802	1.6161	0.0000	0.2665	1.6100
0.80	0.0000	0.0900	2.8576	0.0000	0.0777	2.6620
0.85	0.0000	0.0218	3.8362	0.0000	0.0173	3.3505
0.90	0.0000	0.0038	3.7147	0.0000	0.0029	3.1528

Table 3.10 Bayes' factor (with adaptive kernel) assuming σ^2 known given $\bar{X}$ = 0.4, 0.6 and 0.8; r=1 & m=10.

TS	Grouped			Ungrouped		
$\bar{Y}$ $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	0.2229	0.0000	0.0000	0.2530	0.0000	0.0000
0.15	1.5087	0.0000	0.0000	1.6373	0.0000	0.0000
0.20	6.8718	0.0000	0.0000	6.8292	0.0000	0.0000
0.25	20.859	0.0004	0.0000	18.264	0.0004	0.0000
0.30	41.055	0.0063	0.0000	30.993	0.0058	0.0000
0.35	49.421	0.0646	0.0000	32.854	0.0498	0.0000
0.40	33.715	0.3724	0.0000	21.635	0.2635	0.0000
0.45	12.658	1.1787	0.0001	8.9732	0.8795	0.0000
0.50	2.7692	2.1688	0.0009	2.3787	1.8777	0.0007
0.55	0.3894	2.5578	0.0072	0.4047	2.5744	0.0067
0.60	0.0384	2.1086	0.0425	0.0450	2.3089	0.0444
0.65	0.0028	1.2880	0.1853	0.0034	1.4207	0.2011
0.70	0.0002	0.5999	0.6162	0.0002	0.6299	0.6550
0.75	0.0000	0.2139	1.5698	0.0000	0.2065	1.5723
0.80	0.0000	0.0575	3.0161	0.0000	0.0519	2.7927
0.85	0.0000	0.0113	4.2294	0.0000	0.0090	3.6360
0.90	0.0000	0.0015	4.1108	0.0000	0.0012	3.4061

Table 3.11 Bayes' factor (with adaptive kernel) assuming σ^2 known given $\bar{X}$ = 0.4, 0.6 and 0.8; r=5 & m=5.

TS	Grouped			Ungrouped		
$\bar{Y}$ $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.20	0.0041	0.0000	0.0000	0.0034	0.0000	0.0000
0.25	0.3612	0.0000	0.0000	0.2926	0.0000	0.0000
0.30	8.4270	0.0000	0.0000	6.4202	0.0000	0.0000
0.35	51.955	0.0000	0.0000	35.795	0.0000	0.0000
0.40	79.735	0.0001	0.0000	49.507	0.0002	0.0000
0.45	26.665	0.0172	0.0000	16.256	0.0160	0.0000
0.50	1.8818	0.4386	0.0000	1.3419	0.3806	0.0000
0.55	0.0313	2.4958	0.0000	0.0300	2.4348	0.0000
0.60	0.0001	3.4872	0.0001	0.0002	3.9074	0.0001
0.65	0.0000	1.4150	0.0085	0.0000	1.5818	0.0080
0.70	0.0000	0.1900	0.2281	0.0000	0.1964	0.2200
0.75	0.0000	0.0083	2.0740	0.0000	0.0076	1.8945
0.80	0.0000	0.0001	5.8221	0.0000	0.0001	4.8722
0.85	0.0000	0.0000	4.5440	0.0000	0.0000	3.6548
0.90	0.0000	0.0000	0.8543	0.0000	0.0000	0.7454

Table 3.12 Bayes' factor (with adaptive kernel) assuming σ^2 known given $\bar{X}$ = 0.4, 0.6 and 0.8; r=5 & m=10.

TS	Grouped			Ungrouped		
$\bar{Y}$ \ $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.20	0.0002	0.0000	0.0000	0.0002	0.0000	0.0000
0.25	0.0663	0.0000	0.0000	0.0606	0.0000	0.0000
0.30	4.6089	0.0000	0.0000	3.8207	0.0000	0.0000
0.35	53.120	0.0000	0.0000	38.307	0.0000	0.0000
0.40	94.486	0.0000	0.0000	59.203	0.0000	0.0000
0.45	22.300	0.0035	0.0000	13.393	0.0028	0.0000
0.50	0.6659	0.2574	0.0000	0.4692	0.1999	0.0000
0.55	0.0028	2.6139	0.0000	0.0028	2.4363	0.0000
0.60	0.0000	4.0741	0.0000	0.0000	4.5760	0.0000
0.65	0.0000	1.1815	0.0009	0.0000	1.3384	0.0009
0.70	0.0000	0.0744	0.0831	0.0000	0.7675	0.0843
0.75	0.0000	0.0010	1.7043	0.0000	0.0009	1.5762
0.80	0.0000	0.0000	6.9294	0.0000	0.0000	5.7008
0.85	0.0000	0.0000	4.9857	0.0000	0.0000	3.8777
0.90	0.0000	0.0000	0.5426	0.0000	0.0000	0.4552

Table 3.13 Bayes' factor (with adaptive kernel) assuming σ^2 known given $\bar{X}$ = 0.4, 0.6 and 0.8; r=10 & m=10.

TS	Grouped			Ungrouped		
$\bar{Y}$ \ $\bar{X}$	0.4	0.6	0.8	0.4	0.6	0.8
0.10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0008	0.0000	0.0000	0.0007	0.0000	0.0000
0.30	0.6951	0.0000	0.0000	0.5470	0.0000	0.0000
0.35	37.325	0.0000	0.0000	26.517	0.0000	0.0000
0.40	119.47	0.0000	0.0000	75.750	0.0000	0.0000
0.45	19.718	0.0000	0.0000	11.977	0.0000	0.0000
0.50	0.1576	0.0361	0.0000	0.1094	0.0308	0.0000
0.55	0.0001	1.8055	0.0000	0.0001	1.7141	0.0000
0.60	0.0000	5.0692	0.0000	0.0000	5.7109	0.0000
0.65	0.0000	0.9603	0.0000	0.0000	1.0824	0.0000
0.70	0.0000	0.0148	0.0179	0.0000	0.0153	0.0170
0.75	0.0000	0.0000	1.4703	0.0000	0.0000	1.3088
0.80	0.0000	0.0000	8.8215	0.0000	0.0000	7.1030
0.85	0.0000	0.0000	3.4354	0.0000	0.0000	2.6683
0.90	0.0000	0.0000	0.0730	0.0000	0.0000	0.0633

μ is Normally distributed with mean ξ and variance η^2. For the case of known σ^2, the numerator of (3.3), $f(y|x,C)$ may be written as

$$\int f(y|\mu,C)\ f(\mu|x,C)\ d\mu.$$

The second term in the integrand is the posterior p.d.f. of μ, given x, and as was shown in Section 2.2.1, $f(\mu|x,C)$ is also Normally distributed with mean and variance

$$\mu_p = [\{\bar{x}/(\sigma^2/m)\}+\{\xi/\eta^2\}]/[\{1/(\sigma^2/m)\}+\{1/\eta^2\}]$$

and

$$\sigma_p^2 = 1/[\{1/(\sigma^2/m)\}+\{1/\eta^2\}],$$

respectively. $\bar{X}$ is defined as in Section 3.5. The maximum of this distribution will not be at $\bar{x}$ unless η^2 is large relative to σ^2; the so-called 'shrinkage' effect in which the maximum of $f(\mu|x,C)$ is shrunk towards ξ. Lindley (1977) assumed that σ^2 was small in comparison with η^2 and this ensured that $f(\mu|x,C)$ had a maximum at the control value $\bar{x}$.

To illustrate this phenomenon, using Lindley's model (i.e. $f(\mu) \sim N(\xi,\eta^2)$) the numerator of the Bayes' factor, $f(Y|X,C)$, given X = 0.0(0.2)1.0 is plotted in Fig. 3.2 for various values of r and m. The density function is plotted in range between 0.0 and 1.0, since the medullary fraction is restricted to this range. As can be clearly seen as m increases the maximum of $f(Y|X,C)$ is shifted towards X for X = 0.0(0.2)1.0. Whereas, the increment of r reduces the variability of $f(Y|X,C)$. Similar features have also appeared when the numerator of the Bayes' factor under the kernel known σ^2 model (see Fig. 3.3). The most significant difference between the

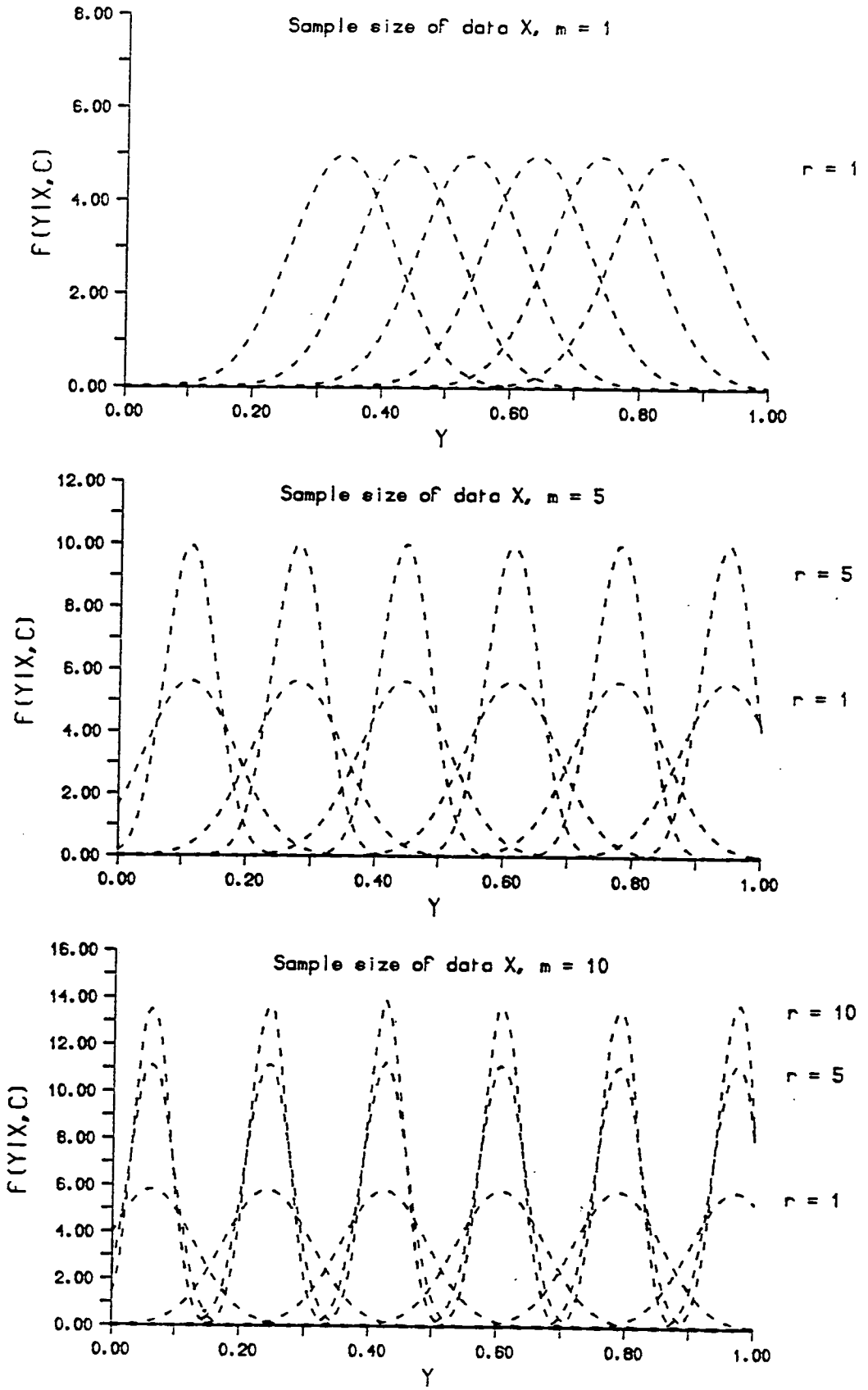


Fig. 3.2 Numerator of the Bayes' Factor, $f(Y|X,C)$ given X = 0.0(0.2)1.0 -- Normal and known σ^2 model.

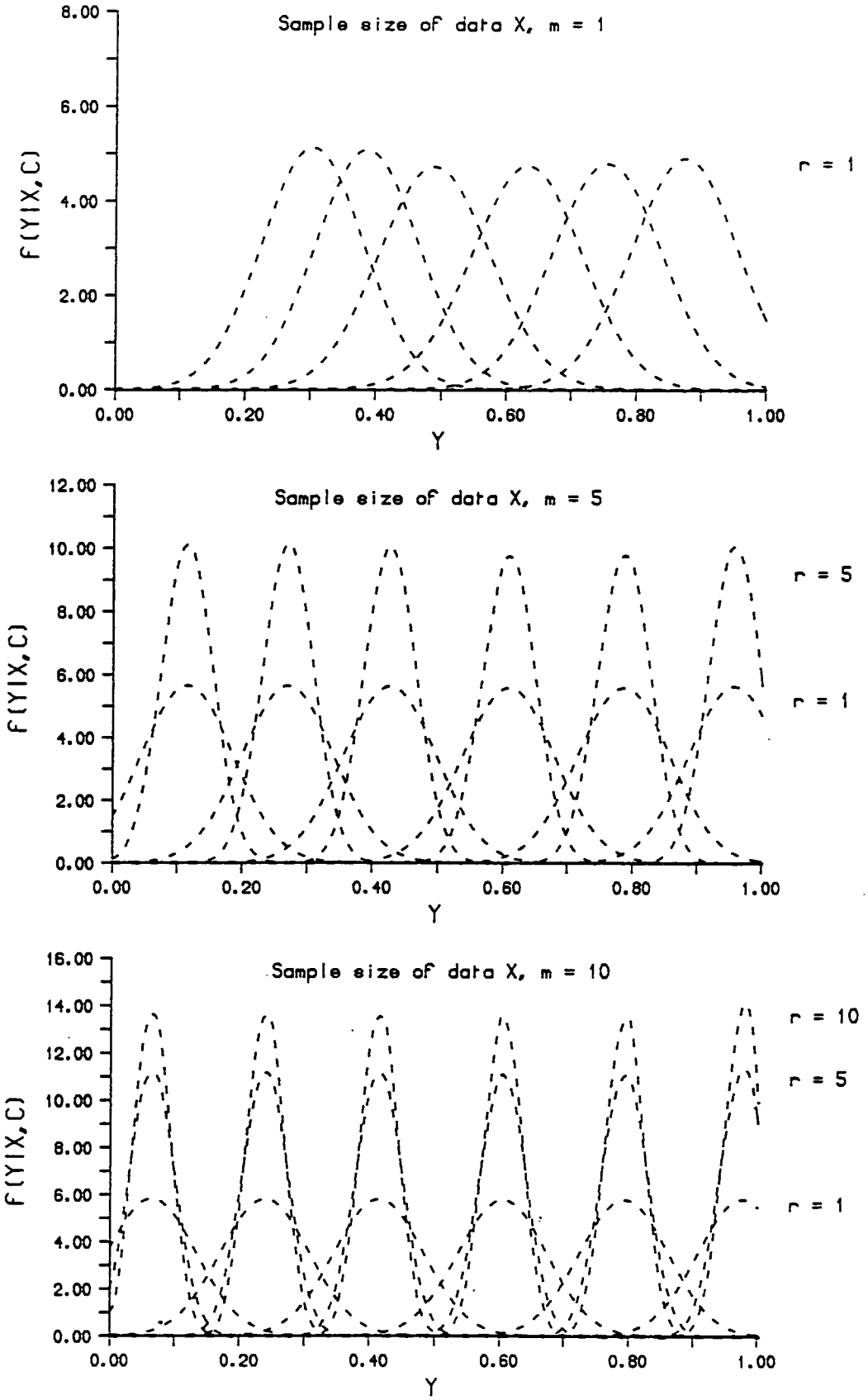


Fig. 3.3 Numerator of the Bayes' Factor, F(Y|X,C) given X = 0.0(0.2)1.0 -- Kernel grouped and known σ^2 model.

two models is when r=m=1. The denominators of the Bayes' factor under the two models (Normal and kernel known σ^2) are shown in Fig 3.4 and 3.5, respectively. In practice, it is preferable that σ^2 be very small in order that there may be good discrimination between two objects or individuals. However, our example is based on a real data set for which $\sigma^2=0.0657^2$ and $\tau^2=0.0659^2$ and are of comparable magnitude. The shrinkage effect may be lessened by increasing the number of control and recovered observations, as illustrated in Table 3.7 when m and r = 10. Since (σ^2/m) will tend to zero as $m \to \infty$ it is not always practical for this to be done.

Graphical representations of the variation in the Bayes' factor as $\bar{X}$ and $\bar{Y}$ vary are shown in Figs. 3.6 and 3.7. Fig. 3.6 illustrates the results from the model developed in Section 3.5.2, representative values of which are given in Tables 3.2 - 3.7. The ranges of $\bar{X}$ and of $\bar{Y}$ as shown in these figures are slightly different from those of the Tables. This emphasises the form of the surface of the BF when $\bar{X}$ and $\bar{Y}$ are both common which would not be illustrated fully if the plotting ranges were extended to rare values. Fig. 3.7 illustrates the case where the within group variance is assumed known and no grouping is assumed.

Figs. 3.8 and 3.9 give a much clearer picture of the behaviour of the Bayes' factor, given the respective training data is grouped or not grouped, for $\bar{X}$ and $\bar{Y}$ ranging from 0.0 to 1.0 with various pairings of r and m. This is achieved by taking the natural logarithm of the Bayes' factor and consequently the enormous value of the BF when $\bar{X}$ and $\bar{Y}$ are uncommon is removed. In Figs 3.8 and 3.9

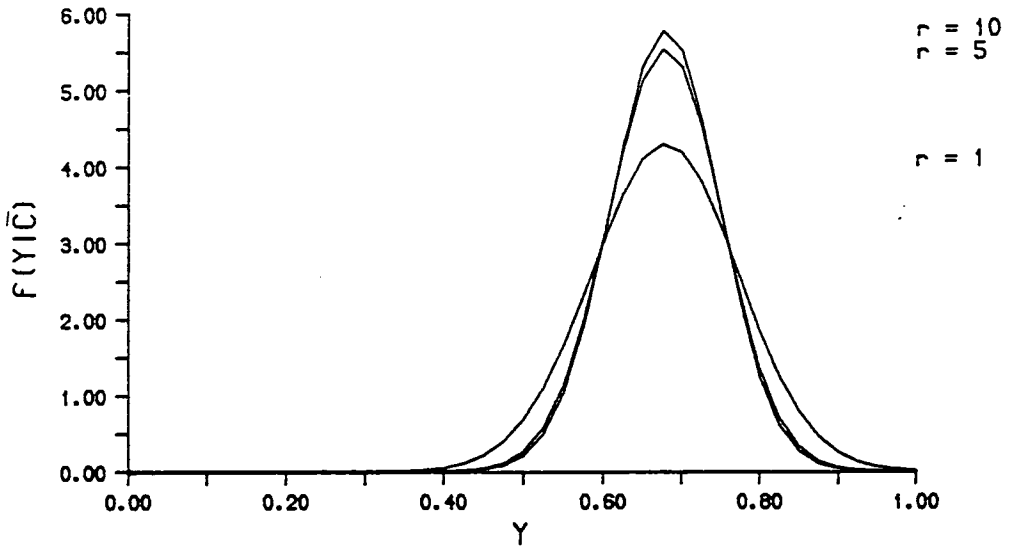


Fig. 3.4 Denominator of the Bayes' factor, $f(Y|\bar{C})$ -- Normal and known σ^2 model.

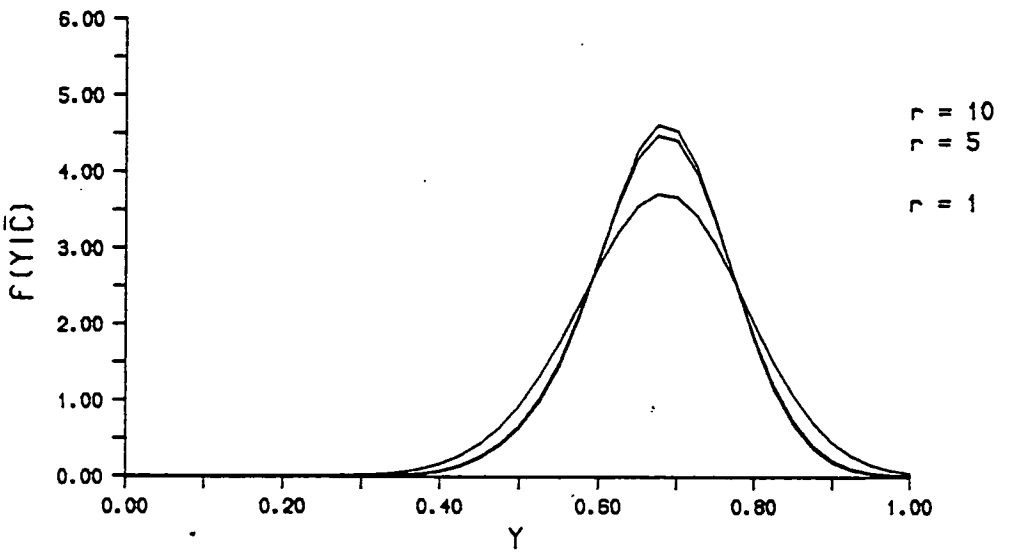


Fig. 3.5 Denominator of the Bayes' factor, $f(Y|\bar{C})$ -- Kernel grouped and known σ^2 model.

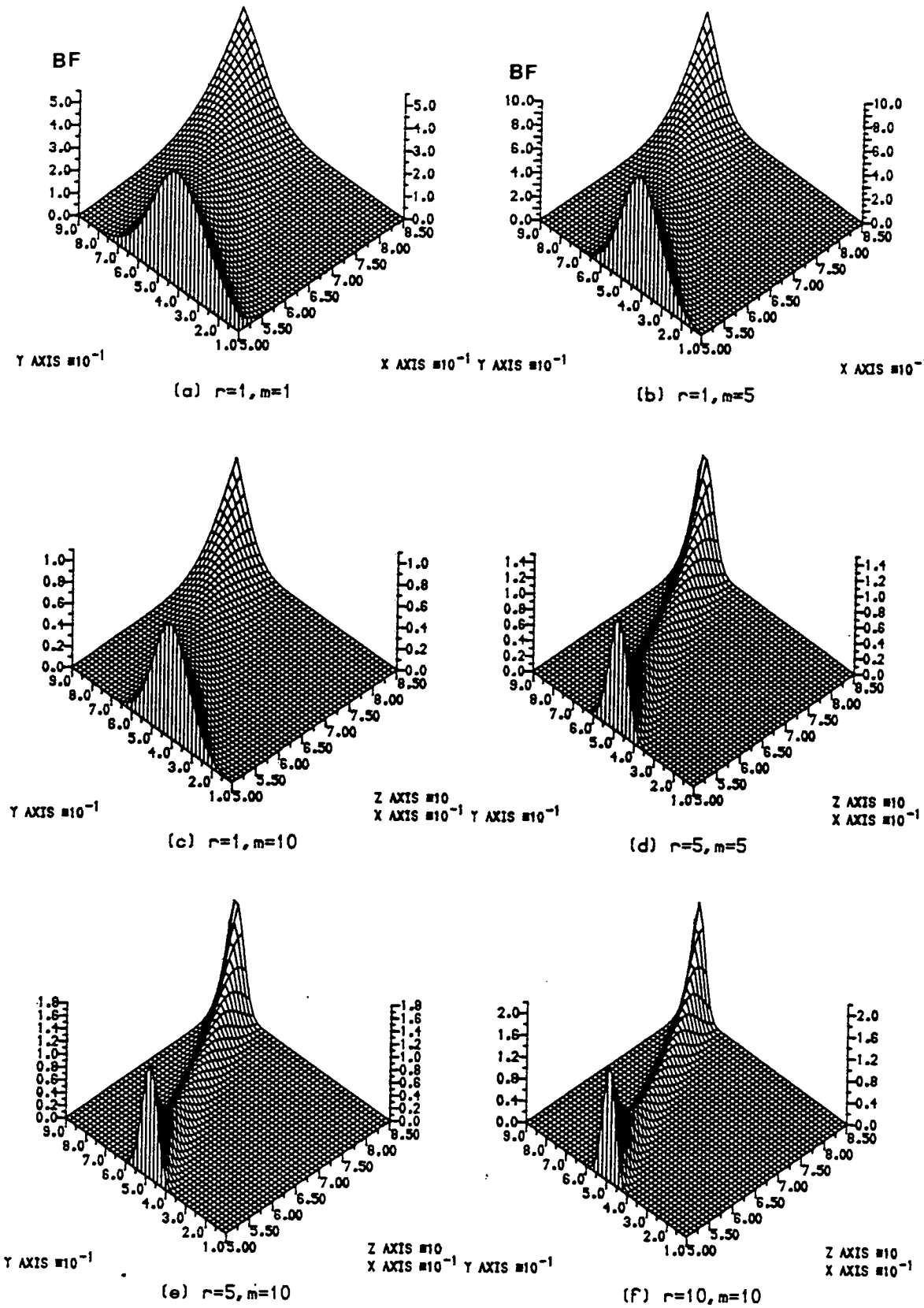


Fig. 3.6 Bivariate plot of the Bayes' factor of $\bar{X}$ in the range between 0.50 and 0.85, $\bar{Y}$ between 0.1 and 0.90 for various pairings of r and m. The within-group variance is assumed known. Grouped data.

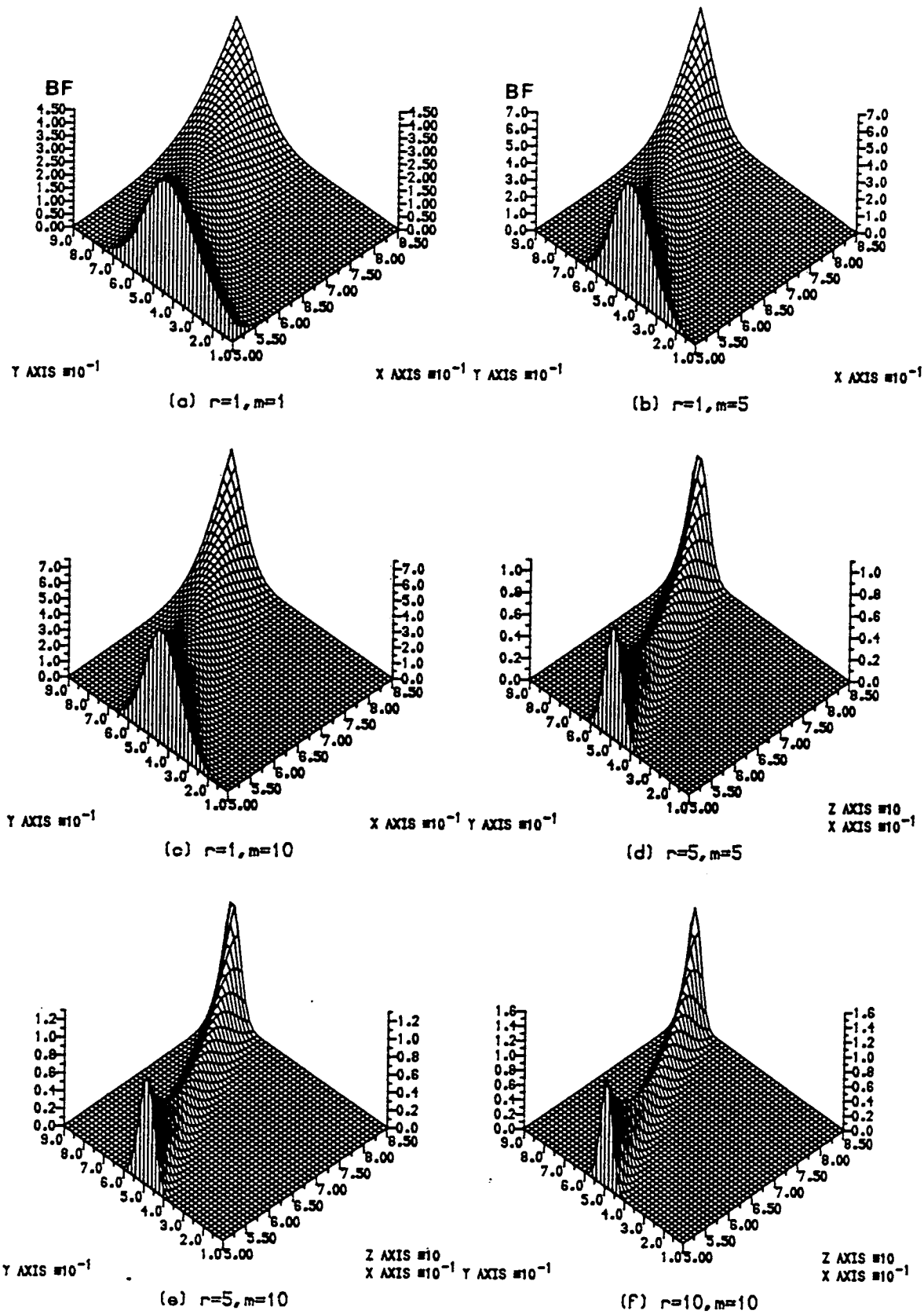


Fig. 3.7 Bivariate plot of the Bayes' factor of $\bar{X}$ in the range between 0.50 and 0.85, $\bar{Y}$ between 0.1 and 0.90 for various pairings of r and m. The within-group variance is assumed known. No grouped data.

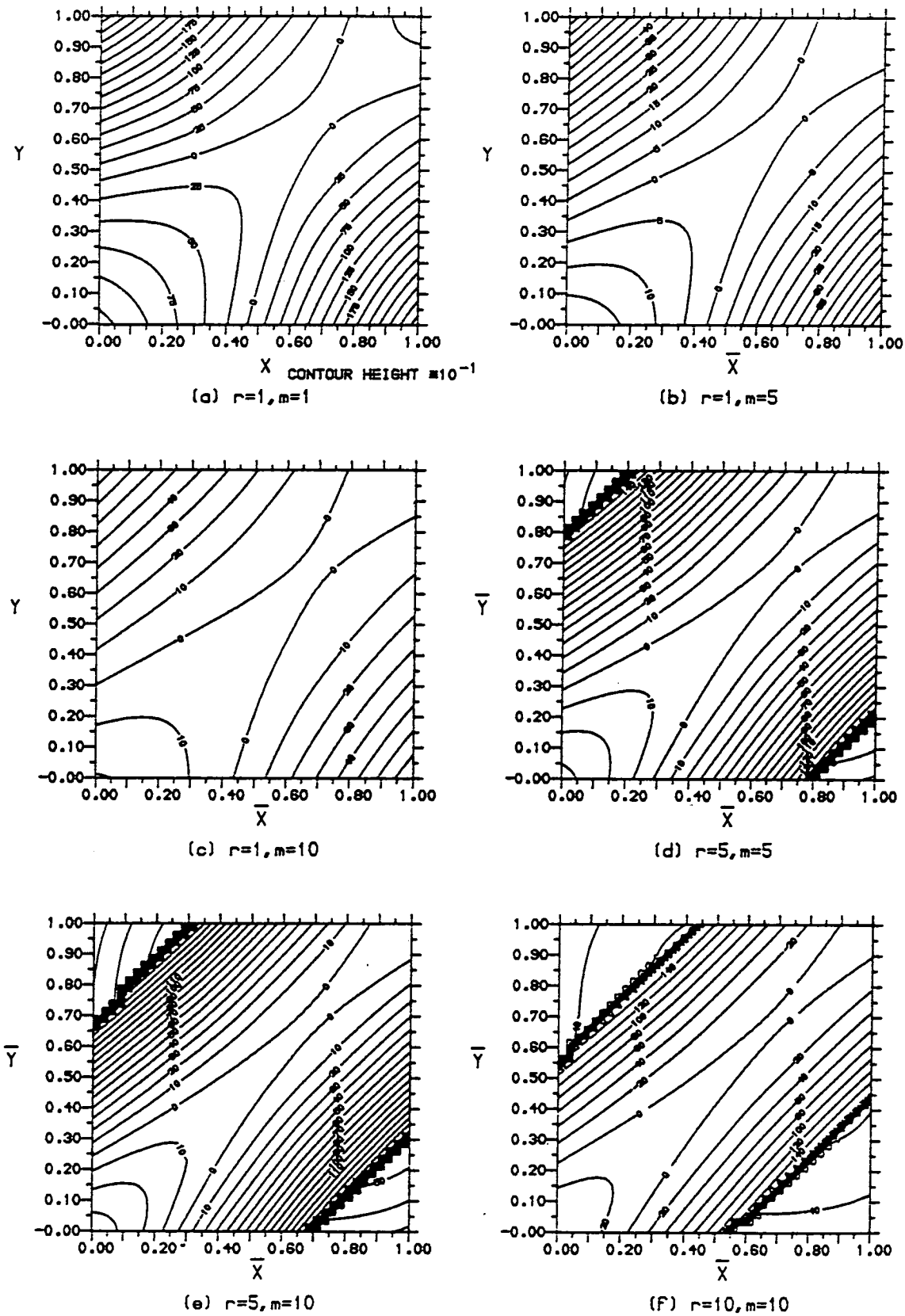


Fig. 3.8 Contours of the logarithm of the Bayes' Factor - Known σ^2 and grouped data.

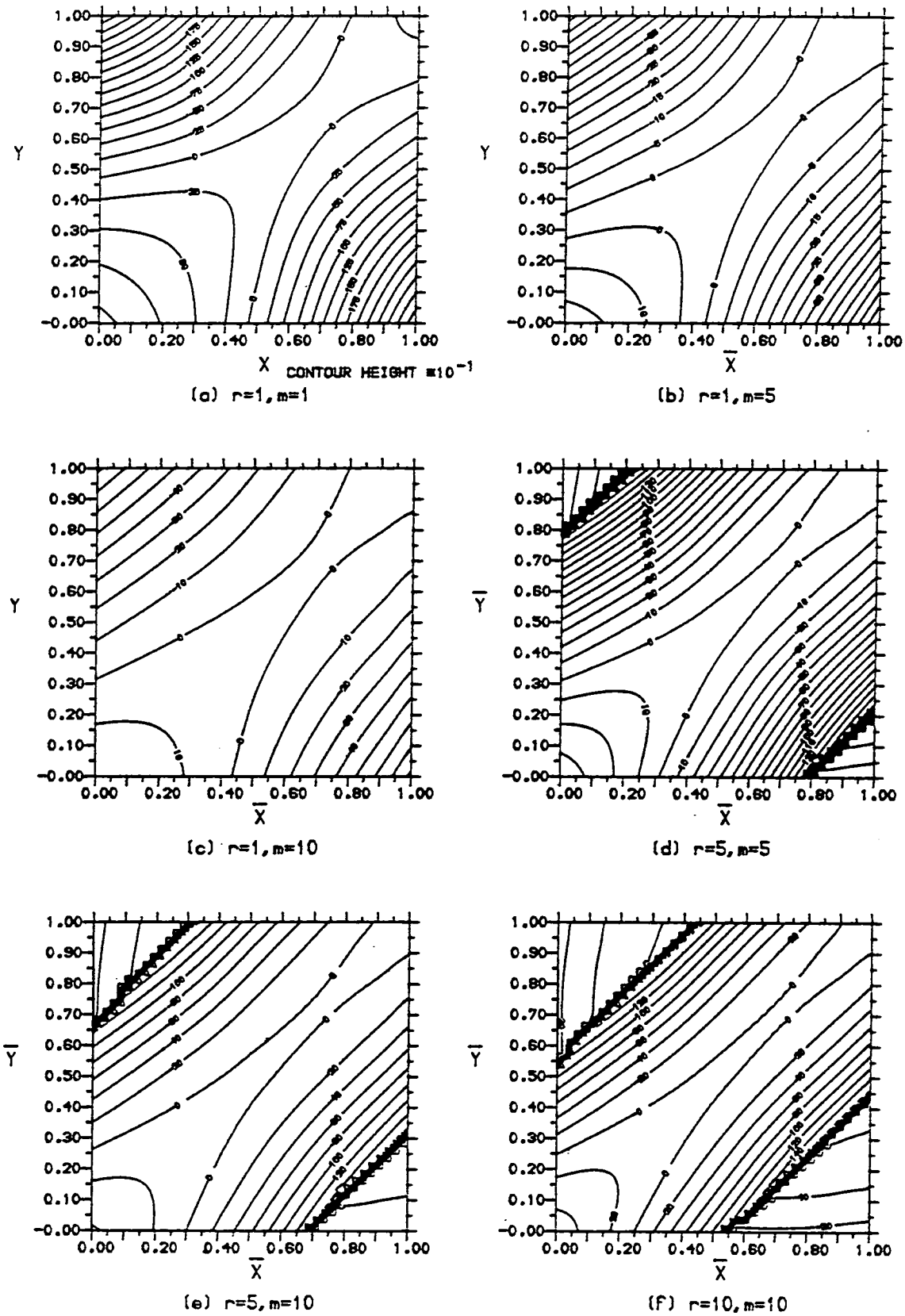


Fig. 3.9 Contours of the logarithm of the Bayes' Factor - Known σ^2 and Nogrouped data.

(d) - (f), a very small number (1.0×10^{-30}) was added to the BF to avoid taking logarithm of zero. The region where the Bayes' factor is greater than one is the area within the contour of value 0. The area where $\bar{X}$ and $\bar{Y}$ indicate a support to the hypothesis C runs diagonally from the bottom left hand to the top right hand corner. This graphical presentation of the logarithm of the BF resembles a saddle shape. Also notice that the area for BF is greater than one which becomes narrower and the value of $\log_e(BF)$ increases as r and m increase.

3.7 Sensitivity analysis of the models derived in Sections 3.5.2 and 3.5.3

We now consider the sensitivity of the Bayes' factor with respect to the smoothing parameter λ, the training data Z and the value of σ^2 assumed known for the kernel group and nogroup models. The group and nogroup models are so called because of the assumption concerning the structure of the training data (see Section 3.2 for the definition). The BF is calculated under the assumption that the distribution of the unknown true mean μ takes a kernel density form and the training data are assumed to be grouped for the group model or ungrouped for the nogroup model. We are interested in how the BF is affected when λ, Z and σ depart from their 'true' values. In other words, how sensitive is the value of the BF to changes in these values? This sensitivity analysis is carried out based on the example given in Section 3.6. Thus the 'true' values of λ, Z and σ are the values stated in Section 3.6.

3.7.1 Sensitivity of the Bayes' factor to changes in values of the smoothing parameter λ

The Bayes's factor is calculated under the kernel group and nogroup models, with the smoothing parameter perturbed by a factor of two which gives 0.5λ, λ and 2λ. The values of natural logarithms of $(BF_{k\lambda}/BF_{\lambda})$ for k = 0.5,2 are plotted in Figs. 3.10 - 3.12 given some fixed value of $\bar{X}$ for r=1,m=1; r=1,m=10 and r=10, m=10 respectively. BF_{λ} is the Bayes's factor calculated using the 'true' values of λ. The 'true' value of the smoothing parameter λ is estimated using the pseudo-maximum likelihood method and its value is found to be 0.7855 and 0.4079 for the kernel group and nogroup model in the example given in Section 3.6. When $\bar{X}$ and $\bar{Y}$ are far apart, greater sensitivities at the extremes become apparent as the sample sizes of the control data (m) and recovered data (r) increase. The value of the BF is most sensitive when λ is reduced by a factor of 2 and r and m are equal to 10. This reflects the smoothness of the kernel prior in (3.10) and (3.11), that is if λ is small the functions of (3.10) and (3.11) are rough and are particularly sensitive to small values of λ.

The sensitivity of BF when small changes in λ are made is studied by adding a small random number, ϵ, onto λ to give λ^*, say. The number of simulations of this sensitivity analysis carried out for the group and nogroup model was 50 and 25 respectively. The reason the number of simulations for the nogroup model is less than the group model is to reduce the amount of computational time involved under the nogroup model. The random numbers were generated from a Normal distribution with mean 0 and standard deviation

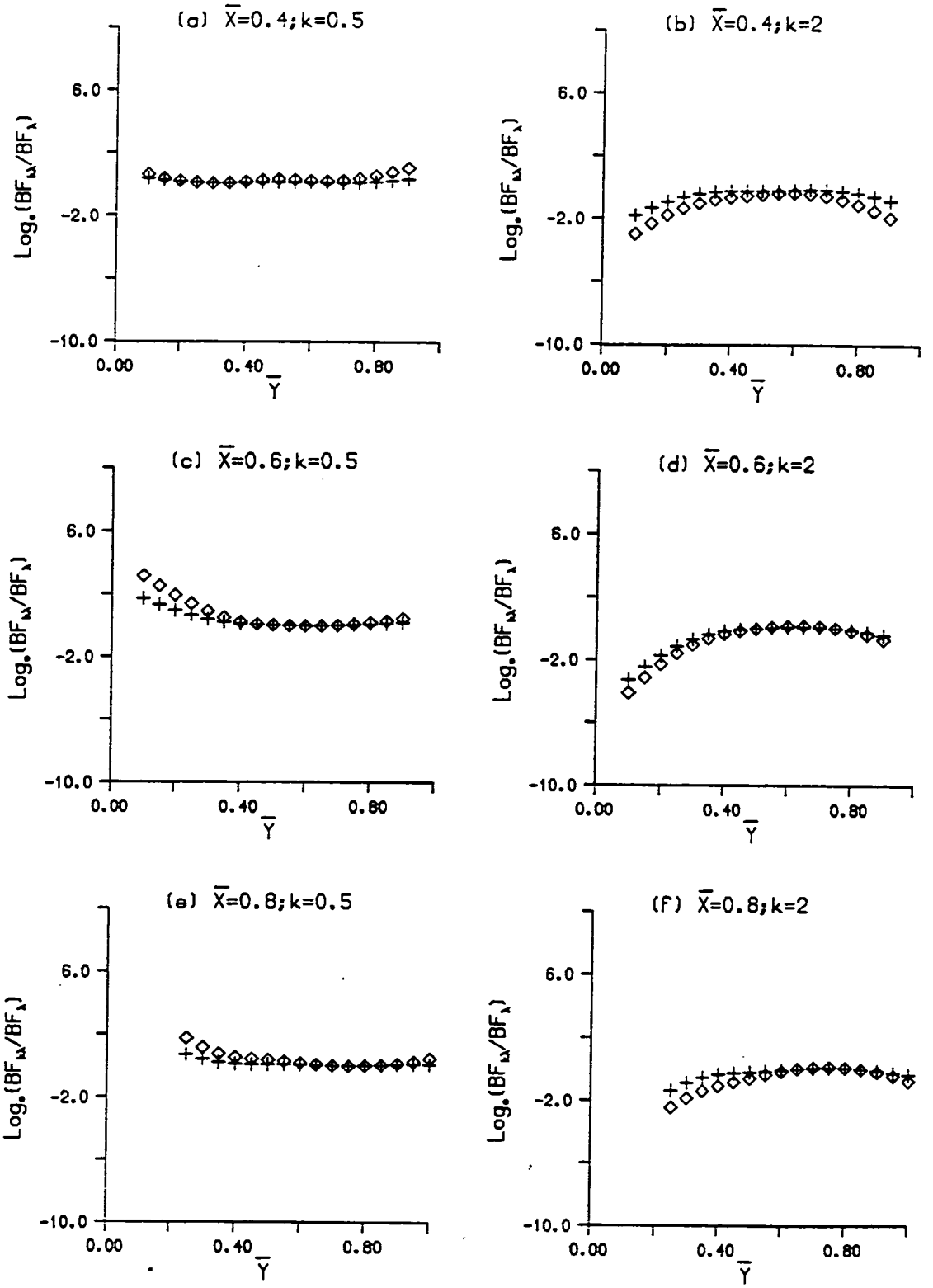


Fig. 3.10 Sensitivity of BF to smoothing parameter λ given some values of $\bar{X}$ when r=1, m=1 assuming the training data are not grouped (+) and grouped (◇).

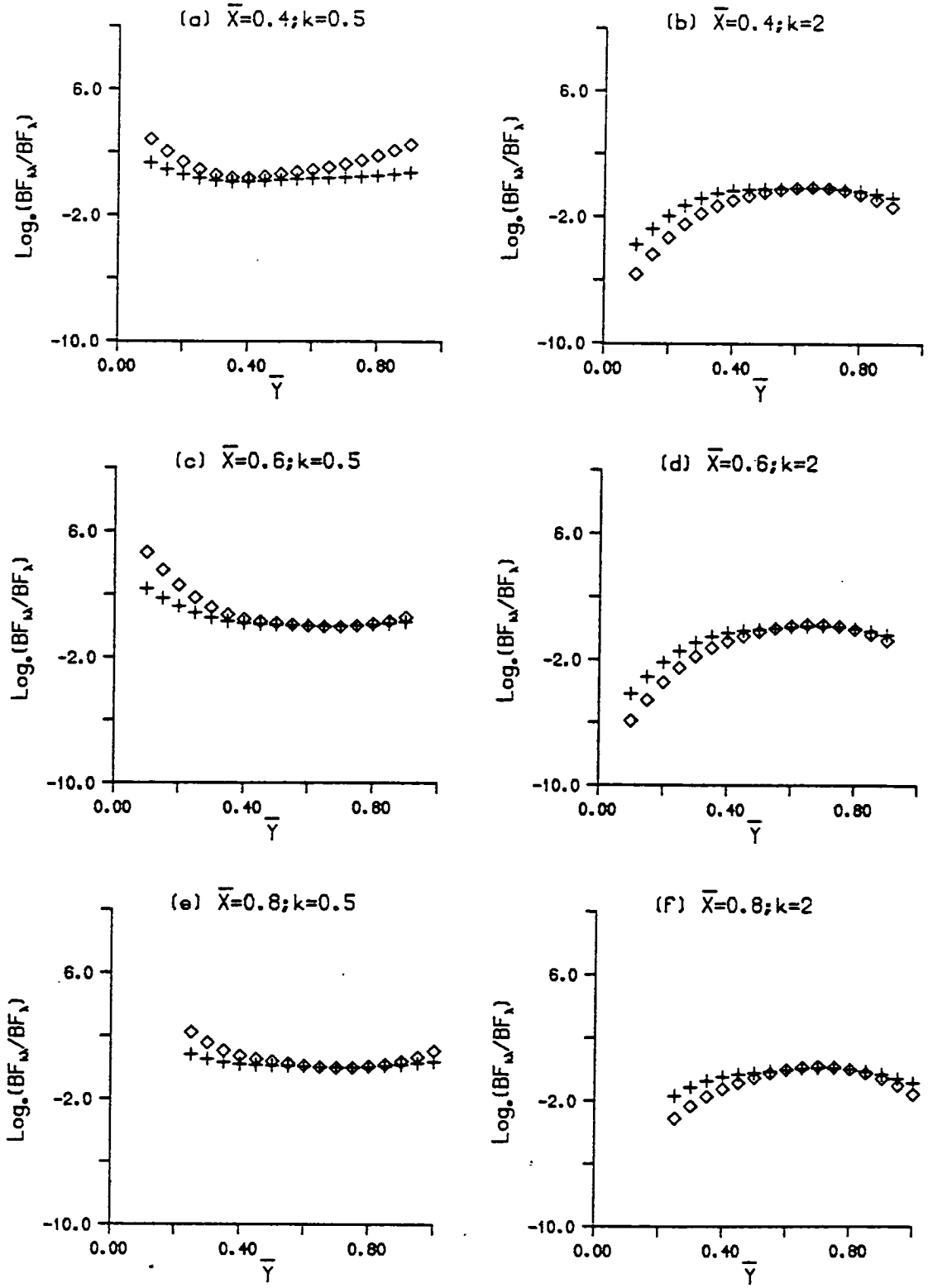


Fig. 3.11 Sensitivity of BF to smoothing parameter λ given some values of $\bar{X}$ when r=1, m=10 assuming the training data are not grouped (+) and grouped (◇).

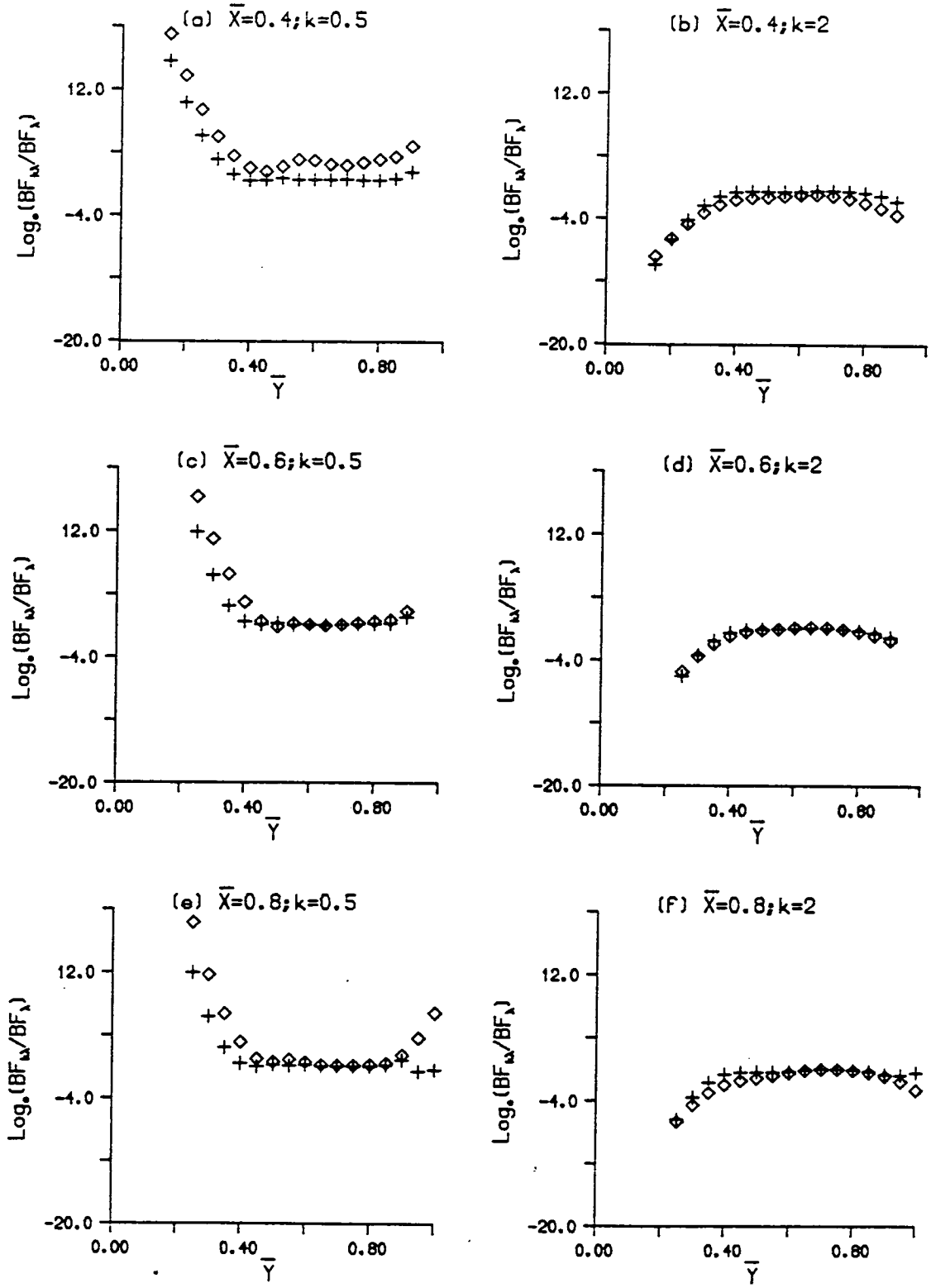


Fig. 3.12 Sensitivity of BF to smoothing parameter λ given some values of $\bar{X}$ when r=10, m=10 assuming the training data are not grouped (+) and grouped (◇).

(0.01×k) where the value of k ranges from 1 to 5 in steps of 2 which help in determining how great the changes in λ is. The larger the value of k the further the value of λ^* is away from the 'ideal' λ. The Bayes' factors evaluated using these contaminated smoothing parameters are denoted by BF(k). Results are shown in Tables 3.14 - 3.16 for r=1,m=1; r=1,m=10 and r=10,m=10 respectively. There are no apparent differences in the values of BF due to small changes in the smoothing parameter for both grouped and nogrouped models.

3.7.2. Sensitivity of the Bayes' factor to changes in the training data Z

The training data $\bar{z}_{i.}$ and z_{ij} for the kernel group and nogrouped model, respectively, are perturbed by adding a small value. Take the group case, for instance

$$\bar{z}_i^* = \bar{z}_{i.} + \gamma_i$$

where $\bar{z}_i^*$ (i=1,2,...,n) is the 'new', or contaminated, training data and γ_i is generated from a Normal distribution with mean 0 and standard deviation 0.01×k, where k takes values from 1(2)5. A large value of k means greater changes in $\bar{z}_{i.}$. We then calculate the Bayes' factor using the $\bar{z}_i^*$ as the training data and repeat the procedure 50 times. The average of the BFs, $\overline{BF(k)}$ over these 50 runs, the standard error of the estimated BFs and values of the $|\log_{10}\{BF/BF(k)\}|$ are shown in Table 3.17 and 3.18 given values of $\bar{X}$ and of $\bar{Y}$ for r=1,m=1 and r=10,m=10 respectively.

The same procedure is also applied to the nogrouped model. That is

$$z_{\ell}^{*} = z_{\ell} + \gamma_{\ell}$$

where γ_{ℓ} is as in the grouped case and $\ell = 1,2,\ldots,N$.

Because of the large sample (N = 220) in the nogrouped model, the average of BF(k)s, for each k is obtained over 25 runs. Results are also tabulated in Tables 3.17 and 3.18, given fixed values of $\bar{X}$ and of $\bar{Y}$, again for r=1,m=1 and r=10,m=10 respectively, and are shown in the second column of the tables. The results are similar under the two models. The values of the Bayes' factor vary slightly as k increases but do not change much overall when r and m are small. When r and m are equal to 10, and the BF becomes more sensitive to changes in the training data Z when $\bar{X}$ and $\bar{Y}$ are least common and far apart. This again happened at the extremes but the BF remains insensitive otherwise.

3.7.3 Sensitivity of the Bayes' factor to changes in the value of σ which is assumed known

As in Section 3.7.2, σ is perturbed by adding a small factor ν. This time ν is chosen to be distributed as $N(0.0,(0.001xk)^2)$ and k is taken to be ranging from 1(2)5. Tables 3.19 and 3.20 show the absolute logarithm to base 10 of the ratio of the standard BF to the contaminated BF(k) when r=1,m=1 and r=10,m=10 respectively. There are no drastic changes in BF(k) as k increases, though greater differences occur when $\bar{X}$ and $\bar{Y}$ are less common and different when r and m equal 10, than when r=m=1.

Table 3.14 Sensitivity of BF to small changes of the smoothing parameter λ for some values of $\bar{X}$ and $\bar{Y}$, R = BF/BF(k) when r=1, m=1.

	Group			Nogroup		
$\bar{X}$=0.4	BF(k)	st.error	$\lvert\log_{10}R\rvert$	BF(k)	st.error	$\lvert\log_{10}R\rvert$
$\bar{Y}$=0.3	28.39	0	-	18.04	0	-
k = 1	28.39	0.0441	.00000	18.04	0.0248	.00000
3	28.34	0.1206	.00077	17.97	0.0623	.00169
5	27.92	0.2162	.00725	18.07	0.1012	.00072
$\bar{Y}$=0.4	16.71	0	-	12.02	0	-
k = 1	16.69	0.0156	.00052	12.00	0.0093	.00072
3	16.79	0.0607	.00207	12.03	0.0315	.00036
5	16.68	0.0965	.00078	11.97	0.0469	.00181
$\bar{Y}$=0.5	5.023	0	-		0	-
k = 1	5.017	0.0054	.00052	4.239	0.0040	.00020
3	5.019	0.0148	.00035	4.219	0.0148	.00226
5	5.010	0.0289	.00113	4.219	0.0234	.00226
$\bar{X}$=0.6						
$\bar{Y}$=0.5	1.520	0	-	1.484	0	-
k = 1	1.520	.00007	.00000	1.484	.00000	.00000
3	1.519	.00019	.00029	1.484	.00002	.00000
5	1.519	.00035	.00029	1.484	.00010	.00000
$\bar{Y}$=0.6	1.601	0	-	1.725	0	-
k = 1	1.601	.00036	.00000	1.724	.00048	.00025
3	1.600	.00098	.00027	1.725	.00123	.00000
5	1.603	.00200	.00504	1.723	.00234	.00050
$\bar{Y}$=0.7	.9284	0	-	.9629	0	-
k = 1	.9285	.00010	.00005	.9629	.00019	.00000
3	.9287	.00032	.00014	.9631	.00060	.00009
5	.9289	.00048	.00023	.9628	.00080	.00005
$\bar{X}$=0.8						
$\bar{Y}$=0.7	1.056	0	-	1.069	0	-
k = 1	1.055	.00007	.00041	1.069	.00011	.00000
3	1.056	.00016	.00000	1.068	.00041	.00041
5	1.056	.00046	.00000	1.068	.00066	.00041
$\bar{Y}$=0.8	2.065	0	-	2.030	0	-
k = 1	2.065	.00021	.00000	2.030	.00033	.00000
3	2.064	.00056	.00021	2.032	.00067	.00043
5	2.064	.00105	.00021	2.030	.00104	.00000
$\bar{Y}$=0.9	2.361	0	-	2.200	0	-
k = 1	2.361	.00087	.00000	2.200	.00112	.00000
3	2.354	.00253	.00129	2.196	.00418	.00079
5	2.371	.00515	.00184	2.198	.00648	.00039

Table 3.15 Sensitivity of BF to small changes of the smoothing parameter λ for some values of X and Y, R = BF/BF(k) when r=1, m=10.

	Group			Nogroup		
$\bar{X}$=0.4	BF(k)	st.error	$\lvert\log_{10}R\rvert$	BF(k)	st.error	$\lvert\log_{10}R\rvert$
$\bar{Y}$=0.3	116.1	0	-	53.97	0	-
k = 1	116.2	0.3866	.00037	54.14	0.1950	.00137
3	115.6	1.2775	.00187	53.22	0.4826	.00608
5	117.9	1.9485	.00668	54.34	0.7858	.00297
$\bar{Y}$=0.4	35.23	0	-	21.10	0	-
k = 1	35.35	0.0636	.00148	21.09	0.0258	.00021
3	35.53	0.1757	.00368	21.12	0.0821	.00041
5	35.24	0.3025	.00012	21.23	0.1422	.00267
$\bar{Y}$=0.5	2.884	0	-	2.433	0	-
k = 1	2.888	0.0039	.00060	2.428	0.0026	.00089
3	2.875	0.0104	.00136	2.419	0.0140	.00251
5	2.890	0.0215	.00090	2.429	0.0176	.00071
$\bar{X}$=0.6						
$\bar{Y}$=0.5	1.946	0	-	1.737	0	-
k = 1	1.945	.00116	.00022	1.737	.00055	.00000
3	1.949	.00329	.00067	1.738	.00176	.00025
5	1.940	.00533	.00134	1.735	.00353	.00050
$\bar{Y}$=0.6	2.101	0	-	2.308	0	-
k = 1	2.101	.00029	.00000	2.308	.00029	.00000
3	2.104	.00107	.00062	2.308	.00091	.00000
5	2.103	.00181	.00041	2.308	.00137	.00000
$\bar{Y}$=0.7	.6082	0	-	.6455		
k = 1	.6085	.00018	.00021	.6456	.00027	.00007
3	.6085	.00062	.00021	.6457	.00075	.00013
5	.6085	.00090	.00021	.6466	.00130	.00074
$\bar{X}$=0.8						
$\bar{Y}$=0.7	.6330	0	-	.6654	0	-
k = 1	.6327	.00016	.00021	.6651	.00026	.00020
3	.6340	.00058	.00069	.6659	.00068	.00033
5	.6328	.00098	.00014	.6660	.00101	.00039
$\bar{Y}$=0.8	2.900	0	-	2.729	0	-
k = 1	2.900	.00028	.00000	2.729	.00020	.00000
3	2.900	.00112	.00000	2.729	.00051	.00000
5	2.902	.00182	.00030	2.730	.00073	.00016
$\bar{Y}$=0.9	3.902	0	-	3.232	0	-
k = 1	3.902	.00588	.00000	3.234	.00414	.00027
3	3.892	.01398	.00111	3.227	.01062	.00067
5	3.880	.02494	.00246	3.239	.02600	.00094

Table 3.16 Sensitivity of BF to small changes of the smoothing parameter λ for some values of X and Y. R = BF/BF(k) when r=10 & m=10.

	Group			Nogroup		
$\bar{X}$=0.4	BF(k)	st.error	$\lvert\log_{10}R\rvert$	BF(k)	st.error	$\lvert\log_{10}R\rvert$
$\bar{Y}$=0.3	9.006	0	-	5.083	0	-
k = 1	8.950	0.1007	.00271	5.084	0.1070	.00009
3	8.936	0.3016	.00339	5.124	0.2500	.00349
5	10.31	0.8377	.05873	5.545	0.2299	.03778
$\bar{Y}$=0.4	215.5	0	-	87.38	0	-
k = 1	217.1	0.8900	.00321	87.15	0.3755	.00114
3	215.3	2.7473	.00040	87.01	0.8041	.00184
5	219.3	5.5148	.00759	87.51	1.5255	.00065
$\bar{Y}$=0.5	.2581	0	-	.1375		
k = 1	.2586	0.0011	.00084	.1370	.00045	.00158
3	.2604	0.0030	.00385	.1387	.00138	.00377
5	.2704	0.0067	.02022	.1402	.00303	.00845
$\bar{X}$=0.6						
$\bar{Y}$=0.5	.0323	0	-	.0289	0	-
k = 1	.0323	.00001	.00000	.0289	.00004	.00000
3	.0322	.00004	.00135	.0290	.00012	.00150
5	.0322	.00005	.00135	.0287	.00011	.00320
$\bar{Y}$=0.6	4.778	0	-	5.514	0	-
k = 1	4.778	.00026	.00000	5.516	.00217	.00016
3	4.780	.00098	.00018	5.520	.00668	.00047
5	4.781	.00152	.00027	5.515	.01572	.00008
$\bar{Y}$=0.7	.0139	0	-	.0151		
k = 1	.0139	.00000	.00000	.0151	.00001	.00000
3	.0139	.00000	.00000	.0151	.00002	.00000
5	.0139	.00000	.00000	.0152	.00004	.00287
$\bar{X}$=0.8						
$\bar{Y}$=0.7	.0173	0	-	.0163	0	-
k = 1	.0173	.00000	.00000	.0163	.00000	.00000
3	.0173	.00001	.00000	.0163	.00001	.00000
5	.0173	.00001	.00000	.0163	.00001	.00000
$\bar{Y}$=0.8	7.593	0	-	6.443	0	-
k = 1	7.581	.00381	.00069	6.443	.00075	.00000
3	7.610	.01378	.00097	6.444	.00223	.00007
5	7.602	.02577	.00051	6.442	.00368	.00007
$\bar{Y}$=0.9	.0848	0	-	.0751	0	-
k = 1	.0846	.00022	.00103	.0751	.00034	.00000
3	.0846	.00066	.00103	.0787	.00169	.02033
5	.0879	.00133	.01559	.0778	.00284	.01534

Table 3.17 Sensitivity of BF to small changes of the training data Z for some values of X and Y, R = BF/BF(k) when r=1,m=1.

TS	Group			Nogroup		
$\bar{X}$=0.4	BF(k)	st.error	$\lvert\log_{10}R\rvert$	BF(k)	st.error	$\lvert\log_{10}R\rvert$
$\bar{Y}$=0.3	28.39	0	-	18.04	0	-
k = 1	28.17	0.2949	.00338	17.87	0.1000	.00411
3	23.79	0.6822	.07677	16.19	0.1304	.04699
5	21.54	1.1438	.11992	14.20	0.3609	.01039
$\bar{Y}$=0.4	16.71	0	-	12.02	0	-
k = 1	16.15	0.1196	.01480	11.97	0.0571	.00181
3	15.22	0.2808	.04056	11.29	0.1245	.02721
5	13.00	0.3902	.10903	10.39	0.1631	.06329
$\bar{Y}$=0.5	5.023	0	-	4.241	0	-
k = 1	4.981	0.0306	.00365	4.206	0.0146	.00360
3	4.761	0.0686	.02326	3.952	0.0330	.03065
5	4.462	0.1006	.05143	3.662	0.0356	.06375
$\bar{X}$=0.6						
$\bar{Y}$=0.5	1.520	0	-	1.484	0	-
k = 1	1.523	.00336	.00086	1.485	.00180	.00029
3	1.496	.01057	.00691	1.486	.00622	.00058
5	1.508	.01906	.00344	1.490	.00768	.00175
$\bar{Y}$=0.6	1.601	0	-	1.725	0	-
k = 1	1.599	.00334	.00054	1.725	.00208	.00000
3	1.617	.00980	.00432	1.760	.00403	.00872
5	1.690	.01634	.02350	1.814	.00966	.02185
$\bar{Y}$=0.7	.9284	0	-	.9629	0	-
k = 1	.9287	.00091	.00014	.9650	.00078	.00095
3	.9389	.00263	.00488	.9739	.00187	.00493
5	.9620	.00443	.01544	.9930	.00308	.01337
$\bar{X}$=0.8						
$\bar{Y}$=0.7	1.056	0	-	1.069	0	-
k = 1	1.057	.00148	.00041	1.070	.00074	.00041
3	1.068	.00598	.00491	1.076	.00249	.00283
5	1.071	.00679	.00613	1.090	.00386	.00845
$\bar{Y}$=0.8	2.065	0	-	2.030	0	-
k = 1	2.069	.00504	.00084	2.028	.00209	.00043
3	2.064	.01587	.00021	2.055	.00834	.00532
5	2.089	.02341	.00502	2.114	.00753	.01761
$\bar{Y}$=0.9	2.361	0	-	2.200	0	-
k = 1	2.367	.00969	.00110	2.192	.00392	.00158
3	2.299	.02483	.01156	2.140	.01045	.01201
5	2.227	.04109	.02538	2.027	.01344	.03557

Table 3.18 Sensitivity of BF of small changes of the training data Z for some values of $\bar{X}$ and $\bar{Y}$. R = BF/BF(k) when r=10, m=10.

TS	Group			Nogroup		
$\bar{X}$=0.4	BF	st.error	$\lvert\log_{10}R\rvert$	BF	st.error	$\lvert\log_{10}R\rvert$
$\bar{Y}$=0.3	9.006	0	-	5.083	0	-
k = 1	11.01	0.7450	.08726	4.704	0.2385	.03365
3	16.30	3.4516	.25766	3.029	0.2952	.48130
5	12.28	3.5568	.13467	1.621	0.2339	.49634
$\bar{Y}$=0.4	215.5	0	-	87.38	0	-
k = 1	233.9	11.735	.03558	85.08	1.2911	.01158
3	212.4	23.064	.00629	76.18	2.7647	.05957
5	300.0	56.806	.14367	65.75	2.9744	.12352
$\bar{Y}$=0.5	.2581	0	-	.1375	0	-
k = 1	.2602	0.0070	.00352	.1335	.00253	.01282
3	.2333	0.0148	.04387	.1191	.00381	.06239
5	.1865	0.0198	.01411	.0968	.00407	.01524
$\bar{X}$=0.6						
$\bar{Y}$=0.5	.0323	0	-	.0289	0	-
k = 1	.0324	.00017	.00134	.0285	.00018	.00605
3	.0312	.00065	.01505	.0283	.00037	.00911
5	.0302	.00072	.02920	.0275	.00042	.02157
$\bar{Y}$=0.6	4.778	0	-	5.514	0	-
k = 1	4.758	.02126	.00182	5.530	.01558	.00126
3	4.818	.06263	.00362	5.468	.05558	.00364
5	4.900	.10757	.01095	5.347	.05002	.01336
$\bar{Y}$=0.7	.0139	0	-	.0151	0	-
k = 1	.0139	.00003	.00000	.0151	.00004	.00000
3	.0141	.00008	.00620	.0151	.00009	.00000
5	.0144	.00016	.01535	.0151	.00013	.00000
$\bar{X}$=0.8						
$\bar{Y}$=0.7	.0173	0	-	.0163	0	-
k = 1	.0173	.00006	.00000	.0163	.00006	.00000
3	.0171	.00014	.00505	.0166	.00016	.00792
5	.0173	.00021	.00000	.0169	.00015	.01570
$\bar{Y}$=0.8	7.593	0	-	6.443	0	-
k = 1	7.634	0.0508	.00234	6.460	.02107	.00114
3	7.208	0.1075	.02260	6.393	.06242	.00338
5	7.226	0.1809	.02152	6.418	.08253	.00169
$\bar{Y}$=0.9	.0848	0	-	.0751	0	-
k = 1	.0839	.00127	.00463	.0724	.00073	.01590
3	.0824	.00271	.01247	.0565	.00103	.12359
5	.0729	.00523	.06567	.0461	.00104	.21193

Table 3.19 Sensitivity of BF to small changes of the known σ for some values of X and Y, R = BF/BF(k) when r=1, m=1.

TS	Group			Nogroup		
$\bar{X}$=0.4	BF(k)	st.error	$\lvert\log_{10}R\rvert$	BF(k)	st.error	$\lvert\log_{10}R\rvert$
$\bar{Y}$=0.3	28.39	0	-	18.04	0	-
k = 1	28.55	0.1499	.00244	17.99	0.1523	.00121
3	29.26	0.5461	.01311	18.18	0.3463	.00336
5	28.68	0.8379	.00441	18.55	0.7334	.01211
$\bar{Y}$=0.4	16.71	0	-	12.02	0	-
k = 1	16.71	0.1179	.00000	12.08	0.0867	.00216
3	17.06	0.3425	.00900	12.56	0.3005	.01909
5	17.50	0.5118	.02006	12.19	0.2847	.00610
$\bar{Y}$=0.5	5.023	0	-	4.241	0	-
k = 1	5.025	0.0180	.00017	4.257	0.0107	.00164
3	4.992	0.0470	.00269	4.226	0.0346	.00154
5	4.939	0.0662	.00732	4.104	0.0592	.01426
$\bar{X}$=0.6						
$\bar{Y}$=0.5	1.520	0	-	1.484	0	-
k = 1	1.520	.00031	.00000	1.484	.00015	.00000
3	1.517	.00110	.00086	1.478	.00260	.00176
5	1.511	.00178	.00258	1.478	.00258	.00176
$\bar{Y}$=0.6	1.601	0	-	1.725	0	-
k = 1	1.600	.00253	.00027	1.731	.00329	.00151
3	1.597	.00839	.00109	1.751	.01749	.00650
5	1.604	.01324	.00081	1.723	.01618	.00050
$\bar{Y}$=0.7	.9284	0	-	.9629	0	-
k = 1	.9286	.00096	.00009	.9622	.00138	.00032
3	.9276	.00283	.00037	.9618	.00346	.00050
5	.9177	.00450	.00503	.9503	.00611	.00572
$\bar{X}$=0.8						
$\bar{Y}$=0.7	1.056	0	-	1.069	0	-
k = 1	1.055	.00065	.00041	1.066	.00155	.00122
3	1.053	.00179	.00124	1.064	.00380	.00204
5	1.045	.00427	.00455	1.073	.00446	.00162
$\bar{Y}$=0.8	2.065	0	-	2.030	0	-
k = 1	2.072	.00429	.00147	2.026	.00462	.00086
3	2.068	.01445	.00063	2.073	.01747	.00910
5	2.048	.02544	.00359	2.031	.03167	.00021
$\bar{Y}$=0.9	2.391	0	-	2.200	0	-
k = 1	2.364	.00308	.00493	2.197	.00320	.00059
3	2.350	.00980	.00751	2.191	.00906	.00178
5	2.357	.01728	.00622	2.184	.01659	.00317

Table 3.20 Sensitivity of BF to small changes of the known σ for some values of X and Y. R = BF/BF(k) when r=10, m=10.

TS	Group			Nogroup		
$\bar{X}$=0.4	BF	st.error	$\lvert\log_{10}R\rvert$	BF	st.error	$\lvert\log_{10}R\rvert$
$\bar{Y}$=0.3	9.006	0	-	5.083	0	-
k = 1	9.276	0.1523	.01283	5.165	0.1347	.00695
3	9.772	0.5761	.03545	4.983	0.3959	.00863
5	11.62	0.8670	.11067	5.124	0.6323	.00349
$\bar{Y}$=0.4	215.5	0	-	87.38	0	-
k = 1	215.2	0.7211	.00061	87.83	0.4161	.00223
3	217.4	2.2657	.00381	86.66	1.1537	.00359
5	221.9	4.2531	.01271	87.31	1.6543	.00348
$\bar{Y}$=0.5	.2581	0	-	.1375	0	-
k = 1	.2713	.00591	.02166	.1335	.00449	.01282
3	.2851	.01989	.04321	.1460	.01358	.02605
5	.2667	.03120	.01424	.1549	.01965	.05175
$\bar{X}$=0.6						
$\bar{Y}$=0.5	.0323	0	-	.0289	0	-
k = 1	.0314	.00061	.01227	.0281	.00091	.01219
3	.0365	.00233	.05309	.0319	.00305	.02340
5	.0375	.00333	.06483	.0308	.00453	.02765
$\bar{Y}$=0.6	4.778	0	-	5.514	0	-
k = 1	4.768	.01067	.00091	5.515	.01807	.00008
3	4.730	.02781	.00439	5.512	.04967	.00016
5	4.891	.06945	.01015	5.438	.08716	.00603
$\bar{Y}$=0.7	.0139	0	-	.0151	0	-
k = 1	.0138	.00034	.00314	.0150	.00063	.00289
3	.0158	.00114	.05564	.0160	.00153	.02514
5	.0150	.00142	.03308	.0173	.00273	.05907
$\bar{X}$=0.8						
$\bar{Y}$=0.7	.0173	0	-	.0163	0	-
k = 1	.0184	.00039	.02677	.0155	.00056	.02186
3	.0192	.00114	.04526	.0145	.00198	.05082
5	.0241	.00200	.14397	.0188	.00333	.06197
$\bar{Y}$=0.8	7.593	0	-	6.443	0	-
k = 1	7.607	.01863	.00080	6.446	.01948	.00020
3	7.655	.05482	.00353	6.459	.05135	.00108
5	7.653	.10534	.00342	6.657	.09201	.01419
$\bar{Y}$=0.9	.0848	0	-	.0751	0	-
k = 1	.0854	.00194	.00306	.0752	.00155	.00058
3	.1041	.00645	.08905	.0775	.00669	.01366
5	.0965	.00895	.05613	.0839	.01568	.04812

3.8 Simulation Studies

3.8.1 Comparison between kernel and Normal priors

This simulation study is designed to compare the Normal prior and kernel prior with respect to the performance of the model developed in Section 3.5.2. Lindley (1977) stated that the assumption of a Normal distribution for the group means when dealing with data concerning the refractive index of glass was unsatisfactory and replaced the Normal distribution with a Taylor series expansion. The aim is to compare the Lindley model with the kernel model when the grouped means are known to be non-Normal.

Group means were generated from each of the following distributions which were used as representatives of the between group distributions. The appropriate probability density function $f(\mu)$ and the number (n) of group means used in the simulation is also given.

(1). Normal distribution:

$$f(\mu) = (2\pi)^{-1/2} \exp(-\mu^2/2)$$

n = 300.

(2). Gamma distribution:

$$f(\mu) = \beta^{\alpha} \mu^{\alpha-1} \exp\{-\beta\mu\}/\Gamma(\alpha) \qquad \alpha > 0,\ \beta > 0,\ \mu > 0$$

with $\alpha = 2$, $\beta = 1$; n = 100.

(3). Uniform distribution:

$$f(\mu) = \begin{cases} 1/6 & -3 < \mu < 3, \\ 0 & \text{elsewhere}, \end{cases}$$

n = 500.

(4). Mixture of two Normal distributions:

$$f(\mu) = \frac{0.5}{(2\pi)^{1/2}} \left[\exp\left\{ -\frac{(\mu+1.5)^2}{2} \right\} + (0.33)^{-1} \exp\left\{ -\frac{(\mu-1.5)^2}{2\times 0.33^2} \right\} \right]$$

n = 300; (150 from each of the two component distributions).

Notice that the methods are independent of the number of observations J in the groups in the training set.

For each of the distributions the Bayes' factor was calculated under each of the three assumptions for the between group distributions.

(a) It is Normal.

(b) It is unknown.

(c) It is the true distribution.

Note that for (a) and (b) the parameters of the distributions are estimated from the training data. The true, known, values are not used.

Let the simulated data be denoted by $(\bar{z}_1,\ldots,\bar{z}_n)$ where n = 300, 100, 500 and 300 for each of (1), (2), (3) and (4), respectively. For the estimation of the Bayes' factor in (a) an estimate of the between group variance is required. This estimate is taken to be

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{z}_i - \bar{z}_.)^2,$$

rather than the usual analysis of variance estimate of the between group variance. This is because only group means were generated in the simulation study, not actual observations. The within group variance σ^2 is taken to be 1.0..

Under the assumption (a) that the group means were distributed Normally with mean $\bar{z}$ and variance s^2, the estimate of the numerator of (3.4) is

$$\frac{1}{a\sigma(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{(\bar{x}-\bar{y})^2}{2\sigma^2 a^2}\right\} \times$$

$$\frac{1}{[(2\pi)\{s^2+\sigma^2/(r+m)\}]^{\frac{1}{2}}} \exp\left\{-\frac{(w-\bar{z})^2}{2[s^2+\sigma^2/(r+m)]}\right\}$$

where a and w are as defined after (3.13) and σ^2 may be taken to be 1. Similarly the estimate of the denominator of (3.4) is given by

$$\frac{1}{[(2\pi)\{s^2+\sigma^2/m\}]^{\frac{1}{2}}} \exp\left\{-\frac{(\bar{x}-\bar{z})^2}{2(s^2+\sigma^2/m)}\right\} \times$$

$$\frac{1}{[(2\pi)\{s^2+\sigma^2/r\}]^{\frac{1}{2}}} \exp\left\{-\frac{(\bar{y}-\bar{z})^2}{2(s^2+\sigma^2/r)}\right\}.$$

The purpose of the study was to compare the performance of the kernel based estimate assuming grouping and known within group variance of the Bayes' factor with that of the estimate obtained assuming a Normal distribution for the random effects. The "correct" result was taken to be the Bayes' factor obtained from assumption (c) where the true, or, standard distribution was used. The Bayes' factor based on this standard distribution is denoted BF_S. Note that numerical integration is required for distribution (2) when BF is evaluated.

Comparison between the kernel and Normal based estimates of the Bayes' factor is made using a statistic, which is called the maximum

absolute $\log_e$ ratio statistic (MALR), defined by

$$\text{MALR} = \max_{(x,y)} \left| \log_e \left[\frac{BF_A}{BF_S} \right] \right|$$

for all (x,y) in the interval $[\{\min_i(\bar{z}_i)-s, \max_i(\bar{z}_i)+s\},\ i=1,2,\ldots,n]$ and BF_A denotes the Bayes' factor based on the kernel or Normal model. Values of MALR close to zero are good in the sense that the assumed model is close to the true model. The results of the simulation study are given in Table 3.21. The kernel method does better than the Normal model when they are compared with the non-Normal and the kernel method is comparable with the Normal model when the Normal model is the correct one.

Table 3.21 Comparison between Normal and kernel models using maximum absolute $\log_e$ ratio (MALR) statistics in the case where r=10, m=10 and $\sigma^2=1$.

	Assumed Model	
Standard True Model	Kernel (3.5.2)	Normal
N(0,1)	2.2765	1.0131
Ga(2,1)	1.1267	4.4906
Un(-3,3)	0.9251	2.5201
½{N(-1.5,1)+ N(1.5,0.33²)}	2.9715	12.1601

3.8.2 Aspects of comparison between models in terms of error probabilities

The idea of this experiment is to estimate probabilities of Type

I and Type II errors. These are the probabilities of rejecting the null hypothesis, (i.e. the control and recovered data are of the same source) given it is true and of not rejecting the null hypothesis given it is not true, respectively. A test set is generated and the Bayes' factor is calculated based on these data to obtain an estimate of the probabilities over 20 and 50 runs of 100 simulations under the known σ^2 kernel group, nogroup and normal models, see Sections 3.5.2, 3.5.3 and 3.8.1 respectively. The kernel group and nogrouped models are the models for which the BF is obtained under the assumption that the distribution of the unknown true mean μ of the group means takes a kernel density form with the training data assumed to have a grouping and nogrouping structure, respectively. The normal model assumes μ to be Normally distributed with some known parameters.

Control and recovered test data are generated from a Normal distribution with different values of mean and variance which determine whether they are from the same or different sources. The control and recovered test data are regarded as coming from the same source if they were generated from a distribution with the same mean and as coming from a different source otherwise. This experiment is carried out in a view of the example given in Section 3.6. The distributions used are:

(i) Normal $N(\mu_X, \sigma_X^2)$ for the control test data with μ_X = 0.6789, 0.8187 or 0.5095 and σ_X = 0.0656609;

(ii) Normal $N(\mu_Y, \sigma_Y^2)$. For the recovered test data in the same source case, μ_Y is as in (i) and σ_Y = 0.0656609 or 0.01. In the different

source case μ_Y = 0.4923 or 0.9558, 0.7887 or 0.5395 and 0.6957 or 0.6325: μ_X and σ_X take the values as in (i). The μ_Y's are chosen such that the are approximately 2 or 3 standard deviations from μ_X, respectively.

Notice that the choice of 0.01 for σ_Y in (ii) above is entirely arbitary. The reason and purpose of its choice is to show that the error probabilities should be smaller than those obtained using σ_Y which equals 0.0656609 since the sample generated from a smaller σ_Y will be much closer together than those generated from a bigger σ_Y. Hence the chance of X and Y are being regarded coming from different source is minute, after having being generated from the same mean.

There are a total of 12 comparisons between the control and recovered simulated samples. Various sample sizes of the control and recovered data are also under consideration. Under each model, the numbers of misclassified cases, in terms of BF less than or greater than a threshold value out of the 100 simulations in each run for the same and different sources examples, respectively, are counted. The numbers of runs are 20 and 50. The estimated probabilities of Type I and Type II errors are then obtained. The estimated error probabilities under the three models are tabulated in Tables 3.22 and 3.23, with 20 and 50 runs respectively, for a threshold value of one. Standard errors of the estimated probabilities are shown in parentheses. Since results from the two sets of independent runs are similar, only the results from 20 runs are discussed. The estimated error probabilities of both Type I and II under the Normal model are relatively higher than the two kernel based models. The error probabilities appear to be higher when $\bar{X}$ and $\bar{Y}$ are both common. As

Table 3.22 Estimated probabilities (shown in **bold**) of Type I and Type II errors of 20 runs of size 100 for various sample sizes of X and of Y under i) kernel group, ii) Normal and iii) kernel nogroup models; standard error ($\times 10^{-3}$) of the estimate shown in brackets.

