Statistical uncertainties in posterior probabilities

A.W. Ambergen

1980 Mathematics Subject Classification: 63H30, 62H10, 62H12, 62F15.
ISBN 90 6196 422 9
NUGI-code: 811

PREFACE

This tract is a reprint of my thesis. A few typing errors have been corrected.

I am grateful to all those persons from whose comments and suggestions I have benefited in carrying out this research and in writing this monograph. In particular I would like to acknowledge Prof.dr. W. Schaafsma for his enthusiastic and stimulating support, his many valuable advices, and his improvements on preliminary versions of the manuscript, Prof.dr. R.D. Gill for suggesting some alternative proofs, Prof.dr. A.J. Stam for detailed comments on a preliminary version of chapter two and for some references, Prof.dr. I. Ford for reading a preliminary version of the manuscript and indicating relevant literature, Dr. G.N. van Vark for providing the data for the application of the Border Cave cranium, Drs. D.M. van der Sluis for incorporating theory of this monograph in the computer program POSCON, Mickey Schim-van der Loeff for improving the English text, and Mini Middelberg for her conscientious typing of the manuscript.

I am grateful to the Editorial Board of the Stichting Mathematisch Centrum for the publication of this monograph in the Series CWI Tracts.

CONTENTS

Chapter 1

Introduction

1.1. INTRODUCTION, SCOPE, AND BACKGROUND

In this monograph some statistical aspects of classifications are considered. Two different meanings of the word classification should be distinguished: first, that where structures are imposed on an hitherto unstructured class of objects, e.g. by defining subclasses, designing categorical, hierarchical, or taxonomic systems, secondly, that where individual objects are allocated to the different classes of such predetermined systems.

We shall concern ourselves mainly with problems associated with the second meaning. Many activities result in such classifications. We are subconsiously recognizing, i.e. classifying, objects around us such as books, chairs, cats, dogs, etc. Taking action on the basis of such classifications is often completely interwoven with our everyday behaviour. Spoons, forks, knives are recognized as such and put into different boxes. Patterns are recognized as letters from which words and sentences are made. In these situations the observer feels no uncertainty about the true group membership of the observed objects. An assessment of the consequences of corresponding actions can be made without taking into account the fact that the objects could have been misinterpreted.

However, there are many practical situations in which considerable uncertainty exists about a specific classification, and consequently the observer has to often reckon with a wide range of possible consequences of corresponding actions. For example, pathologists are often concerned with situations in which there is much uncertainty about the true disease a patient is suffering from.

Much scientific work deals with the classification of objects, in both mentioned meanings of the world. In disciplines like archeology, biology, psychology, medicine, etc., the task may be to design a categorical system, or, if such

a system has become available, to allocate objects or individuals to the different classes.

Classification methods are studied from an abstract, i.e. not subject matter related point of view by techniques like discriminant analysis, pattern recognition, statistical decision functions, cluster analysis, fuzzy sets, etc. Each of these techniques focuses, from its own point of view, on a smaller or greater part of the process in which relations are established between objects and classes.

An ideal system such as a categorical system which partitions the overall population of objects into mutually exclusive subpopulations, may not always be expected to fully describe the real situation. Definitions may be more or less vague, or not enough knowledge is available. However, in this monograph we shall assume that a well-defined system with mutually exclusive classes, which are together exhaustive, is available.

The classification or assignment of a specific object to a class or population will usually appear as a forced decision at the end of the process of investigating relations between objects and classes. We shall consider models in which previous to the stage in which the classification is made, a set of probabilities has been computed for the object. One such probability indicates the probability that the object belongs to a specific population. Such probabilities are always conditional on the model specified. This means that in the statistical models we shall consider, these probabilities depend on the definitions of the populations, number of the populations, features selected, probability densities of these features for the populations, vector of scores of the objects on these features, and prior probabilities.

If the above-mentioned model parts are all known then the probabilities, which we shall call posterior probabilities from now on, can be computed without statistical uncertainty. In practice, however, not all model parts are completely known. For example, the population probability densities of the features have to be estimated from past observations. The consequence is that the posterior probabilities are affected by statistical uncertainties. In this monograph we shall study, among other things, some aspects of these statistical uncertainties in the posterior probabilities by means of the sampling distributions of the estimators of posterior probabilities.

Most of the statistical models in this monograph can be described as follows. The object under investigation is known to originate from one of $k(\geqslant 2)$ populations $\Pi_1,...,\Pi_k$. By means of the scores on p measurements, a p dimensional vector of scores x, characterizing the object, is obtained. We regard x as the outcome of a p dimensional random vector X. The probability distribution of this random vector X depends on the population the object comes from. Because it is unknown which population the object comes from, it is therefore unknown by which one of the k probability distributions the random vector X is described. Let these k probability distributions by which the populations $\Pi_1,...,\Pi_k$ are characterized, have probability densities $f_1,...,f_k$ on the p dimensional outcome space $\mathcal{X}$ of X. Then the object under investigation produces the probability density values $f_1(x),...,f_k(x)$ for the k populations. Often k so-

called prior probabilities $\rho_1,...,\rho_k$, $\Sigma_{h=1}^k \rho_h = 1$ can be assigned to the object. This assignment should be based on information which is conditionally independent of X given the population number. The meaning of these prior probabilities can be made clear by looking at the situation in which only the object and the k populations are given and no vector of scores of the object is available. In that situation the prior probabilities contain all the information we have about the group membership of the object. If we introduce the random variable T with

$$P(T=t) = \rho_t \quad , \quad t=1,...,k$$

where T describes the number of the population to which x belongs, then $\mathscr{L}X|T=t$ has density f_t, $t=1,...,k$. The conditional probabilities $\rho_{t|x}$ $t=1,...,k$ are determined by the theorem of Bayes. We have

$$\begin{aligned} \rho_{t|x} &= P(T=t|X=x) \\ &= P(T=t)f_t(x)/\sum_{h=1}^{k} P(T=h)f_h(x) \\ &= \rho_t f_t(x)/\sum_{h=1}^{k} \rho_h f_h(x) \quad t=1,...,k. \end{aligned} \tag{1.1.1}$$

It is assumed that the numerical values of $x, \rho_1,...,\rho_k$ are given. If $f_1,...,f_k$ would also be known, then also $f_1(x),...,f_k(x)$, and hence the posterior probabilities $\rho_{t|x}$, $t=1,...,k$, can be computed.

However, in practice, the population densities $f_1,...,f_k$ are unknown. Often, only their functional form as an element of a class or family of densities can be postulated. Accordingly, let us assume that the k-tuple of population densities is fully specified once the value is given of the underlying parameter θ which can be any element of the parameter set Θ. Write $f_{1,\theta},...,f_{k,\theta}$ for these population densities. As θ is unknown, the values $f_{1,\theta}(x),...,f_{k,\theta}(x)$ in the observation vector x are unknown. Hence the posterior probabilities

$$\rho_{t|x}(\theta) = \rho_t f_{t,\theta}(x)/\sum_{h=1}^{k} \rho_h f_{h,\theta}(x) \quad t=1,...,k \tag{1.1.2}$$

(see (1.1.1)) are functions of the unknown parameter θ. This implies that they themselves have to be regard as unknown parameters.

Throughout this monograph it is assumed that for each of the populations $\Pi_1,...,\Pi_k$ a random sample of objects is given. Each object is characterized by its p dimensional vector of scores. Thus, data of the form $x_{h1},...,x_{hn_h} \in \mathbb{R}^p$, $h=1,...k$, the so-called training samples, are available to us and regarded as outcomes of the independent random variables $X_{h1},...,X_{hn_h}$, $h=1,...,k$, where X_{hj} has density $f_{h,\theta}$, $h=1,...,k; j=1,...,n_h$. Further, let $X|T=t$ have density $f_{t,\theta}$. From the training samples the unknown parameter θ, the unknown values $f_{h,\theta}(x)$, $h=1,...,k$, and the unknown posterior probabilities $\rho_{h|x}(\theta)$, $h=1,...,k$ can be estimated. If $R_{t|x}$ denotes an estimator for $\rho_{t|x}(\theta)$, then $R_{t|x} = r_{t,x,\rho_1,...,\rho_k}(X_{11},...,X_{kn_k})$ where r is a suitable function.

Now, let $n = n_1 + \cdots + n_k$, and

$$R_{\cdot|x} = (R_{1|x},\ldots,R_{k|x})^T \text{ and } \rho_{\cdot|x}(\theta) = (\rho_{1|x}(\theta),\ldots,\rho_{k|x}(\theta)^T.$$

The asymptotic distribution of $n^{1/2}(R_{\cdot|x} - \rho_{\cdot|x}(\theta))$ for various models will be one of the main objects of study of this monograph.

Strong emphasis on the statistical uncertainties in estimates of posterior probabilities can be found in the following publications. SCHAAFSMA (1976) presents the asymptotic distributions of some statistics useful in obtaining such results. MCLACHLAN (1977) studied the asymptotic bias of posterior probabilities in a $k = 2$ model, but ignored the asymptotic variance. SCHAAFSMA and VAN VARK (1977) present asymptotic distributions of estimators for posterior probabilities in the case $k = 2, p = 1$ under assumptions of normality of the population distributions and equality of the variances. In SCHAAFSMA and VAN VARK (1979) results for the multivariate case $p \geqslant 1$ can be found. They assume $k = 2$ and equality of covariance matrices. In AMBERGEN (1981) their results have been extended to the case $k \geqslant 2, p \geqslant 1$ and equality of the covariance matrices. This latter publication also presents variances and covariances of some related statistics. AMBERGEN and SCHAAFSMA (1982) consider the case $k \geqslant 2, p \geqslant 1$, normality of the population densities, both for the situation of equality of covariance matrices and for the situation in which no assumptions about the equality of covariance matrices are made. This latter publication also theorizes about the nonparametric approach where kernel estimators are used for the estimation of posterior probabilities. AMBERGEN (1984) gives asymptotic distributions of estimators for posterior probabilities in the situation that both continuous and discrete variables are involved. AMBERGEN and SCHAAFSMA (1983, 1984) studied, as an application of the theory of estimating posterior and typicality probabilities, the origin of the Border Cave cranium, a subject well-known in physical anthropology. AMBERGEN and SCHAAFSMA (1985) compare four estimators of posterior probabilities under the assumption that the populations have normal densities.

Further relevant articles are those of SCHAAFSMA (1973, 1982, 1983, 1985a, 1985b). A credibility interval for the posterior probabilities has been derived by RIGBY (1982). CRITCHLEY and FORD (1984a, 1984b, 1985) studied uncertainties in certain statistics, e.g. estimators of log-odds, which are basic to the estimation of posterior probabilities, in $k = 2$ models. They introduced plots displaying the discriminant scores together with the statistical uncertainties in the corresponding log-odds. RIJAL (1984) extended their work to the case of unequal covariance matrices. AITCHISON, HABBEMA, and KAY (1977) compared the estimative method with the predictive method (see section 3.5). HABBEMA, HILDEN, and BJERREGAARD (1978, 1981) studied the measurement of performance in probabilistic diagnosis. Various types of scoring rules played a major role in their research. HERMANS and HABBEMA (1975) compared five methods to estimate posterior probabilities. HERMANS et al. (1981) evaluated several methods of discriminant analysis by means of posterior probabilities and penalty scores. They studied the reliability of the estimated posterior probabilities by means of equation (9) of HABBEMA, HILDEN, and

BJERREGAARD (1978,II). GANESALINGAM and MCLACHLAN (1979) analysed data of HABBEMA, HERMANS and VAN DEN BROEK (1974) by computing discriminant functions and posterior probabilities. They tried to obtain some idea of the reliability of the estimated posterior probabilities by comparing the results of the two methods.

Recent publications closely related to the subject of this monograph, are the following ones. CRITCHLEY, FORD, and RIJAL (1988) present methods based on the profile log-likelihood with the aim of improving the small-sample properties of interval estimates for posterior log-odds. CRITCHLEY, FORD, and RIJAL (1987) review a number of methods. CRITCHLEY, FORD, and HIRST (1988) present an evaluation of methods of interval estimation for log-odds by applying them to a set of medical data. CRITCHLEY, FORD, and HIRST (1987) deal with a possible simplification of the profile log-likelihood method.

Some of the results presented in this monograph have been implemented in the computer program POSCON, that has been programmed by D.M. van der Sluis. The program computes estimates of posterior probabilities, standard deviations, and correlations between them. An important facility, which has been implemented in this program, is that the set of variables is allowed to be partitioned into a number of subsets which are regarded as independent. To each of the subset one of the models, to be considered in the next chapters, can be applied. See the manuals of the computer program by VAN DER SLUIS, SCHAAFSMA and AMBERGEN (1985, 1986), section 5.2 of this monograph for a short description, VAN DER SLUIS and SCHAAFSMA (1984), and SCHAAFSMA and AMBERGEN (1987).

The subject of selecting variables has not been treated in this monograph. For a recent study and survey we refer to STEERNEMAN (1987).

In many practical discrimination problems both continuous and discrete variables are involved. For example in medicine a differential diagnosis is often based on discrete variables, e.g. symptoms which are present or absent, as well as based on continuous variables, e.g. clinical tests. A method often used in this situation is logistic discrimination. See COX (1966, 1970), DAY and KERRIDGE (1967) and ANDERSON (1972, 1973). The inverse of the Fisher-information matrix of corresponding maximum likelihood estimators can be used to give standard errors of the logistic posterior probabilities. By means of iterative algorithms the maximum likelihood parameters are computed. The logistic approach is also used if all variables are of one type, discrete or continuous. The logistic posterior probabilities are equal to the classical posterior probabilities for example when the variables are (1) multivariate normal with equal covariance matrices, (2) multivariate dichotomous, (3) multivariate dichotomous following the log-linear model with equal second and higher order effects, (4) a combination of (1) and (3), (see ANDERSON (1973)). Many have applied the logistic approach. For example, KRUSIŃSKA (1981, 1982) uses iterative algorithms for the maximum likelihood estimators of the parameters of the logistic discriminant functions in order to compute posterior probabilities in a $k=2$ ($k \geqslant 2$, respectively) model with data from exponential families of distributions. The measure of the goodness of

discrimination is the percentage of correctly classified individuals.

1.2. CHAPTER OUTLINES

In this section a short survey will be given of the subjects to be studied in the following chapters.

In chapter 2, the main result of this research, the asymptotic distribution of the estimator of the vector of posterior probabilities is derived for a wide class of models. Each of these models has $k \geqslant 2$ populations and $p \geqslant 1$ variables. We distinguish between models in which all variables are discrete, those where all variables are continuous, and models in which some variables are discrete, the other ones being continuous. If all variables are continuous, then it is assumed that they follow a multivariate normal distribution. If both continuous and discrete variables are involved, then the normality of the continuous variables is assumed to be conditional on the discrete ones. Further, some special cases are specified by assuming equality of covariance matrices of multivariate normal distributions for some special groups of outcomes of the discrete variables.

Chapter 3 starts with the standard construction of Bayes rules in situations where the parameters of the population densities are known. Some applications of Bayes rules to relevant problems with normal densities are given. Next, in situations where parameters of the population densities are unknown, the asymptotic distribution of the posterior probabilities is used in various ways. Further, attention is paid to situations where a forced decision has to be taken. The fully Bayesian approach is also considered.

In chapter 4 various miscellaneous results are given. For models with multivariate normal distributions, the non-existence of unbiased estimators of the posterior probablities is established. Unbiased estimators for various quantities such as (1) the value of the multivariate normal density in a specified point, (2) the value of the log-odds, and (3) the logarithm of the multivariate normal density, are derived. Variances and covariances of the last two estimators are obtained. Further, various methods of constructing confidence intervals for posterior probabilities are compared.

In chapter 5 a numerical application is studied, a short description of the computer program POSCON is given, and a simulation study is carried out. The application concerns a case from physical anthropology, namely a cranium found in Border Cave in South Africa. The computations were performed by a computer program written by the author. The computer program POSCON is a new, extended version of this earlier program. Section 5.3 deals with a simulation study where both discrete and continuous variables are involved.

Chapter 2

Derivation of the asymptotic distribution of the estimator for the vector of posterior probabilities

2.1. INTRODUCTION AND SUMMARY

In this chapter the asymptotic distribution of the estimator of the vector of posterior probabilities is derived for various models. The models have in common that $k \geqslant 2$ populations are involved about which information is available through so-called training samples. The vector of scores of the individual or object we want to study, as well as the prior probabilities for each of the k populations for this individual or object are given. The posterior probabilities belonging to the individual or object are considered as parameters which are estimated from the training samples.

In this chapter we shall use the notation in which X refers to a continuous random vector, D to a discrete one, and (X,D) to a random vector consisting of continuous as well as discrete components. Y will be used in the discussion of the general situation covering each of the three cases. Let the vector of scores of the individual or object under investigation be denoted by y, which is considered as a realisation of the random vector Y. Let T be the random variable which describes the number of the population the vector of scores y comes from, and let

$$P(T=t) = \rho_t \quad , \quad t=1,...,k$$

denote the prior probabilities which are assumed to be given. The conditional distribution $\mathcal{L}(Y|T=t)$ depends on the unknown parameter $\theta \in \Theta$, usually through some subvector $\theta_t, t=1,...,k$. Let $\mathcal{L}(Y|T=t)$ have the Radon-Nikodym derivative $f_{t,\theta}$ with respect to some measure σ on the outcome space of $Y, (t=1,...,k)$. In our models the measure σ will be (1) the Lebesgue measure, or (2) the counting measure, or (3) the product of (1) and (2). The posterior probabilities $\rho_{t|y}(\theta)$ are given by

$$\rho_{t|y}(\theta) = \rho_t f_{t,\theta}(y) / \sum_{h=1}^{k} \rho_h f_{h,\theta}(y) \quad t=1,...,k \tag{2.1.1}$$

(see (1.1.2)). These posterior probabilities depend on the unknown parameter θ. We shall estimate $\rho_{t|y}(\theta), t=1,...,k$ from the training samples $y_{11},...,y_{kn_k}$ which are considered outcomes of the independent random variables $Y_{11},...,Y_{kn_k}$. The random variable Y_{hi} has the same distribution as Y given $T=h$ $(i=1,...,n_h; h=1,...,k)$. Let $R_{t|y}$ be an estimator of $\rho_{t|y}(\theta), t=1,...,k$. Of course, $R_{t|y}$ is a function of $Y_{11},...,Y_{kn_k}$. We shall use the notation

$$R_{\cdot|y} = (R_{1|y},...,R_{k|y})^T \text{ and } \rho_{\cdot|y}(\theta)=(\rho_{1|y}(\theta),...,\rho_{k|y}(\theta))^T.$$

Let θ_0 be the real but unknown parameter point. Further, let $n=n_1+\cdots+n_k$. We shall derive that

$$\mathcal{L} n^{1/2}(R_{\cdot|y}-\rho_{\cdot|y}(\theta_0) \to N_k(0,\Psi_{\theta_0} M_{\theta_0} \Psi_{\theta_0}) \tag{2.1.2}$$

where Ψ_{θ_0} is the $k \times k$ matrix of partial derivatives of $\rho_{\cdot|y}(\theta)$ with respect to $\log f_{t,\theta}(y), t=1,...,k$ evaluated at the point $(\log f_{1,\theta_0}(y),...,\log f_{k,\theta_0}(y))^T$. From

$$\rho_{t|y}(\theta) = \rho_t \exp(\log f_{t,\theta}(y)) / \sum_{h=1}^{k} \rho_h \exp(\log f_{h,\theta}(y))$$

$t=1,...,k$, we have that Ψ_{θ_0} is specified by

$$\Psi_{\theta_0;tt} = \rho_{t|y}(\theta_0)(1-\rho_{t|y}(\theta_0)) \quad t=1,...,k \tag{2.1.3}$$

$$\Psi_{\theta_0;ts} = -\rho_{t|y}(\theta_0)\rho_{s|y}(\theta_0) \quad t,s=1,...,k;\ t \neq s. \tag{2.1.4}$$

The matrix M_{θ_0} has size $k \times k$ and depends on the model assumptions to be imposed. Note that Ψ_{θ_0} can also be written as

$$\Psi_{\theta_0} = diag(\rho_{\cdot|y}(\theta_0) - \rho_{\cdot|y}(\theta_0)\rho_{\cdot|y}^T(\theta_0)$$

where *diag* means diagonal matrix. Further, note that Ψ_{θ_0} is symmetric and that $|\Psi_{\theta_0}|=0$.

Instead of $\rho_{\cdot|y}(\theta_0)$, Ψ_{θ_0}, and M_{θ_0}, expressing that the vector of posterior probabilities and the two matrices depend on θ_0, we shall usually use the shorter notations $\rho_{\cdot|y}$, Ψ, and M, respectively.

The derivation of the asymptotic result of (2.1.2) for various models will be the subject of this chapter.

Three different kinds of model will be studied. These differ from each other in the types of components of the random variables Y and $Y_{hi}, i=1,...,n_h; h=1,...,k$. Let there be q components of discrete type and p components of continuous type. We distinguish between

(1) all discrete; $q>0, p=0$,
(2) all continuous; $q=0, p>0$,
(3) both continuous and discrete; $q>0, p>0$.

For the reader's convenience, the corresponding results are mentioned below.

The proofs will be given in the sections 2.3 up to 2.6.

Ad(1). (All components are discrete). Let us, for notational convenience, transform the q discrete random variables into a new one. Let the categories of this new random variable be numbered from 1 up to $d(\leqslant\infty)$. Usually, this number d is the product of the numbers of categories of the q original random variables, but in practice some recoding may be applied reducing the number of categories. Accordingly, we work with the independent discrete-valued random variables

$$D,\ D_{11},\ldots,D_{1n_1},\ldots,\ D_{k1},\ldots,D_{kn_k}$$

in which

$$\begin{aligned}
&P(D=\ell|T=t)=p_{t\ell}, && \ell=1,\ldots,d; \quad t=1,\ldots,k,\\
&P(T=t)=\rho_t\ , && t=1,\ldots,k\\
&P(D_{hi}=\ell)=P_{h\ell}\ , && \ell=1,\ldots,d; \quad i=1,\ldots,n_h; \quad h=1,\ldots,k,
\end{aligned}$$

and

$$\sum_{\ell=1}^{d} p_{h\ell}=1, \qquad h=1,\ldots,k.$$

Let j denote the outcome of the random variable D. Then, using the natural estimator of the $\rho_{t|j}$'s, i.e. that to be defined in (2.3.2), the elements of the matrix M in (2.1.2) are

$$\begin{aligned}
&M_{tt} = b_t^{-1}p_{tj}^{-1}(1-p_{tj}) && t = 1,\ldots,k\\
&M_{ts} = 0 && t,s = 1,\ldots,k; \quad t\neq s
\end{aligned}$$

where b_t is defined by $n_t/n\to b_t$, $t=1,\ldots,k$, and y has been replaced by j.

Ad(2). (All components are continuous.) We assume that the p continuous variables follow a multivariate normal distribution. So the independent random variables

$$X, X_{11},\ldots,X_{1n_1},\ldots,X_{k1},\ldots,X_{kn_k}$$

are distributed as follows

$$\begin{aligned}
&X|T = t \sim N_p(\mu_t, \Sigma_t) && t = 1,\ldots,k\\
&P(T = t) = \rho_t && t = 1,\ldots,k\\
&X_{hi} \sim N_p(\mu_h, \Sigma_h) && i=1,\ldots,n_h; \quad h = 1,\ldots,k.
\end{aligned}$$

Let $f_t(x)$ denote the $f_{t,\theta}(y)$ of (2.1.1) then

$$f_t(x) = (2\pi)^{-\frac{1}{2}p}|\Sigma_t|^{-1/2}\exp(-\tfrac{1}{2}\Delta_{x;t}^2)$$

where

$$\Delta^2_{x;t} = (x-\mu_t)^T\Sigma_t^{-1}(x-\mu_t).$$

It is natural to distinguish between situations where equality of the dispersion matrices $\Sigma_1,\ldots,\Sigma_k$ is imposed, case (A), and those where this assumption is not made, case (B). As a consequence, different estimators will appear each having its own asymptotic distribution.

(A). (Equality of dispersion matrices: $\Sigma_1 = \cdots = \Sigma_k = \Sigma$.) We distinguish between two ways of estimating the posterior propabilities.
(A.1). Because of the equality of dispersion matrices the factors $(2\pi)^{-\frac{1}{2}p}|\Sigma_t|^{-\frac{1}{2}}$ can be cancelled in numerator and denominator of (2.1.1). Then using the estimator to be defined in (2.4.16) the matrix M of (2.1.2) becomes

$$M_{tt} = b_t^{-1}\Delta^2_{x;t} + \tfrac{1}{2}\Delta^4_{x;t} \qquad t = 1,\ldots,k$$

$$M_{ts} = \tfrac{1}{2}\{(x-\mu_t)^T\Sigma^{-1}(x-\mu_s)\}^2 \qquad t,s = 1,\ldots,k;\quad t\neq s.$$

(A.2). The factors mentioned under A.1 are not cancelled. Using the estimators to be defined in A.2 of section 2.4 the matrix M becomes

$$M_{tt} = \tfrac{1}{2}p + (b_t^{-1}-1)\Delta^2_{x;t} + \tfrac{1}{2}\Delta^4_{x;t}$$

$$M_{ts} = \tfrac{1}{2}p - \tfrac{1}{2}\Delta^2_{x;t} - \tfrac{1}{2}\Delta^2_{x;s} + \tfrac{1}{2}\{(x-\mu_s)^T\Sigma^{-1}(x-\mu_t)\}^2$$

for $t,s = 1,\ldots,k;\ t\neq s$.

REMARK. The matrices M presented under A.1 and A.2 are different. The first one is the covariance matrix of the asymptotic distribution of

$$n^{1/2}(\tfrac{1}{2}\hat{\Delta}^2_{x;t} - \tfrac{1}{2}\Delta^2_{x;t}),\ t = 1,\ldots,k.$$

Whereas the M under A.2 is the covariance matrix of the asymptotic distribution of

$$n^{1/2}(-\tfrac{1}{2}\log(|2\pi\hat{\Sigma}|) - \tfrac{1}{2}\hat{\Delta}^2_{x;t} - (-\tfrac{1}{2}\log(|2\pi\Sigma|) - \tfrac{1}{2}\Delta^2_{x;t})),\quad t = 1,\ldots,k.$$

However, if the M's are premultiplied and postmultiplied by the matrix Ψ specified in (2.1.3) and (2.1.4), the composed matrices $\Psi M\Psi$ are equal for both approaches. Therefore, the asymptotic distributions of the estimator of the vector of posterior probabilities are equal for both approaches.

(B). (No assumptions about the dispersion matrices.) With the estimators to be defined in B of section 2.4 matrix M becomes

$$M_{tt} = \tfrac{1}{2}b_t^{-1}(p+\Delta^4_{x;t}) \qquad t = 1,\ldots,k$$

$$M_{ts} = 0 \qquad t,s = 1,\ldots,k;\quad t\neq s.$$

REMARK. If $\Sigma_1 = \cdots = \Sigma_k$ is assumed then a comparison between the matrices M under A.2 and B is of interest as a check on the validity of the formulas. The matrices M under B and A.2, which are the covariance matrices of the asymptotic distribution of $n^{1/2}(\log \hat{f}(x) - \log f(x))$, have the property that their difference is positive definite. For this verification see remark after theorem 2.4.4.

Ad(3). (Both continuous and discrete components.) The discrete variables are combined into one discrete variable. We assume further that the p continuous variables have conditional on the discrete variable a multivariate normal distribution. We work with the independent random variables

$$(X,D),\ (X_{11},D_{11}),\ldots,(X_{1n_1},\ D_{1n_1}),\ldots,(X_{k1},\ D_{k1}),\ldots,(X_{kn_k},D_{kn_k})$$

in which

$$\begin{aligned} &X|D = \ell,\ T=t \sim N_p(\mu_{t\ell},\Sigma_{t\ell}) && t=1,\ldots,k; \quad \ell=1,\ldots,d \\ &P(D = \ell|T=t) = p_{t\ell} && t=1,\ldots,k; \quad \ell=1,\ldots,d \\ &P(T=t) = \rho_t && t=1,\ldots,k \end{aligned}$$

and

$$\begin{aligned} &X_{hi}|D_{hi} = \ell \sim N_p(\mu_{h\ell},\ \Sigma_{h\ell}) && i=1,\ldots,n_h;\ h=1,\ldots,k;\ \ell = 1,\ldots,d \\ &P(D_{hi} = \ell) = p_{h\ell} && i=1,\ldots,n_h;\ h=1,\ldots,k;\ \ell = 1,\ldots,d. \end{aligned}$$

Again different cases are generated by making different assumptions about the dispersion matrices $\Sigma_{h\ell}, h=1,\ldots,k; \ell=1,\ldots,d$. In each of the cases a special estimator for the posterior probabilities will appear. These estimators differ in the way in which the dispersion matrices $\Sigma_{hj}, h=1,\ldots,k$ are estimated; here j comes from the realisation (x,j) of (X,D). The four cases presented below are special situations of the more general results to be presented in theorem 2.5.2 and 2.5.3. We have for the matrix M of (2.1.2) that the following results hold true.

First. (No assumption about $\Sigma_{1j},\ldots,\Sigma_{kj}$.)

$$\begin{aligned} M_{tt} &= b_t^{-1}\left\{\tfrac{1}{2}p_{tj}^{-1}(p+\Delta^4_{x;tj}) + p_{tj}^{-1}(1-p_{tj})\right\} \\ M_{ts} &= 0 \end{aligned}$$

for $t,s=1,\ldots,k;\ t\neq s$.

Second. (Assumption $\Sigma_{h1} = \cdots = \Sigma_{hd},\ h=1,\ldots,k$.)

$$\begin{aligned} M_{tt} &= b_t^{-1}\left\{\tfrac{1}{2}(p+\Delta^4_{x;tj}) + p_{tj}^{-1}(1-p_{tj})(1+\Delta^2_{x;tj})\right\} \\ M_{ts} &= 0 \end{aligned}$$

for $t,s=1,\ldots,k;\ t\neq s$.

Third. (Assumption $\Sigma_{1j} = \cdots = \Sigma_{kj}$.)

$$M_{tt} = b_t^{-1}\left\{p_{tj}^{-1}(1-p_{tj}) + p_{tj}^{-1}\Delta^2_{x;tj}\right\} + (\sum_{h=1}^{k} p_{hj}b_h)^{-1}\left\{\frac{1}{2}p - \Delta^2_{x;tj} + \frac{1}{2}\Delta^4_{x;tj}\right\}$$

$$M_{ts} = \frac{1}{2}(\sum_{h=1}^{k} p_{hj}b_h)^{-1}\left\{p - \Delta^2_{x;tj} - \Delta^2_{x;sj} + \Delta^4_{x;tjsj}\right\}$$

for $t,s = 1,\ldots,k$; $t \neq s$.

Fourth. (Assumption $\Sigma_{11} = \cdots = \Sigma_{kd}$.)

$$M_{tt} = b_t^{-1}p_{tj}^{-1}(1-p_{tj}+\Delta^2_{x;tj}) + \frac{1}{2}p - \Delta^2_{x,tj} + \frac{1}{2}\Delta^4_{x;tj}$$

$$M_{ts} = \frac{1}{2}p - \frac{1}{2}\Delta^2_{x;tj} - \frac{1}{2}\Delta^2_{x;sj} + \frac{1}{2}\Delta^4_{x;tjsj}$$

for $t,s = 1,\ldots,k$; $t \neq s$.
In the above mentioned results the following notations were used: $\Delta^2_{x;tjsj} = (x-\mu_{tj})^T\Sigma_{uj}^{-1}(x-\mu_{sj})$ with $u = t$ or $u = s$, and $\Delta^2_{x;tj} = \Delta^2_{x;tjtj}$.

REMARK. The results of the discrete case under $ad(1)$, and the continuous cases A.2 and B under $ad(2)$ can be derived as special situations of these formulas. The discrete case is obtained from each of these four cases by taking $p = 0$, and $\Delta^2_{x;tj} = \Delta^2_{x;sj} = \Delta^4_{x;tjsj} = 0$. If the discrete variable has only one possible outcome then the continuous variables are the only interesting ones. Therefore, with $p_{tj} = 1$, tj replaced by t, and $tjsj$ replaced by ts, we obtain that case *first* becomes B, *second* becomes B, *third* becomes A.2, and *fourth* becomes A.2. For more of such special derivations, especially those who give A.1, we refer to the remarks after theorem 2.5.2 and 2.5.3.

In many applications one can throw doubt on the assumption of normality and especially that of equality of covariance matrices. In such situations a suitable approach is to do tests about the normality and equality of covariance matrices in order to get a decisive answer. Possible transformations of the data may precede these tests.

If the normality is rejected, and eventually other types of parametrized densities are rejected too, one can always fall back upon nonparametric methods with kernel and window estimators. Such methods for obtaining the population densities are often based on the minimization of the integrated mean squared error, which is the variance plus the squared bias. See the early articles of ROSENBLATT (1956), PARZEN (1962), and CACOULLOS (1966). A review article about the subject is BEAN and TSOKOS (1980). The sensitivity of the estimated population densities for the width parameter can cause large standard errors in the estimates of the posterior probabilities. Asymptotic distribution results for posterior probability estimates based on nonparametric

methods with kernel estimators can be found in e.g. AMBERGEN and SCHAAFSMA (1982).

Simulation-based comparative studies on the quality of the linear, quadratic, and kernel discrimination have been done by e.g. REMME et al. (1980) for multinormal, lognormal, etc. distributions. Their criterion for the goodness of discrimination is the percentage of correct classifications. See also articles of HABBEMA et al. (1974, 1978).

The effects of dimension on the percentage of correct classification for linear, quadratic, and kernel discrimination were studied by e.g. VAN NESS and SIMPSON (1976) and VAN NESS (1979).

By means of the bootstrap method, standard errors or confidence intervals for the posterior probability can also be obtained. As references to the bootstrap method we mention EFRON (1977, 1981, and 1982). The percentile method, described in EFRON (1981) section 4, can be used for obtaining a confidence interval for the posterior probabilities for each of the earlier mentioned models.

Dealing with both discrete and continuous variables, KRUSIŃSKA (1984) transforms the discrete variables into linguistic variables using a transformation method of SAITTA and TORASSO (1981). Thereafter linear or quadratic discrimination functions can be applied. For linguistic variables and related membership functions of fuzzy sets, see ZADEH (1965, 1975, and 1976).

The type of distribution used for the case of both continuous and discrete variables in *ad*(3) and in section 2.5, is a special one of the class of conditional Gaussian distributions presented in LAURITZEN and WERMUTH (1984). They studied so-called mixed-interaction-models for a set of continuous and discrete variables using conditional Gaussian distributions. If there are q discrete and p continuous variables, i belongs to the outcome space of the discrete ones, and y to the outcome space of the continuous ones, then the probability function of a conditional Gaussian distribution is given by

$$f(i,y) = \exp\left\{g(i) + h(i)^T y - \tfrac{1}{2} y^T \Lambda(i) y\right\}$$

where g is a real valued function of i, h is a vector valued function of i taking values in $\mathbb{R}^p$, $\Lambda(i)$ is a $p \times p$ matrix valued function of i taking values in the set of positive definite symmetric matrices. Easy to see that the continuous variables given the discrete ones follow a multivariate normal distribution. In LAURITZEN and WERMUTH (1984) a mixed-interaction-model is represented by a graph. A vertex of the graph corresponds to a variable. Connections between vertices by line segments have the following meaning. If for any pair of vertices, not directly connected with each other, the corresponding variables are conditionally independent given the remaining variables, then the model has the G-Markov property. Interesting is that a conditional Gaussian distribution has the G-Markov property if and only if it is G-Gibbsian. This latter property means that the logarithm of the probability function has an expansion in which the index runs over those subsets of the discrete variables whose

corresponding vertices are all directly connected with each other. The functions used in such an expansion depend only on the variables of the index element. For the conditional Gaussian distributions this becomes, for example

$$\log f(i,y) = \sum_{d \subseteq \Delta} \phi_d(i) + \sum_{d \subseteq \Delta} \eta_d(i)^T y - \frac{1}{2} \sum_{d \subseteq \Delta} y^T \Psi_d(i) y$$

where Δ is the set of all discrete variables. See also DARROCH, LAURITZEN, and SPEED (1980), EDWARDS (1986), and GILL (1985). A recent publication about this and related subjects is WERMUTH and LAURITZEN (1987).

2.2. MAIN LINE OF THE DERIVATION

A central role in the derivation of the asymptotic distribution of the estimator of the vector of posterior probabilities is played by the "δ-method" which is formulated in lemma 2.2.1. The dispersion matrix of the asymptotic distribution of the estimator of the vector of posterior probabilities is given in (2.1.2) as the product of three matrices. This is a consequence of the δ-method.

LEMMA 2.2.1
If

$$\mathcal{L} n^{1/2}(Y_n - \eta) \to N_p(0,\Sigma)$$

for some sequence of p-dimensional random vectors Y_n, $n = 1,2,...$ *and* $g = (g_1,...,g_q)^T$: $\mathbb{R}^p \to \mathbb{R}^q$ *is differentiable in* η *while*

$$\frac{\partial}{\partial x} = \left(\frac{\partial}{\partial x_1},...,\frac{\partial}{\partial x_p}\right)^T$$

and

$$\nabla_{g(\eta)} = \left(\frac{\partial}{\partial x} g_1(\eta),...,\frac{\partial}{\partial x} g_q(\eta)\right)^T$$

then

$$\mathcal{L} n^{1/2}(g(Y_n) - g(\eta)) \to N_q(0, \nabla_{g(\eta)} \Sigma (\nabla_{g(\eta)})^T).$$

This lemma can be found, for example, in SERFLING (1980), section 3.3, theorem A. Application of this lemma in the derivation of the asymptotic distribution of the estimator of posterior probabilities for the cases of the following sections, will be along the following lines. First, the asymptotic distribution of the estimator of the vector θ of unknown parameters is derived. Let $\hat{\theta}_n$ be the estimator of θ and let $n = n_1 + \cdots + n_k$. We shall see that, if $n_h/n \to b_h \in (0,\infty)$, $h = 1,...,k$, in all cases

$$\mathcal{L} n^{1/2}(\hat{\theta}_n - \theta) \to N_k(0,\Lambda) \tag{2.2.1}$$

where Λ depends on the model specified. The next step is the derivation of the asymptotic distribution of the estimator of the k logarithms of the Radon-

Nikodym derivatives in the score vector y of the individual under investigation by applying lemma 2.2.1. Let $\log f_\theta(y)=(\log f_{1,\theta}(y),\ldots,\ \log f_{k,\theta}(y))^T$ and let $\log \hat{f}_n(y)$ be the estimator of the vector $\log f_\theta(y)$, then

$$\mathcal{L}n^{1/2}(\log \hat{f}_n(y)-\log f_\theta(y)) \to N_k(0,M) \tag{2.2.2}$$

where M depends on the model specified. M can be written as the product of three matrices, say $M=A\Lambda A^T$, where Λ is given in (2.2.1) and A is the matrix of partial derivatives of $\log f_\theta(y)$ in the point θ. An exception to this step is the derivation to be given in section 2.4 for the case of equality of dispersion matrices. In that case the asymptotic distribution of the estimator of the vector of squared Mahalanobis distances $\Delta^2_{x,t}$, $t=1,\ldots,k$ is derived, instead of the asymptotic distribution of the estimator of the vector of the logarithms of densities. See also the remark under *ad*(2) in the previous section. The third step is the derivation of the asymptotic distribution of the estimator $R_{\cdot|y}$ of the vector of posterior probabilities $\rho_{\cdot|y}$. With lemma 2.2.1 we get

$$\mathcal{L}n^{1/2}(R_{\cdot|y}-\rho_{\cdot|y}) \to N_k(0,\Psi M\Psi) \tag{2.2.3}$$

where M is given in (2.2.2) and Ψ is the matrix given in (2.1.3) and (2.1.4).

There are some advantages in the availability of the asymptotic distribution of the estimators of the logarithm of the Radon-Nikodym derivatives in (2.2.2). Suppose that the set of component variables of the random vector Y is partitioned into s subsets which henceforth are regarded as mutually independent. Let $y=(y^{(1)},\ldots,y^{(s)})$ denote the score vetor, in which $y^{(j)}$ is the score vector of the j-th subset. The postulated independence implies that

$$f_{h,\theta}(y) = \prod_{j=1}^{s} f^{(j)}_{h,\theta}(y^{(j)})$$

and

$$\log f_{h,\theta}(y) = \sum_{j=1}^{s} \log f^{(j)}_{h,\theta}(y^{(j)})$$

where $f^{(j)}_{h,\theta}$ is the Radon-Nikodym derivative which belongs to the j-th subset of variables. The independence between the subsets implies that the dispersion matrix of the asymptotic distribution of

$$n^{1/2}(\log \hat{f}_n(y)-\log f_\theta(y))$$

is the sum of the dispersion matrices of the asymptotic distribution of the

$$n^{1/2}(\log \hat{f}^{(j)}_n(y^{(j)})-\log f^{(j)}_\theta(y^{(j)}))$$

$j=1,\ldots,s$. This obvious but useful result has been implemented in the POSCON computer program to be discussed in section 5.2. More precisely, it is the basis of the incorporated update system: as the log-density estimates and the corresponding dispersion matrices are sums, a recursive implementation is allowed.

2.3. DISCRETE CASE

In this section the asymptotic distribution of the estimator of the vector of the k posterior probabilities will be derived in the case that there are only discrete variables. Let the discrete variables be recoded into one discrete variable with d cells or categories as possible outcomes. Let $P_{h\ell}$ be the probability mass in cell ℓ of population $h, \ell=1,...,d; h=1,...,k$. The object or individual for which we want to compute the posterior probabilities has such scores on the original discrete variables that in terms of the recoded variable its outcome is in cell j. Let D denote the recoded variable which generates this observation, then

$$P(D=j|T=t)=p_{tj} \quad t=1,...,k; \quad j=1,...,d$$

where T is the variable which describes the number of the population the observation comes from. This T has distribution

$$P(T=t)=\rho_t \quad t=1,...,k.$$

These ρ_t's are the prior probabilities. The posterior probabilities are given by

$$\begin{aligned}\rho_{t|j} &= P(T=t|D=j) = P(T=t,D=j)/\sum_{h=1}^{k} P(T=h,\ D=j) \\ &= P(T=t)P(D=j|T=t)/\sum_{h=1}^{k} P(T=h)P(D=j|T=h) \\ &= \rho_t p_{tj}/\sum_{h=1}^{k} \rho_h p_{hj} \quad t=1,...,k. \end{aligned} \tag{2.3.1}$$

We shall use the notation $\rho_{\cdot|j}=(\rho_{1|j},...,\rho_{k|j})^T$.

Estimators of the posterior probabilities are function of the independently distributed random variables

$$D_{11},...,D_{1n_1},...,D_{k1},...,D_{kn_k}$$

where

$$P(D_{hj}=\ell) = p_{h\ell} \quad \ell=1,...,d; \ h=1,...,k; \quad j=1,...,n_h.$$

We shall use the notation $R_{t|j}$ for the estimator of $\rho_{t|j}$, $t=1,...,k$ and write $R_{\cdot|j}=(R_{1|j},...,R_{k|j})^T$. Define

$$N_{h\ell} = \sum_{i=1}^{n_h} I\{D_{hi}=\ell\} \quad h=1,...,k; \quad \ell=1,...,d,$$

for the number of observations in cell ℓ of training sample h. Write

$$\hat{p}_{h\ell} = n_h^{-1} N_{h\ell} \quad h=1,...,k; \ \ell=1,...,d,$$

and let the estimator of the posterior probabilities be defined by

$$R_{t|j} = \rho_t \hat{p}_{tj}/\sum_{h=1}^{k} \rho_h \hat{p}_{hj} \quad t=1,...,k. \tag{2.3.2}$$

The variables $N_{h\ell}$, $h=1,...,k$; $\ell=1,...,d$; $d<\infty$, are components of the random vector $(N_{h1},...,N_{hd})$ following the multinomial distribution $M(n_h;p_{h1},...,p_{hd})$, $h=1,...,d$. The expectations, variances and covariances are given by

$$EN_{h\ell} = p_{h\ell}n_h$$
$$VARN_{h\ell} = p_{h\ell}(1-p_{h\ell})n_h$$
$$COV(N_{h\ell},N_{hm}) = -p_{h\ell}p_{hm}n_h$$

with as a consequence that

$$E\hat{p}_{h\ell} = p_{h\ell}$$
$$VAR\hat{p}_{h\ell} = p_{h\ell}(1-p_{h\ell})n_h^{-1}$$
$$COV(\hat{p}_{h\ell},\hat{p}_{hm}) = -p_{h\ell}p_{hm}n_h^{-1}$$

where, of course, $h=1,...,k$; $\ell=1,...,d$, and $m=1,...,d$. For n_h tending to infinity the following result can be formulated for $h=1,...,k$

$$\mathscr{L}n_h^{1/2}\begin{bmatrix}\hat{p}_{h1}-p_{h1}\\ \cdot\\ \cdot\\ \cdot\\ \hat{p}_{hd}-p_{hd}\end{bmatrix} \to N_d(0,V_h) \tag{2.3.3}$$

where V_h is defined by

$$V_{h,tt} = p_{ht}(1-p_{ht}) \qquad t=1,...,d$$

and

$$V_{h,ts} = -p_{hs}p_{ht} \qquad t,s=1,...,d;\ t\neq s.$$

See, for example, CRAMÉR (1946) p. 318 or BISHOP et al. (1975) p. 470. It can be proved by establishing convergence of characteristic or moment generating functions. Another way to prove result (2.3.3) is by means of the multivariate central limit theorem to be formulated in section 2.4, see (2.4.8). From (2.3.3) and the independence between $\hat{p}_{hi}$, $h=1,...,k$ for i fixed, it follows that

$$\mathscr{L}n^{1/2}\begin{bmatrix}\hat{p}_{1j}-p_{1j}\\ \cdot\\ \cdot\\ \cdot\\ \hat{p}_{kj}-p_{kj}\end{bmatrix} \to N_k(0,W) \tag{2.3.4}$$

where W is the diagonal matrix defined by

$$W_{tt} = b_t^{-1}p_{tj}(1-p_{tj}) \qquad t=1,...,k$$

and where $n_t/n \to b_t$, $t=1,...,k$, $n=n_1+\cdots+n_k$ and, as indicated earlier, j is the cell containing the score of the individual under investigation. With lemma 2.2.1 we obtain

$$\mathscr{L}n^{1/2}\begin{bmatrix}\log\hat{p}_{1j} - \log p_{1j}\\ \vdots \\ \log\hat{p}_{kj} - \log p_{kj}\end{bmatrix} \to N_k(0,M) \tag{2.3.5}$$

where M is the diagonal matrix specified by

$$M_{tt} = b_t^{-1}p_{tj}^{-1}(1-p_{tj}) \quad t=1,...,k. \tag{2.3.6}$$

Once again applying the δ-method (lemma 2.2.1) we obtain

THEOREM 2.3.1.

$$\mathscr{L}n^{1/2}(R_{\cdot|j}-\rho_{\cdot|j}) \to N_k(0,\Psi M\Psi)$$

where M is defined in (2.3.6) and Ψ in (2.1.3) and (2.1.4) (y has been replaced by j).

2.4. CONTINUOUS CASE

In this section we shall derive the asymptotic distribution of the estimator of the vector of posterior probabilities if the k populations have multivariate normal densities in $\mathbb{R}^p$. Two different cases will be considered. Namely, (A) the case with the assumption of equality of dispersion matrices, and (B) the case in which no assumption about the dispersion matrices is made. Moreover, we shall give two different approaches for the first case. As a consequence we shall consider the derivations of the approaches A.1, A.2, and B, supresent But first we shall present some theory for each of the three approaches.

Let x be the vector of scores for the p variables of the individual under investigation. We shall consider x as the realisation of the random variable X with distribution

$$X|T=t \sim N_p(\mu_t,\Sigma_t) \quad t=1,...,k$$

and use the postulated values

$$P(T=t) = \rho_t \quad t=1,...,k$$

for the prior probabilities. The posterior probabilities are given by (see (2.1.1))

$$\rho_{t|x} = \rho_t f_t(x)/\sum_{h=1}^{k}\rho_h f_h(x) \quad t=1,...,k \tag{2.4.1}$$

where the value $f_h(x)$ of the density of the h-th population at the vector x is

given by

$$f_h(x) = |2\pi\Sigma_h|^{-1/2}\exp(-\tfrac{1}{2}\Delta^2_{x;h}) \tag{2.4.2}$$

in which

$$\Delta^2_{x;h} = (x-\mu_h)^T\Sigma_h^{-1}(x-\mu_h). \tag{2.4.3}$$

For the vector of the k posterior probabilities we shall use the notation $\rho_{\cdot|x}=(\rho_{1|x},...,\rho_{k|x})^T$.

Estimators of the posterior probabilities are functions of the independently distributed random vectors

$$X_{11},...,X_{1n_1},...,X_{k1},...,X_{kn_k}$$

in which $X_{hi}\sim N_p(\mu_h, \Sigma_h)$, $h=1,...,k$; $i=1,...,n_h$. We shall use the notation $R_{t|x}$ for the estimator of $\rho_{t|x}$, $t=1,...,k$ and write $R_{\cdot|x}=(R_{1|x},...,R_{k|x})^T$. The sample mean vector $\hat{\mu}_h$ and the matrix of the corrected sums of squares and cross products S_h are defined by

$$\hat{\mu}_h = n_h^{-1}\sum_{i=1}^{n_h} X_{hi} \qquad h=1,...,k \tag{2.4.4}$$

and

$$S_h = \sum_{i=1}^{n_h}(X_{hi}-\hat{\mu}_h)(X_{hi}-\hat{\mu}_h)^T \qquad h=1,...,k \tag{2.4.5}$$

respectively. They are distributed as follows

$$\hat{\mu}_h \sim N_p(\mu_h, n_h^{-1}\Sigma_h) \tag{2.4.6}$$

and

$$S_h \sim W_p(n_h-1,\Sigma_h) \tag{2.4.7}$$

in which W is the notation for a Wishart distribution. Definition and properties of a Wishart distribution are given in, for example, RAO (1973) p. 533, ANDERSON (1958), p. 154 and MUIRHEAD (1982), p. 85.

A theorem which lies at the basis of many or our results is the multivariate central limit theorem. It can be formulated as follows (see RAO (1973), p. 128). If $Y_1,...,Y_n$ are independent identically distributed (i.i.d.) random variables which assume values in $\mathbb{R}^p$ with $EY_i=\mu$ and $VARCOV(Y_i)=T$ for $i=1,...,n$, then

$$\mathscr{L}n^{1/2}(n^{-1}\sum_{i=1}^{n} Y_i-\mu) \to N_p(0,T). \tag{2.4.8}$$

If we apply this theorem to the estimator $\hat{\mu}_h$ defined in (2.4.4) we obtain

$$\mathscr{L}n_h^{1/2}(\hat{\mu}_h-\mu_h) \to N_p(0,\Sigma_h) \tag{2.4.9}$$

where in fact the equality sign is valid because of result (2.4.6).

The asymptotic distribution of $\hat{\Sigma}_h=(n_h-1)^{-1}S_h$ can also be derived. We shall use a slightly more general notation by replacing n_h-1 by f and Σ_h by Σ. Let A be a $m\times n$ matrix. Write $A_{.j}$ for the j-th column of A, $j=1,...,n$. Then vec(A) is the $mn\times 1$ vector defined by

$$\text{vec}\,(A) = (A_{.1}^T,...,A_{.n}^T)^T.$$

This means that vec(A) is obtained by placing the columns of A below each other.

If $f\hat{\Sigma}\sim W_p(f,\Sigma)$, then

$$\mathcal{L}f^{1/2}(\text{vec}(\hat{\Sigma})-\text{vec}(\Sigma)) \to N_{p^2}(0,T) \tag{2.4.10}$$

where the $p^2\times p^2$ matrix T is specified by

$$T_{(j-1)p+i,(\ell-1)p+k}=\Sigma_{ik}\Sigma_{j\ell}+\Sigma_{i\ell}\Sigma_{jk}$$

for $i,j,k,\ell=1,...,p$.

PROOF: By definition of the Wishart distribution

$$\mathcal{L}f\hat{\Sigma} = \mathcal{L}\sum_{s=1}^{f} Z_sZ_s^T = \mathcal{L}\sum_{s=1}^{f} W_s$$

where $Z_1,...,Z_f$ are i.i.d. with $Z_s\sim N_p(0,\Sigma)$ and $W_1,...,W_f$ are i.i.d. with $W_s\sim W_p(1,\Sigma)$ for $s=1,...,f$. Thus

$$\mathcal{L}f^{1/2}(\text{vec}(\hat{\Sigma})-\text{vec}(\Sigma)) = \mathcal{L}f^{1/2}(f^{-1}\sum_{s=1}^{f}\text{vec}(W_s)-\text{vec}(\Sigma)).$$

By means of the multivariate central limit theorem, see (2.4.8), it is proven that the distribution at the right-hand side tends to $N_{p^2}(0,T)$. Therefore, note that

$$E\text{vec}(W_s) = \text{vec}(EW_s) = \text{vec}(EZ_sZ_s^T) = \text{vec}(\Sigma)$$

for $s=1,...,f$, and

$$\begin{aligned} &VARCOV(\text{vec}(W_s)) \\ &= E(\text{vec}(W_s)(\text{vec}(W_s))^T)-E\text{vec}(W_s)(E\text{vec}(W_s))^T \\ &= E(\text{vec}(Z_sZ_s^T)(\text{vec}(Z_sZ_s^T))^T)-E\text{vec}(Z_sZ_s^T)(E\text{vec}(Z_sZ_s^T))^T. \end{aligned}$$

The $((j-1)p+i,(\ell-1)p+k)$-th element of $VARCOV(\text{vec}(W_s))$ is

$$\begin{aligned} &EZ_{s,i}Z_{s,j}Z_{s,k}Z_{s,\ell}-EZ_{s,i}Z_{s,j}EZ_{s,k}Z_{s,\ell} \\ &= \Sigma_{ij}\Sigma_{k\ell} + \Sigma_{ik}\Sigma_{j\ell} + \Sigma_{i\ell}\Sigma_{jk}-\Sigma_{ij}\Sigma_{k\ell} \\ &= \Sigma_{ik}\Sigma_{j\ell} + \Sigma_{i\ell}\Sigma_{jk} \end{aligned}$$

for $i,j,k,\ell=1,...,p$ and $s=1,...,f$. This completes the proof of (2.4.10).

Now, we shall present some definitions and notations about matrices. Some

properties, which will be used frequently later on, are also given. MAGNUS and NEUDECKER (1979, 1986, 1988), and MUIRHEAD (1982) are suitable references.

Let A be a $m \times n$ and B a $s \times t$ matrix. The Kronecker product $A \otimes B$ is the $ms \times nt$ matrix with $a_{ij}B$ as its (i,j)-th submatrix, $i = 1,...,m$ and $j = 1,...,n$.