	Type I error					
$\mu_X=\mu_Y=$	0.6785		0.8187		0.5095	
$\sigma_Y=$	0.0657	0.01	0.0657	0.01	0.0657	0.01
a) i)	**0.3130**	**0.1520**	**0.1185**	**0.0455**	**0.0755**	**0.0270**
	(7.4020)	(6.6332)	(8.9228)	(4.6154)	(4.7833)	(4.1738)
ii)	**0.3645**	**0.1865**	**0.0945**	**0.0420**	**0.0430**	**0.0210**
	(8.8398)	(8.0549)	(6.8622)	(4.6791)	(3.7062)	(3.4717)
iii)	**0.2925**	**0.1315**	**0.1220**	**0.0455**	**0.0840**	**0.0280**
	(7.2864)	(6.6992)	(9.1075)	(4.6154)	(5.7308)	(4.2051)
b) i)	**0.2535**	**0.0140**	**0.1105**	**0.0000**	**0.0750**	**0.0000**
	(9.2984)	(2.3395)	(7.6940)	(0.0000)	(4.1988)	(0.0000)
ii)	**0.2975**	**0.0370**	**0.1000**	**0.0000**	**0.0640**	**0.0000**
	(9.2016)	(3.8458)	(7.2909)	(0.0000)	(3.7975)	(0.0000)
iii)	**0.2305**	**0.0095**	**0.1120**	**0.0005**	**0.0750**	**0.0000**
	(8.5062)	(1.8460)	(7.9006)	(0.5000)	(3.9403)	(0.0000)
c) i)	**0.2350**	**0.0015**	**0.1055**	**0.0000**	**0.0965**	**0.0000**
	(10.649)	(0.8191)	(6.6679)	(0.0000)	(6.1248)	(0.0000)
ii)	**0.2765**	**0.0105**	**0.1045**	**0.0000**	**0.0900**	**0.0000**
	(9.7137)	(1.6976)	(6.3796)	(0.0000)	(7.3244)	(0.0000)
iii)	**0.2080**	**0.0005**	**0.1050**	**0.0000**	**0.0935**	**0.0000**
	(10.198)	(0.5000)	(6.7862)	(0.0000)	(5.7708)	(0.0000)
d) i)	**0.1720**	**0.0615**	**0.0400**	**0.0085**	**0.0225**	**0.0025**
	(8.0328)	(4.3695)	(4.1675)	(1.6663)	(3.7608)	(0.9935)
ii)	**0.2230**	**0.0905**	**0.0230**	**0.0040**	**0.0070**	**0.0015**
	(8.6175)	(5.5476)	(3.1705)	(1.3376)	(2.0647)	(0.8191)
iii)	**0.1400**	**0.0485**	**0.0455**	**0.0085**	**0.0290**	**0.0030**
	(6.6885)	(3.9918)	(4.3815)	(1.6663)	(3.8319)	(1.0513)
e) i)	**0.1355**	**0.0165**	**0.0380**	**0.0000**	**0.0195**	**0.0000**
	(7.3081)	(2.8353)	(4.7903)	(0.0000)	(3.8713)	(0.0000)
ii)	**0.1745**	**0.0290**	**0.0200**	**0.0000**	**0.0080**	**0.0000**
	(9.6374)	(3.8319)	(2.9911)	(0.0000)	(2.0000)	(0.0000)
iii)	**0.1200**	**0.0110**	**0.0440**	**0.0000**	**0.0265**	**0.0000**
	(6.6885)	(2.5026)	(5.0991)	(0.0000)	(4.2471)	(0.0000)
f) i)	**0.1210**	**0.0355**	**0.0270**	**0.0020**	**0.0145**	**0.0010**
	(8.7629)	(5.3050)	(4.2364)	(1.1696)	(3.2016)	(0.6883)
ii)	**0.1540**	**0.0560**	**0.0100**	**0.0010**	**0.0025**	**0.0005**
	(8.9561)	(6.8211)	(2.2942)	(0.6883)	(1.2300)	(0.5000)
iii)	**0.1045**	**0.0285**	**0.0350**	**0.0020**	**0.0200**	**0.0020**
	(7.5558)	(4.7169)	(3.9403)	(1.1696)	(3.3245)	(1.1696)

Table 3.22 cont'd.

	Type II error					
μ_X=	0.6785		0.8187		0.5095	
μ_Y=	0.4923	0.9558	0.6325	0.5395	0.6957	0.7887
a) i)	**0.2675**	**0.1205**	**0.1870**	**0.0225**	**0.2290**	**0.0250**
	(8.0091)	(7.1992)	(6.8479)	(3.0672)	(10.049)	(2.7624)
ii)	**0.3040**	**0.1620**	**0.1865**	**0.0220**	**0.2530**	**0.0245**
	(7.6227)	(7.6983)	(8.6837)	(2.7720)	(8.4945)	(3.6617)
iii)	**0.2610**	**0.1050**	**0.1935**	**0.0260**	**0.2270**	**0.0275**
	(8.6723)	(7.0149)	(6.6203)	(2.8469)	(9.8436)	(3.0672)
b) i)	**0.1255**	**0.0130**	**0.0865**	**0.0030**	**0.0890**	**0.0030**
	(7.0140)	(2.9109)	(5.2453)	(1.6383)	(6.3204)	(1.0513)
ii)	**0.1455**	**0.0265**	**0.0850**	**0.0030**	**0.0970**	**0.0030**
	(7.7280)	(2.9267)	(5.1042)	(1.6383)	(7.4020)	(1.2772)
iii)	**0.1260**	**0.0105**	**0.0915**	**0.0040**	**0.0925**	**0.0040**
	(7.1598)	(2.1119)	(6.1248)	(1.8353)	(6.3193)	(1.5218)
c) i)	**0.0980**	**0.0040**	**0.0565**	**0.0015**	**0.0920**	**0.0020**
	(5.8759)	(1.5218)	(6.1675)	(0.8191)	(5.6939)	(0.9177)
ii)	**0.1020**	**0.0060**	**0.0565**	**0.0015**	**0.0950**	**0.0015**
	(7.5950)	(1.9735)	(5.7250)	(0.8191)	(5.8714)	(0.8191)
iii)	**0.1035**	**0.0025**	**0.0580**	**0.0015**	**0.0970**	**0.0020**
	(6.3358)	(1.4281)	(5.9647)	(0.8191)	(6.1601)	(0.9177)
d) i)	**0.0035**	**0.0000**	**0.0015**	**0.0000**	**0.0020**	**0.0000**
	(1.0943)	(0.0000)	(0.8191)	(0.0000)	(0.9177)	(0.0000)
ii)	**0.0055**	**0.0000**	**0.0010**	**0.0000**	**0.0020**	**0.0000**
	(1.3523)	(0.0000)	(0.6883)	(0.0000)	(0.9177)	(0.0000)
iii)	**0.0030**	**0.0000**	**0.0020**	**0.0000**	**0.0020**	**0.0000**
	(1.0513)	(0.0000)	(0.0918)	(0.0000)	(0.0918)	(0.0000)
e) i)	**0.0005**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	(0.5000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
ii)	**0.0010**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	(0.6883)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
iii)	**0.0005**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	(0.5000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
f) i)	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
ii)	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
iii)	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

Note : a) for r=1,m=1; b) for r=1,m=5; c) for r=1,m=10;
d) for r=5,m=5; e) for r=5,m=10 and f) for r=10,m=10

Table 3.23 Estimated probabilities (shown in **bold**) of Type I and Type II errors of 50 runs of size 100 for various sample sizes of X and of Y under i) kernel group, ii) Normal and iii) kernel nogroup models; standard error ($\times 10^{-3}$) of the estimate shown in brackets.

Type I error

$\mu_X=\mu_Y=$	0.6785		0.8187		0.5095	
$\sigma_Y=$	0.0657	0.01	0.0657	0.01	0.0657	0.01
a) i)	**0.3116**	**0.1582**	**0.1072**	**0.0416**	**0.0800**	**0.0290**
	(6.0160)	(5.5305)	(4.9324)	(2.9187)	(3.4522)	(2.6380)
ii)	**0.3548**	**0.2020**	**0.0852**	**0.0366**	**0.0522**	**0.0214**
	(6.4158)	(5.7142)	(4.4834)	(2.6590)	(3.0276)	(2.3905)
iii)	**0.2924**	**0.1388**	**0.1110**	**0.0412**	**0.0936**	**0.0328**
	(5.6609)	(5.6397)	(4.8001)	(2.8376)	(4.1379)	(2.9002)
b) i)	**0.2518**	**0.0200**	**0.1306**	**0.0000**	**0.0800**	**0.0000**
	(6.2184)	(2.3733)	(4.0883)	(0.0000)	(4.6422)	(0.0000)
ii)	**0.2980**	**0.0406**	**0.0942**	**0.0000**	**0.0652**	**0.0000**
	(6.8570)	(3.1610)	(3.6932)	(0.0000)	(3.9095)	(0.0000)
iii)	**0.2288**	**0.0124**	**0.1048**	**0.0000**	**0.0826**	**0.0000**
	(6.1587)	(1.8857)	(4.1031)	(0.0000)	(4.5875)	(0.0000)
c) i)	**0.2320**	**0.0022**	**0.1104**	**0.0000**	**0.0926**	**0.0000**
	(5.2450)	(0.7715)	(4.0300)	(0.0000)	(4.3967)	(0.0000)
ii)	**0.2770**	**0.0114**	**0.1100**	**0.0000**	**0.0878**	**0.0000**
	(5.8011)	(1.3999)	(3.9795)	(0.0000)	(4.6448)	(0.0000)
iii)	**0.2106**	**0.0012**	**0.1110**	**0.0000**	**0.0902**	**0.0000**
	(5.6814)	(0.4643)	(4.0025)	(0.0000)	(4.4947)	(0.0000)
d) i)	**0.1808**	**0.0522**	**0.0366**	**0.0062**	**0.0248**	**0.0020**
	(5.6267)	(3.0942)	(2.5170)	(1.0257)	(2.4946)	(0.6389)
ii)	**0.2258**	**0.0810**	**0.0174**	**0.0030**	**0.0070**	**0.0008**
	(6.1881)	(3.8571)	(1.7097)	(0.8690)	(1.3170)	(0.3875)
iii)	**0.1409**	**0.0422**	**0.0452**	**0.0090**	**0.0310**	**0.0028**
	(5.2235)	(2.8755)	(2.6379)	(1.2856)	(2.6991)	(0.8102)
e) i)	**0.1614**	**0.0156**	**0.0326**	**0.0004**	**0.0220**	**0.0006**
	(4.9979)	(1.7879)	(2.2822)	(0.2799)	(2.1571)	(0.3393)
ii)	**0.2042**	**0.0316**	**0.0202**	**0.0002**	**0.0094**	**0.0002**
	(5.4592)	(2.5771)	(2.0302)	(0.2000)	(1.2907)	(0.2000)
iii)	**0.1352**	**0.0102**	**0.0402**	**0.0008**	**0.0282**	**0.0006**
	(4.4834)	(1.3847)	(2.6261)	(0.4815)	(2.6448)	(0.3393)
f) i)	**0.1204**	**0.0292**	**0.0264**	**0.0030**	**0.0198**	**0.0010**
	(4.7462)	(2.3183)	(2.5834)	(0.7693)	(1.9057)	(0.4285)
ii)	**0.1568**	**0.0492**	**0.0146**	**0.0016**	**0.0068**	**0.0002**
	(4.9109)	(3.0080)	(1.6711)	(0.5237)	(1.0475)	(0.2000)
iii)	**0.1022**	**0.0224**	**0.0314**	**0.0034**	**0.0252**	**0.0018**
	(4.5110)	(1.7510)	(2.5716)	(0.7881)	(2.1235)	(0.5489)

Table 3.23 cont'd.

	Type II error					
μ_X=	0.6785		0.8187		0.5095	
μ_Y=	0.4923	0.9558	0.6325	0.5395	0.6957	0.7887
a) i)	**0.2532**	**0.1176**	**0.1880**	**0.0208**	**0.2254**	**0.0296**
	(7.1340)	(4.9780)	(4.4447)	(1.6621)	(6.5851)	(2.2849)
ii)	**0.2976**	**0.1654**	**0.1822**	**0.0194**	**0.2470**	**0.0310**
	(7.3710)	(5.9261)	(4.8763)	(1.5494)	(6.7142)	(2.5912)
iii)	**0.2400**	**0.1044**	**0.1976**	**0.0234**	**0.2232**	**0.0316**
	(6.5215)	(4.8332)	(4.5315)	(1.8208)	(6.4551)	(2.2741)
b) i)	**0.1236**	**0.0140**	**0.0780**	**0.0032**	**0.1046**	**0.0032**
	(4.2161)	(2.0404)	(4.3799)	(0.8300)	(4.4281)	(0.7251)
ii)	**0.1416**	**0.0272**	**0.0774**	**0.0032**	**0.0116**	**0.0036**
	(4.3925)	(2.6807)	(4.3687)	(0.7794)	(4.8936)	(0.7959)
iii)	**0.1232**	**0.0106**	**0.0812**	**0.0034**	**0.1078**	**0.0040**
	(4.7931)	(1.6759)	(4.5860)	(0.8857)	(4.4656)	(0.7559)
c) i)	**0.0956**	**0.0052**	**0.0610**	**0.0014**	**0.0834**	**0.0026**
	(4.1616)	(0.9578)	(4.0532)	(0.6395)	(4.1562)	(0.7456)
ii)	**0.1038**	**0.0106**	**0.0576**	**0.0012**	**0.0872**	**0.0028**
	(4.3037)	(1.3525)	(4.0159)	(0.5450)	(4.1111)	(0.7022)
iii)	**0.1030**	**0.0034**	**0.0626**	**0.0018**	**0.0882**	**0.0034**
	(4.3542)	(0.7345)	(4.0688)	(0.6815)	(4.2637)	(0.8383)
d) i)	**0.0030**	**0.0000**	**0.0014**	**0.0000**	**0.0016**	**0.0000**
	(0.7693)	(0.0000)	(0.6395)	(0.0000)	(0.5967)	(0.0000)
ii)	**0.0038**	**0.0000**	**0.0012**	**0.0000**	**0.0022**	**0.0000**
	(0.8026)	(0.0000)	(0.6155)	(0.0000)	(0.6572)	(0.0000)
iii)	**0.0030**	**0.0000**	**0.0014**	**0.0000**	**0.0020**	**0.0000**
	(0.7143)	(0.0000)	(0.5722)	(0.0000)	(0.7000)	(0.0000)
e) i)	**0.0002**	**0.0000**	**0.0002**	**0.0000**	**0.0002**	**0.0000**
	(0.2000)	(0.0000)	(0.2000)	(0.0000)	(0.2000)	(0.0000)
ii)	**0.0004**	**0.0000**	**0.0002**	**0.0000**	**0.0002**	**0.0000**
	(0.3999)	(0.0000)	(0.2000)	(0.0000)	(0.2000)	(0.0000)
iii)	**0.0002**	**0.0000**	**0.0002**	**0.0000**	**0.0002**	**0.0000**
	(0.2000)	(0.0000)	(0.2000)	(0.0000)	(0.2000)	(0.0000)
f) i)	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
ii)	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
iii)	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

Note : a) for r=1,m=1; b) for r=1,m=5; c) for r=1,m=10;
d) for r=5,m=5; e) for r=5,m=10 and f) for r=10,m=10

the sample sizes of X and of Y increase, error probabilities are reduced greatly, especially for the Type II error.

3.8.3 Validation of kernel density as an estimate for the prior density of μ

The simulation study is carried out to validate the assumption of training data in estimating the distribution of the between group random factor. It also examines whether the use of the group means as the data points to construct a kernel density estimate, $\hat{f}(\mu)$, will give a better fit to $f(\mu)$ if the training data has an obvious random structure. If $f(\mu)$ is non-Normal, will assuming the training data is grouped be beneficial in estimating $f(\mu)$? Thus, this simulation study is designed not only to measure the goodness of fit of the kernel density estimate, using the training data, but also the effect of $f(\mu)$ being non-Normal indicated from the training data. Then the kernel density estimate obtained by the ungrouped training data might not be the best to represent the true density $f(\mu)$, especially when there is an apparent random structure in the training data.

The kernel density estimate, $\hat{f}(\mu)$, for $f(\mu)$ obtained by assuming the training data is either grouped or not grouped (see Section 3.5.1 for details). The goodness of fit of the kernel density estimate is measured by Mean Integrated Square Error (MISE). In this study, like the entire thesis, I restrict myself only to consider the training data consisted of equal within-group size, i.e every group has an equal number of observations. Various group sizes and numbers of observations within each group are also considered in this study.

Training data are generated from the model (3.1) with μ_i generated from the following mixture of two Normal distributions:

$$p(\mu_i|p,\mu_1,\mu_2,\sigma_1,\sigma_2) = p\, f_N(\mu_i|\mu_1,\sigma_1) + (1-p) f_N(\mu_i|\mu_2,\sigma_2),$$

$-\infty < \mu_1,\mu_2 < \infty$, $0 \leq p \leq 1$, $\sigma_1,\sigma_2 > 0$.

Random samples of μ_i generated from the above distribution are said to be from

a) a Normal distribution with mean 5 and unit variance, if p is set to 1, and $\mu_1 = 5.0$ and variance $\sigma_1^2 = 1.0$,

b) a Skewed distribution if $p = 0.95$, $\mu_1 = -0.1$, $\mu_2 = 1.9$, $\sigma_1 = 1.0$ and $\sigma_2 = 3.0$, and

c) a Bimodal distribution if $p = 0.5$, $\mu_1 = -1.5$, $\mu_2 = 1.5$, $\sigma_1 = 1.0$ and $\sigma_2 = 0.33$.

The density functions of the above three distributions are plotted in Fig. 3.13. Formulae for the mean and variance of the distributions (b) and (c) are shown in Appendix 1. The usual error term ϵ_{ij} in the model (3.1) is generated from a Normal distribution with mean zero and variance σ^2. The within-group variance, σ^2, is set to be either equal to the between-group variance, σ_a^2 or one-hundredth of σ_a^2.In the former case the training data will not represent an obvious or strong random effects structure. The group sizes are chosen to be 10, 20, 50 or 100 and the number of within-group observations is 1, 5 or 10. When the number of within-group observation is one, the assumption of the training data being grouped is same as the training data being ungrouped. Two

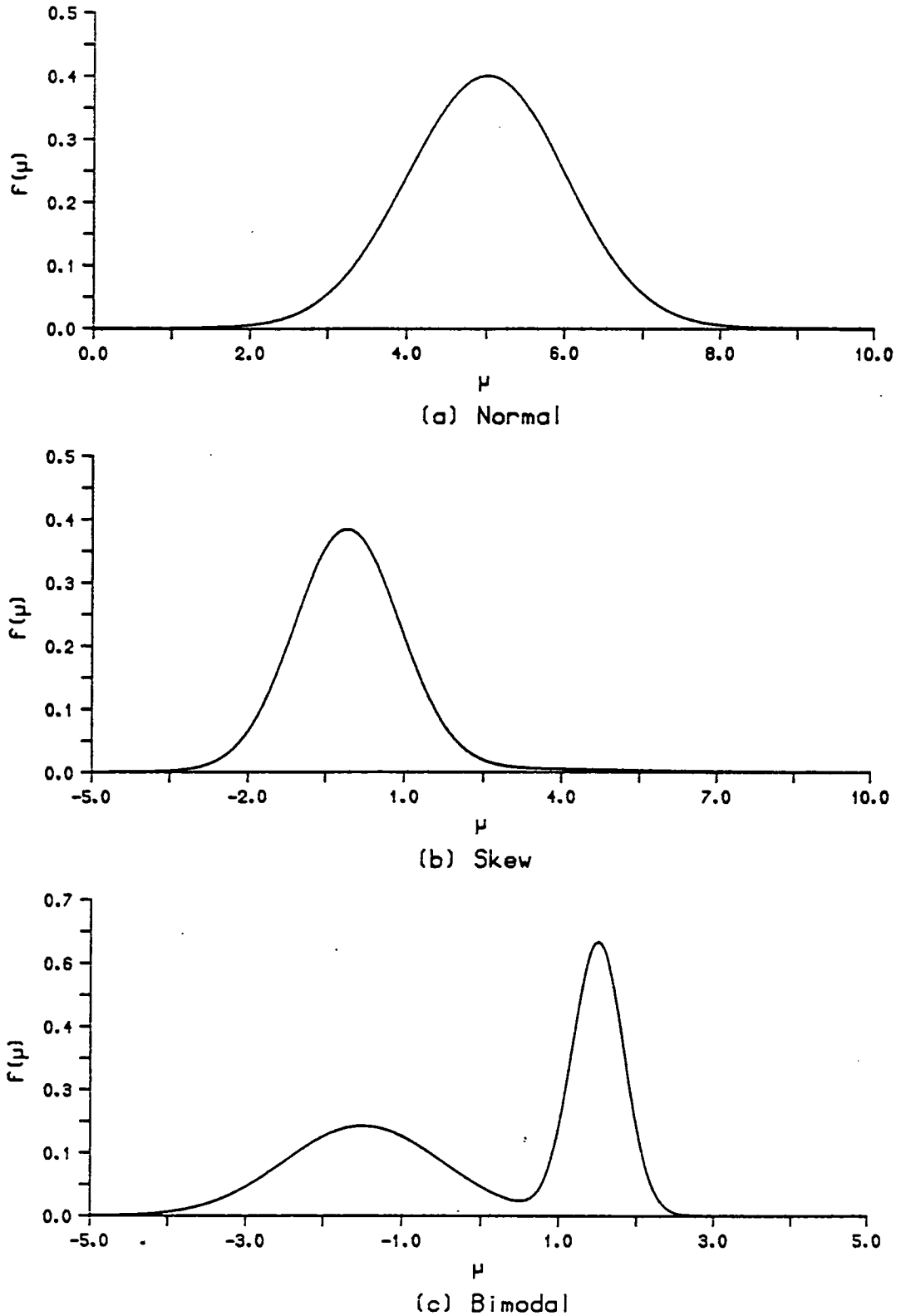


Fig. 3.13 Density plot of the true underlying distribution used in Section 3.7.3 for a simulation study

kernel density estimation methods are used: they are the ordinary and adaptive method described in Section 2.1.

Results of the study are shown in Tables 3.24 and 3.25 for $\sigma = \sigma_B$ and $\sigma = 0.1 \times \sigma_B$, respectively. Because of the computational time, only one run of the simulation for each of the above cases is carried out. Although one should not read too much into one set of simulations, the results shown in these Tables do indicate, as one might expect, that using the grouped sample means to estimate the between group distribution is more adequate when there is an obvious random structure in the training data (see Table 3.25). From both Tables, it is clear that the MISE increases as the group size increases. Moreover MISE, in general, decreases as the number of within-group observations increases, when the training data has an obvious random structure (Table 3.25). In this simulation study, the adaptive kernel method does not perform as well as the ordinary kernel method in terms of MISE, though the adaptive kernel method seems to improve the behaviour of the Bayes' factor. This confirmed the findings by Breiman et al (1977), see remarks by Abramson (1982).

Table 3.24 shows the results where the training data do not have a sufficient grouped structure since $\sigma = \sigma_B$. Hence the use of ungrouped data to estimate $f(\mu)$ seems to be better in terms of MISE. However if the grouped structure in the training data is strong, as showed in Table 3.25, then using the group means to construct $f(\mu)$ is more superior to the ungrouped data. This becomes even more so when the underlying true distribution for μ is non-Normal.

Table 3.24 MISE of the estimate for $f(\mu)$ given different group size (n) and within group size (J) assuming the training data is either grouped or ungrouped using the ordinary or the adaptive kernel method. $\sigma = \sigma_B$

Between group distribution: Normal

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
10	1	0.3590	0.3671	0.3590	0.3671
	5	0.7716	0.8323	1.2501	1.2496
	10	0.4347	0.4489	0.5803	0.5885
20	1	0.2688	0.2769	0.2688	0.2769
	5	0.4352	0.4569	0.5304	0.5391
	10	0.3189	0.3434	0.3640	0.3889
50	1	0.2961	0.3229	0.2961	0.3229
	5	0.3314	0.3516	0.3922	0.4198
	10	0.3211	0.3415	0.4281	0.4437
100	1	0.3742	0.4073	0.3742	0.4073
	5	0.3111	0.3207	0.3960	0.4102
	10	0.3154	0.3313	0.4075	0.4332
Between group distribution: Skewed					
10	1	0.3176	0.3542	0.3176	0.3542
	5	0.3892	0.4196	0.4164	0.4605
	10	0.6599	0.7119	0.8140	0.8822
20	1	0.5343	0.5334	0.5343	0.5334
	5	0.5766	0.6193	0.7339	0.7402
	10	0.6003	0.6492	0.7426	0.8319
50	1	0.8107	0.8651	0.8107	0.8651
	5	0.7258	0.7900	0.9934	1.0802
	10	0.7120	0.7516	1.0016	1.0952
100	1	0.5505	0.6322	0.5505	0.6322
	5	0.6184	0.6454	0.7474	0.8219
	10	0.6348	0.6722	0.9474	1.0063
Between group distribution: Bimodal					
10	1	0.2954	0.3048	0.2954	0.3048
	5	0.4041	0.4036	0.5113	0.5032
	10	0.3854	0.4061	0.5302	0.5628
20	1	0.3956	0.4047	0.3956	0.4047
	5	0.3450	0.3551	0.3860	0.3765
	10	0.3675	0.3811	0.5299	0.5410
50	1	0.3508	0.3711	0.3508	0.3711
	5	0.3516	0.3664	0.4364	0.4588
	10	0.3869	0.3936	0.6134	0.6542
100	1	0.3458	0.3593	0.3458	0.3593
	5	0.3543	0.3509	0.6062	0.6421
	10	0.3593	0.3644	0.7461	0.8057

Table 3.25 MISE of the estimate for $f(\mu)$ given different group size (n) and within group size (J) assuming the training data is either grouped or ungrouped using the ordinary or the adaptive kernel method. $\sigma=0.1\times\sigma_B$

Between group distribution: Normal

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordianry	Adaptive
10	1	0.4375	0.4613	0.4375	0.4613
	5	0.4614	0.4627	0.4062	0.4156
	10	0.2953	0.2961	0.2871	0.2956
20	1	0.5312	0.5779	0.5312	0.5779
	5	0.5868	0.5877	0.5510	0.5631
	10	0.5756	0.5761	0.4596	0.4901
50	1	0.5062	0.5426	0.5062	0.5426
	5	0.5054	0.5069	0.4779	0.4962
	10	0.5010	0.5022	0.4818	0.5042
100	1	0.4718	0.4948	0.4718	0.4948
	5	0.5778	0.5804	0.5260	0.5539
	10	0.5895	0.5904	0.5522	0.5709
Between group distribution: Skewed					
10	1	0.6374	0.6794	0.6374	0.6794
	5	1.0112	1.0523	0.7156	0.7766
	10	1.4931	1.4703	1.3269	1.3499
20	1	1.1075	1.1417	1.1075	1.1417
	5	1.1951	1.2069	1.1865	1.1958
	10	1.3460	1.3621	1.1449	1.2099
50	1	0.8914	0.9585	0.8914	0.9585
	5	0.9625	0.9690	0.8633	0.9220
	10	0.9534	0.9618	0.7941	0.8750
100	1	0.9033	0.9943	0.9033	0.9943
	5	1.0083	1.0123	0.9416	0.9813
	10	0.9123	0.9137	0.8242	0.8896
Between group distribution: Bimodal					
10	1	0.7093	0.7905	0.7093	0.7905
	5	0.9291	0.9348	0.6779	0.6901
	10	2.0732	2.1610	0.6646	0.7997
20	1	0.8628	0.8973	0.8628	0.8973
	5	2.0845	2.2018	1.2963	1.7182
	10	0.9245	0.9608	0.6161	0.6693
50	1	0.8617	0.9529	0.8617	0.9529
	5	1.3158	1.3784	1.2039	1.3769
	10	0.9943	1.0193	0.9842	1.0844
100	1	1.1103	1.2669	1.1103	1.2669
	5	1.2719	1.3456	1.2534	1.3766
	10	1.2816	1.3086	1.2126	1.3454

3.9 Conclusions

Four methods of evaluation of the Bayes' factor have been developed based on different assumptions about the structure of the training data and the within-group variance. The training data in the example given in Section 3.6 do not represent the size of the underlying population of the training data since only 22 groups are considered. The computational time can be reduced greatly when the grouped model is used. The effect of ignoring the grouping structure in the training data is that the value of Bayes' factor is reduced slightly. If the random structure in the training data is apparent then the model developed in Section 3.5.2 should be used. There is not a lot to choose between the assumed known and unknown variance models as far as the grouped data model is concerned when one is dealing with a large training data set. It reflects the large degrees of freedom in estimating the within-group variance.

In view of the paradoxical phenomenon discussed in Section 3.6, an alternative approach might be to estimate the ratio of the densities nonparametrically rather than the densities themselves. This is problem because with the kernels involved, one would have the ratio of two sums of functions that may not be smooth. Simple approximations to these sums may merely lead us back to parametric densities.

The simulation study in Section 3.8.1 shows that an improvement over the Normality assumption of the Bayes' factor estimates is obtained by using a kernel method when the random effects are not Normally distributed.

In Section 3.5.1 part a, the sample group means of the training data were used to construct the distribution of μ. Other estimates for μ_i may be used as the data points, for instance, one could use the EB posterior estimator for μ_i (see Maritz (1970), Rao (1973)), namely

$$(1-\delta)\ \bar{z}_{i.} + \delta\ \bar{z}_{..}$$

where $\delta = \hat{\sigma}_e^2/(\hat{\sigma}_e^2+J\hat{\sigma}_a^2)$, $\hat{\sigma}_e^2$ = WMS, $\hat{\sigma}_a^2$ = max{ (BMS - $\hat{\sigma}_e^2$), 0} (see Section 2.3 for example). It will be interesting to see how the Bayes' factor behaves using above instead of the $\bar{z}_i$'s.

The models developed here could be extended to take into account other factors and situations such as it could be possible for a suspect to be present at the crime scene, not to have picked up cat hairs there and to have picked up cat hairs from some other source. Also the discussion so far has centred mainly on the occurrence of a single item of transfer evidence. In practice, several types of transfer evidence will be present, for example apart from cat hair there might be fibres or glass fragments involved as well. Then the functions in (3.3) should be estimated on the basis of this combination of evidence. Each type of transfer evidence will require the determination of a set in which the origin will belong. If there are q types of transfer evidence, there may be q separate sets. For each set, the probalilities of observing the control and recovered data giving they have come from the same or from different sources are estimated. However, the final probability required is that of a random selection being in the intersection of the q sets. Let D be

the intersection of the D_i sets (i=1,...,q). Then we have

$$P(D) = P(D_1, D_2, \ldots, D_q)$$
$$= P(D_1 | D_2 \ldots D_q) \times P(D_2 | D_3, \ldots D_q) \times \ldots \times P(D_q) \qquad (*)$$

where $D_1, D_2, \ldots D_q$ is the intersection of the D_i sets (i=1,...q), etc. If the properties defining the q sets are statistically independent, then (*) can be reduced to

$$P(D) = P(D_1) \times P(D_2) \times \ldots\ldots \times P(D_q). \qquad (**)$$

Further work is required to incorporate such possibilities into a measure of the strength of the evidence.

In view of the paper given by Makov (1987) the models developed here could also be extended to multi-suspects or even missing suspect problem. The extension of the group and assumed known within-group variance model to the multivariate case is explored in Chapter 6.

CHAPTER 4

ESTIMATION OF VARIANCE COMPONENTS

4.1 Introduction

In this chapter, I consider from a Bayesian viewpoint some aspects of a balanced one-way random effects model

$$z_{ij} = A_i + \epsilon_{ij} \quad (i=1,\ldots,n; j=1,\ldots,J) \tag{4.1}$$

where z_{ij} are the observations, μ is a location parameter, A_i and ϵ_{ij} are independently distributed random variables with means zero and variances σ_a^2 and σ_e^2 respectively. Thus,

$$E(Z_{ij}) = \mu, \quad \text{Var}(Z_{ij}) = \sigma_a^2 + \sigma_e^2.$$

In the usual analysis of the model (4.1), A_i and ϵ_{ij} are assumed Normal, and interest is usually centred on the estimation of the two variance components (σ_a^2, σ_e^2). The problems of estimation and hypothesis testing concerning the variances have already been outlined in Chapter 2. Here, I consider the situation in which A_i is not Normally distributed. In addition, unlike Tiao and Ali (1971) who specified the distribution of A_i to be a known mixture of two Normal distributions I assume that the distribution of A_i has an unknown distribution and model the sample group means by a kernel density. The effect of this on the inferences about the variance σ_a^2 is investigated. This effect is studied from a Bayesian viewpoint, and a comparison between the Normal and kernel models is made by a simulation study. The maximum likelihood (M.L.) and Bayesian estimates of the variance components under the two models are also

compared.

4.2 The likelihood function

To derive the likelihood function for a random effects model, it is convenient to work with the group means. Under the non-Normality assumption about the group means, the proposed density for the group means takes a kernel density form, namely

$$f(t|\lambda,\sigma_e^2,\sigma_a^2,\underset{\sim}{\bar{z}}_{.}) = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{[2\pi\sigma_3^2/J]^{\frac{1}{2}}}\exp\left\{-\frac{J(t-\bar{z}_{k.})^2}{2\sigma_3^2}\right\} \qquad (4.2)$$

where $t = \bar{z}_{i.}$ for $i=1,2,\ldots,n$, $\sigma_3^2 = \lambda^2\sigma_{1z}^2$, $\sigma_{1z}^2 = J\sigma_a^2 + \sigma_e^2$ and λ is the 'standardised' smoothing parameter and $\underset{\sim}{\bar{z}}_{.} = (\bar{z}_{1.},\ldots,\bar{z}_{n.})'$. With the similar manner due to Tiao and Ali (1971), it follows from (4.2) and the assumption of Normality of ϵ_{ij} and independence of (A_i,ϵ_{ij}), that

(i) $\bar{z}_{i.}$ and $(z_{ij}-\bar{z}_{i.})$ are independent,

(ii) $\nu_1 m_1$ is distributed as $\sigma_e^2\chi_{\nu_1}^2$.

Thus the likelihood function of the parameters is

$$L(\sigma_e^2,\sigma_a^2,\lambda|Z) \propto \frac{1}{(\sigma_e^2)^{\nu_1/2}}\exp\left\{-\frac{\nu_1 m_1}{2\sigma_e^2}\right\}\prod_{i=1}^{n} f^*(\bar{z}_{i.}|\lambda,\sigma_a^2,\sigma_e^2,\underset{\sim}{\bar{z}}_{.}) \qquad (4.3)$$

where Z denotes the entire data, ν_1 is the within group degrees of freedom as defined in (ii) above, m_1 is the sample within mean square and f^* is as in (4.2) except the summation over k does not include k=i (The reason for this is to avoid the returning of zeros (see Chapter 2 for details)) and n is replaced by n^* (= n-1).

The likelihood above is a product of two factors. The first factor represents information coming from the residuals about the parameter σ_e^2 only, whereas the second factor provides information about all the parameters $(\sigma_e^2, \sigma_a^2, \lambda)$ coming from the group means $\bar{z}_{i.}$'s. Note that estimation of the smoothing parameter λ is not of much interest as far as the variance components estimation problem is concerned.

Before I proceed to discuss the estimation of the variance components, I would like to give some references to several data sets which later will be used to illustrate the method of analysis of the problem.

There are six data sets, some were generated from known distributions and some are published data sets. The latter provide a direct comparison between the model I propose and models used by others. The simulated data are generated from the model (4.1) with the A_i from (a) Normal and (b) Gamma distributions. The distribution of ϵ_{ij} is assumed to be Normal. The published data were taken from Tiao and Ali (1971) and are reproduced in Table 4.1. (N.B. it is denoted as Tiao's data hereafter and the model which Tiao and Ali derived and applied to these data is called Tiao's model) The other two published data sets, taken from Tables 5.1.2 and 5.1.4 of Box and Tiao (1973), are tabulated in Table 4.2 and 4.3 respectively. Details of all these data sets are summarised in Table 4.4.

Table 4.1 The Ordered Group Means $\bar{z}_{i.}$ (Tiao and Ali (1971))

Group i	$\bar{z}_{i.}$	Group i	$\bar{z}_{i.}$
1	-3.682	11	-0.378
2	-2.057	12	0.000
3	-1.780	13	0.112
4	-1.238	14	0.791
5	-0.797	15	0.923
6	-0.671	16	1.571
7	-0.646	17	1.712
8	-0.471	18	4.223
9	-0.436	19	6.415
10	-0.401	20	7.072

Table 4.2 The Ordered Group Means (Dye Data) taken from Table 5.1.2 of Box and Tiao (1973)

Group i	1	2	3	4	5	6
$\bar{z}_{i.}$	1470	1498	1505	1528	1564	1600

Table 4.3 The Ordered Group Means (Generated Data) taken from Table 5.1.4 of Box and Tiao (1973)

Group i	1	2	3	4	5	6
$\bar{z}_{i.}$	3.8252	4.6560	5.6848	6.0796	6.2268	7.5212

Table 4.4 Summary of the data sets used in the analysis of variance components problem

Data	σ^2_{eT}	σ^2_{aT}	BMS	WMS(m_1)	$\hat{\lambda}$	n	J	N	ν_1
Normal	1	4	21.883	0.9998	0.5043	100	5	500	400
Gamma	1	2	27.042	0.9920	0.3872	100	10	1000	900
Tiao$^{\nabla}$	1	4	21.368	1.1525	0.5073	20	3	60	40
Dye$^{\Delta}$	-	-	1.1×10^{4}	2.5×10^{3}	0.9572	6	5	30	24
Generated$^{\Delta}$	16	4	8.3363	14.946	1.0186	6	5	30	24
Cats	-	-	4.7×10^{-2}	4.3×10^{-3}	0.7855	22	10	220	198

∇ Data taken from Tiao and Ali (1971);
Δ " " " Box and Tiao (1973);
σ^2_{eT} & σ^2_{aT} denote the 'True' values of σ^2_e and σ^2_a resp.

Two methods were used to obtain estimates for σ_a^2 and σ_e^2. they are the Maximum likelihood and Bayesian methods.

4.3 Maximum Likelihood (M.L.) Method

Searle (1971) derived maximum likelihood estimates for variance components in a balanced one-way random effects model under a Normality assumption about the distribution of the random factor A_i. For the kernel model, we obtain the estimates for the variance components and smoothing parameter by maximising the likelihood function of (4.3). The likelihood function is maximised with respect to all three unknown parameters simultaneously using the NAG maximisation routine E04JAF. There is a problem that different starting values of σ_a^2 and λ yield different estimates whereas the estimate for σ_e^2 remains unaltered. This leads to a suspicion that σ_a^2 and λ might be somehow related. To pursue it further I obtained a set of estimates for σ_a^2 and λ, fixing σ_e^2, by using different sets of starting values. Then, for the examples given in Table 4.4, the values of $\hat{\sigma}_a^2$ and $\hat{\lambda}$ are plotted and the relationship is shown in Fig. 4.1.

An analytical result confirmed the relationship implied by the graphs. Take the logarithm to the base e of the likelihood function (4.3) giving

$$\log L \propto -\left[\frac{v_1}{2}\right]\ell n(\sigma_e^2)-\left[\frac{v_1 m_1}{2\sigma_e^2}\right]-\left[\frac{n}{2}\right]\ell n(\sigma_{1z}^2)-$$

$$\left[\frac{n}{2}\right]\ell n(\lambda^2) + \sum_{i=1}^{n} \ell n\left[\sum_{k\neq i} \exp\left\{-\frac{J(\bar{z}_{i.}-\bar{z}_{k.})^2}{2\lambda^2\sigma_{1z}^2}\right\}\right].$$

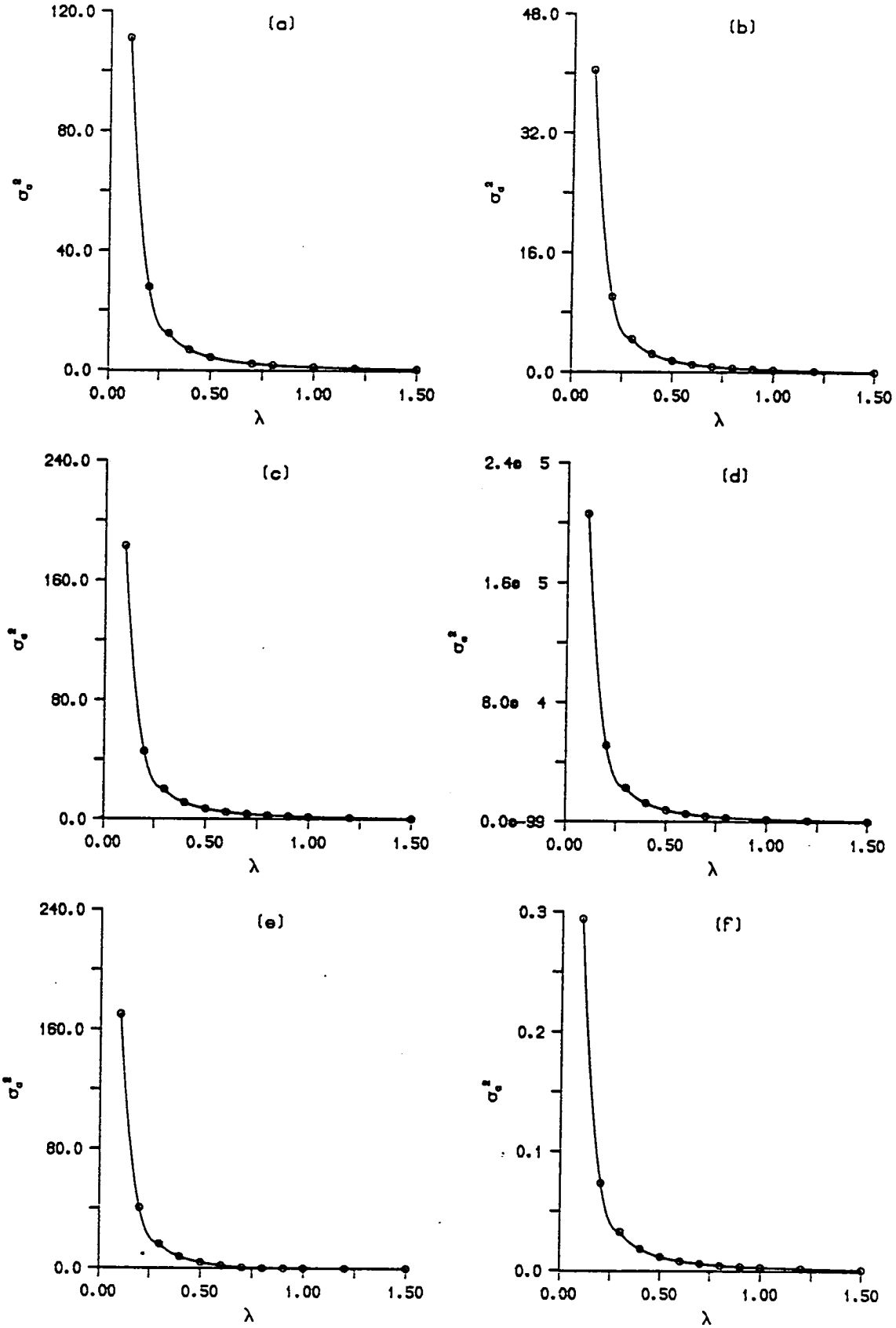


Fig. 4.1 Relationship between σ_e^2 and λ for (a) Normal; (b) gamma (c) Tiao; (d) Dye; (e) Generated and (f) cat data under kernel model.

Differentiate log L with respect to σ_e^2, σ_a^2 and λ^2 in turn to yield three differential equations, namely

$$(\dagger) \qquad \frac{d\log L}{d\sigma_e^2} = -\left[\frac{v_1}{2\sigma_e^2}\right] + \left[\frac{v_1 m_1}{2\sigma_e^4}\right] - \left[\frac{n}{2\sigma_{i2}^2}\right] + \sum_{i=1}^{n}\left[\sum_{\ell\neq i}(D_{i\ell}b_{i\ell})/(\sigma_{i2}^2\sum_{k\neq i}D_{ik})\right] = 0,$$

$$(\dagger\dagger) \qquad \frac{d\log L}{d\sigma_a^2} = -\left[\frac{nJ}{2\sigma_{i2}^2}\right] + \sum_{i=1}^{n}\left[J\sum_{\ell\neq i}(D_{i\ell}b_{i\ell})/(\sigma_{i2}^2\sum_{k\neq i}D_{ik})\right] = 0, \text{ and}$$

$$(\dagger\dagger\dagger) \qquad \frac{d\log L}{d\lambda^2} = -\left[\frac{n}{2\lambda^2}\right] + \sum_{i=1}^{n}\left[\sum_{\ell\neq i}(D_{i\ell}b_{i\ell})/(\lambda^2\sum_{k\neq i}D_{ik})\right] = 0;$$

where

$D_{ij} = \exp\{-b_{ij}\}$ and

$$b_{ij} = \frac{J(\bar{z}_{i.}-\bar{z}_{j.})^2}{2\lambda^2\sigma_{i2}^2} \qquad \text{for } j = \ell \text{ or } k.$$

Multiply $(\dagger)$ by J/λ^2 to yield,

$$\left[\frac{v_1 J}{2\lambda^2\sigma_e^2}\right] + \left[\frac{v_1 m_1 J}{2\lambda^2\sigma_e^4}\right] - \left[\frac{nJ}{2\lambda^2\sigma_{i2}^2}\right] + D = 0, \qquad (4.4)$$

where $D = \sum_{i=1}^{n}\left[J\sum_{\ell\neq i}(D_{i\ell}b_{i\ell})/(\lambda^2\sigma_{i2}^2\sum_{k\neq i}D_{ik})\right]$, and D_{ij} and b_{ij} are as after $(\dagger\dagger\dagger)$ for $j = \ell$ or k. Multiplying $(\dagger\dagger)$ by $1/\lambda^2$ and $(\dagger\dagger\dagger)$ by J/σ_{i2}^2, yields

$$-\left[\frac{nJ}{2\lambda^2\sigma_{i2}^2}\right] + D = 0. \qquad (4.5)$$

Substitute (4.5) in (4.4) to give a solution for σ_e^2 as m_1, namely the

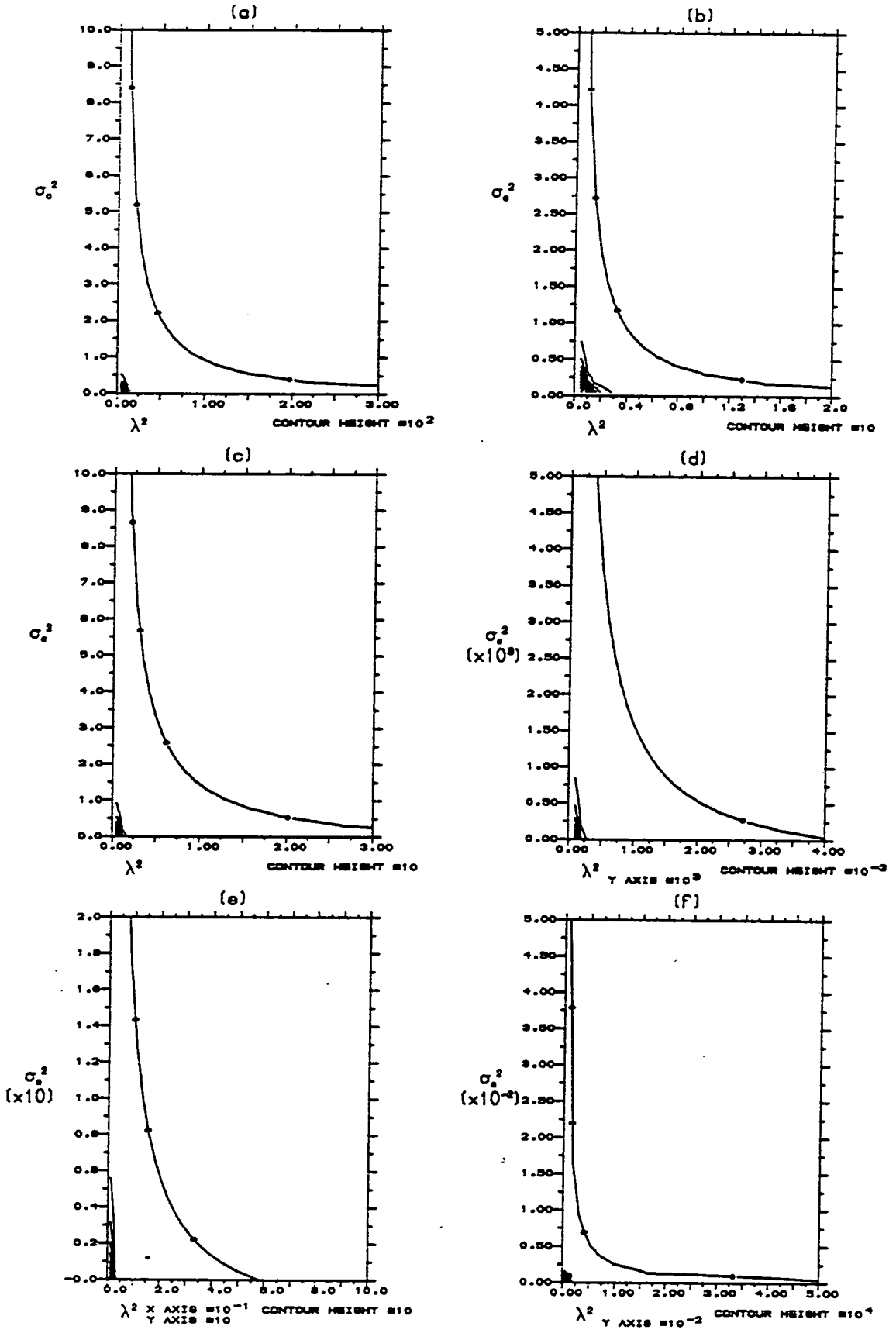


Fig. 4.2 Contour plot of relationship between σ_e^2 and λ^2 given by l.h.s. of equation (4.5) with σ_*^2 fixed equal to WMS for (a) Normal, (b) Gamma, (c) Tiao, (d) Dye, (e) Generated and (f) cat data.

within group mean square, which confirmed the finding earlier. Whereas the solutions for σ_a^2 and λ^2 are given by (4.5), hence the hyperbolic relationship between the two parameters. Contour plots of (4.5) as a function of σ_a^2 and λ^2 are shown in Fig. 4.2 for the examples in Table 4.4. The contours are the same as those shown in Fig. 4.1 as far as the relationship is concerned. The line in Fig. 4.2 represents a set of values of σ_a^2 and λ^2 which satisfied the equation (4.5).

However this relationship is not entirely surprising since, in the equation (4.2), there is basically one parameter. So if we let $\eta = \lambda\sigma_{12}$, the likelihood can be maximized with respect to σ_e^2 and η^2 and a unique solution can be obtained for σ_e^2 and η^2 separately. Furthermore, let the m.l.e. of η^2 be $\hat{\eta}^2$, then we have $\hat{\lambda}^2(J\hat{\sigma}_a^2+\hat{\sigma}_e^2) = \hat{\eta}^2$. Substitution of $\hat{\sigma}_e^2$ from the maximisation of the first component in (4.3) will give one equation in two unknowns $\hat{\lambda}$ and $\hat{\sigma}_a^2$. These two will be related by

$$\hat{\lambda}^2 = \frac{\hat{\eta}^2}{(J\hat{\sigma}_a^2 + \hat{\sigma}_e^2)} .$$

If $\hat{\sigma}_e^2$ is small relative to $J\hat{\sigma}_a^2$, this will be approximately represented as $\hat{\lambda}^2 = \hat{\eta}^2/J\hat{\sigma}_a^2$. The hyperbola in Fig. 4.1 agrees with this relationship.

There seems to be no unique solution as far as these two parameters are concerned. One possible solution is to fix λ, then maximise the likelihood function with respect to σ_e^2 and σ_a^2 only, to obtain M.L. estimates for them. An objective choice of λ, as before, is determined by maximum likelihood leave-one-out method. Thus the $\hat{\lambda}$

for each data set is shown in Table 4.4 also. Conditioning on $\hat{\lambda}$, the maximum likelihood estimates for σ_a^2 and σ_e^2 under the kernel model are tabulated in Tables 4.5 and 4.6. Also shown in the table is the MLE under the Normal model, and the ANOVA estimate. Note that the maximum likelihood estimate for σ_a^2 always has a downward bias since it fails to take into account the loss of degrees of freedom. Notice also the similarity between the ANOVA estimate and the kernel estimate. This is as expected since when $\hat{\lambda}$ was obtained, the group means are first standardised by their sample deviation s. The sample variance $s^2 = (n-1)^{-1} \sum_{i=1}^{n} (\bar{z}_{i.} - \bar{z}_{..})^2$ is an unbiased estimator of $(\sigma_a^2 + \sigma_e^2/J)$ since

$$s^2 = \text{BMS}/J \text{ and } E(\text{BMS}) = \sigma_e^2 + J\sigma_a^2.$$

Thus, this is equivalent to obtaining $\hat{\lambda}$ by fixing the values of σ_e^2 and σ_a^2 and hence of $E(\text{BMS})$. Notice that the fixed values of σ_e^2 and σ_a^2 are the ANOVA estimates by the above properties.

Table 4.5 Estimates for σ_a^2 using Maximum likelihood and Bayesian method under the Normal and Kernel model.

Data	ANOVA	Normal model		Kernel Model	
		MLE	Post'r mode	MLE	Post'r mode
Normal	4.1765	4.1328	4.08989	4.1766	3.11645
Gamma	2.6049	2.5779	2.55142	2.6050	2.28426
Tiao	6.7385	6.3824	6.06015	6.7386	5.28223
Dye	1764.1	1388.3	1119.96	1764.1	578.446
Generated	-1.322	0.0000	0.00000	-*	0.00000
Cats	0.0043	0.0041	0.00425	0.0043	0.00325

* NAG routine fails to find a maximum

Table 4.6 Estimates for σ_e^2 using Maximum likelihood and Bayesian method under the Normal and Kernel model.

Data	ANOVA	Normal model		Kernel Model	
		MLE	Post'r mode	MLE	Post'r mode
Normal	0.9998	0.9998	0.99480	0.9998	0.99480
Gamma	0.9920	0.9920	0.98977	0.9920	0.98977
Tiao	1.1525	1.1525	1.09762	1.1525	1.09763
Dye	2451.3	2451.3	2262.58	2451.2	2252.88
Generated	14.946	13.346	12.4144	14.034	12.5234
Cats	0.0043	0.0043	0.00427	0.0043	0.00427

4.4 Bayesian method

4.4.1 Prior and Posterior distribution

For a more general case, let the prior distribution for these parameters be

$$p(\sigma_e^2,\sigma_a^2,\lambda) = p(\sigma_e^2,\sigma_a^2|\lambda)p(\lambda). \qquad (4.6)$$

Suppose that we are in a situation that for any given λ, little is known about σ_e^2 and σ_a^2. We shall take the same prior distribution as used by Tiao and Ali (1971), namely

$$p(\sigma_e^2,\sigma_a^2|\lambda) \propto \sigma_e^{-2}(\sigma_e^2+J\sigma_a^2)^{-1} \qquad (4.7)$$

subject to the restriction of $\sigma_a^2 > 0$. Note that this choice of prior distribution was initially suggested by Tiao and Tan (1965). This choice of prior distribution (4.7) was criticized by Stone and Springer (1965). Box and Tiao (1973) showed that the vague prior on the expected mean squares of the analysis of variance is equivalent to a Jeffreys' type vague non-informative prior on the variance components. Nevertheless, continue with this prior meanwhile. Later

a more general prior is considered.

Combining the prior distributions in (4.6) and (4.7) with the likelihood function in (4.3) yields the joint posterior distribution of $(\sigma_e^2,\sigma_a^2,\lambda)$ as

$$p(\sigma_e^2,\sigma_a^2,\lambda|Z) \propto p(\lambda)\times\frac{1}{(\sigma_e^2)^{(v_1/2)+1}}\exp\left\{-\frac{v_1m_1}{2\sigma_e^2}\right\} Q(\sigma_e^2,\sigma_a^2,\lambda|Z) \qquad (4.8)$$

where

$$Q(\sigma_e^2,\sigma_a^2,\lambda|Z) = (\sigma_e^2+J\sigma_a^2)^{-1}\prod_{i=1}^{n} f^*(\bar{z}_{i.}|\sigma_e^2,\sigma_a^2,\lambda,\bar{\underset{\sim}{z}}_{..}), \qquad \sigma_e^2>0,\ \sigma_a^2>0,\ \lambda>0,$$

and the prior, $p(\lambda)$, for λ is independent of σ_e^2, σ_a^2. The joint posterior distribution of σ_e^2 and σ_a^2, conditional on λ and Z can be written as

$$p(\sigma_e^2,\sigma_a^2|\lambda,Z) \propto \left[\frac{1}{\sigma_e^2}\right]^{(v_1/2)+1}\exp\left\{-\frac{v_1m_1}{2\sigma_e^2}\right\} Q(\sigma_e^2,\sigma_a^2,\lambda|Z). \qquad (4.9)$$

Inference about σ_a^2 may be obtained from the marginal posterior distribution of σ_a^2, conditional on λ and Z, which may be obtained by integrating (4.9) over σ_e^2, yielding

$$p(\sigma_a^2|\lambda,Z) \propto \int_0^\infty \left[\frac{1}{\sigma_e^2}\right]^{(v_1/2)+1}\exp\left\{-\frac{v_1m_1}{2\sigma_e^2}\right\} Q(\sigma_e^2,\sigma_a^2,\lambda|Z)\, d\sigma_e^2 \qquad (4.10)$$

where Q is given after (4.8). In general, we may write (4.10) as the expectation

$$p(\sigma_a^2|\lambda,Z) \propto E_x Q(x,\sigma_a^2,\lambda|Z) \qquad (4.11)$$

where v_1m_1/x is distributed as x^2 with v_1 d.f. When v_1 is large, the

density of x is sharp around $x = m_1$, so that approximately,

$$p(\sigma_a^2|\lambda,Z) \propto Q(m_1,\sigma_a^2,\lambda|Z). \qquad (4.12)$$

It does not seem possible to express (4.9) and (4.12) in a simpler form. Numerical integration is required to obtain the appropriate normalising constants for the distributions (4.9), (4.10) and (4.12), which is done by using NAG routine F01GAF.

4.4.2 Examples

First of all examine the inferences concerning σ_a^2. For the examples in Table 4.4, Fig. 4.3 shows the posterior distribution of σ_a^2, conditional on $\hat{\lambda}$, calculated from the Kernel model (4.12) (solid line). Since in most cases, ν_1 is fairly large, the use of (4.10) would give nearly the same results. The dashed curve is obtained from the assumption that the random factor A_i is Normally distributed. Full details of the model under the Normality assumption about the random factor A_i can be obtained from Box and Tiao (1973).

To start with, I compared my result with the result of the example given by Tiao & Ali (1971). The Tiao data are positively skewed so it is not surprising that the distributions of σ_a^2 under the Normal and kernel model are different. However, the two distributions obtained under the kernel and Tiao model are quite different too, in a sense that the peak of the distribution of σ_a^2 from the kernel model is considerably lower and it has a longer tail than the Tiao's (see Fig. 4.3(c)). This difference could be due to the fact that Tiao's model makes use of the theoretical distribution,

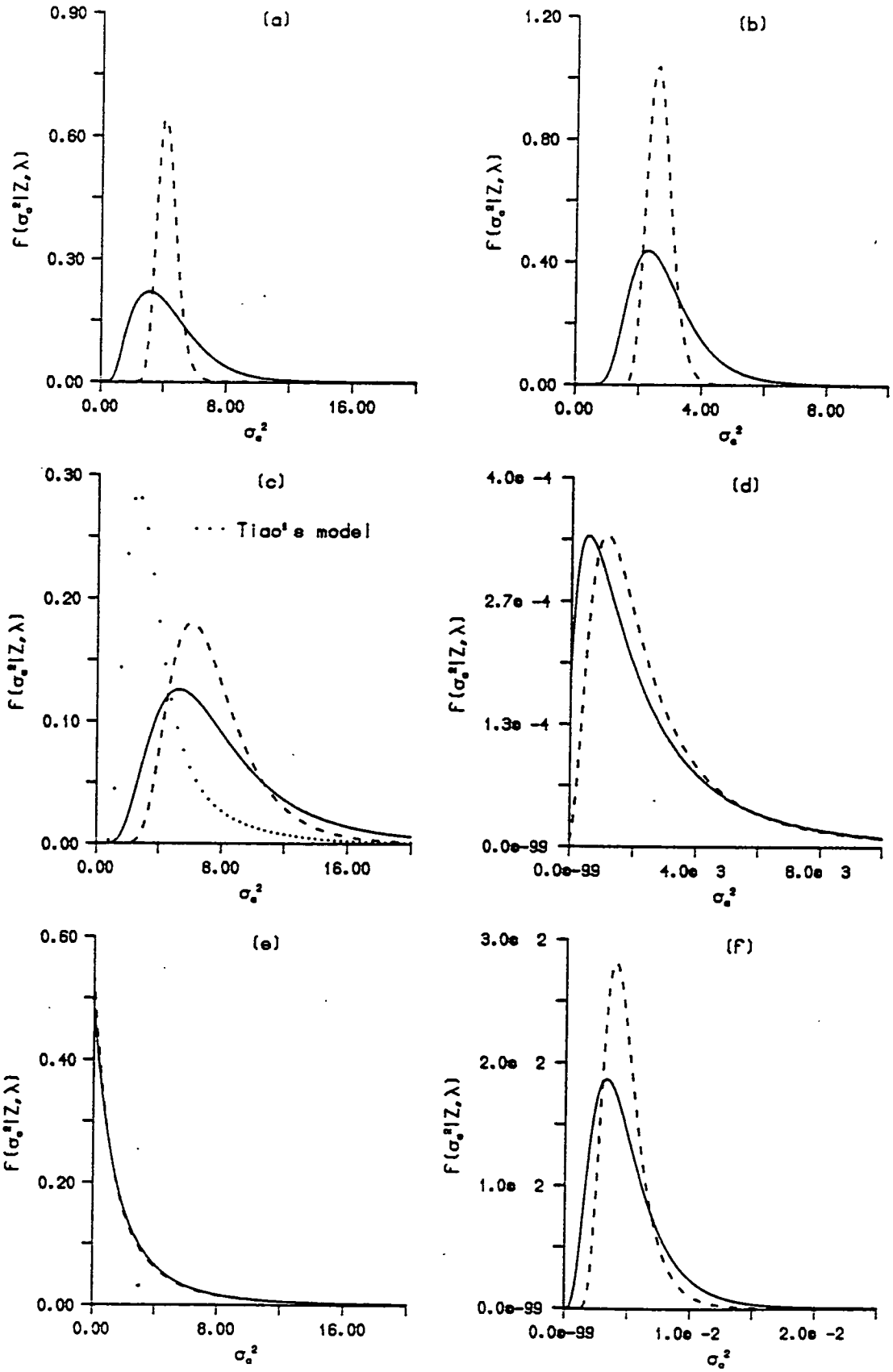


Fig. 4.3 Approximate posterior density of σ_e^2 for (a) Normal; (b) gamma; (c) Tiao; (d) Dye; (e) Generated and (f) cat data, —— kernel model; - - - - Normal model.

i.e. the true distribution of the underlying assumption of the A_i. The kernel model is data dependent so it will show more variability than the other models. Further the distribution of σ_a^2, under the Normal and kernel model for the Normal data, are also quite different. Again ,as shown in Fig. 4.3, the approximate posterior distribution of σ_a^2 under the kernel model has a longer tail than the distribution obtained from a Normal model, in general. The data for which the results under the kernel and Normal model are comparable, are the generated data taken from Box and Tiao (1973). These data were generated from the model (4.1) with within group variance σ_e^2 considerably greater than the between group variance σ_a^2 (see Table 4.4). So the ANOVA estimate for σ_a^2 is negative (see Table 4.5) and could well be treated as zero. Therefore the entire set of observations of this data set may be regarded as arising from a Normal distribution. This will explain why the approximate posterior distribution of σ_a^2 under the kernel and Normal models are so close to each other.

The posterior mode is chosen to be the Bayesian estimate for the parameter concerned. Another estimate which is used by other investigators is the posterior mean. However, in most cases the posterior distributions of the σ_a^2 are known to be skewed, so it does not seem sensible to use it. The posterior mode for each of the data sets is also shown in Table 4.5. The estimate for σ_a^2 is generally smaller under the kernel model than the Normal. However, the posterior mode is closer to the 'true' value under the kernel than the Normal model, when the data are from a Gamma distribution. To determine the reliability of the Kernel estimator for σ_a^2, I generated

a further 25 simulations from the model (4.1), of size 500 and 1000 for the Normal and gamma data respectively. Then I calculated the posterior mode under the kernel and Normal model. The position of the modes under the two models ranged from 0.84671 to 3.86937 and from 2.87588 to 4.68669 for the Normal data, and from 0.94141 to 2.42506 and 1.42067 to 2.83299 for the gamma data. Recall that the 'true' value of σ_a^2 is 4.0 and 2.0 for the Normal and gamma data, respectively. It appears that the estimates obtained under the kernel model vary a lot compared with those obtained under the Normal model for both Normal and gamma data. It shows the kernel estimator is slightly less reliable and tends to be downward biased. The complicated form of (4.12) makes if difficult to establish the reason for the occurrence of the large variabilities under the kernel model.

With regard to inference about σ_e^2, the results I obtained confirm the Tiao and Ali's finding that the estimates of σ_e^2 are insensitive to non-Normality of the distribution of σ_a^2. This can be seen in Table 4.7 which shows that the estimates for σ_e^2 under the Normal and kernel models are very similar even when the group means are not Normally distributed.

Table 4.7 Mode of the joint posterior distribution of σ_e^2 and σ_a^2 as described in Section 4.4.1 of (4.9)

Under Model		Data					
		Normal	Gamma	Tiao	Dye	Generate	Cats
Normal	σ_e^2	0.9948	0.9898	1.0976	2262.69	12.1328	$4.2\text{x}10^{-3}$
	σ_a^2	4.0909	2.5516	6.0784	1157.67	0.0000	$3.9\text{x}10^{-3}$
Kernel	σ_e^2	0.9948	0.9898	1.0976	2262.72	12.2927	$4.2\text{x}10^{-3}$
	σ_a^2	3.1175	2.2845	5.3005	616.201	0.0000	$3.2\text{x}10^{-3}$

Another way to estimate the parameters is with the mode of the joint posterior distribution of σ_e^2 and σ_a^2. Contour plots of the joint distribution of σ_e^2 and σ_a^2, conditional on λ and Z, for the examples shown in Table 4.4 are sketched in Figs. 4.4 - 4.9 under the Normal and Kernel model. And the resultant posterior modes of such joint distributions, are also obtained and tabulated in Table 4.8.

4.4.3 Vague prior for λ

Instead of estimating λ separately as was done in Section 4.2 using the Bayesian method, consider a vague prior for λ. The prior distribution of the parameters in (4.4) is taken to be

$$(\sigma_e^2)^{-1}(\sigma_e^2+J\sigma_a^2)^{-1}\lambda^{-1}.$$

Then the joint posterior distribution of the three parameters σ_e^2, σ_a^2 and λ can be expressed as (4.6) with

$$p(\lambda) \propto \lambda^{-1}.$$

Then the joint posterior distribution of σ_e^2 and σ_a^2 only, upon integrating out λ, is given by

$$p(\sigma_e^2,\sigma_a^2|Z) \propto \left[\frac{1}{\sigma_e^2}\right]^{(\nu_1/2)+1} \exp\left\{-\frac{\nu_1 m_1}{2\sigma_e^2}\right\} \left[\frac{1}{\sigma_e^2+J\sigma_a^2}\right]^{(n/2)+1} \times$$

$$\int_0^\infty \prod_{i=1}^{n} \sum_{k\neq i} \left[\frac{1}{\lambda^2}\right]^{\frac{1}{2}[1+(1/n)]} \exp\left\{-\frac{J(\bar{z}_{i.}-\bar{z}_{k.})^2}{2\lambda^2\sigma_{12}^2}\right\} d\lambda. \qquad (4.13)$$

Again numerical integration is required to solve the integral above. The posterior modes of the joint distribution of σ_e^2 and σ_a^2 are given in Table 4.8 for the examples shown in Table 4.4. Again the estimate for σ_e^2 remains unaltered whereas the estimate for σ_a^2

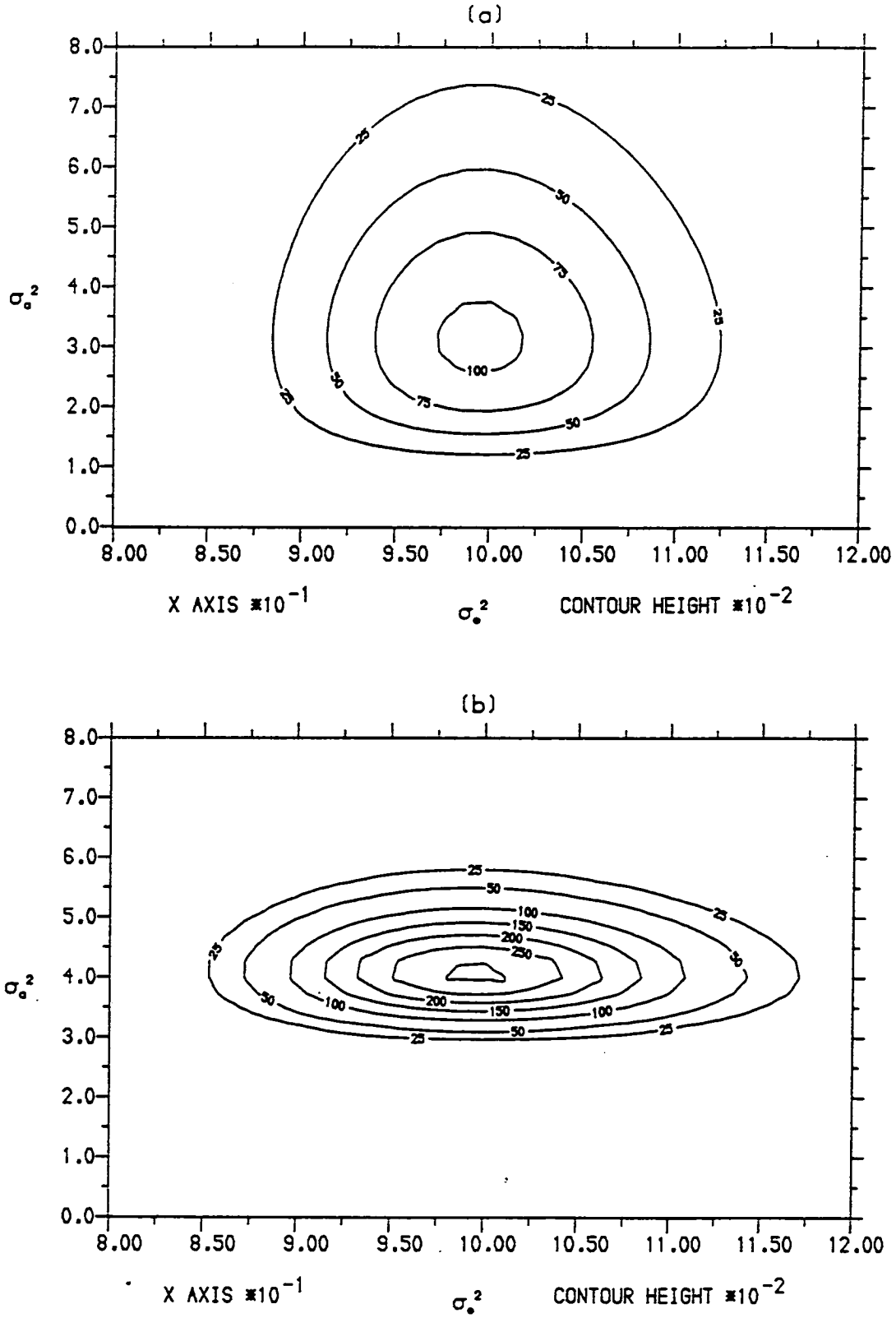


Fig. 4.4 Contours of the joint distribution of the variance components (σ_e^2, σ_a^2) under (a) Kernel model with λ = 0.50433 and (b) Normal model: the Normal data.

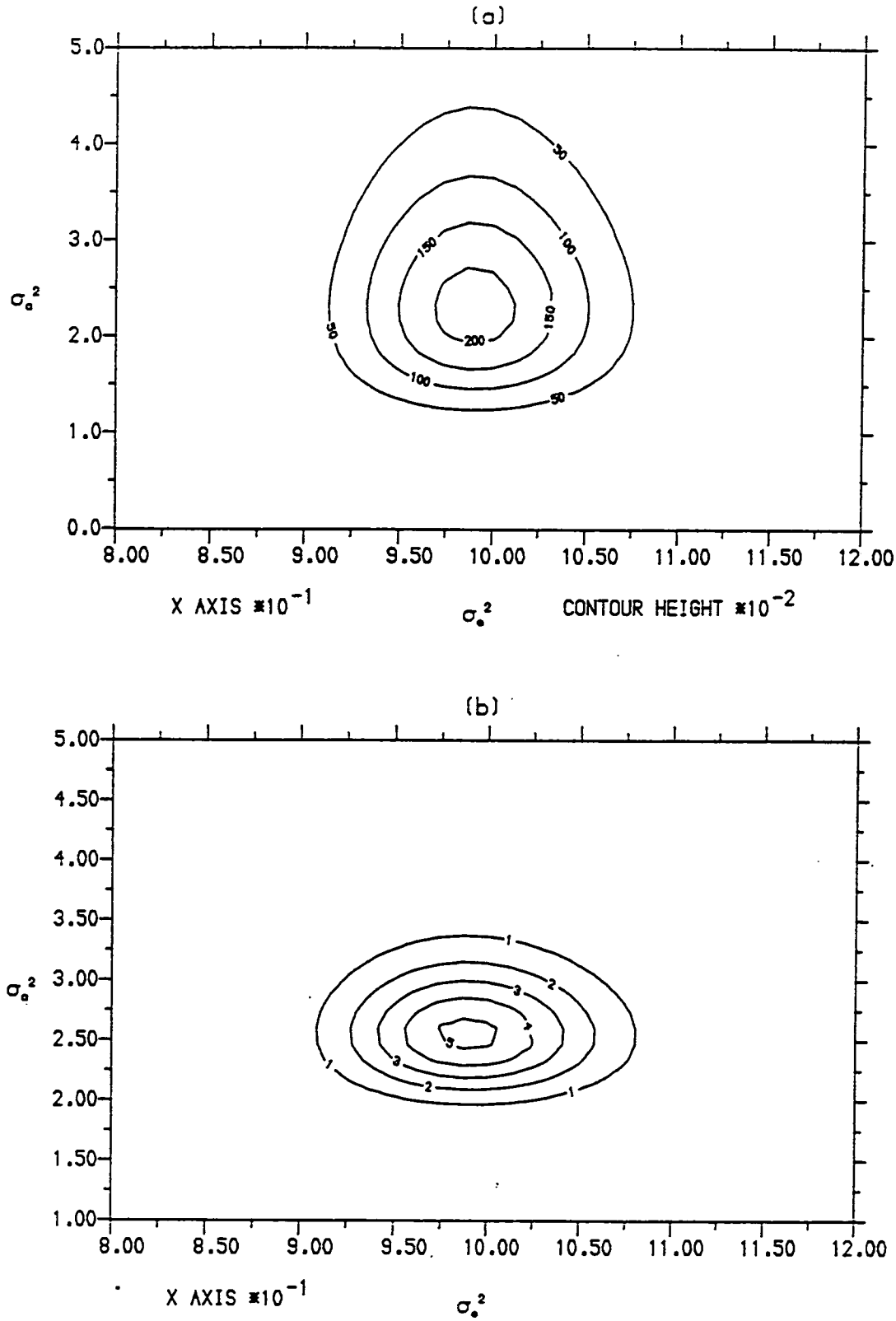


Fig. 4.5 Contours of the joint distribution of the variance components (σ_e^2, σ_a^2) under (a) Kernel model with λ = 0.38722 and (b) Normal model: the Gamma data.

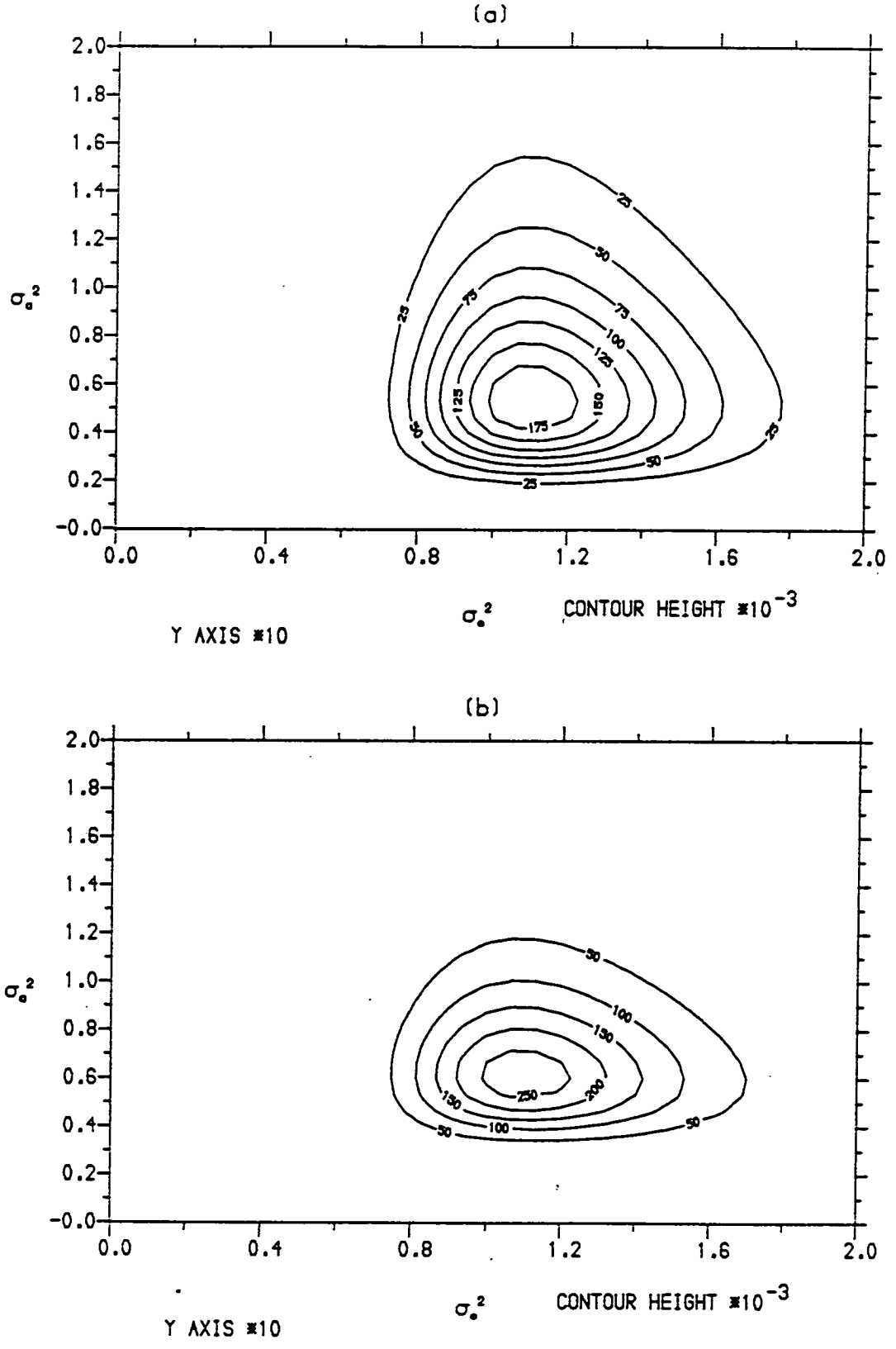


Fig. 4.6 Contours of the joint distribution of the variance components (σ_s^2, σ_e^2) under (a) Kernel model with λ = 0.50727 and (b) Normal model: the Tiao data.

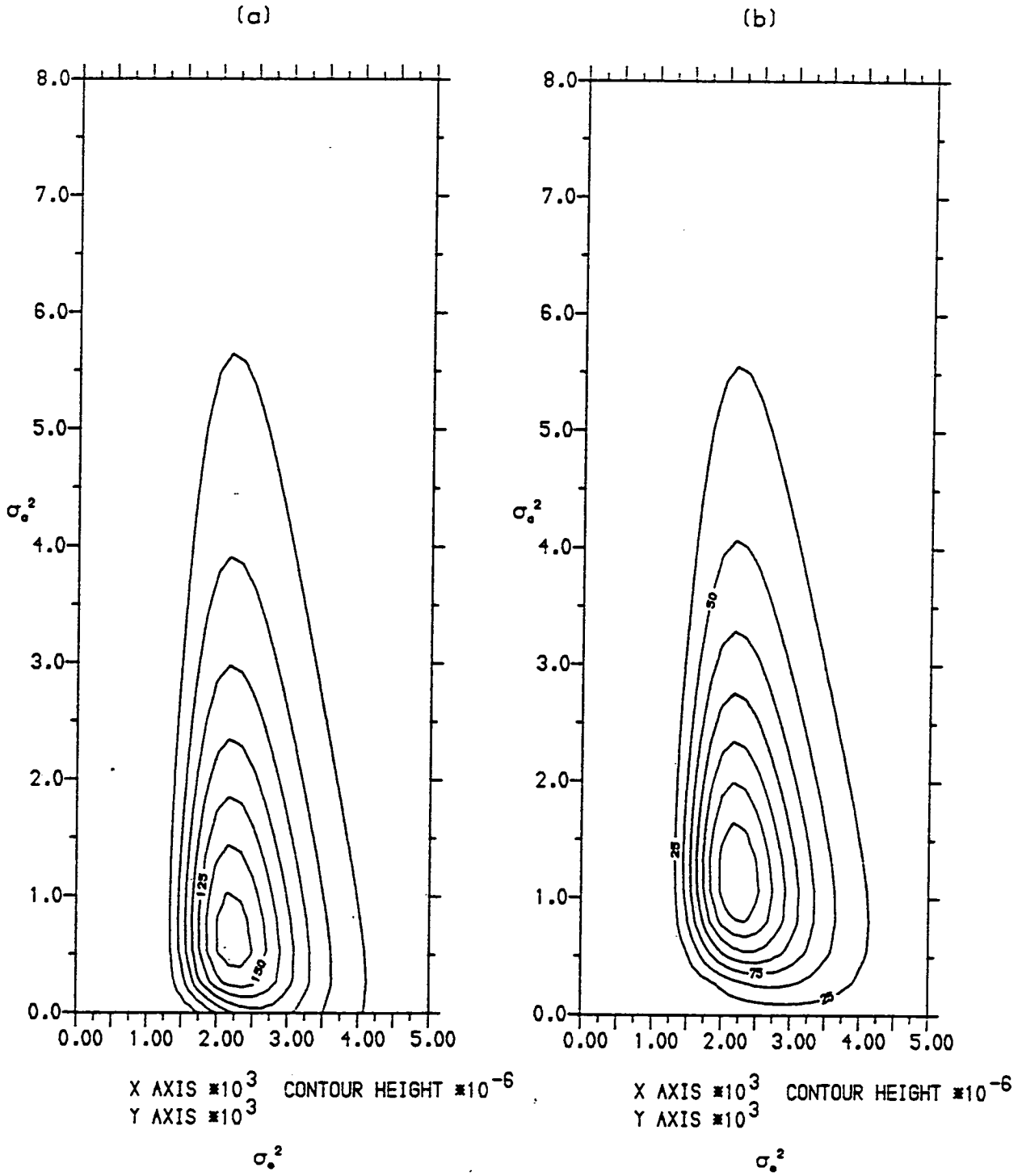


Fig. 4.7 Contours of the joint distribution of the variance components (σ_e^2, σ_e^2) under (a) Kernel model with λ = 0.95715 and (b) Normal model: the Dyestuff data.