The following properties are useful. It is assumed that the size of the matrices is such that the sums and products exist.

$$
\begin{aligned}
&(a) \quad (A_1 \otimes B_1)(A_2 \otimes B_2) = A_1 A_2 \otimes B_1 B_2 \\
&(b) \quad (A \otimes B)^T = A^T \otimes B^T \\
&(c) \quad (A \otimes B) \otimes C = A \otimes (B \otimes C) \\
&(d) \quad (A + B) \otimes C = (A \otimes C) + (B \otimes C) \\
&(e) \quad A \otimes (B + C) = (A \otimes B) + (A \otimes C) \\
&(f) \quad A \otimes 1 = 1 \otimes A = A \\
&(g) \quad (\alpha A) \otimes (\beta B) = \alpha\beta (A \otimes B), \alpha \text{ and } \beta \text{ scalars.}
\end{aligned}
\tag{2.4.11}
$$

If A is a $m \times m$ matrix then trace (A) is defined by

$$
\operatorname{trace}(A) = \sum_{i=1}^{m} A_{ii}.
$$

If the matrix A has size $m \times n$, B $n \times m$, C $s \times t$, D $t \times u$, and E $u \times v$, then

$$
\begin{aligned}
&(a) \quad \operatorname{trace}(AB) = \operatorname{vec}^T(A^T)\operatorname{vec}(B) \\
&(b) \quad \operatorname{vec}(CDE) = (E^T \otimes C)\operatorname{vec}(D).
\end{aligned}
\tag{2.4.12}
$$

Let $e_i = \mathrm{I}_{\cdot i}$ denote the i-th column of the identity matrix I of dimension p. We define

$$
E_{ij} = e_i e_j^T \quad \text{and} \quad K_p = \sum_{i=1}^{p} \sum_{j=1}^{p} E_{ij} \otimes E_{ji}.
$$

For the $p \times p$ matrices A and B we have

$$
\begin{aligned}
&(a) \quad K_p \operatorname{vec}(A) = \operatorname{vec}(A^T) \\
&(b) \quad K_p = K_p^T \\
&(c) \quad K_p (A \otimes B) = (B \otimes A) K_p \\
&(d) \quad \operatorname{vec}(I)\operatorname{vec}^T(I) = \sum_{i=1}^{p} \sum_{j=1}^{p} E_{ij} \otimes E_{ij}.
\end{aligned}
\tag{2.4.13}
$$

With the above definitions the convergence result (2.4.10) can also be expressed as follows. If $f\hat{\Sigma} \sim W_p(f, \Sigma)$ then

$$
\mathscr{L} f^{1/2}(\operatorname{vec}(\hat{\Sigma}) - \operatorname{vec}(\Sigma)) \to N_{p^2}(0, (I_{p^2} + K_p)(\Sigma \otimes \Sigma)) \tag{2.4.14}
$$

(see also MUIRHEAD (1982), p. 90). MAGNUS and NEUDECKER (1980) define $v(\Sigma) = L_p \operatorname{vec}(\Sigma)$ where L_p is the linear transformation from $\mathbb{R}^{p \times p}$ to $\mathbb{R}^{\frac{1}{2}p(p+1)}$ which eliminates those elements of $\operatorname{vec}(\Sigma)$ which originate form the supradiagonal part of Σ. Using this, (2.4.14) can be written as

$$\mathscr{L}f^{1/2}(v(\hat{\Sigma})-v(\Sigma)) \to N_{\frac{1}{2}p(p+1)}(0,\tfrac{1}{2}L_p(I_{p^2}+K_p)(\Sigma\otimes\Sigma)(I_{p^2}+K_p)L_p^T)$$

where, of course, the covariance matrix is equal to $L_p(I_{p^2}+K_p)(\Sigma\otimes\Sigma)L_p^T$. The advantage is that the covariance matrix is non-singular. However, we shall use the result (2.4.14).

A.1. Equality of dispersion matrices

It is assumed that the k dispersion matrices are equal, i.e. $\Sigma_1 = \cdots = \Sigma_k = \Sigma$. Hence, in the formula of the posterior probabilities (2.4.1) the factors $|2\pi\Sigma|^{-1/2} = (2\pi)^{-p/2}|\Sigma|^{-1/2}$ can be cancelled. This implies that

$$\rho_{t|x} = \rho_t \exp(-\tfrac{1}{2}\Delta_{x;t}^2) / \sum_{h=1}^{k} \rho_h \exp(-\tfrac{1}{2}\Delta_{x;h}^2) \tag{2.4.14}$$

where

$$\Delta_{x;h}^2 = (x-\mu_h)^T\Sigma^{-1}(x-\mu_h) \quad h=1,\ldots,k. \tag{2.4.15}$$

For the estimation of the common dispersion matrix Σ the observations of each of the k populations can be used. So, let us define

$$S = \sum_{h=1}^{k} S_h$$

where S_h is defined in (2.4.5), then

$$S \sim W_p(\sum_{h=1}^{k}(n_h-1),\ \Sigma)$$

(see Rao (1973), section 8b). Sometimes additional information is available in the form of extra independent observations which can be used for improving the estimate of Σ. Therefore we shall write $S \sim W_p(f,\Sigma)$ where $f = \Sigma_{h=1}^{k}(n_h-1)$ in most applications but larger values of f are also allowed. By replacing the parameters μ_h by $\hat{\mu}_h$, $h=1,\ldots,k$ and Σ by $\hat{\Sigma} = f^{-1}S$ in (2.4.15), the plug-in estimator

$$\hat{\Delta}_{x;h}^2 = (x-\hat{\mu}_h)^T(\hat{\Sigma})^{-1}(x-\hat{\mu}_h) \tag{2.4.16}$$

of $\Delta_{x;h}^2$, $h=1,\ldots,k$ is obtained. Next, by replacing $\Delta_{x;h}^2$ by $\hat{\Delta}_{x;h}^2$ in (2.4.14) the plug-in estimators $R_{t|x}$ of $\rho_{t|x}$, $t=1,\ldots,k$, and $R_{\cdot|x}$ or $\rho_{\cdot|x}$ are obtained. Let us define

$$\Delta_x^2 = (\Delta_{x;1}^2,\ldots,\Delta_{x;k}^2)^T$$

and

$$\hat{\Delta}_x^2 = (\hat{\Delta}_{x;1}^2,\ldots,\hat{\Delta}_{x;k}^2)^T.$$

Now, the asymptotic distribution of $n^{1/2}(R_{\cdot|x}-\rho_{\cdot|x})$ will be derived by means of the δ-method (lemma 2.2.1) from the asymptotic distribution of

$n^{1/2}(\hat{\Delta}^2_{x;h}-\Delta^2_{x;h})$. This latter distribution will be derived in its turn from the $k+1$ independent asymptotic distributional results

$$\mathscr{L}n^{1/2}(\hat{\mu}_h-\mu_h) \to N_p(0,b_h^{-1}\Sigma) \quad h=1,...,k \tag{2.4.17}$$

and

$$\mathscr{L}f^{1/2}(\text{vec}(\hat{\Sigma})-\text{vec}(\Sigma)) \to N_{p^2}(0,(I_{p^2}+K_p)(\Sigma\otimes\Sigma)) \tag{2.4.18}$$

by applying the δ-method. For that purpose and later derivations it is useful to have some matrix derivatives available. RAO (1973), ANDERSON (1958) and ROGERS (1980) are suitable references. If A is a $p\times p$ matrix and b a column vector of length p, then

$$\begin{aligned}
&(a) \quad \frac{\partial|A|}{\partial A} = |A|(A^{-1})^T \\
&(b) \quad \frac{\partial b^TAb}{\partial A} = bb^T \\
&(c) \quad \frac{\partial b^TAb}{\partial b} = 2Ab \quad , \quad A \text{ symmetric.}
\end{aligned} \tag{2.4.19}$$

Let Σ denote a covariance matrix of size $p\times p$, and x and μ column vectors of length p (see ROGERS (1980) p. 85). If $g=(x-\mu)^T\Sigma^{-1}(x-\mu)$ then

$$\begin{aligned}
&(a) \quad \frac{\partial g}{\partial \mu} = -2\Sigma^{-1}(x-\mu) \\
&(b) \quad \frac{\partial g}{\partial(\Sigma^{-1})} = (x-\mu)(x-\mu)^T \\
&(c) \quad \frac{\partial g}{\partial \Sigma} = -\Sigma^{-1}(x-\mu)(x-\mu)^T\Sigma^{-1}.
\end{aligned} \tag{2.4.20}$$

If $g=|\Sigma^{-1}|^{-r}$ then

$$\begin{aligned}
&(a) \quad \frac{\partial g}{\partial(\Sigma^{-1})} = -r|\Sigma^{-1}|^{-r}\Sigma \\
&(b) \quad \frac{\partial g}{\partial \Sigma} = r|\Sigma^{-1}|^{-r}\Sigma^{-1}.
\end{aligned} \tag{2.4.21}$$

If $g=\ln|\Sigma|$ then

$$\begin{aligned}
&(a) \quad \frac{\partial g}{\partial \Sigma} = \Sigma^{-1} \\
&(b) \quad \frac{\partial g}{\partial(\Sigma^{-1})} = -\Sigma.
\end{aligned} \tag{2.4.22}$$

Now, applying formula (2.4.20,a) and (2.4.20,c) to $\Delta^2_{x;h}$ we obtain

$$\frac{\partial\Delta^2_{x;h}}{\partial\mu_\ell} = -2\Sigma^{-1}(x-\mu_h)\delta_{h\ell} \qquad h,\ell=1,...,k$$

and

$$\frac{\partial \Delta^2_{x;h}}{\partial \text{vec}(\Sigma)} = -\Sigma^{-1}(x-\mu_h)\otimes\Sigma^{-1}(x-\mu_h) \qquad h=1,\ldots,k$$

where $\delta_{h\ell}=0$ if $h\neq\ell$ and $\delta_{h\ell}=1$ if $h=\ell$ and where vec $(bb^T)=b\otimes b$, $b\in\mathbb{R}^p$, has been used. This latter formula follows easily from the definitions of vec and $\otimes$, and also from formula (2.4.12,b). Results about vector and matrix differentiation, as in above formulas, can also be found in MAGNUS and NEUDECKER (1980, 1985, 1986, 1988).

If $n_h/n \to b_h>0$ and $n/f \to 1$ then

$$\mathscr{L}n^{1/2}\begin{bmatrix} \hat{\mu}_1-\mu \\ \cdot \\ \cdot \\ \cdot \\ \hat{\mu}_k-\mu_k \\ \text{vec}(\hat{\Sigma})-\text{vec}(\Sigma) \end{bmatrix} \to N_{kp+p^2}(0,B)$$

where B is the block-diagonal matrix specified by the blocks

$$b_1^{-1}\Sigma,\ldots,b_k^{-1}\Sigma, \quad (I_{p^2}+K_p)(\Sigma\otimes\Sigma).$$

Now, by using lemma 2.2.1 we obtain that

$$\mathscr{L}n^{1/2}(\hat{\Delta}^2_x-\Delta^2_x) \to N_k(0,T)$$

where $T=D+Q$ with D the diagonal matrix specified by

$$\begin{aligned} D_{tt} &= \left(\frac{\partial \Delta^2_x}{\partial \mu_t}\right)^T (b_t^{-1}\Sigma)\left(\frac{\partial \Delta^2_x}{\partial \mu_t}\right) \\ &= (-2(x-\mu_t)^T\Sigma^{-1})(b_t^{-1}\Sigma)(-2\Sigma^{-1}(x-\mu_t)) \\ &= 4b_t^{-1}\Delta^2_{x;t} \end{aligned}$$

and Q specified by

$$\begin{aligned} Q_{st} &= \left(\frac{\partial \Delta^2_x}{\partial \text{vec}(\Sigma)}\right)^T (I_{p^2}+K_p)(\Sigma\otimes\Sigma)\left(\frac{\partial \Delta^2_x}{\partial \text{vec}(\Sigma)}\right) \\ &= (-(x-\mu_s)^T\Sigma^{-1}\otimes(x-\mu_s)^T\Sigma^{-1})(I_{p^2}+K_p) \\ &\quad \cdot(\Sigma\otimes\Sigma)(-\Sigma^{-1}(x-\mu_t)\otimes\Sigma^{-1}(x-\mu_t)) \\ &= (x-\mu_s)^T\Sigma^{-1}(x-\mu_t)(x-\mu_s)^T\Sigma^{-1}(x-\mu_t) \\ &\quad + ((x-\mu_s)^T\Sigma^{-1}\otimes(x-\mu_s)^T\Sigma^{-1})K_p(x-\mu_t)\otimes(x-\mu_t) \\ &= 2\{(x-\mu_s)^T\Sigma^{-1}(x-\mu_t)\}^2 , \end{aligned}$$

where we have used that $K_p(b\otimes b)=K_p$ vec $(bb^T)=\text{vec}(bb^T)=b\otimes b$.

So, the variance-covariance matrix T can be written, with $\Delta^4_{x;t}=(\Delta^2_{x;t})^2$, as

$$T_{tt} = 4b_t^{-1}\Delta^2_{x;t} + 2\Delta^4_{x;t} \qquad t=1,\ldots,k$$
$$T_{ts} = 2\{(x-\mu_s)^T\Sigma^{-1}(x-\mu_t)\}^2 \qquad t,s=1,\ldots,k;\quad t\neq s.$$

Finally, the asymptotic distribution of the estimator of the vector of posterior probabilities is given in the following theorem.

THEOREM 2.4.1.

$$\mathscr{L}n^{1/2}(R_{\cdot|x}-\rho_{\cdot|x}) \to N_k(0,\Psi M\Psi)$$

where Ψ has been specified in (2.1.3) and (2.1.4) and matrix M by

$$M_{tt} = b_t^{-1}\Delta^2_{x;t} + \tfrac{1}{2}\Delta^4_{x;t} \qquad t=1,\ldots,k$$

$$M_{ts} = \tfrac{1}{2}\{(x-\mu_s)^T\Sigma^{-1}(x-\mu_t)\}^2 \qquad t,s=1,\ldots,k;\quad t\neq s.$$

A.2. ALTERNATIVE DERIVATION IN THE CASE OF EQUALITY OF DISPERSION MATRICES

It is also possible to derive the asymptotic distribution of the estimator of posterior probabilities in the case of equality of dispersion matrices without cancelling the factors $(2\pi)^{-p/2}|\Sigma|^{-1/2}$, in contrast with (2.4.14). The motivation for this extra derivation is that the intermediate results concerning the asymptotic distributions of $n^{1/2}(\hat{f}_h(x)-f_h(x))$ and $n_h^{1/2}(\log\hat{f}_h(x)-\log f_h(x))$ are interesting by themselves and did not appear in the foregoing derivation. In addition, assuming $\Sigma_1=\cdots=\Sigma_k$, it is interesting to compare the variance of these asymptotic distributions with the corresponding variance which will be derived in case B where no assumption about the dispersion matrices is made. See further remark after theorem 2.4.4.

The estimators $R_{t|x}$, $t=1,\ldots,k$ are functions of the estimators $\hat{f}_h(x)$ of the densities $f_h(x),h=1,\ldots,k$. We get

$$R_{t|x} = \rho_t\hat{f}_t(x)/\sum_{h=1}^{k}\rho_h\hat{f}_h(x) \qquad t=1,\ldots,k$$

where $\hat{f}_h(x)$, $h=1,\ldots,k$ are obtained by replacing in $f_h(x)$ the μ_h and Σ by $\hat{\mu}_h$ and $\hat{\Sigma}$, respectively, where $\hat{\Sigma}=f^{-1}S$ and

$$f_h(x) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-\tfrac{1}{2}(x-\mu_h)^T\Sigma^{-1}(x-\mu_h)\}$$

for $h=1,\ldots,k$. By using the results of (2.4.19) and (2.4.20) the following derivatives are obtained

$$\frac{\partial f_h(x)}{\partial\mu_\ell} = f_h(x)\Sigma^{-1}(x-\mu_h)\delta_{h\ell} \qquad h,\ell=1,\ldots,k \tag{2.4.23}$$

$$\frac{\partial f_h(x)}{\partial \text{vec}(\Sigma)} = f_h(x)\{-\tfrac{1}{2}\text{vec}(\Sigma^{-1}) \tag{2.4.24}$$

$$+\tfrac{1}{2}(\Sigma^{-1}(x-\mu_h))\otimes(\Sigma^{-1}(x-\mu_h))\}.$$

With the derivatives in (2.4.23) and (2.4.24) and with the asymptotic distributions of $\hat{\mu}_1,\ldots,\hat{\mu}_k$, and vec $(\hat{\Sigma})$ in (2.4.17) and (2.4.18), we can derive, using the δ-method of lemma 2.2.1, that

$$\mathscr{L}n^{1/2}\begin{bmatrix}\hat{f}_1(x)-f_1(x)\\ \cdot\\ \cdot\\ \cdot\\ \hat{f}_k(x)-f_k(x)\end{bmatrix} \to N_k(0,\Gamma) \tag{2.4.25}$$

where Γ is defined by

$$\Gamma_{tt} = (f_t(x))^2\{\tfrac{1}{2}p+(b_t^{-1}-1)\Delta^2_{x;t}+\tfrac{1}{2}\Delta^4_{x;t}\} \tag{2.4.26}$$

$$\Gamma_{st} = f_s(x)f_t(x)\{\tfrac{1}{2}p-\tfrac{1}{2}\Delta^2_{x;t}-\tfrac{1}{2}\Delta^2_{x;s} \tag{2.4.27}$$

$$+\tfrac{1}{2}\{(x-\mu_s)^T\Sigma^{-1}(x-\mu_t)\}^2\}$$

for $t, s=1,\ldots,k$; $t\neq s$ and $\Delta^2_{x;h}$ in (2.4.15).

PROOF. The dispersion matrix Γ can be written as $\Gamma=D+Q$ where D is a diagonal matrix. We have

$$D_{tt} = \left(\frac{\partial f_t(x)}{\partial \mu_t}\right)(b_t^{-1}\Sigma)\left(\frac{\partial f_t(x)}{\partial \mu_t}\right)$$

$$= (f_t(x)(x-\mu_t)^T\Sigma^{-1})(b_t^{-1}\Sigma)(f_t(x)\Sigma^{-1}(x-\mu_t))$$

$$= (f_t(x))^2 b_t^{-1}\Delta^2_{x;t} \qquad t=1,\ldots,k.$$

The matrix Q is specified by

$$Q_{st} = \left(\frac{\partial f_s(x)}{\partial \text{vec}(\Sigma)}\right)^T(I_{p^2}+K_p)(\Sigma\otimes\Sigma)\left(\frac{\partial f_t(x)}{\partial \text{vec}(\Sigma)}\right)$$

$$= f_s(x)\{-\tfrac{1}{2}\text{vec}^T(\Sigma^{-1})+\tfrac{1}{2}((x-\mu_s)^T\Sigma^{-1})\otimes((x-\mu_s)^T\Sigma^{-1})\}$$

$$\cdot (I_{p^2}+K_p)(\Sigma\otimes\Sigma)f_t(x)\{-\tfrac{1}{2}\text{vec}(\Sigma^{-1})$$

$$+\tfrac{1}{2}(\Sigma^{-1}(x-\mu_t))\otimes(\Sigma^{-1}(x-\mu_t))\}$$

$$= f_s(x)f_t(x)\{\tfrac{1}{4}\mathrm{vec}^T(\Sigma^{-1})(\Sigma\otimes\Sigma)\mathrm{vec}(\Sigma^{-1})$$

$$-\tfrac{1}{4}\mathrm{vec}^T(\Sigma^{-1})(\Sigma\otimes\Sigma)((\Sigma^{-1}(x-\mu_t))\otimes(\Sigma^{-1}(x-\mu_t)))$$

$$-\tfrac{1}{4}((x-\mu_s)^T\Sigma^{-1})\otimes((x-\mu_s)^T\Sigma^{-1})(\Sigma\otimes\Sigma)(\mathrm{vec}(\Sigma^{-1}))$$

$$+\tfrac{1}{4}((x-\mu_s)^T\Sigma^{-1})\otimes((x-\mu_s)\Sigma^{-1})$$

$$\cdot(\Sigma\otimes\Sigma)(\Sigma^{-1}(x-\mu_t)\otimes\Sigma^{-1}(x-\mu_t))$$

$$+ \text{ the last four terms again but with } \Sigma\otimes\Sigma \text{ replaced by } K_p(\Sigma\otimes\Sigma)\}.$$

Now, with (2.4.12.b) and (2.4.12.a) we find that

$$\mathrm{vec}^T(\Sigma^{-1})(\Sigma\otimes\Sigma)\mathrm{vec}(\Sigma^{-1}) = \mathrm{vec}^T(\Sigma^{-1})\mathrm{vec}(\Sigma) = \mathrm{trace}(I) = p.$$

Further, using (2.4.12.b)

$$\mathrm{vec}^T(\Sigma^{-1})((x-\mu_t)\otimes(x-\mu_t))$$
$$= \{((x-\mu_t)^T\otimes(x-\mu_t)^T)\mathrm{vec}(\Sigma^{-1})\}^T$$
$$= \{\mathrm{vec}((x-\mu_t)^T\Sigma^{-1}(x-\mu_t))\}^T = \Delta^2_{x;t}.$$

Hence

$$Q_{st} = f_s(x)f_t(x)\{\tfrac{1}{4}p-\tfrac{1}{4}\Delta^2_{x;t}-\tfrac{1}{4}\Delta^2_{x;s}+\tfrac{1}{4}\{(x-\mu_s)^T\Sigma^{-1}(x-\mu_t)\}^2$$

$$+ \text{ the four terms mentioned above}\}.$$

Now, with (2.4.13.c) and (2.4.13.a) we get

$$K_p(\Sigma\otimes\Sigma)\mathrm{vec}(\Sigma^{-1}) = (\Sigma\otimes\Sigma)K_p\mathrm{vec}(\Sigma^{-1}) = (\Sigma\otimes\Sigma)\mathrm{vec}(\Sigma^{-1})$$

and with (2.4.13.b) and (2.4.13.a) we derive that

$$\mathrm{vec}^T(\Sigma^{-1})K_p = (K_p^T\mathrm{vec}(\Sigma^{-1}))^T = (K_p\mathrm{vec}(\Sigma^{-1}))^T = \mathrm{vec}^T(\Sigma^{-1}).$$

Further, $K_p(b\otimes b)=K_p\mathrm{vec}(bb^T)=\mathrm{vec}(bb^T)=b\otimes b$. Hence, the last four terms in Q_{st} with $K_p(\Sigma\otimes\Sigma)$ are the same as the four terms with $\Sigma\otimes\Sigma$. So that

$$Q_{st} = f_s(x)f_t(x)\{\tfrac{1}{2}p-\tfrac{1}{2}\Delta^2_{x;t}-\tfrac{1}{2}\Delta^2_{x;s}+\tfrac{1}{2}\{(x-\mu_s)^T\Sigma^{-1}(x-\mu_t)\}^2\}$$

for $s,t=1,\ldots,k$.
The matrix Γ is obtained by putting $\Gamma_{tt}=D_{tt}+Q_{tt}$, $t=1,\ldots,k$ and $\Gamma_{st}=Q_{st}$, $t,s=1,\ldots,k$; $s\neq t$. This completes the proof of (2.4.25).

For the logarithm of the estimator of the density we can easily derive that

$$\mathscr{L}n^{1/2}\begin{bmatrix}\log\hat{f}_1(x)-\log f_1(x)\\ \cdot\\ \cdot\\ \cdot\\ \log\hat{f}_L(x)-\log f_L(x)\end{bmatrix} \to N_k(0,M) \qquad (2.4.28)$$

where M is defined by

$$M_{tt} = \tfrac{1}{2}p + (b_t^{-1}-1)\Delta_{x;t}^2 + \tfrac{1}{2}\Delta_{x;t}^4 \tag{2.4.29}$$

$$M_{ts} = \tfrac{1}{2}p - \tfrac{1}{2}\Delta_{x;t}^2 - \tfrac{1}{2}\Delta_{x;s}^2 + \tfrac{1}{2}\{(x-\mu_s)^T\Sigma^{-1}(x-\mu_t)\}^2. \tag{2.4.30}$$

For the estimator of the vector of posterior probabilities we obtain

THEOREM 2.4.2.

$$\mathscr{L}n^{1/2}(R_{\cdot|x}-\rho_{\cdot|x}) \to N_k(0,\Psi M\Psi) \tag{2.4.31}$$

where Ψ is defined in (2.1.3) and (2.1.4) and M in (2.4.29) and (2.4.30), respectively.

REMARK. At first sight the dispersion matrices $\Psi M\Psi$ of the asymptotic distributions in theorems 2.4.1 and 2.4.2 look different because the M in theorem 2.4.2 has a few terms more than the M in theorem 2.4.1. However, the matrix of these extra terms premultiplied and postmultiplied by Ψ is zero. Hence the $\Psi M\Psi$ of theorem 2.4.1 and 2.4.2 are equal. See also the remark in section 2.1 under A.2.

B. NO ASSUMPTION ABOUT THE DISPERSION MATRICES

In this case there are k dispersion matrices Σ_h, $h=1,\ldots,k$ about which no further assumptions are made. The estimator for $\rho_{t|x}$ is

$$R_{t|x} = \rho_t\hat{f}_t(x)/\sum_{h=1}^{k}\rho_h\hat{f}_h(x) \quad t=1,\ldots,k$$

where $\hat{f}_h(x)$, $h=1,\ldots,k$ are the estimators for the densities

$$f_h(x) = (2\pi)^{-p/2}|\Sigma_h|^{-1/2}\exp\{-\tfrac{1}{2}(x-\mu_h)^T\Sigma_h^{-1}(x-\mu_h)\}$$

$h=1,\ldots,k$ and which are obtained by replacing μ_h by $\hat{\mu}_h$ and Σ_h by $f_h^{-1}S_h$ where $f_h=n_h-1$. For these estimators we have

$$\mathscr{L}n_h^{1/2}(\hat{\mu}_h-\mu_h) \to N_p(0,\Sigma_h) \quad h=1,\ldots,k \tag{2.4.32}$$

and

$$\mathscr{L}n_h^{1/2}(\mathrm{vec}(\hat{\Sigma}_h)-\mathrm{vec}(\Sigma_h)) \to N_{p^2}(0,(I_{p^2}+K_p)(\Sigma_h\otimes\Sigma_h)) \tag{2.4.33}$$

and these $k+1$ estimators are independent of each other.

LEMMA 2.4.3.

$$\mathscr{L}n_h^{1/2}(\hat{f}_h(x)-f_h(x)) \to N(0,\tfrac{1}{2}(p+\Delta_{x;h}^4)(f_h(x))^2) \tag{2.4.34}$$

and

$$\mathscr{L}n_h^{1/2}(\log\hat{f}_h(x)-\log f_h(x)) \to N(0,\tfrac{1}{2}(p+\Delta^4_{x;h})) \tag{2.4.35}$$

for $h=1,\ldots,k$.