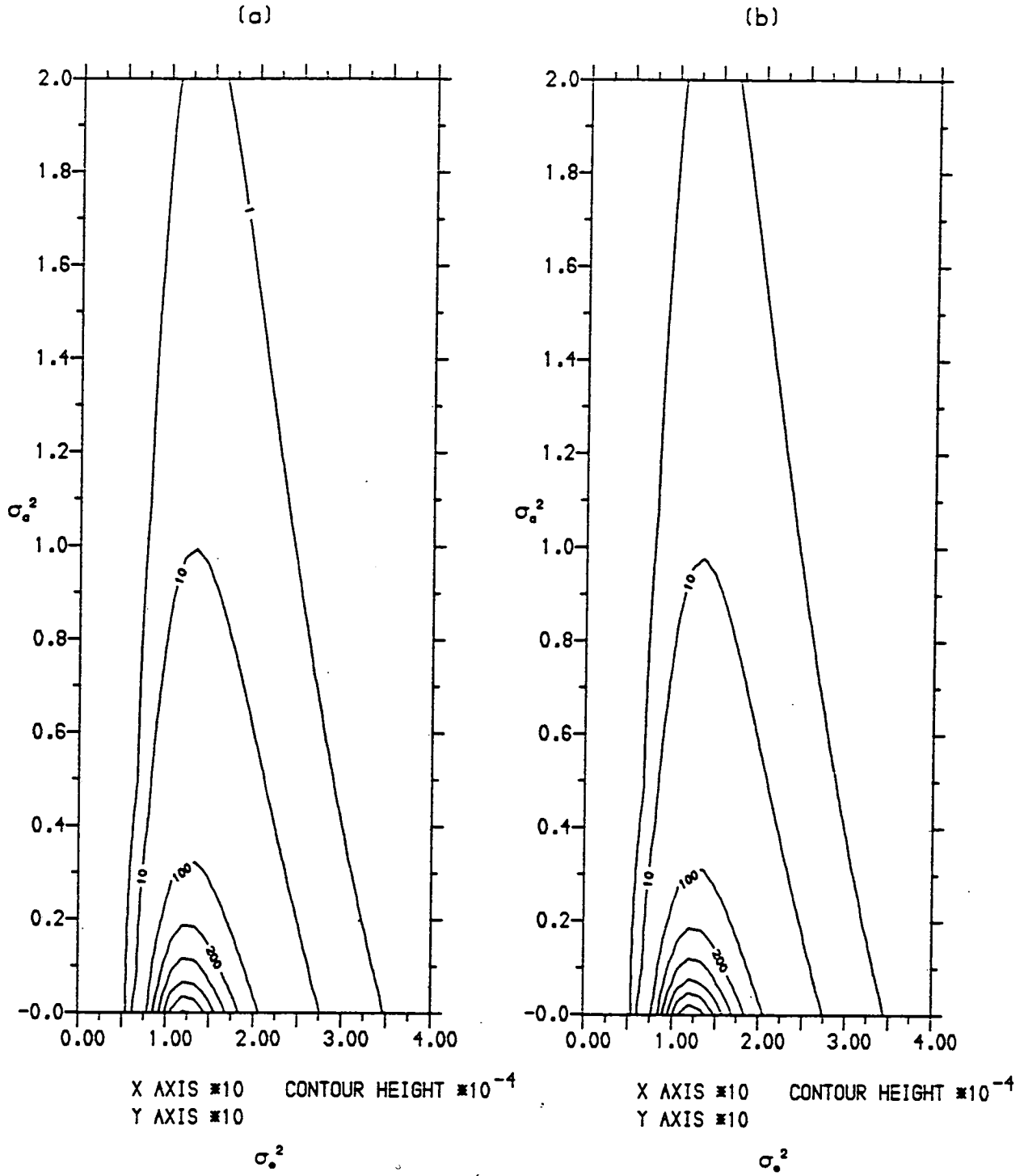
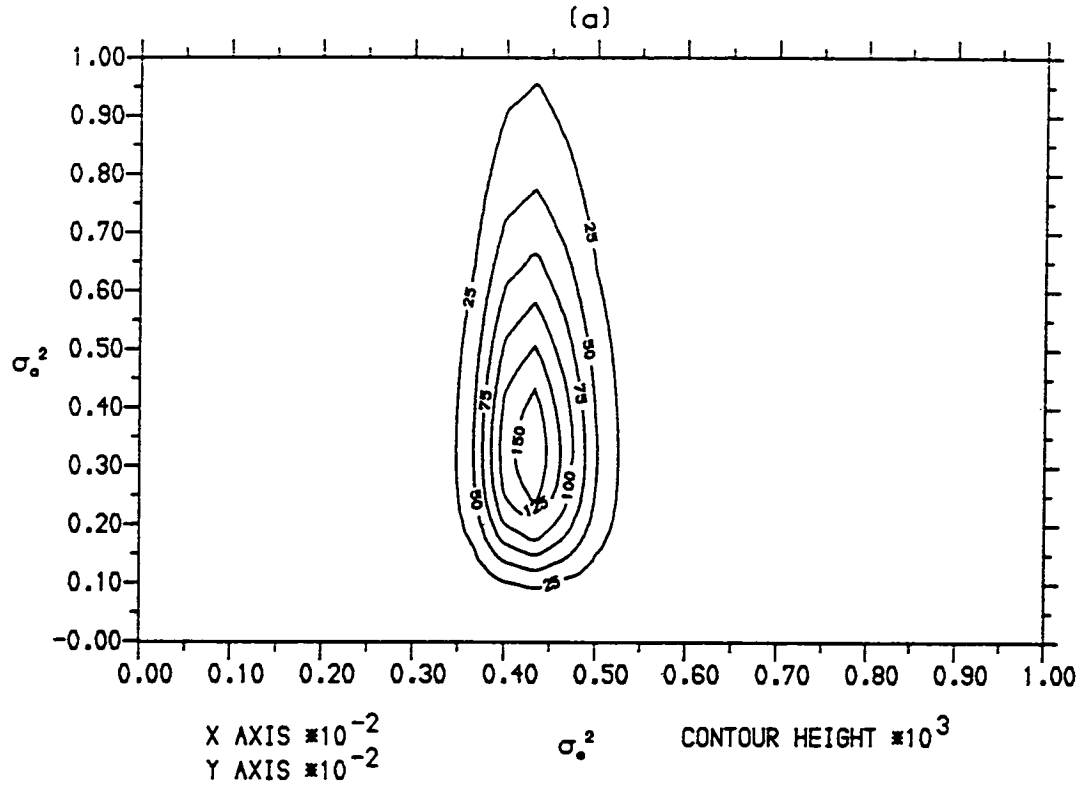


Fig. 4.8 Contours of the joint distribution of the variance components (σ_e^2, σ_a^2) under (a) Kernel model with λ = 1.01861 and (b) Normal model: the Generated data.



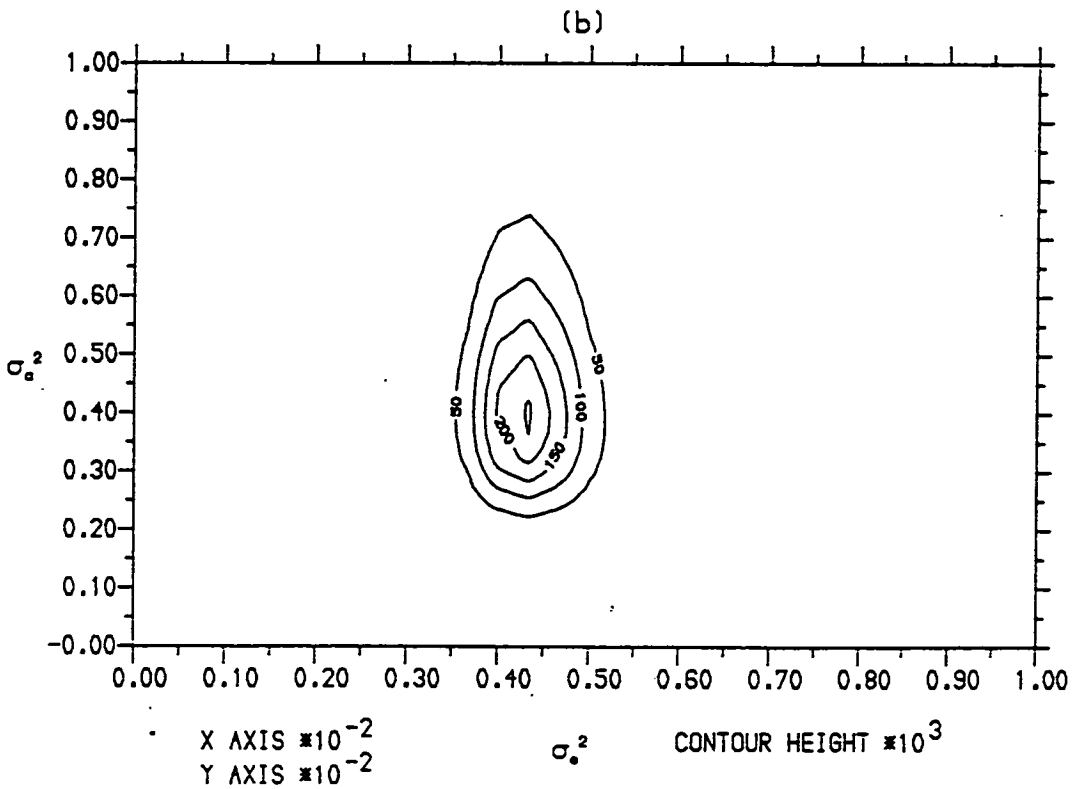


Fig. 4.9 Contours of the joint distribution of the variance components (σ_e^2, σ_a^2) under (a) Kernel model with λ = 0.78550 and (b) Normal model: the Cats data.

becomes zero. This is either due to the relationships between λ and σ_a^2 or is a result of the improper prior being used. Contour plots of the unconditional joint distribution of σ_e^2 and σ_a^2 using vague prior for both λ and σ_a^2 are shown in Fig. 4.10 (a) and Fig. 4.10 (b) for the Normal and Tiao data ,respectively.

Table 4.8 Mode of the joint distribution for σ_e^2 and σ_a^2 with vague prior for λ and σ_a^2.

	Data					
	Normal	Gamma	Tiao	Dye	Generate	Cats
σ_e^2	0.9617	1.0152	1.0521	2081.6	12.812	$.42\times10^{-3}$
σ_a^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

4.4.4 Informative prior for σ_a^2 and vague prior for λ

In Section 4.4.3, we saw that, because of the existence of a relationship between λ and σ_a^2, the introduction of a vague prior for both σ_a^2 and λ did not produce a reasonable result. Consider the adoption of an informative prior for the between group variance σ_a^2. Thus let the prior distribution of σ_a^2 be proportional to

$$(\sigma_a^2)^{-[(\alpha/2)+1]} \exp\left\{-\frac{\beta}{2\sigma_a^2}\right\}, \qquad (4.14)$$

where α and β are unknown. This prior is suggested by Hill (1965). The priors for σ_e^2 and λ are as in Section 4.4.3. Upon combining the likelihood function of (4.3) with the prior distibutions and integrating over λ we obtain

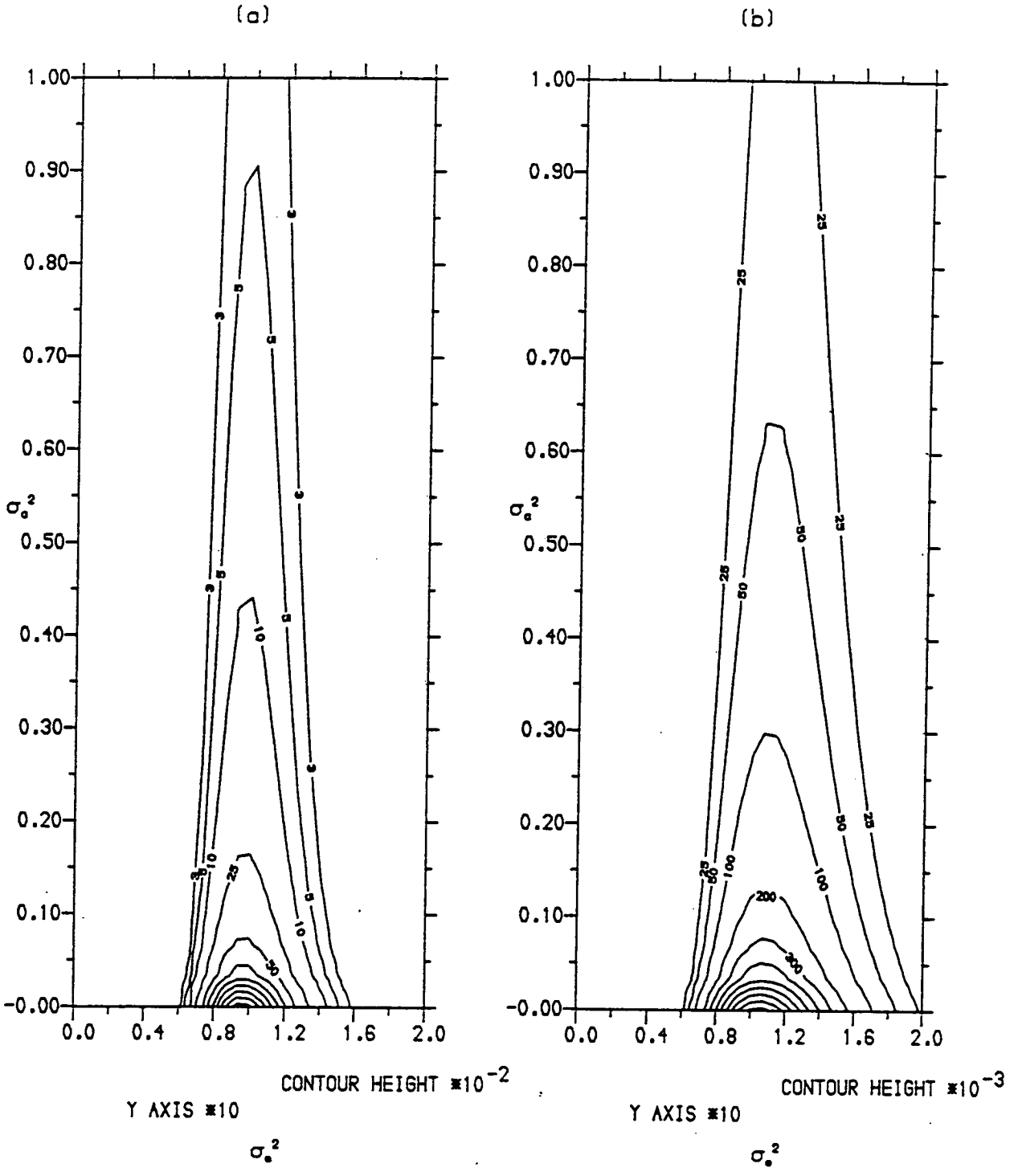


Fig. 4.10 Contours of the joint distribution of the variance components (σ_e^2, σ_a^2) under Kernel model with vague prior for λ as described in Section 4.3.3 for (a) the Normal and (b) the Tiao data.

$$p(\sigma_e^2,\sigma_a^2|Z) = \omega_1 \left[\frac{1}{\sigma_e^2}\right]^{(v_1/2)+1} \exp\left\{-\frac{v_1 m_1}{2\sigma_e^2}\right\} \left[\frac{1}{\sigma_{1z}^2}\right]^{(n-1)/2} \times$$

$$\left[\frac{1}{\sigma_a^2}\right]^{(\alpha/2)+1} \exp\left\{-\frac{\beta}{2\sigma_a^2}\right\} \times$$

$$\int_0^\infty \prod_{i=1}^{n} \sum_{k\neq i} \left[\frac{1}{\lambda}\right]^{1+(1/n)} \exp\left\{-\frac{J(\bar{z}_{i.}-\bar{z}_{k.})^2}{2\lambda^2\sigma_{1z}^2}\right\} d\lambda, \qquad (4.15)$$

where ω_1 is the normalising constant.

This formula is applied to Tiao's data and the Normal data. However, for the Normal data the number of groups has to be reduced to 20 groups and 5 observations in each group. This is because of the enormous computational time involved (the normalised constant is evaluated via numerical integration over the three parameters). The numerical integration is done by using the NAG routine F01GAF.

Results of the joint posterior modes of σ_e^2 and σ_a^2 for the Tiao's and Normal data are shown in Tables 4.9 and 4.10, respectively, given different values of α and β. The choicesof α and β are arbitrary. The contours of the joint posterior distribution of the variance components under (a) the Kernel and (b) the Normal models are plotted in Figs. 4.11 - 4.13 and Figs. 4.14 - 4.16 for the Normal and the Tiao data, respectively given three sets of selected hyperparameters. The results obtained under the Normal model are computed with the same prior distributions for σ_e^2 and σ_a^2 as in the kernel model, and in addition a vague prior for μ is also used. The nuisance parameter μ is integrated out to leave the joint posterior distribution of the variance components,

$$p(\sigma_e^2,\sigma_a^2|Z) = \omega_2 \left[\frac{1}{\sigma_e^2}\right]^{(\nu_1/2)+1} \exp\left\{-\frac{\nu_1 m_1}{2\sigma_e^2}\right\} \left[\frac{1}{\sigma_{1z}^2}\right]^{\nu_2/2} \times$$

$$\exp\left\{-\frac{\nu_2 m_2}{2\sigma_{1z}^2}\right\} \exp\left\{-\frac{\beta}{2\sigma_a^2}\right\} \left[\frac{1}{\sigma_a^2}\right]^{(\alpha/2)+1},$$

where ω_2 is the normalising constant, ν_2 is the between-group degrees of freedom and m_2 is the between-group mean square. Note that results obtained under the Normal model are shown in *italics*.

Table 4.9 Mode of the joint distribution of σ_e^2 and σ_a^2, with vague prior for λ and informative prior for σ_a^2; Kernel model: Normal data. Figures shown in *italic* are obtained under the Normal model.

α	β	σ_e^2	σ_a^2	$f(\sigma_e^2,\sigma_a^2)$
4	0.1	0.980325	0.016796	11480.916
		0.978558	*2.977850*	*1.018985*
4	0.5	0.975924	0.084378	122.1364
		0.978424	*2.996057*	*1.013918*
4	1	0.974513	0.167830	20.68925
		0.978262	*3.018777*	*1.007679*
4	2	0.973938	0.333805	4.515323
		0.977950	*3.064088*	*0.995503*
4	4	0.973880	0.666720	1.543080
		0.977374	*3.154226*	*0.972261*
4	5	0.973879	0.833352	1.227689
		0.977109	*3.199069*	*0.961142*
5	10	0.973879	1.428564	0.932395
		0.976682	*3.276051*	*0.966671*
5	20	0.973878	2.856775	0.472703
		0.974876	*3.697524*	*0.874121*
10	50	0.973878	4.166097	0.492735
		0.973756	*4.092371*	*0.891840*
10	100	0.973877	8.352436	0.248821
		0.971752	*5.770662*	*0.643084*
20	50	0.973878	2.272690	1.353088
		0.978133	*3.037165*	*1.355918*
20	100	0.973878	4.545188	0.673330
		0.973273	*4.322678*	*0.999269*
50	100	0.973879	1.923071	2.790225
		0.985052	*2.432314*	*2.152473*
50	200	0.973878	3.846030	1.303797
		0.974263	*3.894596*	*1.501666*
40	200	0.973878	4.761800	0.934263
		0.972887	*4.550114*	*1.187946*
80	400	0.973878	4.878009	1.307147
		0.972633	*4.730647*	*1.498238*

Table 4.10 Mode of the joint distribution of σ_e^2 and σ_a^2, with vague prior for λ and informative prior for σ_a^2: Kernel model: Tiao data. Figures shown in *italic* are obtained under the Normal model.

α	β	σ_e^2	σ_a^2	$f(\sigma_e^2,\sigma_a^2)$
4	0.1	1.098427	0.016671	7680.0894
		1.109413	*4.952661*	*0.374500*
4	0.5	1.097979	0.083377	77.48322
		1.109207	*4.971240*	*0.373461*
4	1	1.097764	0.166737	12.87128
		1.108953	*4.994432*	*0.372169*
4	2	1.097648	0.333389	2.789022
		1.108459	*5.040721*	*0.369604*
4	4	1.097621	0.666682	0.952257
		1.107520	*5.132921*	*0.364558*
4	5	1.097620	0.833340	0.757603
		1.107075	*5.178842*	*0.363530*
5	10	1.097619	1.428568	0.035404
		1.107150	*5.170998*	*0.372085*
5	20	1.097613	2.856450	0.291587
		1.013537	*5.605152*	*0.348914*
10	50	1.097606	4.164628	0.303701
		1.103046	*5.674948*	*0.386139*
10	100	1.097703	8.384539	0.153474
		1.095680	*7.403836*	*0.305404*
20	50	1.097617	2.272676	0.835103
		1.121568	*4.156973*	*0.587107*
20	100	1.097605	4.544087	0.415250
		1.104331	*5.498424*	*0.469876*
50	100	1.097618	1.923072	1.721904
		1.164478	*2.994077*	*0.990543*
50	200	1.097607	3.845796	0.804736
		1.114838	*4.537219*	*0.757649*
40	200	1.097606	4.761158	0.576437
		1.105732	*5.326740*	*0.607099*
80	400	1.097607	4.877668	0.806643
		1.106938	*5.193237*	*0.818796*

First consider the example concerning the Normal data: Note the ill-conditioning which occurred in the previous section has disappeared. The joint posterior mode for the parameter σ_e^2 with various choices of α and β is similar under the two models. It suggests that the joint posterior mode for σ_e^2 is insensitive to the

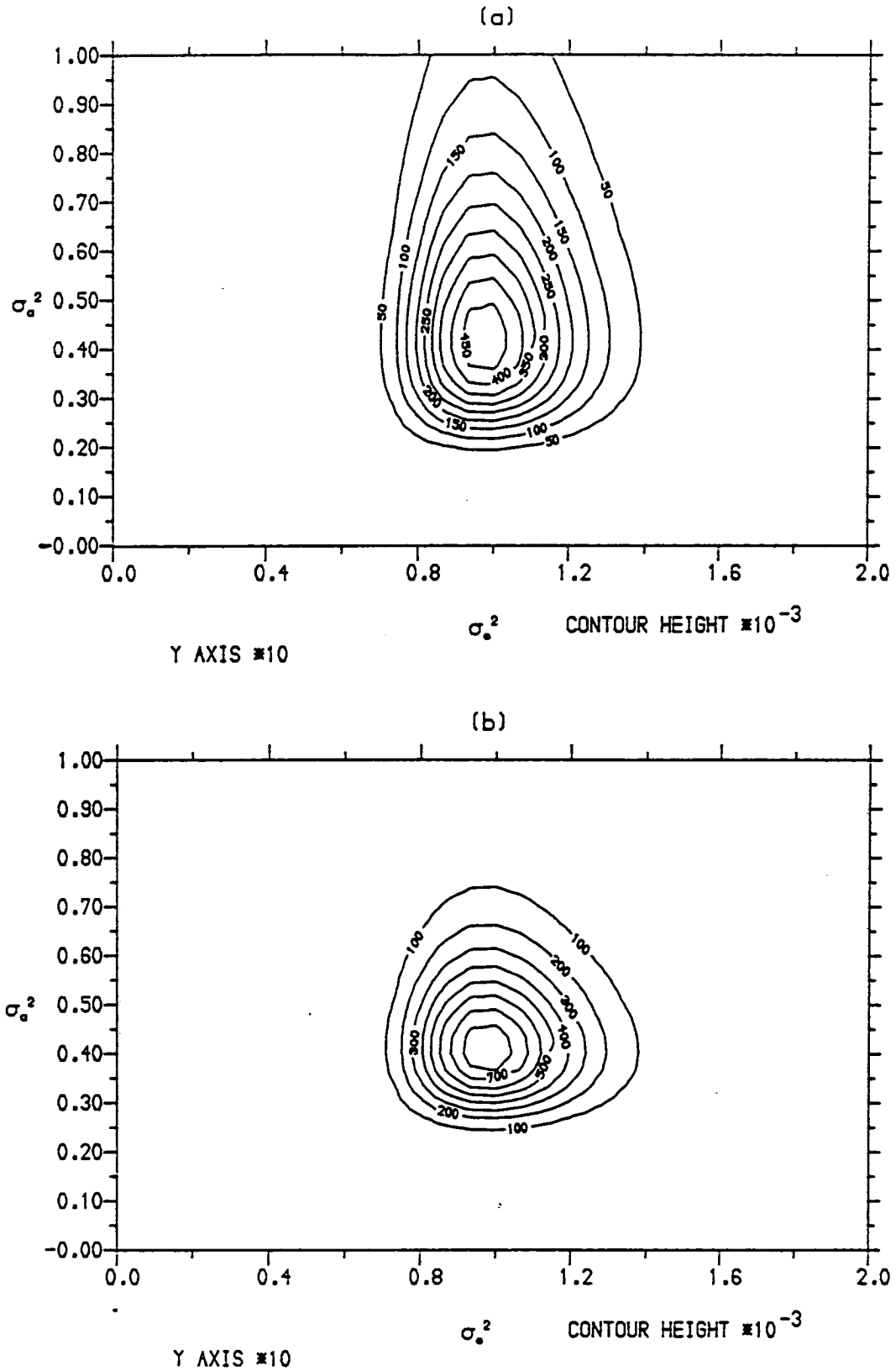


Fig. 4.11 Contours of the joint distribution of the variance components (σ_e^2, σ_a^2) under (a) Kernel model as in Section 4.3.4 with α = 10 & β = 50 and (b) Normal model; the Normal data.

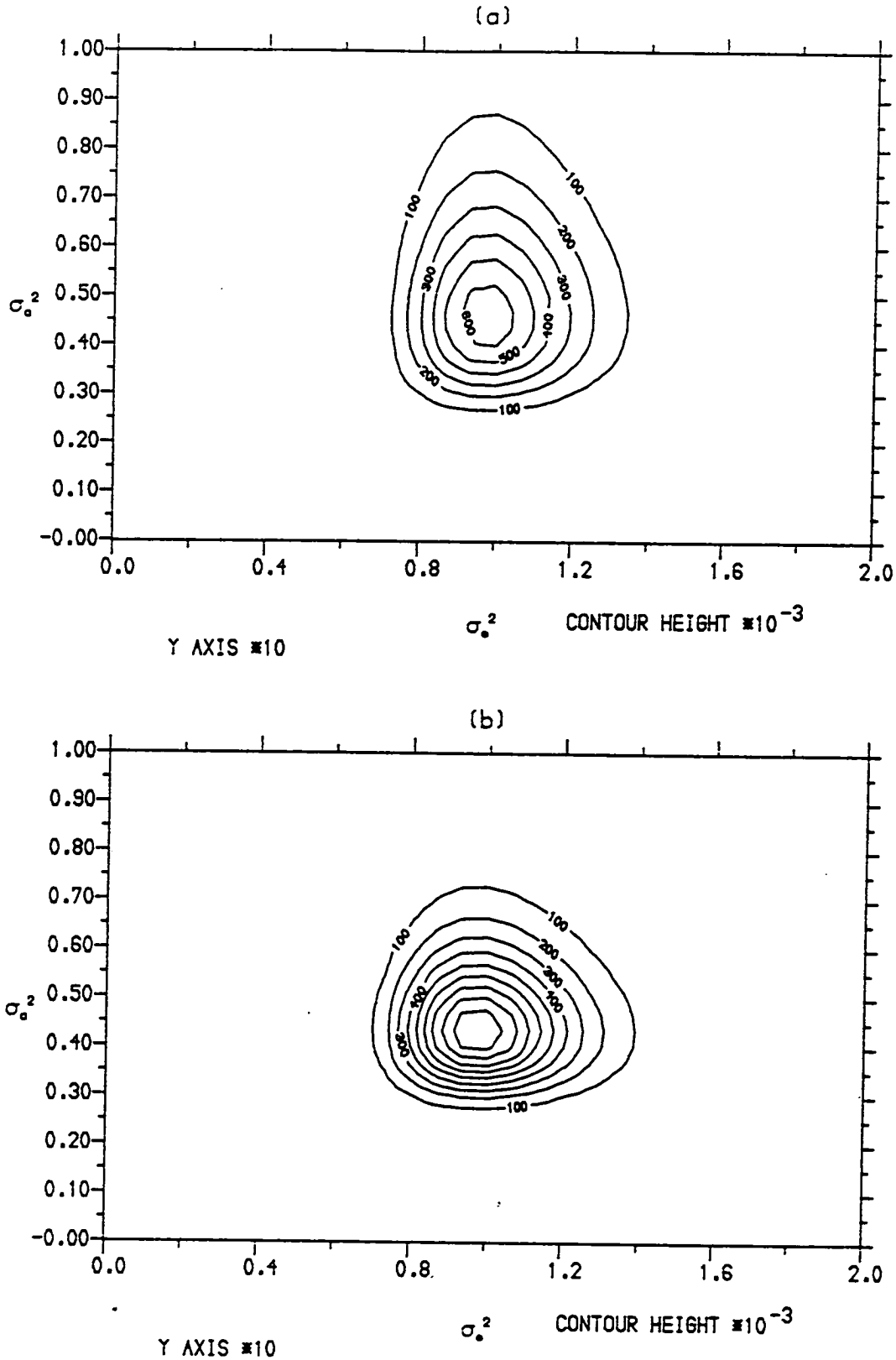


Fig. 4.12 Contours of the joint distribution of the variance components (σ_e^2, σ_a^2) under (a) Kernel model as in Section 4.3.4 with $\alpha = 20$ & $\beta = 100$ and (b) Normal model; the Normal data.

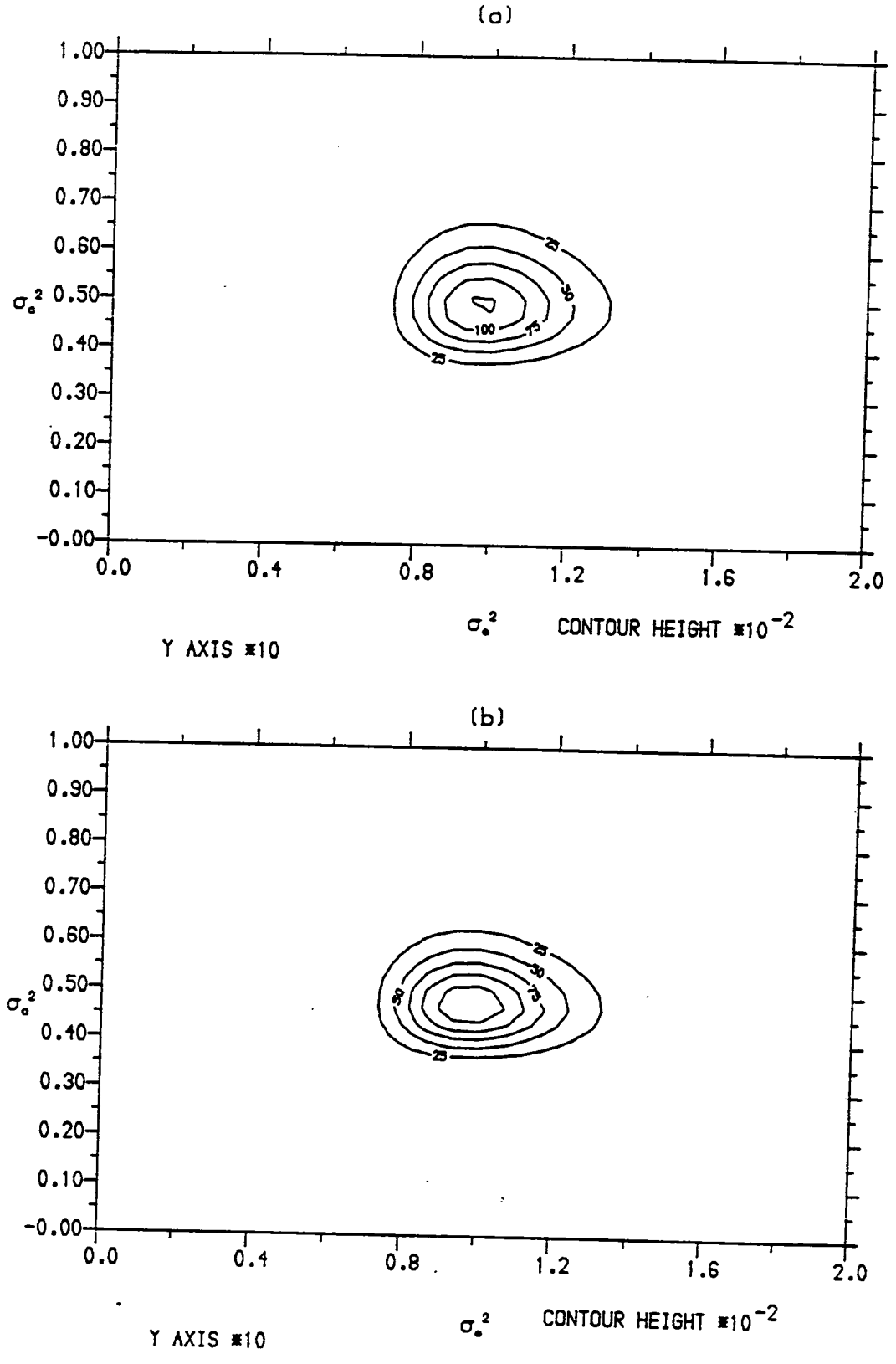


Fig. 4.13 Contours of the joint distribution of the variance components $(\sigma_e^2, \sigma_\beta^2)$ under (a) Kernel model as in Section 4.3.4 with α = 80 & β = 400 and (b) Normal model; the Normal data.

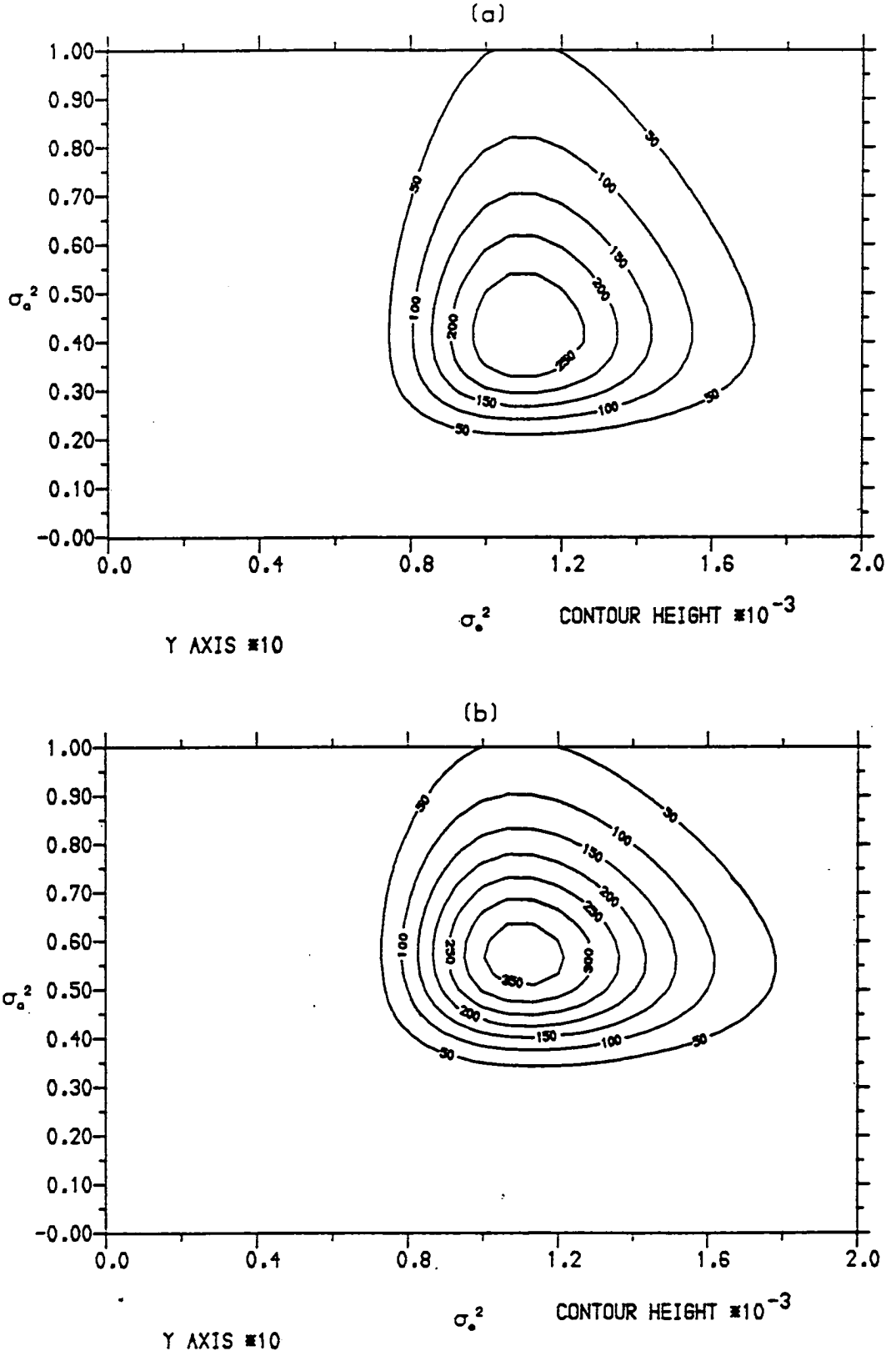


Fig. 4.14 Contours of the joint distribution of the variance components (σ_e^2, σ_a^2) under (a) Kernel model as in Section 4.3.4 with $\alpha = 10$ & $\beta = 50$ and (b) Normal model; the Tiao data.

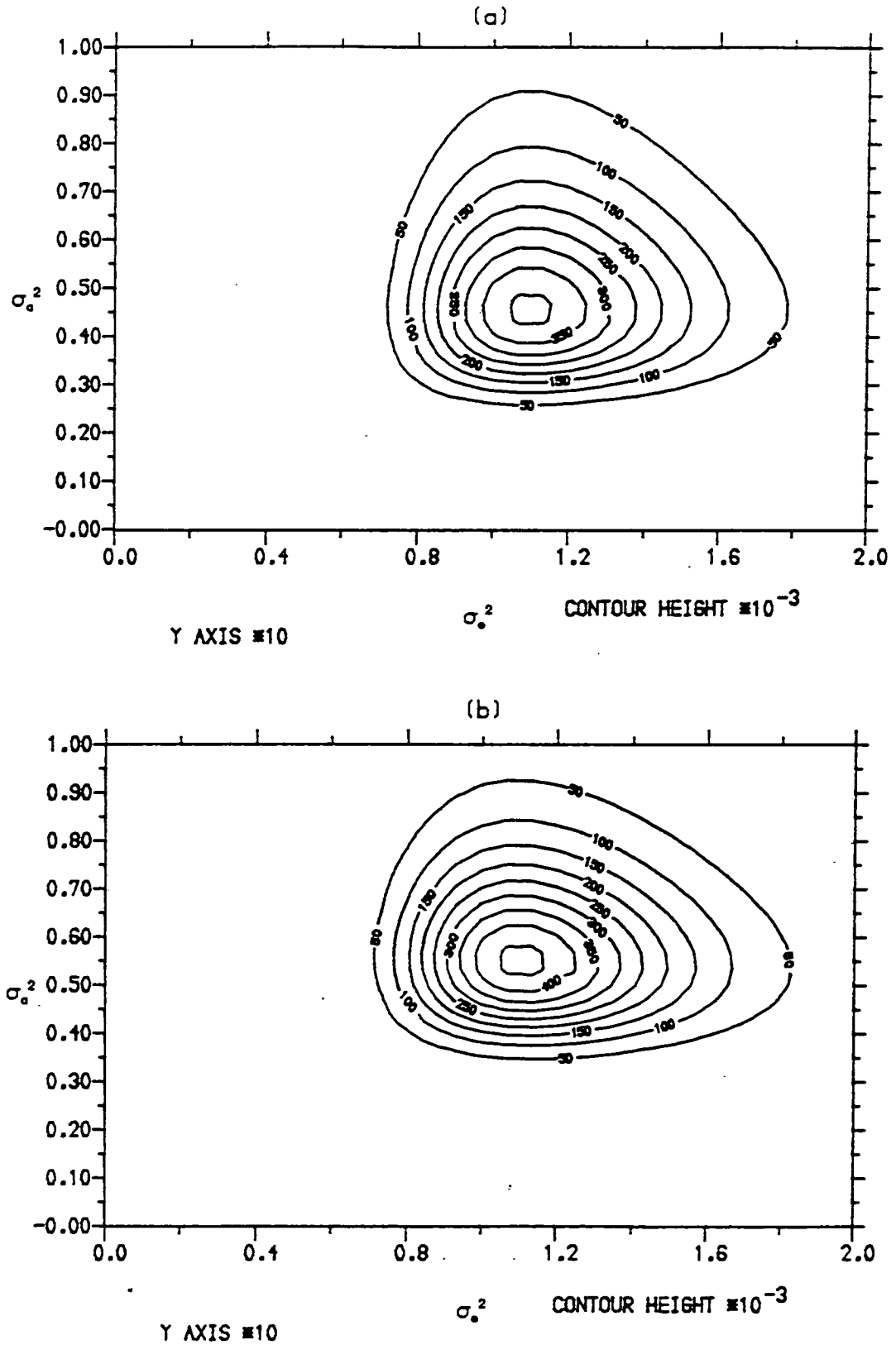


Fig. 4.15 Contours of the joint distribution of the variance components (σ_e^2, σ_a^2) under (a) Kernel model as in Section 4.3.4 with α = 20 & β = 100 and (b) Normal model; the Tiao data.

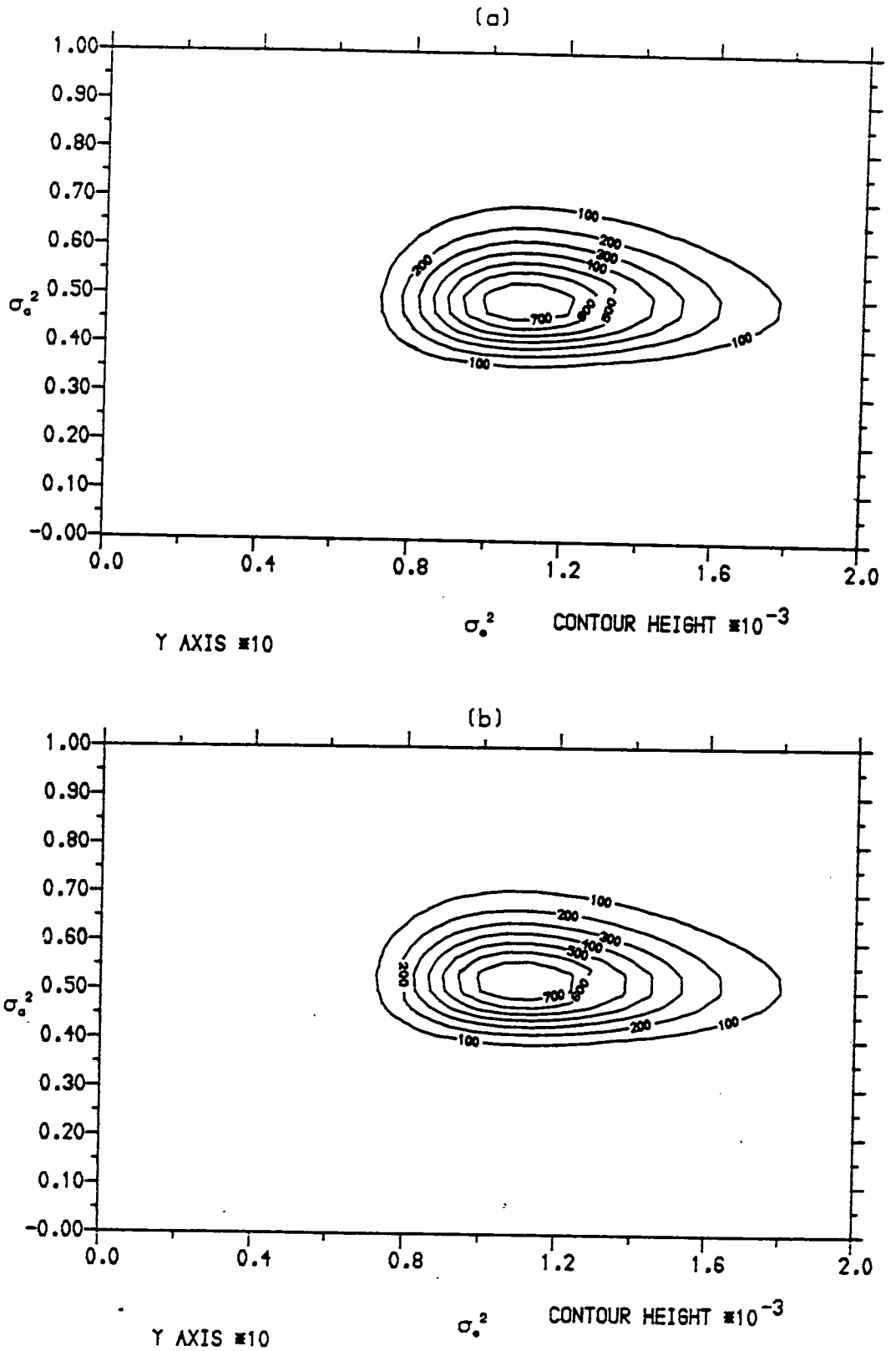


Fig. 4.16 Contours of the joint distribution of the variance components (σ_e^2, σ_a^2) under (a) Kernel model as in Section 4.3.4 with α = 80 & β = 400 and (b) Normal model; the Tiao data.

prior distribution for σ_a^2. However the posterior mode for σ_a^2 varies with α and β and is quite different under the two models. The posterior modes for σ_a^2 under the Normal and Kernel models are closest when α and β equal to 10 and 50, respectively. Also notice that the shapes of the joint distribution of the variance components are dissimilar. When α and β equal to 20 and 100, respectively, which are twice the former chosen values, the posterior modes under the two models are still close. The shape of the joint posterior distribution changes in the Y-direction and becomes very similar under the two models when $\alpha = 80$ & $\beta = 400$.

4.5 Discussion

The result obtained from Section 4.3 shows that the ML estimators for (σ_e^2, σ_a^2) obtained from the Kernel model by estimating the smoothing parameter λ in advance are equivalent to the ANOVA estimators.

In Section 4.4, the results show that the modes of the joint or marginal posterior distributions of the variance component parameters (σ_e^2, σ_a^2) from the noninfomative prior (4.7) are viable estimators, though the posterior distribution of σ_a^2 under the kernel model shows more uncertainty. This is understandable as mentioned before since the kernel method is data dependent. Although one should not read too much into one or two results, the estimates obtained under the kernel model, for the Tiao and Gamma data, are closer to the 'true' value. This study also suggests the Bayesian estimators, under either the Normal or the Kernel model, are generally better than the ML and the ANOVA estimators.

One problem which is not completely resolved is the relationship between the smoothing parameter, λ and the between group variance, σ_a^2. I suspect that during the computation of the posterior distribution of σ_a^2 by numerical integration, some values of σ_a^2 have to be assigned and it is taken to be ranging from 0.0 upward. This affects the density function of (4.2) with λ estimated by the M.L. leave-one-out method. Then I investigated the differences between the density function of the group means under the Normal and kernel model. The density function, (4.2), for the Normal data is plotted in Fig. 4.17 with fixed λ and σ_e^2, and σ_a^2 taken to be ranging from 0.0 and 10.0 in a step of 2.0. It can be seen that the density function, (4.2), becomes extremely rough as σ_a^2 is close to zero, in contrast to the density function (see Fig. 4.18) under the Normal model. And, as σ_a^2 approaches 10, the differences between the density function under the two models is negligible. So the kernel density of the group means is no longer a smooth function when σ_a^2 is around zero. This may explain why the posterior modes of σ_a^2 under the kernel model are generally smaller than those under the Normal model. Further investigation into this area is needed.

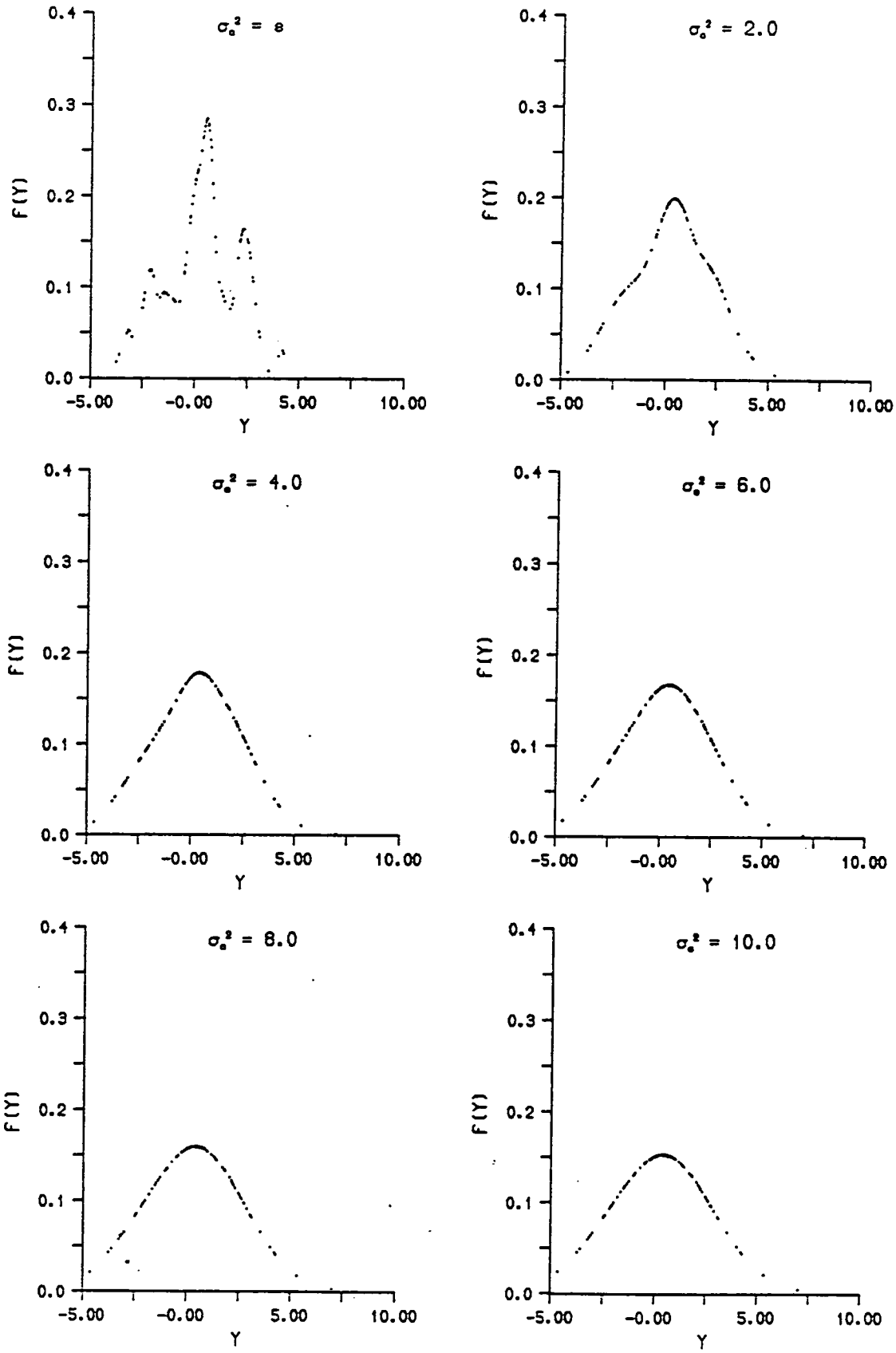


Fig. 4.17 Density of the Normal data given different values of σ_e^2 -- Kernel Model. $\varepsilon = 10^{-10}$

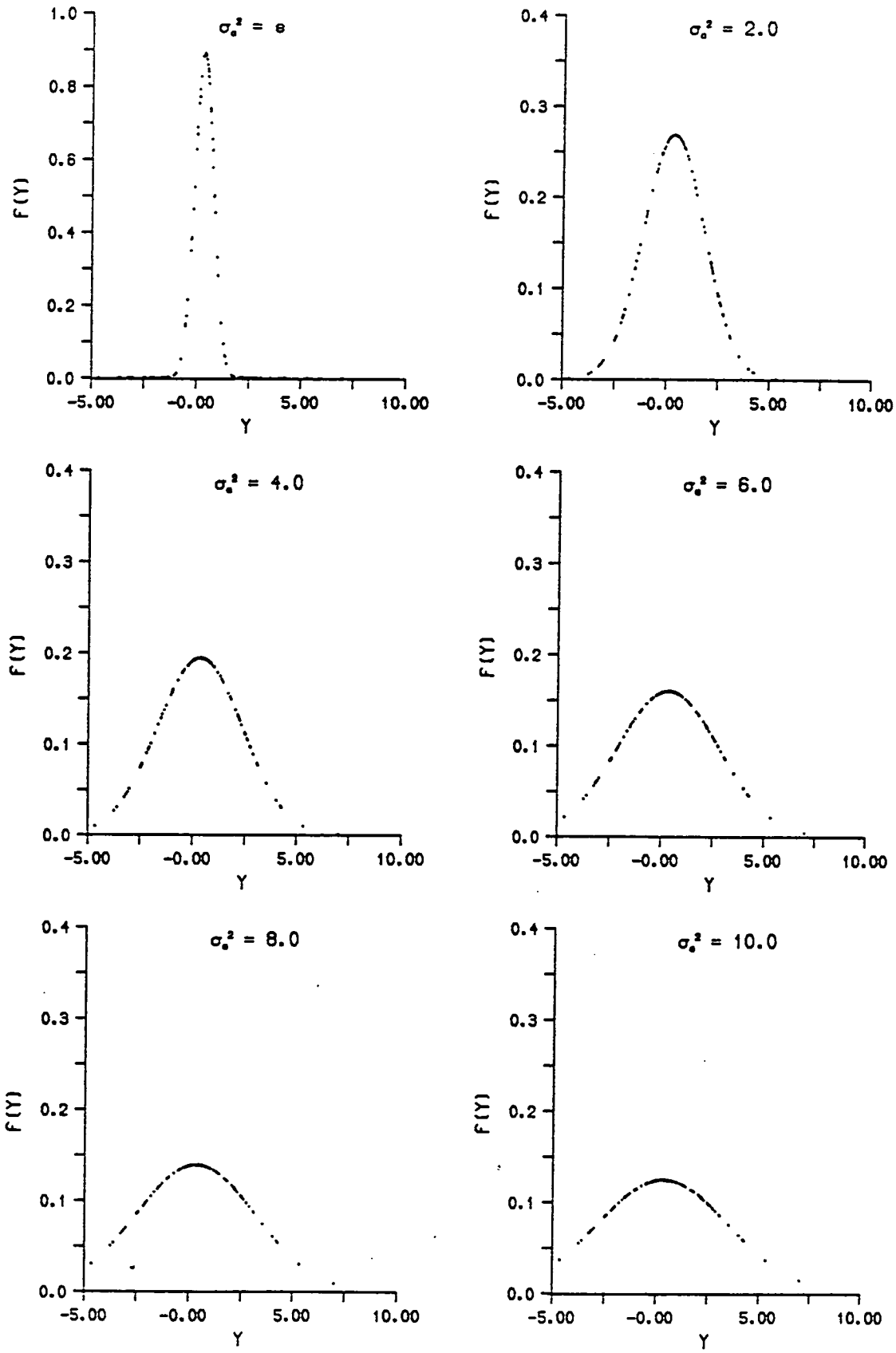


Fig. 4.18 Density of the Normal data given different values of σ_e^2 -- Normal Model. $\varepsilon = 10^{-10}$

CHAPTER 5

MODELLING THE BAYES' FACTOR FOR A PARTICULAR FORM OF MIXTURE DATA

5.1 Introduction

The problem considered in this chapter is motivated by a set of data which has a finite probability at a particular value and is a continuous positive random variable otherwise. For the example discussed in this Chapter, the particular value is 'zero'. The modelling of the Bayes' factor in a forensic context is to take into account this special feature. Two models are suggested - one is adapted from Evett el al (1987), which does not utilise the random structure in the data. The other one is an adaptation of the model developed in Chapter 3 to modelling this particular feature in measuring the weight of evidence.

5.2 Distribution and structure of the mixture data

The data set in question is similar to the cat hairs data (see Section 1.3) which consists of hair measurements, namely hair width and medullary fraction (medullary width/hair width). However, some hairs obtained from dogs have no medulla in which case the medullary fraction takes the value zero. As in Chapter 3, we consider a univariate problem so the hair width variable is not included in the formulation of the problem.

So we are faced with the problem of a distribution specified by

i) a non-zero probability that the variable assumes a zero value, together with

ii) a conditional distribution for the positive value of the variable.

Aitchison (1955) discussed unbiased estimators for the parameters concerned with this type of data. Our aim is not so much concerned with obtaining unbiased estimators for the unknown parameters but to model the distribution of this type of data.

Such problems lead us to consider a random variable U with the following properties. There is a non-zero probability θ that U is zero and hence a probability $1-\theta$ that U is not. The distribution of the non-zero part of U is in this chapter that of a positive variable, either continuous or discrete. Thus we may write:

$$\Pr\{U = 0\} = \theta, \quad \Pr\{U > 0\} = 1 - \theta.$$

and for the continuous case,

$$P\{U \subset (u,u+du) \mid u>0\} = g(u)du.$$

where g(u) is the conditional probability density function; and so

$$P\{U \subset (u,u+du)\} = (1-\theta)g(u)\,du, \quad u>0.$$

Suppose we have a random sample S which consists of t zero values and (n-t) other 'positive' values $u_1, u_2, \ldots, u_{n-t}$. And suppose α and β are the mean and variance (respectively) of the non-zero positive part of U. Then assume an unbiased estimator $a_{(n-t)}$ of α exists for a sample of n-t values. If the distribution of U depends on parameter β in addition to θ and α, then the likelihood function L of the sample may be written in the form

$$L(S|\theta,\alpha,\beta) = \binom{n}{t}\theta^{t}(1-\theta)^{n-t} \times h(a_{n-t}|\alpha,\beta) \times f(S|\beta)$$

where h and f are probability density functions containing the sample values only in the form of t/n, $a_{(n-t)}$ and S. Thus t/n, $a_{(n-t)}$ and S are jointly sufficient estimators of θ, α and β, respectively. See Appendix 5 for details.

5.3 Estimation of the Bayes' factor: single hair problem

Recall from Chapter 3 that the Bayes' factor is a ratio of two probability (density) functions. The numerator of the Bayes' factor is a predictive distribution of Y given X and the denominator is an unconditional marginal distribution of Y. Unlike Chapter 3 where we allowed more than one observation in the recovered data Y, here we consider only one observation from the recovered data, i.e. r=1. The model developed here can easily be extended to the case where there is more than one observation from the recovered data.

In a preliminary investigation of modelling the Bayes' factor of this particular mixture data, we adapt a model originated from Evett et al (1987), who estimated the denominator of the Bayes' factor by a kernel form density, i.e no distributional assumptions were made about the recovered data Y given $\bar{C}$. Effectively their model involved using all individual observations in the training data, hence the random structure in the training data is not utilised. Later in this section a more thorough treatment will be discussed to make use of the random structure in the training data in modelling this type of mixture data. This model is known as the Kernel model (see Chapter 3). Before we proceed, some notation and assumptions are required

modelling the Bayes' factor.

5.3.1 Notation and assumptions

In addition to the notation used for the cat data analysis in Chapter 3, let T(x) and T(y) be random variables representing the number of zeros found in the sample X and Y and let t(x) and t(y) denote the realisation of T(x) and T(y); $\bar{x}_{m-t(x)}$ denotes sample means of the m-t(x) non-zero positive observations from the control sample; and $g(.|\alpha,\beta)$ denotes a density function with parameters α and β. It is assumed that g is defined by two parameters. Further note that if $g(.|.)$ is a Normal density then the parameters α and β are the mean and variance respectively.

When deriving the numerator of the Bayes' factor, we assume the non-zero positive observations are Normally distributed with unknown mean μ and known variance σ^2. If X consists of m items of which m-t(x) are not equal to zero, then the sufficient statistic for the true mean μ (conditional on $x > 0$) is the sample mean $\bar{X}_{m-t(x)}$. This is also Normally distributed about the unknown true μ with variance $\sigma^2/[m-t(x)]$. Given one observation from the recovered data Y, and if Y is not equal to zero, then under C, Y is also Normally distributed with μ and variance σ^2. Using an informative prior for the unknown mean μ, we assume for the present that μ is also Normally distributed with mean ξ and variance η^2. The parameters ξ and η^2 are so-called hyper-parameters and are assumed known. Note that Evett et al (1987) used vague priors for the group population mean μ and within-group variance σ^2. Here we assume throughout this chapter that the within-group variance σ^2 is known.

5.3.2 Preliminary analysis - ECA model

Here we modify the ECA model to incorporate the possibility that Y and some, or all, of the X's may take a value zero. The probability density functions of X and of Y are written in a two-fold definition. We now consider the denominator and numerator in turn.

A. The denominator of the BF

Under $\bar{C}$, that is the situation in which the recovered hair comes from an unknown source other than the control data, the distribution of the random variable Y can be summarised in the following two-fold definition,

$$f(Y|\bar{C}) = \begin{cases} \Theta, & y=0 \qquad (5.1) \\ (1-\Theta)\times k(y) & y\neq 0. \qquad (5.2) \end{cases}$$

In (5.1) and (5.2), Θ is the probability that Y is zero. Suppose that the training data $Z = \{z_1,\ldots,z_N\}$ where N is the total sample size of the training data Z. And let $Z^* = \{z_1^*,\ldots,z_{N-t(z)}^*\}$ be the training data after extracting the zero values from Z, where t(z) is the total number of zeros in the training data Z. Modifying the ECA model, the density function, $k(y)$, in (5.2) may be written as

$$k(y) = \frac{1}{N-t(z)} \sum_{\ell=1}^{N-t(z)} \frac{1}{(2\pi)^{\frac{1}{2}}\lambda s} \exp\left\{ -\frac{(y-z_\ell^*)^2}{2\lambda^2 s^2} \right\},$$

where λ is the 'standardised' smoothing parameter,

s^2 is the sample variance of the altered training data Z^*.

B. The numerator of the BF

Here we consider the formulation of the numerator of the Bayes'

factor for the following two cases:

Case (i) : Both X and Y consist of one measurement only (i.e. r=1; m=1), and

Case (ii) : Y consists of one measurement but X has m measurements with a non-zero probability there are t(x) zeros in the sample.

5.3.3 Case (i) Both X and Y consist of one observation

Random variables X and Y can take the values zero or non-zero. Thus X and Y can be summarised by the random variables T(x) and T(y) which denote the number of zeros in the sample. In this particular case, T(x) and T(y) can only take the values zero or one. Thus, T(x) and T(y) have a binomial distribution with parameter θ, the probability of obtaining a zero (i.e $Pr(T(x)=1)=\theta$ etc). There are four possibilities from this case, that is (a) both x and y are zero; (b) y is zero but x is not; (c) y is not zero but x is and (d) both x and y are not zero.

First of all let us specify a prior distribution for the unknown parameter θ. The conjugate prior for θ is a beta function with parameters a and b, namely

$$f(\theta) = [Be(a,b)]^{-1}\ \theta^{a-1}\ (1-\theta)^{b-1} \qquad (5.3)$$

where a and b are assumed to be known. From Section 5.3.1, assuming the between group random factors are identically Normally distributed, then the probability density function $f(\mu)$ can be specified as

$$f(\mu) = \frac{1}{(2\pi)^{\frac{1}{2}}\eta} \exp\left\{-\frac{(\mu-\xi)^2}{2\eta^2}\right\}. \qquad (5.4)$$

Let us now consider these four possibilities in turn.

(a) Both x and y are zero

We may consider the numerator of BF as a discrete predictive distribution with random variables T(x) and T(y). From Table 2.2 of Aitchison and Dunsmore (1975), the predictive distribution of T(y) given T(x) is a Beta-Binomial distribution. Hence $Pr(T(y)=1|T(x)=1,C)$ is

$$\frac{Be(a+2,b)}{Be(a+1,b)} = \frac{a+1}{a+b+1}. \qquad (5.5)$$

And the denominator of BF is defined in (5.1).

(b) y is zero but x is not

Here we have,

$$Pr(T(y)=1|C) = \theta$$

and

$$f(x|C) = (1-\theta)\times g(x|\mu,\sigma^2),$$

where $g(x|\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\{(x-\mu)^2/(-2\sigma^2)\}$. Combining with the priors, which are assumed independent namely $f(\theta)\times f(\mu)$, and integrating over θ and μ the numerator of BF can be written as

$$Pr(T(y)=1|X,C) = \frac{a}{a+b+1}. \qquad (5.6)$$

The denominator of BF is given as in (a).

(c) y is not zero but x is zero

Here we have the opposite case of (b), the density of Y under C is

$$f(y|C) = (1-\theta) \times g(y|\mu,\sigma^2).$$

And the probability of X being zero is given by

$$Pr(T(x)=1) = \theta.$$

Then the numerator of BF can be written as

$$f(y|t(x)=1,C) = \frac{b}{a+b+1} \times g(y|\xi,(\eta^2+\sigma^2)). \qquad (5.7)$$

The denominator of BF is given in (5.2).

(d) Both x and y are not zero

Now both X and Y have the density function described earlier, they are

$$f(x|C) = (1-\theta) \times g(x|\mu,\sigma^2)$$

and

$$f(y|C) = (1-\theta) \times g(y|\mu,\sigma^2).$$

Then, after some simplification , the numerator of BF is given as

$$f(y|x,C) = \frac{b+1}{a+b+1} \times \frac{g(x-y|0,\sigma_1^2) \times g(w|\xi,\sigma_2^2)}{g(x|\xi,\sigma_3^2)} \qquad (5.8)$$

where $\sigma_1^2 = 2\sigma^2$,

$\sigma_2^2 = \eta^2+\sigma^2/2$,

$\sigma_3^2 = \eta^2+\sigma^2$ and

$$w = (x+y)/2.$$

And the denominator of BF is given as in case (c).

There are some interesting features from the above result. First, note that the predictive distribution of Y given X can be summarised in the following two fold definition, i.e.,

$$f(y|x=0) = \begin{cases} \psi, & y=0 \\ (1-\psi)\times g(y|.) & y\neq 0. \end{cases} \qquad (5.9)$$

and

$$f(y|x\neq 0) = \begin{cases} \psi', & y=0 \\ (1-\psi')\times g'(y|x,.) & y\neq 0. \end{cases} \qquad (5.10)$$

where ψ and ψ' are the probabilities of y being zero given x is and is not zero respectively. The functions g and g' are the respective predictive functions of the non-zero samples.

5.3.4 Case (ii) Y consists of one measurement and X consists of m measurements with a non-zero probability of t(x) zeros in the sample

Let us now consider case (ii), for r=1 there are two possibilities that is (a) y is zero (i.e. t(y)=1) and (b) y is not zero (i.e. t(y)=0). If X consists of t(x) zeros and (m-t(x)) other values $x_1,\ldots,x_{m-t(x)}$, then the likelihood function L of this sample is

$$f(X|\theta,\mu) = \begin{cases} \binom{m}{t(x)}\, \theta^{t(x)}\, (1-\theta)^{m-t(x)} \times \\ \qquad g\{\bar{x}_{m-t(x)}|\mu,\sigma^2/(m-t(x))\}, & t(x)<m \\ \theta^m, & t(x)=m. \end{cases}$$

Note that the two-fold definition is required since $\bar{x}_{m-t(x)}$ is not

defined when t(x)=m. The prior density for the unknown parameters θ and μ are given as in case (i). And all the hyper-parameters are assumed known and as before their values are obtained from the training data. Let us first consider case (a), that is y is zero, then $Pr(T(y)=1)=\theta$ and the predictive distribution of T(y) given X can be written as

$$Pr(T(y)=1|X) = \frac{t(x)+a}{a+b+m}. \qquad (5.11)$$

When y is not zero the probability operator becomes a density function, so under C the conditional density function of Y is given by

$$(1-\theta) \times g(y|\mu,\sigma^2).$$

Then the predictive distribution of Y given X can be written as

$$\frac{b+m-t(x)}{a+b+m} \times \frac{g(\bar{x}_{m-t(x)}-y|0,\sigma_1^2) \times g(w|\xi,\sigma_2^2)}{g(\bar{x}_{m-t(x)}|\xi,\sigma_3^2)}, \qquad (5.12)$$

where $\sigma_1^2 = \sigma^2\{1+[1/(m-t(x))]\}$;

$\sigma_2^2 = \tau^2+\sigma^2/[1+m-t(x)]$;

$\sigma_3^2 = \tau^2+\sigma^2/[m-t(x)]$ and

$w = \{[m-t(x)]\bar{x}+y\}/[m-t(x)+1]$.

And the denominator of BF is given by (5.1) and (5.2) for case (a) and (b), respectively.

For the degenerate case when t(x)=m, a general form of the predictive distribution of T(y) given T(x)=m can be obtained from

Aitchson and Dunsmore (1975) which is given as

$$f(T(y)=t(y)\,|\,T(x)=t(x)) =$$

$$\begin{bmatrix} r \\ t(y) \end{bmatrix} \times \frac{Be\{\alpha+t(x)+t(y),\,b+r-t(y)+m-t(x)\}}{Be(\alpha+t(x),\,b+m-t(x))} . \qquad (5.13)$$

Hence the predictive numerators of BF for case (a) and (b) are

$$Pr(T(y)=1\,|\,T(x)=m) = \frac{Be(\alpha+m+1,\,b)}{Be(\alpha+m,\,b)} = \frac{\alpha+m}{\alpha+b+m} \qquad (5.14)$$

and

$$f(y\,|\,T(x)=m) = \frac{Be(\alpha+m,\,b+1)}{Be(\alpha+m,\,b)} \times g(y\,|\,\xi,\eta^2+\sigma^2)$$

$$= \frac{b}{\alpha+b+m} \times g(y\,|\,\xi,\eta^2+\sigma^2), \qquad (5.15)$$

respectively.

Note that for m=1, (5.11) and (5.12) reduce to (5.5) or (5.6) and (5.7) or (5.8), respectively.

Similar to Section 5.3.1, the predictive distribution of Y given X can be summarised in the following two fold definitions, namely

$$f(y\,|\,t(x)<m) = \begin{cases} \gamma, & y=0 \\ (1-\gamma)\times g(y\,|\,\bar{x}) & y\neq 0. \end{cases} \qquad (5.16)$$

and

$$f(y\,|\,t(x)=m) = \begin{cases} \gamma', & y=0 \\ (1-\gamma')\times g'(y\,|\,.) & y\neq 0. \end{cases} \qquad (5.17)$$

where γ and γ' are the probabilities that y is zero given the number

of zeros t(x) in the sample of X is less than or is equal to m, respectively. The functions g and g' are the respective predictive functions of the non-zero positive observations.

Notice the similarity of this analysis with the cat data analysis, so it is possible to use the kernel prior instead of the Normal prior for the unknown parameter μ as in the cat data analysis. This is to be considered in the following section.

5.4 Kernel model: Single hair problem

In view of the results of the preliminary formulation of the Bayes' factor in the previous section, we can now adapt the model developed in Chapter 3 of Section 3.5.1 to utilise the grouping structure in the training data. The model adapted from Section 3.5.1 in Chapter 3 is the simplest one where the within-group variance is assumed known. The effect of ignoring the grouping in the training data is not investigated here but it is worth investigating in future research.

Here, instead of assuming the population group means are independent identically Normal distributed with hyper-parameters estimated from the training data as assumed in Section 5.3, we use a kernel prior for the unknown population group mean, which is a (non-zero) positive random variable, namely

$$\hat{f}(\mu) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{(2\pi)}s'\lambda'} \exp\left\{ - \frac{(\mu-\bar{z}_i^*)^2}{2s'^2\lambda'^2} \right\}, \qquad (5.18)$$

where n is the number of groups in the training data,

$\bar{z}_i^* = \frac{1}{J_i}\sum_{j=1}^{J_i} z_{ij}^*$, is the i^{th} sample group mean of the training data Z^* from Section 5.3.2 part A.

s' is the standard deviation of the sample group means;

λ' is the 'standardised' smoothing parameter.

5.4.1 Case (i) of Section 5.3

This section provides formulae of the Bayes' factor for the case where X and Y only consist of one observation. From equations (5.5), (5.6), (5.7) and (5.8) the formulae for the numerator of the Bayes' factor using a kernel prior shown in (5.18) for the unknown group population mean, can be summarised in the Table 5.1.

Table 5.1 The numerator of the Bayes' factor given various cases of X and of Y:

	Y is zero	Y is not zero
X is zero	$\frac{a+1}{a+b+1}$	$\frac{b}{a+b+1}\, k\{y \mid \bar{z}_i^*\text{'s}, \sigma_1^2\}$
X is not zero	$\frac{a}{a+b+1}$	$\frac{b+1}{a+b+1} \times \frac{\Phi\{(x-y)/\sigma_2\}\, k(w \mid \bar{z}_i^*\text{'s}, \sigma_3^2\}}{k\{x \mid \bar{z}_i^*\text{'s}, \sigma_1^2\}}$

Notes: a and b are the hyper-parameters of the prior for the parameter Θ, and their values are obtained from the training data;
k is the kernel density;
Φ is the standard Normal density;
λ is the 'standardised' smoothing parameter;
s is the standard deviation of the group sample means of the training data;
$\bar{z}_i^*$'s are the sample group means of the training data
$w = (x+y)/2$
$\sigma_1^2 = s^2\lambda^2+\sigma^2$
$\sigma_2 = \sigma\sqrt{2}$
$\sigma_3^2 = s^2\lambda^2+(\sigma^2/2)$

Similary, the denominator of the Bayes' factor using the kernel prior for the unknown population group mean is summarised in the Table 5.2.

Table 5.2 The denominator of the Bayes' factor given Y is zero or not zero:

Y is zero	Y is not zero
$\frac{a}{a+b}$	$\frac{b}{a+b}\, k\{y\|\bar{z}_i^{*}\text{'s}, \sigma_i^2\}$

Notes: a, b, k, $\bar{z}_i^{*}$'s and σ_i^2 are as in Table 5.1

5.4.2 Case (ii) of Section 5.3

Here we suppose that the recovered data Y consist of one observation and that there are t(x) non-zero positive values out of m observations in the control data X. The denominator of the Bayes' factor is the same as Case (i) above. The formulae for the numerator of the Bayes' factor using a kernel prior for the unknown group population mean, can be summarised in the Table 5.3.

There are some features of interest which are worth a mention from the Tables illustrated. First of all, in Case (i) given X is zero, the Bayes' factor is constant over all non-zero positive values of Y since the kernel density functions in the numerator and the denominator of the Bayes' factor cancel each other out. Similarly, in Case (ii) given all m observations of the control data X are zero,

the Bayes' factor is also constant when y is not zero. This means if a non-zero positive value of Y is observed, it does not matter what value y takes, the Bayes' factor is the same for all non-zero positive values of Y given X is zero.

Table 5.3 The numerator of the Bayes' factor for $T(x)<m$ and $T(x)=m$, where $T(x)$ is the number of zeros in the control data X

	Y is zero	Y is not zero
$T(x)<m$	$\frac{a+t(x)}{a+b+m}$	$\frac{b+m-t(x)}{a+b+m} \times \frac{\Phi\{(\bar{x}'-y)/\sigma_2\}k\{w'\|\bar{z}_i^{*}\text{'s},\sigma_3^2\}}{k\{\bar{x}'\|\bar{z}_i^{*}\text{'s},\sigma_4^2\}}$
$T(x)=m$	$\frac{a+m}{a+b+m}$	$\frac{b}{a+b+m} \times k\{y\|\bar{z}_i^{*}\text{'s},\sigma_1^2\}$

Notes: $\bar{x}' = \bar{x}_{[m-t(x)]}$
$w' = \{y+[m-t(x)]\bar{x}_{[m-t(x)]}\}/[1+m-t(x)]$
$\sigma_2 = \sigma\sqrt{[1+(m-t(x))^{-1}]}$
$\sigma_3^2 = s^2\lambda^2+\{\sigma^2/[1+(m-t(x))]\}$
$\sigma_4^2 = s^2\lambda^2+[\sigma^2/m-t(x)]$
and a, b, k, Φ, $\bar{z}_i^{*}$'s, σ_i^2, λ and s are as in Table 5.1.

5.5 Determination of the hyper-parameters and within-group variance values from the training data

From equation (5.3), i.e. Θ follows a beta distribution and from Appendix 1, then the expectation of Θ is given by

$$E(\Theta) = \frac{a}{a+b} .$$

Also, from the training data the average number of zeros occurring can be evaluated as below,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{t_i}{J} = \frac{1}{nJ}\sum_{i=1}^{n} t_i = \frac{t(z)}{N} .$$

Thus we may take $a = t(z)$ and $b = N - t(z)$.

In evaluating the Bayes' factor described in Section 5.3 certain values for the hyper-parameters ξ and η^2, and the within-group variance σ^2 are required. We cannot use the usual formulation for the ANOVA estimates as in the balanced case as we did in Chapter 3, since the training data are no longer balanced.

The formula of the estimates for ξ can be taken as either

$$\frac{1}{N-t(z)}\sum_{\ell=1}^{N-t(z)} z_\ell^* \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n}\bar{z}_i . \qquad (5.19)$$

The latter is known as the unweighted mean. In the example below the unweighted mean is used. Formulae of the ANOVA estimates for σ^2 and η^2 may be obtained from the ANOVA table shown in Table 2.1. The only change necessary in Table 2.1 is accomplished by replacing the EMS for the between groups by $\sigma^2+J_0\sigma_a^2$ where

$$J_0 = \frac{(N-t(z)) - J_1/(N-t(z))}{n-1} \qquad (J_0=J \text{ if each } J_i=J)$$

where $N-t(z) = \sum_{i=1}^{n} J_i$ and $J_1 = \sum_{i=1}^{n} J_i^2$.

Hence σ^2 is estimated as usual by MSE,

$$\hat{\sigma}^2 = WSS/[n(J-1)]$$

and

$$\hat{\sigma}^{*2} = \frac{BMS - WMS}{J_0} . \qquad (5.20)$$

An alternative estimate for σ^{*2} exists, which is derived from the unweighted sum of squares of deviations between groups, namely

$$\sum_{i=1}^{n} (\bar{z}_{i.} - \bar{z}_u)^2$$

where $\bar{z}_u$ is given as in the R.H.S. of (5.19). If BMS' = $\Sigma(\bar{z}_{i.}-\bar{z}_u)^2/(n-1)$, its mean value is

$$E(BMS') = \sigma_a^2 + (\sigma^2/n)(\Sigma(1/J_i) = \sigma_a^2 + \sigma^2/J_h.$$

Hence, the quantity BMS'$-\hat{\sigma}^2/J_h$ is an alternative unbiased estimate of σ_a^2, where the harmonic mean J_h = $n/[\Sigma(1/J_i)]$. Note that these estimates of σ_a^2 , as stated in Section 2.3, have the awkward feature that they can take negative values; biased estimators that are always positive may be superior.

These estimators are chosen because they do not require a Normality assumption. Before showing the results of the models just described, first consider the problem of kernel density estimation in the dog data.

5.6 Problem of kernel density estimation for the dog data

The medullary fraction is a ratio of two measurements, and like the cat data mentioned earlier in Chapter 3 is restricted to lie between 0 and 1. In the cat data, the measurements of medullary fraction lie well away from the boundaries, and thus it does not pose

a problem as far as the kernel density estimation is concerned. However, unlike the cat data, the measurement of the medullary fractions of the dog data are close to the lower boundary, namely near zero. The problem is that the density estimate obtained when this boundary condition is ignored, will give weight to negative numbers which is unacceptable. One possible way of ensuring that $f(\nu)$ is zero for negative ν is simply to calculate the estimate for positive ν ignoring the boundary conditions, and then to set $f(\nu)$ to zero for negative ν. A drawback of this approach is that if we use a method, which usually produces estimates which are probability densities, the estimates obtained will no longer integrate to unity. To make matters worse, the contribution of the points near zero will be much less than that of the points well away from the boundary. So the weight of the distribution near zero will be underestimated. Silverman (1986) suggested several ways to tackle this problem. One of which is the reflection method, the argument of which is as follows. Suppose we augment the data by adding the reflections of all the points in the boundary, to give the set $\{v_1,-v_1,v_2,-v_2,\ldots\}$. If a kernel estimate f^* is constructed from this data set of size 2n, then an estimate based on the original data can be given by putting

$$f(\nu) = \begin{cases} 2\times f^*(\nu) & \text{for } \nu \geqslant 0 \\ 0 & \text{for } \nu < 0. \end{cases}$$

Fig. 5.1(a) and 5.1(b) shows the kernel density estimate using the original method and the reflection method, respectively. In Fig. 5.1(a), the non-zero density estimate is clearly shown when the x-axis is extended beyond zero. Note that the density estimates near zero are slightly higher when the reflection method was used, than

the original method. The dotted line is fitted by an adaptive kernel method (see Section 2.1 for details). The reason that the adaptive method is used also is that the Bayes' factor obtained using the ordinary kernel density to estimate the denominator in the ECA model of Section 5.3.2 might produce a problem at the tails. The problem is that the ordinary kernel estimate tends to zero much faster than the function on the numerator of the BF. That is, the BF behaves reasonably well when we observe measurements well within our previous experience. The problem posed a question, 'Can we make a reasonable judgement when we observed a measurement which is even slightly outside our previous experience?' This leads to deriving a new kernel density function which will be discussed in Chapter 7.

In the example given in the next section, the reflection method is used.

5.7 Illustration

The models developed in Sections 5.3 and 5.4 are applied to the dog data. The numbers of observations in each group as well as the ordered sample group means are tabulated in Table 5.4. Histograms of the 20 dogs, ten hairs from each dog are plotted in Fig. 5.2. It can be seen that Dogs 1, 6, 10 and 20 have zero values of medullary fraction.

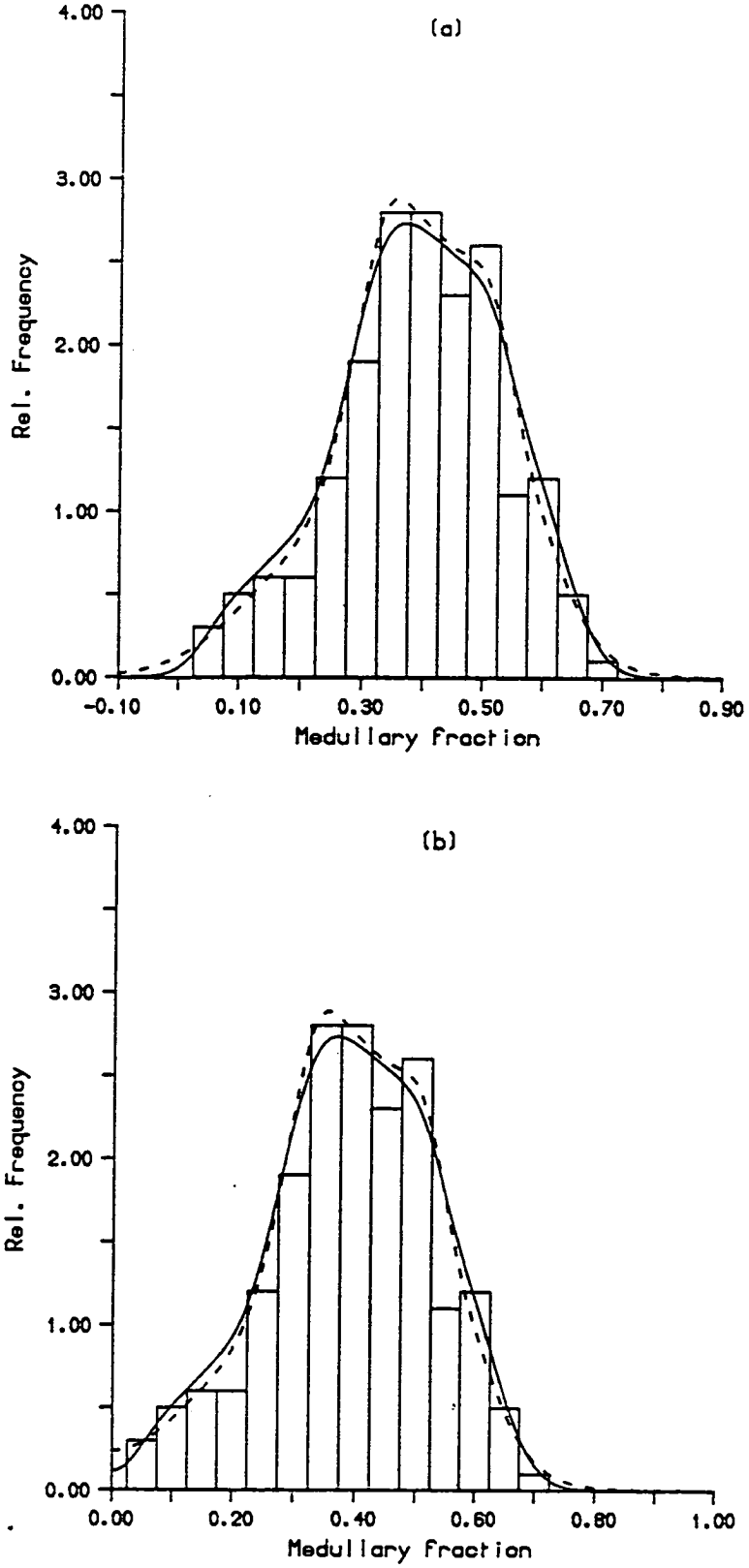


Fig. 5.1 Histogram and density plot of the non-zero measurement of the dog data; fitted by the ordinary (——) and adaptive (- - -) kernels using (a) the original and (b) the reflection methods.

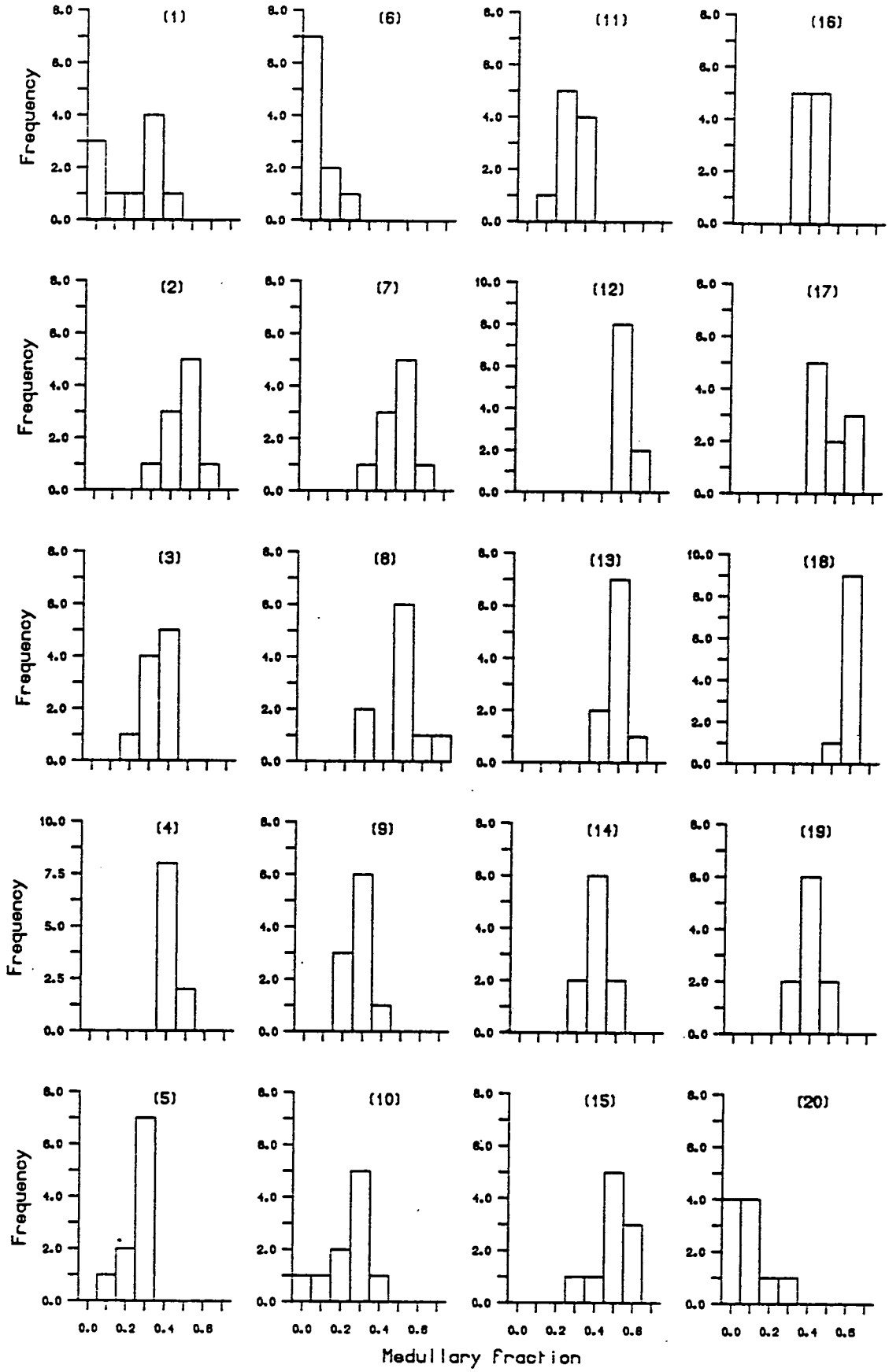


Fig. 5.2 Histogram of 20 sets of dog hairs, 10 hairs from each dog.

Table 5.4 Ordered sample group means and number of observations J_i in each group

Group	$\bar{z}^*_{i.}$	J_i	Group	$\bar{z}^*_{i.}$	J_i
1	0.1182	3	11	0.4081	10
2	0.1482	6	12	0.4253	10
3	0.2349	10	13	0.4614	10
4	0.2564	9	14	0.4633	10
5	0.2599	10	15	0.4666	10
6	0.2822	7	16	0.4880	10
7	0.2846	10	17	0.4908	10
8	0.3422	10	18	0.4990	10
9	0.3597	10	19	0.5263	10
10	0.3884	10	20	0.6032	10

$$WMS = \hat{\sigma}^2 = 0.00506; \quad BMS = 0.13169.$$

5.7.1 Results of the modified ECA model

Results from the model established in Section 5.3.2. are tabulated in Table 5.5 (i) and (ii) for r=1,m=1 and for r=1,m=5 with different values of t(x), respectively. The results shown here are obtained by using the ordinary kernel estimate and the reflection method. Results from the adaptive kernel with the reflection method are shown in Table 5.6 (i) and (ii) for r=1,m=1 and r=1,m=5, respectively. There is little difference between the two kernel density estimate methods. The values of BF are slightly smaller when the ordinary kernel is used.

Because of the reason given in Section 5.6, we only concentrate on the results obtained from the adaptive and reflection methods. If $\bar{x}$ and $\bar{y}$ are equal, the value of BF decreases as the number of zeros in the sample of X increases. For instance, when $\bar{x}$ and $\bar{y}$ are both 0.2, then BF decreases from 6.0958 to 4.9673. Other interesting

points from the Table 5.6(ii) are that BF appears to be larger when $\bar{x}$ and $\bar{y}$ are close to zero and the t(x) is small. This suggests if most of the control sample are non-zero, the least common value of the medullary fraction is 0.2 and the next least common is 0.6 which is clearly shown in Fig. 5.2. However, when all of the control sample are zeros, the least common value is 0.6.

Graphical presentation of the BF given some values of X (either zero or non-zero) over the range of non-zero part of Y is shown in Fig. 5.3 for r=1,m=1. Figs. 5.4 and 5.5 show the case when t(x)=0 and t(x)=3 respectively, for r=1,m=5 of the non-zero part of X. The red solid line in these Figures represents the Bayes' factor equal to one. Notice that in Fig. 5.3(b) there is a slight irregularity in the behaviour of BF when x=0. This is the case (i)c of Section 5.3, of which BF is a ratio of a normal function and the slightly less smooth kernel density estimate. We have already seen from Fig. 5.2 that the kernel density estimate is slightly irregular. So it is not surprising that the irregular behaviour occurred. However, one might expect this irregular behaviour would disappear if a kernel prior were used. So further research should be done in this aspect. A similar feature also happened in the ordinary kernel case but it does not show in the graph because of the dominant part at the tail. Figs. (a) of 5.3, 5.4 and 5.5 confirmed fears of the unreasonable behaviour of BF at the tail when the ordinary kernel was used. This problem does not arise when the adaptive kernel estimate is used, as shown in Figs. (b) of 5.3, 5.4, and 5.5. Therefore it is advisable to use the adaptive kernel estimate to avoid complications when we observe an extremely rare value of X and of Y.

5.7.2 Results of the Kernel model

Results from the Kernel model derived in Section 5.4 are shown here in Tables 5.7 and 5.8 using the respective ordinary and adaptive kernel methods. For instance, Table 5.7 (i), (ii) and (iii) are for the case where r=1&m=1, r=1&m=5 and r=1&m=10, respectively. The Bayes' factor is plotted as a function of (positive, non-zero) Y given X = 0.2 (0.2) 0.6 and are shown in Fig. 5.6 for r=1,m=1. Figs. 5.7 - 5.8 illustrate the case where r=1,m=5 for t(x)=0 and t(x)=3, respectively.

There is not much difference between the ordinary and adaptive kernel methods. Though the adaptive kernel method produces slight larger values of BF when both positive non-zero values of X and Y are observed. The values of BF increase slightly as m increases.

5.8 Conclusion and discussion

The modified ECA and Kernel models are entirely different since they are based upon different assumptions about the recovered data and the training data. So direct comparison between the two models will be inappropriate. A simulation study will be required to distinguish between the merit of the two models. However, the kernel model is more desirable than the modified ECA model because it makes use of the random structure in the data. Also the distribution of the population group mean is relevant in modelling the forensic data.

The ordinary and adaptive kernel methods produce similar results of the Bayes' factor in the Kernel model but from the results of the

modified ECA model, the adaptive kernel method appears to be better in terms of the behaviour of the BF. That is, the Kernel model does not pose a problem when a rare value of X or of Y is observed as illustrated in Figs. 5.3 - 5.5.

Further work could be done on the extension to the r-samples case (see Appendix 5) and examining the effect of ignoring the zero features in the sample. The effect, if any, of the degree of unbalancedness of group sizes in the training data should also be investigated in the context of modelling the Bayes' factor.

Table 5.5(i) Bayes' factor of the analysis of dog data given some values of X and of Y for r=1,m=1. Modified ECA model. Reflection and ordinary kernel method.

Y \ X	0.00	0.20	0.40	0.60
0.0	1.0614	0.9950	0.9950	0.9950
0.2	1.1763	3.9705	0.4949	0.0054
0.4	1.0767	0.4530	1.5873	0.4862
0.6	0.8015	0.0037	0.3619	3.1168

Table 5.5(ii) Bayes' factor of the analysis of dog data given some values of X and of Y for r=1,m=5 with t(x) = 0(1)5, number of zeros in the data X. Modified ECA model. Reflection and ordinary kernel method.

		$\bar{X}$			
Y	t(x)	0.00	0.20	0.40	0.60
0.0	0	-	1.4387	0.0001	0.0000
	1	-	1.4777	0.0002	0.0000
	2	-	1.5372	0.0003	0.0000
	3	-	1.6505	0.0009	0.0000
	4	-	2.0178	0.0063	0.0000
	5	1.3008	-	-	-
0.2	0	-	5.6145	0.2015	0.0000
	1	-	5.4935	0.2206	0.0000
	2	-	5.3291	0.2516	0.0001
	3	-	5.0767	0.3084	0.0004
	4	-	4.5751	0.4350	0.0055
	5	1.1533	-	-	-
0.4	0	-	0.1064	1.8987	0.1106
	1	-	0.1264	1.8455	0.1323
	2	-	0.1624	1.7680	0.1719
	3	-	0.2398	1.6370	0.2585
	4	-	0.4639	1.3473	0.5178
	5	1.0557	-	-	-
0.6	0	-	0.0000	0.1566	4.5352
	1	-	0.0000	0.1704	4.4424
	2	-	0.0000	0.1926	4.3188
	3	-	0.0003	0.2329	4.1342
	4	-	0.0036	0.3223	3.7839
	5	0.7858	-	-	-

Table 5.6(i) Bayes' factor of the analysis of dog data given some values of X and of Y for r=1,m=1. Kernel model. Reflection and adaptive kernel method.

Y \ X	0.00	0.20	0.40	0.60
0.0	1.0614	0.9950	0.9950	0.9950
0.2	1.2771	4.3109	0.5373	0.0059
0.4	1.0548	0.4438	1.5550	0.4763
0.6	0.9454	0.0043	0.4269	3.6767

Table 5.6(ii) Bayes' factor of the analysis of dog data given some values of X and of Y for r=1,m=5 with t(x) = 0(1)5, number of zeros in the data X. Kernel model. Reflection and adaptive kernel method.

		$\bar{X}$			
Y	t(x)	0.00	0.20	0.40	0.60
0.0	0	-	1.4387	0.0000	0.0000
	1	-	1.4777	0.0002	0.0000
	2	-	1.5372	0.0003	0.0000
	3	-	1.6505	0.0009	0.0000
	4		2.0178	0.0063	0.0000
	5	1.3008	-	-	-
0.2	0	-	6.0958	0.2187	0.0000
	1	-	5.9644	0.2395	0.0000
	2	-	5.7860	0.2732	0.0001
	3	-	5.5119	0.3348	0.0005
	4	-	4.9673	0.4723	0.0059
	5	1.2522	-	-	-
0.4	0	-	0.1043	1.8601	0.1083
	1	-	0.1239	1.8079	0.1297
	2	-	0.1591	1.7320	0.1684
	3	-	0.2349	1.6037	0.2533
	4	-	0.4545	1.3199	0.5073
	5	1.0342	-	-	-
0.6	0	-	0.0000	0.1848	5.3497
	1	-	0.0000	0.2011	5.2403
	2	-	0.0001	0.2272	5.0945
	3	-	0.0004	0.2748	4.8768
	4	-	0.0043	0.3802	4.4636
	5	0.9270	-	-	-

Table 5.7(i) Bayes' factor of the analysis of dog data given some values of X and of Y for r=1,m=1. Kernel model. Reflection and ordinary kernel method.

Y \ X	0.00	0.20	0.40	0.60
0.0	1.0614	0.9950	0.9950	0.9950
0.2	0.9950	2.9539	0.3533	0.0024
0.4	0.9950	0.3533	1.7970	0.4695
0.6	0.9950	0.0024	0.4695	3.5626

Table 5.7(ii) Bayes' factor of the analysis of dog data given some values of X and of Y for r=1,m=5 with t(x) = 0(1)5, number of zeros in the data X. Kernel model. Reflection and ordinary kernel method.

		$\bar{X}$			
Y	t(x)	0.00	0.20	0.40	0.60
0.0	0	-	1.8624	0.0001	0.0000
	1	-	2.0032	0.0002	0.0000
	2	-	2.2223	0.0003	0.0000
	3	-	2.6176	0.0011	0.0000
	4		3.6273	0.0102	0.0000
	5	1.3008	-	-	-
0.2	0	-	3.7983	0.1339	0.0000
	1	-	3.7010	0.1472	0.0000
	2	-	3.5627	0.1698	0.0001
	3	-	3.3373	0.2145	0.0002
	4	-	2.8673	0.3290	0.0024
	5	0.9756	-	-	-
0.4	0	-	0.0980	2.2356	0.1182
	1	-	0.1119	2.1765	0.1387
	2	-	0.1362	2.0913	0.1746
	3	-	0.1873	1.9494	0.2493
	4	-	0.3395	1.6382	0.4554
	5	0.9756	-	-	-
0.6	0	-	0.0000	0.1814	4.8230
	1	-	0.0000	0.2006	4.7114
	2	-	0.0000	0.2320	4.5560
	3	-	0.0002	0.2904	4.3079
	4	-	0.0023	0.4217	3.7879
	5	0.9756	-	-	-

Table 5.7(iii) Bayes' factor of the analysis of dog data given some values of X and of Y for r=1,m=10 with t(x) = 0(2)10, number of zeros in the data X. Kernel model. Reflection and ordinary kernel method.

		$\bar{X}$			
Y	t(x)	0.00	0.20	0.40	0.60
0.0	0	-	1.5502	0.0000	0.0000
	2	-	1.6363	0.0000	0.0000
	4	-	1.7682	0.0001	0.0000
	6	-	2.0080	0.0002	0.0000
	8	-	2.6233	0.0011	0.0000
	10	1.5873	-	-	-
0.2	0	-	3.9762	0.1064	0.0000
	2	-	3.8911	0.1121	0.0000
	4	-	3.7808	0.1222	0.0000
	6	-	3.6128	0.1437	0.0000
	8	-	3.2578	0.2094	0.0002
	10	0.9524	-	-	-
0.4	0	-	0.0711	2.3432	0.0795
	2	-	0.0768	2.2925	0.0878
	4	-	0.0870	2.2264	0.1027
	6	-	0.1092	2.2125	0.1354
	8	-	0.1828	1.9030	0.2434
	10	0.9524	-	-	-
0.6	0	-	0.0000	0.1408	5.0281
	2	-	0.0000	0.1495	4.9249
	4	-	0.0000	0.1647	4.7934
	6	-	0.0000	0.1958	4.5992
	8	-	0.0002	0.2835	4.2054
	10	0.9524	-	-	-

Table 5.8(i) Bayes' factor of the analysis of dog data given some values of X and of Y for r=1,m=1. Kernel model. Reflection and adaptive kernel method.

Y \ X	0.00	0.20	0.40	0.60
0.0	1.0614	0.9950	0.9950	0.9950
0.2	0.9950	3.0948	0.3558	0.0029
0.4	0.9950	0.3558	1.7178	0.4971
0.6	0.9950	0.0029	0.4971	3.6589

Table 5.8(ii) Bayes' factor of the analysis of dog data given some values of X and of Y for r=1,m=5 with t(x) = 0(1)5, number of zeros in the data X. Kernel model. Reflection and adaptive kernel method.

		$\bar{X}$			
Y	t(x)	0.00	0.20	0.40	0.60
0.0	0	-	1.8008	0.0001	0.0000
	1	-	1.9312	0.0002	0.0000
	2	-	2.1400	0.0003	0.0000
	3	-	2.5394	0.0010	0.0000
	4	-	3.6544	0.0090	0.0000
	5	1.3008	-	-	-
0.2	0	-	4.0372	0.1389	0.0000
	1	-	3.9354	0.1520	0.0000
	2	-	3.7917	0.1743	0.0001
	3	-	3.5601	0.2182	0.0003
	4	-	3.0870	0.3295	0.0030
	5	0.9756	-	-	-
0.4	0	-	0.0958	2.1150	0.1221
	1	-	0.1100	2.0582	0.1453
	2	-	0.1348	1.9763	0.1860
	3	-	0.1876	1.8389	0.2708
	4	-	0.3481	1.5358	0.5001
	5	0.9756	-	-	-
0.6	0	-	0.0000	0.1968	5.2065
	1	-	0.0000	0.2174	5.0989
	2	-	0.0001	0.2506	4.9528
	3	-	0.0002	0.3106	4.7248
	4	-	0.0028	0.4391	4.2441
	5	0.9756	-	-	-

Table 5.8(iii) Bayes' factor of the analysis of dog data given some values of X and of Y for r=1.m=10 with t(x) = 0(2)10, number of zeros in the data X. Kernel model. Reflection and adaptive kernel method.

		$\bar{X}$			
Y	t(x)	0.00	0.20	0.40	0.60
0.0	0	-	1.5176	0.0000	0.0000
	2	-	1.5956	0.0000	0.0000
	4	-	1.7151	0.0001	0.0000
	6	-	1.9359	0.0002	0.0000
	8	-	2.5449	0.0010	0.0000
	10	1.5873	-	-	-
0.2	0	-	4.2238	0.1115	0.0000
	2	-	4.1339	0.1171	0.0000
	4	-	4.0177	0.1272	0.0000
	6	-	3.8417	0.1484	0.0000
	8	-	3.4753	0.2130	0.0003
	10	0.9524	-	-	-
0.4	0	-	0.0686	2.2178	0.0792
	2	-	0.0744	2.1700	0.0884
	4	-	0.0848	2.1067	0.1050
	6	-	0.1073	2.0092	0.1418
	8	-	0.1831	1.7952	0.2644
	10	0.9524	-	-	-
0.6	0	-	0.0000	0.1525	5.4045
	2	-	0.0000	0.1621	5.2985
	4	-	0.0000	0.1787	5.1663
	6	-	0.0000	0.2122	4.9775
	8	-	0.0002	0.3032	4.6123
	10	0.9524	-	-	-

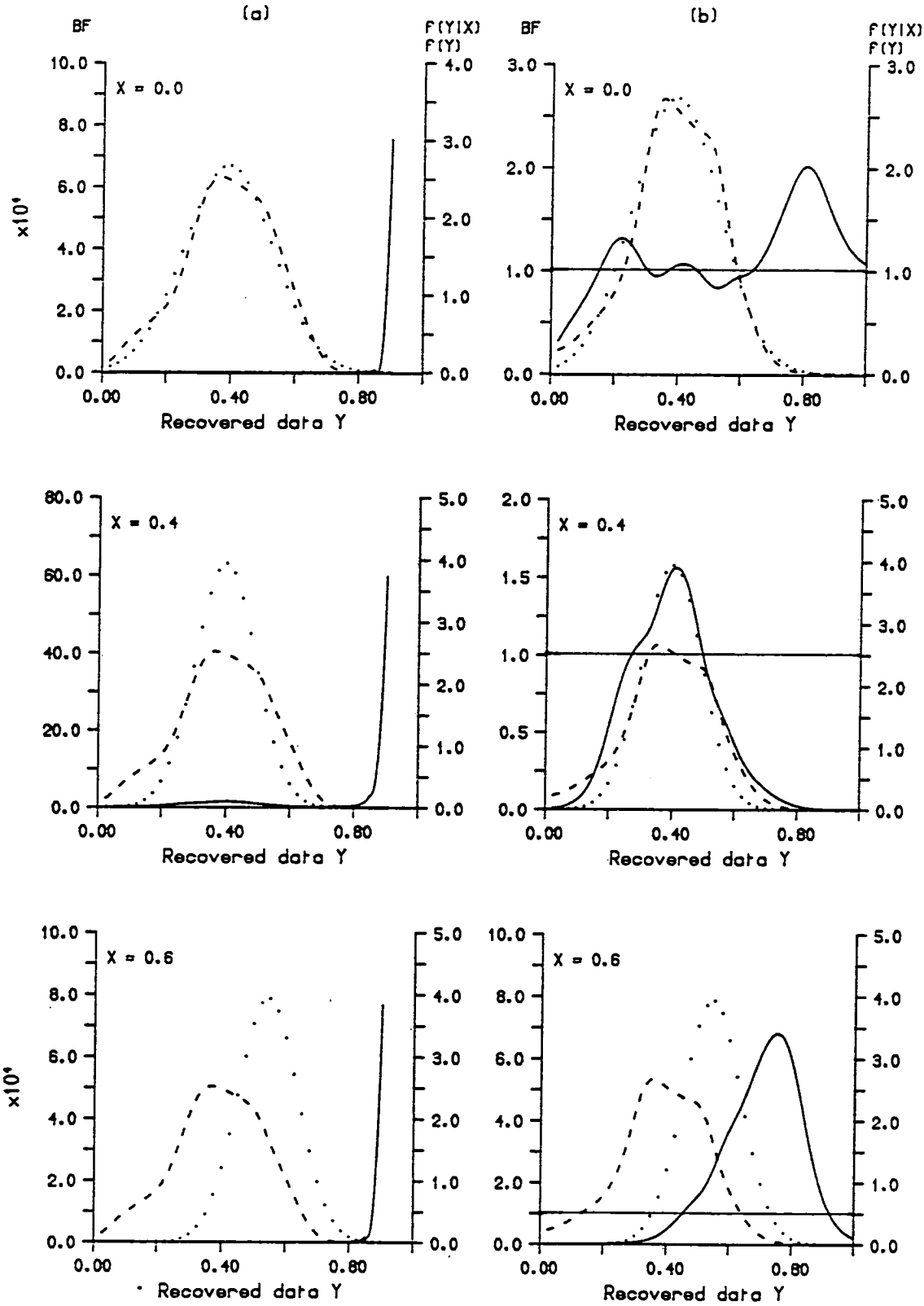


Fig. 5.3 Plot of Bayes' factor (——) together with the predictive density of Y given some values X (····) and marginal density of Y (- - -) by (a) Ordinary and (b) Adaptive kernel methods for r=1 & m=1. Modified ECA model.

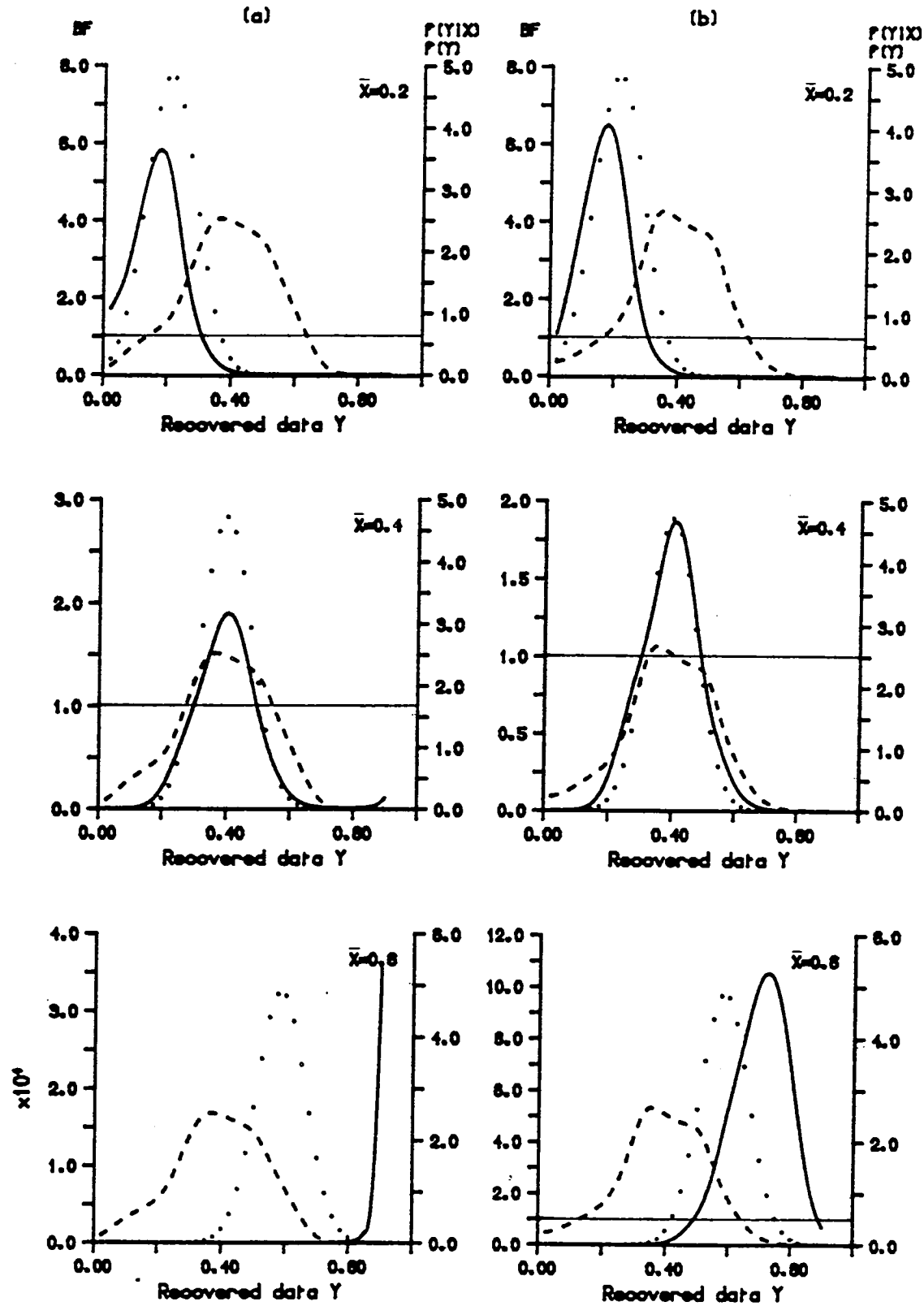


Fig. 5.4 Plot of Bayes' factor (——) together with the predictive density of Y given some values X̄ (····) and marginal density of Y (- - -) by (a) Ordinary and (b) Adaptive kernel methods for r=1, m=5 and t(x)=0. Modified ECA model.

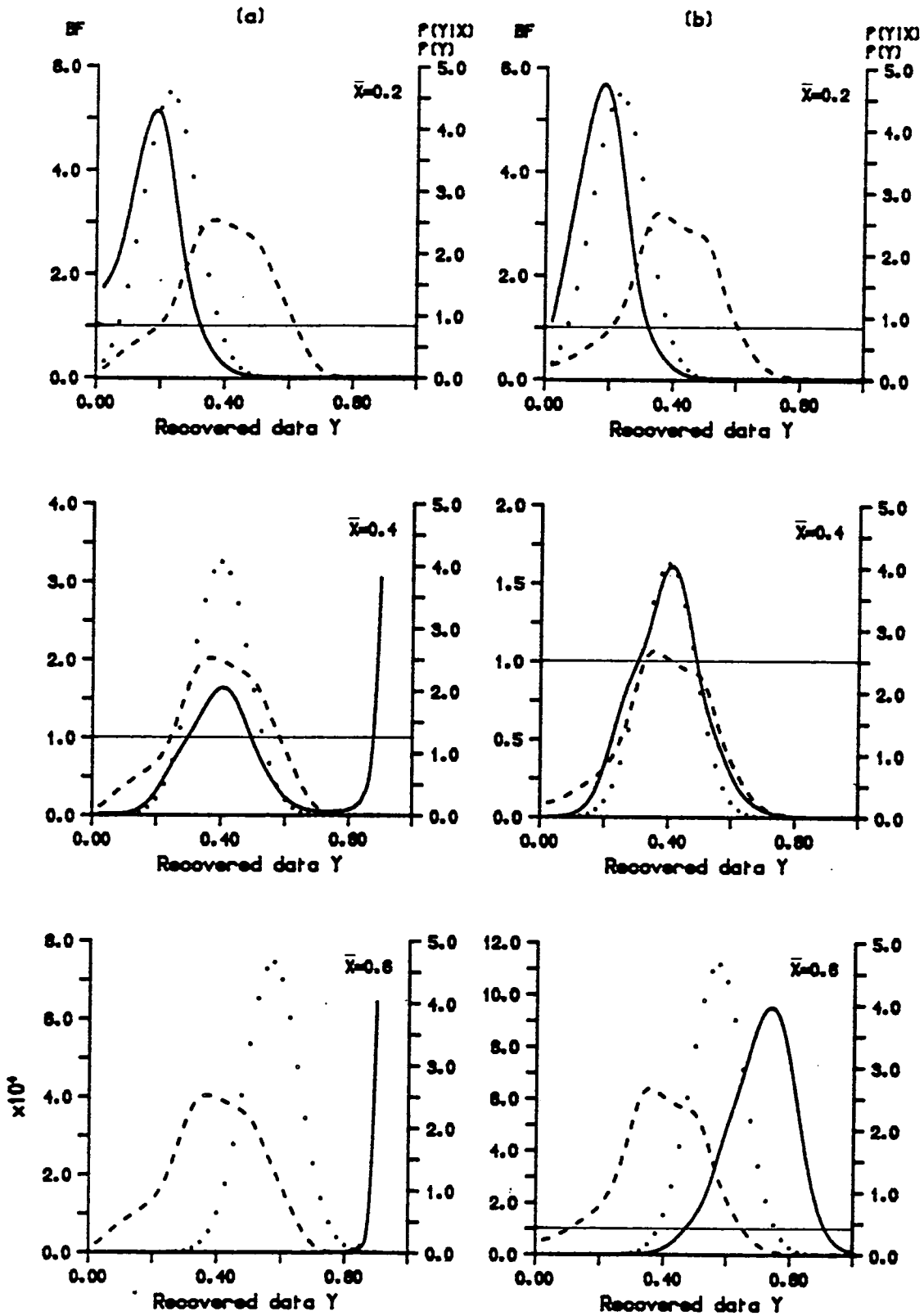


Fig. 5.5 Plot of Bayes' factor (——) together with the predictive density of Y given some values X̄ (····) and marginal density of Y (- - -) by (a) Ordinary and (b) Adaptive kernel methods for r=1, m=5 and t(x)=3. Modified ECA model.

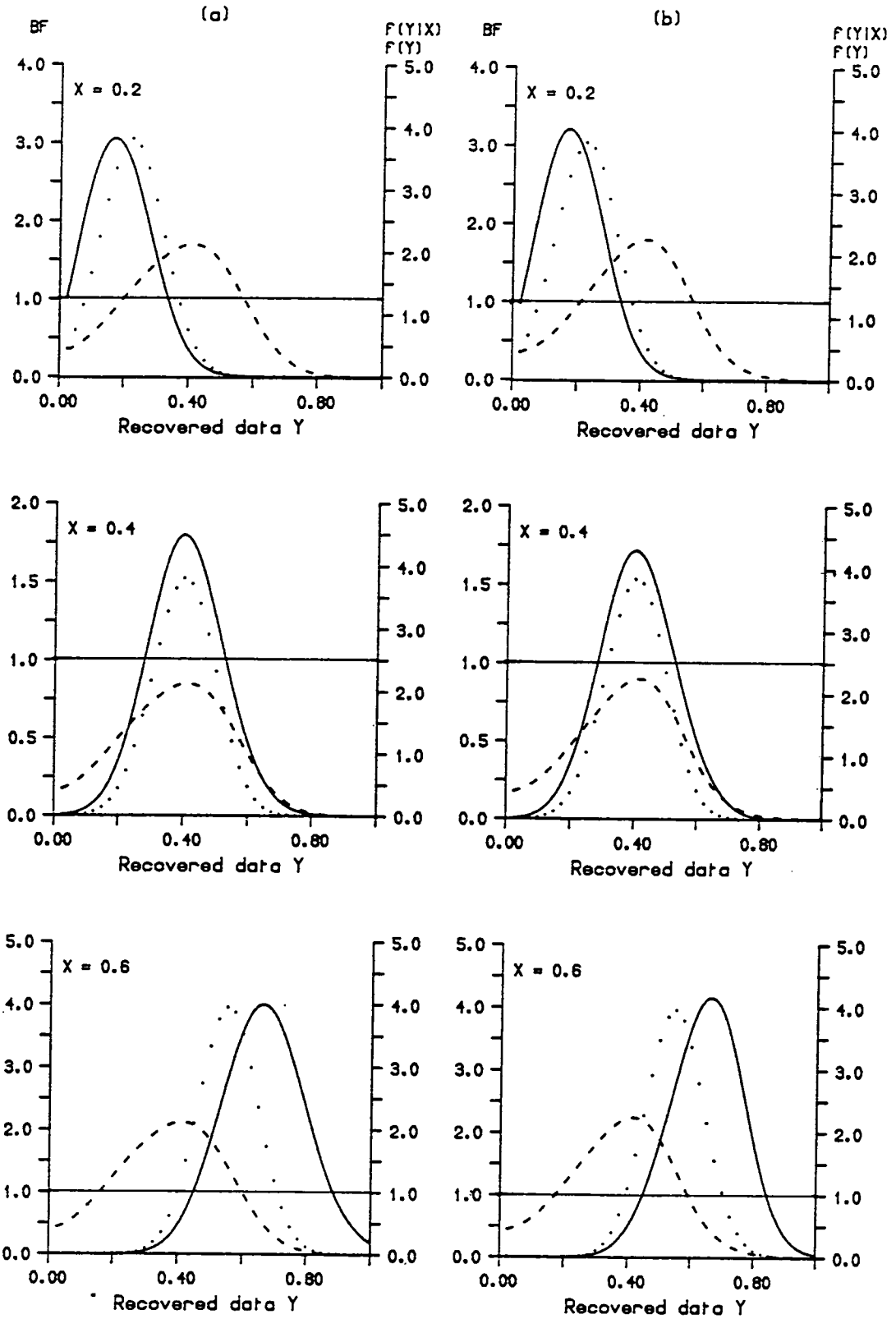


Fig. 5.6 Plot of Bayes' factor (——) together with the predictive density of Y given some values X (····) and marginal density of Y (- - -) by (a) Ordinary and (b) Adaptive kernel methods for r=1 & m=1. Kernel model.

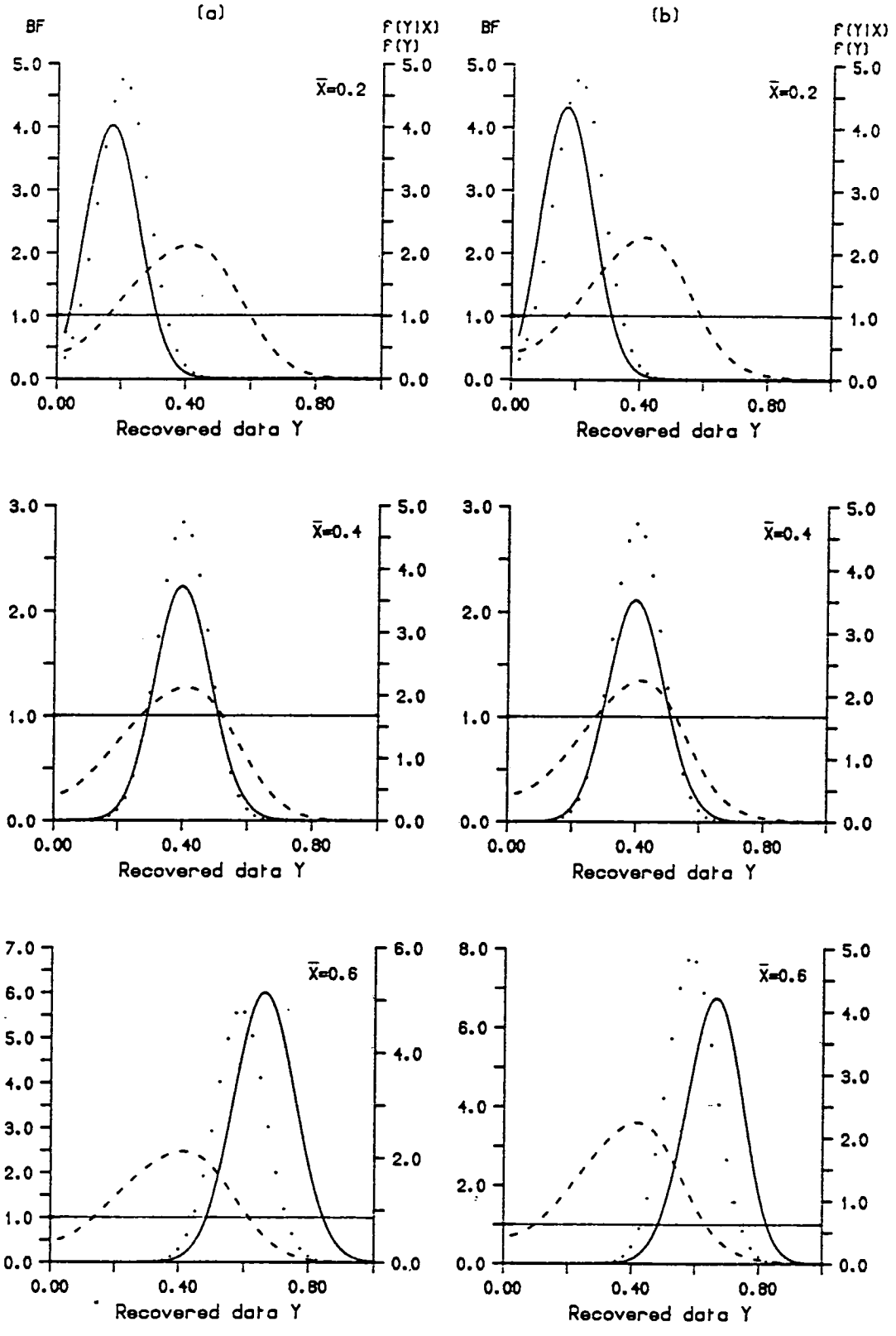


Fig. 5.7 Plot of Bayes' factor (——) together with the predictive density of Y given some values $\bar{X}$ (····) and marginal density of Y (- - -) by (a) Ordinary and (b) Adaptive kernel methods for r=1, m=5; and t(x)=0. Kernel model.

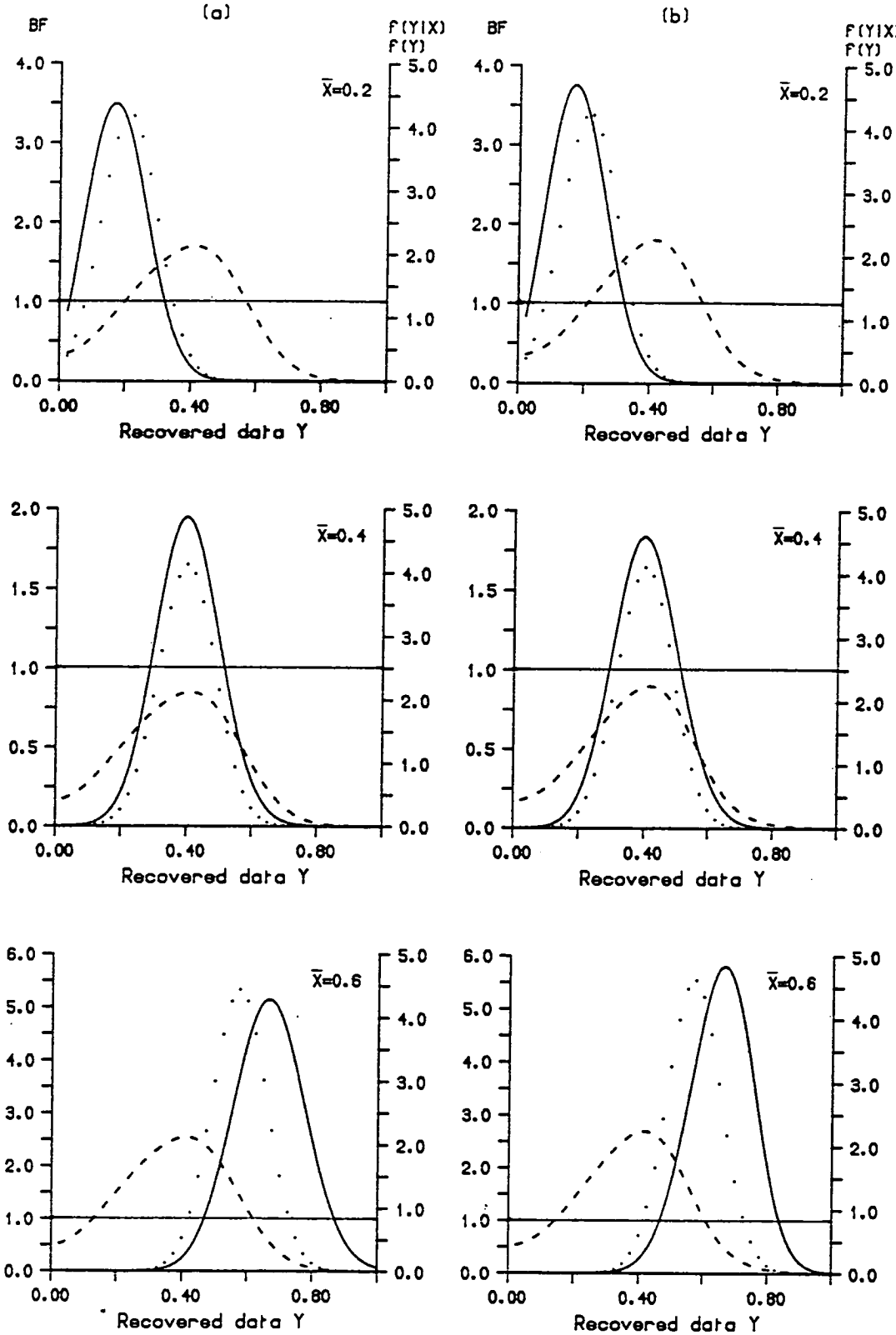


Fig. 5.8 Plot of Bayes' Factor (——) together with the predictive density of Y given some values $\bar{X}$ (····) and marginal density of Y (- - -) by (a) Ordinary and (b) Adaptive kernel methods for r=1, m=5 and t(x)=3. Kernel model.

CHAPTER 6

MODELLING THE BAYES FACTOR FOR MULTIVARIATE DATA

6.1 Introduction

In Chapter 3 we considered the case where observations were univariate. In this chapter the modelling and estimation of the Bayes' factor assuming the within-group variance is known is extended to the multivariate case of p-dimensional observations.

Multivariate data consist of observations on several different variables for a number of individuals or objects. In the spirit of Chapter 3 the measurements taken from a material of interest found at the scene of a crime or on the suspect will be multivariate, consisting of two or more dimensions. For example, as in Chapter 3, in addition to the variable medullary fraction measurements on the cat hair we could also have the measurements for the hair width or indeed medullary width. Using the same notation as in Chapter 3, a formula for the Bayes' factor is now derived below.

First consider the data structure of the training data. Suppose we have n populations or groups from which observations are drawn independently. Evett et al (1987) considered a bivariate problem in which the population means $\underset{\sim}{\mu}_i$ are assumed to have come from a noninformative prior distribution. This chapter deals with the case where the population means $\underset{\sim}{\mu}_i$ have been randomly selected from the same parent population in advance.

More specifically, assume that the elements within any randomly selected population i are Normally distributed with mean vector $\underset{\sim}{\mu}_i$ and common covariance matrix Σ, denoted by $N_p(\underset{\sim}{\mu}_i,\Sigma)$, and that each $\underset{\sim}{\mu}_i$, for i=1,2,...n, is drawn independently from the same parent population. In the usual random effects model, the $\underset{\sim}{\mu}_i$ has a Normal distribution with mean vector $\underset{\sim}{\xi}$ and covariance matrix T. However, the assumption of the between populations $\underset{\sim}{\mu}_i$ being Normally distributed is often unrealistic. Also, the multivariate data are usually multimodal. In high dimensions, non-Normality of the data is extremely difficult to detect. Thus the initial interest in this chapter is in the estimation of the distribution of $\underset{\sim}{\mu}_i$ using the training data available.

In the next two sections a method of estimating the probability density function of $\underset{\sim}{\mu}_i$ is proposed and expressions for the predictive and marginal densities will be derived under the random effects model. The method, involving the use of the past or selected representative data to estimate the probability density function of an unknown parameter is called the empirical Bayes method. Details of such a method have been described in Section 2.2. In Section 6.4 the computational aspects of evaluating the Bayes' factor are discussed. In the final section the results are applied to the cat data to evaluate the weight of evidence in a forensic context.

6.2 Probability density function of $\underset{\sim}{\mu}_i$

A set of p-dimensional training data $\underset{\sim}{Z}_{ij}$ (i=1,2,...,n; j=1,2,...,J) is taken to be a representative sample from some population of interest. The training data consists of n groups with

J observations within each group. Adopting the empirical Bayes' method the distribution of $\underset{\sim}{\mu}$ is estimated by the ordinary and adaptive kernel methods under the assumptions that the training data are grouped or ungrouped. For the definition of 'grouped' and 'ungrouped' see Chapter 3. A multivariate Gaussian kernel is used for mathematical convenience.

6.2.1 The ordinary kernel method

Analogously with Section 2.1.2, the ordinary kernel density estimate for $f(\underset{\sim}{\mu})$ using the group sample means as the data point for the assumed grouped training data case is given by

$$\hat{f}_p(\underset{\sim}{\mu}) = \frac{1}{(2\pi)^{p/2}\,|\lambda^2 S|^{\frac{1}{2}}\,n} \sum_{i=1}^{n} \exp\left\{-\frac{1}{2\lambda^2}(\underset{\sim}{\mu}-\underset{\sim}{\bar{z}}_i)'S^{-1}(\underset{\sim}{\mu}-\underset{\sim}{\bar{z}}_i)\right\} \qquad (6.1)$$

where S is the $p \times p$ sample covariance matrix of the group mean vectors with the jk^{th} $(j,k = 1,2,\ldots,p)$ entries defined as

$$s_{jk}^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (\bar{z}_{ij}-\bar{z}_{.j})(\bar{z}_{ik}-\bar{z}_{.k}). \qquad (6.2)$$

As in Chapter 3, λ is the 'standardised' smoothing parameter which determines the smoothness of the density function. The kernel density estimate shown in (6.1) is slightly different from the one used by Habbema et al (1974) who used a robust version of the sample covariance matrix. Habbema et al (1974), standardising the variables by a simple tranformation and assuming the variables are independent, obtained an estimate for λ. I will denote this estimate by λ_D. This is equivalent to assuming the sample covariance matrix S in (6.1) is diagonal. Silverman (1986) quoted a suggestion of Tukey and Tukey

(1981), that a robust sample covariance matrix should be used. Let matrix D be such a diagonal matrix so that from (6.1) above another kernel density estimate of $f_p(\underset{\sim}{\mu})$ may be given by

$$\hat{f}_p(\underset{\sim}{\mu}) = \frac{1}{(2\pi)^{p/2}\ |\lambda_D^2 D|^{\frac{1}{2}}\ n} \sum_{i=1}^{n} \exp\left\{ -\frac{1}{2\lambda_D^2}(\underset{\sim}{\mu}-\underset{\sim}{\bar{z}}_i)'D^{-1}(\underset{\sim}{\mu}-\underset{\sim}{\bar{z}}_i) \right\}. \qquad (6.3)$$

The diagonal entries of D, d_{kk}^2 (for k = 1, 2,..., p) are the sample variance of the k^{th} variable as defined in (6.2). The smoothing parameters λ and λ_D are estimated using the pseudo-maximum likelihood leave-one-out method as described in Section 2.1.3. The formulae of the Bayes' factor derived from (6.1) and (6.3) will be denoted by BF and BF_D, respectively.

Similarly the distribution of $\underset{\sim}{\mu}$ for the ungrouped training data case can be obtained by replacing $\underset{\sim}{\bar{z}}_i$ with z_ℓ and n with N (= n×J) in (6.1) and (6.3) for the diagonal and non-diagonal sample covariance matrix case respectively. The sample covariance matrix of the ungrouped data is given by

$$S' = \frac{1}{(N-1)} \sum_{\ell=1}^{N} (\underset{\sim}{z}_\ell - \underset{\sim}{\bar{z}}_.)(\underset{\sim}{z}_\ell - \underset{\sim}{\bar{z}}_.)'. \qquad (6.4)$$

6.2.2 The adaptive kernel method

A general derivation of the adaptive kernel method has already been described in Section 2.1.2. From the experience of Chapter 3, the adaptive kernel method may be used to improve the evaluation of the Bayes' factor. The ordinary kernels (6.1) and (6.3) are used as a pilot estimate to obtain the smoothing factors λ_i for the diagonal and non-diagonal case. The sensitivity parameter δ in Section 2.1.2

is set to a half. Thus, from (6.1) and (6.3) adaptive kernel estimates for the distribution of $\underset{\sim}{\mu}$, may be written as

$$\hat{f}_p(\underset{\sim}{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \frac{|S^{-1}|^{\frac{1}{2}}}{(2\pi\lambda^2\kappa_i^2)^{p/2}} \exp\left\{-\frac{1}{2\lambda^2\kappa_i}(\underset{\sim}{\mu}-\underset{\sim}{\bar{z}}_i)'S^{-1}(\underset{\sim}{\mu}-\underset{\sim}{\bar{z}}_i)\right\} \quad (6.5)$$

and

$$\hat{f}_p(\underset{\sim}{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \frac{|D^{-1}|^{\frac{1}{2}}}{(2\pi\lambda^2\kappa_{Di}^2)^{p/2}} \exp\left\{-\frac{1}{2\lambda^2\kappa_{Di}^2}(\underset{\sim}{\mu}-\underset{\sim}{\bar{z}}_i)'D^{-1}(\underset{\sim}{\mu}-\underset{\sim}{\bar{z}}_i)\right\}. \quad (6.6)$$

A similar expression for the ungrouped training data case is shown in Appendix 6.

6.2.3 A simulation study

A Simulation study was carried out to examine the assumption that the training data are grouped or not grouped when they are used to estimate the distribution of the unknown between groups random factor $\underset{\sim}{\mu}$ proposed in the previous sections. Here a bivariate situation is considered. Random bivariate observations $\underset{\sim}{z}_{ij}$ are generated from the following model

$$\underset{\sim}{z}_{ij} = \underset{\sim}{\mu}_i + \underset{\sim}{\epsilon}_{ij} \qquad ,i = 1,\ldots,n;\ j=1,\ldots,J.$$

Random vectors generated from the underlying distribution of $\underset{\sim}{\mu}_i$ are either to be a bivariate Normal or a bivariate non-Normal distribution. The random error vector $\underset{\sim}{\epsilon}_{ij}$ is distributed as a bivariate Normal with mean vector zero and identity covariance matrix.

The between group random observations are generated from the following mixture of two Normals distribution:

$$f(\underset{\sim}{\mu}) = q\,[f_2(\underset{\sim}{\mu}|\underset{\sim}{\xi}_1,\Sigma_1)] + (1-q)\,[f_2(\underset{\sim}{\mu}|\underset{\sim}{\xi}_2,\Sigma_2)], \qquad (6.7)$$

where f_2 is a bivariate Normal density function with appropriate mean vectors $\underset{\sim}{\xi}_1$, $\underset{\sim}{\xi}_2$ and covariance matrices Σ_1, Σ_2. The covariance matrices Σ_1 and Σ_2 are denoted by

$$\begin{bmatrix} \sigma_{11}^2 & \sigma_1 \\ \sigma_1 & \sigma_{12}^2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \sigma_{21}^2 & \sigma_2 \\ \sigma_2 & \sigma_{22}^2 \end{bmatrix},$$

respectively, where $\sigma_1=\rho_1\sigma_{11}\sigma_{12}$ and $\sigma_2=\rho_2\sigma_{21}\sigma_{22}$.

Different choices of the values for q and covariance matrices determine whether the distribution of $\underset{\sim}{\mu}_i$ is Normal or non-Normal. In the Normal case, the value of q is set to 1. Without loss of generality, the mean vector $\underset{\sim}{\xi}_1$ is chosen to be (0.0,0.0)'. The values of the elements in the covariance matrix Σ_1 are given in Table 6.1 below. The density of the five bivariate Normal distirbutions are plotted in Fig. 6.1. In the non-Normal case, q is set to 0.5 and the respective mean vectors $\underset{\sim}{\xi}_1$ and $\underset{\sim}{\xi}_2$ are set to (-1.5,-1.5)' and (1.5,1.5)'. The values of the covariance matrices Σ_1 and Σ_2 are also given in Table 6.1. Different values for Σ_1 and Σ_2 determine different shapes for $f(\underset{\sim}{\mu})$. The first three bivariate densities shown in Table 6.1 are plotted in Fig. 6.2 which show that they are bimodal. The last three densities (d) - (f) are plotted in Fig. 6.3 which show that $f(\underset{\sim}{\mu})$ is skewed. Note that the values for the covariance matrix given in the Table are arbitrary. The main theme is to investigate the assumption that the training data are grouped or not grouped in estimating the distribution of $\underset{\sim}{\mu}$. Further the choices of different covariance matrices provide a measure of the effect of non-Normality of the distribution of $\underset{\sim}{\mu}$ on the test of the

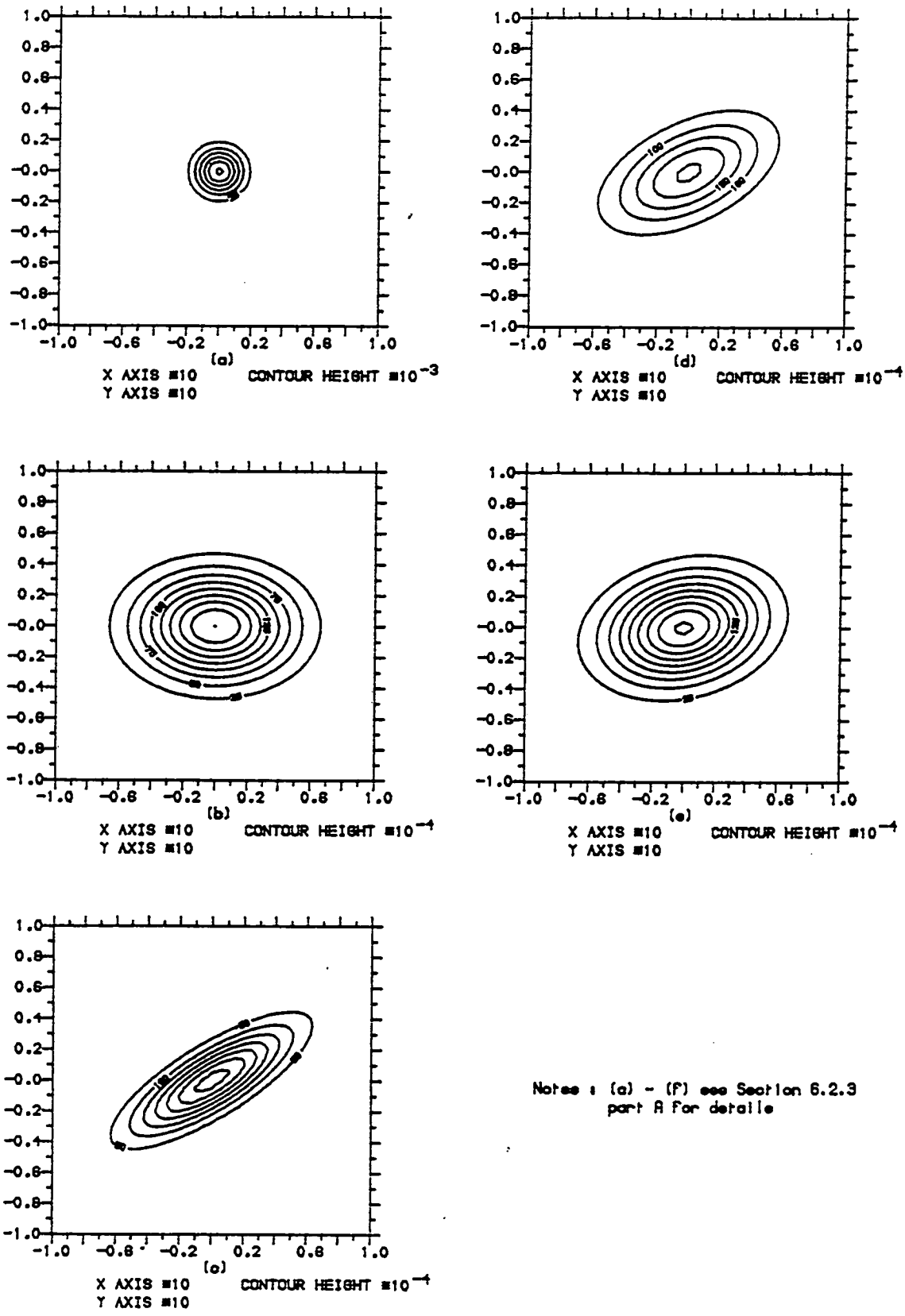
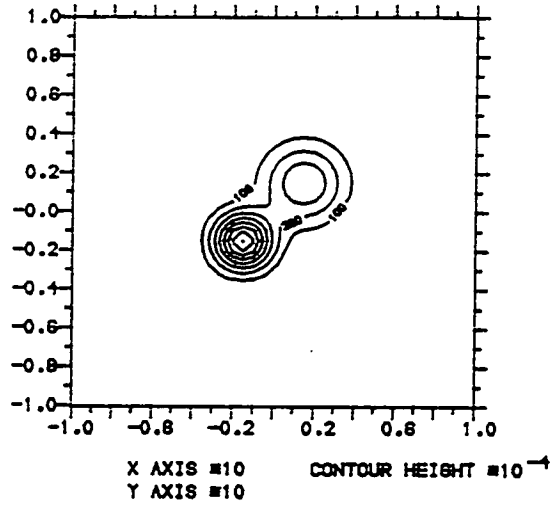
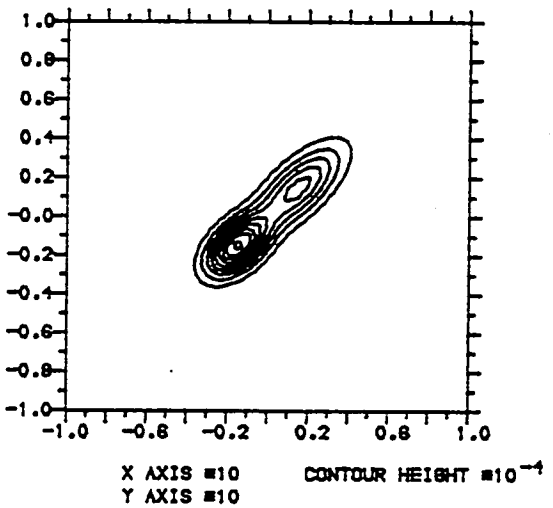


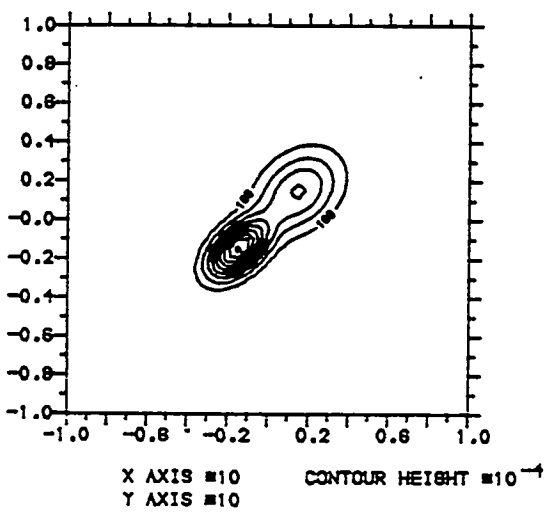
Fig. 6.1. Contour of the bivariate Normal density function of the true underlying between group distributions used in a simulation study in Section 6.2.3.



Case (a) of Section 6.2.3 part B

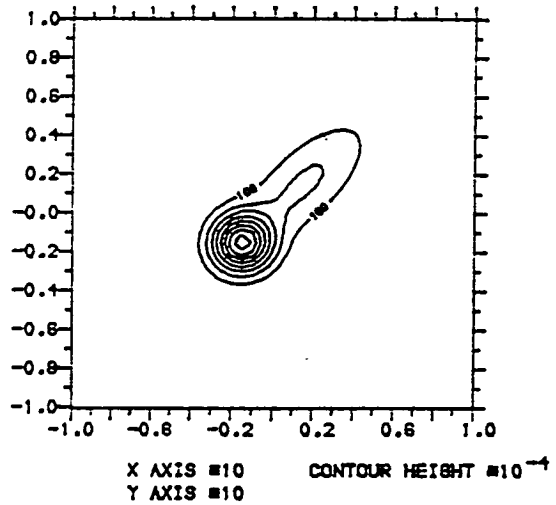


Case (b) of Section 6.2.3 part B

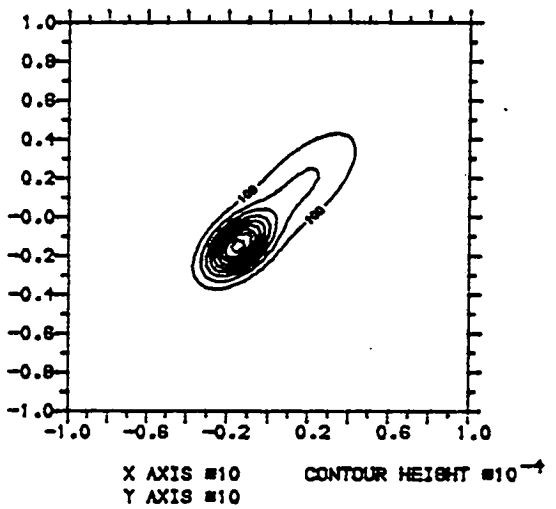


Case (c) of Section 6.2.3 part B

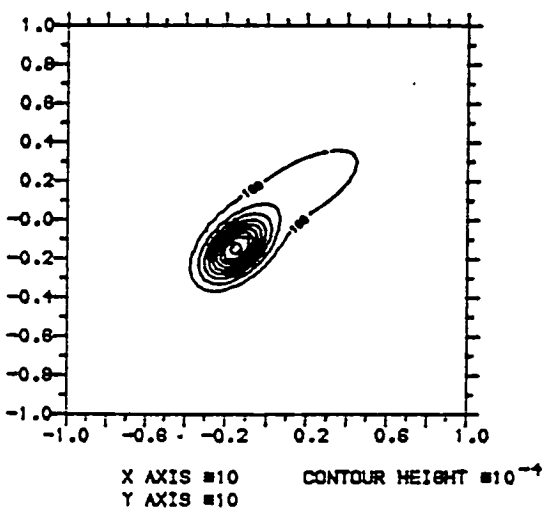
Fig. 6.2 Contour of the bivariate non-Normal density function of the true underlying between group distributions used in a simulation study in Section 6.2.3.



Case (d) of Section 6.2.3 part B



Case (e) of Section 6.2.3 part B



Case (f) of Section 6.2.3. part B

Fig. 6.3 Contour of the bivariate non-Normal density function of the true underlying between group distributions used in a simulation study in Section 6.2.3.

assumption.

Table 6.1 Summary of the choices of covariance matrices given in (6.7)

Part A. The distribution of $\underset{\sim}{\mu}_i$ is Normal

Case	Σ_1			Σ_2			Remarks
	σ_{i1}^2	σ_{i2}^2	ρ_1	σ_{21}^2	σ_{22}^2	ρ_1	
a	1	1	0.0	-	-	-	⎫
b	1	5	0.0	-	-	-	⎪
c	1	5	0.8	-	-	-	⎬ Unimodal
d	1	5	0.5	-	-	-	⎪
e	1	5	0.2	-	-	-	⎭

Part B. The distribution of $\underset{\sim}{\mu}_i$ is non-Normal

a	1	1	0.0	2	2	0.0	⎫
b	1	1	0.5	2	2	0.7	⎬ Bimodal
c	1	1	0.5	2	2	0.3	⎭
d	1	1	0.0	5	5	0.7	⎫
e	1	1	0.5	5	5	0.7	⎬ Skewed
f	1	1	0.5	10	5	0.7	⎭

Similar to the simulation study carried out in Chapter 3, the number of group sizes (n) are chosen to be 20, 50 and 100. The chosen number of within-group observations (J) are 1, 5 and 10. Integrated Square Error (ISE) is used as a measure of goodness of fit of the kernel density estimates. The ISE is chosen in favour of MISE (see Chapter 3) to reduce excessive computational time. Again, one simulation from each case described above is carried out.

The ISE of the kernel density estimate using the ordinary kernel and the adaptive kernel methods given the training data are grouped

and not grouped for the Normal case, are shown in Tables 6.2 - 6.6. Although one should not read too much into one set of simulation the Tables do suggest that in most cases the adaptive kernel method outshone the ordinary kernel method. In general using the group means to estimate $f(\underset{\sim}{\mu})$ gives smaller ISE, especially when n is large. The values of ISE decrease as number of within-group observations increases. Also that ISE decreases as group size increases. Note that when J = 1, the values of ISE are the same for the training data are grouped and not grouped.

Table 6.2 ISE ($\times 10^{-2}$) of the kernel density estimate given the training data is grouped and ungrouped, with diagonal covariance matrix using the ordinary and adaptive kernel methods. The underlying between groups distribution is a bivariate Normal with parameters as shown in Case (a) of Section 6.2.3.

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
20	1	2.5712	2.2107	2.5712	2.2107
	5	2.4920	1.8711	1.7647	1.4479
	10	2.2915	1.9549	2.0022	2.0922
50	1	1.7385	1.2206	1.7385	1.2206
	5	2.6722	2.2474	1.7243	1.2737
	10	2.3280	1.9194	1.0947	0.7471
100	1	2.2862	1.5953	2.2862	1.5953
	5	2.0037	1.4805	0.8317	0.5396
	10	1.8798	1.4539	0.8247	0.7582

Table 6.3 ISE ($\times 10^{-2}$) of the kernel density estimate given the training data is grouped and ungrouped, with diagonal covariance matrix using the ordinary and adaptive kernel methods. The underlying between groups distribution is a bivariate Normal with parameters as shown in Case (b) of Section 6.2.3.

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
20	1	0.4397	0.3520	0.4397	0.3520
	5	0.3313	0.3650	0.3421	0.2741
	10	0.1997	0.2705	0.1113	0.1294
50	1	0.0411	0.0573	0.0411	0.0573
	5	0.1316	0.1258	0.1310	0.1040
	10	0.1034	0.1271	0.0788	0.0623
100	1	0.1175	0.1284	0.1175	0.1284
	5	0.0442	0.0653	0.0279	0.0731
	10	0.0691	0.0639	0.0732	0.0508

Table 6.4 ISE ($\times 10^{-2}$) of the kernel density estimate given the training data is grouped and ungrouped, with diagonal covariance matrix using the ordinary and adaptive kernel methods. The underlying between groups distribution is a bivariate Normal with parameters as shown in Case (c) of Section 6.2.3.

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
20	1	0.7725	0.8146	0.7725	0.8146
	5	0.4279	0.3760	0.5033	0.3812
	10	0.3807	0.3293	0.6224	0.5028
50	1	0.4737	0.3416	0.4737	0.3416
	5	0.2149	0.1920	0.1782	0.1783
	10	0.3144	0.2589	0.3155	0.2359
100	1	0.4838	0.4111	0.4838	0.4111
	5	0.3213	0.2255	0.2895	0.1762
	10	0.1894	0.1627	0.1503	0.1435

Table 6.5 ISE ($\times 10^{-2}$) of the kernel density estimate given the training data is grouped and ungrouped, with diagonal covariance matrix using the ordinary and adaptive kernel methods. The underlying between groups distribution is a bivariate Normal with parameters as shown in Case (d) of Section 6.2.3.

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
20	1	0.3265	0.3267	0.3265	0.3267
	5	0.2519	0.2947	0.2646	0.1872
	10	0.3873	0.5217	0.2743	0.3209
50	1	0.0934	0.0615	0.0934	0.0615
	5	0.1611	0.1805	0.2141	0.3912
	10	0.0909	0.0960	0.0838	0.0819
100	1	0.2110	0.1551	0.2110	0.1551
	5	0.0621	0.0571	0.0619	0.0551
	10	0.1309	0.1346	0.1272	0.1189

Table 6.6 ISE ($\times 10^{-2}$) of the kernel density estimate given the training data is grouped and ungrouped, with diagonal covariance matrix using the ordinary and adaptive kernel methods. The underlying between groups distribution is a bivariate Normal with parameters as shown in Case (e) of Section 6.2.3.

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
20	1	0.2292	0.3217	0.2292	0.3217
	5	0.2575	0.2885	0.2180	0.2195
	10	0.3193	0.3553	0.2700	0.2558
50	1	0.1432	0.0852	0.1432	0.0852
	5	0.1351	0.1429	0.1250	0.1031
	10	0.1816	0.1943	0.1742	0.1799
100	1	0.1154	0.0707	0.1154	0.0707
	5	0.0880	0.0778	0.0896	0.0593
	10	0.0549	0.0759	0.0493	0.1072

The results of the simulation study for the non-Normal case are shown in Tables 6.7 - 6.12. The superiority of the adaptive kernel

method over the ordinary kernel method is clearly shown in most cases. The use of group means to estimate $f(\underset{\sim}{\mu})$ produces a smaller ISE when the group size is large. When n is small, the ungrouped model is better than the grouped model. This is probably due to the difference in the number of data points in constructing the kernel density estimate for $f(\underset{\sim}{\mu})$. For example, when n=20 and J=10 there are only twenty data points to estimate $f(\underset{\sim}{\mu})$ in the grouped model, comparing 200 data points used in the ungrouped model.

Table 6.7 ISE ($\times 10^{-2}$) of the kernel density estimate given the training data is grouped and ungrouped, with diagonal covariance matrix using the ordinary and adaptive kernel methods. The underlying between groups distribution is non-Normal as shown in Case (a) of Section 6.2.3.

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
20	1	1.2562	1.0492	1.2562	1.0492
	5	0.4881	0.4183	0.3509	0.6676
	10	0.7233	0.6895	0.6870	0.6714
50	1	0.7987	0.7467	0.7987	0.7467
	5	0.7657	0.5931	0.6003	0.4050
	10	0.4373	0.3140	0.2059	0.1540
100	1	0.7563	0.5108	0.7563	0.5108
	5	0.6671	0.5839	0.4315	0.3939
	10	0.4632	0.3293	0.2177	0.1794

Table 6.8 ISE ($\times 10^{-2}$) of the kernel density estimate given the training data is grouped and ungrouped, with diagonal covariance matrix using the ordinary and adaptive kernel methods. The underlying between groups distribution is non-Normal as shown in Case (b) of Section 6.2.3.

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
20	1	2.2577	2.0224	2.2577	2.0224
	5	1.3120	1.1187	1.0104	1.2547
	10	1.9345	1.9327	1.8439	2.1272
50	1	1.8202	1.5782	1.8202	1.5782
	5	1.2173	1.0170	0.9856	0.9735
	10	1.0767	0.9393	0.5166	0.4258
100	1	1.3346	1.1913	1.3346	1.1913
	5	1.0350	0.8105	0.3630	0.2723
	10	1.0701	0.9142	0.5936	0.5210

Table 6.9 ISE ($\times 10^{-2}$) of the kernel density estimate given the training data is grouped and ungrouped, with diagonal covariance matrix using the ordinary and adaptive kernel methods. The underlying between groups distribution is non-Normal as shown in Case (c) of Section 6.2.3.

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
20	1	1.2438	1.1948	1.2438	1.1948
	5	1.2295	1.0624	1.0694	1.0319
	10	0.7478	0.6790	0.8649	1.1449
50	1	1.5687	1.4843	1.5687	1.4843
	5	1.2725	1.1709	0.8929	0.8465
	10	0.9810	0.8130	0.4872	0.3201
100	1	1.2540	1.1557	1.2540	1.1557
	5	0.7652	0.5505	0.4159	0.2733
	10	0.7089	0.5569	0.3802	0.3329

Table 6.10 ISE ($\times 10^{-2}$) of the kernel density estimate given the training data is grouped and ungrouped, with diagonal covariance matrix using the ordinary and adaptive kernel methods. The underlying between groups distribution is non-Normal as shown in Case (d) of Section 6.2.3.

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
20	1	1.1636	1.0038	1.1636	1.0038
	5	0.9956	0.7641	1.0341	0.7059
	10	1.1489	1.0909	1.2002	1.1168
50	1	1.2881	0.9251	1.2881	0.9251
	5	0.8712	0.7121	0.8796	0.6433
	10	0.6280	0.4906	0.4546	0.3544
100	1	0.6609	0.4103	0.6609	0.4103
	5	0.6402	0.5136	0.4410	0.4372
	10	0.5756	0.4696	0.2098	0.1523

Table 6.11 ISE ($\times 10^{-2}$) of the kernel density estimate given the training data is grouped and ungrouped, with diagonal covariance matrix using the ordinary and adaptive kernel methods. The underlying between groups distribution is non-Normal as shown in Case (e) of Section 6.2.3.

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
20	1	1.4533	1.1390	1.4533	1.1390
	5	1.4808	1.4216	1.4260	1.4504
	10	1.1965	1.0940	1.3046	1.1918
50	1	1.2815	1.0737	1.2815	1.0737
	5	1.0452	0.7214	0.5088	0.2630
	10	1.1598	0.9671	0.9257	0.8844
100	1	1.1182	0.8526	1.1182	0.8526
	5	1.0059	0.7642	0.9169	0.5460
	10	0.7844	0.6341	0.4184	0.1823

Table 6.12 ISE ($\times 10^{-2}$) of the kernel density estimate given the training data is grouped and ungrouped, with diagonal covariance matrix using the ordinary and adaptive kernel methods. The underlying between groups distribution is non-Normal as shown in Case (f) of Section 6.2.3.

Assume TS		Ungrouped		Grouped	
n	J	Ordinary	Adaptive	Ordinary	Adaptive
20	1	1.4577	1.5653	1.4577	1.5653
	5	1.0329	0.7964	1.7628	1.4650
	10	0.7685	0.6709	1.0320	0.7552
50	1	1.3003	1.1834	1.3003	1.1834
	5	0.8935	0.6784	0.8412	0.7412
	10	0.7530	0.5755	0.5362	0.2769
100	1	1.2405	0.9351	1.2405	0.9351
	5	0.9014	0.6095	0.8257	0.4388
	10	0.8027	0.6406	0.5356	0.2369

6.3 Predictive distribution

Estimation of the Bayes' factor (3.3) in a multivariate case involves the derivation of a p-dimensional predictive distribution. In this section the predictive distribution will be derived for the grouped training data case using the ordinary kernel method under the random effects model with non-Normal random factor.

The predictive distributions under the 'fixed effects' model have been derived by Geisser (1964,1966) under various assumptions and are widely applied in Discriminant analysis. Fatti (1982) adopted the Geisser method and derived expressions for the predictive distribution under the random effects model in a discrimination problem. He applied the usual assumptions of the random effects model, that is, the within group population is characterised by a p-dimensional Normal distribution $N_p(\underset{\sim}{\mu}_i,\Sigma)$, where $\underset{\sim}{\mu}_i$ has been

randomly selected from a $N_p(\underset{\sim}{\xi},T)$ distribution. This is in contrast to the "fixed effects" model where the distribution of $\underset{\sim}{\mu}_i$ is assumed to be vague. Fatti also assumed that Σ and the two "hyper-parameters" $\underset{\sim}{\xi}$ and T are unknown and have a joint noninformative prior distribution. In the following section the predictive distribution of an observation or a summary of a set of observations is derived under the random effects model without the assumption that $f(\underset{\sim}{\mu}_i)$ comes from a specific parametric family. The formulae for the Bayes' factor for the other cases such as ungrouped training data and adaptive kernel method will be given in Appendix 6.

6.3.1 Derivation of the predictive distribution

The predictive distribution is derived without the assumption that the random factor $\underset{\sim}{\mu}_i$ is Normally distributed. The distribution of $\underset{\sim}{\mu}_i$ is estimated by the method discussed in Section 6.2 under various assumptions about the training data. Estimation or determination of the distribution of the hyper-parameters $\underset{\sim}{\xi}$ and T, arising from the Normality assumption about the distribution of the between population means, is no longer required. The within group covariance matrix is assumed known and its known value is obtained from the training data and is given by

$$\Sigma = \frac{1}{n(J-1)} \sum_{i=1}^{n} \sum_{j=1}^{J} (\underset{\sim}{z}_{ij}-\underset{\sim}{\bar{z}}_{i.})(\underset{\sim}{z}_{ij}-\underset{\sim}{\bar{z}}_{i.})'.$$

With a similar situation and terminology as in Chapter 3, suppose that the control and recovered data consist of a set of p-dimensional vector observations. Let $\mathbf{Y} = \{\underset{\sim}{y}_1', \underset{\sim}{y}_2', \ldots, \underset{\sim}{y}_r'\}'$ and $\mathbf{X} = \{\underset{\sim}{x}_1', \underset{\sim}{x}_2', \ldots, \underset{\sim}{x}_m'\}'$ denote $r \times p$ and $m \times p$ data matrices of the recovered

and control samples, respectively. Suppose the $\underset{\sim}{Y}_i$'s and $\underset{\sim}{X}_j$'s are independent identically distributed p-dimensional Normal with unknown mean vector $\underset{\sim}{\mu}_y$ and $\underset{\sim}{\mu}_x$, and covariance matrices Σ_y and Σ_x respectively.