PROOF. The asymptotic variance in (2.4.34) is given by

$$(\frac{\partial f_h(x)}{\partial \mu_h})^T(\Sigma_h)(\frac{\partial f_h(x)}{\partial \mu_h})^T+(\frac{\partial f_h(x)}{\partial \mathrm{vec}(\Sigma_h)})^T(I_{p^2}+K_p)(\Sigma_h\otimes\Sigma_h)(\frac{\partial f_h(x)}{\partial \mathrm{vec}(\Sigma_h)})$$

where we have used the δ-method of lemma 2.2.1 and the asymptotic variances in (2.4.32) and (2.4.33). The derivatives are specified by

$$\frac{\partial f_h(x)}{\partial \mu_h} = f_h(x)\Sigma_h^{-1}(x-\mu_h)$$

and

$$\frac{\partial f_h(x)}{\partial \mathrm{vec}(\Sigma_h)} = f_h(x)\{-\tfrac{1}{2}\mathrm{vec}(\Sigma_h^{-1})+\tfrac{1}{2}(\Sigma_h^{-1}(x-\mu_h))\otimes(\Sigma_h^{-1}(x-\mu_h))\}.$$

The computation is straightforward. The second statement, (2.4.35), follows by once again using lemma 2.2.1.

Result (2.4.34) has been published earlier in AMBERGEN and SCHAAFSMA (1982, 1984). It will also be presented under (1) of theorem (4.5.1) of this thesis.

For the estimator of the vector of posterior probabilities we obtain

THEOREM 2.4.4.

$$\mathscr{L}n^{1/2}(R_{\cdot|x}-\rho_{\cdot|x}) \to N_k(0,\Psi M\Psi)$$

where M *is defined by*

$$M_{tt} = \tfrac{1}{2}b_t^{-1}(p+\Delta^4_{x;h}) \qquad t=1,\ldots,k$$

$$M_{ts} = 0 \qquad t,s=1,\ldots,k;\quad t\neq s$$

and Ψ *is given in (2.1.3) and (2.1.4).*

REMARK. A comparison between the covariance matrices of the asymptotic distributions of $n^{1/2}(\log\hat{f}_n(x)-\log f(x))$ in the cases A.2 and B if $\Sigma_1=\cdots=\Sigma_k$ is assumed, is of interest because it provides a check on the validity of the formulas. The matrix V of the differences between on the one hand (2.4.35) and on the other (2.4.29) and (2.4.30) is

$$V_{tt} = \tfrac{1}{2}(b_t^{-1}-1)\{(p-1)+(\Delta^2_{x;t}-1)^2\} \qquad t=1,\ldots,k$$

$$V_{ts} = -\tfrac{1}{2}\{(p-1)+(\Delta^2_{x;t}-1)(\Delta^2_{x;s}-1)+(\Delta^4_{x;ts}-\Delta^2_{x;t}\Delta^2_{x;s})\}$$

for $t,s=1,...,k$; $t\neq s$. We shall verify that V is positive definite. This result is obvious because the estimation method A.2 in which all observations are used for the estimation of the common dispersion matrix Σ, will have smaller variance for the estimation of $\Sigma_{t=1}^{k} a_t \log f_t(x)$, $a_t \in \mathbb{R}$, $t=1,...k$, than method B will have. The verification is given first for the case $p=1$, next for $p \geqslant 2$. Let $a=(a_1,...,a_k)^T$.

Case $p=1$. We have $\Delta_{x;ts}^4 - \Delta_{x;t}^2 \Delta_{x;s}^2 = 0$, hence

$$a^T V a = \tfrac{1}{2}\sum_t b_t^{-1}(\Delta_{x;t}^2 - 1)^2 a_t^2 - \tfrac{1}{2}(\sum_t (\Delta_{x;t}^2 - 1)a_t)^2.$$

With the Cauchy-Schwarz inequality we derive that

$$\begin{aligned}(\sum_t (\Delta_{x;t}^2 - 1)a_t)^2 &= (\sum_t (b_t^{1/2})(b_t^{-1/2}(\Delta_{x;t}^2 - 1)a_t))^2 \\ &\leqslant (\sum_t b_t)(\sum_t b_t^{-1}(\Delta_{x;t}^2 - 1)^2 a_t^2) \qquad (2.4.36) \\ &= \sum_t b_t^{-1}(\Delta_{x;t}^2 - 1)^2 a_t^2.\end{aligned}$$

Using this result we find $a^T V a \geqslant 0$.

Case $p \geqslant 2$. Define $V = C + D$ where

$$C_{tt} = \tfrac{1}{2}(p-2)(b_t^{-1} - 1) \qquad t=1,...,k$$

$$C_{ts} = -\tfrac{1}{2}(p-2) \qquad t,s=1,...,k; \quad t\neq s$$

and

$$D_{tt} = \tfrac{1}{2}(b_t^{-1} - 1)(\Delta_{x;t}^2 - 1)^2 + \tfrac{1}{2}(b_t^{-1} - 1) \qquad t=1,...,k$$

$$D_{ts} = -\tfrac{1}{2} - \tfrac{1}{2}(\Delta_{x;t}^2 - 1)(\Delta_{x;s}^2 - 1) + \tfrac{1}{2}\Delta_{x;t}^2 \Delta_{x;s}^2 \sin^2 \gamma_{t,s}$$

for $t,s=1,...,k$, $t\neq s$, where we use $\Delta_{x;ts}^4 = \Delta_{x;t}^2 \Delta_{x;s}^2 \cos^2 \gamma_{t,s}$ with $\gamma_{t,s}$ being the angle between $(x-\mu_t)^T \Sigma^{-1/2}$ and $(x-\mu_s)^T \Sigma^{-1/2}$. Now, we have that

$$a^T C a = \tfrac{1}{2}(p-2)\sum_t b_t^{-1} a_t^2 - \tfrac{1}{2}(p-2)(\sum_t a_t)^2.$$

Again, Cauchy-Schwarz gives that

$$\begin{aligned}(\sum_t a_t)^2 &= (\sum_t (b_t^{1/2})(b_t^{-1/2} a_t))^2 \\ &\leqslant (\sum_t b_t)(\sum_t b_t^{-1} a_t^2) \\ &= \sum_t b_t^{-1} a_t^2. \qquad (2.4.37)\end{aligned}$$

With this inequality we find $a^T Ca \geqslant 0$. In order to verify that $a^T Da \geqslant 0$, note that

$$\begin{aligned} a^T Da &= \tfrac{1}{2}\sum_t b_t^{-1}(\Delta_{x;t}^2 - 1)^2 a_t^2 - \tfrac{1}{2}(\sum_t (\Delta_{x;t}^2 - 1)a_t)^2 \\ &+ \tfrac{1}{2}\sum_t b_t^{-1} a_t^2 - \tfrac{1}{2}(\sum_t a_t)^2 \\ &+ \tfrac{1}{2}\sum_{t,s} \Delta_{x;t}^2 \Delta_{x;s}^2 a_t a_s \sin^2 \gamma_{t,s} \,. \end{aligned} \tag{2.4.38}$$

With the inequalities (2.4.36) and (2.4.37) we find

$$a^T Da \geqslant \tfrac{1}{2}\sum_{t,s} \Delta_{x;t}^2 \Delta_{x;s}^2 a_t a_s \sin^2 \gamma_{t,s} \,.$$

If the a_t's $t = 1,\ldots,k$ all have the same sign then $a^T Da \geqslant 0$. The situation remains that there are positive as well as negative a_t's. From now on, let t denote the index of positive and s the index of negative a_i's. Split the summations in (2.4.38) into a part with positive and a part with negative a_i's. Apply inequalities (2.4.36) and (2.4.37) to these parts. Use the lower bound $\Sigma_{t,s} \Delta_{x;t}^2 \Delta_{x;s}^2 a_t a_s$ for the last term of (2.4.38), take terms together, then

$$\begin{aligned} a^T Da &\geqslant \tfrac{1}{2}(\sum_s b_s)(\sum_t b_t^{-1}(\Delta_{x;t}^2 - 1)^2 a_t)^2 + \tfrac{1}{2}(\sum_t b_t)(\sum_s b_s^{-1}(\Delta_{x;s}^2 - 1)^2 a_s)^2 \\ &+ \tfrac{1}{2}(\sum_s b_s)(\sum_t b_t^{-1} a_t^2) + \tfrac{1}{2}(\sum_t b_t)(\sum_s b_s^{-1} a_s^2) \\ &+ \sum_{t,s}((\Delta_{x;t}^2 - 1) + (\Delta_{x;s}^2 - 1))a_t a_s \\ &= \tfrac{1}{2}\sum_{t,s} b_t^{-1} b_s^{-1}\{((\Delta_{x;t}^2 - 1)b_s a_t + b_t a_s)^2 + ((\Delta_{x;s}^2 - 1)b_t a_s + b_s a_t)^2\} \geqslant 0. \end{aligned}$$

So, it is established that V is a positive definite matrix. This reassures us that no computational errors are made.

2.5. BOTH CONTINUOUS AND DISCRETE VARIABLES

In this section the asymptotic distribution of the estimator of the vector of the k posterior probabilities will be derived in the case that there are both continuous and discrete variables. As mentioned in section 2.1 the discrete variables are combined into one discrete variable which can take on d different values. By these d different values each of the k populations is divided into d subpopulations. Let (h,l) be the l-th subpopulation of population h; $h = 1,\ldots,k$; $l = 1,\ldots,d$. The p continuous variables, given the discrete variable, follow a multivariate normal distribution, or in other words, the subpopulations can be associated with a multivariate normal density in $\mathbb{R}^p$. The score (x,j) on the

continuous and discrete variable of the individual or object we are interested in, is considered as a realisation of the random variable (X,D) which has distribution

$$\begin{aligned} X|D = \ell, T=t &\sim N_p(\mu_{t\ell}, \Sigma_{t\ell}) \quad \ell=1,...,d; \quad t=1,...,k \\ P(D=\ell|T=t) &= p_{t\ell} \qquad\qquad \ell=1,...,d; \quad t=1,...,k \\ P(T=t) &= \rho_t \qquad\qquad t=1,...,k \end{aligned}$$

where the ρ_t, $t=1,...,k$ are the given prior probabilities. The posterior probabilities are defined by

$$\begin{aligned} \rho_{t|(x,j)} &= P(T=t)f_t(x,j)/\sum_{h=1}^{k} P(T=h)f_h(x,j) \\ &= \rho_t p_{tj} f_{tj}(x)/\sum_{h=1}^{k} \rho_h p_{hj} f_{hj}(x) \qquad t=1,...,k \end{aligned}$$

where

$$f_h(x,\ell) = P(D=\ell|T=h)f_{h\ell}(x)$$

and

$$f_{h\ell}(x) = (2\pi)^{-p/2}|\Sigma_{h\ell}|^{-1/2}\exp\{-\tfrac{1}{2}(x-\mu_{h\ell})^T\Sigma_{h\ell}^{-1}(x-\mu_{h\ell})\}$$

for $\ell=1,...,d$; $h=1,...k$, which is the density of the (h,ℓ)-th subpopulation, i.e. the conditional density of X given $T=h, D=\ell$.
For the vector of the k posterior probabilities we shall use the notation

$$\rho_{\cdot|(x,j)} = (\rho_{1|(x,j)},...,\rho_{k|(x,j)})^T.$$

Estimators of the posterior probabilities are functions of the random variables

$$(X_{11},D_{11}),...,(X_{1n_1},D_{1n_1}),...,(X_{k1},D_{k1}),...,(X_{kn_k},D_{kn_k})$$

which yield the training samples. They are independent of each other and distributed according to

$$\begin{aligned} X_{hi}|D_{hi} = \ell &\sim N_p(\mu_{h\ell}, \Sigma_{h\ell}) \quad i=1,...,n_h; \quad h=1,...,k; \quad \ell=1,...,d \\ P(D_{hi}=\ell) &= p_{h\ell} \qquad\qquad i=1,...,n_h; \quad h=1,...,k; \quad \ell=1,...,d. \end{aligned}$$

Let us define the random variable

$$N_{hs} = \sum_{i=1}^{n_h} I(D_{hi}=s) \qquad h=1,...,k; \qquad s=1,...,d$$

wich gives the number of observations in cell s of sample h. Here I is the indicator function. Let n_{hs} denote the outcome of N_{hs}. Hence $n_{h1}+\cdots+n_{hd}=n_h$ where $n_1+\cdots+n_k=n$. We shall assume that $n_{hs}>1$, $h=1,...,k$, $s=1,...,d$ without explicitly mention this each time. Further, let us

define

$$\hat{p}_{hs} = n_h^{-1} N_{hs} \quad h=1,\ldots,k; \quad s=1,\ldots,d$$

as the estimator of p_{hs} and

$$\hat{\mu}_{hs} = N_{hs}^{-1} \sum_{i=1}^{n_h} I(D_{hi}=s) X_{hi}$$

and

$$\hat{\Sigma}_{hs} = (N_{hs}-1)^{-1} \sum_{i=1}^{n_h} I(D_{hi}=s)(X_{hi}-\hat{\mu}_{hs})(X_{hi}-\hat{\mu}_{hs})^T$$

as estimators for μ_{hs} and Σ_{hs}, respectively, for $h=1,\ldots,k$ and $\ell=1,\ldots,d$. The estimator $\hat{f}_{hj}(x)$, $h=1,\ldots,k$, where j is given by the score (x,j), is defined by replacing in $f_{hj}(x)$ the paramater μ_{hj} by $\hat{\mu}_{hj}$ and Σ_{hj} by a suitable estimator which will be defined explicitly in the various cases which follow. The estimator $R_{t|(x,j)}$ of $\rho_{t|(x,j)}$ for $t=1,\ldots,k$ is defined by

$$R_{t|(x,j)} = \rho_t \hat{p}_{tj} \hat{f}_{tj}(x) / \sum_{h=1}^{k} \rho_h \hat{p}_{hj} \hat{f}_{hj}(x)$$

and the notation for the k estimators together is

$$R_{\cdot|(x,j)} = (R_{1|(x,j)},\ldots,R_{k|(x,j)})^T.$$

For the derivation of the asymptotic distribution of the estimator of the vector of posterior probabilities we shall use the following lemma.

LEMMA 2.5.1.

$$\mathscr{L} n_h^{1/2} \begin{bmatrix} \hat{p}_{h1}-p_{h1} \\ \cdot \\ \cdot \\ \cdot \\ \hat{p}_{hd}-p_{hd} \\ \hat{\mu}_{h1}-\mu_{h1} \\ \mathrm{vec}(\hat{\Sigma}_{h1})-\mathrm{vec}(\Sigma_{h1}) \\ \cdot \\ \cdot \\ \cdot \\ \hat{\mu}_{hd}-\mu_{hd} \\ \mathrm{vec}(\hat{\Sigma}_{hd})-\mathrm{vec}(\Sigma_{hd}) \end{bmatrix} \to N(0,M_h)$$

for $h=1,\ldots,k$ and where M_h is a $d+d(p+p^2)$ square block-diagonal matrix with blocks

$$D_h - p_h p_h^T \; ; \; p_{hs}^{-1} \begin{bmatrix} \Sigma_{hs} & 0 \\ 0 & (I_{p^2}+K_p)(\Sigma_{hs}\otimes\Sigma_{hs}) \end{bmatrix}, \quad s=1,\ldots,d$$

with D_h a diagonal matrix with diagonal elements $p_{h1},\ldots,p_{hd}$, and where $p_h=(p_{h1},\ldots,p_{hd})^T$ and K_p defined just after (2.4.12).

PROOF. The proof of this lemma is given in section 2.6.

As in the situation of only continuous variables in section 2.4, we shall also consider two cases in this section. However, instead of assumptions about dispersion matrices of populations themselves, the assumptions about the dispersion matrices of the subpopulations lead to the two cases. With j the score on the discrete variable, j considered fixed, we shall distinguish between the following situations.

A. No assumption about the dispersion matrices Σ_{1j} up to Σ_{kj}.
B. Equality of the dispersion matrices Σ_{1j} up to Σ_{kj}.

A. NO ASSUMPTION ABOUT THE DISPERSION MATRICES Σ_{1j} UP TO Σ_{kj}

Let us use the notation $\Sigma_h=\Sigma_{hj}$, $h=1,\ldots,k$, where j is the score on the discrete variable. Now, define the sets of indices

$$G_h = \{s;\ 1\leqslant s\leqslant d \text{ for which } \Sigma_{hs} = \Sigma_h\}, \quad h=1,\ldots,k.$$

The value $f_{hj}(x)$ of density f_{hj} at the score vector x is estimated by $\hat{f}_{hj}(x)$ which is obtained by replacing in $f_{hj}(x)$ the parameter μ_{hj} by $\hat{\mu}_{hj}$ and Σ_{hj} by $\hat{\Sigma}_h$, defined as follows

$$\begin{aligned}\hat{\Sigma}_h &= \frac{1}{\sum_{s\in G_h}(n_{hs}-1)} \sum_{s\in G_h} \sum_{i=1}^{n_h} I(D_{hi}=s)(X_{hi}-\hat{\mu}_{hs})(X_{hi}-\hat{\mu}_{hs})^T \\ &= \sum_{s\in G_h} \frac{(n_{hs}-1)}{\sum_{t\in G_h}(n_{ht}-1)} \hat{\Sigma}_{hs}.\end{aligned}$$

Thus we assume that the observations of those subpopulations which have the same dispersion matrix as Σ_{hj}, are used for the estimation of the latter.
For large n_h we find

$$\hat{\Sigma}_h \simeq \sum_{s\in G_h} \frac{\hat{p}_{hs}}{\sum_{t\in G_h}\hat{p}_{ht}} \hat{\Sigma}_{hs}$$

and, lemma 2.5.1 implies that

$$\mathcal{L} n_h^{1/2}(\mathrm{vec}(\hat{\Sigma}_h)-\mathrm{vec}(\Sigma_h)) \to N_{p^2}(0,(\sum_{t\in G_h} p_{ht})^{-1}(I_{p^2}+K_p)(\Sigma_h\otimes\Sigma_h))$$

where the notation Σ_h is used for Σ_{hj}. Together with the result

$$\mathcal{L}n_h^{1/2}(\hat{\mu}_{hj}-\mu_{hj}) \to N_p(0,p_{hj}^{-1}\Sigma_h),$$

mentioned in lemma 2.5.1 and with the derivatives

$$\frac{\partial f_{hj}(x)}{\partial \mu_{hj}} \quad \text{and} \quad \frac{\partial f_{hj}(x)}{\partial \text{vec}(\Sigma_h)},$$

evaluated at the point $(\mu_{hj}^T,\text{vec}^T(\Sigma_h))^T$, see also (2.4.23) and (2.4.24), we can easily derive with lemma 2.2.1 that

$$\mathcal{L}n_h^{1/2}(\hat{f}_{hj}(x)-f_{hj}(x)) \to N(0,\Gamma_h)$$

where

$$\Gamma_h = (f_{hj}(x))^2\{p_{hj}^{-1}\Delta_{x;hj}^2+\tfrac{1}{2}p(\sum_{s\in G_h}p_{hs})^{-1}-(\sum_{s\in G_h}p_{hs})^{-1}\Delta_{x;hj}^2$$

$$+\tfrac{1}{2}(\sum_{s\in G_h}p_{hs})^{-1}\Delta_{x;hj}^4\}$$

and in which

$$\Delta_{x;hj}^2 = (x-\mu_{hj})^T\Sigma_{hj}^{-1}(x-\mu_{hj}).$$

This is proved in the same way as the analogue result in lemma 2.4.3. For the variance Γ_h we can write $\Gamma_h = D_h + Q_h$ where

$$D_h = (f_{hj}(x))^2 p_{hj}^{-1}\Delta_{x;hj}^2$$

and

$$Q_h = (f_{hj}(x))^2(\sum_{s\in G_h}p_{hs})^{-1}\{\tfrac{1}{2}p-\Delta_{x;hj}^2+\tfrac{1}{2}\Delta_{x;hj}^4\}.$$

Now, using the asymptotic distribution of $n_h^{1/2}(\hat{p}_{hj}-p_{hj})$ mentioned in lemma 2.5.1, it is easy to derive that

$$\mathcal{L}n_h^{1/2}(\hat{p}_{hj}\hat{f}_{hj}(x)-p_{hj}f_{hj}(x)) \to N(0,\Theta_h)$$

where

$$\Theta_h = p_{hj}^2\Gamma_h + (f_{hj}(x))^2 p_{hj}(1-p_{hj}).$$

Further

$$\mathcal{L}n_h^{1/2}(\log(\hat{p}_{hj}\hat{f}_{hj}(x))-\log(p_{hj}f_{hj}(x))) \to N(0,T_h)$$

where

$$T_h = \tfrac{1}{2}p(\sum_{s\in G_h}p_{hs})^{-1}+(p_{hj}^{-1}-(\sum_{s\in G_h}p_{hs})^{-1})\Delta_{x;hj}^2+\tfrac{1}{2}(\sum_{s\in G_h}p_{hj})^{-1}\Delta_{x;hj}^4$$

$$+ p_{hj}^{-1}(1-p_{hj}).$$

The following theorem can now be formulated.

THEOREM 2.5.2.

$$\mathscr{L}n^{1/2}(R_{\cdot|(x,j)} - \rho_{\cdot|(x,j)}) \to N_k(0, \Psi M \Psi)$$

where Ψ *is defined in (2.1.3) and (2.1.4) and* M *by*

$$M_{tt} = b_t^{-1}\{\tfrac{1}{2}p(\sum_{s\in G_t} p_{ts})^{-1} + (p_{tj}^{-1} - (\sum_{s\in G_t} p_{ts})^{-1})\Delta^2_{x;tj}$$

$$+\tfrac{1}{2}(\sum_{s\in G_t} p_{ts})^{-1}\Delta^4_{x;tj} + p_{tj}^{-1}(1-p_{tj})\} \quad t=1,\ldots,k$$

$$M_{ts} = 0 \quad t,s = 1,\ldots,k;\ t\neq s$$

and where $b_t = \lim_{n\to\infty} n_t n^{-1}$, $t=1,\ldots,k$.

REMARK. From theorem 2.5.2 a number of special cases can be derived.

(1). Take $G_t=\{j\}$, this means that only the observations from subpopulation (t,j) are used for the estimation of Σ_{tj}. Hence $\Sigma_{s\in G_t} p_{ts} = p_{tj}$ and

$$M_{tt} = b_t^{-1}\{\tfrac{1}{2}pp_{tj}^{-1} + \tfrac{1}{2}p_{tj}^{-1}\Delta^4_{x;tj} + p_{tj}^{-1}(1-p_{tj})\}.$$

(2). Take $G_t=\{s;\ s=1,\ldots,d\}$, this means that the subpopulations $(t, 1)$ up to (t,d) are assumed to have the same dispersion matrix. We obtain $\Sigma_{s\in G_t} p_{ts} = 1$ and

$$M_{tt} = b_t^{-1}\{\tfrac{1}{2}p + (p_{tj}^{-1} - 1)\Delta^2_{x;tj} + \tfrac{1}{2}\Delta^4_{x;tj} + p_{tj}^{-1}(1-p_{tj})\}.$$

(3). Take $d=1$, hence with probability 1 the discrete variable has a given value. The only nontrivial random variables are the continuous ones. The asymptotic distributions are the same as those of the model with no assumptions about the dispersion matrices in section 2.4. Now, $p_{tj}=1$, $\Sigma_{s\in G_t} p_{ts}=1$, and $(t,j)\to(t,1)\to t$, then

$$M_{tt} = \tfrac{1}{2}b_t^{-1}\{p + \Delta^4_{x;t}\} \quad t=1,\ldots,k$$

and this formula can also be found in theorem 2.4.4.

(4). By taking $p=0$ we obtain the case of only discrete variables presented in section 2.3. With $\Delta^2_{x;hj}=\Delta^4_{x;hj}=0$ the diagonal elements of M become

$$M_{tt} = b_t^{-1}p_{tj}^{-1}(1-p_{tj}) \quad t=1,\ldots,k$$

see also theorem 2.3.1.

REMARK. The above-mentioned cases correspond with cases mentioned earlier in section 2.1 under *ad*(1), *ad*(2) and *ad*(3). We have that (1) is (*ad*(3), *first*), (2) is (*ad*(3), *second*), (3) is (*ad*(2), B), and (4) is *ad*(1).

REMARK. Note that the continuous case B, see section 2.1, can be derived as a special case of above-mentioned cases (1) and (2).

B. EQUALITY OF THE DISPERSION MATRICES Σ_{1j} UP TO Σ_{kj}

Let us use the notation $\Sigma=\Sigma_{1j}=\cdots=\Sigma_{kj}$ and define the set G of double indices as follows

$$G = \{(h,s);\ 1\leqslant h\leqslant k;\ 1\leqslant s\leqslant d \text{ for which } \Sigma_{hs}=\Sigma\}.$$

The densities $f_{hj}(x), h=1,\ldots,k$ at the score vector x are estimated by $\hat{f}_{hj}(x), h=1,\ldots,k$. The estimated densities are obtained by replacing μ_{hj} by $\hat{\mu}$ and Σ_{hj} by $\hat{\Sigma}$ in the formula of the densities. The estimator $\hat{\Sigma}_{hj}$ of Σ is based on the observation vectors of the subpopulations with double index in the set G. We define

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{\sum\limits_{(t,u)\in G}(n_{tu}-1)}\sum_{(h,s)\in G}\sum_{i=1}^{n_{hs}} I(D_{hi}=s)(X_{hi}-\hat{\mu}_{hs})(X_{hi}-\hat{\mu}_{hs})^T \\ &= \sum_{(h,s)\in G}\frac{(n_{hs}-1)}{\sum\limits_{(t,u)\in G}(n_{tu}-1)}\hat{\Sigma}_{hs}.\end{aligned}$$

For large n this can be written as

$$\hat{\Sigma}\simeq\sum_{(h,s)\in G}\left[\frac{\hat{p}_{hs}b_h}{\sum\limits_{(t,u)\in G}\hat{p}_{tu}b_t}\right]\hat{\Sigma}_{hs}.$$

With use of lemma 2.5.1 it can derived that

$$\mathcal{L}n^{1/2}(\mathrm{vec}(\hat{\Sigma})-\mathrm{vec}(\Sigma)) \to N_{p^2}(0,(\sum_{(h,s)\in G}p_{hs}b_h)^{-1}(I_{p^2}+K_p)(\Sigma\otimes\Sigma)).$$

Now, using

$$\mathcal{L}n^{1/2}(\hat{\mu}_{hj}-\mu_{hj}) \to N(0,p_{hj}^{-1}b_h^{-1}\Sigma)$$

from lemma 2.5.1 and the expressions for

$$\frac{\partial f_{hj}(x)}{\partial \mu_{si}} \quad \text{and} \quad \frac{\partial f_{hj}(x)}{\partial \mathrm{vec}(\Sigma)},$$

evaluated at the vector $(\mu_{hj}^T,\mathrm{vec}^T(\Sigma))^T$ and given earlier in (2.4.23) and (2.4.24), we can derive with the δ-method of lemma 2.2.1 that

$$\mathcal{L}n^{1/2}\begin{bmatrix}\hat{f}_{1j}(x)-f_{1j}(x)\\ \cdot\\ \cdot\\ \cdot\\ \hat{f}_{kj}(x)-f_{kj}(x)\end{bmatrix} \to N_k(0,\Gamma)$$

where Γ is given by $\Gamma = D + Q$ and D a diagonal matrix with

$$D_{tt} = (\frac{\partial f_{tj}(x)}{\partial \mu_{tj}})^T (p_{tj}^{-1} b_t^{-1} \Sigma)(\frac{\partial f_{tj}(x)}{\partial \mu_{tj}})$$

$$= (f_{tj}(x))^2 p_{tj}^{-1} b_t^{-1} \Delta^2_{x;tj} \quad t = 1,...,k$$

and the matrix Q is defined by

$$Q_{ts} = (\frac{\partial f_{tj}(x)}{\partial \text{vec}(\Sigma)})^T ((\sum_{(h,u)\in G} p_{hu} b_h)^{-1} (I_{p^2} + K_p)(\Sigma \otimes \Sigma))(\frac{\partial f_{sj}(x)}{\partial \text{vec}(\Sigma)})$$

$$= f_{tj}(x) f_{sj}(x) (\sum_{(h,u)\in G} p_{hu} b_h)^{-1} \frac{1}{2} \{p - \Delta^2_{x;tj} - \Delta^2_{x;sj} + \Delta^4_{x;tjsj}\}$$

for $t,s = 1,...,k$, hence

$$\Gamma_{tt} = (f_{tj}(x))^2 \{\frac{1}{2} p (\sum_{(h,u)\in G} p_{hu} b_h)^{-1} + ((p_{tj} b_t)^{-1} - (\sum_{(h,u)\in G} p_{hu} b_h)^{-1}) \Delta^2_{x;tj}$$

$$+ \frac{1}{2} (\sum_{(h,u)\in G} p_{hu} b_h)^{-1} \Delta^4_{x;tj}\} \quad t = 1,...,k$$

and

$$\Gamma_{ts} = f_{tj}(x) f_{sj}(x) (\sum_{(h,u)\in G} p_{hu} b_h)^{-1} \frac{1}{2} \{p - \Delta^2_{x;tj} - \Delta^2_{x;sj} + \Delta^4_{x;tjsj}\}$$

for $t,s = 1,...,k$; $t \neq s$ and in which

$$\Delta^2_{x;tjsj} = (x - \mu_{tj})^T \Sigma^{-1} (x - \mu_{sj})$$

and $\Delta^2_{x;tj} = \Delta^2_{x;tjtj}$.