Under the hypothesis C, that $\underset{\sim}{Y}$ and $\underset{\sim}{X}$ are from the same source, $\underset{\sim}{Y}_i$'s and $\underset{\sim}{X}_j$'s are p-dimensional Normally distributed with unknown common mean vector $\underset{\sim}{\mu}_C$, say and covariance matrix Σ, say. Let $\bar{\underset{\sim}{Y}}$ and $\bar{\underset{\sim}{X}}$ denote the mean vector of samples of sizes r and m of the recovered and control data respectively. Extension of the model developed in Section 3.5.2 shows that the numerator of the BF in (3.3) is a p-dimensional predictive distribution of $\bar{\underset{\sim}{Y}}$ given $\bar{\underset{\sim}{X}}$ and is proportional to

$$\int f_p(\bar{\underset{\sim}{Y}}|\underset{\sim}{\mu}_C)\ f_p(\bar{\underset{\sim}{X}}|\underset{\sim}{\mu}_C)\ f_p(\underset{\sim}{\mu}_C)\ d\underset{\sim}{\mu}_C$$

where, assuming the within group covariance matrix Σ is known,

$$f_p(\bar{\underset{\sim}{Y}}|\underset{\sim}{\mu}) = \frac{|r\Sigma^{-1}|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{r}{2}(\bar{\underset{\sim}{y}}-\underset{\sim}{\mu})'\Sigma^{-1}(\bar{\underset{\sim}{y}}-\underset{\sim}{\mu})\right\} \qquad (6.8)$$

and

$$f_p(\bar{\underset{\sim}{X}}|\underset{\sim}{\mu}) = \frac{|m\Sigma^{-1}|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{m}{2}(\bar{\underset{\sim}{x}}-\underset{\sim}{\mu})'\Sigma^{-1}(\bar{\underset{\sim}{x}}-\underset{\sim}{\mu})\right\}. \qquad (6.9)$$

are the sampling distributions of $\bar{\underset{\sim}{Y}}$ and of $\bar{\underset{\sim}{X}}$ respectively. Combining (6.1), (6.8) and (6.9), using the identity (which is reproduced here in Appendix 6) given by Box and Tiao (1973) for combining two quadratic forms and the fact that S and Σ are positive definite, and by integrating over $\underset{\sim}{\mu}$ the predictive distribution of $\bar{\underset{\sim}{Y}}$ given $\bar{\underset{\sim}{X}}$, $f_p(\bar{\underset{\sim}{Y}}|\bar{\underset{\sim}{X}},C)$, is proportional to

$$\frac{1}{|a^2\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2a^2}(\bar{\underset{\sim}{x}}-\bar{\underset{\sim}{y}})'\Sigma^{-1}(\bar{\underset{\sim}{x}}-\bar{\underset{\sim}{y}})\right\} \times$$

$$\frac{1}{|A_w|^{\frac{1}{2}}\, n} \sum_{i=1}^{n} \exp\left\{-\frac{1}{2}(\underset{\sim}{w}-\bar{\underset{\sim}{z}}_i)'A_w^{-1}(\underset{\sim}{w}-\bar{\underset{\sim}{z}}_i)\right\} \qquad (6.10)$$

where $A_w = (\Sigma_w + S_\lambda)$, $a^2 = (m^{-1}+r^{-1})$, $S_\lambda = \lambda^2 S$, $\Sigma_w = (m+r)^{-1}\Sigma$ and $\underset{\sim}{w} = (m\bar{\underset{\sim}{x}} + r\bar{\underset{\sim}{y}})/(m+r)$.

The normalising constant can be obtained by combining (6.1) and (6.9) and integrating over $\underset{\sim}{\mu}$ to give

$$f_p(\bar{\underset{\sim}{x}}|C) = \frac{1}{(2\pi)^{p/2}\, |A_x|^{\frac{1}{2}}\, n} \sum_{i=1}^{n} \exp\left\{-\frac{1}{2}(\bar{\underset{\sim}{x}}-\bar{\underset{\sim}{z}}_i)'A_x^{-1}(\bar{\underset{\sim}{x}}-\bar{\underset{\sim}{z}}_i)\right\} \qquad (6.11)$$

where $A_x = (\Sigma_x + S_\lambda)$, $\Sigma_x = m^{-1}\Sigma$ and S_λ as above.

Similarly, the denominator of the BF, $f(\bar{\underset{\sim}{y}}|\bar{C})$ is given by

$$f_p(\bar{\underset{\sim}{y}}|\bar{C}) = \frac{1}{(2\pi)^{p/2}\, |A_y|^{\frac{1}{2}}\, n} \sum_{i=1}^{n} \exp\left\{-\frac{1}{2}(\bar{\underset{\sim}{y}}-\bar{\underset{\sim}{z}}_i)'A_y^{-1}(\bar{\underset{\sim}{y}}-\bar{\underset{\sim}{z}}_i)\right\} \qquad (6.12)$$

where $A_y = (\Sigma_y + S_\lambda)$, $\Sigma_y = r^{-1}\Sigma$ and S_λ as above.

Note that if (6.3) is used in place of (6.1) then the matrix S_λ after the equations (6.10), (6.11) and (6.12) will be replaced by D_λ as defined in Section 6.2. It is easily seen that a similar expression of the Bayes' factor for the assumed 'ungrouped' training data case can be obtained simply by replacing the group means $\bar{z}_i$, $(i=1,\ldots,n)$, by z_ℓ $(\ell=1,\ldots,N)$, S_λ by S'_λ with the appropriate smoothing parameter estimate.

6.4 Interpretation of the Bayes' factor

In Chapter 3 the behaviour of the Bayes' factor could easily be presented as a function of the control and recovered data. However in the multivariate situation inspection of the behaviour of the Bayes' factor is difficult even in the bivariate case.

Since the Bayes' factor is a measure of evidence provided by the data for the hypothesis C against $\bar{C}$ we can use a similar 'order of magnitude' suggested by Jeffrey (1939/83) to interpret the Bayes' factor based on the logarithm scale. The ordering below shows a verbal scale for the order of magnitude in favour of C, a converse of the one given by Jeffery (see Section 2.2.6) which showed the order of magnitude against the hypothesis C:

$BF < 1$	evidence against C
$1 < BF < 10^{\frac{1}{2}}$	very slight evidence for C
$10^{\frac{1}{2}} < BF < 10$	moderate evidence for C
$10 < BF < \sqrt[3]{10}$	strong evidence for C
$\sqrt[3]{10} < BF < 10^2$	very strong evidence for C
$10^2 < BF$	decisive evidence for C

These orderings are used to represent the scale of support for the hypothesis C implied by the evidence in the example below.

6.5 Examples

Suppose there is available a set of bivariate hair measurements from 22 cats with 10 samples from each cat. The two variables of

interest are hair width and medullary fraction. Fig. 6.4 shows a bivariate dot plot of (a) the 220 observations and (b) the 22 group means. Figs. 6.5 and 6.6 show contours of the bivariate density estimate for the observations displayed in Fig. 6.4 using the respective ordinary and adaptive kernel methods described in Section 6.2 with the two separate assumptions that the sample covariance matrix S is non-diagonal or diagonal. From Figs 6.5 and 6.6, the estimated distribution of the parameter vector $\underset{\sim}{\mu}$ have slight differences when the training data are assumed grouped and then not grouped. There are also small differences in the estimated distribution of $\underset{\sim}{\mu}$ when assuming the covariance matrix S is diagonal or non-diagonal. There are not significant differences between the ordinary and adaptive kernel methods. Note that the contour heights in Fig. 6.5 is multiplied by 10^{-3} and in Fig. 6.6, the multiplier is 10^{-4}.

In the next two sections the values of the Bayes' factor are shown as a function of a vector $\underset{\sim}{\bar{Y}}$ given some vector values of $\underset{\sim}{\bar{X}}$. The values for the first and second variables of $\underset{\sim}{\bar{X}}$ are chosen to be in this case 10.0(30.0)100.0 and 0.4(0.2)0.8, respectively.

6.5.1 The ordinary kernel method

This section describes the results of the evaluation of the Bayes' factor presented in Section 6.3 using the ordinary method. The cases when the training data are grouped and not grouped are considered and also when the covariance matrix is diagonal and non-diagonal.

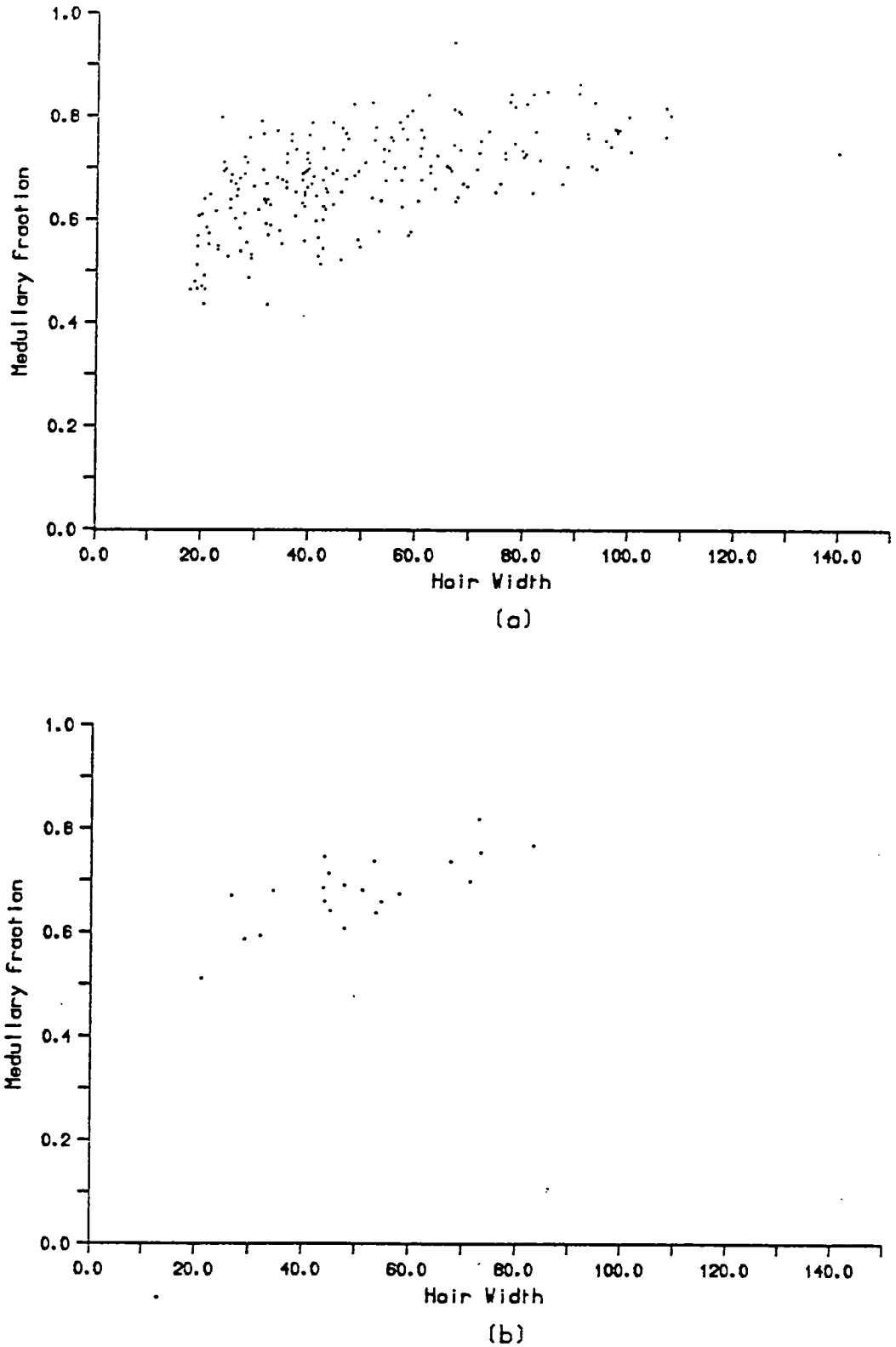


Fig. 6.4 A bivariate dot plot of (a) 220 cat hairs and (b) 22 group means

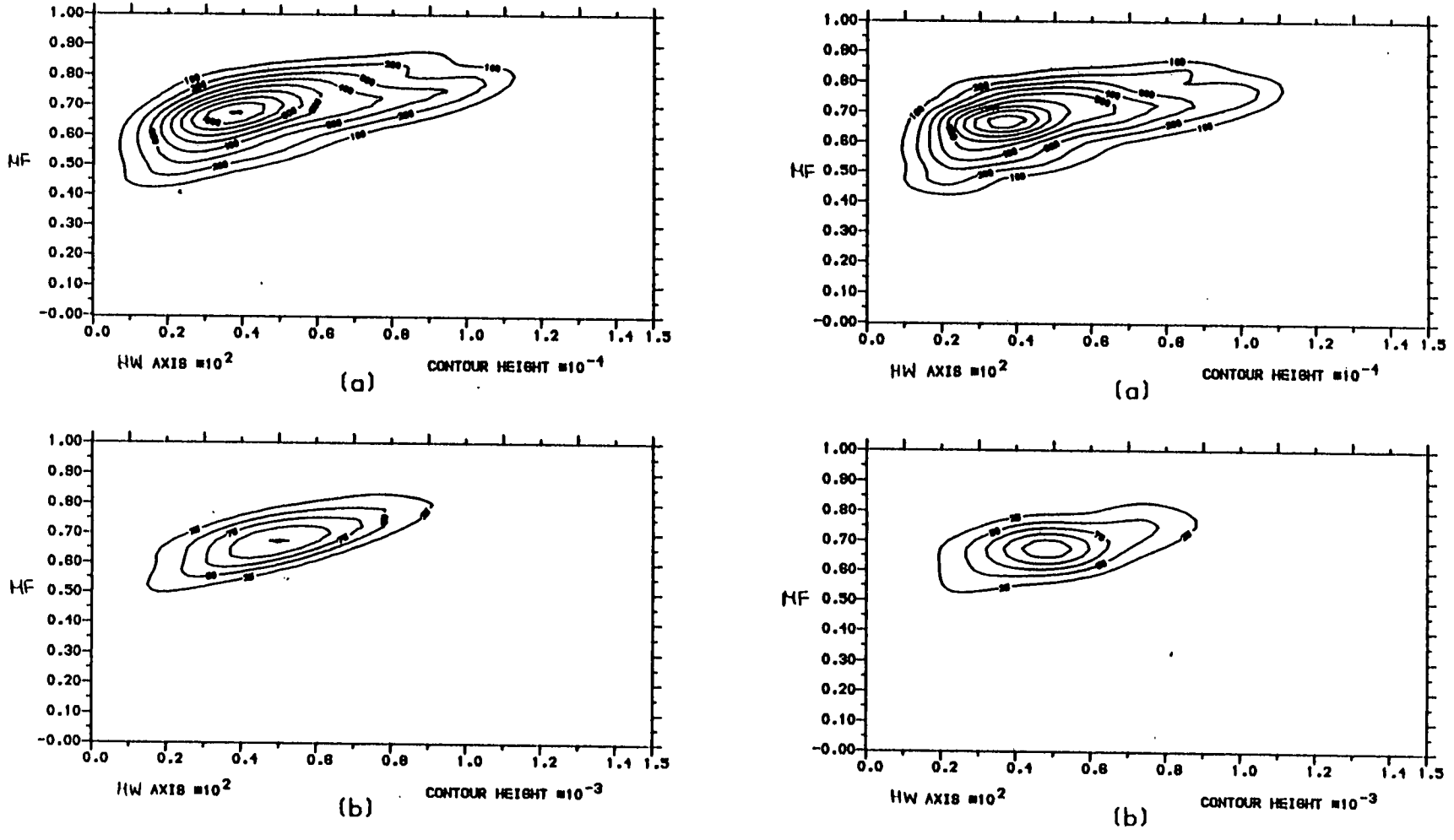


Fig. 6.5 Contour of $\hat{f}_a$ bivariate ordinary kernel density estimate for (a) 220 cat hairs and (b) 22 group means assuming the sample covariance matrix S Non-diagonal(left) or Diagonal(right).

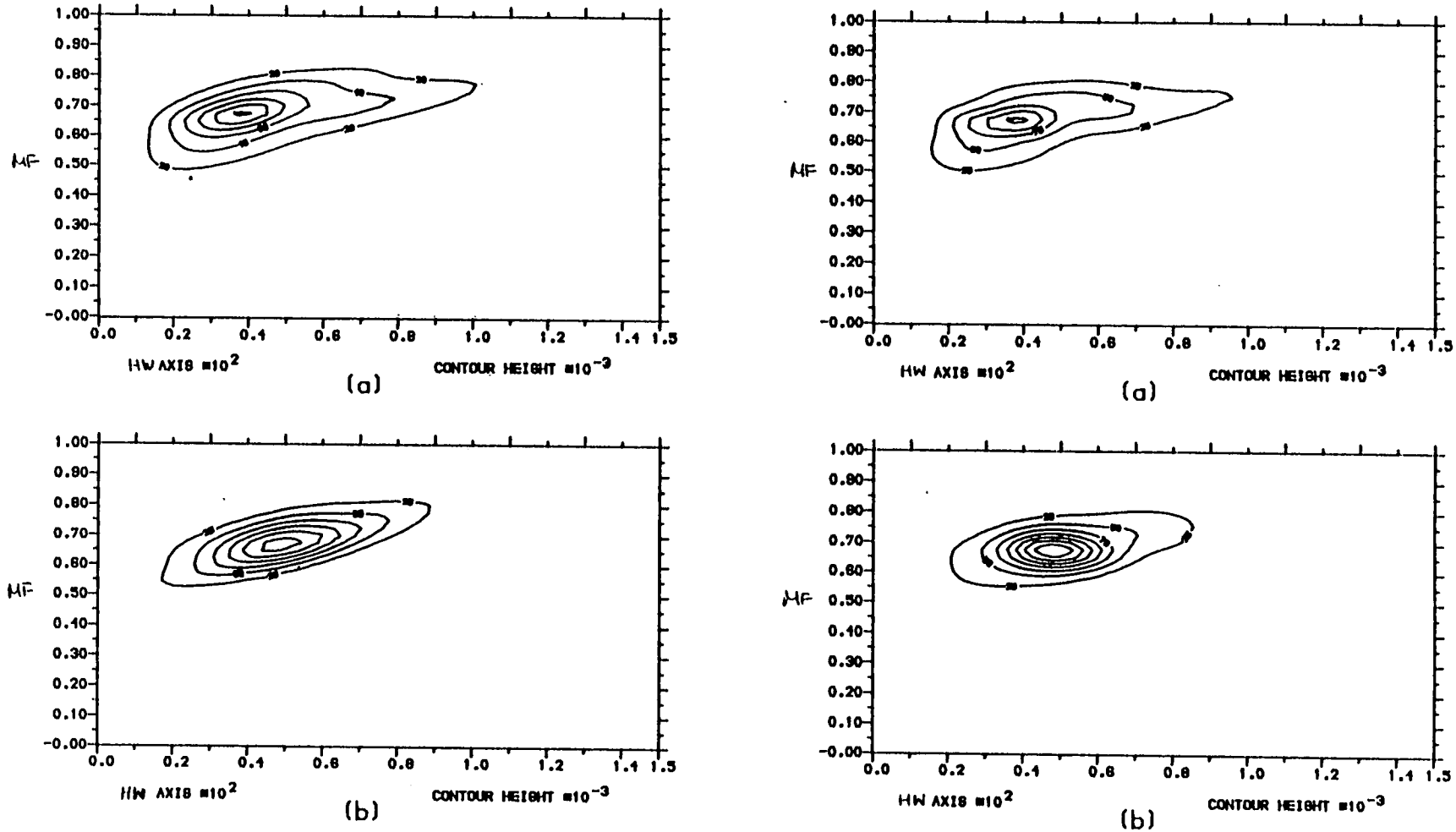


Fig. 6.6 Contour of a bivariate adaptive kernel density estimate for (a) 220 cat hairs and (b) 22 group means assuming the sample covariance matrix S Non-diagonal(left) or Diagonal(right).

Using the grouping of the BF as in Section 6.4 the values of the Bayes' factor given the training data is grouped and assuming that the covariance matrix S is non-diagonal (Fig. 6.7) and diagonal (Fig. 6.8) are illustrated in a graphical form for r=1,m=1. Figs 6.9 and 6.10 show the behaviour of the Bayes' factor given the training data are ungrouped and assuming that the covariance matrix S is non-diagonal and diagonal respectively for r=1,m=1. From Fig. 6.7, it appears that only slight or moderate evidence in support of the hypothesis C (i.e. there is a contact between the suspect and the crime scene) occurs when $\underset{\sim}{X}'$ = (10.0,0.6), (40.0,0.6), (40.0,0.8), (70.0,0.8), (70.0,0.6) and (100.0,0.8). From Fig. 6.5(b), these values are relatively common, especially $\underset{\sim}{X}'$ = (40.0,0.6) which is most common. There is little difference between the cases when the covariance matrix S non-diagonal and diagonal. Moreover, the assumption of matrix S being diagonal seems to improve the behaviour of the Bayes' factor as far as the location of the maximum of the BF is concerned. There also seems to be more indecisive support in the hypothesis C when common values of $\underset{\sim}{X}$ and of $\underset{\sim}{Y}$ are observed. The area of weak evidence can be seen clearly from Figs. 6.7 and 6.8.

When the number of control and recovered data increase to ten each, given the training data are grouped, Figs 6.11 and 6.12 show the behaviour of the Bayes' factor assuming the covariance matrix S is non-diagonal or diagonal, respectively. The most significant feature between the different values of r and m is that it is difficult to find evidence in favour of the hypothesis C. But when the evidence in favour of the hypothesis C is found the values of the Bayes' factor are extremely large. The region of doubt in supporting

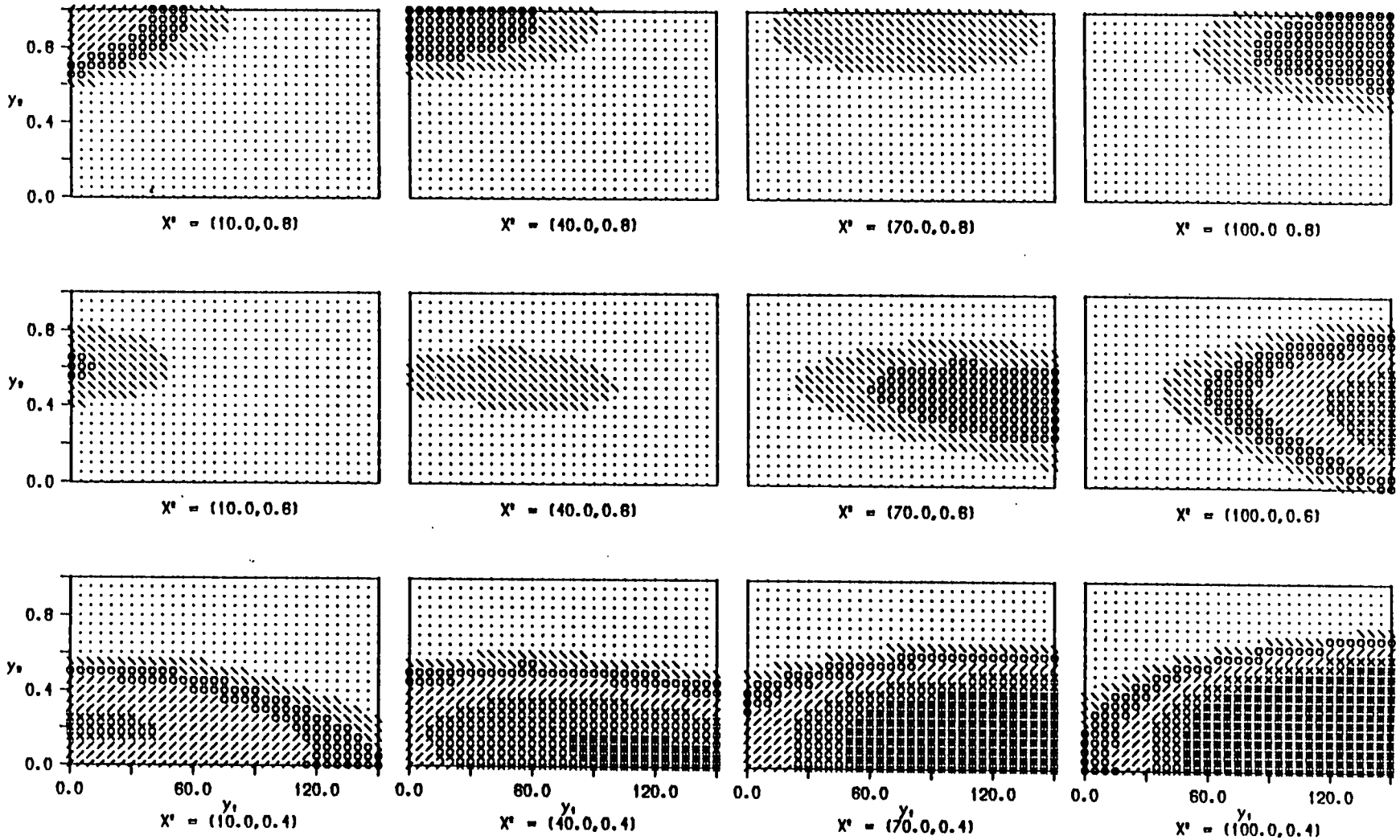


Fig. 6.7 Contour of Bayes' factor as a function of $\underline{Y}$, given $\underline{X}^t = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; for r=1, m=1 - Grouped model.
(· -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and ■ -- BF>10.0^2)

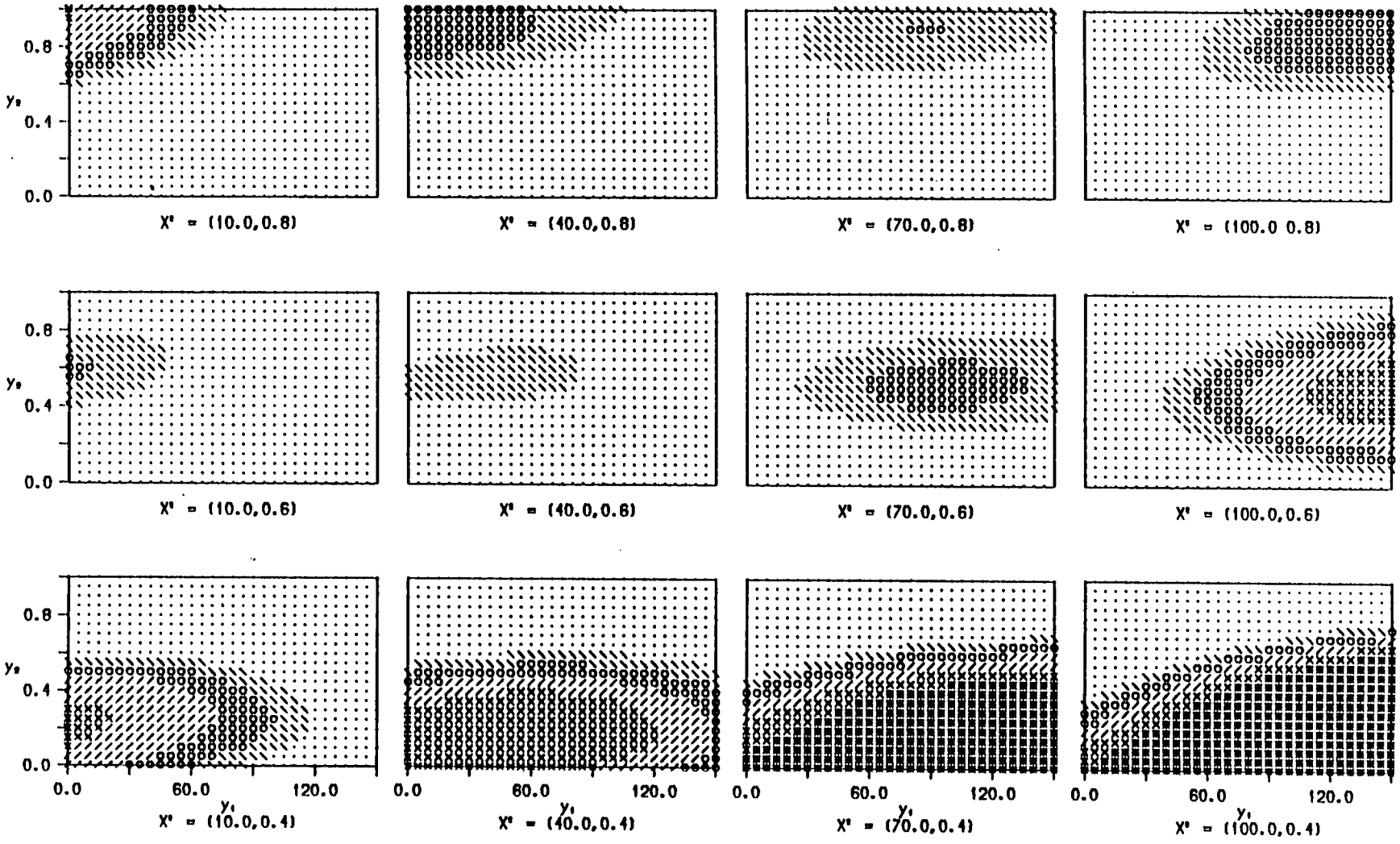


Fig. 6.8 Contour of Bayes' factor as a function of $\underline{Y}$, given $\underline{X}^0 = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; for r=1, m=1 - Grouped model.
(· -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and ■ -- BF>10.0^2)

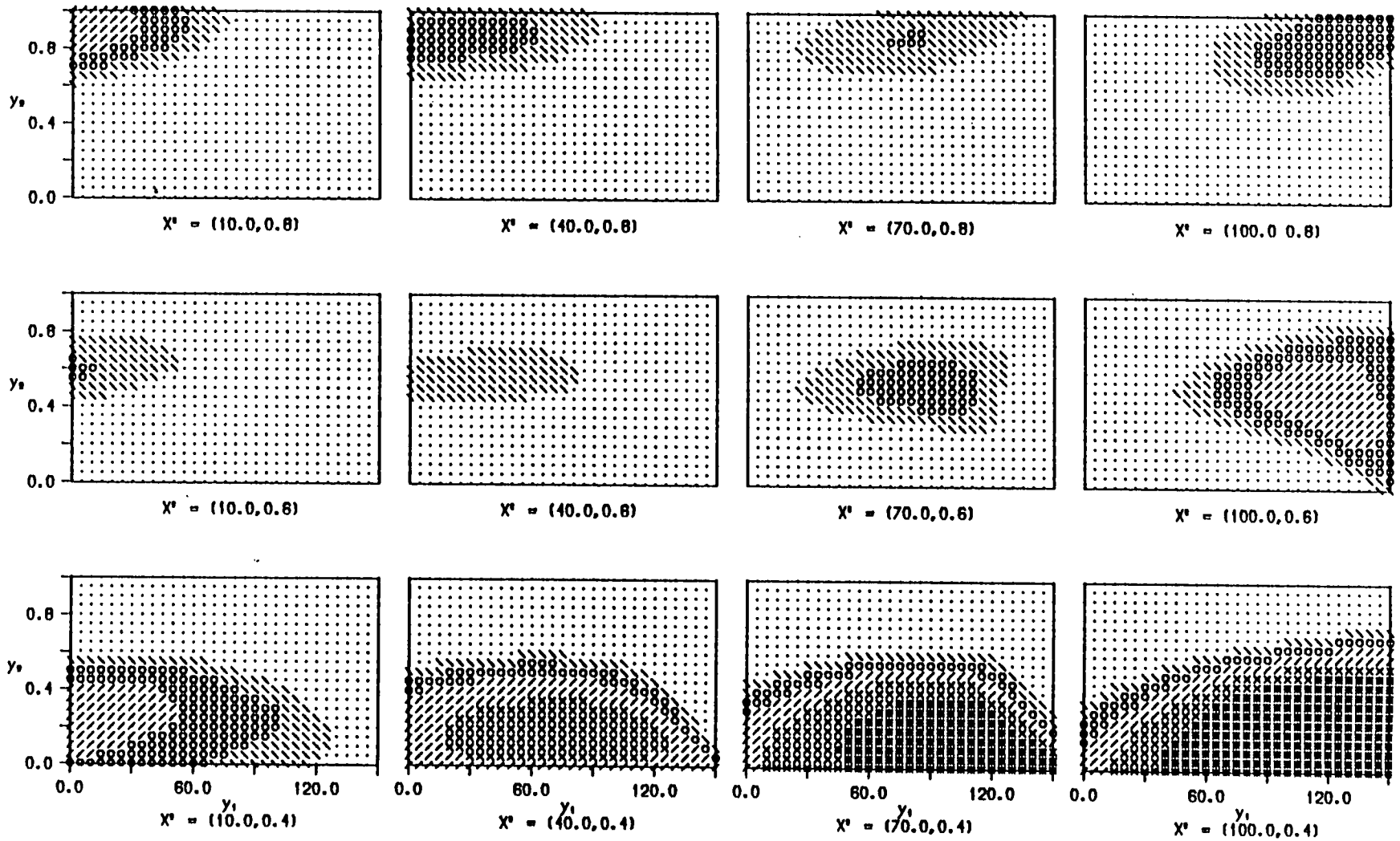


Fig. 6.9 Contour of Bayes' Factor as a function of $\underline{Y}$, given $\underline{X}^0 = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; For r=1, m=1 - Nogrouped model.
(· -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{1/2}$, x -- $10.0^{1/2}$<BF<10.0^{2} and ■ -- BF>10.0^{2})

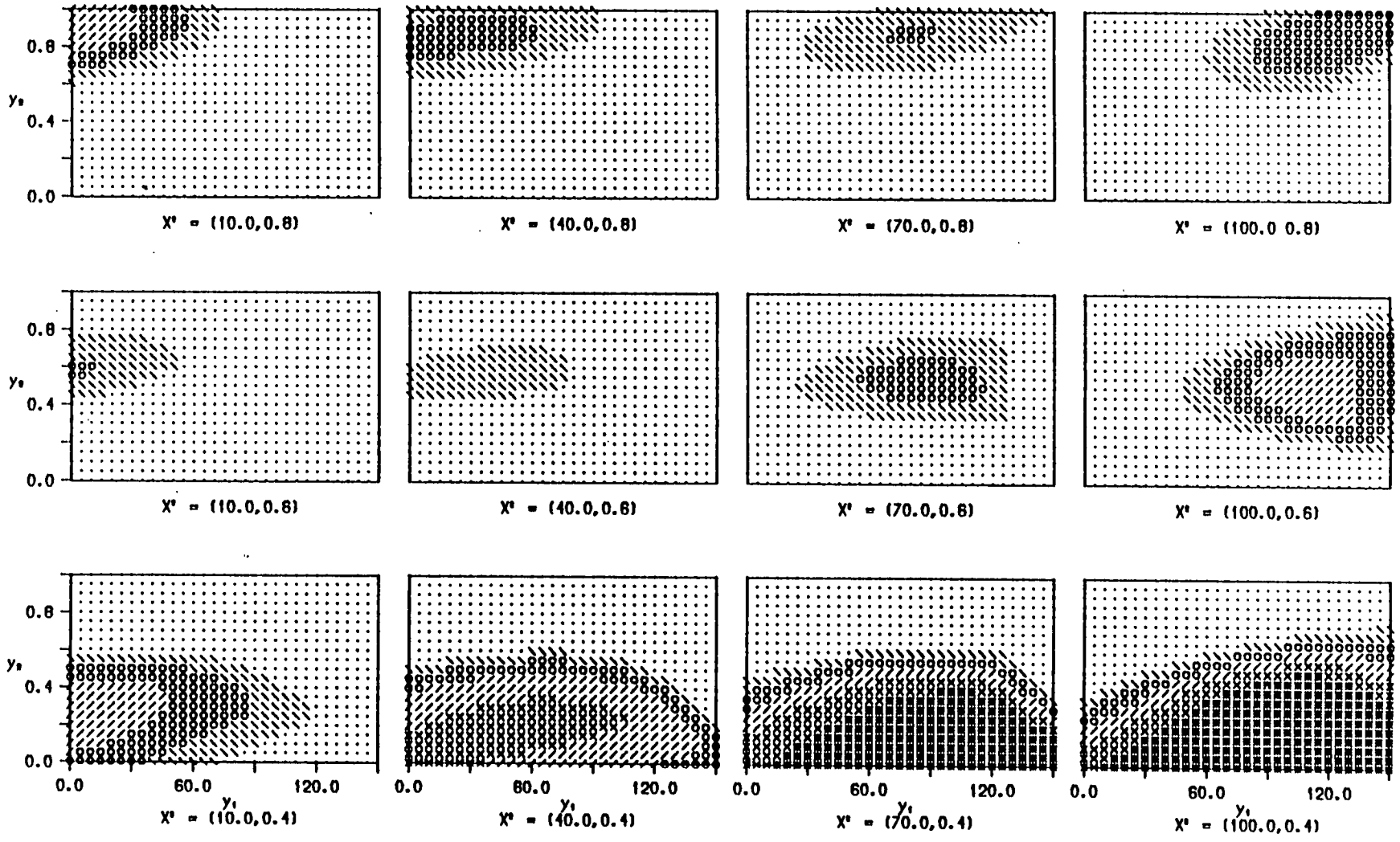


Fig. 6.10 Contour of Bayes' Factor as a Function of $\underline{Y}$, given $\underline{X}^0 = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; For r=1, m=1 - Nogrouped model.
(. -- BF<1.0, \ -- $1.0<BF<10^{1/2}$, o -- $10^{1/2}<BF<10.0$, / -- $10.0<BF<10.0^{3/2}$, x -- $10.0^{3/2}<BF<10.0^2$ and ■ -- $BF>10.0^2$)

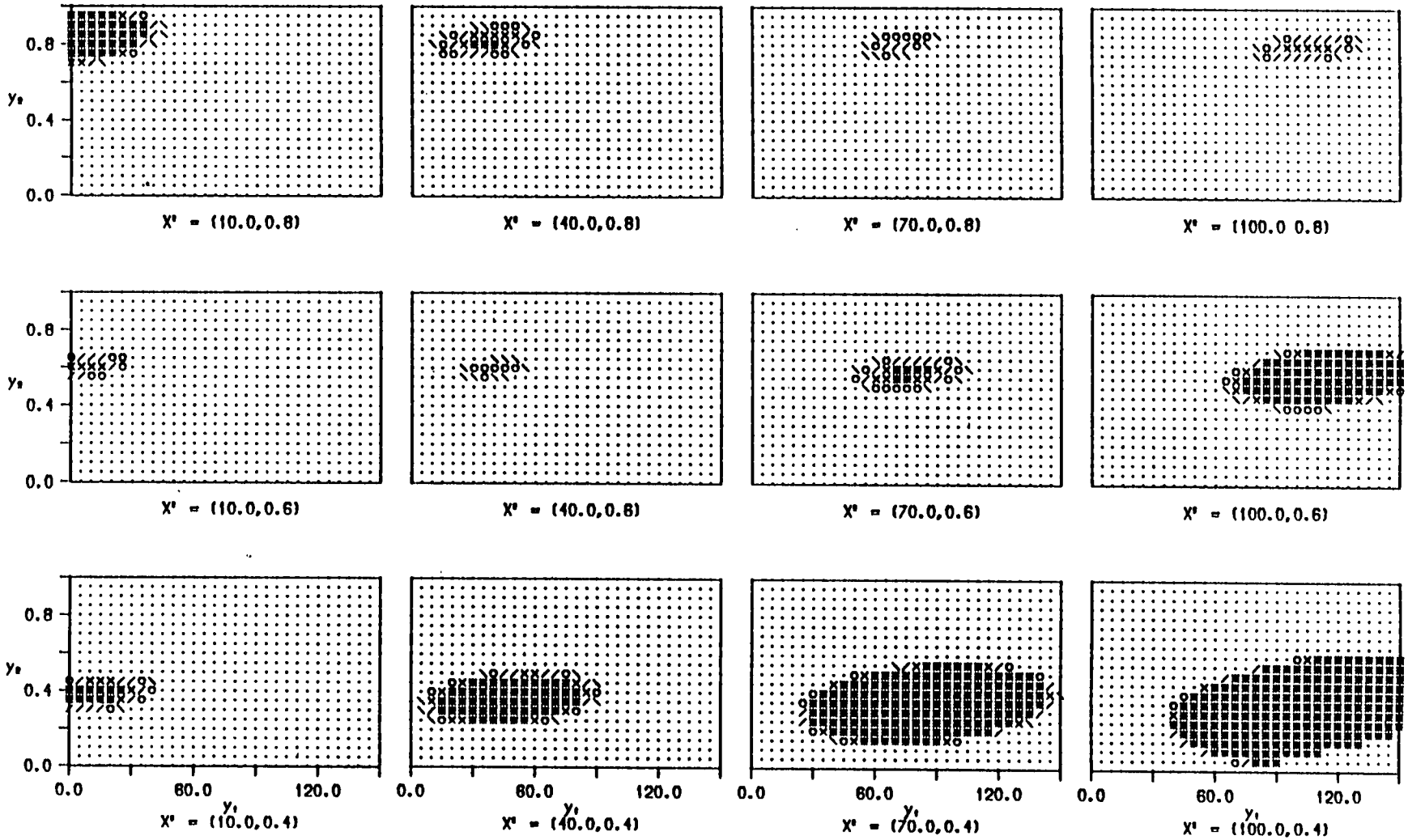


Fig. 6.11 Contour of Bayes' factor as a function of $\underline{Y}$, given $\underline{X}^0 = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; for r=10, m=10 - Grouped model.
(· -- BF<1.0, \ -- $1.0<BF<10^{1/2}$, o -- $10^{1/2}<BF<10.0$, / -- $10.0<BF<10.0^{3/2}$, x -- $10.0^{3/2}<BF<10.0^{2}$ and ■ -- $BF>10.0^{2}$)

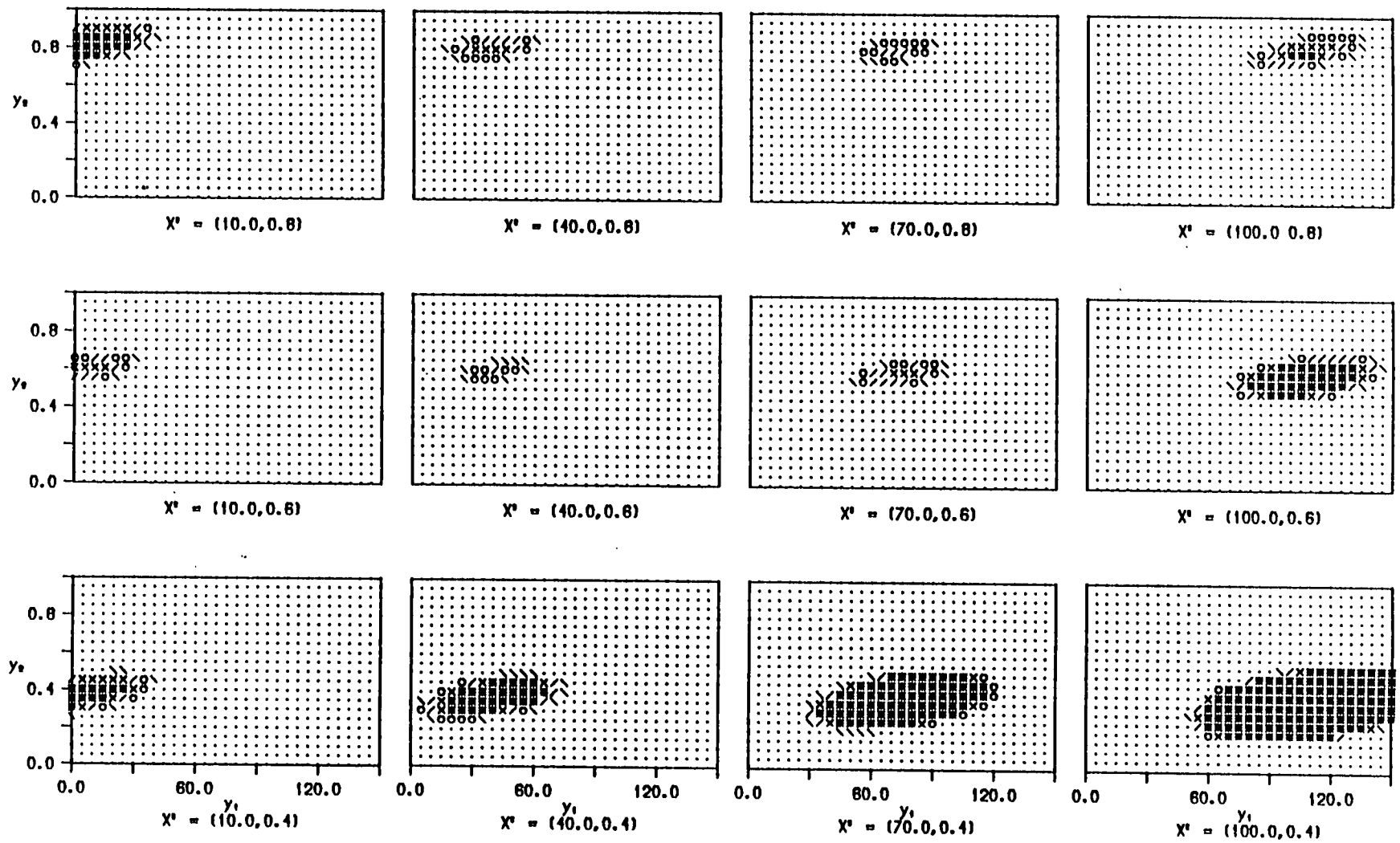


Fig. 6.12 Contour of Bayes' factor as a function of $\underline{Y}$, given $\underline{X}^0 = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; For r=10, m=10 - Grouped model.
(· -- BF<1.0, \ -- $1.0<BF<10^{1/2}$, o -- $10^{1/2}<BF<10.0$, / -- $10.0<BF<10.0^{3/2}$, x -- $10.0^{3/2}<BF<10.0^{2}$ and ■ -- BF>10.0^{2})

the hypothesis C is almost non-existent. The results suggest that when r and m are large, the evidence implied by the control and recovered observations is conclusive either for or against the hypothesis C. Figs 6.13 - 6.14 show the behaviour of the BF, given the training data are not grouped, assuming the covariance matrix S is non-diagonal and diagonal, respectively.

6.5.2 The adaptive kernel method

This section describes the results of the evaluation of the Bayes' factor described in Section 6.3 using the adaptive method. The cases when the training data are grouped and not grouped are considered and also when the covariance matrix is diagonal and non-diagonal.

Again using the ordering presented in the Section 6.4, the behaviour of the Bayes' factor given the training data are grouped and assuming that the covariance matrix S is non-diagonal (Fig. 6.15) and diagonal (Fig. 6.16) are illustrated in a graphical form for r=1,m=1. When these graphs are compared with Figs 6.7 and 6.8 for Section 6.5.1 it is found that the adaptive kernel method improves the behaviour of the Bayes' factor with regard to the position of the maximum of the BF. Similarly the adaptive kernel method provides improvement for the behaviour of the Bayes' factor when the training data is assumed ungrouped. See Figs 6.17 and 6.18 for this case.

Figs 6.19 and 6.20 show the behaviour of the Bayes' factor given the training data are grouped and assuming the covariance matrix S is non-diagonal and diagonal respectively for r=10,m=10. Comparing

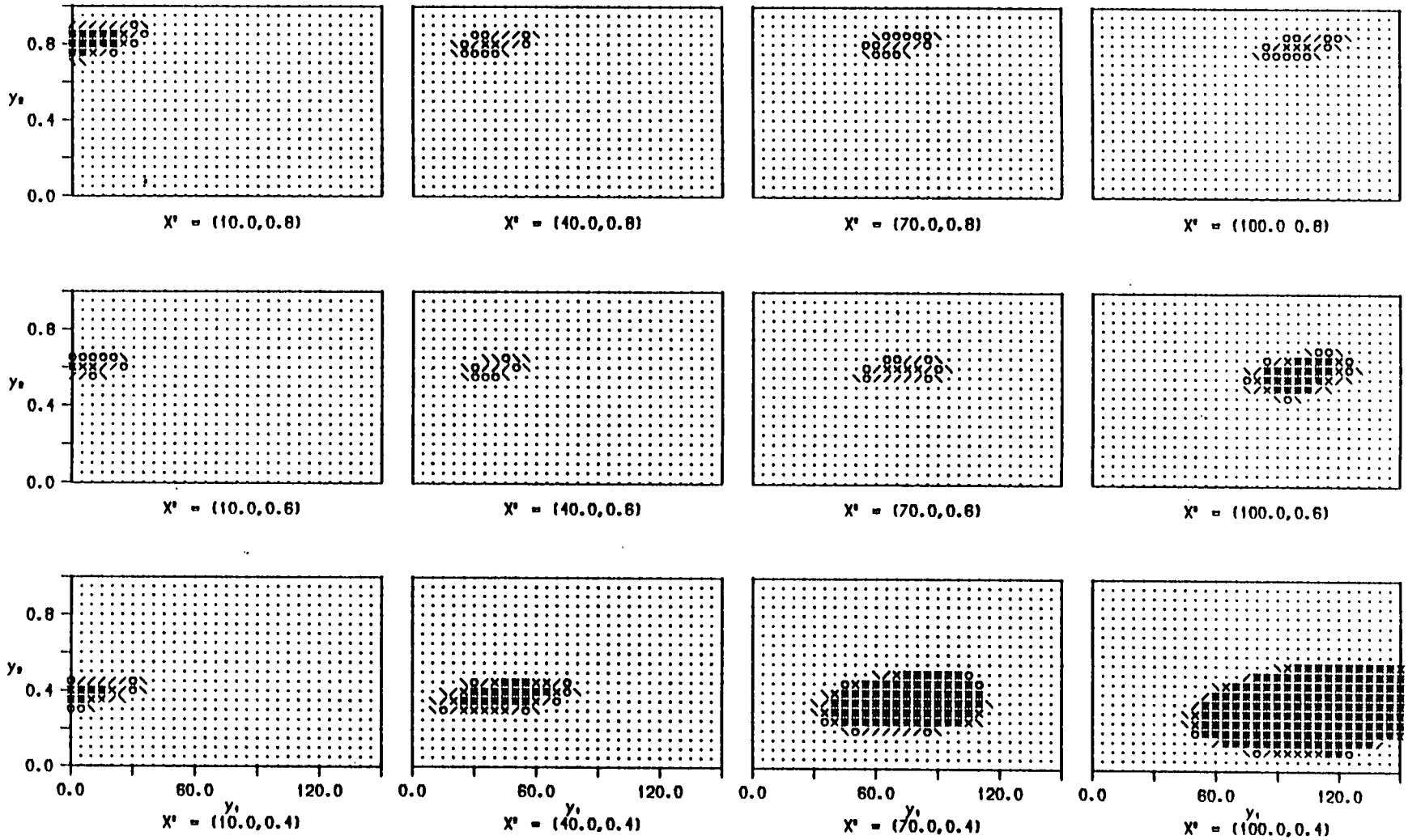


Fig. 6.13 Contour of Bayes' Factor as a function of $\underline{Y}$, given $\underline{X}' = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; For r=10, m=10 - Nogrouped model.

(· -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^{2} and ■ -- BF>10.0^{2})

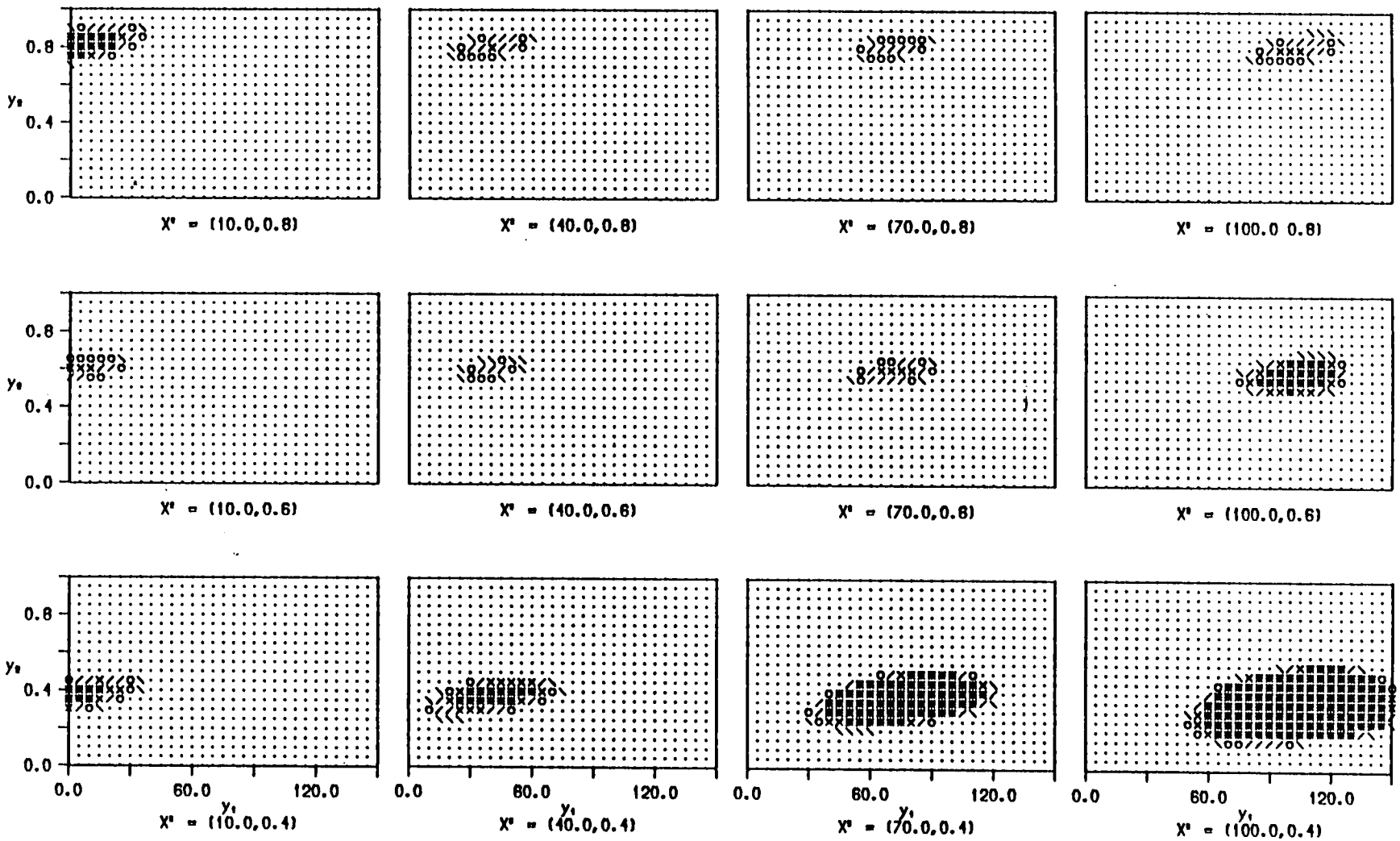


Fig. 6.14 Contour of Bayes' factor as a function of $\underline{Y}$, given $\underline{X}' = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; for r=10, m=10 - Nogrouped model.

(. -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and ■ -- BF>10.0^2)

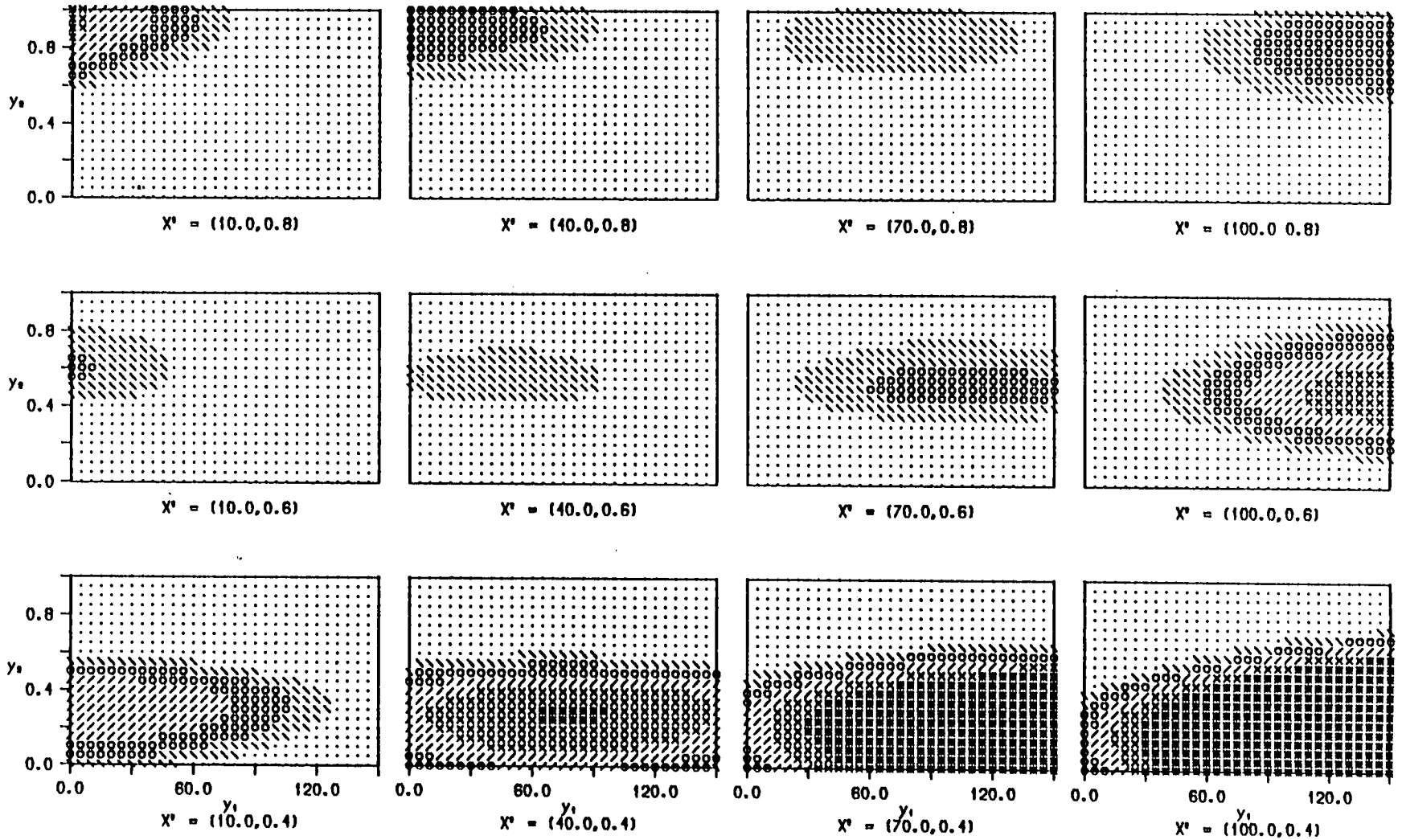


Fig. 6.15 Contour of Bayes' Factor as a function of $\underline{Y}$, given $\underline{X}^T = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; For r=1, m=1 - Grouped model.
(. -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and * -- BF>10.0^2)

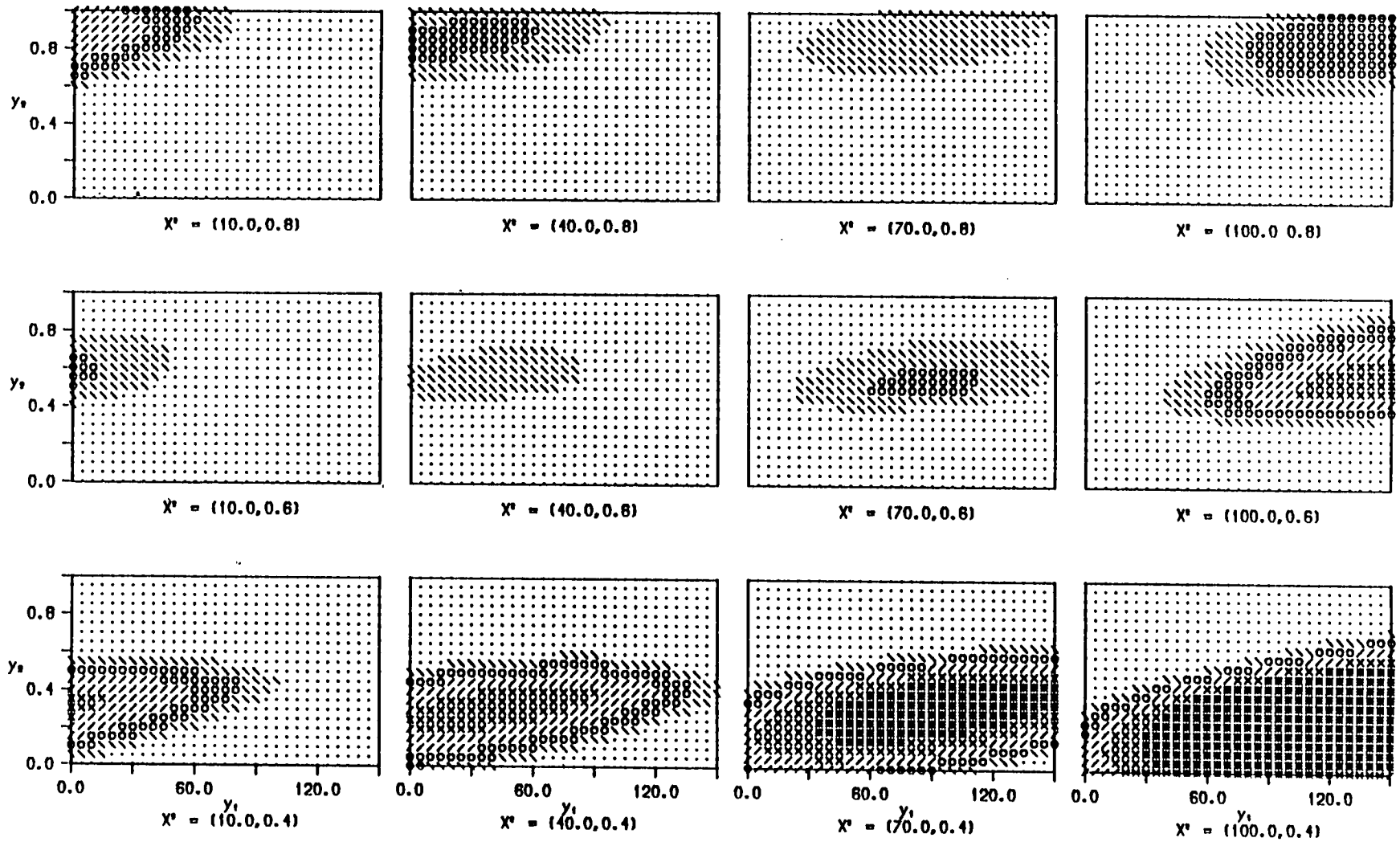


Fig. 6.16 Contour of Bayes' Factor as a function of $\underline{Y}$, given $\underline{X}^t = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; For r=1, m=1 - Grouped model.
(. -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and ■ -- BF>10.0^2)

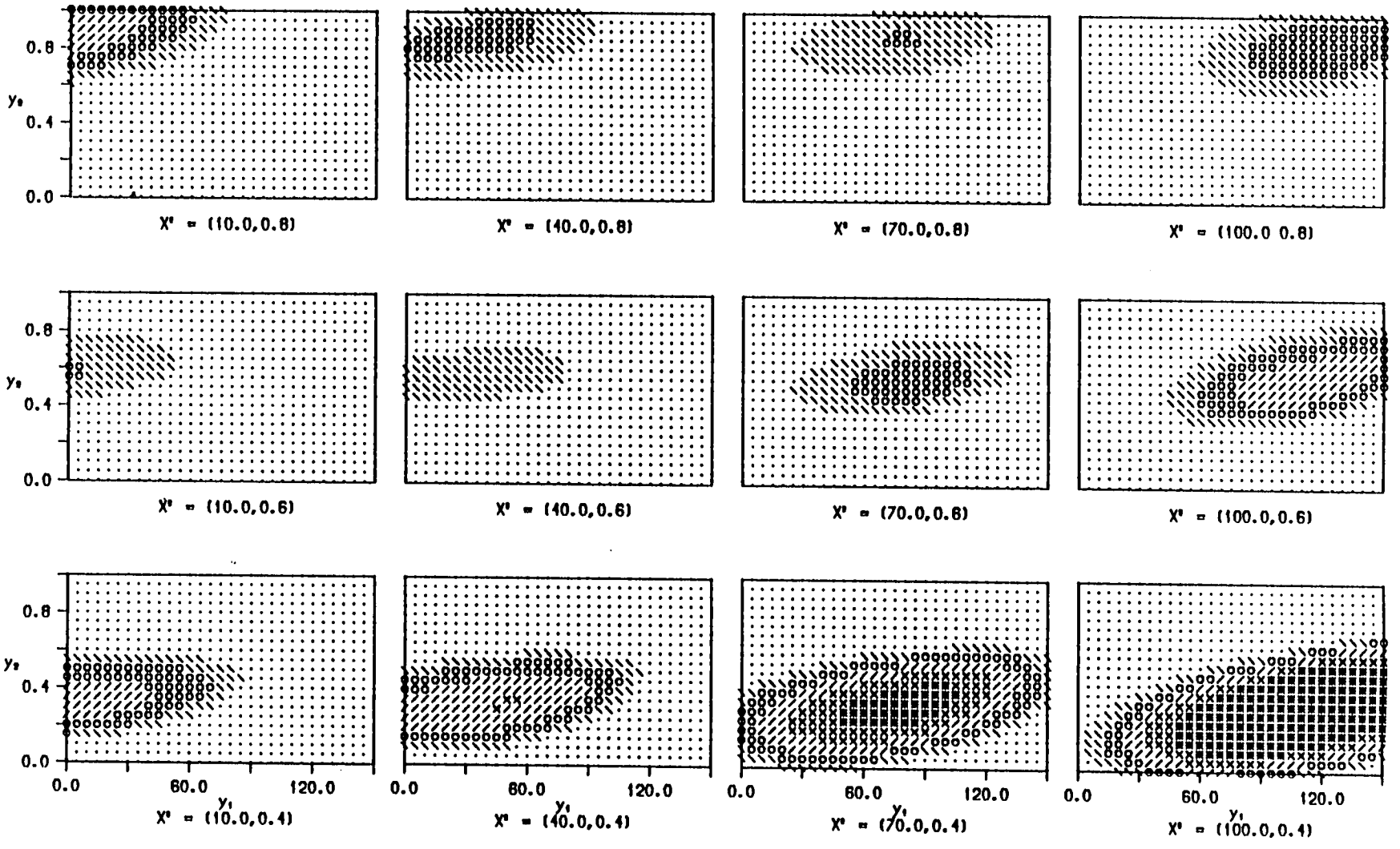


Fig. 6.17 Contour of Bayes' Factor as a function of $\underline{Y}$, given $\underline{X}^0 = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; For r=1, m=1 - Ungrouped model.
(· -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and ■ -- BF>10.0^2)

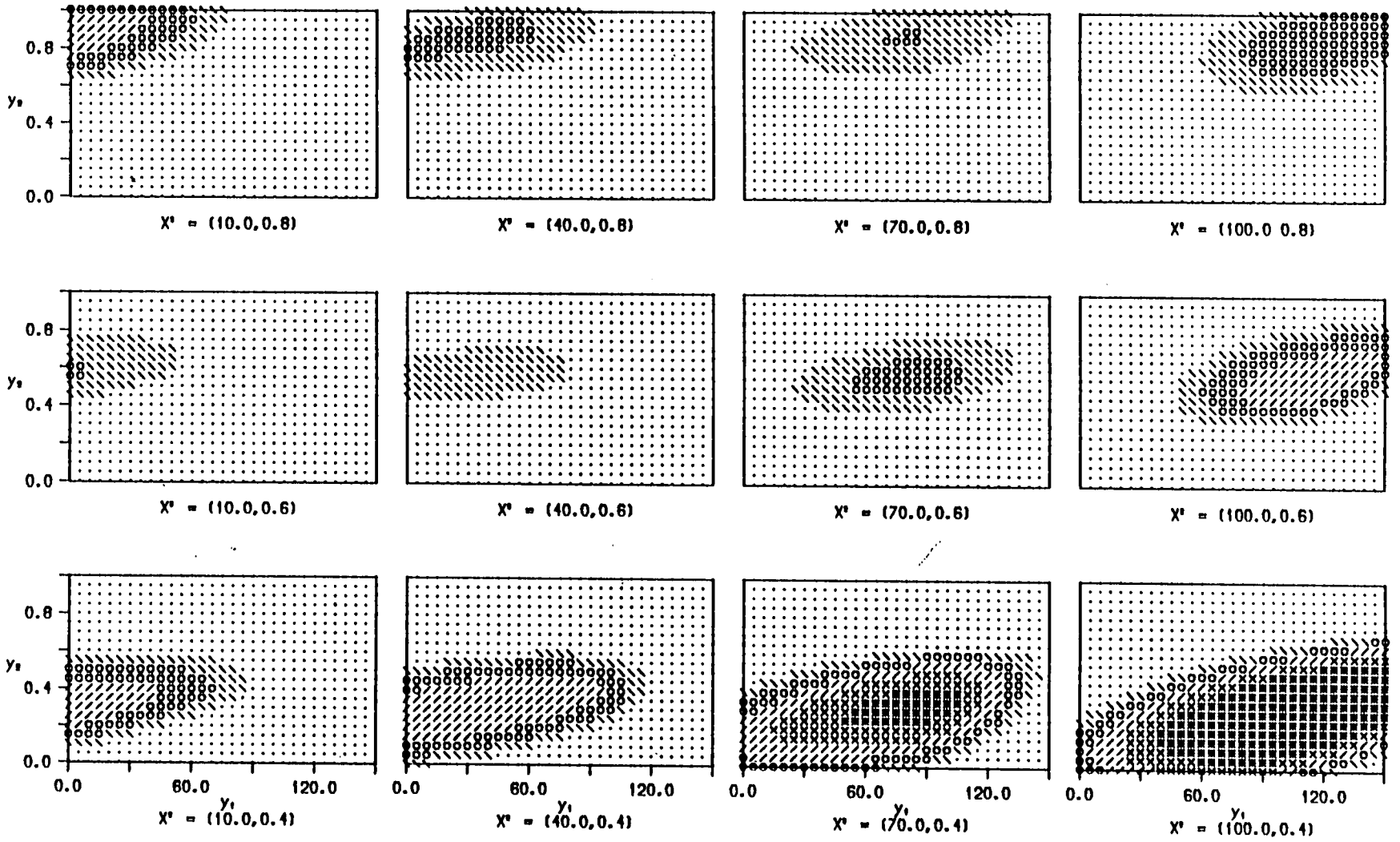


Fig.6.18 Contour of Bayes' Factor as a function of $\underline{Y}$, given $\underline{X}^t = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; For r=1, m=1 - Ungrouped model.
(. -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and ■ -- BF>10.0^2)

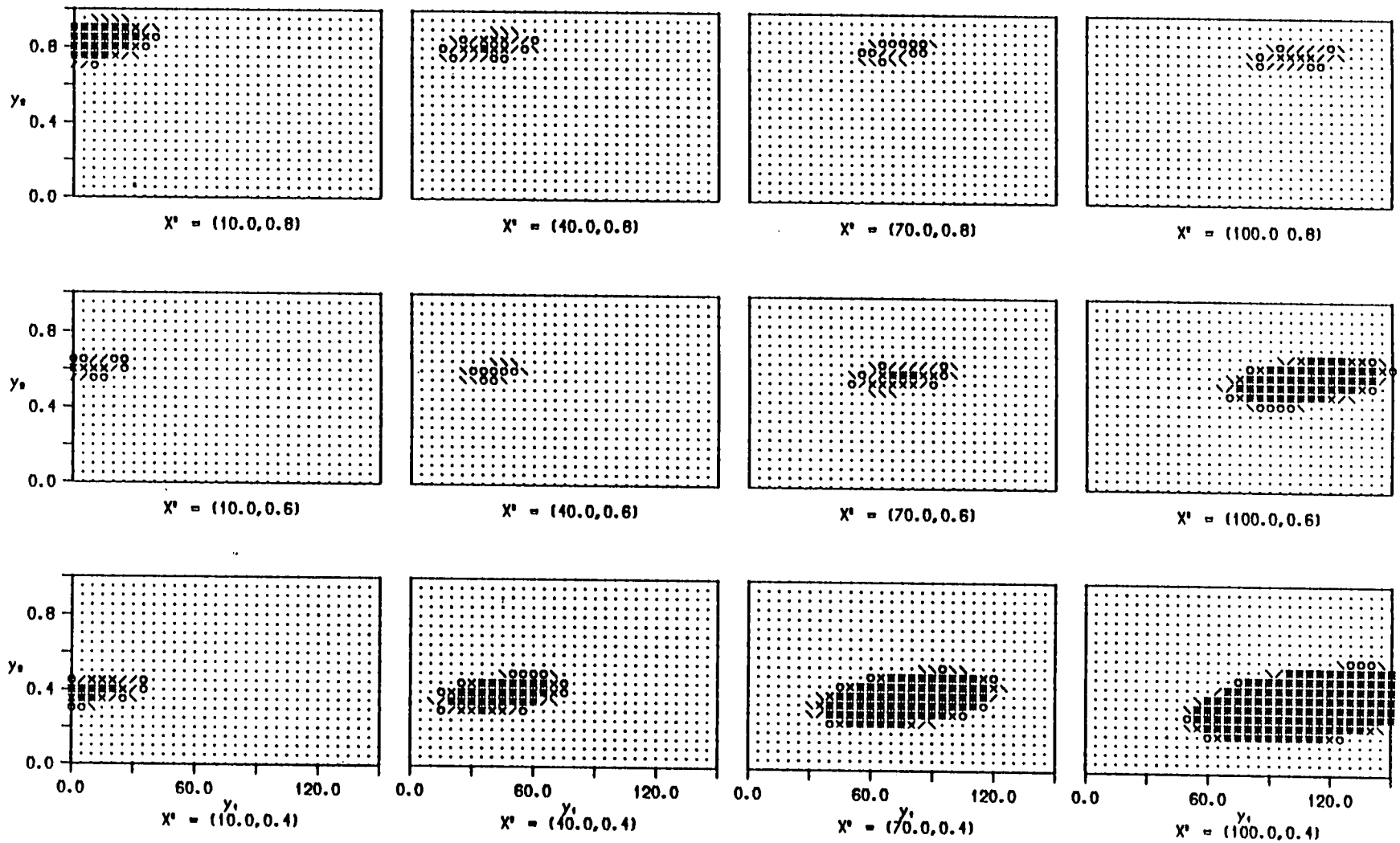


Fig. 6.19 Contour of Bayes' Factor as a function of $\underline{Y}$, given $\underline{X}^t = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; for r=10, m=10 - Grouped model.
(· -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and ■ -- BF>10.0^2)

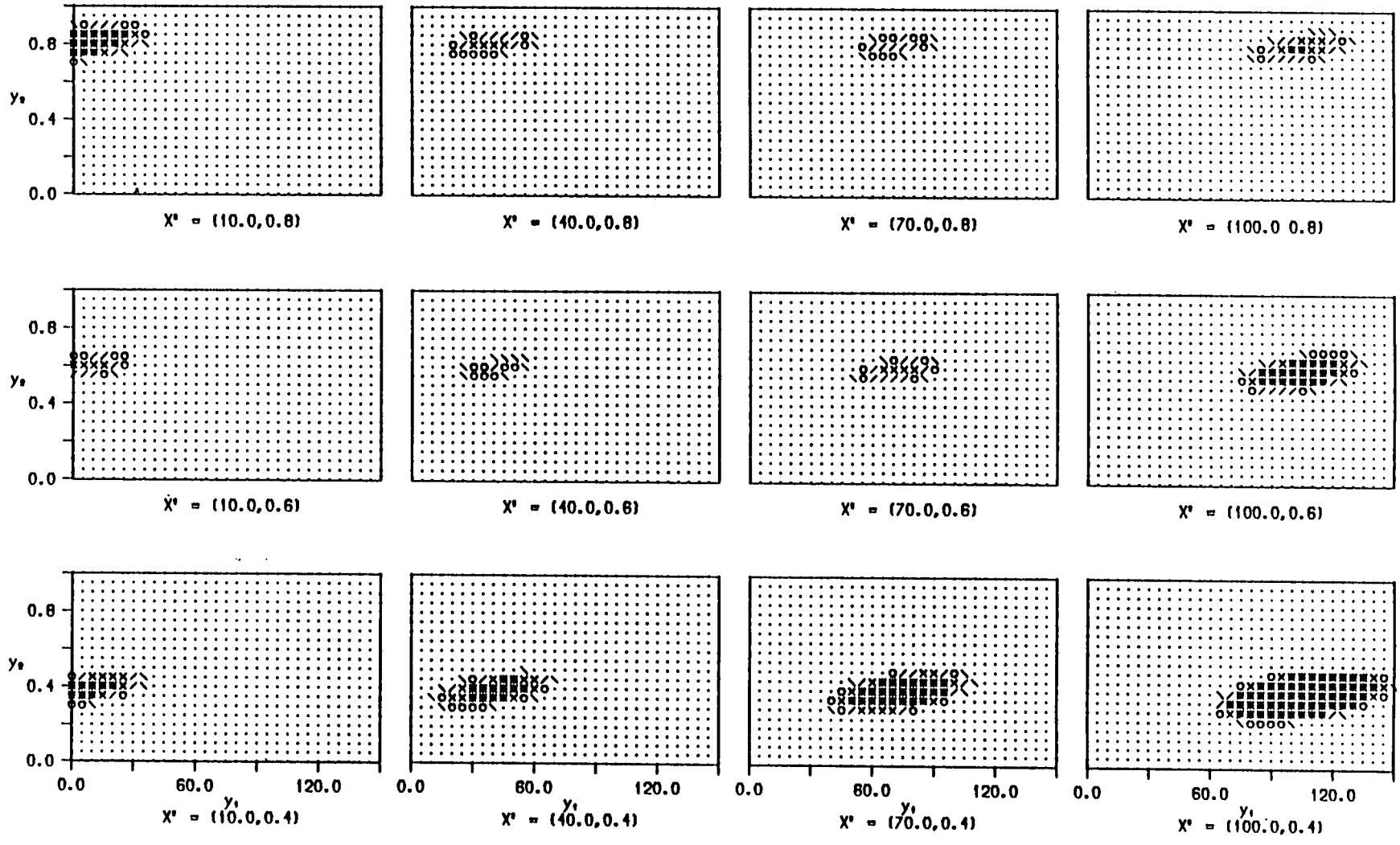


Fig. 6.20 Contour of Bayes' Factor as a function of $\underline{Y}$, given $\underline{X}' = (x_1, x_2)$ where x_1 and x_2 are 10.0, 40.0, 70.0 & 100.0 and 0.4, 0.6 & 0.8, respectively; for r=10, m=10 - Grouped model.

(· -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^{2} and ■ -- BF>10.0^{2})

Figs. 6.19 and 6.20 with Figs. 6.11 and 6.12 respectively, there is no significant difference between the two kernel methods except when the value of $\underset{\sim}{X}'$ is least common. When $\underset{\sim}{X}'$ = (40.0,0.6) and (70.0,0.8), the evidence implied by the control and recovered observations shown in the graphs is least convincing.

Generally, the area which shows slight or moderate evidence for supporting the hypothesis C does not appear to exist when r=10 and m=10. Finally, note that when r=10 and m=10 there are many values of the Bayes' factor of less than one. This suggests that it is difficult to find evidence to support the hypothesis C. In a forensic context, this implies that the probability of the material found at the crime scene and also on the suspect originating from the same source is small. But when such an event occurs, the weight of the evidence suggested by the control and recovered data is considerable. The weight of evidence depends on the rarity of the control and/or of the recovered data.

The values of $\underset{\sim}{Y}$ where the maximum of the Bayes factor occurred for the grouped model corresponding to the Figs. 6.7, 6.8, 6.11, 6.12, 6.15, 6.16, 6.19 and 6.20 are tabulated in Tables 6.13 and 6.14. Ideally one would prefer that the maximum occur where $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ coincide. Note that the values $\underset{\sim}{X}$ takes such as (10.0, 0.4), (10.0, 0.6), (10.0, 0.8), (40.0, 0.4), (70.0, 0.4) and (100.0, 0.4) are not in the training data, and hence are outside our experience. Thus, it is not surprising that the maximum does not occur where $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ are equal.

Table 6.13 The values of Y where the maximum of the Bayes' factor occurred, given some values of X. The value of the maximum is given in **bold**. Grouped model; r=1,m=1.

Method:	Ordinary		Adaptive	
Case	Non-diagonal	Diagonal	Non-diagonal	Diagonal
a	15.0 0.20	0.0 0.20	10.0 0.30	0.0 0.30
	3.4×10^{1}	**3.8×10^{1}**	**3.0×10^{1}**	**3.8×10^{1}**
b	0.0 0.60	0.0 0.60	0.0 0.60	0.0 0.55
	3.6	**3.6**	**3.6**	**3.8**
c	0.0 1.00	0.0 0.95	0.0 0.95	0.0 0.90
	3.1×10^{1}	**3.3×10^{1}**	**4.0×10^{1}**	**3.1×10^{1}**
d	120.0 0.05	50.0 0.15	80.0 0.30	40.0 0.30
	1.3×10^{2}	**9.4×10^{1}**	**1.1×10^{2}**	**5.5×10^{1}**
e	40.0 0.55	35.0 0.55	40.0 0.30	35.0 0.55
	1.9	**1.9**	**1.9**	**1.9**
f	5.0 0.95	25.0 0.90	10.0 0.90	25.0 0.85
	6.8	**5.1**	**6.3**	**4.7**
g	150.0 0.00	130.0 0.15	145.0 0.25	100.0 0.35
	2.7×10^{3}	**1.1×10^{3}**	**3.0×10^{3}**	**3.9×10^{2}**
h	125.0 0.45	95.0 0.50	110.0 0.50	90.0 0.55
	6.2	**4.6**	**5.3**	**4.1**
i	75.0 0.90	90.0 0.90	75.0 0.85	85.0 0.90
	2.7	**3.1**	**2.7**	**3.0**
j	150.0 0.00	150.0 0.10	150.0 0.20	150.0 0.40
	4.3×10^{4}	**7.0×10^{4}**	**2.3×10^{5}**	**1.0×10^{4}**
k	150.0 0.40	150.0 0.50	150.0 0.50	150.0 0.55
	7.2×10^{1}	**6.3×10^{1}**	**9.0×10^{1}**	**5.2×10^{1}**
l	150.0 0.85	150.0 0.90	140.0 0.80	145.0 0.90
	7.9	**9.6**	**7.9**	**9.5**

Notes: a. $X' = (10.0,0.4)$, b. $X' = (10.0,0.6)$, c. $X' = (10.0,0.8)$
d. $\tilde{X}' = (40.0,0.4)$, e. $\tilde{X}' = (40.0,0.6)$, f. $\tilde{X}' = (40.0,0.8)$
g. $\tilde{X}' = (70.0,0.4)$, h. $\tilde{X}' = (70.0,0.6)$, i. $\tilde{X}' = (70.0,0.8)$
j. $\tilde{X}' = (100.0,0.4)$, k. $\tilde{X}' = (100.0,0.6)$, l. $\tilde{X}' = (100.0,0.8)$

Table 6.14 The values of Y where the maximum of the Bayes' factor occurred, given some values of X. The value of the maximum is given in **bold**. Grouped model; r=10,m=10.