Next we derive that

$$\mathscr{L}n^{1/2} \begin{bmatrix} \hat{p}_{1j}\hat{f}_{1j}(x) - p_{1j}f_{1j}(x) \\ \cdot \\ \cdot \\ \cdot \\ \hat{p}_{kj}\hat{f}_{kj}(x) - p_{kj}f_{kj}(x) \end{bmatrix} \to N_k(0,\Theta)$$

where Θ is defined as

$$\Theta = diag\{b_h^{-1} p_{hj}(1-p_{hj})(f_{hj}(x))^2,\ h = 1,...,k\}$$

$$+ diag\{p_{hj},\ h = 1,...,k\} \Gamma\ diag\{p_{hj},\ h = 1,...,k\}$$

and where $diag\{d_h,\ h = 1,...k\}$ denotes the diagonal matrix with $d_1,...,d_k$ as its diagonal elements. Further we find

$$\mathscr{L}n^{1/2} \begin{bmatrix} \log(\hat{p}_{1j}\hat{f}_{1j}(x)) - \log(p_{1j}f_{1j}(x)) \\ \cdot \\ \cdot \\ \cdot \\ \log(\hat{p}_{kj}\hat{f}_{kj}(x)) - \log(p_{kj}f_{kj}(x)) \end{bmatrix} \to N_k(0,M)$$

where M is defined by

$$M_{tt} = b_t^{-1}p_{tj}^{-1}\{1-p_{tj}+\Delta^2_{x;tj}\} + (\sum_{(h,u)\in G} p_{hu}b_h)^{-1}\{\tfrac{1}{2}p-\Delta^2_{x;tj}+\tfrac{1}{2}\Delta^4_{x;tj}\}$$

$$M_{ts} = \tfrac{1}{2}(\sum_{(h,u)\in G} p_{hu}b_h)^{-1}\{p-\Delta^2_{x;tj}-\Delta^2_{x;sj}+\Delta^4_{x;tjsj}\}$$

for $t,s=1,\ldots,k$; $t\neq s$. Hence

THEOREM 2.5.3.

$$\mathcal{L}n^{1/2}(R_{\cdot|(x,j)}-\rho_{\cdot|(x,j)}) \to N_k(0,\Psi M\Psi)$$

where Ψ is defined in (2.1.3) and (2.1.4) and M just above this theorem.

Because of the property that the sum of the elements of a row or column of the matrix Ψ is zero, see also the remark after theorem 2.4.2, theorem 2.5.3 can also be formulated with M replaced by M', where M' is defined by

$$M'_{tt} = b_t^{-1}p_{tj}^{-1}\{1-p_{tj}+\Delta^2_{x;tj}\} + (\sum_{(h,u)\in G} p_{hu}b_h)^{-1}\tfrac{1}{2}\Delta^4_{x;tj}$$

$$M'_{ts} = (\sum_{(h,u)\in G} p_{hu}b_h)^{-1}\tfrac{1}{2}\Delta^4_{x;tjsj}$$

REMARK. The following special cases can be derived.

(1). $G=\{(h,j);\ h=1,\ldots,k\}$. This means that only the observations of the subpopulations $(1,j),\ldots,(k,j)$ are used for the estimation of Σ. Cases (1a) and (1b) are obtained by replacing $\Sigma_{(h,u)\in G} p_{hu}b_h$ by $\Sigma_{h=1}^{k} p_{hj}b_h$ in M and M', respectively.

(2). $G=\{(h,s);\ h=1,\ldots,k;\ s=1,\ldots,d\}$. The assumption is that all subpopulations have the same dispersion matrix. We obtain that $\Sigma_{(h,u)\in G} p_{hu}b_h=1$. Substituting this into the formulas for M and M' gives case (2a):

$$M_{tt} = b_t^{-1}p_{tj}^{-1}(1-p_{tj}+\Delta^2_{x;tj})+\tfrac{1}{2}p-\Delta^2_{x;tj}+\tfrac{1}{2}\Delta^4_{x;tj}$$

$$M_{ts} = \tfrac{1}{2}(p-\Delta^2_{x;tj}-\Delta^2_{x;sj}+\Delta^4_{x;tjsj}),$$

and case (2b):

$$M'_{tt} = b_t^{-1}p_{tj}^{-1}(1-p_{tj}+\Delta^2_{x;tj})+\tfrac{1}{2}\Delta^4_{x;tj}$$

$$M'_{ts} = \tfrac{1}{2}\Delta^4_{x;tjsj}.$$

(3). Assume that $d=1$. Hence we have only continuous variables. The results are the same as those of equality of dispersion matrices in section 2.4, see

theorems 2.4.1 and 2.4.2. Putting $d=1, p_{tj}=1, \Sigma_{(h,u)\in G} p_{hu}b_h=1$, and $(h,j)=h$ in the formulas for M and M' gives case (3a):

$$M_{tt} = \tfrac{1}{2}p + (b_t^{-1}-1)\Delta^2_{x;t} + \tfrac{1}{2}\Delta^4_{x;t}$$

$$M_{ts} = \tfrac{1}{2}p - \tfrac{1}{2}\Delta^2_{x;t} - \tfrac{1}{2}\Delta^2_{x;s} + \tfrac{1}{2}\{(x-\mu_t)^T\Sigma^{-1}(x-\mu_s)\}^2$$

and case (3b):

$$M'_{tt} = b_t^{-1}\Delta^2_{x;tj} + \tfrac{1}{2}\Delta^4_{x;t}$$

$$M'_{ts} = \tfrac{1}{2}\{(x-\mu_t)^T\Sigma^{-1}(x-\mu_s)\}^2$$

which are the same as the M of theorem 2.4.2 and the M of the theorem 2.4.1, respectively.

(4). Take $p=0$, hence only discrete variables are involved. We obtain $M=M'$ where

$$M_{tt} = b_t^{-1}p_{tj}^{-1}(1-p_{tj})$$

$$M_{ts} = 0$$

which correspond with the elements of matrix M in theorem 2.3.1.

REMARK. These above-mentioned cases correspond with cases mentioned earlier in section 2.1 under *ad*(1), *ad*(2), and *ad*(3). We have that (1a) is (*ad*(3), *third*), (2a) is (*ad*(3), *fourth*), (3a) is (*ad*(2), A.2), (3b) is (*ad*(2), A.1), and (4) is *ad*(1).

REMARK. The continuous case A.1, see section 2.1, can be obtained as a special case of above-mentioned cases (1b), (2b), and (3b). The continuous case A.2 can be derived from (1a), (2a) and (3a).

2.6. PROOF OF LEMMA 2.5.1

In this section we shall give three different methods of proving lemma 2.5.1. This lemma deals with the situation of the n_h independently distributed random variables (X_{hi}, D_{hi}), $i=1,\ldots,n_h$, h fixed, $1\leqslant h\leqslant k$ and

$$X_{hi}|D_{hi}=\ell \sim N_p(\mu_{h\ell},\Sigma_{h\ell}) \quad i=1,\ldots,n_h;\ h=1,\ldots,k;\ \ell=1,\ldots,d$$

and

$$P(D_{hi}=\ell) = p_{h\ell} \quad i=1,\ldots,n_h;\ h=1,\ldots,k;\ \ell=1,\ldots,d.$$

The lemma gives the asymptotic distribution of

$$(\hat{p}_{h1},\ \hat{\mu}_{h1}^T,\ \mathrm{vec}^T(\hat{\Sigma}_{h1}),\ldots,\hat{p}_{hd},\ \hat{\mu}_{hd}^T,\ \mathrm{vec}^T(\hat{\Sigma}_{hd}))^T,$$

see lemma 2.5.1 for the precise formulation.

METHOD 1

This method of proof is based on lemma 2.6.1, which gives a relation between asymptotic distributions of a fixed and a random sampling scheme. The randomness of the sampling scheme is brought about by the discrete component of the random variables. The k different classes in lemma 2.6.1 correspond with the d classes of the training set variables of section 2.5.

Let $w_{\ell,m}: \mathbb{R}^{q\times m}\to\mathbb{R}^q$, $\ell=1,...,k$ and $m=1,2,...$ be continuous functions. Let the i.i.d. random variables $Y_i^{(\ell)}, i=1,2,...$ each with density function f_{θ_ℓ}, $\theta_\ell\in\mathbb{R}^q$ assume values in $\mathbb{R}^q$, $\ell=1,...,k$. The random variables $W_m^{(\ell)}$, defined by

$$W_m^{(\ell)} = w_{\ell,m}(Y_1^{(\ell)},...,Y_m^{(\ell)}) \quad \ell=1,...,k; \quad m=1,2,...$$

are assumed to satisfy

$$\mathcal{L}m^{1/2}(W_m^{(\ell)}-\theta_\ell)\to N_q(0,\Sigma_\ell) \quad \ell=1,...,k.$$

Let (X_i,D_i), $i=1,...n$, be i.i.d. random variables whose distributions are given by

$$P(D_i=d) = p_d>0, \quad d=1,...,k\ , \ \sum_{d=1}^{k} p_d=1$$

and $X_i|D_i=d$ has density f_{θ_d}, $d=1,...,k$; $i=1,...,n$. Let $N_{\ell,n}$ be defined by

$$N_{\ell,n} = \sum_{i=1}^{n} I(D_i=\ell) \quad \ell=1,...,k$$

where I is the indicator function, and let $\hat{p}_{\ell,n}$ be defined by

$$\hat{p}_{\ell,n} = n^{-1}N_{\ell,n} \quad \ell=1,...,k.$$

Further, let $W_{\ell,n}$ be defined by

$$W_{\ell,n} = w_{\ell,n_\ell}(X_{\ell_1},...,X_{\ell_{n_\ell}}) \quad \ell=1,...,k$$

where n_ℓ is the value assumed by $N_{\ell,n}$ and where X_{ℓ_j} (X with subindex $\ell_j, j=1,...,n_\ell$) is the j-th element of the subsequence of those X_i's with $D_i=\ell$. Now, using the notation $\hat{p}_n=(\hat{p}_{1,n},...,\hat{p}_{k,n})^T$ and $p=(p_1,...,p_k)^T$, we have the following result.

LEMMA 2.6.1.

$$\mathcal{L}n^{1/2}\begin{bmatrix} \hat{p}_n-p \\ W_{1,n}-\theta_1 \\ \vdots \\ W_{k,n}-\theta_k \end{bmatrix} \to N_{k+kq}(0,M)$$

where M is a blockdiagonal matrix with the blocks

$$D-pp^T,\ p_1^{-1}\Sigma_1,\ldots,p_k^{-1}\Sigma_k$$

and where D is the diagonal matrix $D=diag\{p_1,\ldots,p_k\}$.

PROOF. We shall prove the convergence in distribution by the pointwise convergence of the corresponding characteristic function. So, let $t_0\in\mathbb{R}^k$, $t_i\in\mathbb{R}^q$, $t=1,\ldots,k$ then we define

$$\begin{aligned}&\phi_n(t_0,t_1,\ldots,t_k)=\\&E\{\exp(it_0^Tn^{1/2}(\hat{p}_n-p)+it_1^Tn^{1/2}(W_{1,n}-\theta_1)+\cdots\\&+it_k^Tn^{1/2}(W_{k,n}-\theta_k))\}.\end{aligned}\tag{2.6.1}$$

By conditioning to $D_1,\ldots,D_n$ formula (2.6.1) can be written as

$$E_{D_1,\ldots,D_n}\{E\{\exp(it_0^Tn^{1/2}(\hat{p}_n-p)+\sum_{j=1}^{k}it_j^Tn^{1/2}(W_{j,n}-\theta_j))|D_1=d_1,\ldots,D_n=d_n\}\}=$$

$$E_{D_1,\ldots,D_n}\{\exp(it_0^Tn^{1/2}(\hat{p}_n-p))E\{\exp(\sum_{j=1}^{k}it_j^Tn^{1/2}(W_{j,n}-\theta))|D_1=d_1,\ldots,D_n=d_n\}\}=$$

$$\sum_{(d_1,\ldots,d_n)}\exp(it_0^Tn^{1/2}(\hat{p}_n-p))E\{\exp(\sum_{j=1}^{k}it_j^Tn^{1/2}(W_{j,n}-\theta_j))|D_1=d_1,\ldots,D_n=d_n\}$$

$$\cdot P(D_1=d_1,\ldots,D_n=d_n)\tag{2.6.2}$$

where $P(D_1=d_1,\ldots,D_n=d_n)=p_{d_1}\cdots p_{d_n}$. The $W_{1,n},\ldots,W_{k,n}$ are conditional on $D_1,\ldots,D_n$ independent because of their definitions in which the (X_i,D_i) 's appear as i.i.d. random variables. Hence the expression for the characteristic function, formula (2.6.2), becomes

$$\sum_{(d_1,\ldots,d_n)}\exp(it_0^Tn^{1/2}(\hat{p}_n-p))\prod_{j=1}^{k}E\{\exp(it_j^Tn^{1/2}(W_{j,n}-\theta_j))|D_1=d_1,\ldots,D_n=d_n\}$$

$$\cdot P(D_1=d_1,\ldots,D_n=d_n).\tag{2.6.3}$$

Now, note that

$$\mathcal{L}(W_{j,n}|(D_1=d_1,\ldots,D_n=d_n;N_{j,n}=n_j))=\mathcal{L}W_{n_j}^{(j)}$$

hence

$$\begin{aligned}&\mathcal{L}(n^{1/2}(W_{j,n}-\theta_j)|(D_1=d_1,\ldots,D_n=d_n;N_{j,n}=n_j))\\&=\mathcal{L}(\frac{n}{n_j})^{1/2}n_j^{1/2}(W_{n_j}^{(j)}-\theta_j)\end{aligned}$$

so the formula (2.6.3) becomes

$$\sum_{(n_1,\ldots,n_k)}\exp(it_0^Tn^{1/2}(\hat{p}_n-p))\prod_{j=1}^{k}E\{\exp(it_j^T(\frac{n}{n_j})^{1/2}n_j^{1/2}(W_{n_j}^{(j)}-\theta_j))\}$$

$$\cdot P(N_{1,n}=n_1,\ldots,N_{k,n}=n_k)=$$

$$\sum_{(n_1,\ldots,n_k)} \exp(it_0^T n^{1/2}(\hat{p}_n-p))\prod_{j=1}^{k}\phi_{n_j}^{(j)}(t_j(\frac{n}{n_j})^{1/2})$$

$$\cdot P(N_{1,n}=n_1,\ldots,N_{k,n}=n_k) \tag{2.6.4}$$

where $\phi_{n_j}^{(j)}$ is the characteristic function of $n_j^{1/2}(W_{n_j}^{(j)}-\theta_j)$. Now, let us write for $j=1,\ldots,k$

$$\phi_{n_j}^{(j)}(t_j(\frac{n}{n_j})^{1/2}) = \phi_\infty^{(j)}(t_j p_j^{-1/2})+R_{n_j}^{(j)}(t_j,p_j,n) \tag{2.6.5}$$

where $\phi_\infty^{(j)}$ is the characteristic function of the $N_q(0,\Sigma_j)$ distribution, which is the limit distribution of $m^{1/2}(W_m^{(j)}-\theta_j)$. The fact that characteristic functions are bounded in absolute value by one and hence the remainder terms $R_{n_j}^{(j)}(t_j,p_j,n)$ bounded in absolute value by two, implies that

$$|\prod_{j=1}^{k}(\phi_\infty^{(j)}(t_j p_j^{-1/2})+R_{n_j}^{(j)}(t_j,p_j,n))-\prod_{j=1}^{k}\phi_\infty^{(j)}(t_j p_j^{-1/2})|$$

$$\leqslant f(k)\sum_{j=1}^{k}|R_{n_j}^{(j)}(t_j,p_j,n)| \tag{2.6.6}$$

in which $f(k)$ is a suitable constant which depends only on k. Using (2.6.5) and (2.6.6), formula (2.6.4) can be written as

$$\sum_{(n_1,\ldots,n_k)} \exp(it_0^T n^{1/2}(\hat{p}_n-p))\prod_{j=1}^{k}\phi_\infty^{(j)}(t_j p_j^{-1/2})$$

$$\cdot P(N_{1,n}=n_1,\ldots,N_{k,n}=n_k)+R \tag{2.6.7}$$

where

$$|R|\leqslant \sum_{(n_1,\ldots,n_k)} f(k)\sum_{j=1}^{k}|R_{n_j}^{(j)}(t_j,p_j,n)|P(N_{1,n}=n_1,\ldots,N_{k,n}=n_k)$$

$$= f(k)\sum_{j=1}^{k}E|R_{N_j}^{(j)}(t_j,p_j,n)|. \tag{2.6.8}$$

Now, because $n^{1/2}(\hat{p}_n-p_n)$ converges in distribution to a $N_k(0,D-pp^T)$ distribution, (see CRAMÉR (1946) p. 419), the first term of formula (2.6.7) tends to

$$\exp(-\tfrac{1}{2}t_0^T(D-pp^T)t_0-\sum_{j=1}^{k}\tfrac{1}{2}t_j^T p_j^{-1}\Sigma_j t_j)$$

which is the desired characteristic function. The proof is finished if we show that the remainder term R tends to zero. Now,

$$E|R_{N_j}^{(j)}(t_j,p_j,n)|\leqslant E|\phi_{N_j}^{(j)}(t_j(\frac{n}{N_j})^{1/2})-\phi_\infty^{(j)}(t_j(\frac{n}{N_j})^{1/2})|$$

$$+ E|\phi_\infty^{(j)}(t_j(\frac{n}{N_j})^{1/2})-\phi_\infty^{(j)}(t_j p_j^{-1/2})|. \tag{2.6.9}$$

For the first term at the right-hand side of formula (2.6.9) we have that

$$N_j \sim Bin(n,p_j)$$

$$t_j(\frac{n}{N_j})^{1/2} \xrightarrow{P} t_j p_j^{-1/2}$$

$$\phi_m^{(j)}(t) \to \phi_\infty^{(j)}(t),\ \forall t \in \mathbb{R}^q,\ \text{uniform on compact sets}$$

hence

$$\phi_{N_j}^{(j)}(t_j(\frac{n}{N_j})^{1/2}) - \phi_\infty^{(j)}(t_j(\frac{n}{N_j})^{1/2}) \xrightarrow{P} 0. \tag{2.6.10}$$

Further, because of

$$|\phi_{N_j}^{(j)}(t_j(\frac{n}{N_j})^{1/2}) - \phi_\infty^{(j)}(t_j(\frac{n}{N_j})^{1/2})| \leqslant 2 \tag{2.6.11}$$

we conclude from (2.6.12) and (2.6.11) (see CHUNG (1968) th. 4.1.4) that

$$E|\phi_{N_j}^{(j)}(t_j(\frac{n}{N_j})^{1/2}) - \phi_\infty^{(j)}(t_j(\frac{n}{N_j})^{1/2})| \to 0. \tag{2.6.12}$$

For the second term at the right-hand side of formula (2.6.9) we obtain from

$$N_j \sim Bin(n,p_j)$$

$\phi_\infty^{(j)}$ is a continuous function

$$t_j(\frac{n}{N_j})^{1/2} \xrightarrow{P} t_j p_j^{-1/2}$$

that

$$\phi_\infty^{(j)}(t_j(\frac{n}{N_j})^{1/2}) \xrightarrow{P} \phi_\infty^{(j)}(t_j p_j^{-1/2})$$

and

$$\phi_\infty^{(j)}(t_j(\frac{n}{N_j})^{1/2}) - \phi_\infty^{(j)}(t_j p_j^{-1/2}) \xrightarrow{P} 0. \tag{2.6.13}$$

Further

$$|\phi_\infty^{(j)}(t_j(\frac{n}{N_j})^{1/2}) - \phi_\infty^{(j)}(t_j p_j^{-1/2})| \leqslant 2. \tag{2.6.14}$$

Formula (2.6.13) and (2.6.14) imply that

$$E\ |\phi_\infty^{(j)}(t_j(\frac{n}{N_j})^{1/2}) - \phi_\infty^{(j)}(t_j p_j^{-1/2})| \to 0. \tag{2.6.15}$$

This completes the proof of lemma 2.6.1.

The proof of lemma 2.5.1 can easily be obtained by using lemma 2.6.1. For the estimators $\hat{\mu}_{h\ell}$ and $\text{vec}(\hat{\Sigma}_{h\ell})$ of the mean and dispersion matrix respectively

of a multivariate normal distribution with parameters $\mu_{h\ell}$ and $\Sigma_{h\ell}$ we have

$$\mathcal{L}m^{1/2}\begin{bmatrix}\hat{\mu}_{h\ell} & -\mu_{h\ell} \\ \text{vec}(\hat{\Sigma}_{h\ell}) & -\text{vec}(\Sigma_{h\ell})\end{bmatrix} \to N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\Sigma_{h\ell} & 0 \\ 0 & (I_{p^2}+K_p)(\Sigma_{h\ell}\otimes\Sigma_{h\ell})\end{bmatrix}\right)$$

see section 2.4, and MUIRHEAD (1982) p. 90. Now, let $(\hat{\mu}_{h\ell}, \text{vec}^T(\hat{\Sigma}_{h\ell}))^T$ play the role of $W_m^{(\ell)}$ in lemma 2.6.1, then the correctness of lemma 2.5.1 follows immediately.

METHOD 2

In this proof of lemma 2.5.1 we begin with the definition of the following random variables, which are functions of (X_{hi}, D_{hi}), $i=1,\ldots,n_h$, namely

$$\begin{aligned} R_{hs,i} &= I(D_{hi}=s) \\ S_{hs,i} &= X_{hi}I(D_{hi}=s) \\ T_{hs,i} &= (X_{hi}\otimes X_{hi})I(D_{hi}=s) \end{aligned}$$

for $s=1,\ldots,d$ and $i=1,\ldots,n_h$. We formulate the following lemma.

LEMMA 2.6.2.

$$\mathcal{L}n_h^{1/2}\begin{bmatrix} n_h^{-1}\sum_{i=1}^{n_h} R_{h1,i} - p_{h1} \\ n_h^{-1}\sum_{i=1}^{n_h} S_{h1,i} - p_{h1}\mu_{h1} \\ n_h^{-1}\sum_{i=1}^{n_h} T_{h1,i} - p_{h1}(\mathrm{vec}(\Sigma_{h1}) + \mu_{h1}\otimes\mu_{h1}) \\ \vdots \\ n_h^{-1}\sum_{i=1}^{n_h} R_{hd,i} - p_{hd} \\ n_h^{-1}\sum_{i=1}^{n_h} S_{hd,i} - p_{hd}\mu_{hd} \\ n_h^{-1}\sum_{i=1}^{n_h} T_{hd,i} - p_{hd}(\mathrm{vec}(\Sigma_{hd}) + \mu_{hd}\otimes\mu_{hd}) \end{bmatrix} \to N(0, Q_h)$$

where Q_h is a square matrix of size $d(1+p+p^{2)}$ and partitioned as

$$Q_h = \begin{bmatrix} Q_{h,11} & \cdots & Q_{h,1d} \\ \vdots & & \vdots \\ Q_{h,d1} & \cdots & Q_{h,dd} \end{bmatrix}$$

with $Q_{h,st}, s = 1,\ldots,d,\ t = 1,\ldots,d$ square matrices of size $1+p+p^2$. These $Q_{h,st}$ are specified by the following submatrices, for $s = 1,\ldots,d$:

$$\begin{aligned}
Q_{h,ss;11} &= p_{hs}(1-p_{hs}) \\
Q_{h,ss;12} &= p_{hs}(1-p_{hs})\mu_{hs}^T \\
Q_{h,ss;13} &= p_{hs}(1-p_{hs})\ \{\mathrm{vec}^T(\Sigma_{hs}) + \mu_{hs}^T\otimes\mu_{hs}^T\} \\
Q_{h,ss;22} &= p_{hs}\Sigma_{hs} + p_{hs}(1-p_{hs})\mu_{hs}\mu_{hs}^T \\
Q_{h,ss;23} &= p_{hs}\{\mu_{hs}^T\otimes\Sigma_{hs} + \Sigma_{hs}\otimes\mu_{hs}^T\} + \\
&\quad p_{hs}(1-p_{hs})\{\mu_{hs}\ \mathrm{vec}^T(\Sigma_{hs}) + \mu_{hs}(\mu_{hs}^T\otimes\mu_{hs}^T)\} \\
Q_{h,ss;33} &= p_{hs}\ \{(\Sigma_{hs}\otimes\Sigma_{hs})(I_{p^2} + K_p) + (\mu_{hs}\mu_{hs}^T)\otimes\Sigma_{hs}\ + \\
&\quad \mu_{hs}^T\otimes\Sigma_{hs}\otimes\mu_{hs} + \mu_{hs}\otimes\Sigma_{hs}\otimes\mu_{hs}^T + \Sigma_{hs}\otimes\mu_{hs}\mu_{hs}^T\} + \\
&\quad p_{hs}(1-p_{hs})\{\mathrm{vec}(\Sigma_{hs})\mathrm{vec}^T(\Sigma_{hs})\ +
\end{aligned}$$

$$\text{vec}(\Sigma_{hs})(\mu_{hs}^T \otimes \mu_{hs}^T) + (\mu_{hs} \otimes \mu_{hs})\text{vec}^T(\Sigma_{hs}) +$$
$$(\mu_{hs} \otimes \mu_{hs})(\mu_{hs}^T \otimes \mu_{hs}^T)\}$$

and the other submatrices of $Q_{h,ss}$ *follow from the symmetry. For* $s=1,\ldots,d$, $t=1,\ldots,d$, $s \neq t$ *we have*

$$\begin{aligned}
Q_{h,ts;11} &= -p_{ht}p_{hs} \\
Q_{h,ts;12} &= -p_{ht}p_{hs}\mu_{hs}^T \\
Q_{h,ts;13} &= -p_{ht}p_{hs}\{\text{vec}^T(\Sigma_{hs}) + \mu_{hs}^T \otimes \mu_{hs}^T\} \\
Q_{h,ts;21} &= -p_{ht}p_{hs}\mu_{ht} \\
Q_{h,ts;22} &= -p_{ht}p_{hs}\mu_{ht}\mu_{hs}^T \\
Q_{h,ts;23} &= -p_{ht}p_{hs}\mu_{ht}\{\text{vec}^T(\Sigma_{hs}) + \mu_{hs}^T \otimes \mu_{hs}^T\} \\
Q_{h,ts;31} &= -p_{ht}p_{hs}\{\text{vec}(\Sigma_{ht}) + \mu_{ht} \otimes \mu_{ht}\} \\
Q_{h,ts;32} &= -p_{ht}p_{hs}\{\text{vec}(\Sigma_{ht}) + \mu_{ht} \otimes \mu_{ht}\}\mu_{hs}^T \\
Q_{h,ts;33} &= -p_{ht}p_{hs}\{\text{vec}(\Sigma_{ht})\text{vec}^T(\Sigma_{hs}) + \\
&\quad (\mu_{ht} \otimes \mu_{ht})\text{vec}^T(\Sigma_{hs}) + \text{vec}(\Sigma_{ht})(\mu_{hs}^T \otimes \mu_{hs}^T) + \\
&\quad (\mu_{ht} \otimes \mu_{ht})(\mu_{hs}^T \otimes \mu_{hs}^T)\}.
\end{aligned}$$

The proof of this lemma will be given after lemma 2.6.4. We shall first give a few equalities which are frequently used in the fortcoming derivations. Let μ be a p-dimensional vector and Σ a $p \times p$ symmetrical matrix, then

(a) $\text{vec}(\mu\mu^T) = \mu \otimes \mu = I\mu \otimes \mu 1 = (I \otimes \mu)(\mu \otimes 1) = (I \otimes \mu)\mu$

(b) $(\mu \otimes I)\Sigma = \mu \otimes \Sigma$

PROOF. $((\mu \otimes I)\Sigma)^T = \Sigma(\mu^T \otimes I) = (1 \otimes \Sigma)(\mu^T \otimes I) = \mu^T \otimes \Sigma = (\mu \otimes \Sigma)^T$

(c) $(I \otimes \mu)\Sigma = (\Sigma \otimes \mu)^T$

PROOF.