Method:	Ordinary		Adaptive	
Case	Non-diagonal	Diagonal	Non-diagonal	Diagonal
a	10.0 0.40	0.0 0.35	10.0 0.40	10.0 0.40
	3.9×10^{2}	**1.1×10^{3}**	**3.1×10^{2}**	**4.5×10^{2}**
b	5.0 0.60	5.0 0.60	5.0 0.60	5.0 0.60
	6.0×10^{1}	**8.4×10^{1}**	**6.1×10^{1}**	**6.9×10^{1}**
c	0.0 0.85	5.0 0.80	5.0 0.80	0.0 0.80
	4.0×10^{5}	**3.7×10^{3}**	**4.1×10^{4}**	**2.1×10^{3}**
d	45.0 0.35	45.0 0.40	45.0 0.40	40.0 0.40
	4.3×10^{3}	**2.1×10^{5}**	**6.7×10^{3}**	**6.4×10^{2}**
e	40.0 0.60	40.0 0.60	40.0 0.60	40.0 0.60
	9.6	**1.1×10^{1}**	**9.3**	**1.2×10^{1}**
f	35.0 0.80	40.0 0.80	35.0 0.80	40.0 0.80
	1.2×10^{2}	**5.3×10^{1}**	**1.1×10^{2}**	**5.8×10^{1}**
g	85.0 0.35	75.0 0.40	80.0 0.40	75.0 0.40
	1.7×10^{11}	**3.6×10^{6}**	**7.3×10^{6}**	**1.0×10^{4}**
h	75.0 0.60	75.0 0.60	75.0 0.60	75.0 0.60
	1.8×10^{2}	**5.4×10^{1}**	**1.8×10^{2}**	**6.4×10^{1}**
i	70.0 0.80	70.0 0.80	70.0 0.80	70.0 0.80
	1.8×10^{1}	**2.1×10^{1}**	**2.2×10^{1}**	**2.7×10^{1}**
j	125.0 0.35	105.0 0.35	115.0 0.40	105.0 0.40
	1.4×10^{20}	**8.2×10^{10}**	**4.2×10^{11}**	**1.6×10^{6}**
k	110.0 0.55	110.0 0.60	110.0 0.60	105.0 0.60
	3.5×10^{6}	**1.2×10^{4}**	**3.6×10^{5}**	**2.8×10^{3}**
l	105.0 0.80	105.0 0.80	105.0 0.80	105.0 0.80
	8.5×10^{1}	**1.4×10^{2}**	**8.3×10^{1}**	**1.1×10^{2}**

Notes: a. $X' = (10.0,0.4)$, b. $X' = (10.0,0.6)$, c. $X' = (10.0,0.8)$
d. $\tilde{X}' = (40.0,0.4)$, e. $\tilde{X}' = (40.0,0.6)$, f. $\tilde{X}' = (40.0,0.8)$
g. $\tilde{X}' = (70.0,0.4)$, h. $\tilde{X}' = (70.0,0.6)$, i. $\tilde{X}' = (70.0,0.8)$
j. $\tilde{X}' = (100.0,0.4)$, k. $\tilde{X}' = (100.0,0.6)$, l. $\tilde{X}' = (100.0,0.8)$

6.5.3 Transformation of the variables

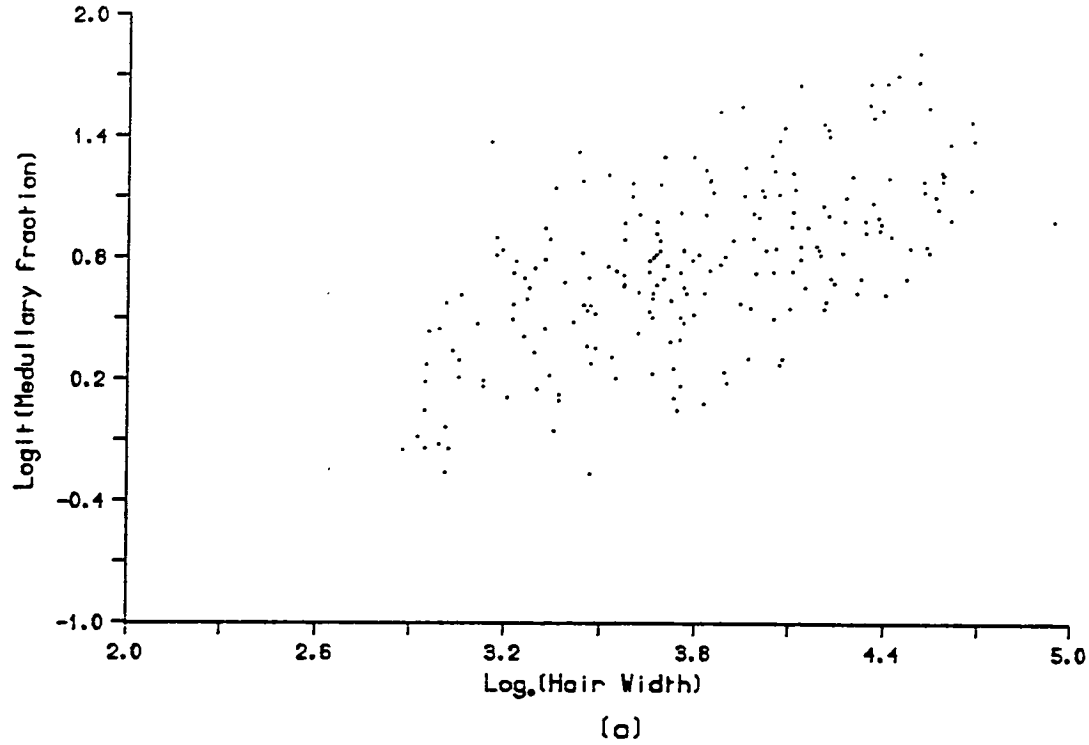
As in Chapter 3, the hair width can only take positive values and the medullary fraction can only take values between 0 and 1. These values conflict with the parameter range of the unknown mean vector $\underset{\sim}{\mu}$ and the support of the kernel density estimate which are

both the real line. Transforming the variables may be more appropriate so that the transformed variables can take values over the whole real line. Natural logarithms of hair width and a $\log_e\{MF/(1-MF)\}$ transformation of medullary fraction are taken. Fig. 6.21 shows the bivariate dot plot of the transformed data for (a) 220 cat hairs and (b) 22 group means. The effect of transforming the data has reduced the skewness of the variables, especially the hair width. Observations which are far distant from the main group of observations before transformation are pulled towards the centre of the group.

The Bayes' factor is calculated under the assumption that the training data are grouped or ungrouped with a diagonal covariance matrix. The ordinary kernel method is used. Figs. 6.22 and 6.23 show the Bayes' factor as a function of $\underset{\sim}{Y}$ given some values of $\underset{\sim}{X}$ under the grouped model for r=1, m=1 and r=10, m=10, respectively. Figs. 6.24 and 6.25 show the Bayes' factor as a function of $\underset{\sim}{Y}$ given some values of $\underset{\sim}{X}$ under the ungrouped model for r=1, m=1 and r=10, m=10, respectively. The range of the axes for these graphs are approximately the same as those shown in Figs 6.7 - 6.20 except for the left hand side of the x-axis. The values of vector $\underset{\sim}{X}$ are the transformed values corresponding to the values of $\underset{\sim}{X}$ shown in Figs 6.7 - 6.20. There is no obvious distinction between the grouped and ungrouped models. This is due to the effect of the transformation as discussed above.

6.6 Conclusions and Discussion

From the results found in the previous section it is advisable



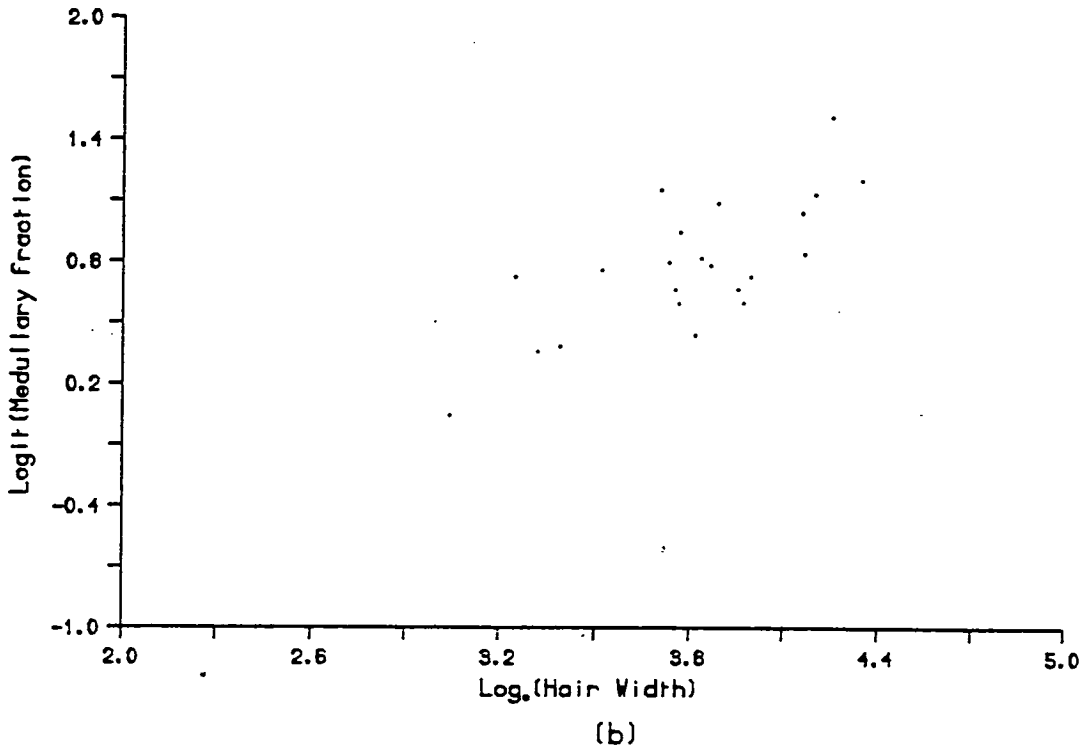


Fig. 6.21 A bivariate dot plot of the transformed data for (a) 220 cat hairs and (b) 22 group means

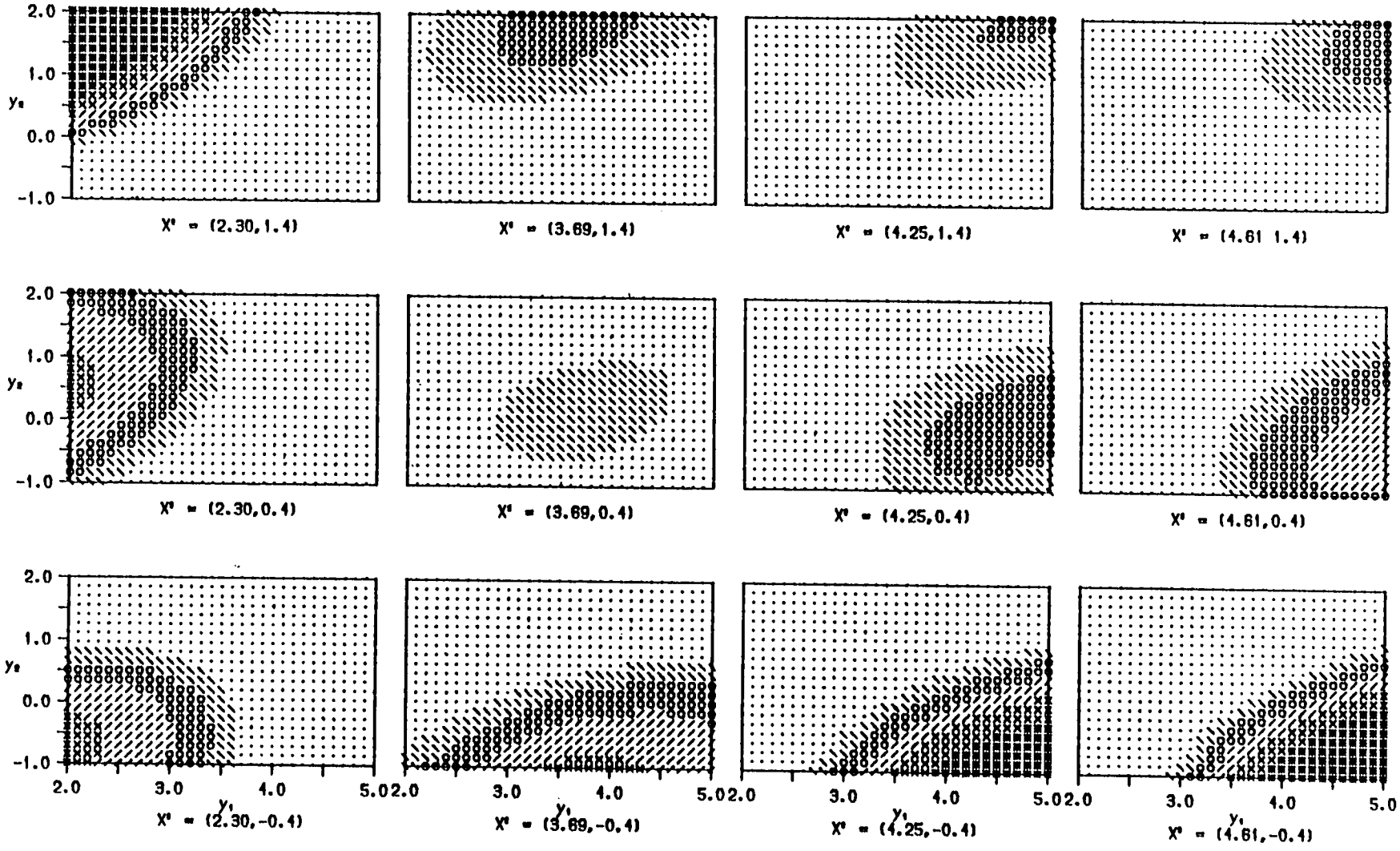


Fig. 6.22 Contour of Bayes' factor as a function of $\underline{Y}$, given $\underline{X}^t = (x_1, x_2)$ where x_1 and x_2 are 2.30, 3.69, 4.25 & 4.61 and -0.4, 0.4 & 1.4, respectively; for r=1, m=1 - Grouped model. Transformed data.
(· -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and ■ -- BF>10.0^2)

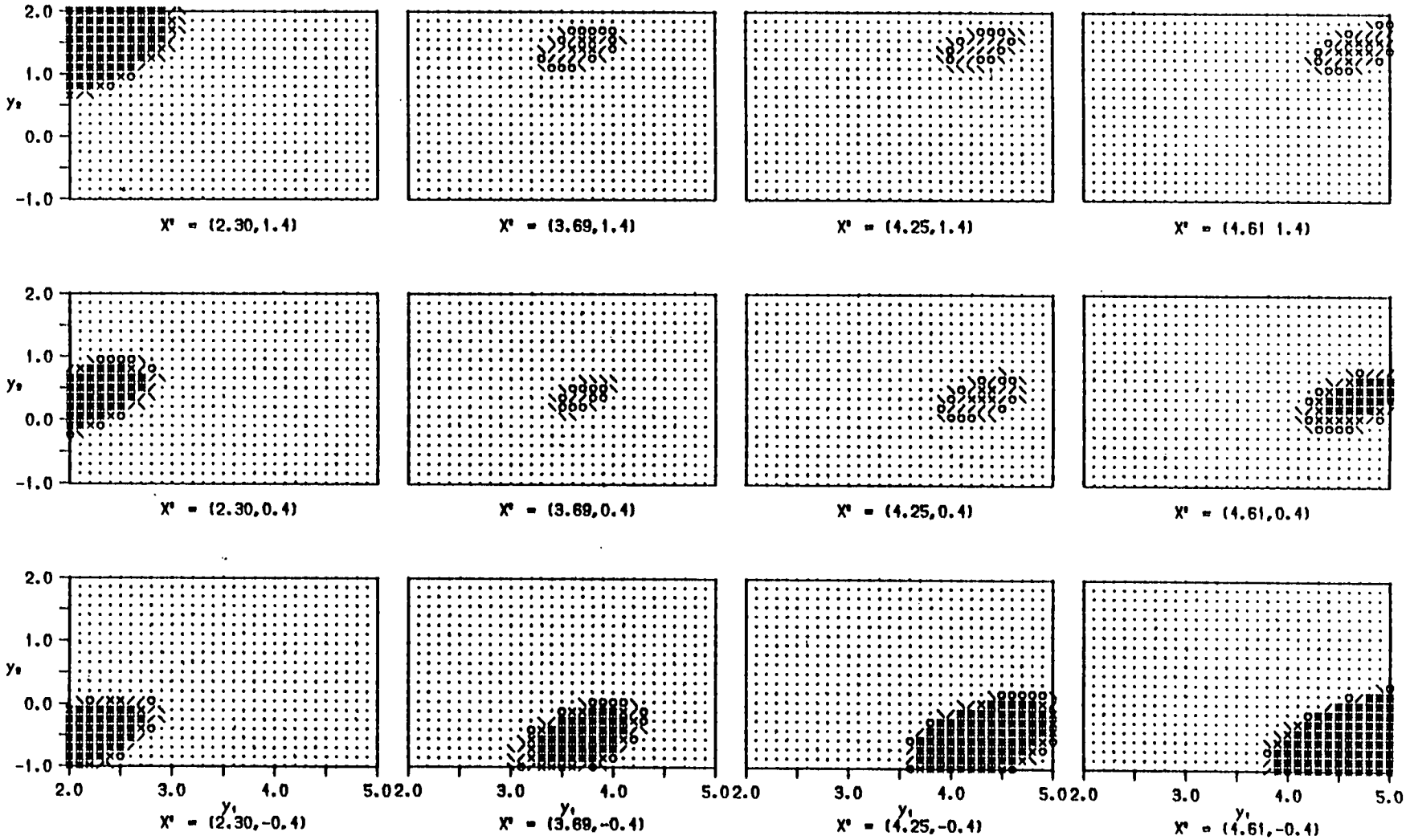


Fig. 6.23 Contour of Bayes' factor as a function of $\underline{Y}$, given $\underline{X}^0 = (x_1, x_2)$ where x_1 and x_2 are 2.30, 3.69, 4.25 & 4.61 and -0.4, 0.4 & 1.4, respectively; for r=10, m=10 - Grouped model. Transformed data.
(. -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and ■ -- BF>10.0^2)

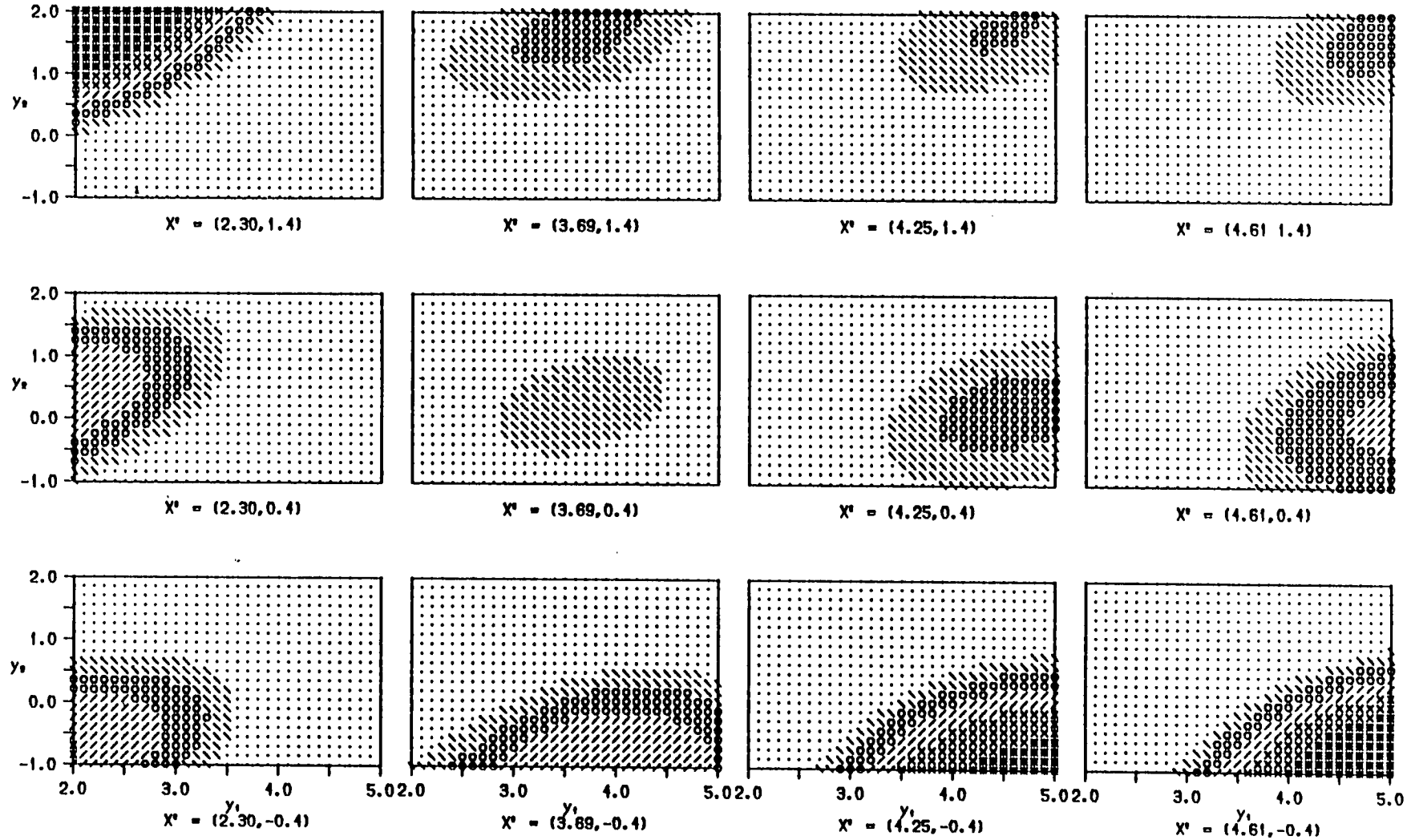


Fig. 6.24 Contour of Bayes' factor as a function of $\underline{Y}$, given $\underline{X}^0 = (x_1, x_2)$ where x_1 and x_2 are 2.30, 3.69, 4.25 & 4.61 and -0.4, 0.4 & 1.4, resp. for r=1, m=1 - Nogrouped model. Transformed data.
(· -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and ■ -- BF>10.0^2)

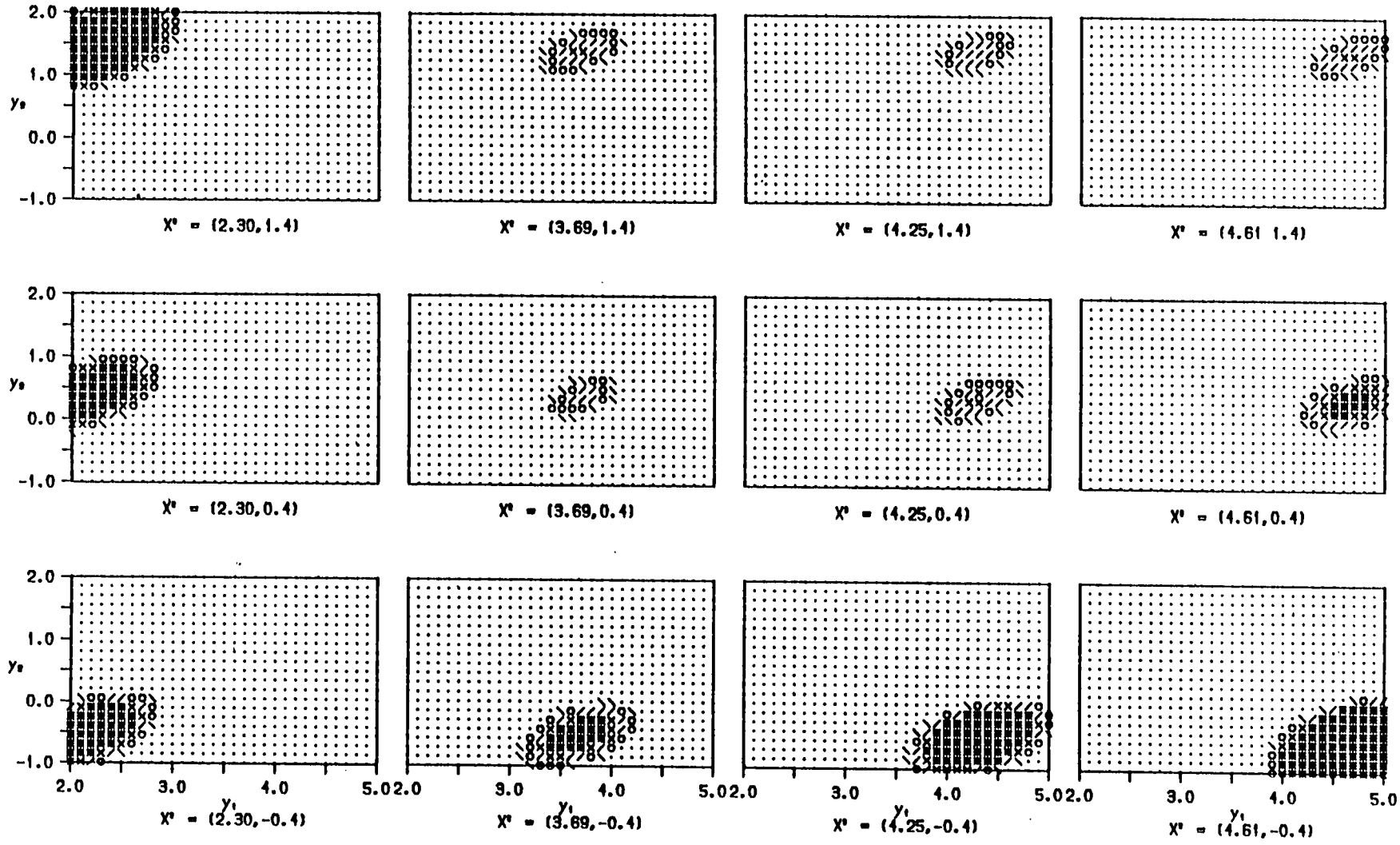


Fig. 6.25 Contour of Bayes' Factor as a function of $\underline{Y}$, given $\underline{X}^0 = (x_1, x_2)$ where x_1 and x_2 are 2.30, 3.69, 4.25 & 4.61 and -0.4, 0.4 & 1.4, resp.; For r=10, m=10 - Nogrouped model. Transformed data.
(· -- BF<1.0, \ -- 1.0<BF<$10^{1/2}$, o -- $10^{1/2}$<BF<10.0, / -- 10.0<BF<$10.0^{3/2}$, x -- $10.0^{3/2}$<BF<10.0^2 and ■ -- BF>10.0^2)

to use the adaptive kernel method. The use of a robust covariance matrix is also favourable. Other robust forms of covariance matrix can be used to replace S or D, such as instead of the use of (n-1) degrees of freedom in the equating S in (6.2) one would use (n-p-1) degrees of freedom. Like Chapter 3, it is required to validate the assumption of homogenous within-group variance among the groups. Alternatively, one could modify the model to accommodate such a possibility.

CHAPTER 7

STUDENT-t KERNEL

7.1 Introduction

It is generally accepted that the choice of kernel function in a density estimation problem is less crucial than the value given to scale parameter λ. The latter is usually better known as the smoothing parameter or window width. There is a considerable literature devoted to estimating the smoothing parameter λ. Many studies involved the use of a Gaussian kernel to compare different methods of estimating the smoothing parameter. In this chapter, a Bayesian method is employed to estimate the smoothing parameter by introducing a prior distribution for λ. As the result a new kernel is formed. Methods of estimating the parameters involved are discussed.

7.2 Student-t kernel

A general background of the kernel function has already been outlined in Chapter 1. In brief, given a data set $D = \{u_1, u_2, \ldots, u_N\}$ a Gaussian kernel is used for mathematical convenience, then the kernel density estimate for the data $u_1, u_2, \ldots, u_N$, is given by

$$\hat{f}(u|u_i,\lambda) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(2\pi\lambda^2)^{\frac{1}{2}}} \exp\left\{-\frac{(u-u_i)^2}{2\lambda^2}\right\}. \qquad (7.1)$$

Without ambiguity, the dependence on sample size N of the density estimate (7.1) is dropped. Note that, unlike the previous chapters, this kernel density estimate of (7.1) has not been standardised.

This enables us to adapt some results given by Silverman (1986) which will be used later in this Chapter.

As remarked by Cover (1972), the method of Loftsgaarden and Quesenberry (1965), later better known as the nearest neighbour method, allows the smoothing parameter, λ, to depend on the data. Loftsgaarden and Quesenberry 's method, required λ to be the distance to the κ_nth nearest point to u among the samples u_1, u_2,..., u_N. Thus λ is a random variable depending on the data. In addition it is obvious that different samples drawn from the unknown density f(u), will yield different values of λ. From a Bayesian viewpoint, one could assign a prior density to λ. So if we let $\tau = \lambda^{-2}$, then an intuitive choice of prior for τ is

$$g(\tau|\alpha,\beta) = \frac{(\beta/2)^{\alpha/2}\ \tau^{(\alpha/2)-1}}{\Gamma(\alpha/2)} \exp\left\{-\frac{\beta\tau}{2}\right\}, \qquad (7.2)$$

where α and β are usually unknown. The choice of prior for λ of the form (7.2) is a natural choice from a Bayesian point of view since, from Chapter 2, λ is the variance of a kernel located at a value u_i and here the kernel is chosen to be Gaussian. Of course, other choices of prior for λ can be used, but the form of (7.2) leads to simple form of a new kernel density estimate.

Hence, the kernel density of u given the data D only, can be obtained by combining (7.1) and (7.2), and integrating over τ, namely

$$\hat{f}_g(u|D) = \int f(u|\tau,D)\ g(\tau)\ d\tau$$

$$= \frac{1}{Be(\alpha/2,1/2)\ N\ \sqrt{\beta}} \sum_{i=1}^{N} \frac{1}{[1+\beta^{-1}(u-u_i)^2]^{(\alpha+1)/2}}. \qquad (7.3)$$

Since, in general, the Bayes solution depends on $g(\tau)$, the Bayes kernel density estimate is denoted by $\hat{f}_g(u)$ to indicate this dependence. Essentially, we have replaced one parameter, namely λ, by two parameters which are α and β. The form of (7.3) involves the choice of α and β. They are so-called hyperparamters in a Bayesian context. Now the density estimate of (7.3) may be written as

$$\hat{f}_g(u) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\beta^{1/2}} K \left[\frac{(u-u_i)}{\sqrt{\beta}} \right]$$

where

$$K(t) = \frac{1}{Be(\alpha/2,1/2)\ (1+t^2)^{(\alpha+1)/2}} . \qquad (7.4)$$

Then $K(t)$ is a scaled Student-t distribution so let us denote it by K_t and call it, a Student-t kernel. The kernel K_t has a longer tail than the Gaussian kernel and the peakedness is determined by α. A small value of α gives a flat kernel and a large value of α gives a spiky kernel. So it is in a sense we can choose the choice of peakedness of the kernel to be placed over each observation initially. Then further adjustment may made when β is determined

As an illustration, there are measurements on 185 dog hairs. This data set has already been described in detail in Chapters 1 and 5, where there are 200 dogs hairs instead of 185. Fifteen zeros were excluded because they caused problems when the parameters were estimated later. The measurements are in the form of the ratio of the width of a central core, known as the medullary width and the width of the hair. Figs. 7.1, 7.2 and 7.3 show the estimated

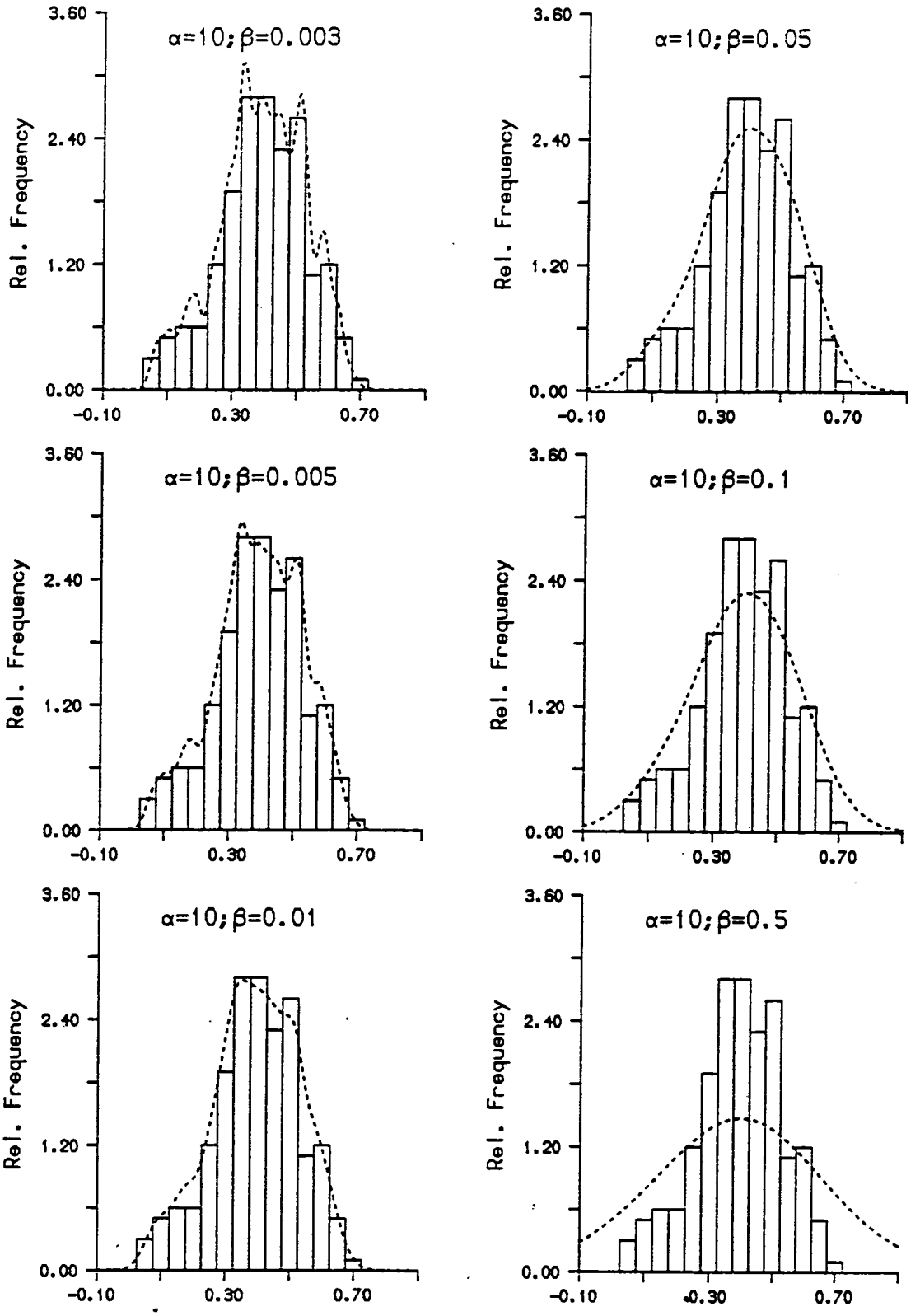


Fig. 7.1 Histogram and density plot of the dog data (ex. zeros); (----) fitted by Student-t kernel with informative prior for the smoothing parameter λ.

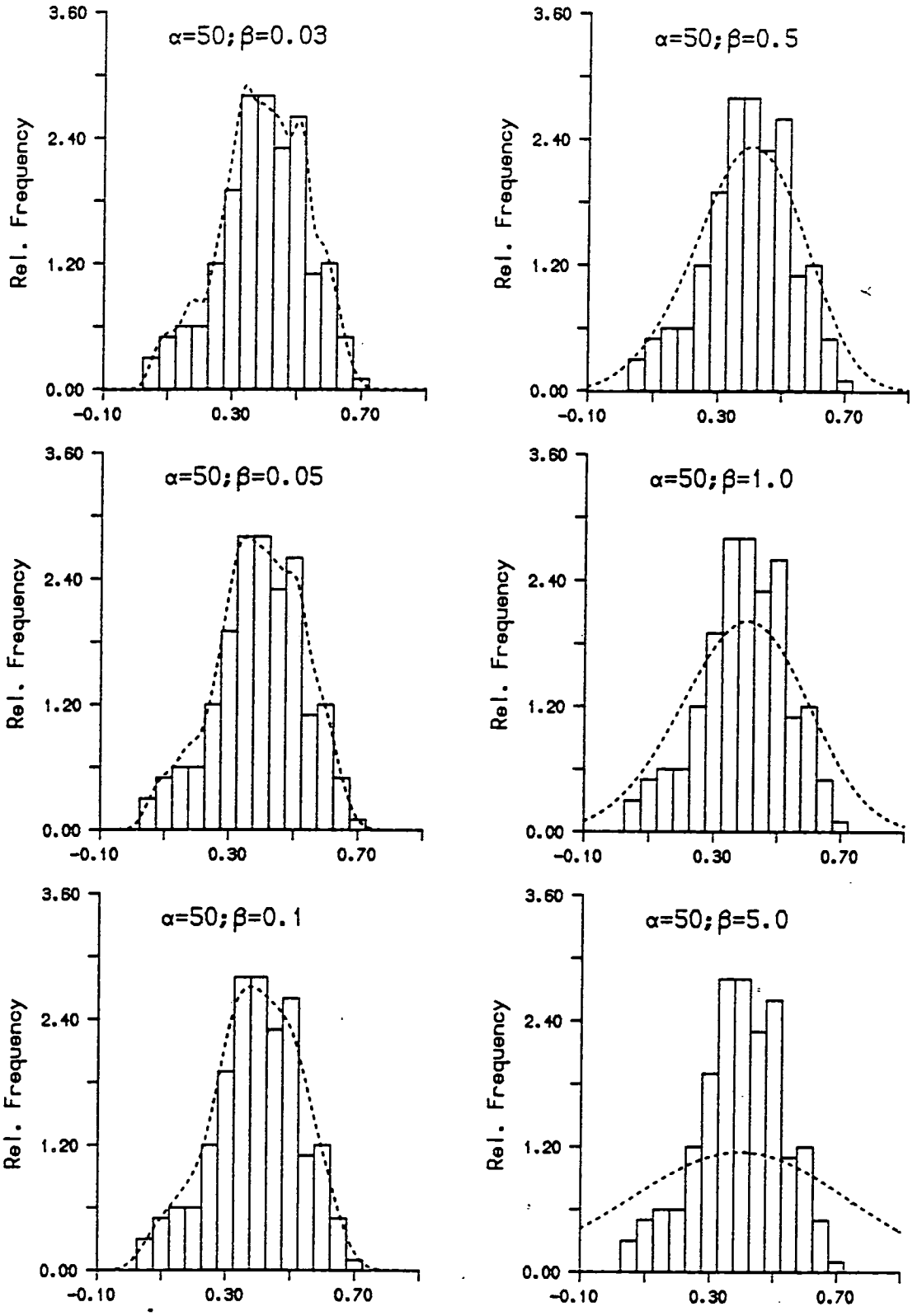


Fig. 7.2 Histogram and density plot of the dog data (ex. zeros); (----) fitted by Student-t kernel with informative prior for the smoothing parameter λ.

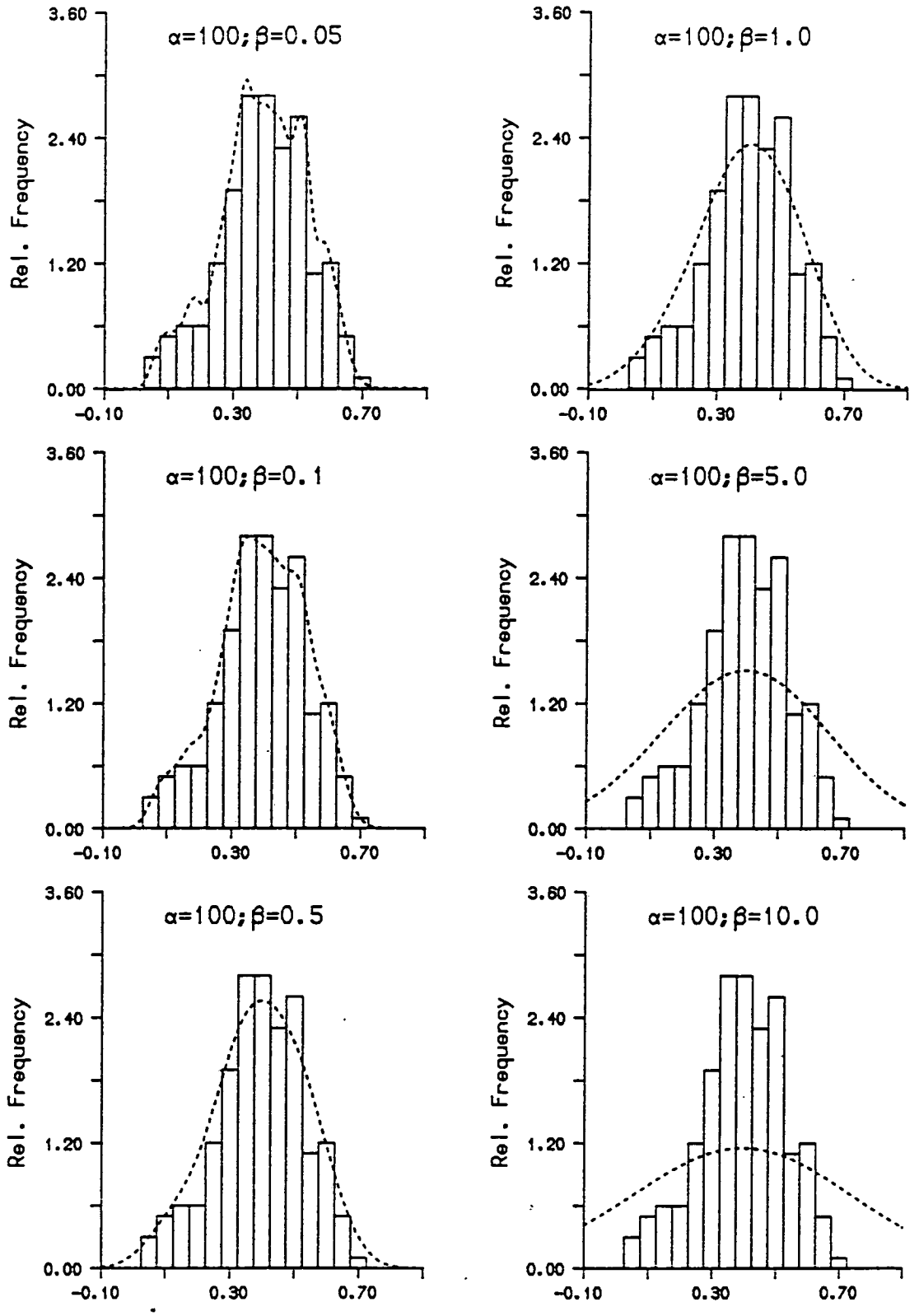


Fig. 7.3 Histogram and density plot of the dog data (ex. zeros); (----) fitted by Student-t kernel with informative prior for the smoothing parameter λ.

density of the dog data (excluding the zeros where no central core was present) with α equal to 10, 50 and 100 respectively, and with different values of β. It is quite clear from the plots that the degree of roughness of the density estimate increase as β decreases when α is fixed. Similarly, when β is fixed the degree of smoothing decrease as α increases. Also, notice that the density estimate with α=50 & β=0.5 and α=100 & β=1.0 yield a similar fit to the data. This could be explained by noting the fact that from (7.2) the expected value of τ, the precision of λ^2, is β/α. Thus in the above cases, they both yield the same value of β/α. For the present if one ignores the Bayesian context, one could obtain a density estimate of the data u_i's by putting in appropriate values of α and β. Since (7.3) is a p.d.f in its own right, one could use an empirical Bayes procedure to estimate the values of α and β. In fact (7.3) can be seen as a predictive distribution of x, given the 'past' data D. Later we shall discuss these objective ways of choosing α and β.

7.3 Efficiency of the Student-t kernel

In this section the efficiency of the 'new' kernel function obtained in the previous section is compared with the other kernels. Silverman (1986) gives a table of some existing kernels and their efficiencies relative to the Epanechnikov kernel (K_e) defined as

$$K_e(t) = \begin{cases} [3(1-(t^2/5))]/(4\sqrt{5}) & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & \text{otherwise.} \end{cases}$$

Define the efficiency of K to be

$$\text{eff}(K) = \{C(K_e)/C(K)\}^{5/4}$$

$$= 3\,[5\sqrt{5}]^{-1}\,\{\int t^2K(t)dt\}^{-1/2}\,\{\int K(t)^2dt\}^{-1}. \qquad (7.5)$$

In (7.5) C(K) is a constant and, for any kernel K, is given by

$$C(K) = k_2^{2/5}\,\{\int K(t)^2\,dt\}^{4/5},$$

where k_2 is defined as $\int t^2K(t)dt$ and is not equal 0. For $K = K_t$, the first integral in (7.5) is just the variance of a scaled Student-t distribution defined in (7.4) and is equal to α-1. The second integral can be simplified as

$$\int K_t(t)^2dt = \frac{Be(1/2,(2\alpha+1)/2)}{[Be(\alpha/2,1/2)]^2}.$$

Thus the efficiency of the Student-t kernel, K_t, relative to the Epanechnikov kernel is

$$eff(K_t) = \frac{3\sqrt{\pi}\sqrt{(\alpha-2)}\Gamma(\alpha+1)\Gamma(\alpha/2)^2}{5\sqrt{5}\;\Gamma[(\alpha+1)/2]^2\Gamma[(2\alpha+1)/2]}.$$

The efficiency of the kernel K_t is dependent on the value α but independent of β. The value of α represents how much we know about the smoothing parameter λ. One will expect the $eff(K_t)$ will increase as α increases. $Eff(K_t)$ is plotted in Fig. 7.4 as a function α. The scale along the x-axis is in logarithm to base 10. The value of $eff(K_t)$ increases sharply when α jumps from 3 to 20 and then it levels off at around 0.9512, which suggested that the limit of $eff(K_t)$ as α tends to infinity, is around 0.9512. Incidentally the Gaussian kernel also has an efficiency value of 0.9512. This is not surprising at all since a Student-t with degree of freedom ν, say, tends to Normality as $\nu \to \infty$. The value of α, when $eff(K_t)$ is approximate 0.9512, is 10000. It gives an indication that, for the

Student-t kernel to be at least as efficient as the Gaussian kernel, we do require a lot of information about the smoothing parameter prior to an experiment. This seems reasonable since the kernel density estimate of the form (7.3) does not involve the estimation of the smoothing parameter λ at all.

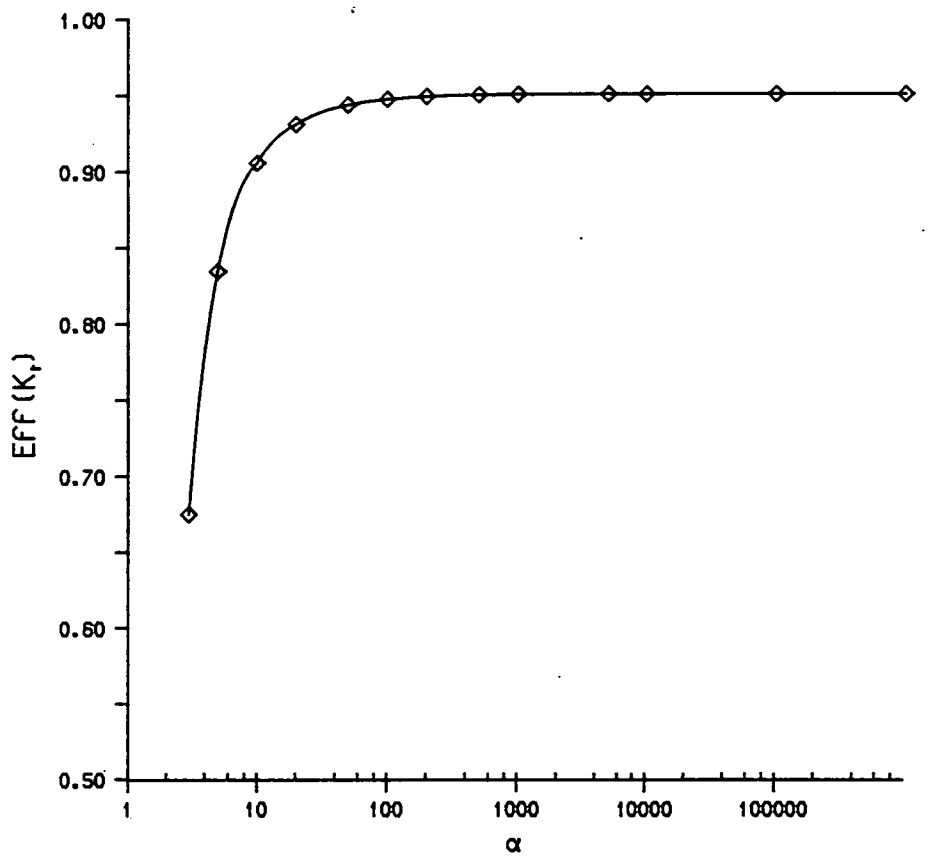


Fig. 7.4 Efficiency of the Student-t kernel, K_t relative to Epanechnikov kernel K_e.

7.4 Estimation of the hyperparameters

The problem we now face is the determination of appropriate values for α and β. The estimation procedure is based on an

empirical Bayes approach (See Maritz (1970)), since the hyperparameters are chosen objectively by the data under consideration. Two methods are considered: the method of moments and a maximum likelihood method.

7.4.1 The method of moments

The idea of the method of moments is simple and is based on the assumption of the prior distribution of λ. That is, if the prior distribution of τ is of the form shown in (7.2), then after reparameterisation, λ^2 has a prior density proportion to

$$[\ 1/\lambda^2\ ]^{(\alpha/2)+1}\ \exp\ \{-\beta/2\lambda^2\}.$$

The expectation and variance of λ^2 are easily found by remarking that since β/λ^2 is x^2 with α degrees of freedom. And if $x = \beta/2\lambda^2$ then X is $\Gamma(\alpha/2,1)$ with density $e^{-x}x^{(\alpha/2)-1}/[(\alpha/2)-1]!$ and hence

$$E(\lambda^2) = \beta/(\alpha-2) \tag{7.6}$$

and

$$\mathrm{Var}(\lambda^2) = 2\beta^2/(\alpha-2)^2(\alpha-4), \tag{7.7}$$

where $E(.)$ is the expectation operator. Therefore, if one knows the expected value and variance of λ^2, then one could solve the above two equations to yield an objective choice for α and β. Now suppose $E(\lambda^2)$ and $\mathrm{Var}(\lambda^2)$ are known and are taken to be a and b, respectively. From (7.6), it is easily seen that

$$\beta = a(\alpha-2). \tag{7.8}$$

Thus, α can be obtained by substituting β of (7.8) into (7.7) namely,

$$\alpha = 2a^2/b + 4.$$

In practice, a and b are not known and have to be estimated. Since we do not have any previous data as such to estimate the moments of this problem, one could obtain such a sample by a bootstrap technique. A sample is generated by successively selecting uniformly with replacement from the data set $D = \{u_1, u_2, \ldots, u_N\}$ to construct a bootstrap sample, say $\{u_1^*, u_2^*, \ldots, u_N^*\}$. For each bootstrap sample, the estimate of λ^2 is calculated. There are a number of ways of which λ^2 can be estimated. One of the quick objective ways of obtaining a rough estimate for λ^2 involves the minimum MISE criterion.

Recall that the Gaussian kernel involves only one parameter, namely λ. An optimal value of λ, λ_{opt} say, can be obtained using the minimum Mean Integrated Squared Error (MISE) criterion. Silverman (1986) gives an ideal value of λ, based on minimising the approximate mean integrated squared error under suitable conditions, namely that

$$\tfrac{1}{4}\lambda^4 k_2^2 \int f''(u)^2 du + N^{-1}\lambda^{-1}\int K(t)^2 dt, \qquad (7.9)$$

is, by simple calculus, to be equal to λ_{opt}, where λ_{opt} is given as

$$k_2^{-2/5} \{\int K(t)^2 dt\}^{1/5} \{\int f''(u)^2 du\}^{-1/5} N^{-1/5}. \qquad (7.10)$$

In both (7.9) and (7.10) k_2 is defined as after (7.5) and f(x) denotes the underlying true density function. However, we are interested in getting a rough estimate for λ^2. So by differentiating (7.9) with respect to λ^2, λ_{opt}^2 is shown to be the square of the expression in (7.10). Unfortunately, λ_{opt}^2 being the square of λ_{opt} depends on the unknown density being estimated. With a Gaussian

kernel, Silverman (1986) provides a simple expression for (7.10) assuming the true underlying density is Normal with variance σ^2, namely,

$$\{1.06\sigma N^{-1/5}\}. \qquad (7.11)$$

Because of the simple form of (7.11), it is used as a pilot estimate for λ^2. When the square of (7.11) is computed to give $\hat{\lambda}^2_{opt}$, σ is replaced by the sample standard deviation of the data concerned. Hence we can now obtain estimates for $E(\lambda^2)$ and $Var(\lambda^2)$ using the following iteration procedure:

Step 1 : Randomly select a sample of size N_{sam} with replacement, from the data $\{u_1,\ldots u_N\}$. Then evaluate $\hat{\lambda}^2_{opt}$ of (8.11) with σ replaced by the standard deviation of the selected sample.

Step 2 : Repeat Step 1 several times, N_{rep} say.

Step 3 : Calculate sample mean and sample variance of the N_{rep} $\hat{\lambda}^2_{opt}$'s. These provide estimates for $E(\lambda^2)$ and $Var(\lambda^2)$. Then solve for α and β as described above.

Note that later on in a simulation study, N_{rep} is taken to be 100.

However, satisfactory results for $\hat{\lambda}^2_{opt}$ will only be expected if the data did come from a Normal distribution, since $\hat{\lambda}^2_{opt}$ is obtained under the assumption that the true density is Normal. Nevertheless other expressions for $\hat{\lambda}^2_{opt}$ may be obtained to safeguard this, for

example, σ may be replaced by a robust measure of spread. For a wide range of distributions, Silverman (1986) suggests the following robust estimate A for σ, given by

$$A = \min \{\text{standard deviation, interquartile range}/1.34\}.$$

Let us call the pilot estimate for λ^2 based on $\hat{A}$, a robust estimate and the former, i.e., the estimate for λ^2 based on s, (s is the sample standard deviation of the n_{sam} observations), a Normal estimate since it is based on the Normality assumption about the true density. Note that these two estimates are obtained from a Gaussian kernel. An optimal λ^2 using a Student-t kernel can be obtained but the expression is complicated because it depends on the parameter α since $\int K_t(t)^2 dt$ of (7.10) is

$$\frac{\Gamma(\alpha_1/2)^2\ \Gamma(\alpha+1/2)}{\Gamma(\alpha/2)\ \Gamma(1/2)\ \Gamma(\alpha_1)}$$

where $\alpha_1 = \alpha+1$. Note also that in order to satisfy certain conditions given by Silverman (1986), the minimum value of α is 3.

7.4.2 Modified Maximum Likelihood method

The Student-t kernel density estimate for the data $u_1,\ldots,u_N$ is given in (7.3). The likelihood function of the parameters, given the data u_i's is

$$L(\alpha,\beta;X) = \prod_{\ell=1}^{N} \hat{f}_g(u_\ell|\alpha,\beta,u_i)$$

$$= \prod_{\ell=1}^{N} \frac{1}{Be(\alpha/2,1/2)\sqrt{\beta}\ N} \sum_{i=1}^{N} \frac{1}{\{1+(u_\ell-u_i)^2/\beta\}^{(\alpha+1)/2}}. \qquad (7.12)$$

As in Chapter 2, to get rid of the zeros when $\ell=i$, we use a leave-one-out modification of the M.L. method. The density estimator at the point u_ℓ will be based on all the sample except element u_ℓ. Thus, the maximum likelihood estimate for α and β is obtained by maximising (7.12) with N replaced by (N-1) to give $\hat{\alpha}$ and $\hat{\beta}$.

However, we know from (7.6) that α and β are related if we know $E(\lambda^2)$. Thus, maximising the likelihood function with respect to the parameters will not yield a unique solution for α and β. One way round this is to predetermine the value of one parameter, then maximise the likelihood function with respect to the other since the parameter α is interpreted as the degree of knowledge about the smoothing parameter λ. So predetermining α seems reasonable. Any value for α greater than 2 may be chosen. A value of α which represents vague information in λ is one which tends to zero. If we let α equal zero, then certain conditions cannot be satisfied. Perhaps a more automatic way to do this, is to use a two stage procedure namely a combination of the method of moments and the M.L. method and is simply as follows:

Stage 1 : Carry out the method of moments as described in the previous section. Thus obtaining a value for α, say $\hat{\alpha}_m$.

Stage 2 : Maximise the likelihood function with respect to β subject to $\alpha = \hat{\alpha}_m$.

As an example, kernel density estimates for the dog data are shown in Fig. 7.5 using the two estimation methods described above.

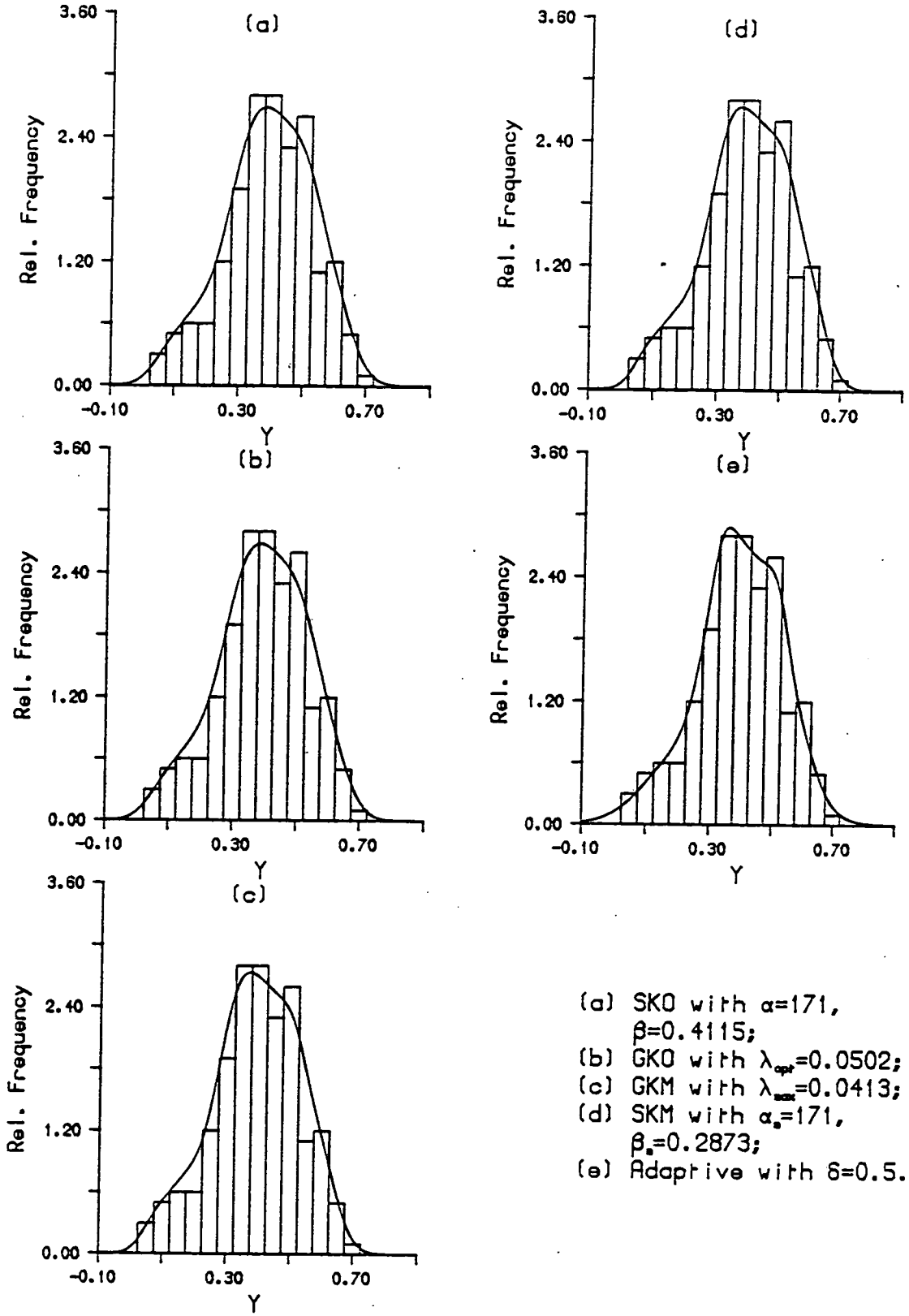


Fig. 7.5 Kernel density estimates of the dog data (excluding zeros) using SKO, GKO, SKM, GKM and adaptive kernel methods. (see Section 7.4 for details)

The Student-t kernel is being compared with the Gaussian kernel. The abbreviations GKO, SKO, GKM and SKM denote, respectively, the Gaussian kernel with the Normal optimum estimation procedure (7.11) for λ, the Student-t kernel with the method of moments estimation procedure described in Section 7.4.1, the Gaussian kernel with the ML estimation procedure mentioned in Chapter 1 and the Student-t kernel with the modified ML estimation procedure described in Section 7.4.2. Note that these abbreviations will be used throughout the rest of this chapter. As mentioned earlier the fifteen zeros from the dog data are removed because ML leave-one-out method will not provide sensible result. From Fig. 7.5 there is little to choose between the Student-t and Gaussian kernels and the different estimation procedures. This is because the data are pretty symmetric. The adaptive kernel method is also employed to fit the data, it seems to be over cautious in the term of having heavy tails and a higher peak.

7.5 A Simulation Study

The small sample performances of the Student-t kernel can be studied by simulating data from a few well-chosen distributions. In a different context, Bowman (1984) compares two procedures of estimating the smoothing parameter and selected four distributions namely, standard Normal, a bimodal mixture of two Normals, a Student-t with 5 degrees of freedom and the standard Cauchy, in a simulation study. The first two are chosen to test for sensitivity to changes of shape in the main body of the distribution. The latter two distributions serve to indicate the relative performance with respect to long-tailed distribution. In order to examine how the Student-t kernel relates to skewness I also include the standard

log-normal distribution.

7.5.1 Three measures of discrepancy

A criterion is required to assess the performance of the Student-t kernel when it is compared with the Gaussian kernel. Various measures have been employed to study the discrepancy of the density estimator $\hat{f}$ from the true density f and were briefly described in Section 2.1. A measure over a range of u values is more appropriate as it gives a general picture of how the two kernels and the methods of estimation procedure compare with each others. So the following three measures are chosen, Mean Integrated Square Error (MISE), Expected Mean Integrated Square Error (EMSE) and Integrated Square Error (ISE). Details of these can be found in Section 2.1.4. Evaluation of the measure ISE is quite straight forward. Whereas, computation of MISE and EMSE is slightly more complicated. These measures are evaluated for the Student-t kernel and the Gaussian kernel in comparison with each of the five distributions listed above. All the integrations involved in evaluating the three measures are done numerically.

7.5.2 Results of the simulation study

Initially, one sample of size 100 (=N) was generated from each of the distributions mentioned above. Four methods of estimation are compared. Tables 7.1a, 7.1b and 7.1c give the MISE, EMSE and ISE, respectively, of the density estimates of the simulated data using the methods GKO, GKM, SKO and SKM with the Normal criteria . Tables 7.2a, 7.2b and 7.2c give the results of the corresponding three measures of the density estimates with optimal smoothing parameter

Table 7.1a MISE ($\times 10^{-2}$) of the density estimates from simulated data using standard deviation as an estimate for σ in (7.11). N=100, $N_{sam}=N\times k$.

Method	k	Normal	Student-t with 5 d.f.	Lognormal	Cauchy	Bimodal
GKO	-	0.5413	0.5636	15.6346	13.6761	0.5748
SKO	1	0.5440	0.5650	14.6579	13.3248	0.5751
	2	0.5430	0.5640	15.0378	13.5220	0.5739
	5	0.5420	0.5628	15.4131	13.6280	0.5738
	10	0.5417	0.5618	15.5282	13.6505	0.5729
	50	0.5415	0.5623	15.5830	13.6548	0.5719
GKM	-	0.5708	0.6789	20.0251	14.3922	0.4967
SKM	1	0.5721	0.6769	6.5583	5.2073	0.4970
	2	0.5714	0.6780	8.0758	6.6125	0.4968
	5	0.5710	0.6785	11.2667	10.5677	0.4968
	10	0.5709	0.6787	14.9542	12.5472	0.4967
	50	0.5708	0.6789	18.9658	13.9685	0.4967

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - " " " the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum criterion procedure
GKM - " " " the ML estimation procedure

Table 7.1b EMSE ($\times 10^{-2}$) of the density estimates from simulated data using standard deviation as an estimate for σ in (7.11). N=100, $N_{sam}=N\times k$.

Method	k	Normal	Student-t with 5 d.f.	Lognormal	Cauchy	Bimodal
GKO	-	0.1403	0.1382	8.7676	3.3146	0.0839
SKO	1	0.1411	0.1386	8.2690	3.2281	0.0839
	2	0.1408	0.1383	8.4656	3.2772	0.0837
	5	0.1405	0.1379	8.6578	3.3037	0.0837
	10	0.1404	0.1375	8.7149	3.3088	0.0836
	50	0.1404	0.1377	8.7409	3.3090	0.0835
GKM	-	0.1492	0.1759	11.0535	3.5004	0.0738
SKM	1	0.1496	0.1752	3.6172	1.2652	0.0738
	2	0.1494	0.1756	4.5301	1.5915	0.0738
	5	0.1493	0.1758	6.3772	2.5178	0.0738
	10	0.1492	0.1758	8.4071	3.0056	0.0738
	50	0.1492	0.1759	10.5127	3.3940	0.0738

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - " " " the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum criterion procedure
GKM - " " " the ML estimation procedure

Table 7.1c ISE ($\times 10^{-2}$) of the density estimates from simulated data using standard deviation as an estimate for σ in (7.11). N=100, $N_{sam}=N\times k$.

Method	k	Normal	Student-t with 5 d.f.	Lognormal	Cauchy	Bimodal
GKO	-	0.7338	0.3761	16.8529	13.7225	0.4174
SKO	1	0.7367	0.3768	16.0125	13.4024	0.4180
	2	0.7353	0.3757	16.3406	13.5826	0.4160
	5	0.7343	0.3731	16.6642	13.6796	0.4159
	10	0.7341	0.3705	16.7623	13.6997	0.4145
	50	0.7339	0.3721	16.8082	13.7027	0.4130
GKM	-	0.7680	0.5822	20.7516	14.4050	0.2322
SKM	1	0.7695	0.5780	8.7266	5.6547	0.2323
	2	0.7687	0.5802	10.1040	6.9978	0.2323
	5	0.7682	0.5814	12.9900	10.8082	0.2323
	10	0.7681	0.5817	16.2578	12.6779	0.2322
	50	0.7680	0.5821	19.7991	13.9992	0.2322

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - " " " the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum criterion procedure
GKM - " " " the ML estimation procedure

Table 7.2a MISE ($\times 10^{-2}$) of the density estimates from simulated data using robust estimate for σ in (7.11). N=100, $N_{sam}=N\times k$.

Method	k	Normal	Student-t with 5 d.f.	Lognormal	Cauchy	Bimodal
GKO	-	0.5413	0.5609	5.9527	0.8391	0.5748
SKO	1	0.5486	0.5565	5.3997	0.8606	0.5751
	2	0.5465	0.5560	5.4691	0.8173	0.5739
	5	0.5428	0.5563	5.6706	0.8059	0.5738
	10	0.5419	0.5565	5.7994	0.8000	0.5729
	50	0.5415	0.5562	5.8755	0.8296	0.5719
GKM	-	0.5708	0.6789	20.0251	14.3922	0.4967
SKM	1	0.5727	0.6739	9.7423	7.2553	0.4970
	2	0.5717	0.6765	12.8286	9.3714	0.4968
	5	0.5710	0.6776	16.9252	12.2474	0.4968
	10	0.5709	0.6782	18.7921	12.9061	0.4967
	50	0.5708	0.6783	19.8193	14.2587	0.4967

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - " " " the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum criterion procedure
GKM - " " " the ML estimation procedure

Table 7.2b EMSE ($\times 10^{-2}$) of the density estimates from simulated data using robust estimate for σ in (7.11). N=100, N_{sam}=N×k.

Method	k	Normal	Student-t with 5 d.f.	Lognormal	Cauchy	Bimodal
GKO	-	0.1403	0.1372	3.2322	0.1858	0.0839
SKO	1	0.1424	0.1344	2.8931	0.1920	0.0839
	2	0.1417	0.1347	2.9340	0.1798	0.0837
	5	0.1407	0.1351	3.0581	0.1765	0.0837
	10	0.1405	0.1352	3.1375	0.1748	0.0836
	50	0.1404	0.1351	3.1845	0.1831	0.0835
GKM	-	0.1492	0.1759	11.0535	3.5004	0.0738
SKM	1	0.1498	0.1742	5.5064	1.7385	0.0738
	2	0.1495	0.1751	7.2494	2.2342	0.0738
	5	0.1493	0.1754	9.4536	2.9279	0.0738
	10	0.1492	0.1756	10.4235	3.1017	0.0738
	50	0.1492	0.1757	10.9489	3.4681	0.0738

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - " " " the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum criterion procedure
GKM - " " " the ML estimation procedure

Table 7.2c ISE $\rceil$($\times 10^{-2}$) of the density estimates from simulated data using robust estimate for σ in (7.11). N=100, N_{sam}=N×k.

Method	k	Normal	Student-t with 5 d.f.	Lognormal	Cauchy	Bimodal
GKO	-	0.7338	0.3681	8.0940	1.0776	0.4174
SKO	1	0.7400	0.3342	7.6094	1.1092	0.4180
	2	0.7378	0.3420	7.6596	1.0501	0.4160
	5	0.7348	0.3475	7.8377	1.0336	0.4159
	10	0.7341	0.3503	7.9539	1.0252	0.4145
	50	0.7339	0.3492	8.0225	1.0648	0.4130
GKM	-	0.7680	0.5822	20.7516	14.4050	0.2322
SKM	1	0.7702	0.5718	11.6163	7.6011	0.2323
	2	0.7690	0.5771	14.3829	9.6427	0.2323
	5	0.7683	0.5794	17.9898	12.4010	0.2323
	10	0.7681	0.5806	19.6439	13.0083	0.2322
	50	0.7680	0.5810	20.5657	14.2765	0.2322

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - " " " the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum criterion procedure
GKM - " " " the ML estimation procedure

λ_{opt} obtained using a robust estimate for σ. In the Tables, N_{sam} is the bootstrap sample size as mentioned in Section 7.4.1, and k is a factor from which N_{sam} is obtained.

All three measures give a similar result. As expected, the method GKO comes out best when the data are Normal. The method SKO comes second when k equals 50. When the data are bimodal, methods GKM and SKM (with k = 10 & 50) perform better than SKO and GKO. However, for the long-tailed and skew distributions the Student-t kernel is superior to the Gaussian kernel. This is illustrated in Fig. 7.6. Note that an adaptive Gaussian kernel method is also employed to fit the data. Details of the adaptive kernel method are discussed in Section 2.1. It is designed to fit long tailed and skewed distributions better. However, it does not perform as well as expected in a univariate dimension (Breiman (1977)). All the methods are compatible with each other when the data come from Normal, Student and Bimodal densities (see Fig. 7.6 (a), (b) and (e) respectively). The Student-t kernel is a better fit for the Lognormal and Cauchy distributions than the adaptive method.

When a robust estimate for σ is used, the values of the ISE do not change much for the Normal and bimodal distributions but are reduced slightly for the Student distribution. However, for the Lognormal and Cauchy distributions, the ISE is reduced considerably and the density estimates for these distributions using SKO and GKO fit better than SKM and GKM which have a slight fluctuation around the tails (see Fig. 7.7 (c) and (d)).

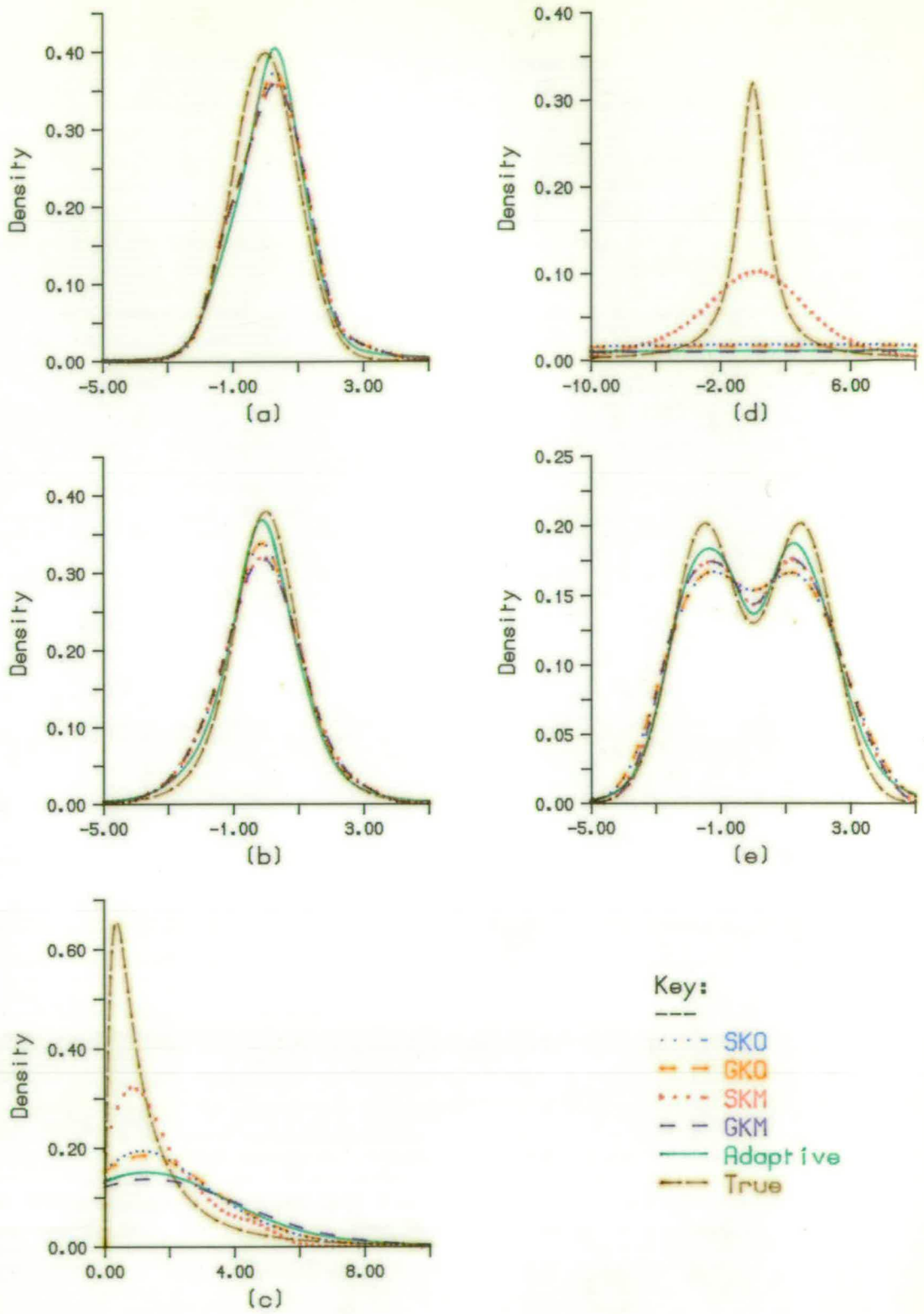


Fig. 7.6 Density plot of some simulated data -- (a) Normal, (b) Student, (c) Lognormal, (d) Cauchy and (e) Bimodal; Normal criteria.

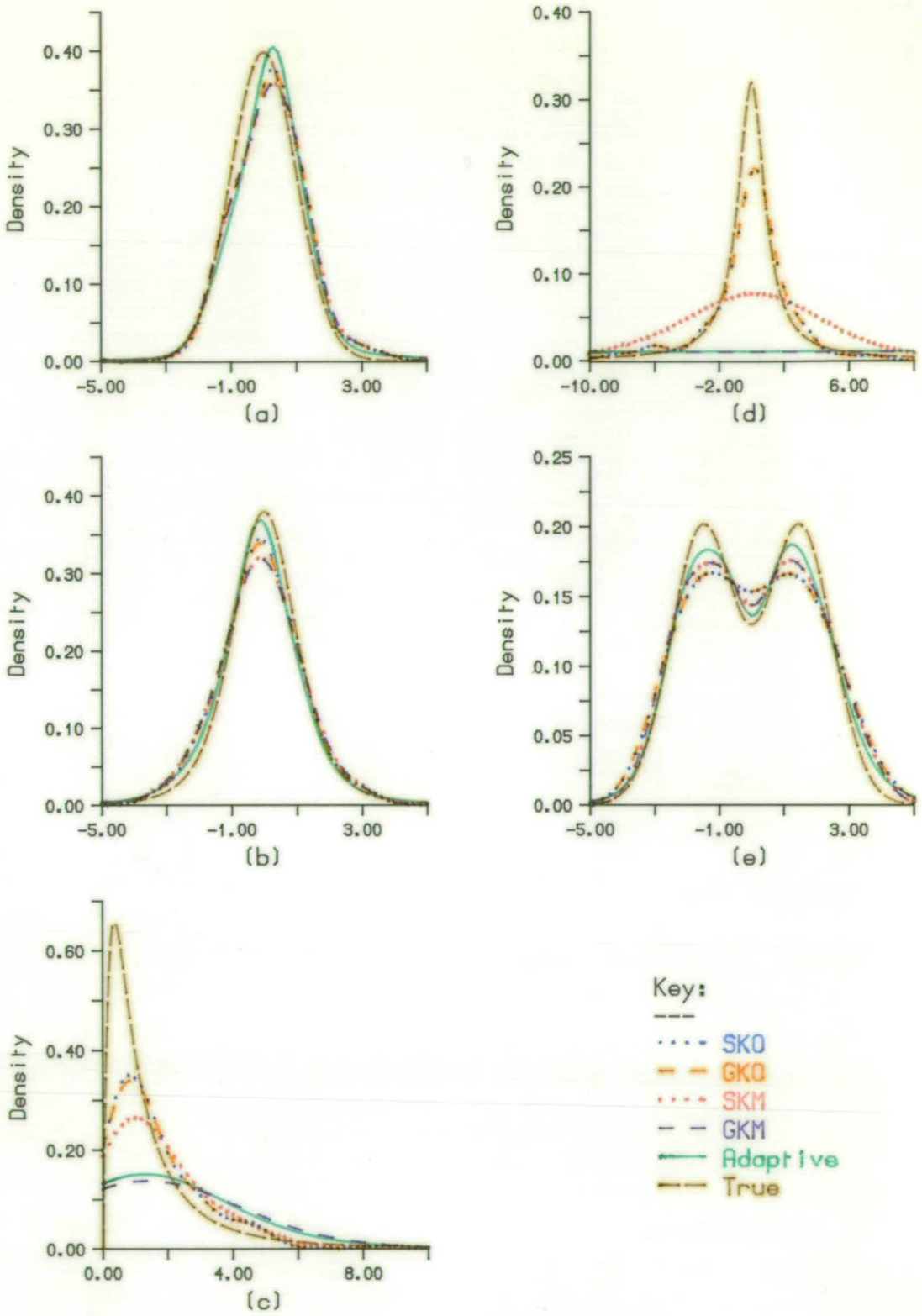


Fig. 7.7 Density plot of some simulated data -- (a) Normal, (b) Student, (c) Lognormal, (d) Cauchy and (e) Bimodal; Robust criteria.

Estimates of the parameters corresponding to the methods are shown in Tables 7.3 and 7.4 for the normal and robust criteria respectively. One interesting feature in the Tables, is that α is considerably smaller for the skewed and long-tailed distributions than the symmetric ones. It seems to suggest that there is uncertainty about these long tailed and skewed distributions since α is interpreted as the degree of knowledge about the smoothing parameter λ, which consequently reflects on the uncertainty about the distribution. However, when the robust criteria is used α increases slightly for the two afore-mentioned skewed and long-tailed distributions. This suggests a slight gain in confidence of λ, since a more cautious estimate is used for σ in (7.11). The prior densities of λ^2 with α and β estimated by the two procedures described in Section 7.4 given k=1 under the Normal and robust criteria are plotted in Figs 7.8 - 7.9 and 7.10 - 7.11, respectively.

To examine the reliability of these methods, more thorough simulations were performed. The sample sizes N were chosen to be 25, 50 and 100, and the number of simulations are 100, 50 and 25 respectively. Again all three measures give similar results, so only those of the ISE are discussed here.

Results of the other two measures are also presented in Tables which are given in the Appendix 7. Results of the ISE under the normal criterion are shown in Table 7.5 with (a) k=1, (b) k=2, (c) k=5, (d) k=10 and (e) k=50. Similar results are also obtained using the robust criterion and are shown in Table 7.6 with (a) k=1, (b) k=2, (c) k=5, (d) k=10 and (e) k=50. In general, SKO with a robust

Table 7.3 Esitmates of the parameters λ_{opt}, λ_{max}, $\hat{E}(\lambda^2)$, $\hat{V}ar(\lambda^2)$, α, β and β_M corresponding to the results shown in Tables 7.1.