$((I \otimes \mu)\Sigma)^T = \Sigma(I \otimes \mu^T) = (\Sigma \otimes 1)(I \otimes \mu^T) = \Sigma \otimes \mu^T = (\Sigma \otimes \mu)^T$

(d) $(\mu \otimes \Sigma)(I \otimes \mu^T) = \mu \otimes \Sigma \otimes \mu^T$

PROOF.

$(\mu \otimes \Sigma)(I \otimes \mu^T) = ((\mu \otimes \Sigma) \otimes 1)(I \otimes \mu^T) = ((\mu \otimes \Sigma)I) \otimes (1\mu^T) =$
$\mu \otimes \Sigma \otimes \mu^T$

(e) $(\Sigma \otimes \mu)(\mu^T \otimes I) = \mu^T \otimes \Sigma \otimes \mu$

PROOF. $((\Sigma \otimes \mu)(\mu^T \otimes I))^T = (\mu \otimes I)(\Sigma \otimes \mu^T) = (\mu \otimes I)(1 \otimes (\Sigma \otimes \mu^T)) =$
$(\mu \cdot 1) \otimes (I(\Sigma \otimes \mu^T)) = \mu \otimes \Sigma \otimes \mu^T = (\mu^T \otimes \Sigma \otimes \mu)^T.$

LEMMA 2.6.3.

(a) $E\ I(D_{hi}=s) = p_{hs}$

(b) $VAR\ I(D_{hi}=s) = p_{hs}(1-p_{hs})$

(c) $E\ I(D_{hi}=s)X_{hi} = p_{hs}\mu_{hs}$
(d) $E\ I(D_{hi}=s)X_{hi}X_{hi}^T = p_{hs}(\Sigma_{hs}+\mu_{hs}\mu_{hs}^T)$
(e) $E\ I(D_{hi}=s)X_{hi}\otimes X_{hi} = p_{hs}(\mathrm{vec}(\Sigma_{hs})+\mu_{hs}\otimes\mu_{hs})$

PROOF. (a) and (b) follow from the $Bin(1,p_{hs})$ distribution, (c), (d) and (e) can be derived with use of conditional expectations.

LEMMA 2.6.4.
Let $U\sim N_p(0,I)$ *and* $X\sim N_p(\mu,\Sigma)$, *then*

(a) $E\ U\otimes U^T = I$
(b) $E\ U\otimes U = \mathrm{vec}(I)$
(c) $E\ UU^T\otimes U = 0$
(d) $E\ UU^T\otimes U^T = 0$
(e) $E\ X(X^T\otimes X^T) = \mu(\mu^T\otimes\mu^T)+\mu\mathrm{vec}^T(\Sigma)+\mu^T\otimes\Sigma+\Sigma\otimes\mu^T$
(f) $E\ UU^T\otimes UU^T = K_p+I\otimes I+\mathrm{vec}(I)\mathrm{vec}^T(I)$
(g) $E(X\otimes X)(X^T\otimes X^T) = \mu\mu^T\otimes\mu\mu^T+\Sigma\otimes\mu\mu^T+(\mu\otimes\mu)\mathrm{vec}^T(\Sigma)+$
$\mu^T\otimes\Sigma\otimes\mu+\mu\otimes\Sigma\otimes\mu^T+\mathrm{vec}(\Sigma)(\mu^T\otimes\mu^T)+$
$\mu\mu^T\otimes\Sigma+(\Sigma\otimes\Sigma)(I_{p^2}+K_p)+\mathrm{vec}(\Sigma)\mathrm{vec}^T(\Sigma)$

where K_p *has been defined after (2.4.12).*

PROOF. Let U_i be the i-th component of $U, i=1,\ldots,p$. Then (a),...,(d) follow directly from

$$E\ U_i = E\ U_i^3 = 0$$

and

$$E\ U_i^2 = 1\,, \qquad E\ U_i^4 = 3\,, \qquad i=1,\ldots,p.$$

PROOF of (e).

$$\begin{aligned} EX(X^T\otimes X^T) &= E(\mu+\Sigma^{1/2}U)(\mu^T+U^T\Sigma^{1/2})\otimes(\mu^T+U^T\Sigma^{1/2}) \\ &= \mu(\mu^T\otimes\mu^T)+E\mu(U^T\Sigma^{1/2}\otimes U^T\Sigma^{1/2})+ \\ &\quad E\Sigma^{1/2}U(\mu^T\otimes U^T\Sigma^{1/2})+E\Sigma^{1/2}U(U^T\Sigma^{1/2}\otimes\mu^T) \end{aligned}$$

where we have deleted terms with first and third moments of U. Now, using $\Sigma^{1/2}U=1\otimes\Sigma^{1/2}U=\Sigma^{1/2}U\otimes 1$ the expression becomes

$$\mu(\mu^T\otimes\mu^T)+\mu E(U^T\otimes U^T)(\Sigma^{1/2}\otimes\Sigma^{1/2})+$$
$$\mu^T\otimes\Sigma^{1/2}EUU^T\Sigma^{1/2}+\Sigma^{1/2}EUU^T\Sigma^{1/2}\otimes\mu^T.$$

With $EUU^T=I$, $EU^T\otimes U^T=\mathrm{vec}^T(I)$ and the property $\mathrm{vec}(ABC)=(C^T\otimes A)\mathrm{vec}(B)$ with the special choice $A=C=\Sigma^{1/2}$ and $B=I$ the proof of (e) is finished.

PROOF of (f). Let $T_{ij}=E_{ij}+E_{ji}$ where E_{ij} is defined after (2.4.12), then

$$EU_iU_jUU^T = T_{ij} + \delta_{ij}I$$

where δ_{ij} is the Kronecker delta: $\delta_{ij}=1$ if $i=j$ and $\delta_{ij}=0$ if $i\neq j$. Now,

$$EUU^T\otimes UU^T = E\sum_{i,j}U_iU_jE_{ij}\otimes UU^T =$$

$$\sum_{i,j}(E_{ij}\otimes(T_{ij}+\delta_{ij}I)) = \sum_{i,j}E_{ij}\otimes T_{ij}+\sum_{i,j}E_{ij}\otimes\delta_{ij}I =$$

$$\sum_{i,j}E_{ij}\otimes E_{ij}+\sum_{i,j}E_{ij}\otimes E_{ji}+(\sum_{i}E_{ii})\otimes I =$$

$$\text{vec}(I)\text{vec}^T(I) + K_p + I\otimes I.$$

PROOF of (g).

$$E(X\otimes X)(X^T\otimes X^T) = EXX^T\otimes XX^T =$$

$$E(\mu+\Sigma^{1/2}U)(\mu^T+U^T\Sigma^{1/2})\otimes(\mu+\Sigma^{1/2}U)(\mu^T+U^T\Sigma^{1/2}).$$

If we delete the terms with first and third moments of U we find

$$\mu\mu^T\otimes\mu\mu^T+\Sigma^{1/2}EUU^T\Sigma^{1/2}\otimes\mu\mu^T+E\mu U^T\Sigma^{1/2}\otimes\mu U^T\Sigma^{1/2} +$$

$$E\Sigma^{1/2}U\mu^T\otimes\mu U^T\Sigma^{1/2} + E\mu U^T\Sigma^{1/2}\otimes\Sigma^{1/2}U\mu^T +$$

$$E\Sigma^{1/2}U\mu^T\otimes\Sigma^{1/2}U\mu^T+\mu\mu^T\otimes\Sigma^{1/2}EUU^T\Sigma^{1/2} +$$

$$E\Sigma^{1/2}UU^T\Sigma^{1/2}\otimes\Sigma^{1/2}UU^T\Sigma^{1/2}.$$

Now, use $EUU^T=I$, $EU^T\otimes U^T=\text{vec}^T(I)$ and $\text{vec}^T(I)(\Sigma^{1/2}\otimes\Sigma^{1/2})=\text{vec}^T(\Sigma)$, then the first three terms become

$$\mu\mu^T\otimes\mu\mu^T+\Sigma\otimes\mu\mu^T+(\mu\otimes\mu)\text{vec}^T(\Sigma).$$

Using $EU\otimes U^T=I$ the fourth term becomes

$$(\Sigma^{1/2}\otimes\mu)I(\mu^T\otimes\Sigma^{1/2}) = (1\otimes(\Sigma^{1/2}\otimes\mu))(\mu^T\otimes\Sigma^{1/2}) =$$

$$(1\cdot\mu^T)\otimes((\Sigma^{1/2}\otimes\mu)\Sigma^{1/2}) = \mu^T\otimes(\Sigma^{1/2}\otimes\mu)(\Sigma^{1/2}\otimes 1)=\mu^T\otimes\Sigma\otimes\mu.$$

The fifth term is the transpose of the fourth term and becomes $\mu\otimes\Sigma\otimes\mu^T$. Further, the sixth and seventh term can be evaluated into

$$(\Sigma^{1/2}\otimes\Sigma^{1/2})\text{vec}(I)(\mu^T\otimes\mu^T)=\text{vec}(\Sigma)(\mu^T\otimes\mu^T)$$

and

$$\mu\mu^T\otimes\Sigma^{1/2}EUU^T\Sigma^{1/2} = \mu\mu^T\otimes\Sigma$$

respectively. The last term can be written as

$$(\Sigma^{1/2}\otimes\Sigma^{1/2})EUU^T\otimes UU^T(\Sigma^{1/2}\otimes\Sigma^{1/2}) =$$

$$(\Sigma^{1/2}\otimes\Sigma^{1/2})K_p(\Sigma^{1/2}\otimes\Sigma^{1/2})+(\Sigma^{1/2}\otimes\Sigma^{1/2})(I\otimes I)(\Sigma^{1/2}\otimes\Sigma^{1/2}) +$$

$$(\Sigma^{1/2}\otimes\Sigma^{1/2})\text{vec}(I)\text{vec}^T(I)(\Sigma^{1/2}\otimes\Sigma^{1/2}).$$

Using $(\Sigma^{1/2}\otimes\Sigma^{1/2})K_p=K_p(\Sigma^{1/2}\otimes\Sigma^{1/2})$ and $(\Sigma^{1/2}\otimes\Sigma^{1/2})\text{vec}(I)=\text{vec}(\Sigma)$ this becomes

$$(\Sigma\otimes\Sigma)(I_{p^2}+K_p) + \text{vec}(\Sigma)\text{vec}^T(\Sigma).$$

This completes the proof of (g) and hence the proof of lemma 2.6.4 has been finished. The results of this lemma can also be found in, for example, MAGNUS and NEUDECKER (1979).

PROOF OF LEMMA 2.6.2. The proof follows from the multivariate central limit theorem, see (2.4.8), and the lemmas 2.6.3 and 2.6.4. The expectations $ER_{ht\ell,i}, ES_{ht\ell,i}$ and $ET_{ht\ell},i$ are given in lemma 2.6.3. The submatrices $Q_{h,ss}$ and $Q_{h,ts}$ of the variance-covariance matrix Q_h are obtained in the following way

$$Q_{h,ss;11} = VAR(R_{hs,i})=VAR(I(D_{hi}=s)) = p_{hs}(1-p_{hs}).$$

Further, without a detailed derivation, we summarize

$$\begin{aligned}
Q_{h,ss,12} &= EI(D_{hi}=s)X_{hi}^T - EI(D_{hi}=s)EI(D_{hi}=s)X_{hi}^T \\
Q_{h,ss;13} &= EI(D_{hi}=s)X_{hi}^T\otimes X_{hi}^T - EI(D_{hi}=s)EI(D_{hi}=s)X_{hi}^T\otimes X_{hi}^T \\
Q_{h,ss;22} &= EI(D_{hi}=s)X_{hi}X_{hi}^T - EI(D_{hi}=s)X_{hi}EI(D_{hi}=s)X_{hi}^T \\
Q_{h,ss;23} &= EI(D_{hi}=s)X_{hi}(X_{hi}^T\otimes X_{hi}^T) - EI(D_{hi}=s)X_{hi}EI(D_{hi}=s)X_{hi}^T\otimes X_{hi}^T \\
Q_{h,ss,33} &= EI(D_{hi}=s)(X_{hi}\otimes X_{hi})(X_{hi}^T\otimes X_{hi}^T) - EI(D_{hi}=s)X_{hi}\otimes X_{hi} \\
&\quad \cdot EI(D_{hi}=s)X_{hi}^T\otimes X_{hi}^T.
\end{aligned}$$

The elements of the submatrix $Q_{h,ts}$ are derived in a similar way. We have

$$\begin{aligned}
Q_{h,ts;11} &= COV(R_{ht,i}, R_{hs,i}) = COV(I(D_{hi}=t), I(D_{hi}=s)) = \\
&= -p_{ht}p_{hs}
\end{aligned}$$

and it is easy to see that

$$\begin{aligned}
Q_{h,ts;12} &= -EI(D_{hi}=t)EI(D_{hi}=s)X_{hi}^T \\
Q_{h,ts;13} &= -EI(D_{hi}=t)EI(D_{hi}=s)X_{hi}^T\otimes X_{hi}^T \\
Q_{h,ts;21} &= -EI(D_{hi}=t)X_{hi}EI(D_{hi}=s) \\
Q_{h,ts;22} &= -EI(D_{hi}=t)X_{hi}EI(D_{hi}=s)X_{hi}^T \\
Q_{h,ts;23} &= -EI(D_{hi}=t)X_{hi}EI(D_{hi}=s)X_{hi}^T\otimes X_{hi}^T \\
Q_{h,ts;31} &= -EI(D_{hi}=t)X_{hi}\otimes X_{hi}EI(D_{hi}=s) \\
Q_{h,ss,32} &= -EI(D_{hi}=t)X_{hi}\otimes X_{hi}EI(D_{hi}=s)X_{hi}^T \\
Q_{h,ss,33} &= -EI(D_{hi}=t)X_{hi}\otimes X_{hi}EI(D_{hi}=s)X_{hi}^T\otimes X_{hi}^T.
\end{aligned}$$

Now, with the results of lemma 2.6.3 and 2.6.4 the elements of $Q_{h,ss}$ and $Q_{h,ts}$ of lemma 2.6.2 can be obtained immediately. This completes the proof of lemma 2.6.2.

The $\hat{p}_{hs}, \hat{\mu}_{hs}$ and $\text{vec}(\hat{\Sigma}_{hs}), s=1,...,d$ can be expressed as functions of the $R_{hs,i}$, $S_{hs,i}$ and $T_{hs,i}$ $s=1,...,d$, $i=1,...,n_h$. We have

$$\hat{p}_{hs} = \frac{1}{n_h}\sum_{i=1}^{n_h} R_{hs,i}$$

$$\hat{\mu}_{hs} = \frac{\frac{1}{n_h}\sum_{i=1}^{n_h} S_{hs,i}}{\frac{1}{n_h}\sum_{i=1}^{n_h} R_{hs,i}}$$

$$\text{vec}(\hat{\Sigma}_{hs}) = \frac{1}{N_{hs}-1}\sum_{i=1}^{n_h} I(D_{hi}=s)(X_{hi}-\hat{\mu}_{hs})\otimes(X_{hi}-\hat{\mu}_{hs})^T$$

$$= \frac{N_{hs}}{N_{hs}-1}\left\{\frac{\frac{1}{n_h}\sum_{i=1}^{n_h} T_{hs,i}}{\frac{1}{n}\sum_{i=1}^{n_h} R_{hs,i}} - \frac{\frac{1}{n_h}\sum_{i=1}^{n_h} S_{hs,i}}{\frac{1}{n_h}\sum_{i=1}^{n_h} R_{hs,i}}\otimes\frac{\frac{1}{n_h}\sum_{i=1}^{n_h} S_{hs,i}}{\frac{1}{n_h}\sum_{i=1}^{n_h} R_{hs,i}}\right\}$$

where $N_{hs}=\Sigma_{i=1}^{n_h} I(D_{hi}=s)=\Sigma_{i=1}^{n_h} R_{hs,i}$.

We shall reformulate lemma 2.5.1 in a way which corresponds better with the formulation of lemma 2.6.2.

LEMMA 2.6.5.

$$\mathcal{L}n_h^{1/2}\begin{bmatrix} \hat{p}_{h1}-p_{h1} \\ \hat{\mu}_{h1}-\mu_{h1} \\ \text{vec}(\hat{\Sigma}_{h1})-\text{vec}(\Sigma_{h1}) \\ \cdot \\ \cdot \\ \cdot \\ \hat{p}_{hd}-p_{hd} \\ \hat{\mu}_{hd}-\mu_{hd} \\ \text{vec}(\hat{\Sigma}_{hd})-\text{vec}(\Sigma_{hd}) \end{bmatrix} \to N(0,B_h)$$

where B_h is partitioned as $B_h=(B_{h,ts},\ t=1,...,d; s=1,...,d)$, with $B_{h,ts}$ square matrices of size $1+p+p^2$, specified as follows. For $s=1,...,d$ the $B_{h,ss}$ are the blockdiagonal matrices

$$B_{h,ss} = diag\{p_{hs}(1-p_{hs}), \frac{1}{p_{hs}}\Sigma_{hs}, \frac{1}{p_{hs}}(I_{p^2}+K_p)(\Sigma_{hs}\otimes\Sigma_{hs})\}$$

and for $s=1,...,d; t=1,...,d; t\neq s$:

$$B_{h,ts} = -p_{ht}p_{hs}\epsilon_1\epsilon_1^T$$

where ϵ_1 *is the unit vector of size* $1+p+p^2$.

PROOF. For the proof of the asymptotic result of this lemma it is allowed to replace the factor $N_{hs}/(N_{hs}-1)$ in $\text{vec}(\hat{\Sigma}_{hs})$ by 1 because this factor is $1+o(1)$. After this replacement the $\hat{p}_{hs}$, $\hat{\mu}_{hs}$ and $\text{vec}(\hat{\Sigma}_{hs})$ are functions of

$$\frac{1}{n_h}\sum_{i=1}^{n_h} R_{hs,i}, \quad \frac{1}{n_h}\sum_{i=1}^{n_h} S_{hs,i} \quad \text{and} \quad \frac{1}{n_h}\sum_{i=1}^{n_h} T_{hs,i}.$$

In order to apply the δ-method of lemma 2.2.1 we shall define the function

$$g\colon \mathbb{R}^{d(1+p+p^2)} \to \mathbb{R}^{d(1+p+p^2)}$$

where

$$g=((g_{1,s}, g_{2,s}, g_{3,s}),\ s=1,\ldots,d)$$

with

$$g_{1,s}(u_1, v_1, w_1,\ldots,u_d, v_d, w_d) = u_s$$
$$g_{2,s}(u_1, v_1, w_1,\ldots,u_d, v_d, w_d) = u_s^{-1}v_s$$
$$g_{3,s}(u_1, v_1, w_1,\ldots,u_d, v_d, w_d) = u_s^{-1}w_s - u_s^{-2}v_s\otimes v_s$$

and $u_s\in\mathbb{R}^1$, $v_s\in\mathbb{R}^p$, $w_s\in\mathbb{R}^{p\times p}$, for $s=1,\ldots,d$.

Now, let the vector η_k be defined by

$$\eta_h = (p_{h1}, p_{h1}\mu_{h1}^T, p_{h1}(\text{vec}(\Sigma_{h1}) + \mu_{h1}\otimes\mu_{h1})^T,\ldots$$
$$\ldots,p_{hd}, p_{hd}\mu_{hd}^T, p_{hd}(\text{vec}(\Sigma_{hd}) + \mu_{hd}\otimes\mu_{hd})^T)^T$$

then the matrix of partial derivations of g at the vector η_h are given by $\nabla_{g(\eta_h)}=(\nabla_{g(\eta_h),ts}, t=1,\ldots,d; s=1,\ldots,d)$ with $\nabla_{g(\eta_h),ts}$ square matrices of size $1+p+p^2$, specified by

$$(\nabla_{g(\eta_h),ss})^T = \begin{pmatrix} 1 & -\dfrac{\mu_{hs}^T}{p_{hs}} & -\dfrac{1}{p_{hs}}\text{vec}^T(\Sigma_{hs}) + \dfrac{1}{p_{hs}}\mu_{hs}^T\otimes\mu_{hs}^T \\ 0 & \dfrac{I_p}{p_{hs}} & -\dfrac{1}{p_{hs}}(\mu_{hs}^T\otimes I_p) - \dfrac{1}{p_{hs}}(I_p\otimes\mu_{hs}^T) \\ 0 & 0 & \dfrac{1}{p_{hs}}I_{p^2} \end{pmatrix}$$

for $s=1,\ldots,d$ and $\nabla_{g(\eta_h),ts}=0$ for $t=1,\ldots,d; s=1,\ldots,d; t\neq s$. The variance-covariance matrix B_h can be derive from

$$B_h = \nabla_{g(\eta_h)}Q_h(\nabla_{g(\eta_h)})^T$$

where Q_h is defined in lemma 2.6.2. This completes the proof.

METHOD 3

This method will be based on the asymptotic distribution of maximum likelihood estimators. If the independent identically distributed random variables Y_i, $i=1, 2, 3,...$ have probability density function f_θ, with $\theta=(\theta,...,\theta_k)^T$, and $\hat{\theta}_n$ is the maximum likelihood estimator of θ based on $Y_1,...,Y_n$ then

$$\mathscr{L}n^{1/2}(\hat{\theta}_n-\theta) \to N(0,I_\theta^{-1})$$

where I_θ is the Fisher-information matrix defined by

$$(I_\theta)_{s,t} = E_\theta\left\{-\frac{\partial^2}{\partial\theta_s\partial\theta_t}\log f_\theta(Y_1)\right\} \quad s,\ t=1,...,k.$$

We shall prove that $\hat{p}_{h1},...,\hat{p}_{hd}$, $\hat{\mu}_{h1}$, $n_{h1}^{-1}(n_{h1}-1)\ \hat{\Sigma}_{h1},...,\hat{\mu}_{hd}$, $n_{hd}^{-1}(n_{hd}-1)\hat{\Sigma}_{hd}$ are the maximum likelihood estimators of $p_{h1},...,p_{hd}$, $\mu_{h1}, \Sigma_{h1},...,\mu_{hd}, \Sigma_{hd}$, respectively. For reasons of notational convenience we shall drop the index h. The above-mentioned asymptotic result is also valid if, instead of a probability density function, a Radon-Nikodym derivative with respect to a suitable measure is taken.

Let f be the Radon-Nikodym derivative of the distribution of (X_i, D_i), $i=1,...,n$ with respect to the product measure of Lebesgue measure and counting measure. Let (x_i, d_i) be the outcome of (X_i, D_i) then

$$f(x_i,d_i) = p_{d_i}f_{\mu_{d_i},\Sigma_{d_i}}(x_i)$$

in which the first factor is the probability that the discrete random variable has outcome d_i and in which the second factor is the value of the multivariate normal density with parameters μ_{d_i} and Σ_{d_i} at the vector x_i, $i=1,...,n$. The maximum likelihood estimators yield those parameters for which $\log\Pi_{i=1}^n f(x_i,d_i)$ is maximal. We have that

$$\begin{aligned}\log\prod_{i=1}^{n} f(x_i,\ d_i) &= \log\prod_{s=1}^{d}\prod_{i=1}^{n_s} p_{s_i} + \log\prod_{s=1}^{d}\prod_{i=1}^{n_s} f_{\mu_s,\Sigma_s}(x_{s_i}) \\ &= \log\prod_{s=1}^{d}\prod_{i=1}^{n_s} p_{s_i} + \sum_{s=1}^{d}\log\prod_{i=1}^{n_s} f_{\mu_s,\Sigma_s}(x_{s_i}) \\ &= L(p_1,...,p_d) + \sum_{s=1}^{d} L_s(\mu_s,\ \Sigma_s)\end{aligned}$$

where $s_1,...,s_{n_s}$ are those indices for which $d_{s_i}=s$, $i=1,...,n_s$. Let $t_1,...,t_{n_t}$ be the remaining indices: $d_{t_i}\neq s$, $i=1,...,n_t$, $n_s+n_t=n$. We shall now derive the maximum likelihood estimators $\hat{p}_s$, $\hat{\mu}_s$ and $\hat{\Sigma}_s$ for p_s, μ_s, and Σ_s, respectively, $s=1,...,d$.

$$\begin{aligned}&\frac{\partial}{\partial p_s}\log\prod_{i=1}^{n} f(x_i,\ d_i) = \\ &\quad\overset{(*)}{=}\frac{\partial}{\partial p_s}\log\prod_{s=1}^{d}\prod_{i=1}^{n_s} p_{s_i}\end{aligned}$$

$$= \frac{\partial}{\partial p_s} \sum_{i=1}^{n_s} \log p_{s_i} + \frac{\partial}{\partial p_s} \sum_{i=1}^{n_t} \log(1-p_{s_i})$$

$$= \frac{n_s}{p_s} - \frac{n_t}{1-p_s}.$$

Equating this last expression to zero gives

$$\hat{\hat{p}}_s = \frac{n_s}{n}, \quad s=1,...,d.$$

Note that the expression just after the $(\star)$ also appears when the maximum likelihood estimators of the parameters of a multinomial distribution are derived by means of differentiation. This implies that $L(\hat{p}_1,...,\hat{p}_d)$ is maximal.

$$\frac{\partial}{\partial \mu_s} \log \prod_{i=1}^{n} f(x_i, d_i) =$$

$$\overset{(\star)}{=} \frac{\partial}{\partial \mu_s} \log \prod_{i=1}^{n_s} f_{\mu_s, \Sigma_s}(x_{s_i})$$

$$= \sum_{i=1}^{n_s} \frac{\partial}{\partial \mu_s} \{\log|2\pi\Sigma_s|^{-1/2} - \tfrac{1}{2}(x_{s_i}-\mu_s)^T \Sigma_s^{-1}(x_{s_i}-\mu_s)\}$$

$$= \sum_{i=1}^{n_s} \Sigma_s^{-1}(x_{s_i}-\mu_s).$$

Equating this to zero gives

$$\hat{\hat{\mu}}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} x_{s_i}, \quad s=1,...,d.$$

Further

$$\frac{\partial}{\partial \Sigma_s} \log \prod_{i=1}^{n_s} f(x_i, d_i) =$$

$$\overset{(\star)}{=} \frac{\partial}{\partial \Sigma_s} \log \prod_{i=1}^{n_s} f_{\mu_s, \Sigma_s}(x_{s_i})$$

$$= \sum_{i=1}^{n_s} \frac{\partial}{\partial \Sigma_s} \{\log|2\pi\Sigma_s|^{-1/2} - \tfrac{1}{2}(x_{s_i}-\mu_s)^T \Sigma_s^{-1}(x_{s_i}-\mu_s)\}$$

$$= -\tfrac{1}{2} n_s \Sigma_s^{-1} + \tfrac{1}{2}\Sigma_s^{-1} \Big(\sum_{i=1}^{n_s} (x_{s_i}-\mu_s)(x_{s_i}-\mu_s)^T \Big) \Sigma_s^{-1}.$$

Equating this to zero and substituting $\hat{\hat{\mu}}_s$ for μ_s gives

$$\hat{\hat{\Sigma}}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} (x_{s_i}-\mu_s)(x_{s_i}-\mu_s)^T, \quad s=1,...,d.$$

Now, note that the expressions just after the last two $(\star)$'s also appear when

the maximum likelihood estimators of the mean and covariance matrix of a multivariate normal distribution are derived by means of differentiation. Hence $L_s(\hat{\mu}_s, \hat{\hat{\Sigma}}_s)$ is maximal. See the proof of WATSON (1964), used for example in MUIRHEAD (1982), p.84 and RAO (1973), p. 531. Thus we have that

$$L(\hat{p}_1,...,\hat{p}_d) + \sum_{s=1}^{d} L_s(\hat{\mu}_s, \hat{\hat{\Sigma}}_s)$$

is maximal. Hence $\hat{p}_s, \hat{\mu}_s$, and $\hat{\hat{\Sigma}}_s$, $s=1,...,d$ are maximum likelihood estimators. Because they are equal to $\hat{p}_s$, $\hat{\mu}_s$, and $\hat{\Sigma}_s$, $s=1,....,d$ we have proved that these last-mentioned are maximum likelihood estimators.