Parameter	k	Normal	Student-t with 5 d.f.	Lognormal	Cauchy	Bimodal
λ_{opt}	-	0.4387	0.5071	1.7241	23.8683	0.7730
λ_{max}	-	0.5133	0.6161	2.5203	39.7239	0.5763
$\hat{E}(\lambda^2)$	1	0.1930	0.2583	3.0571	548.072	0.6005
	2	0.1896	0.2571	3.0324	575.654	0.5965
	5	0.1901	0.2552	3.0244	590.616	0.5960
	10	0.1911	0.2534	2.9950	580.328	0.5938
	50	0.1908	0.2544	2.9620	562.147	0.5914
$\hat{V}ar(\lambda^2)$	1	1.2×10^{-3}	1.5×10^{-3}	4.4330	2.4×10^{5}	3.3×10^{-3}
	2	5.3×10^{-4}	7.1×10^{-4}	2.3290	1.1×10^{5}	1.3×10^{-3}
	5	1.6×10^{-4}	2.9×10^{-4}	0.8814	4.6×10^{4}	7.0×10^{-4}
	10	8.6×10^{-5}	1.8×10^{-4}	0.3771	2.2×10^{4}	2.9×10^{-4}
	50	1.8×10^{-5}	2.8×10^{-5}	0.0737	4.9×10^{3}	6.2×10^{-5}
α	1	67	91	8	6	224
	2	139	191	12	10	540
	5	445	455	25	19	1021
	10	850	730	52	35	2463
	50	4122	4686	242	134	11261
β	1	1.25×10^{1}	2.30×10^{1}	1.83×10^{1}	2.19×10^{3}	1.33×10^{2}
	2	2.60×10^{1}	4.86×10^{1}	3.03×10^{1}	4.61×10^{3}	3.21×10^{2}
	5	8.42×10^{1}	1.16×10^{2}	6.96×10^{1}	1.00×10^{4}	6.07×10^{2}
	10	1.62×10^{2}	1.84×10^{2}	1.50×10^{2}	1.92×10^{4}	1.46×10^{3}
	50	7.86×10^{2}	1.19×10^{3}	7.11×10^{2}	7.42×10^{4}	6.66×10^{3}
β_M	1	1.71×10^{1}	3.37×10^{1}	3.32×10^{0}	3.83×10^{1}	7.39×10^{1}
	2	3.61×10^{1}	7.17×10^{1}	7.62×10^{0}	1.13×10^{2}	1.79×10^{2}
	5	1.17×10^{2}	1.72×10^{2}	3.25×10^{1}	1.11×10^{3}	3.39×10^{2}
	10	2.23×10^{2}	2.76×10^{2}	1.35×10^{2}	6.60×10^{3}	8.18×10^{2}
	50	1.09×10^{3}	1.78×10^{3}	1.28×10^{3}	1.11×10^{5}	3.74×10^{3}

Table 7.4 Esitmates of the parameters λ_{opt}, λ_{max}, $\hat{E}(\lambda^2)$, $\hat{V}ar(\lambda^2)$, α, β and β_M corresponding to the results shown in Tables 7.2.

Parameter	k	Normal	Student-t with 5 d.f.	Lognormal	Cauchy	Bimodal
λ_{opt}	-	0.4387	0.5017	0.6252	0.7596	0.7730
λ_{max}	-	0.5133	0.6161	2.5203	39.724	0.5763
$\hat{E}(\lambda^2)$	1	0.1799	0.2282	0.3532	0.6425	0.6005
	2	0.1791	0.2337	0.3491	0.5821	0.5965
	5	0.1862	0.2375	0.3638	0.5580	0.5960
	10	0.1897	0.2394	0.3754	0.5475	0.5938
	50	0.1907	0.2385	0.3823	0.5688	0.5914
$\hat{V}ar(\lambda^2)$	1	1.6×10^{-3}	3.0×10^{-3}	1.8×10^{-2}	1.0×10^{-1}	3.2×10^{-3}
	2	7.0×10^{-4}	1.5×10^{-3}	8.2×10^{-3}	5.2×10^{-2}	1.3×10^{-3}
	5	1.9×10^{-4}	8.8×10^{-4}	3.2×10^{-3}	2.4×10^{-2}	7.0×10^{-4}
	10	8.5×10^{-5}	4.9×10^{-4}	1.4×10^{-3}	1.5×10^{-2}	2.9×10^{-4}
	50	1.8×10^{-5}	3.8×10^{-4}	2.4×10^{-4}	1.8×10^{-3}	6.2×10^{-5}
α	1	46	39	18	12	224
	2	96	75	34	17	540
	5	366	132	86	30	1021
	10	850	237	209	44	2463
	50	4040	307	1208	368	11261
β	1	7.92×10^{0}	8.44×10^{0}	5.65×10^{0}	6.42×10^{0}	1.33×10^{2}
	2	1.68×10^{1}	1.71×10^{1}	1.12×10^{1}	8.73×10^{0}	3.21×10^{2}
	5	6.78×10^{1}	3.09×10^{1}	3.06×10^{1}	1.56×10^{1}	6.07×10^{2}
	10	1.61×10^{2}	5.63×10^{1}	7.77×10^{1}	2.30×10^{1}	1.46×10^{3}
	50	7.70×10^{2}	7.28×10^{1}	4.61×10^{2}	2.08×10^{2}	6.66×10^{3}
β_M	1	1.16×10^{1}	1.40×10^{1}	1.70×10^{1}	1.74×10^{2}	7.39×10^{1}
	2	2.48×10^{1}	2.77×10^{1}	6.00×10^{1}	5.77×10^{2}	1.79×10^{2}
	5	9.59×10^{1}	4.93×10^{1}	3.17×10^{2}	4.55×10^{3}	3.39×10^{2}
	10	2.23×10^{2}	8.92×10^{1}	1.07×10^{3}	1.12×10^{4}	8.18×10^{2}
	50	1.06×10^{3}	1.16×10^{2}	7.40×10^{3}	4.68×10^{5}	3.74×10^{3}

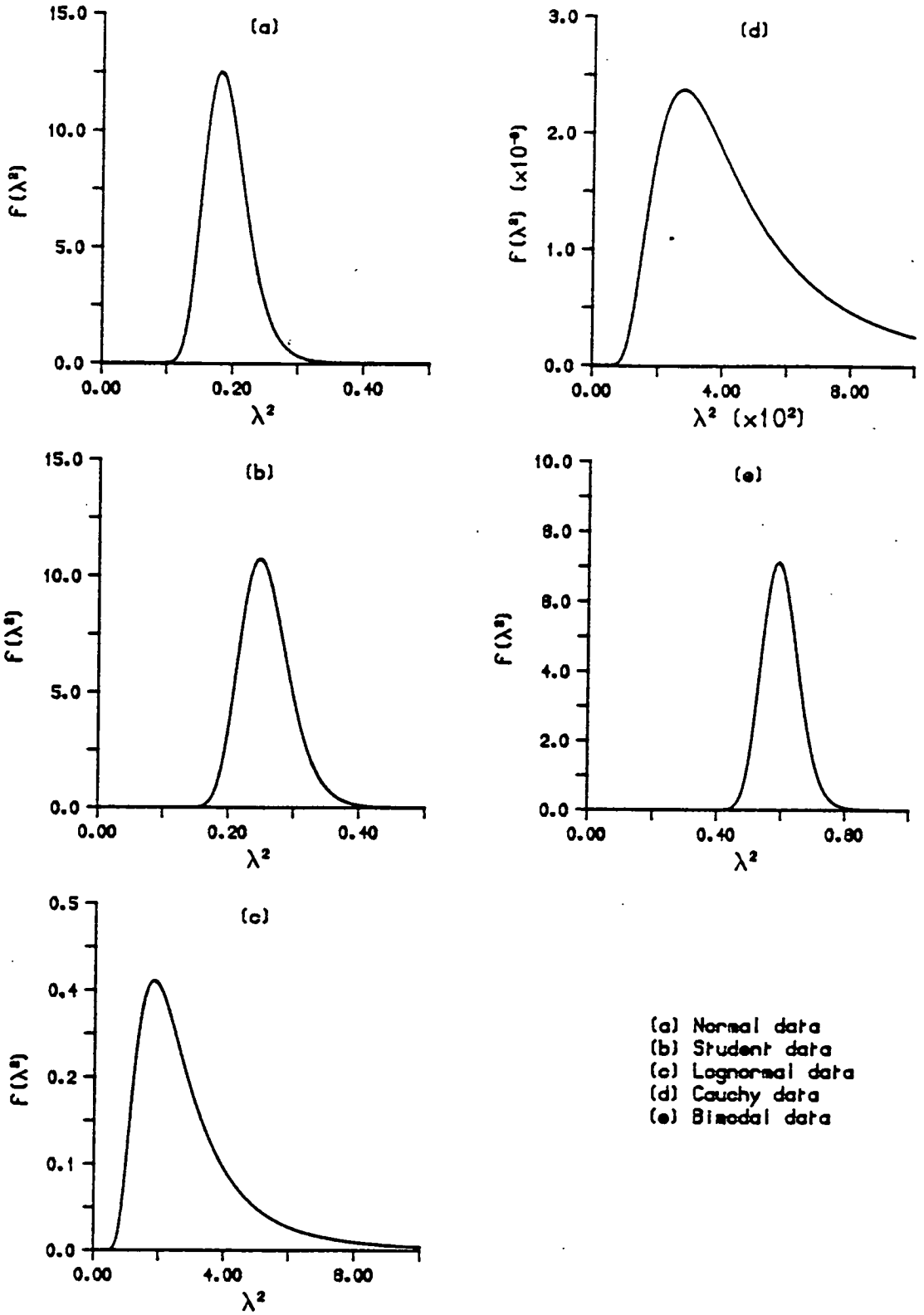


Fig. 7.8 Prior density plot of smoothing parameter λ^2 with hyperparameter β estimated by the moments methods (see Section 7.4.1) under the Normal criteria.

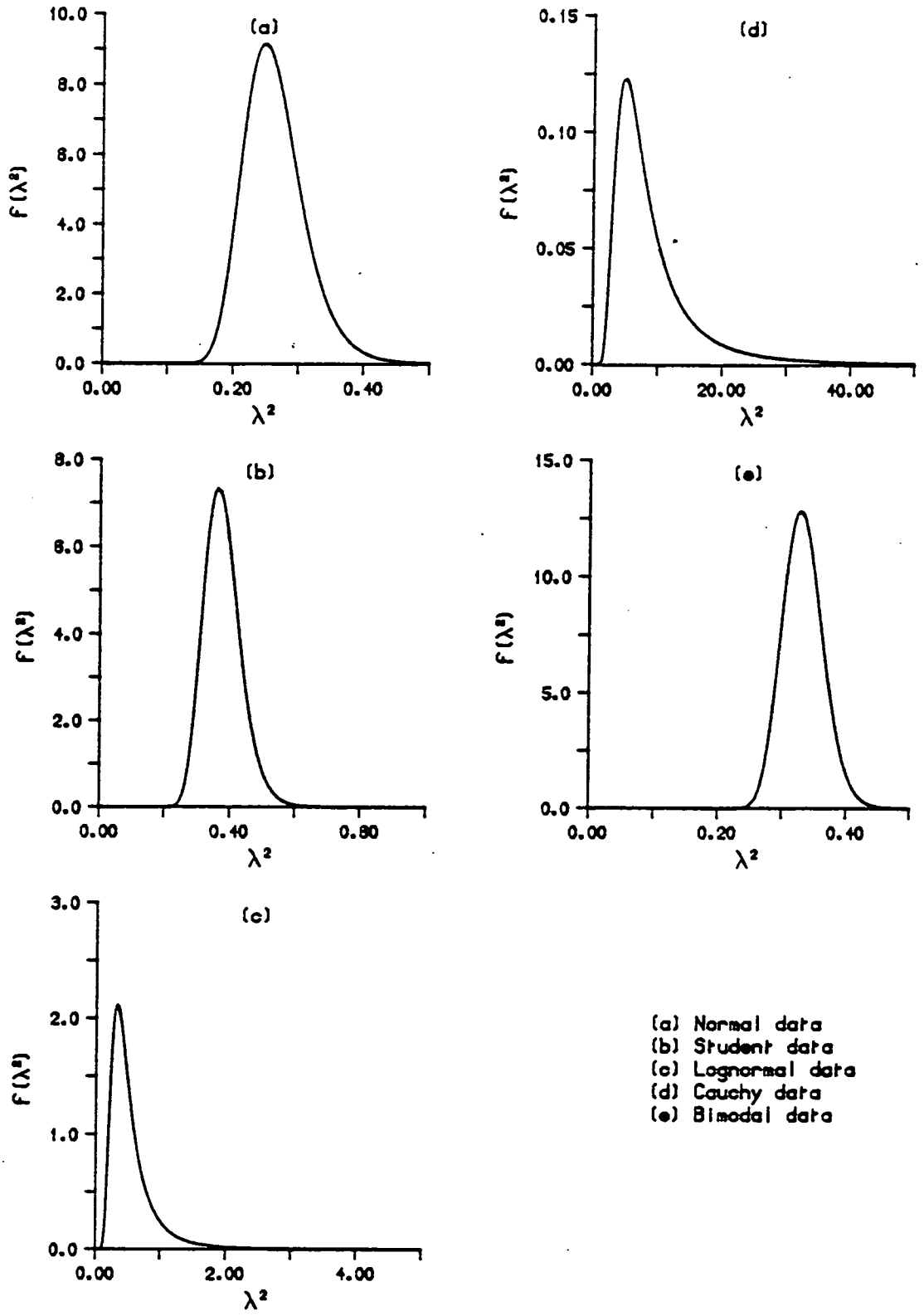


Fig. 7.9 Prior density plot of smoothing parameter λ^2 with hyperparameter β estimated by the modified ML methods (see Section 7.4.1) under the Normal criteria.

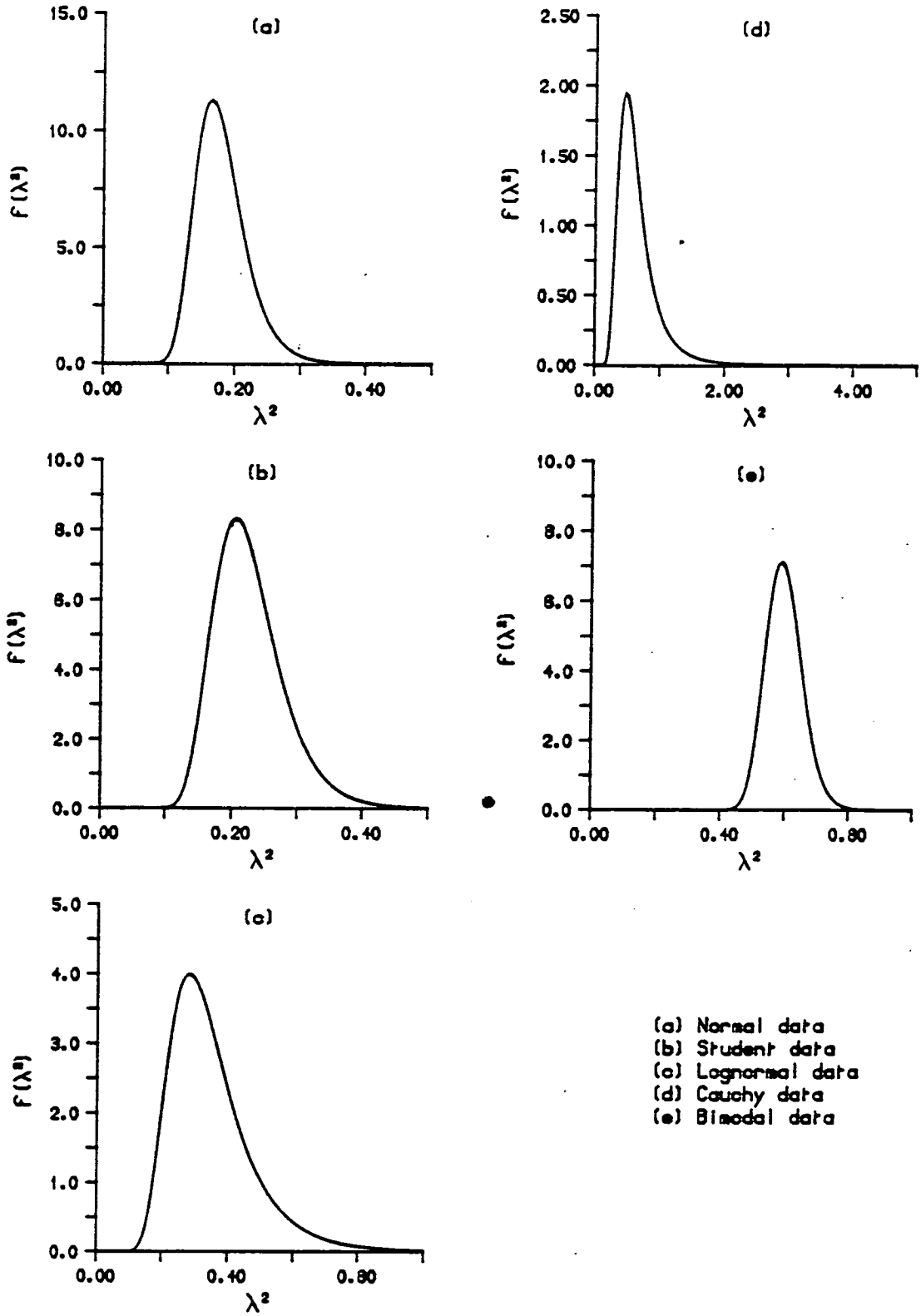


Fig. 7.10 Prior density plot of smoothing parameter λ^2 with hyperparameters β estimated by the moments method (see Section 7.4.1) under the robust criteria.

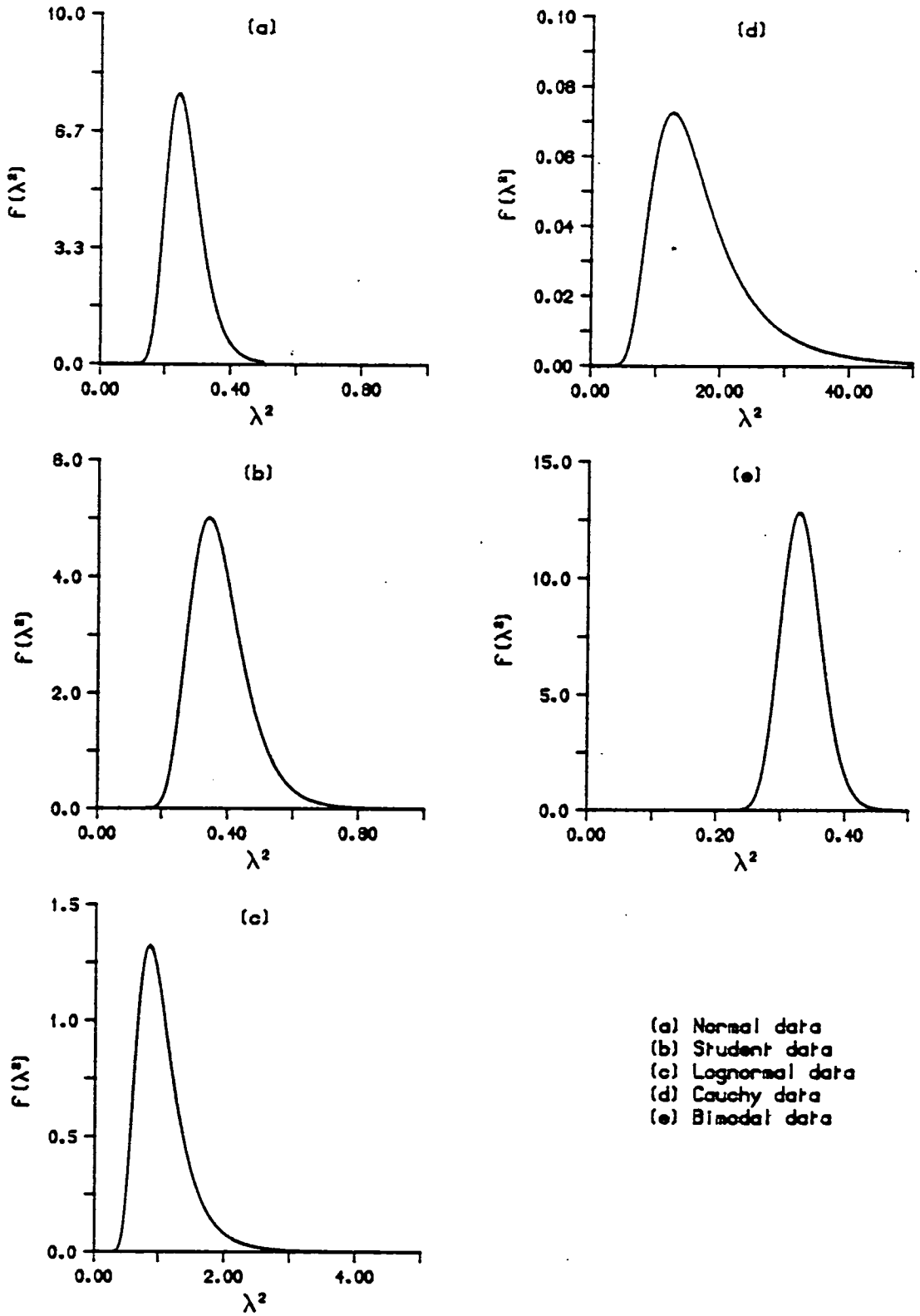


Fig. 7.11 Prior density plot of smoothing parameter λ^2 with hyperparameter β estimated by the modified ML methods (see Section 7.4.1) under the robust criteria.

Table 7.5a ISE ($\times 10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($\times 10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=1.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.6247	0.9290	0.5047
		(0.1354)	(0.0819)	(0.0685)
	GKO	1.5815	0.9149	0.5031
		(0.1337)	(0.0815)	(0.0686)
	SKM	2.2019	1.3037	0.5310
		(0.2145)	(0.1432)	(0.0744)
	GKM	2.1450	1.3046	0.5314
		(0.1921)	(0.1442)	(0.0748)
Student-t with 5 d.f.	SKO	1.5986	1.1658	0.5714
		(0.1288)	(0.1210)	(0.0795)
	GKO	1.6149	1.1767	0.5681
		(0.1346)	(0.1241)	(0.0795)
	SKM	2.0325	1.3391	0.8121
		(0.1980)	(0.1191)	(0.1107)
	GKM	2.1993	1.5792	0.8685
		(0.2307)	(0.2089)	(0.1241)
Lognormal	SKO	9.3680	9.1156	8.1226
		(0.4527)	(0.4750)	(0.4572)
	GKO	9.9444	9.6319	8.3490
		(0.4808)	(0.5139)	(0.4755)
	SKM	7.6655	6.3188	4.3899
		(0.3850)	(0.3909)	(0.4313)
	GKM	9.9869	8.5040	6.6310
		(0.6375)	(0.7023)	(0.8766)
Cauchy	SKO	6.8323	7.1344	7.9671
		(0.3940)	(0.5710)	(0.8115)
	GKO	7.2645	7.5431	8.3391
		(0.3987)	(0.5815)	(0.8080)
	SKM	4.9446	4.5460	3.2142
		(0.2314)	(0.3198)	(0.2559)
	GKM	8.1896	8.9297	8.6324
		(0.4252)	(0.6004)	(0.8277)
Bimodal	SKO	1.1945	0.7538	0.6184
		(0.0771)	(0.0463)	(0.0596)
	GKO	1.2072	0.7614	0.6222
		(0.0749)	(0.0457)	(0.0594)
	SKM	1.4667	0.9881	0.6707
		(0.1072)	(0.1088)	(0.0809)
	GKM	1.4635	1.0642	0.6710
		(0.1063)	(0.1540)	(0.0811)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table 7.5b ISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=2.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.5462	0.9910	0.5857
		(0.1151)	(0.1463)	(0.0887)
	GKO	1.5184	0.9852	0.5841
		(0.1148)	(0.1459)	(0.0886)
	SKM	1.8189	1.2017	0.6654
		(0.1652)	(0.1653)	(0.1105)
	GKM	1.8197	1.2511	0.6650
		(0.1659)	(0.1667)	(0.1107)
Student-t with 5 d.f.	SKO	1.4087	0.8704	0.6670
		(0.1031)	(0.0920)	(0.0940)
	GKO	1.4202	0.8812	0.6734
		(0.1027)	(0.0970)	(0.0936)
	SKM	1.9029	1.1690	1.2501
		(0.1399)	(0.1155)	(0.1597)
	GKM	2.0495	1.2944	1.4062
		(0.1557)	(0.1799)	(0.2111)
Lognormal	SKO	10.1611	9.3413	8.6779
		(0.4079)	(0.5082)	(0.6035)
	GKO	10.5754	9.6181	8.9695
		(0.4220)	(0.5212)	(0.6386)
	SKM	8.5943	6.7365	5.5722
		(0.4202)	(0.4039)	(0.5248)
	GKM	9.7740	8.8232	9.0828
		(0.5423)	(0.7867)	(1.2760)
Cauchy	SKO	6.7638	6.7711	6.9978
		(0.3979)	(0.6296)	(0.7032)
	GKO	7.0665	7.0195	7.2386
		(0.4011)	(0.6334)	(0.7059)
	SKM	5.6628	4.7947	4.7311
		(0.2716)	(0.3246)	(0.3367)
	GKM	8.1661	8.0798	8.0702
		(0.4236)	(0.6145)	(0.6819)
Bimodal	SKO	1.2130	0.8250	0.4694
		(0.0650)	(0.0535)	(0.0436)
	GKO	1.2227	0.8314	0.4719
		(0.0633)	(0.0535)	(0.0436)
	SKM	1.4965	0.9699	0.4491
		(0.1130)	(0.0849)	(0.0510)
	GKM	1.5272	1.0079	0.4490
		(0.1166)	(0.0953)	(0.0510)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table 7.5c ISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=5.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.5354	0.8783	0.5706
		(0.1213)	(0.0993)	(0.0732)
	GKO	1.5116	0.8710	0.5696
		(0.1194)	(0.0993)	(0.0733)
	SKM	1.9031	1.1055	0.6610
		(0.1819)	(0.1468)	(0.0938)
	GKM	1.9025	1.1056	0.6610
		(0.1823)	(0.1468)	(0.0938)
Student-t with 5 d.f.	SKO	1.6119	1.0329	0.6352
		(0.1136)	(0.1018)	(0.0785)
	GKO	1.6199	1.0340	0.6375
		(0.1148)	(0.1024)	(0.0786)
	SKM	2.1740	1.3166	1.0662
		(0.1893)	(0.1073)	(0.1327)
	GKM	2.1953	1.3475	1.0896
		(0.1901)	(0.1075)	(0.1392)
Lognormal	SKO	10.7028	9.6026	8.0979
		(0.4806)	(0.5011)	(0.3980)
	GKO	11.0328	9.7862	8.2103
		(0.4951)	(0.5123)	(0.4049)
	SKM	9.7631	8.1680	6.0121
		(0.4862)	(0.5657)	(0.5868)
	GKM	11.0616	9.1657	6.7202
		(0.6571)	(0.7228)	(0.7211)
Cauchy	SKO	7.0309	8.1700	8.3874
		(0.4500)	(0.5298)	(0.6406)
	GKO	7.2073	8.3547	8.5738
		(0.4508)	(0.5286)	(0.6475)
	SKM	6.8289	6.4511	5.7613
		(0.3517)	(0.3481)	(0.3939)
	GKM	8.0861	9.4782	9.7104
		(0.4514)	(0.5223)	(0.7209)
Bimodal	SKO	1.2325	0.8786	0.4966
		(0.0631)	(0.0673)	(0.0383)
	GKO	1.2421	0.8833	0.4999
		(0.0620)	(0.0665)	(0.0382)
	SKM	1.4872	0.9389	0.5789
		(0.1157)	(0.0730)	(0.0648)
	GKM	1.5068	0.9388	0.5790
		(0.1154)	(0.0730)	(0.0648)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table 7.5d ISE ($\times 10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($\times 10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=10.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.2755	0.9571	0.5633
		(0.1086)	(0.1033)	(0.0691)
	GKO	1.2605	0.9509	0.5616
		(0.1073)	(0.1032)	(0.0692)
	SKM	1.7153	1.1476	0.6143
		(0.1666)	(0.1239)	(0.0778)
	GKM	1.7646	1.1475	0.6142
		(0.1694)	(0.1237)	(0.0777)
Student-t with 5 d.f.	SKO	1.6875	0.8061	0.5968
		(0.1489)	(0.0801)	(0.0708)
	GKO	1.6851	0.8126	0.5996
		(0.1466)	(0.0804)	(0.0712)
	SKM	2.3781	1.2228	0.9275
		(0.1941)	(0.1038)	(0.1062)
	GKM	2.3940	1.2516	0.9341
		(0.1948)	(0.1083)	(0.1078)
Lognormal	SKO	9.4455	8.7484	9.4546
		(0.4331)	(0.4000)	(0.7080)
	GKO	9.6642	8.8779	9.5626
		(0.4386)	(0.4076)	(0.7169)
	SKM	9.0565	7.3645	8.4267
		(0.5057)	(0.5594)	(0.8741)
	GKM	9.4629	7.6904	10.3178
		(0.5571)	(0.6382)	(1.3439)
Cauchy	SKO	7.0196	7.4545	9.4609
		(0.3828)	(0.5674)	(0.7307)
	GKO	7.1566	7.5564	9.5548
		(0.3846)	(0.5679)	(0.7268)
	SKM	7.5021	7.2947	7.0624
		(0.3749)	(0.4681)	(0.3962)
	GKM	8.0578	8.3452	10.4242
		(0.4144)	(0.5943)	(0.6830)
Bimodal	SKO	1.1759	0.8048	0.5645
		(0.0674)	(0.0483)	(0.0567)
	GKO	1.1830	0.8106	0.5670
		(0.0659)	(0.0479)	(0.0565)
	SKM	1.6650	0.9059	0.6623
		(0.1602)	(0.0734)	(0.0997)
	GKM	1.6648	0.9606	0.6623
		(0.1600)	(0.0843)	(0.0997)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table 7.5e ISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=50.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.4949	0.9594	0.5469
		(0.1007)	(0.0903)	(0.0811)
	GKO	1.4738	0.9557	0.5461
		(0.0991)	(0.0902)	(0.0812)
	SKM	1.6794	1.3371	0.6670
		(0.1266)	(0.1545)	(0.0937)
	GKM	1.6831	1.3371	0.6670
		(0.1264)	(0.1545)	(0.0937)
Student-t with 5 d.f.	SKO	1.4532	0.8315	0.6166
		(0.1042)	(0.0913)	(0.0870)
	GKO	1.4594	0.8380	0.6183
		(0.1049)	(0.0924)	(0.0872)
	SKM	2.1031	1.4142	0.9339
		(0.1534)	(0.1843)	(0.1077)
	GKM	2.1084	1.4265	0.9348
		(0.1541)	(0.1894)	(0.1079)
Lognormal	SKO	9.9947	8.8099	7.8561
		(0.4782)	(0.3934)	(0.6794)
	GKO	10.1616	8.9044	7.9040
		(0.4833)	(0.3973)	(0.6823)
	SKM	9.7568	8.3394	6.7553
		(0.5942)	(0.6839)	(0.9584)
	GKM	9.8392	8.4915	6.9071
		(0.6060)	(0.7152)	(1.0314)
Cauchy	SKO	7.5312	7.3573	8.0882
		(0.4170)	(0.5965)	(0.8581)
	GKO	7.6205	7.4102	8.1196
		(0.4160)	(0.5964)	(0.8580)
	SKM	8.4658	7.8428	8.3867
		(0.4253)	(0.6244)	(0.8781)
	GKM	8.5513	7.9678	8.7975
		(0.4285)	(0.6377)	(0.9547)
Bimodal	SKO	1.3019	0.7568	0.5104
		(0.0694)	(0.0589)	(0.0396)
	GKO	1.3082	0.7619	0.5135
		(0.0677)	(0.0584)	(0.0395)
	SKM	1.5006	0.8419	0.5074
		(0.1070)	(0.0827)	(0.0635)
	GKM	1.5007	0.8419	0.5074
		(0.1070)	(0.0827)	(0.0635)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table **7.6a** ISE ($\times 10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($\times 10^{-3}$) in brackets. Robust estimate for σ. Nsam=k*N where k=1.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.6487	1.0169	0.5660
		(0.1324)	(0.1197)	(0.0786)
	GKO	1.6166	1.0329	0.5593
		(0.1325)	(0.1297)	(0.0782)
	SKM	1.7728	1.0715	0.7861
		(0.1645)	(0.1186)	(0.1155)
	GKM	1.7694	1.0737	0.7881
		(0.1637)	(0.1206)	(0.1163)
Student-t with 5 d.f.	SKO	1.6294	0.9362	0.5738
		(0.1031)	(0.0844)	(0.0708)
	GKO	1.5865	0.9253	0.5754
		(0.1064)	(0.0841)	(0.0700)
	SKM	1.9766	1.3341	0.8536
		(0.1959)	(0.1017)	(0.1211)
	GKM	2.0711	1.4397	0.9645
		(0.2054)	(0.1158)	(0.1506)
Lognormal	SKO	7.2876	4.9391	4.7197
		(0.3266)	(0.2860)	(0.3648)
	GKO	7.5031	4.9973	4.7511
		(0.3424)	(0.2922)	(0.3652)
	SKM	8.1321	5.4954	5.1839
		(0.4075)	(0.3469)	(0.4914)
	GKM	11.3421	8.9552	6.8002
		(0.6289)	(0.7628)	(0.8336)
Cauchy	SKO	1.9826	1.2077	0.8793
		(0.1292)	(0.0996)	(0.1205)
	GKO	1.8770	1.1997	0.8726
		(0.1147)	(0.1006)	(0.1214)
	SKM	4.2665	4.9059	5.2300
		(0.1995)	(0.3081)	(0.4030)
	GKM	7.8574	9.0951	10.4384
		(0.3988)	(0.6418)	(0.8154)
Bimodal	SKO	1.2710	0.8784	0.5377
		(0.0723)	(0.0698)	(0.0445)
	GKO	1.2710	0.8852	0.5426
		(0.0678)	(0.0684)	(0.0443)
	SKM	1.6508	1.0432	0.4989
		(0.1500)	(0.0985)	(0.0570)
	GKM	1.6541	1.0447	0.4990
		(0.1531)	(0.0985)	(0.0570)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table 7.6b ISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=2.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.6034	1.0594	0.5886
		(0.1438)	(0.1087)	(0.0827)
	GKO	1.5508	1.0495	0.5925
		(0.1384)	(0.1073)	(0.0829)
	SKM	1.8034	1.2284	0.6807
		(0.1508)	(0.1865)	(0.1071)
	GKM	1.8251	1.2280	0.6812
		(0.1528)	(0.1871)	(0.1075)
Student-t with 5 d.f.	SKO	1.5866	0.9373	0.5263
		(0.1062)	(0.0991)	(0.0644)
	GKO	1.5324	0.9415	0.5203
		(0.1048)	(0.1010)	(0.0630)
	SKM	2.2466	1.3625	0.8895
		(0.2608)	(0.1375)	(0.1179)
	GKM	2.4289	1.4171	0.9431
		(0.2700)	(0.1430)	(0.1324)
Lognormal	SKO	6.9655	5.8697	4.2589
		(0.3447)	(0.3284)	(0.2184)
	GKO	7.2453	6.0474	4.2655
		(0.3601)	(0.3403)	(0.2194)
	SKM	8.1601	7.0197	6.2648
		(0.3932)	(0.5336)	(0.5955)
	GKM	10.3692	8.5683	8.6283
		(0.5875)	(0.6823)	(0.8936)
Cauchy	SKO	2.2237	1.0555	0.8643
		(0.1341)	(0.0900)	(0.0891)
	GKO	2.2327	1.0545	0.8669
		(0.1287)	(0.0880)	(0.0915)
	SKM	5.2997	5.9970	6.6285
		(0.2740)	(0.4272)	(0.4946)
	GKM	7.9823	8.7719	9.3773
		(0.4293)	(0.6221)	(0.7398)
Bimodal	SKO	1.1184	0.7728	0.5606
		(0.0571)	(0.0447)	(0.0386)
	GKO	1.1190	0.7798	0.5636
		(0.0537)	(0.0443)	(0.0383)
	SKM	1.6374	0.9261	0.6439
		(0.1166)	(0.0722)	(0.0814)
	GKM	1.6375	0.9265	0.6441
		(0.1168)	(0.0722)	(0.0816)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table 7.6c ISE ($\times 10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($\times 10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=5.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.4612	1.0083	0.5519
		(0.1181)	(0.1457)	(0.0557)
	GKO	1.4229	0.9790	0.5529
		(0.1141)	(0.1408)	(0.0560)
	SKM	1.6715	1.0987	0.8043
		(0.1378)	(0.1937)	(0.1082)
	GKM	1.6751	1.0972	0.8032
		(0.1383)	(0.1941)	(0.1083)
Student-t with 5 d.f.	SKO	1.6344	1.0408	0.6594
		(0.1203)	(0.0986)	(0.1027)
	GKO	1.6032	1.0391	0.6631
		(0.1175)	(0.0964)	(0.1027)
	SKM	2.1942	1.5198	1.1972
		(0.1913)	(0.1203)	(0.1639)
	GKM	2.2488	1.5722	1.2638
		(0.1970)	(0.1296)	(0.1795)
Lognormal	SKO	6.5934	5.7284	4.7440
		(0.3251)	(0.3717)	(0.3325)
	GKO	6.7981	5.9101	4.8481
		(0.3273)	(0.3763)	(0.3324)
	SKM	8.1107	7.2256	7.4559
		(0.4675)	(0.5684)	(1.1558)
	GKM	9.0780	8.0549	8.1791
		(0.5392)	(0.6636)	(1.3013)
Cauchy	SKO	1.8954	1.1328	0.7205
		(0.1215)	(0.0947)	(0.0873)
	GKO	1.9327	1.1375	0.7336
		(0.1244)	(0.0911)	(0.0884)
	SKM	5.7175	7.9851	7.8657
		(0.2958)	(0.4742)	(0.6588)
	GKM	7.2143	9.8515	9.1775
		(0.3926)	(0.5533)	(0.7864)
Bimodal	SKO	1.2840	0.8248	0.5979
		(0.0770)	(0.0575)	(0.0484)
	GKO	1.2797	0.8292	0.6003
		(0.0738)	(0.0570)	(0.0483)
	SKM	1.6003	1.0280	0.6448
		(0.1418)	(0.0944)	(0.0728)
	GKM	1.5993	1.0282	0.6448
		(0.1419)	(0.0944)	(0.0727)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table 7.6d ISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=10.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.4754	0.9081	0.7002
		(0.1042)	(0.0955)	(0.0867)
	GKO	1.4196	0.8909	0.6989
		(0.1007)	(0.0939)	(0.0867)
	SKM	1.9445	1.1380	0.7160
		(0.2399)	(0.1400)	(0.0898)
	GKM	1.9524	1.1228	0.7153
		(0.2387)	(0.1306)	(0.0897)
Student-t with 5 d.f.	SKO	1.7906	0.9407	0.5424
		(0.1311)	(0.0903)	(0.0504)
	GKO	1.7345	0.9394	0.5434
		(0.1279)	(0.0897)	(0.0510)
	SKM	2.2371	1.3670	1.1593
		(0.1681)	(0.1223)	(0.1814)
	GKM	2.2971	1.3752	1.1847
		(0.1739)	(0.1227)	(0.1890)
Lognormal	SKO	6.5757	5.5660	4.1423
		(0.3101)	(0.3241)	(0.2525)
	GKO	6.8974	5.7437	4.1982
		(0.3207)	(0.3262)	(0.2555)
	SKM	9.3937	7.9390	7.1303
		(0.5688)	(0.6347)	(0.8930)
	GKM	10.4080	8.7325	7.7377
		(0.6522)	(0.7393)	(0.9817)
Cauchy	SKO	1.8795	0.9959	0.8214
		(0.1186)	(0.0858)	(0.0926)
	GKO	1.9261	1.0206	0.8292
		(0.1227)	(0.0893)	(0.0943)
	SKM	7.3734	7.3900	8.8535
		(0.3837)	(0.5616)	(0.8955)
	GKM	8.5460	7.9925	9.5898
		(0.4323)	(0.5830)	(0.9229)
Bimodal	SKO	1.2766	0.8856	0.5438
		(0.0721)	(0.0562)	(0.0453)
	GKO	1.2805	0.8911	0.5471
		(0.0701)	(0.0555)	(0.0454)
	SKM	1.7305	1.0015	0.5481
		(0.1441)	(0.0889)	(0.0646)
	GKM	1.7301	1.0016	0.5481
		(0.1441)	(0.0889)	(0.0646)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table 7.6e ISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=50.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.8165	0.8636	0.5258
		(0.1985)	(0.0969)	(0.0720)
	GKO	1.7269	0.8491	0.5163
		(0.1894)	(0.0950)	(0.0705)
	SKM	2.1409	1.0309	0.6443
		(0.2224)	(0.1125)	(0.1019)
	GKM	2.1446	1.0305	0.6449
		(0.2217)	(0.1121)	(0.1021)
Student-t with 5 d.f.	SKO	1.8123	1.0162	0.6422
		(0.1425)	(0.1114)	(0.0946)
	GKO	1.7429	1.0035	0.6390
		(0.1375)	(0.1120)	(0.0947)
	SKM	2.1724	1.4160	1.2251
		(0.1597)	(0.1685)	(0.2135)
	GKM	2.1921	1.4189	1.2298
		(0.1605)	(0.1690)	(0.2145)
Lognormal	SKO	7.6618	5.5157	4.8877
		(0.3422)	(0.3050)	(0.3911)
	GKO	8.0282	5.6519	4.9415
		(0.3523)	(0.3092)	(0.3952)
	SKM	10.1978	7.6759	7.8797
		(0.5449)	(0.5470)	(1.0330)
	GKM	10.6217	7.9166	7.9957
		(0.5897)	(0.5749)	(1.0501)
Cauchy	SKO	1.9069	1.3050	0.7718
		(0.1307)	(0.1054)	(0.0897)
	GKO	1.9864	1.3300	0.7779
		(0.1336)	(0.1083)	(0.0911)
	SKM	7.4241	8.2291	8.8246
		(0.4265)	(0.5347)	(0.7624)
	GKM	7.9669	8.4056	8.9104
		(0.4477)	(0.5468)	(0.7671)
Bimodal	SKO	1.2486	0.8714	0.6144
		(0.0635)	(0.0703)	(0.0484)
	GKO	1.2396	0.8774	0.6168
		(0.0588)	(0.0699)	(0.0482)
	SKM	1.6607	0.9503	0.6980
		(0.1354)	(0.0942)	(0.0741)
	GKM	1.6638	0.9503	0.6980
		(0.1363)	(0.0942)	(0.0742)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

estimate for σ shows an improvement over the GKO method for skewed and long tail distributions. The Student-t kernel is slightly superior to the Gaussian kernel for most of the distributions considered here, so far as the Maximum likelihood procedure is concerned. However, the optimal criterion procedure based on the Normality assumption is better than the Maximum likelihood leave-one-out (ML for short) method. This is consistent with the simulation study done by Bowman (1985) who also found that the normal optimal solution for the smoothing parameter works exceptionally well for most distributions.

The standard errors of the estimates in the Tables show that Student-t kernel is more reliable than the Gaussian kernel as far as the ML estimation procedure is concerned. And the choice of N_{sam} does not seem to matter too much and generally gives similar results for the Student-t kernel method. Nevertheless, it appears in most cases that MISE, EMSE and ISE all increase as N_{sam} increases for the SKM method. This is in contrast with the results obtained for SKO. Overall when the moments and ML procedures are compared using the Student kernel, the former procedure always gives a smaller MISE, EMSE and ISE.

In summary, the simulation results show that the Student-t kernel fits the data better in terms of MISE, EMSE and ISE for skewed and long-tailed distributions. Also that, the Student-t kernel density estimate is more reliable than the Gaussian kernel when the ML leave-one-out estimation procedure is used. It is easily seen that the Student-t kernel can be extended to the multivariate case.

However, the Gaussian kernel density estimate with λ estimated by the ML leave-one-out procedure is known to be less reliable for the long-tailed distributions. So it would be interesting to compare the Student-t kernel with other kernel density estimates using various methods of estimation for the smoothing parameter λ. These other kernel density estimates with different estimation methods of obtaining $\hat{\lambda}$ were extensively compared by Bowman (1985) in a simulation study.

CHAPTER 8

CONCLUSIONS

8.1 Introduction

This thesis has discussed a kind of nonparametric empirical Bayes modelling method. The main field of application has been in forensic problems and random effects model. Here in this chapter, conclusion of the estimation of the Bayes' factor in a forensic context, Bayesian approach to variance components estimation and density estimation will be drawn. Future research on these topics will also be suggested.

8.2 Conclusions

The following subsections conclude the main topics in the thesis.

8.2.1 Estimation of the Bayes' factor

In a simulation study, the adaptive kernel method for estimation of the density of the random factor in a bivariate case appeared to be better, in terms of ISE, than the ordinary kernel method. However, in a similar study for a univariate case, the performance of the adaptive kernel method was not so good as the ordinary kernel. This confirms the findings by Breiman et al (1977). Results from these simulation studies suggested that the sample means of the groups in the random effects model are preferred to individual observations in the training data for estimation of the distribution of the random factor, especially when there is a clear random

structure in the training data and the random factors are not Normally distributed.

However, in spite of the afore-mentioned simulation studies, the adaptive kernel method seemed to improve the behaviour of the Bayes' factor. The adaptive kernel method allowed for the longer tail, where little information was available. This good property was demonstrated in Chapter 5 using the ECA model.

8.2.2 Analysis of Variance

The problem encountered here is the identifiability between the between-group variance and the smoothing parameter, when the maximum likelihood estimation method was used. However, with the smoothing parameter fixed at an objective pre-determined value, the within- and between-group variance estimates using the proposed kernel model were shown to be equivalent to the usual ANOVA estimates.

Within the Bayesian framework, posterior distributions for the variance components with vague priors for the parameters were derived incorporating kernel density functions. There is no specific analytical form of the posterior distributions under the kernel model. The structure is complicated hence numerical integration was used to carry out most of the evaluations. The modes of these posterior distributions were used as estimates for the variance components. Results from a small simulation study suggested that the proposed kernel model underestimated the between groups variance. The posterior distributions obtained from the proposed kernel model and the estimates derived from them were sensitive to an unknown

factor. Exactly what is causing the problem and further improvement of the model is a matter for further research.

8.2.3 Student-t kernel

A Student-t kernel was derived within a Bayesian framework after introduction of a prior for the smoothing parameter. The efficiency of the Student-t kernel was shown to be equivalent to the Gaussian kernel when the degrees of freedom, one of the hyperparameters, tends to infinity. Two methods of obtaining an objective choice of the hyper-parameters were suggested. They were the modified leave-one-out maximum likelihood and the method of moments which was based on the optimum smoothing parameter determined from Normality assumption. In an extensive simulation study, the Student-t kernel with the modified leave-one-out ML method performed better in terms of MISE, ISE and EMSE than the Gaussian kernel with a similar ML estimation method when the underlying true density was skewed and long-tailed. The simulation studies also suggested that the Student-t kernel with the hyper-parameters estimated by the method of moments performed better, in terms of MISE, ISE and EMSE, than with the modified leave-one-out ML estimation method.

8.3 Further research

In addition to future work suggested in the discussion section of each chapter, the following list of further research is outlined:

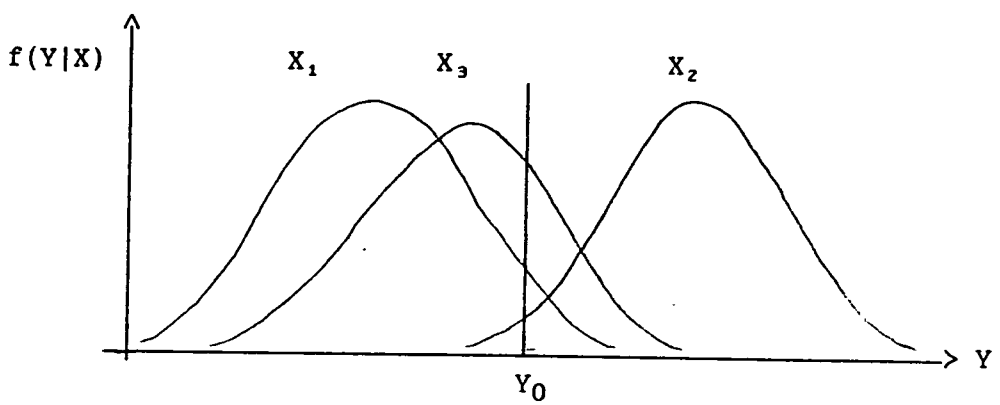
i) The Bayes' factor or the likelihood ratio, $f(Y|X,C)/f(Y|\bar{C})$, may be estimated directly (see Silverman (1978c)). However, interpretation of such method may not be apparent in a Forensic context since the

distribution of the unknown group population means are relevant in modelling the Bayes' factor.

ii) The problem of more than one suspect and one item of material of interest has been dealt with by Makov (1987) using a Bayesian approach. However the case of more than one suspect problem can be viewed as a discrimination problem by considering

$$\frac{f(Y|X_1)}{f(Y|X_2)} > k$$

for some positive value of k. Once the most likely suspect , say suspect 1, is identified, then the evidence relative to the suspect 1 could be weighed using the method developed in Chapter 3 for an univariate case or in Chapter 6 for the multivariate case. For example let suppose there are three suspects and let X_1, X_2 and X_3 represent the control data corresponding to Suspects 1, 2 and 3. In view of Fig. 3.3, the predictive distributions of Y given the three values of X may look something like the following.



Then if Y_0 is observed, for k = 1 a feasible conclusion may be drawn from the diagram that the recovered data Y is most likely coming from the same source as X_3 since

$$f(Y|X_1) > f(Y|X_2),$$

$$f(Y|X_3) > f(Y|X_1),$$

and $$f(Y|X_3) > f(Y|X_2).$$

Then to evaluate the strength of the evidence against suspect 3 simply consider the following ratio

$$\frac{f(Y|X_3,C)}{f(Y|\bar{C})}.$$

iii) Derive the error probabilities theoretically for the proposed model. In Section 3.8.2, Type I and II error probabilities were estimated via a simulation study for an example given in Chapter 3. Lindley (1977) provided the theoretical Type I and II error probabilities given some threshold values under the Normal model. The derivation of the error probabilities under the proposed model would be complicated but it would be worthwhile.

iv) Within-group variances are not homogenous over all groups. In Chapters 3, 5 and 6, when the Bayes' factor was modelled it was under the assumption that the within-group variances are known. So, when the Bayes' factor is evaluated the pooled estimate of the within-group variance from the training data was used to substitute the exact 'known' value. If tests for homogeneity of variance show that there is significant difference among the group variances then perhaps the pooled estimate may not be a sensible estimate to use. A few words about the tests for homogeneity: If there are only two

groups, standard *F* ratio test can be used to test for *HO*: $\sigma_1^2 = \sigma_2^2$ vs *H1*: $\sigma_1^2 \neq \sigma_2^2$. For there are more than two groups, Bartlett (1937) has provided such a test. However, there is a set back of this test, which is that it gives too many significant results with observations that come from a long-tailed distribution, i.e. distribution with positive kurtosis. To avoid this complication, one could use Levene's (1960) approximate test which is much less sensitive to non-Normality in the within-group data. More concise details of the two test statistics can be found in Snedecor and Cochran (1981).

If the test significantly rejects *HO*, one solution to the problem is to use another estimate for the within-group variance, which would take into account of the heterogenity of variance among the groups. Another solution to this is to apply the unknown within-group variance model with hyperparameter of the prior distribution for the within-group variance estimated from the training data using the method of moments described in Chapter 7.

v) Unbalanced nature of the training data. The effect of the unbalanced nature of the training data has not been considered. If the training data were very unbalanced, it could lead to homogenous variance over all groups. Moreover the use of sample group means might not be the best estimate for the random factors since the sample group means $\bar{z}_{i.}$ may no longer be a 'good' estimate for μ_i. Bayesian estimates may be applied, which depend on the values of the within- and between group variances (see Section 3.8 for discussion).

vi) Application of the Student-t kernel in the ECA model. In Chapter 5, we considered a single hair problem for a particular mixture data,

in which the ECA model was modified. The ECA model concerns the use of kernel density to estimate the denominator of the Bayes' factor. The ordinary kernel method is well known to produce density function which tends to zero much quicker than one would like especially in the estimation of a ratio situations and this would lead to spurious behaviour of the Bayes' factor. The adaptive kernel method has provided a solution to this problem. It would be interesting if the Student-t kernel could do the same so that it might establish its value in this already well elaborated density estimation field.

APPENDIX 1

NOTES ON VARIOUS DISTRIBUTIONS

The material in this appendix contains distributions which are used throughout the thesis. Some of which were used as prior in Chapter 3, for instance. Most of them were used in simulation studies. Each of the distributions is defined by a density function, and some of their properties are outlined.

a. The Cauchy distribution

Let U be a real random variable with probability desnity function

$$f(u|a,b) = \frac{1}{\pi b\{1+[(u-a)/b]^2\}}, \qquad u \in \mathbb{R}$$

where a is the location parameter and b is the scale parameter.

b. The Gamma distribution

Let U be a positive random variable with probability density function

$$f(u|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} u^{\alpha-1} e^{-u\beta}, \qquad u > 0,$$

where $\alpha > 0$ and $\beta > 0$, thus U is said to have a Gamma distribution with parameters α and β. This is denoted by $U \sim Ga(\alpha,\beta)$ and it can be shown that

$$E(U) = \alpha/\beta,$$

$$\mathrm{Var}(U) = \alpha/\beta^2,$$

$$\mathrm{Var}(U^{-1}) = \beta/(\alpha-1),\ \alpha > 1$$

and

$$\mathrm{Var}(U^{-1}) = \beta^2/(\alpha-1)^2(\alpha-2),\ \alpha > 2.$$

c. Lognormal distribution

A real random variable U is said to have a lognormal distribution with parameters μ and σ^2 if the density of U is

$$f(u|\mu,\sigma^2) = \frac{1}{u\sigma(2\pi)^{\frac{1}{2}}} \exp\left\{ -\frac{\{\log_e(u)-\mu\}^2}{2\sigma^2} \right\}, \qquad 0 < u < \infty$$

where $\mu \geq 0$, and $\sigma > 0$. It can be shown that

$$E(U) = \exp(\mu)\ \exp(\sigma^2/2),$$

and

$$\mathrm{Var}(U) = \exp(2\mu)\exp(\sigma^2)\{\exp(\sigma^2)-1\}.$$

d. The univariate Normal

Let U be a real random variable with probability density function

$$f(u|\mu,\tau) = \frac{\tau^{\frac{1}{2}}}{\sqrt{2\pi}} \exp\left\{ -\frac{\tau}{2} (u-\mu)^2 \right\}, \qquad u \in \mathbb{R}$$

thus U is said to have a Normal distribution with mean $\mu \in \mathbb{R}$ and precision $\tau > 0$ (variance τ^{-1}) and this relationship is denoted by $U \sim N(\mu,\tau^{-1})$.

e. The Normal-Gamma

Let U be a real random variable and V a positive random variable, then U and V are said to have a Normal-gamma distribution if the density of U and V is

$$f(u,v|\mu,\tau,\alpha,\beta) \propto v^{1/2} \exp\{-\frac{\tau v}{2}(u-\mu)^2\}\, v^{\alpha-1}\, e^{-v\beta}, \quad u \in \mathbb{R},\ v > 0,$$

where $\mu \in \mathbb{R}$, $\tau > 0$, $\alpha > 0$, and $\beta > 0$.

This is a four-parameter density and is a member of the class which is conjugate to the two-parameter normal family. Thus, the prior and predictive analysis of a two-parameters Normal population depends on the Normal-gamma distribution.

f. The univariate Student-t distribution

A real random variable U is said to have a t distribution with parameters ν and μ if the density of U is

$$f(u|\nu,\mu) \propto [1 + (u-\mu)^2/\nu]^{-(\nu+1)/2}, \quad u \in \mathbb{R}$$

where $\mu \in \mathbb{R}$, and $\nu > 0$. It can be shown that

$$E(U) = \mu, \qquad \nu > 1$$

and

$$Var(U) = \nu/(\nu-2), \qquad \nu > 2.$$

g. Mixture of two Normals

A real random variable U is said to have a mixture of two Normal distributions with parameters p, μ_1, μ_2, σ_1 and σ_2 if the density of U is

$$f(u|\mu_1,\mu_2,\sigma_1,\sigma_2) = \frac{p}{\sqrt{(2\pi)}\ \sigma_1} \exp\left\{ - \frac{(u-\mu_1)^2}{2\sigma_1^2} \right\} +$$

$$\frac{(1-p)}{\sqrt{(2\pi)}\ \sigma_2} \exp\left\{ - \frac{(u-\mu_2)^2}{2\sigma_2^2} \right\}, \qquad u \in \mathbb{R}$$

where μ_1, $\mu_2 \in \mathbb{R}$, and σ_1, $\sigma_2 > 0$. It can be shown that

$$E(U) = p\mu_1 + (1-p)\mu_2,$$

and

$$\mathrm{Var}(U) = p(\sigma_1^2+\mu_1^2) + [(1-p)(\sigma_2^2+\mu_2^2)] - [p\mu_1+(1-p)\mu_2]^2$$

Proof of (i):

$$E(U) = \int u\, p(u)\, du$$

$$= \int_{-\infty}^{\infty} u \left[\frac{p}{(2\pi)^{\frac{1}{2}}\sigma_1} \exp\left\{ \frac{-(u-\mu_1)^2}{2\sigma_1^2} \right\} + \frac{(1-p)}{(2\pi)^{\frac{1}{2}}\sigma_2} \exp\left\{ \frac{-(u-\mu_2)^2}{2\sigma_2^2} \right\} \right] du$$

$$= \int_{-\infty}^{\infty} p \left[\frac{u}{(2\pi)^{\frac{1}{2}}\sigma_1} \exp\left\{ \frac{-(u-\mu_1)^2}{2\sigma_1^2} \right\} \right] du +$$

$$\int_{-\infty}^{\infty} (1-p) \left[\frac{u}{(2\pi)^{\frac{1}{2}}\sigma_2} \exp\left\{ \frac{-(u-\mu_2)^2}{2\sigma_2^2} \right\} \right] du$$

$$= p\mu_1 + (1-p)\mu_2 \ (=\mu)$$

Proof of: (ii)

$$\mathrm{Var}(U) = E(u^2) - [E(u)]^2$$

$$E(u^2) = \int u^2\, p(u)\, du$$

$$= \int_{-\infty}^{\infty} u^2 \left[\frac{p}{(2\pi)^{\frac{1}{2}}\sigma_1} \exp\left\{ \frac{-(u-\mu_1)^2}{2\sigma_1^2} \right\} + \frac{(1-p)}{(2\pi)^{\frac{1}{2}}\sigma_2} \exp\left\{ \frac{-(u-\mu_2)^2}{2\sigma_2^2} \right\} \right] du$$

$$= \int_{-\infty}^{\infty} p \left[\frac{u^2}{(2\pi)^{\frac{1}{2}}\sigma_1} \exp\left\{ \frac{-(u-\mu_1)^2}{2\sigma_1^2} \right\} \right] du +$$

$$\int_{-\infty}^{\infty} (1-p) \left[\frac{u^2}{(2\pi)^{\frac{1}{2}}\sigma_2} \exp\left\{ \frac{-(u-\mu_2)^2}{2\sigma_2^2} \right\} \right] du$$

Let $v_1 = (u-\mu_1)/\sigma_1$ and $v_2 = (u-\mu_2)/\sigma_2$, then $u = v_1\sigma_1 + \mu_1$ and $u = v_2\sigma_2 + \mu_2$ and $du = \sigma_1 dv_1$ and $\sigma_2 dv_2$, respectively.

$$= p\left[\int_{-\infty}^{\infty} \frac{(v_1\sigma_1+\mu_1)^2}{(2\pi)^{\frac{1}{2}}\sigma_1} \exp\left\{-\frac{v_1^2}{2}\right\} \sigma_1 dv_1\right] +$$

$$(1-p)\left[\int_{-\infty}^{\infty} \frac{(v_2\sigma_2+\mu_2)^2}{(2\pi)^{\frac{1}{2}}\sigma_2} \exp\left\{-\frac{v_2^2}{2}\right\} \sigma_2 dv_2\right]$$

$$= p\left[\int_{-\infty}^{\infty} \frac{(v_1\sigma_1)^2+2v_1\mu_1\sigma_1+\mu_1^2}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{v_1^2}{2}\right\} dv_1\right] +$$

$$(1-p)\left[\int_{-\infty}^{\infty} \frac{(v_2\sigma_2)^2+2v_2\mu_2\sigma_2+\mu_2^2}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{v_2^2}{2}\right\} dv_2\right]$$

$$= p\left[\int_{-\infty}^{\infty} \frac{(v_1\sigma_1)^2}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{v_1^2}{2}\right\} dv_1 + \mu_1^2\right] +$$

$$(1-p)\left[\int_{-\infty}^{\infty} \frac{(v_2\sigma_2)^2}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{v_2^2}{2}\right\} dv_2 + \mu_2^2\right]$$

$$= p\left[\sigma_1^2 \int_{-\infty}^{\infty} \frac{v_1^2}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{v_1^2}{2}\right\} dv_1 + \mu_1^2\right] +$$

$$(1-p)\left[\sigma_2^2 \int_{-\infty}^{\infty} \frac{v_2^2}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{v_2^2}{2}\right\} dv_2 + \mu_2^2\right]$$

$$= p\,[\,\sigma_1^2 + \mu_1^2\,] + (1-p)\,[\,\sigma_2^2 + \mu_2^2\,]$$

Thus $\mathrm{Var}(U) = p\,[\,\sigma_1^2 + \mu_1^2\,] + (1-p)\,[\,\sigma_2^2 + \mu_2^2\,] - [p\mu_1 + (1-p)\mu_2]^2$

h. The multivariate Normal distribution

We say that a p-dimensional random variable **U** follows the multivariate Normal distribution if its joint p.d.f. is of the form

$$f(u)= \frac{1}{(2\pi)^{p/2}\ |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(u-\mu)'\Sigma^{-1}(u-\mu) \right\}$$

where Σ is any $(p\times p)$ symmetric positive definite matrix. Moreover, if $U_1,\ldots,U_p$ are independent random variables where $U_i \sim N(\mu_i,\sigma_i^2)$, then their joint p.d.f. is simply the product of the appropriate (marginal) density functions, so that

$$f(u_1,\ldots,u_p) = \frac{1}{(2\pi)^{p/2}\prod_{i=1}^{p}\sigma_i} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{p}\left[\frac{u_i-\mu_i}{\sigma_i} \right]^2 \right\}$$

In this case $U' = [U_1,\ldots,U_p]$ has mean $\mu' = [\mu_1,\ldots,\mu_p]$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & 0 & \ldots & 0 \\ \vdots & & & & \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & \sigma_p^2 \end{bmatrix}.$$

But of course the components of U do not generally need to be independent and so Σ does not have to be diagonal, provide that it is symmetric and positive definite. The requirement that Σ be positive definite van be thought of as the multivariate equivalent of the condition that $\sigma^2 > 0$ in the univariate case. It is clear that $f(u) \geq 0$ for every u, and it is also straightforward, thought algebraically tedious, to check that $\int_u f(u)\, du_1\ldots du_p = 1$ for every μ and for every Σ which is symmetric and positive definite. After some algebra, it is also possible to show that $E(U) = \mu$ and that Σ is the covariance matric for U. Thus the parameters μ and Σ have an immediate interpretation, and we write $U \sim N_p(\mu,\Sigma)$, where p denotes the dimension of U, μ denotes the mean vector and Σ denotes the

covariance matrix. The definition of the multivariate Normal distribution via the equation above also requires the covariance matrix to be non-singular so that Σ^{-1} exists.

APPENDIX 2

RESULTS OF THE CONVOLUTION OF NORMAL DENSITY FUNCTIONS

These results are used in Chapter 3.

Define the function $N(.,.)$ by

$$N(a,b^2) = (b\sqrt{2\pi})^{-1} \exp\{-a^2/(2b^2)\}$$

the following identities hold:

(i) $\int N(u-a_1,b_1^2)\times N(u-a_2,b_2^2)\,du$

$= N(a_1-a_2,b_1^2+b_2^2)$

and (ii) $\int N(u-a_1,b_1^2)\times N(u-a_2,b_2^2)\times N(u-a_3,b_3^2)\,du$

$= N(a_1-a_2,b_1^2+b_2^2) \times N\{w-a_3,(b_1^2+b_2^2)/(b_1^2b_2^2+b_2^2b_3^2+b_1^2b_3^2)\}$

where $w = \left[\frac{1}{b_1^2}+\frac{1}{b_2^2}\right]^{-1}\left[\frac{a_1}{b_1^2}+\frac{a_2}{b_2^2}\right]$,

or

$$= (2\pi)^{-\frac{1}{2}}\, N\{(a_1-a_2)^2b_3^2+(a_2-a_3)^2b_1^2+(a_1-a_3)^2b_2^2, b_1^2b_2^2+b_2^2b_3^2+b_1^2b_3^2)\}.$$

Proof of (i):

$$N(u-a_1,b_1^2) \times N(u-a_2,b_2^2)$$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{(u-a_1)^2}{2b_1^2}\right\} \exp\left\{-\frac{(u-a_2)^2}{2b_2^2}\right\}$$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\frac{(u-a_1)^2}{b_1^2}+\frac{(u-a_2)^2}{b_2^2}\right]\right\}$$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\frac{u^2-2ua_1+a_1^2}{b_1^2}+\frac{u^2-2ua_2+a_2^2}{b_2^2}\right]\right\}$$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\left[\frac{1}{b_1^2}+\frac{1}{b_2^2}\right]u^2 - 2u\left[\frac{a_1}{b_1^2}+\frac{a_2}{b_2^2}\right] + \frac{a_1^2}{b_1^2}+\frac{a_2^2}{b_2^2}\right]\right\}$$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\left[\frac{1}{b_1^2}+\frac{1}{b_2^2}\right]u^2 - 2u\left[\frac{a_1}{b_1^2}+\frac{a_2}{b_2^2}\right]\right]\right\}$$

$$\exp\left\{-\frac{1}{2}\left[\frac{a_1^2}{b_1^2}+\frac{a_2^2}{b_2^2}\right]\right\}$$

Let $w = \left[\frac{1}{b_1^2}+\frac{1}{b_2^2}\right]^{-1}\left[\frac{a_1}{b_1^2}+\frac{a_2}{b_2^2}\right]$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2}+\frac{1}{b_2^2}\right](u-w)^2\right\} \times$$

$$\exp\left\{-\frac{1}{2}\left[\left[\frac{a_1^2}{b_1^2}+\frac{a_2^2}{b_2^2}\right] - \left[\frac{1}{b_1^2}+\frac{1}{b_2^2}\right]^{-1}\left[\frac{a_1}{b_1^2}+\frac{a_2}{b_2^2}\right]^2\right]\right\}$$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2}+\frac{1}{b_2^2}\right](u-w)^2\right\} \times$$

$$\exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2}+\frac{1}{b_2^2}\right]^{-1}\left[\frac{a_1^2+a_2^2-2a_1a_2}{b_1^2b_2^2}\right]\right\}$$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2}+\frac{1}{b_2^2}\right](u-w)^2\right\} \times$$

$$\exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2+b_2^2}\right](a_1-a_2)^2\right\}$$

$$= \frac{\sqrt{(b_1^2+b_2^2)}}{(2\pi)^{\frac{1}{2}}b_1 b_2} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2}+\frac{1}{b_2^2}\right](u-w)^2\right\} \times$$

$$\frac{1}{(2\pi)^{\frac{1}{2}}\sqrt{(b_1^2+b_2^2)}} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2+b_2^2}\right](a_1-a_2)^2\right\}$$

Hence (i) follows.

<u>Proof of (ii):</u>

From (i) above, $N(u-a_1,b_1^2) \times N(u-a_2,b_2^2) \times N(u-a_3,b_3^2)$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2+b_2^2}\right](a_1 - a_2)^2\right\} \times$$

$$\frac{1}{(2\pi)^{1/2} b_3} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2} + \frac{1}{b_2^2}\right](u - w)^2\right\} \times \exp\left\{-\frac{(u-a_3)^2}{2b_3^2}\right\}$$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2+b_2^2}\right](a_1 - a_2)^2\right\} \times$$

$$\frac{1}{(2\pi)^{1/2} b_3} \exp\left\{-\frac{1}{2}\left[\left[\frac{1}{b_1^2} + \frac{1}{b_2^2}\right](u^2-2uw+w^2) + \frac{u^2-2ua_3+a_3^2}{2b_3^2}\right]\right\}$$

Let $b = \left[\frac{1}{b_1^2} + \frac{1}{b_2^2}\right]$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2+b_2^2}\right](a_1 - a_2)^2\right\} \times$$

$$\frac{1}{(2\pi)^{1/2} b_3} \exp\left\{-\frac{1}{2}\left[\left[\frac{1}{b_1^2} + \frac{1}{b_2^2} + \frac{1}{b_3^2}\right]u^2 - 2u\left[bw + \frac{a_3}{b_3^2}\right]\right]\right\} \times$$

$$\exp\left\{-\frac{1}{2}\left[bw^2 + \frac{a_3^2}{b_3^2}\right]\right\}$$

Let $v = \left[b + \frac{1}{b_3^2}\right]^{-1}\left[bw + \frac{a_3}{b_3^2}\right]$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2+b_2^2}\right](a_1 - a_2)^2\right\} \times$$

$$\frac{1}{(2\pi)^{1/2} b_3} \exp\left\{-\frac{1}{2}\left[b + \frac{1}{b_3^2}\right][u - v]^2\right\} \times$$

$$\exp\left\{-\frac{1}{2}\left[\left[bw^2 + \frac{a_3^2}{b_3^2}\right] - \left[b + \frac{1}{b_3^2}\right]^{-1}\left[bw + \frac{a_3}{b_3^2}\right]^2\right]\right\}$$

$$= \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2+b_2^2}\right](a_1 - a_2)^2\right\} \times$$

$$\frac{1}{(2\pi)^{1/2} b_3} \exp\left\{-\frac{1}{2}\left[b + \frac{1}{b_3^2}\right][u - v]^2\right\} \times$$

$$\exp\left\{-\frac{1}{2}\left[b+\frac{1}{b_3^2}\right]^{-1}\left[\frac{b(w-a_3)^2}{b_3^2}\right]\right\}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}\surd(b_1^2+b_2^2)} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_1^2+b_2^2}\right](a_1-a_2)^2\right\} \times$$

$$\frac{1}{(2\pi)^{\frac{1}{2}}\surd(b_3^{-2}+b)} \exp\left\{-\frac{1}{2}\left[b+\frac{1}{b_3^2}\right][u \pm v]^2\right\} \times$$

$$\frac{1}{(2\pi)^{\frac{1}{2}}\surd(b_3^2+b^{-1})} \exp\left\{-\frac{1}{2}\left[\frac{1}{b_3^2+b^{-1}}\right](w-a_3)^2\right\}$$

Note $\surd(b_3^{-2}+b) = \dfrac{\surd(b_2^2b_3^2 + b_1^2b_3^2 + b_1^2b_2^2)}{b_1b_2b_3}$

and

$$\surd(b_3{}^2+b^{-1}) = \frac{\surd(b_1^2+b_2^2)}{\surd(b_2^2b_3^2 + b_1^2b_3^2 + b_1^2b_2^2)}$$

Hence (ii) follows.

APPENDIX 3

NOTES ON CONJUGATE PRIOR DENSITY FOR σ^2

The conjugate prior density for the parameter σ^2 is

$$p(\sigma^2) = \left[\frac{\nu_0\sigma_0^2}{\sigma^2}\right]^{\nu_0/2-1}\left[\frac{\nu_0\sigma_0^2}{\sigma^4}\right]\left[\frac{1}{2}\right]^{\nu_0/2}\exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\}\left[\frac{1}{\Gamma(\nu_0/2)}\right].$$

Let $\beta = \nu_0\sigma_0^2$ and $\alpha=\nu_0$, then $p(\sigma^2)$ becomes

$$\left[\frac{\beta}{\sigma^2}\right]^{\alpha/2-1}\left[\frac{\beta}{\sigma^4}\right]\left[\frac{1}{2}\right]^{\alpha/2}\exp\left\{-\frac{\beta}{2\sigma^2}\right\}\left[\frac{1}{\Gamma(\alpha/2)}\right]$$

$$= \left[\frac{\beta}{2}\right]^{\alpha/2}\left[\frac{1}{\sigma^2}\right]^{\alpha/2+1}\exp\left\{-\frac{\beta}{2\sigma^2}\right\}\left[\frac{1}{\Gamma(\alpha/2)}\right]$$

Then let $\tau = \sigma^{-2}$, $d\tau=-(\sigma^2)^{-2}d\sigma^2$ which implies $d\sigma^2=\tau^{-2}d\tau$. Thus

$$p(\tau) = (\beta/2)^{\alpha/2}\,\tau^{\alpha/2-1}\exp\{-\beta\tau/2\}\,[\Gamma(\alpha/2)]^{-1}$$

which is Gamma distribution with parameters $\alpha/2,\beta/2$.

The following properties hold:

i) $E(\sigma^2) = \beta/(\alpha-2)$ since $E(\beta/\sigma^2) = \alpha$ and $\text{var}(\beta/\sigma^2) = 2\alpha$.

ii) $\text{Var}(\sigma^2) = 2\beta^2/[(\alpha-2)^2(\alpha-4)]$

iii) $E(\sigma^4) = \beta^2/[(\alpha-2)(\alpha-4)]$

If α is large, $E(\sigma^2)$ and $\text{var}(\sigma^2)$ are approximately β/α and $2\beta^2/\alpha^3$, respectively. Hence the two numbers at our disposal, β/α and α enable us to alter the mean and variance of the prior distribution: β/α is approximately the mean. (Large values of α correspond to rather precise knowledge of the value of λ prior to the experiment.

The two quantities β/α and α therefore allow considerable variation in the choice of prior distribution within this class of densities. Note that prior distribution of σ^2, like x^2, tends to normality as $\alpha \to \infty$.

APPENDIX 4

MEAN AND VARIANCE OF M WHEN ITS DENSITY TAKES A KERNEL FORM

If a density function of a random variable M takes a kernel form density, namely

$$f(\mu) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}} s\lambda} \exp\left\{ - \frac{1}{2s^2\lambda^2} (\mu-\bar{z}_i)^2 \right\}, \qquad \mu \in \mathbb{R}$$

where $s^2 = (n-1)^{-1} \sum_{i=1}^{n} (\bar{z}_i-\bar{z}_.)^2$

then (i) $E(\mu) = \bar{z}_.$ and (ii) $\mathrm{Var}(\mu) = s^2(\lambda^2+1)$.

Proof of (i):

$$E(\mu) = \int \mu \, f(\mu) \, d\mu$$

$$= \int_{-\infty}^{\infty} \mu \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{(2\pi)}\, s\lambda} \exp\left\{ - \frac{1}{2s^2\lambda^2} (\mu-\bar{z}_i)^2 \right\} d\mu$$

Without lost of generality, we interchange the integral and summation sign

$$= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{\mu}{\sqrt{(2\pi)}\, s\lambda} \exp\left\{ - \frac{1}{2s^2\lambda^2} (\mu-\bar{z}_i)^2 \right\} d\mu$$

$$= \frac{1}{n} \sum_{i=1}^{n} \bar{z}_i \quad (= \bar{z}_.)$$

Proof of (ii):

$$\mathrm{Var}(\mu) = E(\mu^2) - [E(\mu)]^2$$

First we evaluate $E(\mu^2) = \int \mu^2 \, f(\mu) \, d\mu$

$$= \int_{-\infty}^{\infty} \mu^2 \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}} s\lambda} \exp \left\{ - \frac{(\mu-\bar{z}_i)^2}{2s^2\lambda^2} \right\} d\mu \right]$$

W.l.o.g. we interchange the integral and summation sign,

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\int_{-\infty}^{\infty} \frac{\mu^2}{(2\pi)^{\frac{1}{2}} s\lambda} \exp \left\{ - \frac{(\mu-\bar{z}_i)^2}{2s^2\lambda^2} \right\} . d\mu \right]$$

Let $u_i = (\mu-\bar{z}_i)/s\lambda$ then $\mu = u_i s\lambda + \bar{z}_i$ and $d\mu = s\lambda du_i$.

$$= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{(u_i s\lambda+\bar{z}_i)^2}{(2\pi)^{\frac{1}{2}} s\lambda} \exp \left\{ - \frac{u_i^2}{2} \right\} s\lambda du_i$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{[(u_i s\lambda)^2+2u_i s\lambda+\bar{z}_i^2]}{(2\pi)^{\frac{1}{2}}} \exp \left\{ - \frac{u_i^2}{2} \right\} du_i$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\int_{-\infty}^{\infty} \frac{(u_i s\lambda)^2}{(2\pi)^{\frac{1}{2}}} \exp \left\{ - \frac{u_i^2}{2} \right\} du_i + \bar{z}_i^2 \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[(s\lambda)^2 \int_{-\infty}^{\infty} \frac{u_i^2}{(2\pi)^{\frac{1}{2}}} \exp \left\{ - \frac{u_i^2}{2} \right\} du_i + \bar{z}_i^2 \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} [(s\lambda)^2 + \bar{z}_i^2]$$

$$= (s\lambda)^2 + n^{-1} \sum_{i=1}^{n} \bar{z}_i^2$$

$$\text{Var}(\mu) = E(\mu^2) - [E(\mu)]^2$$

$$= (s\lambda)^2 + n^{-1} \sum_{i=1}^{n} \bar{z}_i^2 - \bar{z}^2$$

$$= (s\lambda)^2 + n^{-1} \left[\sum_{i=1}^{n} \bar{z}_i^2 - n\bar{z}^2 \right]$$

$$= (s\lambda)^2 + n^{-1}\left[\sum_{i=1}^{n}(\bar{z}_i - \bar{z}_.)^2\right]$$

$$= (s\lambda)^2 + n^{-1}(n-1)s^2$$

$$= s^2\,[\,\lambda^2 + n^{-1}(n-1)\,]$$

$$\rightarrow s^2\,[\,\lambda^2 + 1\,] \quad \text{as } n \rightarrow \infty$$

So if $\mathrm{Var}(\mu) = \sigma_a^2$, then $\lambda^2 = \sigma_a^2/s^2 - 1$ subject to $\sigma_a^2 > s^2$.

APPENDIX 5

NOTES ON DERIVING THE PREDICTIVE AND MARGINAL DISTRIBUTIONS FOR A PARTICULAR MIXTURE DATA

This appendix provides details of the formulae for the numerator of the Bayes' factor given in Chapter 5. First let T(y) and T(x) denote the number of zeros in the Y and X data and p the probability of X or Y is being zero.

A5.1 Assumptions:

ECA and Kernel models - Positive non-zero values of X or of Y is Normally distributed. So, the probability density of X, for instance, may be represented as

$$f(X|.) = \begin{cases} p & x=0 \\ (1-p)\, g(x|\mu,\sigma) & x\neq 0. \end{cases}$$

where g is assumed to be Normal density function with parameters μ and σ^2. Similarly, two-fold definition is assumed for f(Y|.).

A5.1.1 Priors for p and the unknonw true popluation mean μ:

Assuming the prior densities are independent of each other, then

(i) ***ECA model*** with informative prior for μ — $f(p,\mu) = f(p) \times f(\mu)$

where $f(p) = [Be(a,b)]^{-1}\ p^{a-1}\ (1-p)^{b-1}$, $f(\mu) = g\{\mu|\bar{\mu},\bar{\sigma}^2\}$ and g is a Normal density function with parameters $\bar{\mu}$ and $\bar{\sigma}^2$.

(ii) ***Kernel model*** -- $f(p,\mu) = f(p) \times f(\mu)$

where f(p) as in the ECA model and $f(\mu) = k(\mu|\bar{z}_i^*\text{'s},.)$, kernel density

function with smoothing parameter λ, $\bar{z}_i^*$'s are sample group means of a training data set Z.

A5.2 Predictive distribution (with informative prior for μ and assumed known σ^2

A5.2.1 Single hair problem (see Chapter 5)

(a) Both x and y are zero

$$\Pr(T(y)=1|T(x)=1) = \frac{\int \Pr(T(y)=1|p)\ \Pr(T(x)=1|p)\ f(p)\ dp}{\int \Pr(T(x)=1|p)\ f(p)\ dp}$$

$$= \frac{\int p\ p\ [Be(a,b)]^{-1}\ p^{a-1}\ (1-p)^{b-1}\ dp}{\int p\ [Be(a,b)]^{-1}\ p^{a-1}\ (1-p)^{b-1}\ dp}$$

$$= \frac{\int [Be(a,b)]^{-1}\ p^{(a+2)-1}\ (1-p)^{b-1}\ dp}{\int [Be(a,b)]^{-1}\ p^{(a+1)-1}\ (1-p)^{b-1}\ dp}$$

$$= \frac{Be(a+2,b)}{Be(a+1,b)}$$

$$= \frac{\Gamma(a+2)\ \Gamma(b)}{\Gamma(a+b+2)} \times \frac{\Gamma(a+b+1)}{\Gamma(a+1)\ \Gamma(b)}$$

$$= \frac{(a+1)!\ (a+b)!}{(a+b+1)!\ a!}$$

$$= \frac{a+1}{a+b+1}$$

(b) y is zero but x is not

$$\Pr(T(y)=1|x) = \frac{\int \Pr(T(y)=1|p)\ f(x|\mu)\ f(\mu)\ d\mu}{\int f(x|\mu)\ f(\mu)\ d\mu}$$

$$\propto \iint p(1-p)g\{x|\mu,\sigma^2\}[Be(a,b)]^{-1}\ p^{a-1}(1-p)^{b-1}g\{\mu|\bar{\mu},\bar{\sigma}^2\}\ dpd\mu$$

then,

$$\Pr(T(y)=1|x) \propto \int_0^1 \frac{p^{(a+1)-1}(1-p)^{(b+1)-1}}{Be(a,b)}\,dp \times \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \times$$

$$\frac{1}{(2\pi)^{\frac{1}{2}}\bar{\sigma}} \exp\left\{-\frac{(\mu-\bar{\mu})^2}{2\bar{\sigma}^2}\right\} d\mu$$

(by the results in Appendix 1)

$$= \frac{Be(a+1,b+1)}{Be(a,b)} \times \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2+\bar{\sigma}^2)^{\frac{1}{2}}} \exp\left\{-\frac{(x-\bar{\mu})^2}{2(\sigma^2+\bar{\sigma}^2)}\right\}$$

$$= \frac{ab}{(a+b+1)(a+b)} \times \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2+\bar{\sigma}^2)^{\frac{1}{2}}} \exp\left\{-\frac{(x-\bar{\mu})^2}{2(\sigma^2+\bar{\sigma}^2)}\right\}$$

To obtain the normalised constant we evaluate this quantity,

$$\int f(x|\mu)\, f(\mu)\, d\mu$$

$$= \iint (1-p)g\{x|\mu,\sigma^2\}[Be(a,b)]^{-1}\, p^{a-1}(1-p)^{b-1}g\{\mu|\bar{\mu},\bar{\sigma}^2\}\, dp d\mu$$

$$f(x) = \frac{Be(a,b+1)}{Be(a,b)} \times \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2+\bar{\sigma}^2)^{\frac{1}{2}}} \exp\left\{-\frac{(x-\bar{\mu})^2}{2(\sigma^2+\bar{\sigma}^2)}\right\}$$

$$= \frac{b}{(a+b)} \times \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2+\bar{\sigma}^2)^{\frac{1}{2}}} \exp\left\{-\frac{(x-\bar{\mu})^2}{2(\sigma^2+\bar{\sigma}^2)}\right\}$$

Thus,

$$\Pr(T(y)=1|x) = a/(a+b+1)$$

(c) y is not zero but x is

$$f(y|T(x)=1) = \frac{\int f(y|\mu)\, \Pr(T(x)=1|p)\, f(\mu)\, d\mu}{\int \Pr(T(x)=1|p)\, f(p)\, dp}$$

$$\propto \int_0^1 \int_{-\infty}^{\infty} \frac{(1-p)\ g\{y-\mu,\sigma^2\}\ p\ p^{a-1}\ (1-p)^{b-1}\ g\{\mu|\bar{\mu},\bar{\sigma}^2\}}{Be(a,b)}\ d\mu dp$$

Again, using the result in Appendix 1,

$$f(y|T(x)=1 \propto \frac{Be(a+1,b+1)}{Be(a,b)} \times \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2+\bar{\sigma}^2)^{\frac{1}{2}}} \exp\left\{-\frac{(y-\bar{\mu})^2}{2(\sigma^2+\bar{\sigma}^2)}\right\}$$

$$\equiv \frac{ab}{(a+b+1)(a+b)} \times \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2+\bar{\sigma}^2)^{\frac{1}{2}}} \exp\left\{-\frac{(y-\bar{\mu})^2}{2(\sigma^2+\bar{\sigma}^2)}\right\}$$

Now, $\int Pr(T(x)=1|p)\ f(p)\ dp = \int p\ [Be(a,b)]^{-1}\ p^{a-1}\ (1-p)^{b-1}\ dp$

$$= \frac{Be(a+1,b)}{Be(a,b)} = \frac{a}{a+b}$$

Hence

$$f(y|T(x)=1) = \frac{b}{(a+b+1)} \times \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2+\bar{\sigma}^2)^{\frac{1}{2}}} \exp\left\{-\frac{(y-\bar{\mu})^2}{2(\sigma^2+\bar{\sigma}^2)}\right\}.$$

(d) X and y both are not zero

$$f(y|x) = \frac{\int f(y|\mu)\ f(x|\mu)\ f(\mu)\ d\mu}{\int f(x|\mu)\ f(\mu)\ d\mu}$$

$$\propto \iint \frac{(1-p)g\{y|\mu,\sigma^2\}(1-p)g\{x|\mu,\sigma^2\}p^{a-1}(1-p)^{b-1}g\{\mu|\bar{\mu},\bar{\sigma}^2\}}{Be(a,b)}\ d\mu dp$$

$$\propto \int_0^1 \frac{p^{a-1}(1-p)^{(b+2)-1}}{Be(a,b)}\ dp \times \int_{\infty}^{-\infty} g\{y|\mu,\sigma^2\}g\{x|\mu,\sigma^2\}g\{\mu|\bar{\mu},\bar{\sigma}^2\}\ d\mu$$

$$\equiv \frac{Be(a,b+2)}{Be(a,b)}\ dp \times \int_{\infty}^{-\infty} g\{y|\mu,\sigma^2\}g\{x|\mu,\sigma^2\}g\{(\mu|\bar{\mu},\bar{\sigma}^2\}\ d\mu$$

Using the results in Appendix 1,

$$f(y|x) \propto \frac{b(b+1)}{(a+b+1)(a+b)} \times \frac{1}{(2\pi)^{\frac{1}{2}}\sqrt{2}\sigma} \exp\left\{-\frac{(x-y)^2}{4\sigma^2}\right\} \times$$

$$\frac{1}{(2\pi)^{\frac{1}{2}}\{(\sigma^2/2)+\bar{\sigma}^2\}^{\frac{1}{2}}} \exp\left\{-\frac{(w-\bar{\mu})^2}{2\{(\sigma^2/2)+\bar{\sigma}^2\}}\right\}$$

To obtain the normalised constant, we evaluate $\int f(x|\mu)\, f(\mu)\, d\mu$ which is given as in (b) above, namely

$$\frac{b}{a+b} \times \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2+\bar{\sigma}^2)^{\frac{1}{2}}} \exp\left\{-\frac{(x-\bar{\mu})^2}{2(\sigma^2+\bar{\sigma}^2)}\right\}$$

Thus

$$f(y|x) = \frac{b+1}{a+b+1} \times \frac{1}{(2\pi)^{\frac{1}{2}}\sqrt{2}\sigma} \exp\left\{-\frac{(x-y)^2}{4\sigma^2}\right\} \times$$

$$\frac{1}{(2\pi)^{\frac{1}{2}}\{(\sigma^2/2)+\bar{\sigma}^2\}^{\frac{1}{2}}} \exp\left\{-\frac{(w-\bar{\mu})^2}{2\{(\sigma^2/2)+\bar{\sigma}^2\}}\right\} \times$$

$$\left[\frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2+\bar{\sigma}^2)^{\frac{1}{2}}} \exp\left\{-\frac{(x-\bar{\mu})^2}{2(\sigma^2+\bar{\sigma}^2)}\right\}\right].$$

A5.2.2 Case where m > 1 and r = 1

Controlled data X consists of t(x) zero values and (m-t(x)) other values $x_1,\ldots,x_{m-t(x)}$. Let μ and σ^2 be the mean and variance respectively of the positive non-zero population of X and let $\bar{x}_{[m-t(x)]}$ be a sufficient unbiased estimator of μ for the m-t(x) observations. Then, assuming the variance σ^2 is known $f(X|p,\mu,\sigma^2)$ may be factorised as follow:

$$f(X|p,\mu,\sigma^2) = f(T(x),\bar{X}|p,\mu,\sigma^2) = f(\bar{X}|T(x),\mu,\sigma^2) \times f(T(x)|p).$$

This is assuming $\bar{X}$ and p are conditional independent, i.e.

$$f(\bar{X}|T(x),p,\mu,\sigma^2) = f(\bar{X}|T(x),\mu,\sigma^2).$$

Now, the probability density function of $\bar{X}$ conditional on T(x)=t(x) is taken to be

$$f(\bar{x}_{[m-t(x)]}|\mu,\sigma^2,t(x)),$$

and f(t(x)|p) is taken to be

$$\begin{bmatrix} m \\ t(x) \end{bmatrix} p^{t(x)} (1-p)^{m-t(x)}.$$

Hence, assuming the non-zero positive observations of X are Normally distributed, the likelihood function of the data x is

$$\begin{bmatrix} m \\ t(x) \end{bmatrix} p^{t(x)} (1-p)^{m-t(x)} g(\bar{x}_{[m-t(x)]}|\mu,\sigma^2)$$

where g is a Normal density function containing the non-zero positive values in the form $\bar{x}_{[m-t(x)]}$.

To evaluate the predictive distribution of Y given X in the m >1 case, we consider the two possibilities i.e (a) y is zero and (b) y is not zero.

(a) $Pr(T(y)=1|X) \propto \int Pr(T(y)=1|\underset{\sim}{\theta})\, f(X|\underset{\sim}{\theta})\, f(\underset{\sim}{\theta})\, d\underset{\sim}{\theta}$

where $\underset{\sim}{\theta} = (p,\mu)$ and $f(\underset{\sim}{\theta})$ is given in Section A5.1.

Assuming the true values of the population mean of the positive non-zeros observations has no effects on the phenomenon that Y is zero, i.e. $Pr(T(y)=1|\underset{\sim}{\theta}) = Pr(T(y)=1|p) = p$, and since

$$f(\underset{\sim}{X}|\underset{\sim}{\theta}) = \begin{bmatrix} m \\ t(x) \end{bmatrix} p^{t(x)} (1-p)^{m-t(x)} g\{\bar{x}_{[m-t(x)]}|\mu,\sigma^2_{[m-t(x)]}\}$$

then it can be shown that

$$Pr(T(y)=1|X) = \frac{t(x)+a}{m+a+b} .$$

(b) $f(y|X) \propto \int f(y|\underset{\sim}{\theta})\ f(X|\underset{\sim}{\theta})\ f(\underset{\sim}{\theta})\ d\underset{\sim}{\theta}$

From Section A5.1, the conditional density function of y, $f(y|\underset{\sim}{\theta})$ = $(1-p)\ g\{y|\mu,\sigma\}$. Whereas $f(x|\underset{\sim}{\theta})$ is given as in (a) above, then

$$f(y|X) = \frac{m-t(x)+b}{m+a+b} \times \frac{g\{\bar{x}_{[m-t(x)]}|y,\sigma_1^2\}\ g\{w|\bar{\mu},\sigma_2^2\}}{g\{\bar{x}_{[m-t(x)]}|\bar{\mu},\sigma_3^2\}}$$

where $\sigma_1^2 = \sigma^2[1+(1/(m-t(x)))]$

$\sigma_2^2 = \bar{\sigma}^2 + [\sigma^2/(1+(m-t(x)))]$

$\sigma_3^2 = \bar{\sigma}^2 + [\sigma^2/(m-t(x))]$

$w = [y + (m-t(x))\bar{x}_{(m-t(x))}]/[1+(m-t(x))]$

A5.2.3 Case where m and r are both greater than 1

Here, Y consists of r measurements and X consists of m measurements, so the predictive distribution of Y given X is as follow

$$f(Y|X) \propto \int f(Y|\underset{\sim}{\theta})\ f(X|\underset{\sim}{\theta})\ f(\underset{\sim}{\theta})\ d\underset{\sim}{\theta}$$

where $f(Y|\underset{\sim}{\theta}) = \binom{r}{t(y)} p^{t(y)}(1-p)^{r-t(y)} g\{\bar{y}_{[r-t(y)]}|\mu,\sigma^2_{[r-t(y)]}\}$, and $f(X|\underset{\sim}{\theta})$ is as in Section A5.2.2.

Thus,

$$f(Y|X) \propto \int\int \binom{r}{t(y)} p^{t(y)} (1-p)^{r-t(y)} g\{\bar{y}[r-t(y)]|\mu,\sigma^2[r-t(y)]\} \times$$
$$\binom{m}{t(x)} p^{t(x)} (1-p)^{m-t(x)} g\{(\bar{x}[m-t(x)]|\mu,\sigma^2[m-t(x)]\} \times$$
$$[Be(a,b)]^{-1} p^{a-1} (1-p)^{b-1} g\{\mu|\bar{\mu},\bar{\sigma}^2\} \, dp d\mu$$
$$\equiv \binom{r}{t(y)} \binom{m}{t(x)} [Be(a,b)]^{-1}$$
$$\int p^{t(y)+t(x)+a-1}(1-p)^{r-t(y)+m-t(x)+b-1} \, dp \times$$
$$\int g\{\bar{y}[r-t(y)]|\mu,\sigma^2[r-t(y)]\} \, g\{\bar{x}[m-t(x)]|\mu,\sigma^2[m-t(x)]\} \times$$
$$g\{\mu|\bar{\mu},\bar{\sigma}\} \, d\mu$$
$$= \binom{r}{t(y)} \binom{m}{t(x)} [Be(a,b)]^{-1} Be(a+t(y)+t(x),b+r-t(y)+m-t(x)) \times$$
$$g\{\bar{y}[r-t(y)]^{-}\bar{x}[m-t(x)]|0,[\sigma^2([m-t(x)]^{-1}+[r-t_{(y)}]^{-1})]\} \times$$
$$g\{w|\bar{\mu},[\bar{\sigma}^2+(\sigma^2/(r-t(y)+m-t(x)))]\}$$

The normalised constant is the marignal density function of X, i.e.

$$f(X) = \binom{m}{t(x)} [Be(a,b)]^{-1} Be\{a+t(x),b+m-t(x)\} \times$$
$$g\{\bar{x}[m-t(x)]|\bar{\mu},[\bar{\sigma}^2+(\sigma^2/[m-t(x)])]\}$$

Thus

$$f(Y|X) = \binom{r}{t(y)} Be[a+t(y)+t(x),b+r-t(y)+m-t(x)]$$
$$g\{w|\bar{\mu},[\bar{\sigma}^2+(\sigma^2/(r-t(y)+m-t(x)))]\} \times$$
$$g\{\bar{y}[r-t(y)]^{-}\bar{x}[m-t(x)]|0,[\sigma^2((m-t(x))^{-1}+(r-t(y))^{-1})]\} \times$$
$$[\, Be(a+t(x),b+m-t(x)) \, g\{\bar{x}[m-t(x)]|\bar{\mu},[\bar{\sigma}^2+(\sigma^2/(m-t(x)))]\} \,]^{-1}.$$

The marginal density of Y is given by

$$f(Y|\bar{C}) = \begin{bmatrix} r \\ t(y) \end{bmatrix} \frac{Be(a+t(y),b+r-t(y))}{Be(a,b)} \times$$
$$g\{\bar{y}[r-t(y)]|\bar{\mu},[\bar{\sigma}^2+(\sigma^2/(r-t(y)))]\}$$

A5.3 Evaluation of the Bayes' factor for the r > 1 and m> 1

If t(y)=r, i.e. all observations from Y are zeros, then

$$\Pr(T(y)=r|X) = \frac{Be(a+r+t(x),b+m-t(x))}{Be(a+t(x),b+m-t(x))}$$

$$= \frac{\Gamma(a+r+t(x))\Gamma(b+m-t(x))}{\Gamma(a+b+m)} \times \frac{\Gamma(a+b+m)}{\Gamma(a+t(x))\ \Gamma(b+m+t(x))}$$

$$= \frac{\Gamma(a+r+t(x))\ \Gamma(a+b+m)}{\Gamma(a+b+m+r)\ \Gamma(a+t(x))}$$

$$= \frac{(a+t(x)+r-1)!\ (a+b+m-1)!}{(a+b+m+r-1)!\ (a+t(x)-1)!}$$

$$= \frac{\prod_{s=1}^{r} [a+t(x)+(s-1)]}{\prod_{s=1}^{r} [a+b+m+(s-1)]}, \text{ and}$$

$$\Pr(T(y)=r) = [Be(a,b)]^{-1}\ Be(a+r,b)$$

$$= \frac{\Gamma(a+r)\ \Gamma(b)}{\Gamma(a+b+r)} \times \frac{\Gamma(a+b)}{\Gamma(a)\ \Gamma(b)}$$

$$= \frac{\Gamma(a+r)\ \Gamma(a+b)}{\Gamma(a+b+r)\ \Gamma(a)}$$

$$= \frac{(a+r-1)!\ (a+b-1)!}{(a+b+r-1)!\ (a-1)!}$$

$$= \frac{\prod_{s=1}^{r} [a+(s-1)]}{\prod_{s=1}^{r} [a+b+(s-1)]}.$$

Ratio of the above $\Pr(T(y)=r|X)$ and $\Pr(T(y)=r)$ gives the Bayes' factor for the special case where $t(y) = r$ when $m > 1$.

If $T(y)<r$, i.e. some of the observations of Y are zero, then the Bayes' factor can be simplified as

$$\left[\frac{Be(a+t(y)+t(x),b+r-t(y)+m-t(x))\ Be(a,b)}{Be(a+t(x),b+m-t(x))\ Be(a+t(y),b+r-t(y))} \right] \times$$

$$\left[\frac{g\{w|\bar{\mu},[\bar{\sigma}^2+(\sigma^2/(r-t(y)+m-t(x)))]\}}{g\{\bar{x}_{(m-t(x))}|\bar{\mu},[\bar{\sigma}^2+(\sigma^2/(m-t(x)))]\}} \right] \times$$

$$\left[\frac{g\{\bar{y}_{(r-t(y))}-\bar{x}_{(m-t(x))}|0,[\sigma^2((m-t(x))^{-1}+(r-t(y))^{-1})]\}}{g\{\bar{y}_{(r-t(y))}|\bar{\mu},[\bar{\sigma}^2+(\sigma^2/(r-t(y)))]\}} \right],$$

which can be simplified further as

$$= \left[\frac{\Gamma(a+t(y)+t(x))\Gamma(b+r-t(y)+m-t(x))\Gamma(a+b+m)\Gamma(a)\Gamma(b)\Gamma(a+b+r)}{\Gamma(a+b+r+m)\Gamma(a+t(x))\Gamma(b+m-t(x))\Gamma(a+b)\Gamma(a+t(y))\Gamma(b+r-t(y))} \right] \times$$

$$\left[\frac{g\{w|\bar{\mu},[\bar{\sigma}^2+(\sigma^2/(r-t(y)+m-t(x)))]\}}{g\{\bar{x}_{(m-t(x))}|\bar{\mu},[\bar{\sigma}^2+(\sigma^2/(m-t(x)))]\}} \right] \times$$

$$\left[\frac{g\{\bar{y}_{(r-t(y))}-\bar{x}_{(m-t(x))}|0,[\sigma^2((m-t(x))^{-1}+(r-t(y))^{-1})]\}}{g\{\bar{y}_{(r-t(y))}|\bar{\mu},[\bar{\sigma}^2+(\sigma^2/(r-t(y)))]\}} \right]$$

For the Kernel method, formulae can be easily obtained by substituting a kernel density form in place of the Normal density functions, g, in the formula given above.

APPENDIX 6

NOTES ACCOMPANY CHAPTER 6

6.1 Some useful results in combining quadratic forms

The following two useful lemmas for combining quadratic forms are taken from Box and Tiao (1973):

Lemma 1. Let $\mathbf{u}$, $\mathbf{a}$ and $\mathbf{b}$ be $p \times 1$ vectors, and $\mathbf{A}$ and $\mathbf{B}$ be $p \times p$ symmetric matrices such that the inverse $(\mathbf{A} + \mathbf{B})^{-1}$ exists. Then,

$$(\mathbf{u} - \mathbf{a})^T\mathbf{A}(\mathbf{u} - \mathbf{a}) + (\mathbf{u} - \mathbf{b})^T\mathbf{B}(\mathbf{u} - \mathbf{b}) = (\mathbf{u} - \mathbf{c})^T(\mathbf{A} + \mathbf{B})(\mathbf{u} - \mathbf{c}) + (\mathbf{a} - \mathbf{b})^T\mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}(\mathbf{a} - \mathbf{b})$$

where $\mathbf{c} = (\mathbf{A} + \mathbf{B})^{-1}(\mathbf{Aa} + \mathbf{Bb})$.

Note that if both $\mathbf{A}$ and $\mathbf{B}$ have inverse, then

$$\mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} = (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}.$$

If sometimes happens that we need to combine two quadratic forms for which the matrix $(\mathbf{A} + \mathbf{B})$ has no inverse. In this case, Lemma 1 may be modified as follows:

Lemma 2. Let $\mathbf{u}$, $\mathbf{a}$ and $\mathbf{b}$ be $p \times 1$ vectors, and $\mathbf{A}$ and $\mathbf{B}$ be two $p \times p$ positive symmetric matrices.. Suppose the rank of the matrix $\mathbf{A} + \mathbf{B}$ is q $(< p)$. Then, subject to the constraints $\mathbf{Gu} = \mathbf{0}$,

$$(\mathbf{u} - \mathbf{a})^T\mathbf{A}(\mathbf{u} - \mathbf{a}) + (\mathbf{u} - \mathbf{b})^T\mathbf{B}(\mathbf{u} - \mathbf{b}) = (\mathbf{u} - \mathbf{c}^*)^T(\mathbf{A} + \mathbf{B} + \mathbf{M})(\mathbf{u} - \mathbf{c}^*) + (\mathbf{a} - \mathbf{b})^T\mathbf{A}(\mathbf{A} + \mathbf{B} + \mathbf{M})^{-1}\mathbf{B}(\mathbf{a} - \mathbf{b})$$

where $\mathbf{G}$ is any $(p - q) \times p$ matrix of rank $p - q$ such that the rows of $\mathbf{G}$ are linearly independent of the rows of $\mathbf{A} + \mathbf{B}$, $\mathbf{M} = \mathbf{G}^T\mathbf{G}$ and

$$\mathbf{c}^* = (\mathbf{A} + \mathbf{B} + \mathbf{M})^{-1}(\mathbf{Aa} + \mathbf{Bb}).$$

The proof of these two Lemmas can be found in Box and Tiao (1973)

6.2 Formulae of the Bayes' factor in multivariate case using the kernel models developed in chapter 3

This appendix contains formulae for the predictive and marginal distributions discussed in Chapter 6, assuming ungrouped training data. Similar expressions are also given using the adaptive kernel method assuming the training data is grouped.

6.2.1 Ungrouped training data

The equivalent expression of $f(\underset{\sim}{\bar{Y}}|\underset{\sim}{\bar{X}},C)$ shown in (6.9) for the ungrouped training data is proportional to

$$\frac{1}{|a^2\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2a^2}(\underset{\sim}{\bar{x}}-\underset{\sim}{\bar{y}})'\Sigma^{-1}(\underset{\sim}{\bar{x}}-\underset{\sim}{\bar{y}})\right\} \times \frac{1}{|A_W|^{\frac{1}{2}}\,N} \sum_{\ell=1}^{N} \exp\left\{-\frac{1}{2}(\underset{\sim}{w}-\underset{\sim}{z}_\ell)'A_W^{-1}(\underset{\sim}{w}-\underset{\sim}{z}_\ell)\right\} \qquad (A6.1)$$

where $A_W = (\Sigma_W + S'_\lambda)$, $a^2 = (m^{-1}+r^{-1})$, $S'_\lambda = \lambda^2 S'$, $\Sigma_W = (m+r)^{-1}\Sigma$, $\underset{\sim}{w} = (m\underset{\sim}{\bar{x}} + r\underset{\sim}{\bar{y}})/(m+r)$ and S' are given by (6.4).

Similarly expressions (6.10) and (6.11) $f(\underset{\sim}{\bar{X}}|C)$ and $f(\underset{\sim}{\bar{Y}}|\bar{C})$ for the ungrouped training data is given by

$$f(\underset{\sim}{\bar{X}}|C) = \frac{1}{(2\pi)^{p/2}\,|A_X|^{\frac{1}{2}}\,N} \sum_{\ell=1}^{N} \exp\left\{-\frac{1}{2}(\underset{\sim}{\bar{x}}-\underset{\sim}{z}_\ell)'A_X^{-1}(\underset{\sim}{\bar{x}}-\underset{\sim}{z}_\ell)\right\} \qquad (A6.2)$$

where $A_X = (\Sigma_X + S'_\lambda)$, $\Sigma_X = m^{-1}\Sigma$ and S'_λ is as above, and

$$f(\underset{\sim}{\bar{Y}}|\bar{C}) = \frac{1}{(2\pi)^{p/2}\,|A_y|^{\frac{1}{2}}\,N} \sum_{\ell=1}^{N} \exp\left\{-\frac{1}{2}(\underset{\sim}{\bar{y}}-\underset{\sim}{z}_\ell)'A_y^{-1}(\underset{\sim}{\bar{y}}-\underset{\sim}{z}_\ell)\right\} \qquad (A6.3)$$

where $A_y = (\Sigma_y + S_\lambda')$, $\Sigma_y = r^{-1}\Sigma$ and S_λ' is as above, respectively.

6.2.2 The adaptive kernel method

With $f(\mu)$ estimated by (6.6), the expression of $f(\underset{\sim}{\bar{Y}}|\underset{\sim}{\bar{X}},C)$ is proportional to

$$\frac{1}{|a^2\Sigma|^{1/2}} \exp\left\{-\frac{1}{2a^2}(\underset{\sim}{\bar{x}}-\underset{\sim}{\bar{y}})'\Sigma^{-1}(\underset{\sim}{\bar{x}}-\underset{\sim}{\bar{y}})\right\} \times$$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|A_{wi}|^{1/2}} \exp\left\{-\frac{1}{2}(\underset{\sim}{w}-\underset{\sim}{\bar{z}}_i)'A_{wi}^{-1}(\underset{\sim}{w}-\underset{\sim}{\bar{z}}_i)\right\} \qquad \text{(A6.4)}$$

where $A_{wi} = (\Sigma_w + S_{\lambda i})$, $a^2 = (m^{-1}+r^{-1})$, $S_{\lambda i} = (\lambda\lambda_i)^2 S$, $\Sigma_w = (m+r)^{-1}\Sigma$ and $\underset{\sim}{w} = (m\underset{\sim}{\bar{x}} + r\underset{\sim}{\bar{y}})/(m+r)$.

Then combining (6.1) and (6.5) gives the constant factor namely,

$$f(\underset{\sim}{\bar{X}}|C) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi)^{p/2}\,|A_{xi}|^{1/2}} \exp\left\{-\frac{1}{2}(\underset{\sim}{\bar{x}}-\underset{\sim}{\bar{z}}_i)'A_{xi}^{-1}(\underset{\sim}{\bar{x}}-\underset{\sim}{\bar{z}}_i)\right\} \qquad \text{(A6.5)}$$

where $A_{xi} = (\Sigma_x + S_{\lambda i})$, $\Sigma_x = m^{-1}\Sigma$ and $S_{\lambda i}$ as above.

Similarly, the denominator of the BF, $f(\underset{\sim}{\bar{Y}}|\bar{C})$ is given by

$$f(\underset{\sim}{\bar{Y}}|\bar{C}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi)^{p/2}\,|A_{yi}|^{1/2}} \exp\left\{-\frac{1}{2}(\underset{\sim}{\bar{y}}-\underset{\sim}{\bar{z}}_i)'A_{yi}^{-1}(\underset{\sim}{\bar{y}}-\underset{\sim}{\bar{z}}_i)\right\} \qquad \text{(A6.6)}$$

where $A_{yi} = (\Sigma_y + S_{\lambda i})$, $\Sigma_y = r^{-1}\Sigma$ and $S_{\lambda i}$ is as above.

Note that the determinant of the matrices A_w, A_x and A_y all have index i and so are inside the summation in (A6.4), (A6.5) and (A6.6), respectively.

APPENDIX 7

TABLES TO ACCOMPANY CHAPTER 7

Table A7.1 EMSE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=1.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	0.3889	0.2313	0.1438
		(0.0061)	(0.0017)	(0.0007)
	GKO	0.3775	0.2284	0.1430
		(0.0056)	(0.0015)	(0.0007)
	SKM	0.4982	0.2940	0.1530
		(0.0327)	(0.0232)	(0.0044)
	GKM	0.4843	0.2937	0.1530
		(0.0282)	(0.0234)	(0.0046)
Student-t with 5 d.f.	SKO	0.3771	0.2527	0.1455
		(0.0170)	(0.0138)	(0.0022)
	GKO	0.3837	0.2577	0.1458
		(0.0208)	(0.0169)	(0.0023)
	SKM	0.4878	0.3102	0.2320
		(0.0320)	(0.0156)	(0.0193)
	GKM	0.5338	0.3733	0.2491
		(0.0465)	(0.0508)	(0.0250)
Lognormal	SKO	4.8934	4.8817	4.4661
		(0.2286)	(0.2338)	(0.2150)
	GKO	5.2442	5.1755	4.6056
		(0.2458)	(0.2549)	(0.2261)
	SKM	3.8077	3.0823	2.0484
		(0.2024)	(0.1842)	(0.2032)
	GKM	5.1126	4.3374	3.4581
		(0.3518)	(0.3855)	(0.4969)
Cauchy	SKO	1.6237	1.7200	1.9320
		(0.0938)	(0.1336)	(0.1921)
	GKO	1.7182	1.8076	2.0084
		(0.0948)	(0.1358)	(0.1917)
	SKM	1.1674	1.0963	0.7991
		(0.0510)	(0.0679)	(0.0525)
	GKM	1.9406	2.1494	2.0750
		(0.1024)	(0.1416)	(0.1973)
Bimodal	SKO	0.1747	0.1197	0.0834
		(0.0004)	(0.0004)	(0.0007)
	GKO	0.1752	0.1205	0.0839
		(0.0004)	(0.0005)	(0.0008)
	SKM	0.2145	0.1473	0.0855
		(0.0073)	(0.0101)	(0.0054)
	GKM	0.2143	0.1578	0.0856
		(0.0073)	(0.0160)	(0.0054)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.2 EMSE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=2.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	0.3788	0.2354	0.1457
		(0.0043)	(0.0031)	(0.0015)
	GKO	0.3710	0.2335	0.1452
		(0.0038)	(0.0030)	(0.0014)
	SKM	0.4424	0.2891	0.1555
		(0.0197)	(0.0260)	(0.0034)
	GKM	0.4419	0.3000	0.1554
		(0.0200)	(0.0278)	(0.0034)
Student-t with 5 d.f.	SKO	0.3630	0.2395	0.1617
		(0.0073)	(0.0117)	(0.0066)
	GKO	0.3661	0.2428	0.1635
		(0.0081)	(0.0139)	(0.0070)
	SKM	0.4728	0.3153	0.3111
		(0.0183)	(0.0183)	(0.0364)
	GKM	0.5067	0.3495	0.3520
		(0.0268)	(0.0413)	(0.0518)
Lognormal	SKO	5.2029	4.9901	4.5677
		(0.2023)	(0.2622)	(0.3286)
	GKO	5.4534	5.1482	4.7392
		(0.2113)	(0.2710)	(0.3500)
	SKM	4.2013	3.3564	2.6191
		(0.2119)	(0.2168)	(0.2614)
	GKM	4.9174	4.5378	4.6739
		(0.2975)	(0.4482)	(0.7245)
Cauchy	SKO	1.6002	1.6299	1.6875
		(0.0960)	(0.1491)	(0.1643)
	GKO	1.6661	1.6838	1.7381
		(0.0968)	(0.1499)	(0.1651)
	SKM	1.3387	1.1674	1.1352
		(0.0615)	(0.0721)	(0.0702)
	GKM	1.9377	1.9517	1.9396
		(0.1025)	(0.1470)	(0.1609)
Bimodal	SKO	0.1746	0.1211	0.0826
		(0.0003)	(0.0005)	(0.0006)
	GKO	0.1749	0.1218	0.0829
		(0.0003)	(0.0006)	(0.0006)
	SKM	0.2115	0.1366	0.0819
		(0.0092)	(0.0056)	(0.0034)
	GKM	0.2149	0.1406	0.0819
		(0.0098)	(0.0070)	(0.0034)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A 7.3 EMSE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=5.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	0.3812	0.2330	0.1437
		(0.0056)	(0.0016)	(0.0006)
	GKO	0.3753	0.2315	0.1433
		(0.0051)	(0.0015)	(0.0006)
	SKM	0.4522	0.2777	0.1599
		(0.0257)	(0.0110)	(0.0059)
	GKM	0.4518	0.2776	0.1599
		(0.0259)	(0.0110)	(0.0059)
Student-t with 5 d.f.	SKO	0.3590	0.2265	0.1487
		(0.0066)	(0.0030)	(0.0031)
	GKO	0.3628	0.2276	0.1494
		(0.0075)	(0.0032)	(0.0032)
	SKM	0.5065	0.3200	0.2715
		(0.0398)	(0.0155)	(0.0259)
	GKM	0.5118	0.3291	0.2779
		(0.0394)	(0.0182)	(0.0285)
Lognormal	SKO	5.5825	4.9636	4.3151
		(0.2520)	(0.2546)	(0.1879)
	GKO	5.7727	5.0722	4.3849
		(0.2603)	(0.2606)	(0.1928)
	SKM	4.9392	4.0720	2.8553
		(0.2773)	(0.3015)	(0.3543)
	GKM	5.6099	4.6288	3.2748
		(0.3730)	(0.3994)	(0.4457)
Cauchy	SKO	1.6823	1.9589	2.0032
		(0.1063)	(0.1261)	(0.1517)
	GKO	1.7216	1.9979	2.0424
		(0.1065)	(0.1258)	(0.1533)
	SKM	1.6409	1.5610	1.3813
		(0.0800)	(0.0776)	(0.0909)
	GKM	1.9459	2.2712	2.3218
		(0.1065)	(0.1251)	(0.1744)
Bimodal	SKO	0.1748	0.1211	0.0823
		(0.0003)	(0.0006)	(0.0005)
	GKO	0.1751	0.1218	0.0827
		(0.0003)	(0.0006)	(0.0005)
	SKM	0.2077	0.1334	0.0872
		(0.0081)	(0.0040)	(0.0040)
	GKM	0.2110	0.1334	0.0872
		(0.0087)	(0.0040)	(0.0040)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.4 EMSE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=10.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	0.3747	0.2315	0.1436
		(0.0049)	(0.0017)	(0.0007)
	GKO	0.3700	0.2303	0.1433
		(0.0046)	(0.0016)	(0.0006)
	SKM	0.4448	0.2822	0.1650
		(0.0208)	(0.0195)	(0.0117)
	GKM	0.4617	0.2823	0.1650
		(0.0262)	(0.0194)	(0.0117)
Student-t with 5 d.f.	SKO	0.3603	0.2358	0.1474
		(0.0056)	(0.0050)	(0.0030)
	GKO	0.3629	0.2379	0.1482
		(0.0058)	(0.0053)	(0.0031)
	SKM	0.5135	0.3550	0.2315
		(0.0259)	(0.0239)	(0.0216)
	GKM	0.5176	0.3626	0.2334
		(0.0265)	(0.0262)	(0.0224)
Lognormal	SKO	5.0216	4.5228	5.0797
		(0.2057)	(0.2110)	(0.3629)
	GKO	5.1563	4.6000	5.1406
		(0.2090)	(0.2153)	(0.3681)
	SKM	4.7578	3.8913	4.3753
		(0.2704)	(0.3858)	(0.4851)
	GKM	4.9788	4.0768	5.4050
		(0.3006)	(0.4219)	(0.7492)
Cauchy	SKO	1.6572	1.7855	2.2745
		(0.0908)	(0.1325)	(0.1767)
	GKO	1.6889	1.8078	2.2946
		(0.0911)	(0.1326)	(0.1759)
	SKM	1.7844	1.7429	1.6880
		(0.0879)	(0.1081)	(0.0925)
	GKM	1.9173	2.0015	2.5011
		(0.0983)	(0.1409)	(0.1681)
Bimodal	SKO	0.1749	0.1215	0.0815
		(0.0003)	(0.0005)	(0.0006)
	GKO	0.1750	0.1221	0.0818
		(0.0003)	(0.0006)	(0.0006)
	SKM	0.2370	0.1402	0.0887
		(0.0156)	(0.0067)	(0.0078)
	GKM	0.2369	0.1477	0.0887
		(0.0156)	(0.0101)	(0.0078)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.5 EMSE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=50.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	0.3796	0.2329	0.1428
		(0.0050)	(0.0023)	(0.0005)
	GKO	0.3749	0.2319	0.1426
		(0.0045)	(0.0021)	(0.0005)
	SKM	0.4196	0.2901	0.1652
		(0.0167)	(0.0190)	(0.0107)
	GKM	0.4229	0.2901	0.1652
		(0.0168)	(0.0190)	(0.0107)
Student-t with 5 d.f.	SKO	0.3611	0.2425	0.1484
		(0.0057)	(0.0111)	(0.0035)
	GKO	0.3640	0.2446	0.1492
		(0.0062)	(0.0114)	(0.0036)
	SKM	0.5093	0.4076	0.2330
		(0.0248)	(0.0430)	(0.0135)
	GKM	0.5107	0.4109	0.2333
		(0.0252)	(0.0446)	(0.0186)
Lognormal	SKO	5.2952	4.7594	4.0303
		(0.2403)	(0.2198)	(0.3346)
	GKO	5.3999	4.8146	4.0594
		(0.2426)	(0.2220)	(0.3363)
	SKM	5.1161	4.4198	3.3862
		(0.3247)	(0.3998)	(0.5300)
	GKM	5.1605	4.4977	3.4690
		(0.3310)	(0.4163)	(0.5685)
Cauchy	SKO	1.7766	1.7598	1.9475
		(0.1004)	(0.1425)	(0.2046)
	GKO	1.7985	1.7722	1.9550
		(0.1001)	(0.1424)	(0.2046)
	SKM	1.9985	1.8855	2.0238
		(0.1046)	(0.1498)	(0.2108)
	GKM	2.0187	1.9160	2.1285
		(0.1055)	(0.1532)	(0.2304)
Bimodal	SKO	0.1751	0.1209	0.0825
		(0.0003)	(0.0006)	(0.0007)
	GKO	0.1753	0.1215	0.0829
		(0.0003)	(0.0006)	(0.0007)
	SKM	0.2006	0.1317	0.0835
		(0.0052)	(0.0045)	(0.0029)
	GKM	0.2006	0.1317	0.0835
		(0.0052)	(0.0045)	(0.0029)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.6 MISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=1.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.5499	0.9071	0.5542
		(0.0224)	(0.0061)	(0.0028)
	GKO	1.5101	0.8970	0.5514
		(0.0204)	(0.0056)	(0.0026)
	SKM	1.9349	1.1289	0.5857
		(0.1165)	(0.0826)	(0.0157)
	GKM	1.8879	1.1283	0.5856
		(0.1005)	(0.0833)	(0.0163)
Student-t with 5 d.f.	SKO	1.5912	1.0270	0.5877
		(0.0586)	(0.0463)	(0.0062)
	GKO	1.6094	1.0431	0.5879
		(0.0732)	(0.0585)	(0.0064)
	SKM	1.9776	1.2158	0.8773
		(0.1255)	(0.0522)	(0.0652)
	GKM	2.1419	1.4458	0.9380
		(0.1729)	(0.1826)	(0.0861)
Lognormal	SKO	9.0896	8.8135	8.0011
		(0.3808)	(0.3961)	(0.3650)
	GKO	9.7040	9.3349	8.2480
		(0.4167)	(0.4388)	(0.3861)
	SKM	7.5139	5.8992	4.0723
		(0.3543)	(0.2825)	(0.3223)
	GKM	9.8446	8.0735	6.4504
		(0.6068)	(0.6538)	(0.8416)
Cauchy	SKO	6.8181	7.1628	8.0198
		(0.3825)	(0.5508)	(0.7957)
	GKO	7.2559	7.5798	8.3856
		(0.3884)	(0.5621)	(0.7934)
	SKM	4.9499	4.5687	3.2930
		(0.2037)	(0.2816)	(0.2162)
	GKM	8.1889	8.9893	8.6631
		(0.4157)	(0.5834)	(0.8158)
Bimodal	SKO	1.2361	0.8304	0.5717
		(0.0032)	(0.0037)	(0.0056)
	GKO	1.2451	0.8367	0.5755
		(0.0036)	(0.0040)	(0.0057)
	SKM	1.4741	0.9910	0.5734
		(0.0459)	(0.0641)	(0.0341)
	GKM	1.4740	1.0585	0.5737
		(0.0456)	(0.1019)	(0.0343)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.7 MISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=2.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.5142	0.9224	0.5612
		(0.0160)	(0.0112)	(0.0054)
	GKO	1.4862	0.9154	0.5594
		(0.0139)	(0.0110)	(0.0053)
	SKM	1.7361	1.1105	0.5953
		(0.0705)	(0.0924)	(0.0122)
	GKM	1.7355	1.1493	0.5950
		(0.0715)	(0.0989)	(0.0123)
Student-t with 5 d.f.	SKO	1.5404	0.9839	0.6394
		(0.0240)	(0.0396)	(0.0208)
	GKO	1.5453	0.9945	0.6451
		(0.0267)	(0.0481)	(0.0223)
	SKM	1.9094	1.2488	1.1597
		(0.0625)	(0.0647)	(0.1256)
	GKM	2.0378	1.3741	1.3086
		(0.0959)	(0.1487)	(0.1830)
Lognormal	SKO	9.5605	9.0159	8.2082
		(0.3347)	(0.4541)	(0.5633)
	GKO	9.9961	9.2978	8.5141
		(0.3540)	(0.4735)	(0.6054)
	SKM	8.0427	6.3457	4.9888
		(0.3352)	(0.3411)	(0.4173)
	GKM	9.3283	8.4601	8.6146
		(0.4981)	(0.7819)	(1.2731)
Cauchy	SKO	6.7247	6.8089	7.0303
		(0.3919)	(0.6177)	(0.6828)
	GKO	7.0261	7.0584	7.2723
		(0.3963)	(0.6222)	(0.6869)
	SKM	5.6564	4.8812	4.7156
		(0.2510)	(0.3003)	(0.2950)
	GKM	8.1509	8.1515	8.1146
		(0.4178)	(0.6052)	(0.6670)
Bimodal	SKO	1.2360	0.8419	0.5656
		(0.0027)	(0.0046)	(0.0044)
	GKO	1.2426	0.8479	0.5678
		(0.0031)	(0.0049)	(0.0045)
	SKM	1.4583	0.9236	0.5491
		(0.0581)	(0.0350)	(0.0214)
	GKM	1.4805	0.9489	0.5490
		(0.0616)	(0.0443)	(0.0214)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.8 MISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=5.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.5236	0.9141	0.5537
		(0.0205)	(0.0061)	(0.0024)
	GKO	1.5021	0.9088	0.5524
		(0.0187)	(0.0057)	(0.0022)
	SKM	1.7710	1.0720	0.6111
		(0.0919)	(0.0395)	(0.0211)
	GKM	1.7701	1.0720	0.6110
		(0.0923)	(0.0396)	(0.0211)
Student-t with 5 d.f.	SKO	1.5279	0.9385	0.5959
		(0.0208)	(0.0086)	(0.0094)
	GKO	1.5357	0.9407	0.5978
		(0.0238)	(0.0091)	(0.0096)
	SKM	2.0492	1.2631	1.0106
		(0.1582)	(0.0514)	(0.0903)
	GKM	2.0685	1.2958	1.0338
		(0.1561)	(0.0614)	(0.0999)
Lognormal	SKO	10.2550	8.9710	7.7496
		(0.4348)	(0.4474)	(0.3129)
	GKO	10.5904	9.1613	7.8696
		(0.4530)	(0.4601)	(0.3218)
	SKM	9.4185	7.5786	5.3995
		(0.4632)	(0.4985)	(0.5718)
	GKM	10.6941	8.5825	6.1201
		(0.6525)	(0.6836)	(0.7304)
Cauchy	SKO	7.0846	8.1911	8.3844
		(0.4354)	(0.5231)	(0.6359)
	GKO	7.2597	8.3733	8.5697
		(0.4368)	(0.5218)	(0.6433)
	SKM	6.9258	6.5412	5.7702
		(0.3292)	(0.3232)	(0.3843)
	GKM	8.1780	9.5033	9.7009
		(0.4358)	(0.5155)	(0.7226)
Bimodal	SKO	1.2386	0.8419	0.5632
		(0.0028)	(0.0049)	(0.0037)
	GKO	1.2442	0.8471	0.5662
		(0.0031)	(0.0052)	(0.0038)
	SKM	1.4293	0.9040	0.5822
		(0.0511)	(0.0249)	(0.0255)
	GKM	1.4500	0.9042	0.5823
		(0.0546)	(0.0250)	(0.0256)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.9 MISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=10.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.5004	0.9086	0.5535
		(0.0179)	(0.0062)	(0.0024)
	GKO	1.4829	0.9041	0.5525
		(0.0167)	(0.0057)	(0.0024)
	SKM	1.7440	1.0873	0.6292
		(0.0743)	(0.0692)	(0.0418)
	GKM	1.8046	1.0876	0.6290
		(0.0934)	(0.0691)	(0.0416)
Student-t with 5 d.f.	SKO	1.5383	0.9641	0.5920
		(0.0211)	(0.0155)	(0.0089)
	GKO	1.5420	0.9701	0.5941
		(0.0213)	(0.0165)	(0.0093)
	SKM	2.0834	1.3732	0.8924
		(0.1005)	(0.0843)	(0.0751)
	GKM	2.0986	1.4009	0.8992
		(0.1026)	(0.0926)	(0.0779)
Lognormal	SKO	9.2716	8.2191	9.0941
		(0.3488)	(0.3513)	(0.6350)
	GKO	9.4971	8.3510	9.2030
		(0.3563)	(0.3598)	(0.6458)
	SKM	9.0315	7.4140	7.9587
		(0.4476)	(0.6873)	(0.8155)
	GKM	9.4392	7.7524	9.8840
		(0.5083)	(0.7530)	(1.3270)
Cauchy	SKO	6.9969	7.4781	9.4690
		(0.3716)	(0.5498)	(0.7234)
	GKO	7.1360	7.5796	9.5630
		(0.3734)	(0.5506)	(0.7197)
	SKM	7.5234	7.3232	7.0721
		(0.3611)	(0.4531)	(0.3906)
	GKM	8.0676	8.3755	10.4193
		(0.4025)	(0.5830)	(0.6853)
Bimodal	SKO	1.2377	0.8450	0.5570
		(0.0024)	(0.0045)	(0.0045)
	GKO	1.2422	0.8499	0.5595
		(0.0027)	(0.0048)	(0.0046)
	SKM	1.6180	0.9488	0.5907
		(0.0992)	(0.0424)	(0.0498)
	GKM	1.6180	0.9967	0.5907
		(0.0990)	(0.0639)	(0.0498)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.10 MISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Normal criterion. Nsam=k*N where k=50.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.5185	0.9134	0.5506
		(0.0182)	(0.0084)	(0.0018)
	GKO	1.5007	0.9098	0.5498
		(0.0166)	(0.0079)	(0.0018)
	SKM	1.6547	1.1164	0.6294
		(0.0596)	(0.0679)	(0.0380)
	GKM	1.6666	1.1164	0.6294
		(0.0602)	(0.0679)	(0.0380)
Student-t with 5 d.f.	SKO	1.5398	0.9911	0.5956
		(0.0186)	(0.0374)	(0.0104)
	GKO	1.5446	0.9973	0.5979
		(0.0199)	(0.0387)	(0.0109)
	SKM	2.0580	1.5598	0.8847
		(0.0904)	(0.1534)	(0.0626)
	GKM	2.0629	1.5717	0.8856
		(0.0918)	(0.1592)	(0.0629)
Lognormal	SKO	9.7698	8.6163	7.3224
		(0.4141)	(0.3728)	(0.5760)
	GKO	9.9447	8.7101	7.3714
		(0.4195)	(0.3775)	(0.5797)
	SKM	9.7279	8.2314	6.3528
		(0.5561)	(0.6765)	(0.9149)
	GKM	9.8120	8.3766	6.5100
		(0.5692)	(0.7094)	(0.9942)
Cauchy	SKO	7.4839	7.3613	8.1099
		(0.4113)	(0.5889)	(0.8449)
	GKO	7.5751	7.4143	8.1424
		(0.4103)	(0.5887)	(0.8447)
	SKM	8.4001	7.8764	8.4195
		(0.4265)	(0.6169)	(0.8701)
	GKM	8.4846	8.0010	8.8367
		(0.4302)	(0.6306)	(0.9479)
Bimodal	SKO	1.2417	0.8403	0.5650
		(0.0026)	(0.0050)	(0.0052)
	GKO	1.2464	0.8449	0.5676
		(0.0030)	(0.0052)	(0.0053)
	SKM	1.3863	0.8919	0.5580
		(0.0326)	(0.0286)	(0.0185)
	GKM	1.3864	0.8919	0.5580
		(0.0326)	(0.0286)	(0.0185)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.11 EMSE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=1.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	0.4091	0.2426	0.1460
		(0.0070)	(0.0035)	(0.0009)
	GKO	0.3926	0.2392	0.1438
		(0.0073)	(0.0045)	(0.0007)
	SKM	0.4386	0.2541	0.1896
		(0.0201)	(0.0076)	(0.0199)
	GKM	0.4378	0.2547	0.1896
		(0.0201)	(0.0084)	(0.0201)
Student-t with 5 d.f.	SKO	0.3622	0.2189	0.1362
		(0.0047)	(0.0012)	(0.0010)
	GKO	0.3518	0.2183	0.1358
		(0.0045)	(0.0016)	(0.0011)
	SKM	0.4795	0.3385	0.2262
		(0.0284)	(0.0159)	(0.0233)
	GKM	0.5097	0.3650	0.2583
		(0.0326)	(0.0249)	(0.0362)
Lognormal	SKO	3.3104	2.4384	2.1522
		(0.1142)	(0.0923)	(0.1286)
	GKO	3.4121	2.4727	2.1851
		(0.1213)	(0.0970)	(0.1335)
	SKM	4.0106	2.8306	2.4579
		(0.1860)	(0.1819)	(0.2474)
	GKM	5.8355	4.7965	3.4559
		(0.3424)	(0.4407)	(0.4548)
Cauchy	SKO	0.4223	0.2427	0.1608
		(0.0219)	(0.0095)	(0.0090)
	GKO	0.3678	0.2329	0.1581
		(0.0156)	(0.0110)	(0.0103)
	SKM	1.0275	1.1820	1.2447
		(0.0442)	(0.0702)	(0.0925)
	GKM	1.8602	2.1841	2.5039
		(0.0977)	(0.1547)	(0.2016)
Bimodal	SKO	0.1753	0.1209	0.0831
		(0.0004)	(0.0006)	(0.0007)
	GKO	0.1751	0.1220	0.0837
		(0.0003)	(0.0006)	(0.0007)
	SKM	0.2244	0.1358	0.0821
		(0.0107)	(0.0045)	(0.0033)
	GKM	0.2239	0.1358	0.0821
		(0.0109)	(0.0045)	(0.0033)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.12 EMSE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=2.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	0.4095	0.2371	0.1452
		(0.0090)	(0.0027)	(0.0011)
	GKO	0.3928	0.2329	0.1446
		(0.0086)	(0.0023)	(0.0012)
	SKM	0.4511	0.2789	0.1649
		(0.0214)	(0.0209)	(0.0087)
	GKM	0.4611	0.2787	0.1648
		(0.0229)	(0.0211)	(0.0089)
Student-t with 5 d.f.	SKO	0.3621	0.2191	0.1354
		(0.0056)	(0.0018)	(0.0008)
	GKO	0.3557	0.2199	0.1350
		(0.0050)	(0.0025)	(0.0009)
	SKM	0.5665	0.3440	0.2531
		(0.0431)	(0.0216)	(0.0247)
	GKM	0.6137	0.3593	0.2680
		(0.0466)	(0.0246)	(0.0299)
Lognormal	SKO	3.1710	2.5948	1.9897
		(0.1201)	(0.1295)	(0.0556)
	GKO	3.3567	2.7141	1.9964
		(0.1310)	(0.1407)	(0.0597)
	SKM	4.1196	3.5365	3.2301
		(0.1931)	(0.2838)	(0.3480)
	GKM	5.3546	4.4617	4.6471
		(0.3206)	(0.3848)	(0.5155)
Cauchy	SKO	0.4002	0.2180	0.1567
		(0.0186)	(0.0078)	(0.0069)
	GKO	0.3952	0.2187	0.1569
		(0.0158)	(0.0077)	(0.0080)
	SKM	1.2639	1.4495	1.5819
		(0.0604)	(0.0994)	(0.1149)
	GKM	1.8985	2.1075	2.2455
		(0.1026)	(0.1489)	(0.1789)
Bimodal	SKO	0.1749	0.1206	0.0826
		(0.0004)	(0.0006)	(0.0007)
	GKO	0.1750	0.1213	0.0830
		(0.0003)	(0.0006)	(0.0007)
	SKM	0.2310	0.1411	0.0896
		(0.0098)	(0.0056)	(0.0068)
	GKM	0.2309	0.1412	0.0897
		(0.0098)	(0.0056)	(0.0068)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.13 EMSE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=5.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	0.3927	0.2492	0.1435
		(0.0062)	(0.0053)	(0.0007)
	GKO	0.3789	0.2450	0.1429
		(0.0054)	(0.0048)	(0.0007)
	SKM	0.4310	0.2723	0.1980
		(0.0166)	(0.0183)	(0.0215)
	GKM	0.4317	0.2719	0.1977
		(0.0168)	(0.0184)	(0.0215)
Student-t with 5 d.f.	SKO	0.3632	0.2232	0.1380
		(0.0053)	(0.0035)	(0.0022)
	GKO	0.3580	0.2226	0.1382
		(0.0054)	(0.0033)	(0.0023)
	SKM	0.5130	0.3839	0.3041
		(0.0267)	(0.0276)	(0.0361)
	GKM	0.5265	0.3978	0.3226
		(0.0295)	(0.0316)	(0.0430)
Lognormal	SKO	2.9604	2.5630	2.1149
		(0.1172)	(0.1333)	(0.0873)
	GKO	3.1565	2.6813	2.1893
		(0.1225)	(0.1401)	(0.0867)
	SKM	4.3172	3.6569	3.7839
		(0.2469)	(0.3058)	(0.6296)
	GKM	4.8863	4.1430	4.2077
		(0.2901)	(0.3682)	(0.7150)
Cauchy	SKO	0.3538	0.2266	0.1433
		(0.0136)	(0.0086)	(0.0069)
	GKO	0.3683	0.2303	0.1454
		(0.0153)	(0.0087)	(0.0077)
	SKM	1.3830	1.9104	1.8782
		(0.0662)	(0.1102)	(0.1551)
	GKM	1.7357	2.3614	2.2044
		(0.0919)	(0.1316)	(0.1889)
Bimodal	SKO	0.1762	0.1215	0.0818
		(0.0009)	(0.0006)	(0.0007)
	GKO	0.1758	0.1221	0.0822
		(0.0007)	(0.0006)	(0.0007)
	SKM	0.2101	0.1381	0.0834
		(0.0078)	(0.0052)	(0.0056)
	GKM	0.2100	0.1381	0.0834
		(0.0078)	(0.0052)	(0.0056)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.14 EMSE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=10.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	0.3977	0.2454	0.1483
		(0.0064)	(0.0041)	(0.0021)
	GKO	0.3827	0.2416	0.1477
		(0.0052)	(0.0037)	(0.0021)
	SKM	0.4575	0.2819	0.1694
		(0.0306)	(0.0135)	(0.0128)
	GKM	0.4614	0.2806	0.1693
		(0.0308)	(0.0131)	(0.0128)
Student-t with 5 d.f.	SKO	0.3686	0.2172	0.1347
		(0.0063)	(0.0016)	(0.0007)
	GKO	0.3604	0.2173	0.1351
		(0.0056)	(0.0016)	(0.0009)
	SKM	0.5455	0.3434	0.2992
		(0.0302)	(0.0170)	(0.0446)
	GKM	0.5600	0.3458	0.3060
		(0.0329)	(0.0173)	(0.0468)
Lognormal	SKO	3.0952	2.5333	1.8276
		(0.1077)	(0.1036)	(0.1122)
	GKO	3.3245	2.6549	1.8674
		(0.1184)	(0.1106)	(0.1157)
	SKM	4.9448	4.2391	3.8214
		(0.3105)	(0.3504)	(0.5228)
	GKM	5.5040	4.6708	4.1696
		(0.3592)	(0.4106)	(0.5740)
Cauchy	SKO	0.3435	0.2145	0.1479
		(0.0125)	(0.0090)	(0.0068)
	GKO	0.3638	0.2225	0.1501
		(0.0145)	(0.0102)	(0.0072)
	SKM	1.7467	1.7764	2.1313
		(0.0914)	(0.1333)	(0.2131)
	GKM	2.0336	1.9203	2.3139
		(0.1047)	(0.1386)	(0.2215)
Bimodal	SKO	0.1751	0.1216	0.0832
		(0.0003)	(0.0006)	(0.0006)
	GKO	0.1752	0.1222	0.0836
		(0.0003)	(0.0006)	(0.0006)
	SKM	0.2321	0.1310	0.0824
		(0.0131)	(0.0045)	(0.0030)
	GKM	0.2321	0.1310	0.0824
		(0.0131)	(0.0045)	(0.0030)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.15 EMSE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=50.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	0.4179	0.2409	0.1492
		(0.0107)	(0.0039)	(0.0015)
	GKO	0.4008	0.2376	0.1480
		(0.0092)	(0.0034)	(0.0013)
	SKM	0.4767	0.2748	0.1717
		(0.0262)	(0.0152)	(0.0093)
	GKM	0.4803	0.2748	0.1717
		(0.0263)	(0.0152)	(0.0093)
Student-t with 5 d.f.	SKO	0.3620	0.2209	0.1362
		(0.0049)	(0.0021)	(0.0016)
	GKO	0.3551	0.2195	0.1362
		(0.0042)	(0.0019)	(0.0016)
	SKM	0.5030	0.3301	0.3107
		(0.0228)	(0.0249)	(0.0456)
	GKM	0.5078	0.3306	0.3119
		(0.0235)	(0.0251)	(0.0459)
Lognormal	SKO	3.2502	2.5173	2.0671
		(0.1170)	(0.1115)	(0.1102)
	GKO	3.5584	2.6229	2.1049
		(0.1280)	(0.1135)	(0.1131)
	SKM	5.1512	3.9583	3.9454
		(0.2898)	(0.3154)	(0.5816)
	GKM	5.3918	4.0982	4.0120
		(0.3159)	(0.3327)	(0.5915)
Cauchy	SKO	0.3472	0.2307	0.1505
		(0.0124)	(0.0095)	(0.0085)
	GKO	0.3752	0.2400	0.1529
		(0.0144)	(0.0104)	(0.0088)
	SKM	1.7678	1.9545	2.1046
		(0.1019)	(0.1290)	(0.1821)
	GKM	1.8958	1.9970	2.1255
		(0.1078)	(0.1320)	(0.1834)
Bimodal	SKO	0.1751	0.1217	0.0823
		(0.0003)	(0.0006)	(0.0007)
	GKO	0.1747	0.1224	0.0826
		(0.0003)	(0.0006)	(0.0007)
	SKM	0.2161	0.1354	0.0876
		(0.0072)	(0.0046)	(0.0032)
	GKM	0.2163	0.1354	0.0876
		(0.0072)	(0.0046)	(0.0032)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.16 MISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=1.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.6228	0.9483	0.5621
		(0.0254)	(0.0127)	(0.0035)
	GKO	1.5650	0.9363	0.5541
		(0.0264)	(0.0163)	(0.0027)
	SKM	1.7202	0.9864	0.7163
		(0.0717)	(0.0271)	(0.0709)
	GKM	1.7207	0.9897	0.7163
		(0.0717)	(0.0300)	(0.0715)
Student-t with 5 d.f.	SKO	1.5856	0.9334	0.5692
		(0.0212)	(0.0055)	(0.0032)
	GKO	1.5370	0.9300	0.5678
		(0.0202)	(0.0065)	(0.0039)
	SKM	1.9394	1.3293	0.8603
		(0.1122)	(0.0571)	(0.0798)
	GKM	2.0473	1.4251	0.9750
		(0.1255)	(0.0886)	(0.1271)
Lognormal	SKO	6.5112	4.8491	4.2309
		(0.1682)	(0.1412)	(0.2035)
	GKO	6.7012	4.9259	4.2945
		(0.1803)	(0.1476)	(0.2108)
	SKM	7.7012	5.4917	4.7376
		(0.2913)	(0.2841)	(0.4001)
	GKM	10.9240	8.8871	6.4300
		(0.5882)	(0.7538)	(0.7710)
Cauchy	SKO	2.0317	1.1734	0.7610
		(0.0790)	(0.0329)	(0.0307)
	GKO	1.8776	1.1516	0.7546
		(0.0552)	(0.0383)	(0.0359)
	SKM	4.3669	4.9359	5.1902
		(0.1741)	(0.2914)	(0.3896)
	GKM	7.8571	9.1048	10.4050
		(0.3960)	(0.6358)	(0.8279)
Bimodal	SKO	1.2357	0.8404	0.5694
		(0.0027)	(0.0052)	(0.0051)
	GKO	1.2413	0.8496	0.5734
		(0.0029)	(0.0053)	(0.0053)
	SKM	1.5355	0.9182	0.5496
		(0.0676)	(0.0279)	(0.0205)
	GKM	1.5340	0.9187	0.5495
		(0.0689)	(0.0280)	(0.0207)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.17 MISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=2.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.6256	0.9286	0.5591
		(0.0324)	(0.0097)	(0.0041)
	GKO	1.5662	0.9135	0.5571
		(0.0310)	(0.0084)	(0.0045)
	SKM	1.7651	1.0751	0.6279
		(0.0765)	(0.0745)	(0.0310)
	GKM	1.8034	1.0750	0.6278
		(0.0818)	(0.0750)	(0.0317)
Student-t with 5 d.f.	SKO	1.5835	0.9365	0.5674
		(0.0253)	(0.0070)	(0.0039)
	GKO	1.5461	0.9378	0.5638
		(0.0220)	(0.0087)	(0.0034)
	SKM	2.2555	1.3421	0.9578
		(0.1708)	(0.0773)	(0.0859)
	GKM	2.4280	1.3973	1.0105
		(0.1803)	(0.0875)	(0.1043)
Lognormal	SKO	6.3267	5.1122	3.9763
		(0.1789)	(0.1988)	(0.0872)
	GKO	6.6276	5.3116	3.9929
		(0.1987)	(0.2176)	(0.0934)
	SKM	7.8949	6.6794	6.0046
		(0.3010)	(0.4645)	(0.5797)
	GKM	10.0836	8.2782	8.4325
		(0.5503)	(0.6484)	(0.8877)
Cauchy	SKO	1.9707	1.1043	0.7435
		(0.0663)	(0.0247)	(0.0230)
	GKO	1.9709	1.1054	0.7454
		(0.0553)	(0.0242)	(0.0276)
	SKM	5.3308	6.0789	6.6285
		(0.2469)	(0.4143)	(0.4860)
	GKM	7.9799	8.8165	9.3802
		(0.4204)	(0.6135)	(0.7376)
Bimodal	SKO	1.2346	0.8372	0.5656
		(0.0025)	(0.0049)	(0.0054)
	GKO	1.2412	0.8437	0.5685
		(0.0027)	(0.0053)	(0.0055)
	SKM	1.5736	0.9525	0.5978
		(0.0616)	(0.0353)	(0.0432)
	GKM	1.5736	0.9527	0.5980
		(0.0617)	(0.0353)	(0.0433)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.18 MISE ($\times 10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($\times 10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=5.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.5659	0.9728	0.5530
		(0.0226)	(0.0191)	(0.0027)
	GKO	1.5162	0.9575	0.5510
		(0.0196)	(0.0173)	(0.0025)
	SKM	1.6935	1.0511	0.7464
		(0.0593)	(0.0652)	(0.0765)
	GKM	1.6972	1.0502	0.7454
		(0.0600)	(0.0654)	(0.0765)
Student-t with 5 d.f.	SKO	1.5912	0.9495	0.5791
		(0.0236)	(0.0161)	(0.0097)
	GKO	1.5603	0.9444	0.5789
		(0.0227)	(0.0150)	(0.0103)
	SKM	2.0773	1.4789	1.1333
		(0.0997)	(0.0966)	(0.1267)
	GKM	2.1282	1.5293	1.2004
		(0.1093)	(0.1113)	(0.1522)
Lognormal	SKO	6.0564	5.0725	4.1776
		(0.1704)	(0.2055)	(0.1370)
	GKO	6.3473	5.2616	4.2967
		(0.1804)	(0.2163)	(0.1362)
	SKM	8.2881	6.9077	7.0602
		(0.4030)	(0.5007)	(1.0849)
	GKM	9.2713	7.7523	7.8275
		(0.4865)	(0.6160)	(1.2442)
Cauchy	SKO	1.8346	1.1314	0.7035
		(0.0472)	(0.0286)	(0.0223)
	GKO	1.8829	1.1427	0.7119
		(0.0539)	(0.0290)	(0.0254)
	SKM	5.8462	8.0110	7.8721
		(0.2732)	(0.4596)	(0.6498)
	GKM	7.3184	9.8714	9.1945
		(0.3767)	(0.5432)	(0.7787)
Bimodal	SKO	1.2422	0.8448	0.5596
		(0.0053)	(0.0047)	(0.0054)
	GKO	1.2442	0.8502	0.5621
		(0.0045)	(0.0050)	(0.0055)
	SKM	1.4463	0.9311	0.5592
		(0.0491)	(0.0330)	(0.0355)
	GKM	1.4464	0.9312	0.5592
		(0.0492)	(0.0330)	(0.0355)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.19 MISE ($\times 10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($\times 10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=10.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.5845	0.9592	0.5704
		(0.0233)	(0.0149)	(0.0078)
	GKO	1.5299	0.9455	0.5683
		(0.0191)	(0.0136)	(0.0076)
	SKM	1.7900	1.0879	0.6441
		(0.1091)	(0.0482)	(0.0457)
	GKM	1.8048	1.0835	0.6436
		(0.1096)	(0.0470)	(0.0456)
Student-t with 5 d.f.	SKO	1.6170	0.9264	0.5635
		(0.0284)	(0.0076)	(0.0029)
	GKO	1.5728	0.9233	0.5640
		(0.0249)	(0.0068)	(0.0035)
	SKM	2.1825	1.3368	1.1222
		(0.1115)	(0.0619)	(0.1572)
	GKM	2.2366	1.3450	1.1468
		(0.1210)	(0.0629)	(0.1655)
Lognormal	SKO	6.2239	5.0167	3.7284
		(0.1573)	(0.1568)	(0.1771)
	GKO	6.5709	5.2076	3.7933
		(0.1769)	(0.1687)	(0.1827)
	SKM	9.3899	7.9189	7.0573
		(0.5268)	(0.5931)	(0.8847)
	GKM	10.4071	8.6915	7.6669
		(0.6193)	(0.7046)	(0.9791)
Cauchy	SKO	1.7932	1.0955	0.7153
		(0.0419)	(0.0297)	(0.0223)
	GKO	1.8522	1.1205	0.7230
		(0.0499)	(0.0348)	(0.0239)
	SKM	7.3483	7.4352	8.8875
		(0.3764)	(0.5520)	(0.8853)
	GKM	8.5303	8.0417	9.6218
		(0.4280)	(0.5733)	(0.9146)
Bimodal	SKO	1.2392	0.8460	0.5701
		(0.0025)	(0.0051)	(0.0042)
	GKO	1.2442	0.8512	0.5729
		(0.0029)	(0.0054)	(0.0044)
	SKM	1.5801	0.8883	0.5510
		(0.0829)	(0.0282)	(0.0188)
	GKM	1.5801	0.8883	0.5510
		(0.0829)	(0.0282)	(0.0188)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

Table A7.20 MISE ($x10^{-2}$) of density estimates from simulated data. Averages over simulations, with standard errors ($x10^{-2}$) in brackets. Robust estimate for σ. Nsam=k*N where k=50.

Data	Methods	N=25 Nsim=100	N=50 Nsim=50	N=100 Nsim=25
Normal	SKO	1.6573	0.9431	0.5741
		(0.0384)	(0.0141)	(0.0056)
	GKO	1.5952	0.9389	0.5696
		(0.0330)	(0.0124)	(0.0047)
	SKM	1.8595	1.0611	0.6530
		(0.0935)	(0.0542)	(0.0331)
	GKM	1.8728	1.0610	0.6531
		(0.0940)	(0.0542)	(0.0332)
Student-t with 5 d.f.	SKO	1.5825	0.9464	0.5700
		(0.0214)	(0.0092)	(0.0058)
	GKO	1.5409	0.9372	0.5686
		(0.0171)	(0.0080)	(0.0055)
	SKM	2.0165	1.3049	1.1710
		(0.0804)	(0.0921)	(0.1644)
	GKM	2.0339	1.3070	1.1756
		(0.0828)	(0.0925)	(0.1658)
Lognormal	SKO	6.4496	4.9966	4.1056
		(0.1762)	(0.1712)	(0.1731)
	GKO	6.9114	5.1593	4.1658
		(0.1995)	(0.1750)	(0.1778)
	SKM	9.6432	7.3914	7.2948
		(0.4936)	(0.5208)	(1.0154)
	GKM	10.0852	7.6347	7.4119
		(0.5482)	(0.5536)	(1.0333)
Cauchy	SKO	1.8167	1.1502	0.7273
		(0.0432)	(0.0309)	(0.0287)
	GKO	1.8963	1.1798	0.7351
		(0.0513)	(0.0348)	(0.0297)
	SKM	7.4519	8.1840	8.8004
		(0.4164)	(0.5325)	(0.7570)
	GKM	7.9831	8.3628	8.8868
		(0.4397)	(0.5449)	(0.7620)
Bimodal	SKO	1.2374	0.8472	0.5628
		(0.0028)	(0.0050)	(0.0054)
	GKO	1.2392	0.8526	0.5653
		(0.0030)	(0.0053)	(0.0055)
	SKM	1.4828	0.9133	0.5846
		(0.0452)	(0.0290)	(0.0199)
	GKM	1.4839	0.9133	0.5846
		(0.0453)	(0.0290)	(0.0199)

Notes:
SKO - Student-t kernel with the method of moments procedure
SKM - Student-t kernel with the modified ML estimation procedure
GKO - Gaussian kernel with the Normal optimum estimation procedure
GKM - Gaussian kernel with the ML estimation procedure

REFERENCES

ABRAMSON, I.S. (1982) On Bandwidth Variation in Kernel Estimates - a Square Root Law. Annals of Statistics, **10**, 1217-1223.

AITCHISON, J. (1955) On the Distribution of a Positive Random Variable having a Discrete Probability Mass at the Origin. Journal of the American Statistical Association, **50**, 901-908.

AITCHISON, J. and DUNSMORE, I.R. (1975) Statistical Prediction Analysis. Cambridge: Cambridge University Press.

AITKEN, C.G.G. (1986) Statistical discriminant analysis in forensic science. Journal of the Forensic Science Society, **26**, 237-247.

BARTLETT, M. S. (1937) Some Examples of Statistical Methods of Research in Agriculture and Applied Biology. Journal of Royal Statistical Society Supplement, **4**, 131-137.

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. Journal of Statistical Computation and Simulation, **21**, 313-327.

BOX, G.E.P. and TIAO, G.C. (1973) Bayesian Inference in Statistical Analysis. Reading: Addison Wesley.

BREIMAN, L., MEISEL, W. and PURCELL, E. (1977) Variable Kernel Estimates of Multivariate Densities. Technometrics, **19**, 135-144.

BROEMELING, L.D. (1985) Bayesian Analysis of Linear Models. New York: Marcel Dekker Inc.

CHAN, K.P.S. and AITKEN, C.G.G. (1989) Estimation of the Bayes' Factor in a Forensic Science Problem. Journal of Statistical Computational and Simulation, **33**, 249-264.

COVER, T. (1972) A Hierarchy of Probability Density Function Estimates. Frontiers of Pattern Recognition, (Watanabe, S. (Ed.) 83-98. New York: Academic Press.

COX, D.R. (1966) Notes on the Analysis of Mixed Frequency Distributions. British Journal of Mathematical Statistics Psychology, **19**, 39-47.

DANIELS, H. E. (1939) Ihe Estimation of Components of Variance. Journal of the Royal Statistical Society, **6**, 186-197.

de FINETTI (1930) Sulla Proprieti Conglomerativa delle Probabilita Subordinate. Rend R. Inst. Lombardo (Milano), **63**, 414-418.

DEVROYE, L. and GYÖRFI, L. (1985) Nonparametric Density Estimation: The L_1 View. New York: Wiley.

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.

EFRON, B. (1981) Nonparametric estimates of Standard error: the Jackknife, the Bootstrap and other Methods. Biometrika, 68, 589-599.

EFRON, B. (1982) The Jackknife, the Bootstrap and other Resampling Plans. Philadelphia: SIAM.

EISENHART, C. (1947) The Assumptions Underlying the Analysis of Variance. Biometrics, 3, 1-21.

EVETT, I.W. (1982) What is the Probability that this Blood came from that Person? A Meaningful Question? Journal of Forensic Science Society, 23, 35-39.

EVETT, I.W. (1984) A Quantitative Theory for Interpreting Transfer Evidence in Criminal Cases. Applied Statistics, 33, 25-32.

EVETT, I.W., CAGE, P.E. and AITKEN, C.G.G. (1987) Evaluation of the Likelihood Ratio for Fibre Transfer Evidence in Criminal Cases. Applied Statistics, 36, 174-180.

FATTI, L.P. (1982) Predictive Discrimination Under the Random Effects Models. Journal of South African Statistics 16, 55-77.

FERGUSON, M. (1973) A Bayesian Analysis for Some Non-parametric Problems. Annals of Statistics, 1, 209-230.

FRYER, M.J. (1976) Some Errors Associated with the Nonparametric Methods of Density Estimation. Journal of Institute of Mathematics Applications, 18, 371-380.

FRYER, M.J. (1977) A Review of Some Non-parametric Method of Density Estimation. Journal of Institute of Mathematics Applications, 20, 335-354.

GEISSER, S. (1964) Posterior Odds for Multivariate NMormal Classifications. Journal of Royal Statistical Society, B, 26, 69-76.

GEISSER, S. (1966) Predictive Discrimination. Multivariate Analysis, 149-163.

GOOD, I.J. (1965) The Estimation of Probabilities. Cambridge: MIT Press.

GOOD, I.J. (1985) Weight of evidence: a brief survey. Bayesian Statistics 2 (Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.)). Elsevier Science Publishers B.V. (North - Holland).

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 penalized Likelihood Method Examplified by Scattering and Meteorite data. Journal of American Statistical Association, 75, 42-55.

GRAYBILL, F.A. (1954) On Quardratic estimates of Variance Components. Annals of Mathematical Statistics, 25, 367-372.

GRAYBILL, F.A. and HULTQUIST, R.A. (1961) Theorems on Eisenhart's Model II. Annals of Mathematical Statistics, 32, 261-269.

GRAYBILL, F.A. and WORTHAM, A.W. (1956) A Note on Uniformly Best Unbiased Estimates for Variance Components. Journal of the American Statistical Association, 51, 266-268.

GREGORY, G.G. and SCHUSTER, E.F. (1979) Contributions to Nonparametric Maximum Likelihood Methods of Density Estimation. In 12th Annual Symposium on the Interface of Computer Science and Statistics, (Gentleman, J.F. (Ed.)), 427-431. Waterloo: University of Waterloo.

GROVE, D.M. (1980) The Interpretation of Forensic Evidence Using a Likelihood Ratio. Biometrika, 67, 243-246.

HABBEMA, J.D.F., HERMANS, J. and VAN den BROEK, K. (1974) A Stepwise Discriminant Analysis Program Using Density Estimation. In COMPSTAT 1974, Proceedings in Computational Statistics, (Bruckmann, G. (Ed.)), 101-110, Vienna: Physica Verlag.

HARTIGAN, J.A. (1969) Linear Bayesian Methods. Journal of Royal Statistical Society, B, 31, 446-454.

HARTLEY, H.O. and RAO, J.N.K. (1967) Maximum Likelihood Estimation for the Mixed Analysis of Variance Model. Biometrika, 54, 93-108.

HILL, B.M. (1965) Inference about Variance Components in the One-way Model. Journal of the American Statistical Association, 60, 806-825.

HILL, B.M. (1967) Correlated Errors in the Random Model. Journal of the American Statistical Association, 62, 1387-1400.

HILL, B.M. (1970) Some Contrasts Between Bayesian and Classical Inference in the Analysis of Varaince and in the Testing of Models. In: Bayesian Statistics, (Meyer, D.L. and Collier, R.O. (Eds.)),

HUANG, W.M. (1987) Density Estimation and an adaptive test procedure. Computer Science and Statistics : Proceedings of the 19th Symposium on the Interface, 546-551.

JEFFREYS, H. (1939/1983) The Theory of Probability. Third edition, Oxford: Clarendon Press.

LEONARD, T. (1978) Density Estimation, Stochastic Processes and Prior Information. Journal of Royal Statistical Society, B, 40, 113-146.

LEVENE, H. (1960) In Contributions to Probability and Statistics. Stanford, California: Stanford University Press.

LINDLEY, D.V. (1965) Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2, Inference. Cambridge, England: Cambridge University Press.

LINDLEY, D.V. (1977) A problem in forensic science. Biometrika, 64, 207-213.

LOFTSGAARDEN, F. and QUESENBERY, C. (1965) A nonparametric Estimate of a Multivariate density function. Annals of Mathematical Statistics, 36, 1049-1051.

MAKOV, U.E. (1987) A Bayesian Treatment of the 'missing suspect' Problem in Forensic Science. The Statistician, 36, 251-258.

MARITZ, J.S. (1970) Empirical Bayes Methods. London: Methuen and Co.

NADARAJA, E.A. (1974) On the Integral Mean Square Error of some Nonparametric Estimates of a Probability Density. Theory of Probability and its Applications, 19, 133-141.

PARZEN, E. (1962) On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33, 1065-1076.

PEABODY, A.J., OXBOROUGH, R.J., CAGE, P.E. and EVETT, I.W. (1983) The discrimination of cat and dog hairs. Journal of the Forensic Science Society, 23, 121-129.

PRESS, J.S. (1982) Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference. New York: John Wiley and Sons.

RAATGEVER, J.W., HABBEMA, J.D.F. and HERMANS, J. (1979) Estimation of the Kernel Width in the Variable Kernel Model. Technical Report, Dept. of Medical Statistics, University of Leiden.

RAIFFA, H. and SCHLAIFER, R. (1961) Applied Statistical Decision Theory. Cambridge, Massachusetts.

RAMMSEY, F.P. (1931/1964) Truth and Probability. In The Foundations of Mathematics and other Essays. London: Kegan Paul.

RAO, C.R. (1973) Linear Statistical Inference and Its Applications. 2nd (ed.), New York: Wiley & Sons.

REMME, J., HABBEMA, J.D.F. and HERMANS, J. (1980) A Simulative Comparison of Linear, Quadratic and Kernel Discrimination. Journal of Statistical Computation and Simulation., 11, 87-106.

RIPLEY, B.D. (1983) Computer Generation of Random Variables: a Tutorial. International Statistical Review, 51, 301-319.

ROBBINS, H. (1955) An empirical Bayes approach to statistics. Proceedings of the Third Berkeley Symposium on Mathematics Statistics and Probability, 1, 157-164. University Calif. Press.

ROBBINS, H. (1964) The Empirical Bayes approach to Statistical decision problems. Annal of Mathematical Statistics, 35, 1-20.

ROSENBLATT, M (1956) Remarks on Some Nonparametric Estimates of a Density Function. Annals of Mathematical Statistics, 27, 432-837.

SAHAI, H. (1979) A Bibliography on Variance Components. International Statistical Review, 47, 177-222.

SAVAGE, L.J. (1962) The Foundations of Statistical Inference. New York: Wiley.

SCHEFFÉ, H. (1959) The Analysis of Variance. New York: Wiley & Sons.

SCOTT, D.W., TAPIA, R.A. and THOMPSON, J.R. (1977) Kernel Density Estimation Revisited. Nonlinear Analysis, Theory, Methods and Applications, 1, 339-372.

SEARLE, S.R. (1971) Linear Models. New York: John Wiley & Sons.

SEARLE, S.R. (1977) Variance Components Estimation: A Thumbnail Review. Biometric Unit Mimeo Series, BU-612-M, Ithaca, New York: Cornell University.

SEHEULT, A. (1978) On a Problem in Forensic Science. Biometrika, 65, 646-648.

SILVERMAN, B.W. (1978a) Choosing the Window Width When Estimating a Density. Biometrika, 65, 1-11.

SILVERMAN, B.W. (1978b) Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and its Derivatives. Annals of Statistics, 6, 177-184.

SILVERMAN, B.W. (1978c) Density Ratios, Empirical Likelihood and Cot Death. Applied Statistics, 27, 26-33.

SILVERMAN, B.W. (1981) Using Kernel density Estimates to Investigate Multimodality. Journal of Royal Statistical Society, B, 43, 97-99.

SILVERMAN, B.W. (1986) Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.

SMITH, A.F.M. (1973) Bayes Estimates in One-way and Two-way Models. Biometrika, 60, 319-329.

SMITH, A.F.M. and SPIEGELHALTER, D.J. (1980) Bayes Factors and Choice Criteria for Linear Models. Journal of Royal Statistical Society, B, 42, 213-220.

SNEDECOR and COCHRAN (1980) Statistical Methods. Iowa, U.S.A.: The Iowa State University Press.

SPIEGELHALTER, D.J. and SMITH, A.F.M. (1982) Bayes' factors for linear and log-linear models with vague prior information. Journal of Royal Statistical Society, B, 44, 377-387.

STONE, M. and SPRINGER, B.G.F. (1965) A Paradox Involving Quasi Prior Distributions. Biometrika, 52, 623-627.

TARTER, M.E. and KRONMAL, R.A. (1976) An Introduction to the Implementation of Theory of Nonparametric Density Estimation. The American Statistician, 30, 105-112.

TIAO, G.C. and ALI, M.M. (1971) Effect of Non-normality on Inferences about Variance Components. Technometrics, 13, 635-650.

TIAO, G.C. and TAN, W.Y. (1965) Bayesian Analysis of Random-effect Models in the Analysis of Variance, I. Posterior Distribution of Variance Components. Biometrika, 52, 37-53.

TIAO. G.C. and TAN, W.Y. (1966) Bayesian Analysis of Random-effect Models in the Analysis of Variance, II. Effect of Autocorrelated Errors. Biometrika, 53, 477-495.

TITTERINGTON, D.M. (1985) Common Structure of Smoothing Techniques in Statistics. International Statistical Review, 53, 141-170.

TUKEY, P.A. and TUKEY, J.W. (1981) Graphical Display of Data Sets in 3 or more Dimensions. In Barnett, V. (ed.), Interpreting Multivariate Data. Chichester: Wiley, 189-275.

WAHBA, G. (1976) Optimal Smoothing of Density Estimates. Technical Report No. 469, Department of Statistics, University of Madison.

WAHBA, G. (1983) Bayesian Confidence Intervals for the Cross-validated Smoothing Spline. Journal of Royal Statistical Society, B, 45, 144-150.

WAGNER, T.J. (1975) Nonparametric Estimates of Probability Densities. IEEE Trans. Information Theory, IT-21, 438-440.

WERTZ, W. and SCHNEIDER, B. (1979) Statistical Density Estimation: a Bibliography. International Statistical Review, 47, 155-175.

WHITTLE, P. (1958) On the Smoothing of Probability Density functions. Journal of Royal Statistical Society, B 20, 334-343.

WINSOR, C.P. and CLARKE, G.L. (1940) A Statistical Study of Variation in the Catch of Plankton Nets. Seras Foundation Journal of Marine Research, 3, 1-34.

WOFLGANG, W. and SCHNEIDER, B. (1979) Statistical Density Estimation: a Bibliography. International Statistical Review, **47**, 155-175.

WOODROOFE, M. (1970) On Choosing a Delta Sequence. Annals of Mathematical Statistics, **41**, 1665-1671.

YOUNG, C.W., LEGATES, J.E. and FARTHING, B.R. (1965) Pre- and Post-natal Influences on Growth, Prolificacy and Maternal Performance in Mice. Genetics, **52**, 553-561.