Let us now consider the Fisher-information matrix of a multinomial distribution and of a multivariate normal distribution. For a multinomial distribution with vector of parameters $\mathrm{p}=(p_1,...,p_d)^T$ and vector of maximum likelihood estimators $\hat{\mathrm{p}}_n$ we have

$$\mathscr{L}n^{1/2}(\hat{\mathrm{p}}_n - \mathrm{p}) \to N(0, I_{\mathrm{p}}^{-1})$$

where

$$I_{\mathrm{p}}^{-1} = B - \mathrm{pp}^T \quad \text{with} \quad B = diag(p_1,...,p_d).$$

For a multivariate normal distribution with mean μ_s, covariance matrix Σ_s, using the notation $\lambda_s=(\mu_s^T, \mathrm{vec}^T(\Sigma_s))^T$ and $\hat{\lambda}_{s,n}$, the corresponding vector of maximum likelihood estimators, we have

$$\mathscr{L}n^{1/2}(\hat{\lambda}_{s,n} - \lambda_s) \to N(0, I_{\lambda_s}^{-1})$$

where

$$I_{\lambda_s}^{-1} = \begin{bmatrix} \Sigma_s & 0 \\ 0 & (I_{p^2} + K_p)(\Sigma_s \otimes \Sigma_s) \end{bmatrix}$$

with K_p defined after (2.4.12).

We shall show that the covariance matrix in lemma 2.5.1 is the inverse of the Fisher-information matrix

$$I_\Delta = E_\Delta\{\frac{-\partial^2}{\partial_\Delta \partial_\Delta^T} \log f(X_1, D_1)\}$$

where $\Delta=(\mathrm{p}^T, \lambda_1^T,...,\lambda_d^T)^T$. We shall drop the index 1 in X_1 and D_1. We have that

$$\begin{aligned} I_\Delta &= E_D E_{X|D=s}\left\{\frac{-\partial^2}{\partial_\Delta \partial_\Delta^T}(\log f_{\mu_s, \Sigma_s}(X) + \log p_s)\right\} \\ &= E_D E_{X|D=s}\left\{\frac{-\partial^2}{\partial_\Delta \partial_\Delta^T}\log f_{\mu_s, \Sigma_s}(X)\right\} + E_D\left\{\frac{-\partial^2}{\partial_\Delta \partial_\Delta^T}\log p_D\right\}. \end{aligned}$$

The last term is a matrix with zeros everywhere, except for the upper-left block of size $d \times d$. That block is equal to the Fisher-information matrix I_{p} of the

above-mentioned multinomial distribution. For the first term we need the following four matrices

(a) $$E_D E_{X|D=s}\left\{\frac{-\partial^2}{\partial \mathrm{p}\partial \mathrm{p}^T}\log f_{\mu_s,\Sigma_s}(X)\right\} = 0.$$

(b) $$E_D E_{X|D=s}\left\{\frac{-\partial^2}{\partial \mathrm{p}\partial \lambda_v^T}\log f_{\mu_s,\Sigma_s}(X)\right\} = 0 \quad v = 1,\ldots,d.$$

(c) $$E_D E_{X|D=s}\left\{\frac{-\partial^2}{\partial \lambda_v \partial \lambda_v^T}\log f_{\mu_s,\Sigma_s}(X)\right\}$$

$$= E_D E_{X|D=s}\left\{\delta_{sv}\frac{-\partial^2}{\partial \lambda_v \partial \lambda_v^T}\log f_{\mu_s,\Sigma_s}(X)\right\}$$

$$= E_D \delta_{Dv} E_{X|D=v}\left\{\frac{-\partial^2}{\partial \lambda_v \partial \lambda_v^T}\log f_{\mu_v,\Sigma_v}(X)\right\}$$

$$= p_v E_{X|D=v}\left\{\frac{-\partial^2}{\partial \lambda_v \partial \lambda_v^T}\log f_{\mu_v,\Sigma_v}(X)\right\}$$

$$= p_v I_{\lambda_v} \quad v = 1,\ldots,d$$

where δ is the Kronecker-symbol and I_{λ_v} is the Fisher-information matrix of the multivariate normal distribution with vector of parameters λ_v.

(d) $$E_D E_{X|D=s}\left\{\frac{-\partial^2}{\partial \lambda_w \partial \lambda_v^T}\log f_{\mu_s,\Sigma_s}(X)\right\} = 0, \quad v, w = 1,\ldots,d; v \neq w.$$

So, the Fisher-information matrix I_Δ is the block-diagonal matrix with blocks $I_{\mathrm{p}}, p_1 I_{\lambda_1},\ldots,p_d I_{\lambda_d}$. The inverse I_Δ^{-1} is the covariance matrix in lemma 2.5.1. This completes the proof.

Chapter 3

Incorporating standard errors of posterior probabilities in decision-making processes

3.1. INTRODUCTION

In this chapter we shall describe how the results of chapter 2 can be incorporated in decision-making processes. We suppose that there are k populations with corresponding densities $f_{h,\theta}, h=1,...,k$ where θ is an element of the parameter set Θ. Let X denote the random variable which generates the vector of scores x. We assume that X has density $f_{t,\theta}$ if the vector of scores belongs to population t. Let $X_{11},...,X_{kn_k}$ denote the random variables which generate the training samples $x_{11},...,x_{kn_k}$. For these random variables we assume that X_{hj} has density $f_{h,\theta}$. Moreover $X, X_{11},...,X_{kn_k}$ are supposed to be independent. If the parameter θ is unknown, then the sample space is denoted by $\mathcal{X}^{n+1}$, which is the product of the outcome spaces of $X, X_{11},...,X_{kn_k}$ where $n=\Sigma_{h=1}^{n_k}$. If the parameters are known, then the training samples are not needed with as a consequence that after their deletion, the sample space is denoted by $\mathcal{X}$, i.e. the outcome space of X. In that case we will often write f_h instead of $f_{h,\theta}, h=1,...,k$. Further we assume that an action set $\mathcal{A}=\{a_1,...,a_m\}$ and a loss function $L(t,a)$, which is a function of t and a, have been specified. The unknown number t of the population from which x is a random drawing is usually considered to be the outcome of a random variable T with values in the set $\{1,...,k\}$. The probabilities $P(T=t)=\rho_t, t=1,...,k$ are the prior probabilities. We shall assume that numerical values of $\rho_1,...,\rho_k$ are given. We consider the posterior probabilities

$$\rho_{t|x} = \rho_t f_{t,\theta}(x) / \sum_{h=1}^{k} \rho_h f_{h,\theta}(x) \quad t=1,...,k.$$

If the parameter θ and the prior probabilities $\rho_1,...,\rho_k$ are known, then the

optimal decision rule assigns to x that action a for which the conditional expected loss

$$E\{L(T,a)|X=x\} = \sum_{t=1}^{k} L(t,a)\rho_{t|x}$$

is minimal (non-uniqueness, allowing randomization, appears if there is more than one action for which this minimum is obtained). This is the Bayes rule. More about this approach of constructing multiple decision rules when population densities are known can be found in section 3.2.

In practice, however, the parameter θ is unknown. In section 3.3 we shall consider an approach in which the conditional expected losses $E\{L(T,a_j)|X=x\}, j=1,...,m$ are regarded as unknown parameters which have to be estimated from the training samples. For expressing the statistical uncertainties in the estimates, the theory of chapter 2 can be used.

If one is forced to take a decision and wants to comply with this demand in a rational manner, than one will need a procedure which prescribes the action to be chosen. Such decision rules are studied in section 3.4.

While the theory in the sections 3.2, 3.3 and 3.4 has a classical-statistical touch, that in section 3.5 deals with the so-called fully Bayesian approach. Here priors are not only postulated on $\{1,...,k\}$ but also on Θ. This requires a more or less subjectivistic attitude.

3.2. TAKING DECISIONS IF PARAMETERS ARE KNOWN

Model 1 of this section concerns the situation in which t is regarded as an unknown parameter assuming a value in the parameter set $\{1,...,k\}$. In model 2 class number t is regarded as the outcome of random variable T with $P(T=t)=\rho_t, t=1,...,k$ being given in advance. In any case the specific class densities are assumed to be known.

Model 1. Suppose that the number t has to be considered as an unknown parameter because no real meaning is involved in the randomness of T, or because no reasonable information is available with respect to $\rho_1,...,\rho_k$. Let the vector of scores x be the outcome of the random variable X which has density f_t where t is an unknown element of $\{1,...,k\}$. Let the sample space $\mathscr{X}$ be the set of all possible outcomes of X. Defined $\mathscr{A}$ as the set of m possible actions and let $L(t,a)$ be the loss if action a is taken while t is the true but unknown population number. Let the function $d:\mathscr{X}\rightarrow\mathscr{A}$ be a nonrandomized decision rule. The risk of decision rule d, as a function of the unknown number t, is given by

$$R(t,d) = E_t L(t,d(X)) = \sum_{j=1}^{m} L(t,a_j)P_t(R_j) = \sum_{j=1}^{m} L(t,a_j)\int_{R_j} f_t(x)dx$$

where the $R_j=\{x\in\mathscr{X};\ d(x)=a_j\}$, $j=1,...,m$ satisfy $R_1\cup\cdots\cup R_m=\mathscr{X}$, and

$R_i \cap R_j = \varnothing$ if $i \neq j$.

Model 2. Let t be the outcome of the random variable T where the prior probabilities $P(T=t)=\rho_t$, $t=1,...,k$ are given. The sample space $\mathcal{X}$, the action set $\mathcal{A}$ and the loss function L are suppose to be the same as in model 1. Let d be a nonrandomized decision rule. Use the notation ρ for the prior distribution $\rho_1,...,\rho_k$. The Bayes risk of d with respect to ρ is defined as

$$r(\rho,d) = \sum_{h=1}^{k} \rho_h R(h,d)$$

which is a weighted average of the values of the risk function defined in model 1. An obvious interpretation is that

$$r(\rho,d) = EE\{L(T,d(X))|T\} = EL(T,d(X))$$

is the overall expected loss if d is applied.

Procedures. For both models larger classes of decision rules can be defined by introducing the concept of randomization. For model 2 this is not very useful because, for given ρ, the Bayes risk can be minimized by choosing an appropriate nonrandomized rule (see FERGUSON (1967). p. 43). Thus the following remarks about randomized rules are only of practical interest for model 1. Two different randomization techniques can be distinguished. On the one hand we have the randomized decision rules corresponding with probability distributions on the class of nonrandomized decision rules. On the other hand we can assign to every element $x \in \mathcal{X}$ a distribution over the set $\mathcal{A}$ of possible actions. They are called behavioral decision rules. If the action space $\mathcal{A}$ consists of a finite number of elements, say m, then a behavioral rule can obviously be characterized by $\phi(x)=(\phi(1|x),...,\phi(m|x))$ where $\phi(j|x)$ is the probability that action a_j is taken after $x \in \mathcal{X}$ has been observed and $\Sigma_{j=1}^{m}\phi(j|x)=1$. See BLACKWELL and GIRSHICK (1954) for an extensive treatment and proof of the essential equivalence of both types of randomization.

Comparing procedures for model 1. The following fundamental concepts come from Wald's general theory of statistical decision functions. However, here they are adapted to the specifications of model 1. References are WALD (1950), LEHMANN (1950, 1959), ANDERSON (1958), and FERGUSON (1967).

Let d and d' be two decision rules. We say that d' is *as good as* d if $R(t,d') \leqslant R(t,d)$, $t=1,...,k$, and that d' is *better than* d if $R(t,d') \leqslant R(t,d)$, $t=1,...,k$ while $R(t,d')<R(t,d)$ for at least one $t \in \{1,...,k\}$. A decision rule d is called *admissible* if there is no decision rule better than d. A class of rules is called *complete* if for every rule outside the class there is a rule in the class which is better. A class of rules is called *essentially complete* if for any rule d outside the class there is one in the class which is as good as d.

The class of nonrandomized decision rules can be considered a subset of both the class of nonrandomized decision rules and the class of behavioral decision rules. For most of the situations to be considered, the class of

nonrandomized rules is sufficiently large. According to DVORESTSKY, WALD and WOLFOWITZ (1951) (see FERGUSON (1967) p. 79), the class of nonrandomized rules is essentially complete if the parameter set and the action set are both finite, and the probability distribution of X has no point masses.

Before constructing procedures for model 1, we focus on the following problem.

Constructing the Bayes rule for model 2. It is natural to construct the Bayes rule d_ρ, with respect to the prior distribution ρ, i.e. to construct the rule which minimizes the expected risk $r(\rho,d)$:

$$r(\rho,d_\rho) = \inf_d r(\rho,d).$$

This rule d_ρ can be obtained by conditioning with respect to the observed data. For that purpose note that

$$\begin{aligned} r(\rho,d) &= EL(T,d(X)) \\ &= E\{EL(T,d(X))|X\} \end{aligned}$$

where, for a particular value of x, the integrand

$$EL(T,d(X))|X=x) = \sum_{t=1}^{k} L(t,d(x))\rho_{t|x}$$

is the conditional expected loss. Hence d_ρ is obtained by defining that for any fixed $x \in \mathcal{X}$ the value $d_\rho(x)$ is that action $a \in \mathcal{A}$ which minimizes the conditional expected loss $EL((T,a)|X=x)$. Thus the Bayes rule not only minimizes $r(\rho,d)$, it also minimizes the conditional expected loss given any outcome x of X.

Bayes rules for model 2 are admissible if all prior probabilities are positive (FERGUSON (1967), p. 60).

The Bayes rule constructed by minimizing the conditional expected loss $\Sigma_{t=1}^{k} L(t,d(x))\rho_{t|x}$ can equally well be obtained by minimizing the expression

$$\sum_{t=1}^{k} L(t,d(x)\rho_t f_t(x)$$

because the denominator $f(x)=\Sigma_{h=1}^{k}\rho_h f_h(x)$ of $\rho_{t|x}$ plays no part in the minimization process. The sample space $\mathcal{X}$ is partitioned by the Bayes rule d_ρ into m regions R_j, $j=1,...,m$ where (apart from non-uniqueness complications)

$$R_j = \{x; \sum_{t=1}^{k} L(t,a_j)\rho_{t|x} = \min_i (\sum_{t=1}^{k} L(t,a_i)\rho_{t|x})\}.$$

EXAMPLE. Consider the special case of 0−1 loss, i.e. $m=k$, $L(t,a_j)=1$ if $t\neq j$ and 0 if $t=j$. The Bayes rule is obtained if

$$R_j = \{x; \sum_{t=1;t\neq j}^{k} \rho_t f_t(x) = \min_i (\sum_{t=1;t\neq i}^{k} \rho_t f_t(x))\}$$

$$= \{x;\ \rho_j f_j(x) = \max_i (\rho_i f_i(x))\}$$

$$= \{x;\rho_{j|x} = \max_i (\rho_{i|x})\}.$$

Thus, the Bayes rule assigns the individual under investigation to the population with maximal posterior probability.

EXAMPLE. A special application is that in which the probability densities are those of multivariate normal distributions:

$$f_t(x) = |2\pi\Sigma_t|^{-1/2} \exp\{-\tfrac{1}{2}(x-\mu_t)^T\Sigma_t^{-1}(x-\mu_t)\}.$$

The Bayes rule is given by the regions

$$R_j = \{x;\ \log\rho_j - \tfrac{1}{2}\log(\det(\Sigma_j)) - \tfrac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j) \text{ maximal}\}.$$

If covariance matrices are equal these regions become

$$R_j = \{x;\ \log\rho_j - \tfrac{1}{2}(x-\mu_j)^T\Sigma^{-1}(x-\mu_j) \text{ maximal}\}.$$

If, in addition, the prior probabilities are equal these become

$$R_j = \{x;\ (x-\mu_j)^T\Sigma^{-1}(x-\mu_j) \text{ minimal}\}.$$

This means that the assignment is to that population which has the smallest Mahalanobis distance from observation x. It is easy to see that equality of the covariance matrices implies that the boundaries of the regions $R_j, j = 1,...,k$ are linear. If the covariance matrices are not equal then these boundaries have a quadratic form.

Constructing minimax rules for model 1. For model 1 we shall consider the minimax rule d^*. This rule is defined by

$$\max_t R(t,d^*) = \inf_d \max_t R(t,d)$$

where d runs over the set of all possible randomized rules. Minimax strategies were introduced in game theory by VON NEUMANN and MORGENSTERN (1944). WALD (1950) extended their ideas by regarding the theory of statistics as a game of the statistician (player 2) against nature (player 1). Clear descriptions can be found in ANDERSON (1958), FERGUSON (1967), LEHMANN (1959), etc.

A decision rule which has the same risk $R(t,d)$ for all points of the parameter set is called an equalizer rule. If an equalizer rule is Bayes then it is a minimax rule (see FERGUSON (1967), p. 91). We saw that the Bayes rule d_ρ with respect to some prior ρ can be identified with m regions $R_1(\rho),...,R_m(\rho)$ in the sample space $\mathcal{X}$. If a (least favorable) prior distribution ρ can be found such that

$$R(t,d_\rho) = \sum_{j=1}^{m} L(t,a_j)P_t(R_j(\rho))$$

does not depend on t, then d_ρ is an equalizer rule, Bayes rule, and hence minimax rule.

In the special case of $0-1$ loss ($m=k$) the risk reduces to

$$R(t,d_\rho) = 1-P_t(R_t(\rho)), \quad t=1,...,k.$$

In this case d_ρ is minimax if ρ is such that $P_t(R_t(\rho))$ does not depend on $t\in\{1,...,k\}$.

EXAMPLE. Let $f_1(x)$ be the density of $N(0,1)$ and $f_2(x)$ that of $N(\frac{1}{2},2)$. Let the loss be $0-1$ and $m=k=2$.

(a). The Bayes rule for prior distribution $\rho=(\rho_1,\rho_2)$ is given by the partition $\{R_1(\rho), R_2(\rho)\}$ of $\mathbb{R}$ where

$$R_1(\rho) = \{x;\ \log\rho_1 - \tfrac{1}{2}x^2 > \log\rho_2 - \tfrac{1}{2}\log 2 - \tfrac{1}{2}\cdot\tfrac{1}{2}\cdot(x-\tfrac{1}{2})^2\}$$

$$R_2(\rho) = \mathbb{R}-R_1(\rho).$$

Using the notation

$$\Delta(\rho) = \{\tfrac{1}{2}+2\log 2+4\log(\rho_1/\rho_2)\}^{1/2}$$

we have

$$R_1(\rho) = \{x;\ -\tfrac{1}{2}-\Delta(\rho) < x < -\tfrac{1}{2}+\Delta(\rho)\}.$$

(b). The Bayes rule of (a) is an equalizer, and hence minimax rule if the following probabilities are equal

$$P_1(R_1(\rho)) = \Phi(-\tfrac{1}{2}+\Delta(\rho))-\Phi(-\tfrac{1}{2}-\Delta(\rho))$$

and

$$P_2(R_2(\rho)) = \Phi(-2^{-\frac{1}{2}}(1+\Delta(\rho)))+1-\Phi(-2^{-\frac{1}{2}}(1-\Delta(\rho))).$$

(c). The minimax rule among the rules which have only one cutting point can be obtained as follows. Let d_c be the rule with cutting point c, i.e.

$$R_{1,c} = \{x;\ x<c\} \quad \text{and} \quad R_{2,c} = \{x;x\geqslant c\}.$$

The risks are

$$R(1,d_c) = P_1(R_{2,c}) = 1-P_1(R_{1,c}) = 1-\Phi(c)$$

and

$$R(2,d_c) = P_2(R_{1,c}) = 1-P_2(R_{2,c}) = \Phi(2^{-\frac{1}{2}}(c-\tfrac{1}{2})).$$

The required minimax rule is obtained for that c which is the unique solution of $R(1,d_c)=R(2,d_c)$. This gives $c=.21$.

3.3. DISCUSSING THE CHOICE OF DECISION IF PARAMETERS HAVE TO BE ESTIMATED

In this section we shall consider the realistic situation in which the k densities are unknown while T has the prior distribution defined by $P(T=t)=\rho_t$, $t=1,...,k$, numerical values of the ρ_t being specified. We shall focus on the choice of decision. In section 3.2 this was done of model 2 in which the θ's were known. The present section contains some adequate supplementaries when the θ's are unknown. The sample space $\mathcal{X}^{n+1}$ is the space of outcomes of $X,X_{11},...,X_{kn_k}$ where X generates the vector of scores x of the individual under investigation and X_{hi} generates the vector of scores x_{hi}, $i=1,...,n_h$; $h=1,...,k$. All these $1+\Sigma_{h=1}^{k}n_h$ variables are considered independent given θ and $\{T=t\}$. The random variable X_{hi} has density $f_{h,\theta}$ where θ is the unknown parameter. The density of X is given by the simultaneous distribution of (X,T) where $X|T=t$ has density $f_{t,\theta}$. The action set is denoted by $\mathcal{A}=\{a_1,...,a_m\}$. The loss function is given by $L(t,a)$. As usual, let the posterior probabilities be denoted by

$$\rho_{t|x}(\theta) = \rho_t f_{t,\theta}(x) / \sum_{h=1}^{k} \rho_h f_{h,\theta}(x) \ , \ h=1,...,k.$$

We shall use the notation $\rho_{t|x}$ instead of the longer notation $\rho_{t|x}(\theta)$, although this latter expresses explicitly that the posterior probabilities depend on the unknown parameter θ. The conditional expected loss, given the observation x, if action a_j is chosen, is

$$E\{L(T,a_j)|X=x\} = \sum_{t=1}^{k} L(t,a_j)\rho_{t|x}, \quad j=1,...,m.$$

In practice, the parameter θ has to be estimated from the training samples x_{hi}, $i=1,...,n_h$; $h=1,...,k$. The uncertainty in the estimate to be obtained causes uncertainties in the population densities, the posterior probabilities, and the conditional expected losses. One of the aims of this thesis is to provide means to express these statistical uncertainties. These or similar means should be applied unless the statistical uncertainties are negligible.

Being interested in statistical inference with respect to the $\Sigma_{t=1}^{k}L(t,a_j)\rho_{t|x}$ we focus on the estimates

$$\sum_{t=1}^{k} L(t,a_j)\hat{\rho}_{t|x} \ , \ j=1,...,m.$$

In the previous chapter asymptotic distributions were derived under various model assumptions. These results are of the form

$$\mathcal{L}n^{1/2}(R_{\cdot|x}-\rho_{\cdot|x}) \rightarrow N_k(0,\Sigma_{\cdot|x})$$

where $\Sigma_{\cdot|x}$ depends on the choice of the model. Let L be the $k \times m$ matrix with $L_{tj} = L(t,a_j)$, then

$$\mathcal{L} n^{1/2}(L^T R_{\cdot|x} - L^T \rho_{\cdot|x}) \to N_m(0, L^T \Sigma_{\cdot|x} L)$$

where

$$(L^T R_{\cdot|x})_j = \sum_{t=1}^{k} L(t,a_j)\hat{\rho}_{t|x}$$

is the estimator of the conditional expected loss

$$(L^T \rho_{\cdot|x})_j = \sum_{t=1}^{k} L(t,a_j)\rho_{t|x}$$

if action a_j is chosen ($j = 1,...,m$). It is obvious that the corresponding covariance matrix $n^{-1} L^T \Sigma_{\cdot|x} L$ should play a part in the considerations if the statistical uncertainties expressed by this matrix cast reasonable doubt on statements which one would like to make.

Testing, ranking, and selection techniques. Various kinds of considerations can be based on the asymptotic distribution of $L^T R_{\cdot|x}$. After having obtained realisations of the estimators of the posterior probabilities and the covariance matrix, we can forget the original context and base our considerations entirely on these realisations. We can apply various techniques of testing hypotheses, ranking actions, selecting actions, etc. In our approach the posterior probabilities and the conditional expected losses are regarded as parametric functions to be estimated. Distributions of their estimators have been derived. These results can be used for testing statistical hypotheses about these parametric functions.

The asymptotic distribution of $n^{1/2}(R_{\cdot|x} - \rho_{\cdot|x})$ is the multivariate normal distribution $N_k(0, \Sigma_{\cdot|x})$ on the $k-1$ dimensional hyperplane

$$L_k = \{x;\ x \in \mathbb{R}^k\ ,\ x^T I = 0,\ I = (1,...,1)^T\}$$

in the k dimensional space $\mathbb{R}^k$. This implies that the asymptotic distribution of $n^{1/2}(L^T R_{\cdot|x} - L^T \rho_{\cdot|x})$ is concentrated on a hyperplane of dimension at most $\min(k-1, m)$ in $\mathbb{R}^k$. The actual dimension depends on the structure of the loss matrix L. The just-mentioned multivariate normal distributions can be used to obtain approximations to the exact distributions of $R_{\cdot|x}$ and $L^T R_{\cdot|x}$. We have the approximations:

$$R_{\cdot|x} \sim N(\rho_{\cdot|x}, \frac{1}{n}\Sigma_{\cdot|x})$$

and

$$L^T R_{\cdot|x} \sim N(L^T \rho_{\cdot|x}, \frac{1}{n} L^T \Sigma_{\cdot|x} L).$$

The sets on which the exact distributions are defined are easily indicated as follows. Let $e_j, j = 1,...,k$ be the j-th unit vector in $\mathbb{R}^k$. Let $l_j, j = 1,...,k$

denote the j-th row in the loss matrix L. The distribution of $R_{\cdot|x}$ is concentrated on the so-called unit simplex

$$U_k = \{y;\, y \in \mathbb{R}^k,\, y = \sum_{i=1}^{k} \alpha_i e_i,\, \alpha_i \geqslant 0,\, i = 1,\ldots,k,\, \sum_{i=1}^{k} \alpha_i = 1\}$$

and the distribution of $L^T R_{\cdot|x}$ on

$$H_{K,L} = \{z;\; z \in \mathbb{R}^m,\, z = \sum_{i=1}^{k} \alpha_i l_i,\, \alpha_i \geqslant 0,\, i = 1,\ldots,k,\; \sum_{i=1}^{k} \alpha_i = 1\}.$$

Confidence regions for the unknown parameter $\rho_{\cdot|x}$ can be constructed by means of isodensity ellipsoids. Assume that $\hat{\Sigma}_{\cdot|x}$ is the estimate of $\Sigma_{\cdot|x}$. Note that its rank is equal to $k-1$. Because $\hat{\Sigma}_{\cdot|x}$ is symmetric, there exists a $k \times k$ orthogonal matrix U, i.e. $UU^T = U^T U = I$, such that

$$U^T \hat{\Sigma}_{\cdot|x} U = \begin{bmatrix} \hat{\Lambda}_{\cdot|x} & 0 \\ 0 & 0 \end{bmatrix}$$

where $\hat{\Lambda}_{\cdot|x}$ is a $(k-1) \times (k-1)$ diagonal matrix. Let

$$V = (U^T R_{\cdot|x})^{(1)}$$

be the vector of the first $k-1$ components of $U^T R_{\cdot|x}$. Then

$$S = \{v:\, v \in \mathbb{R}^{k-1},\, n(V-v)^T \hat{\Lambda}_{\cdot|x}^{-1} (V-v) \leqslant \chi^2_{k-1}(\alpha)\}$$

is an approximate confidence ellipsoid for $U^T \rho_{\cdot|x}$ of significance level $1-\alpha$. Here $\chi^2_{k-1}(\alpha)$ is the point with $P(G > \chi^2_{k-1}(\alpha)) = \alpha$ if $G \sim \chi^2_{k-1}$. Hence

$$\{r:\, r \in U_k\,,\, (U^T r)^{(1)} \in S\},$$

where $(U^T r)^{(1)}$ is the vector of the first $k-1$ components of $U^T r$, is an approximate confidence ellipsoid for $\rho_{\cdot|x}$.

In an analogous way confidence ellipsoids for the vector of conditional expected losses $L^T \rho_{\cdot|x}$ can be constructed. Let Y be an orthogonal matrix of size $m \times m$ such that

$$Y^T L^T \hat{\Sigma}_{\cdot|x} L Y = \begin{bmatrix} \hat{\Gamma}_{\cdot|x} & 0 \\ 0 & 0 \end{bmatrix}$$

where $\hat{\Gamma}_{\cdot|x}$ is a diagonal matrix of order $\ell = \min(k-1, m)$. Let

$$Z = (Y^T L^T R_{\cdot|x})^{(1)}$$

be the vector of the first ℓ components of $Y^T L^T R_{\cdot|x}$. Then

$$S_L = \{z:\, z \in \mathbb{R}^{\ell},\, n(Z-z)^T \hat{\Gamma}_{\cdot|x}^{-1} (Z-z) \leqslant \chi^2_{\ell}(\alpha)\}$$

is an approximate confidence ellipsoid for the parameter $Y^T L^T \rho_{\cdot|x}$ of significance level $1-\alpha$. Hence

$$\{\omega:\, \omega \in H_{K,L}\,,\, (Y^T \omega)^{(1)} \in S_L\}$$

where $(Y^T \omega)^{(1)}$ is the vector of the first ℓ components of $Y^T \omega$, is an

approximate confidence ellipsoid for $L^T\rho_{\cdot|x}$.

In practice, instead of the confidence ellipsoids for the whole vector $\rho_{\cdot|x}$ or $L^T\rho_{\cdot|x}$, confidence intervals for the separate posterior probabilities $\rho_{t|x}, t=1,...,k$ or conditional expected losses $(L^T\rho_{\cdot|x})_j, j=1,...,m$ will suffice. One can use

$$R_{t|x} \pm u_{\frac{1}{2}\alpha} n^{-1/2}\{(\hat{\Sigma}_{\cdot|x})_{tt}\}^{1/2}$$

and

$$(L^T R_{\cdot|x})_j \pm u_{\frac{1}{2}\alpha} n^{-1/2}\{(L^T\hat{\Sigma}_{\cdot|x}L)_{jj}\}^{1/2}$$

as confidence intervals of confidence level $1-\alpha$ for $\rho_{t|x}$ and $(L^T\rho_{\cdot|x})_j$, respectively, $t=1,...,k; j=1,...,m$. Here $u_{\frac{1}{2}\alpha}$ is the point with $P(U>u_{\frac{1}{2}\alpha})=\frac{1}{2}\alpha$ if $U\sim N(0,1)$. Because these intervals are derived from the limiting multivariate normal distribution, it can happen that the intervals are not concentrated in the zero-one interval. Various transformations are applicable to overcome this trouble. We present here an adapted version of Fisher's variance stabilizing transformation, $f(\rho)=\frac{1}{2}\ln\frac{1+\rho}{1-\rho}=\operatorname{arctanh}(\rho)$ where $-1\leqslant\rho\leqslant+1$, for the correlation coefficient computed from a sample from a bivariate normal distribution (see WITTING and NÖLLE (1970), p. 52). Let

$$g(x) = \tfrac{1}{2}\ln\frac{x}{1-x} \quad \text{i.e.} \quad x = \tfrac{1}{2}+\tfrac{1}{2}\tanh(g(x))$$

then

$$\mathscr{L}n^{1/2}(g(R_{t|x})-g(\rho_{t|x}))\to N(0,\frac{(\Sigma_{\cdot|x})_{tt}}{4\rho_{t|x}^2(1-\rho_{t|x})^2}).$$

From a $100(1-\alpha)\%$ confidence interval for $g(\rho_{t|x})$ a $100(1-\alpha)\%$ confidence interval for $\rho_{t|x}$ is obtained by

$$\tfrac{1}{2}+\tfrac{1}{2}\tanh\{\operatorname{arctanh}(2R_{t|x}-1)\pm u_{\frac{1}{2}\alpha}\frac{n^{-1/2}\{(\hat{\Sigma}_{\cdot|x})_{tt}\}^{1/2}}{2R_{t|x}(1-R_{t|x})}\}.$$

These intervals have the property that for point estimates $R_{t|x}=\frac{1}{2}$ the intervals are symmetric around $R_{t|x}$. For other point estimates the intervals are asymmetric around $R_{t|x}$, stretching out to either the boundary 0 or 1, whichever one lies farthest away from the point estimate.

If $k=2$ then a confidence interval for the log-odds $\log(f_1(x)/f_2(x))$ can be used, because the posterior probability

$$\rho_{2|x} = \{1+(\rho_1/\rho_2)\exp(\log(f_1(x)/f_2(x)))\}^{-1}$$

is a function of the log-odds. A confidence interval for the posterior probability is a one-to-one transformation of the one for the log-odds and it is a subinterval of $(0,1)$ (see also SCHAAFSMA and VAN VARK (1979). CRITCHLEY, FORD

(1984a, 1984b 1985), CRITCHLEY, FORD and RIJAL (1987, 1988), CRITCHLEY, FORD and HIRST (1987, 1988) and RIJAL (1984) focussed on the interval estimation of the log-odds instead of the interval estimation of the posterior probabilities.

In some situations it is interesting to judge certain specific linear combinations $a^T\rho_{\cdot|x}$ and $b^TL^T\rho_{\cdot|x}$ of posterior probabilities and conditional expected losses. The corresponding $100(1-\alpha)\%$ confidence intervals are

$$a^TR_{\cdot|x} \pm u_{\frac{1}{2}\alpha}n^{-1/2}\{a^T\hat{\Sigma}_{\cdot|x}a\}^{1/2}$$

and

$$b^TL^TR_{\cdot|x} \pm u_{\frac{1}{2}\alpha}n^{-1/2}\{b^TL^T\hat{\Sigma}_{\cdot|x}Lb\}^{1/2},$$

respectively. A special case is testing whether a difference exists between two particular posterior probabilities, or between two conditional expected losses. If, for example, the i-th and j-th posterior probability are considered, then the hypothesis $H:\rho_{i|x}=\rho_{j|x}$ is rejected in favour of $A:\rho_{i|x}\neq\rho_{j|x}$, if $R_{i|x}-R_{j|x}$ is not contained in the interval

$$0 \pm u_{\frac{1}{2}\alpha}n^{-1/2}\{(\hat{\Sigma}_{\cdot|x})_{ii} + (\hat{\Sigma}_{\cdot|x})_{jj} - 2(\hat{\Sigma}_{\cdot|x})_{ij}\}^{1/2}.$$

An interesting question, from a statistical point of view, is whether the population with the largest estimated posterior probability coincides with the population with the largest theoretical, but unknown, posterior probability. First, let us introduce some notation. Let $\{(1),...,(k)\}$ be the permutation of $\{1,...,k\}$ defined by

$$R_{(1)|x} < \cdots < R_{(k)|x}.$$

This means $R_{(h)|x}$, $h=1,...,k$ are the order statistics of $R_{h|x}$, $h=1,....,k$ and (h) is the statistic which gives the number of the population which appears with the h-th smallest of the estimated posterior probabilities. Further, let $\{[1],...,[k]\}$ be the permutation for which

$$\rho_{[1]|x} < \cdots < \rho_{[k]|x}.$$

However, because $\rho_{t|x}$, $t=1,...,k$ are unknown, these $[t]$, $t=1,...,k$ are unknown. Now, the above question, whether $\{(k)=[k]\}$, can be analyzed by considering the theoretical probability

$$P_{\theta,x}((k)=[k])$$

as a function of θ. Let $\sigma_{ij}^2=n^{-1}(\Sigma_{\cdot|x})_{ij}$, $i,j=1,...,k$ and let us for notational convenience assume that $[k]=k$, i.e. (θ,x) is such that the true posterior probability is largest for population k. We have that

$$\begin{pmatrix} R_{1|x}-R_{k|x} \\ \cdot \\ \cdot \\ \cdot \\ R_{k-1|x}-R_{k|x} \end{pmatrix} \text{ is } AN\left(\begin{pmatrix} \rho_{1|x}-\rho_{k|x} \\ \cdot \\ \cdot \\ \cdot \\ \rho_{k-1|x}-\rho_{k|x} \end{pmatrix}, \begin{pmatrix} \Gamma_{11} & \cdots & \Gamma_{1,k-1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \Gamma_{k-1,k-1} & \cdots & \Gamma_{k-1,k-1} \end{pmatrix} \right)$$

where Γ is the $(k-1)\times(k-1)$ matrix with

$$\Gamma_{ii} = \sigma_{ii}^2 + \sigma_{kk}^2 - 2\sigma_{ik}^2 \qquad i=1,...,k-1$$
$$\Gamma_{ij} = \sigma_{ij}^2 + \sigma_{kk}^2 - \sigma_{ik}^2 - \sigma_{kj}^2 \quad i,j=1,...,k-1;\ i\neq j.$$

The probability that $R_{k|x}$ becomes the largest estimated posterior probability, $P_{\theta,x}((k)=k)$, is

$$P\{R_{1|x}-R_{k|x}<0, \ldots, R_{k-1|x}-R_{k|x}<0\},$$

which can obviously be approximated by the value $F(0,...,0)$ of the distribution function of $N_{k-1}(\beta,\Gamma)$ where $\beta_j=\rho_{j|x}-\rho_{k|x}$ $(j=1,...,k-1)$ and where Γ has been defined above. Of course $F(0,\ldots,0)=G(-\beta)$ where G is the distribution function of $N_{k-1}(0,\Gamma)$. Note that $-\beta_j>0$ if $[k]=k$ $(j=1,...,k-1)$. The computation of $F(0_{k-1})$ or $G(-\beta)$ requires numerical integration.

An interesting inequality with respect to the probability that $R_{h|x}$ becomes the largest estimated posterior probability, is as follows. Note that we drop the assumption that $[k]=k$. We use that $P(\cap_i A_i)\geq 1-\Sigma_i P(A_i^c)$ for events $A_1,...,A_n$. Applying this we get

$$P_{\theta,x}((k)=h) = P\Big(\bigcap_{\substack{i\\ i\neq h}}^{k}\{R_{i|x}-R_{h|x}<0\}\Big)$$

$$\geq 1-\sum_{\substack{i=1\\ i\neq h}}^{k} P(R_{i|x}-R_{h|x}>0) \approx 1-\sum_{\substack{i=1\\ i\neq h}}^{k}\Phi\Big(\frac{\rho_{i|x}-\rho_{h|x}}{\{\sigma_{ii}^2+\sigma_{hh}^2-2\sigma_{ih}^2\}^{1/2}}\Big)$$

where Φ is the distribution function of the standard normal distribution. For the probability that the population with the largest estimated posterior probability coincides with the population with the largest theoretical posterior probability we get

$$P_{\theta,x}((k)=[k]) \geq 1-\sum_{\substack{i=1\\ i\neq[k]}}^{k}\Phi\Big(\frac{\rho_{i|x}-\rho_{[k]|x}}{\{\sigma_{ii}^2+\sigma_{[k][k]}^2-2\sigma_{i[k]}^2\}^{1/2}}\Big).$$

In practice this lower bound can be estimated by substituting $R_{(k)|x}$ for $\rho_{[k]|x}$, (k) for $[k]$, and by making similar replacements for the variances and covariances.

The k computed point estimates $R_{t|x}$, $t=1,...,k$ depend on the training samples. Other training samples would have given other point estimates. An interesting question is which population would have its corresponding posterior probability occurring most often as largest in a very long series of training samples when the observation x is the same each time. Therefore, let $\{i_1,...,i_k\}$ be the ordering defined by

$$P_{\theta,x}((k)=i_1)< \cdots <P_{\theta,x}((k)=i_k).$$

The population with the number i_k is the solution of this problem. Assignment to this population can thus be motivated. Interesting is whether the

ordering $\{i_1,...,i_k\}$ coincides with the ordering $\{(1),...,(k)\}$.

There are various rules in the theory of ranking and selection which can be applied to the vectors $R_{\cdot|x}$ and $L^T R_{\cdot|x}$. Most of these rules are derived under the assumptions of normality. We refer to three books about the subjects: BECHHOFER, KIEFER and SOBEL (1968), GIBBONS, OLKIN and SOBEL (1977) and GUPTA and PANCHAPAKESAN (1979). Two ranking ideas for the posterior probabilities are worth mentioning. The first idea is to take a fixed $u \in [0,1]$ and rank the populations according to the increasing order of

$$P_{\theta,x}(R_{t|x} > u) \quad t = 1,...,k.$$

This is approximately as the ranking in increasing order of

$$1 - \Phi\left(\frac{u - R_{t|x}}{\sigma_{tt}}\right) \quad t = 1,...,k$$

or as the ranking in decreasing order of

$$\frac{u - R_{t|x}}{\sigma_{tt}} \quad t = 1,...,k.$$

The second idea is as follows. Rank the populations according to the increasing order of

$$R_{t|x} - u_\alpha \sigma_{tt} \quad t = 1,...,k$$

where α, and thus u_α, is fixed. These points are the lower bounds of one-side confidence intervals for $\rho_{t|x}$, $t = 1,...,k$ with a confidence level of $1 - \alpha$.

The conditional expected losses play a role in the ranking and selection of the actions. Interesting problems are, for example, "selecting a subset containing the best", or, in medical terminology, "making a differential diagnosis consisting of all possible diseases". In these problems a central role is played by the selection of those actions which are not significantly worse than the best, i.e. the action with minimal conditional expected loss. We shall confine ourselves to a reference to the earlier mentioned books about ranking and selection.

3.4. FORCED DECISIONS

In this section we shall consider situations in which one is forced to take a decision. It is not always satisfying for the client or applied statistician to have only estimations and standard errors of posterior probabilities and conditional expected losses. Often he wants a recommendation or specific rule, especially when he is forced to take a decision. In two auxiliary models we shall make some proposals, which incorporate the standard errors of the posterior probabilities.

Let us assume that in the original model the action set is $\mathcal{A} = \{a_1,...,a_m\}$ and that loss function $L(t,a_j)$, $t = 1,...,k; j = 1,...,m$ describes the loss when action

a_j is taken if the observation comes from population t. Further we have realisations r_t of $R_{t|x}$, $t=1,...,k$ and s_{ij} of $n^{-1/2}(\hat{\Sigma}_{\cdot|x})_{ij}^{1/2}$ $i=1,...,k; j=1,...,k$ where we call to mind that the parameters have the following relation to each other

$$\mathscr{L} n^{1/2}(R_{\cdot|x}-\rho_{\cdot|x}) \to N(0,\Sigma_{\cdot|x}).$$

The problem is now that a procedure has to be constructed which prescribes which action $a \in \mathscr{A}$ has to be chosen given the r_t, $t=1,...,k$; s_{ij}, $i=1,...,k; j=1,...,k$ and $L(t,a_j)$ $j=1,...,m; t=1,...,k$. The simple plug-in procedure is a possibility. This rule chooses that action for which $\Sigma_{t=1}^{k} L(t,a) r_t$, which is an estimate for $E\{L(T,a)|X=x\}$, is minimal. When the loss function has $0-1$ structure, this rule chooses that action for which the estimated posterior probability is maximal. This is often done in practice. However, the standard errors of the r_t, $t=1,...,k$ are not taken into account. We shall describe two auxiliary models in which the standard errors s_{tt}, $t=1,...,k$ are not ignored. In the auxiliary models a fictitious decision experiment is done which has a simpler structure than that of the original model. The decision taken in the auxiliary model is used as the decision ultimately taken. The link between the original and the auxiliary model is that the outcomes of the estimators of the posterior probabilities in the original model are taken as outcomes of variables to be used in the auxiliary model.

Auxiliary model 1 (See also SCHAAFSMA (1985)). Let the unknown parameters p_t, $t=1,...,k$ of this auxiliary model correspond with the unknown posterior probabilities $\rho_{t|x}$, $t=1,...,k$ in the original model. For the point $\mathrm{p}=(p_1,...,p_k)$ in the parameter set $\Theta=\{(p_1,...,p_k); p_i \geqslant 0, \Sigma_i p_i = 1\}$ the vector $(N_1,...,N_k)$ of observable random variables follows a multinomial distribution $M(n^*;p_1,...,p_k)$ where n^* is defined later on and which has realisations $N_t=n_t=r_t n^*$, $t=1,...,k$. Thus the estimators N_t/n^* in the auxiliary model have the same realisations as the estimators $R_{t|x}$ of the posterior probabilities in the original model, namely r_t, $t=1,...,k$. In addition we consider the unobservable random variable T, independent of $(N_1,...,N_k)$, and with $P(T=t)=p_t$, $t=1,...,k$. Further, the loss is given by $L'(\mathrm{p},t,a_j)=L(t,a_j)$. For the risk of a decision rule d, which can only be a function of the observable random variables $(N_1,...,N_k)$, we get

$$\begin{aligned} R(\mathrm{p},d) &= E_{\mathrm{p}} L'(\mathrm{p},T,d(N_1,...,N_k)) = E_{\mathrm{p}} L(T,d(N_1,...,N_k)) \\ &= E\{\sum_{t=1}^{k} L(t,d(N_1,...,N_k)) p_t\} \end{aligned}$$

because of the independence between T and $(N_1,...,N_k)$. If we define

$$L^*(\mathrm{p},a_j) = E_{\mathrm{p}} L'(\mathrm{p},T,a_j) = E_{\mathrm{p}} L(T,a_j) = \sum_{t=1}^{k} L(t,a_j) p_t,$$

which corresponds with the conditional expected loss

$$E\{L(T,a_j)|X=x\} = \Sigma_{t=1}^{k} L(t,a_j)\rho_{t|x}$$

in the original model, then the risk of rule d can be written as

$$R(\mathrm{p},d) = E_{\mathrm{p}}L^{*}(\mathrm{p},d(N_1,\ldots,N_k)).$$

Further, let us suppose that on Θ the prior $P(\mathrm{p})$ is given. The problem is to choose that action $a \in \mathscr{A}$ such that the expected risk, i.e. the expected conditional expected loss in the original model, is minimal. Hence, the corresponding Bayes rule is obtained by minimizing

$$\int_{\Theta} R(\mathrm{p},d)dP(\mathrm{p}).$$

The solution is derived by taking for fixed $(n_1,\ldots,n_k)$ that action a which minimizes

$$\int_{\Theta} L^{*}(\mathrm{p},a)dP(\mathrm{p}|n_1,\ldots,n_k).$$

Let the prior $P(\mathrm{p})$ be the distribution with constant density, i.e. the Dirichlet distribution with all parameters equal to one. The density of a Dirichlet distribution is given by (see DEGROOT (1970), p. 50)

$$P(p_1,\ldots,p_k;\ \alpha_1,\ldots,\alpha_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1}$$

where $\alpha_i > 0, i = 1,\ldots,k$ and $p_i \geqslant 0, i = 1,\ldots,k$, $\Sigma_{i=1}^{k} p_i = 1$. The distribution $P(\mathrm{p}|n_1,\ldots,n_k)$ becomes the Dirichlet distribution with parameters $n_1 + 1, \ldots, n_k + 1$. Hence the conditional expected loss becomes

$$\int_{\Theta} \sum_{t=1}^{k} L(t,a)p_t \frac{\Gamma(n^{*} + k)}{\Gamma(n_1 + 1)\cdots\Gamma(n_k + 1)} p_1^{n_1} \cdots p_k^{n_k} dp_1 \ldots dp_k.$$

The action which is assigned to $(n_1,\ldots,n_k)$ is the solution of

$$\min_{a} \sum_{t=1}^{k} L(t,a) \int_{\Theta} p_1^{n_1} \cdots p_t^{n_t + 1} \cdots p_k^{n_k} dp_1 \ldots dp_k.$$

Using the Dirichlet $(n_1 + 1, \ldots, n_t + 2, \ldots, n_k + 1)$ distribution this becomes

$$\min_{a} \sum_{t=1}^{k} L(t,a) \frac{\Gamma(n_1 + 1)\cdots\Gamma(n_t + 2)\cdots\Gamma(n_k + 1)}{\Gamma(n^{*} + k + 1)}$$

or

$$\min_{a} \sum_{t=1}^{k} L(t,a) \frac{\Gamma(n_t + 2)}{\Gamma(n_t + 1)}$$

or

$$\min_{a} \{ \sum_{t=1}^{k} L(t,a) r_t + \frac{1}{n^{*}} \sum_{t=1}^{k} L(t,a) \}.$$

Note that this result differs only from that where the statistical uncertainties are ignored, by the addition of the second term. If n^{*} tends to infinity then

the difference disappears completely.

What remains to be done is to determine n^*. Note that $(N_1,...,N_k)$ has a $M(n_1^*;p_1,...,p_k)$ distribution, hence N_i has variance $n^*p_i(1-p_i)$, $i=1,...,k$ and N_i/n^*, which is an estimator for p_i, has variance $p_i(1-p_i)/n^*$, $1,...,k$. Recall that for the variances of $R_{i|x}$ we got the realisations $s_{ii}^2=n^{-1}(\Sigma_{\cdot|x})_{ii}$, $i=1,...,k$. In order to get a good link between the original and the auxiliary model we choose n^* such that either (1) or (2) holds, where (1) and (2) are defined as follows:

(1). The mean of the variances in the original model is the same as the mean of the variances in the auxiliary model:

$$\sum_{i=1}^{k} s_{ii}^2 = \sum_{i=1}^{k} \frac{p_i(1-p_i)}{n^*}.$$

With r_i as the estimate of p_i this implies that n^* has to be chosen according to

$$n^* = (\sum_{i=1}^{k} \{r_i(1-r_i)\})/\sum_{i=1}^{k} s_{ii}^2.$$

(2). The mean of the variances of the ℓ largest estimated posterior probabilities in the original model are equal to the mean of variances of the corresponding estimates in the auxiliary model. So, let $r_{[1]},...,r_{[k]}$ be the realisations $r_1,...,r_k$ in increasing order and $s_{[ii]}$ the standard derivation belonging to $r_{[i]}$ then

$$n^* = (\sum_{i=k-\ell+1}^{k} \{r_{[i]}(1-r_{[i]})\})/\sum_{i=k-\ell+1}^{k} s_{[ii]}^2.$$

Auxiliary model 2. A drawback of auxiliary model 1 is that the standard errors are modified a bit and that the correlations between estimates are ignored. An alternative model for specifying the actions to be chosen, given r_t,s_{st}, $t,s=1,...k$, is as follows. A random vector Z has outcome z, where Z has the $N_k(\mathrm{p}, n^{-1}\Sigma_{\cdot|x})$ distribution with p unknown and $\Sigma_{\cdot|x}$ known. Z must be compared with the $R_{\cdot|x}$ in the original model in which it has approximately the $N_k(\rho_{\cdot|x},n^{-1}\Sigma_{\cdot|x})$ distribution. The original $\rho_{\cdot|x}$ corresponds with $\mathrm{p}=(p_1,...,p_k)$. We assume that this unknown parameter p belongs to the parameter set $\Theta=\{(p_1,...,p_k); \Sigma_i p_i=1, p_i\geqslant 0, i=1,...,k\}$. Further, there is an unobservable random variable T with $P(T=t)=p_t$, $t=1,...,k$, $\Sigma_i p_i=1$. The loss function is defined by $L'(\mathrm{p},t,a_j)=L(t,a_j)$ in which t is the unobservable realisation of T. On the parameter set Θ the prior $p_d(\mathrm{p};I)$ is specified, where $I=(1,...,1)$. This is a Dirichlet distribution with all parameters one, thus it has the constant density $p_d(\mathrm{p};I)=\Gamma(k)=(k-1)!$. Further let $f(z;\mathrm{p},n^{-1}\Sigma_{\cdot|x})$ be the density of Z. The T and Z are independent given p. So, if we define

$$L^*(\mathrm{p},a_j) = E_{\mathrm{p}}L'(\mathrm{p},T,a_j) = \sum_{t=1}^{k} L(t,a_j)p_t\ ,$$

we find that the risk of a rule d in the parameter point p is given by

$$R(\mathrm{p},d) = E_{\mathrm{p}}L'(\mathrm{p},T,d(Z)) = E_{\mathrm{p}}L^{*}(\mathrm{p},d(Z)).$$

Now

$$P(\mathrm{p}|z) = \frac{f(z;\ \mathrm{p},\frac{1}{n}\Sigma_{\cdot|x})p_d(\mathrm{p};I)}{\int_{\Theta} f(z;p,\frac{1}{n}\Sigma_{\cdot|x})p_d(\mathrm{p};I)d\mathrm{p}}$$

and the Bayes rule is obtained by taking for fixed z that action a for which

$$\int_{\Theta}(\sum_{t=1}^{k} L(t,a)p_t)f(z;\ \mathrm{p},\frac{1}{n}\Sigma_{\cdot|x})p_d(\mathrm{p};I)d\mathrm{p}$$

or

$$\sum_{t=1}^{k} L(t,a)\int_{\Theta} f(\mathrm{p};\ z,\frac{1}{n}\Sigma_{\cdot|x})p_t d\mathrm{p}$$

is minimal. The transformation back to the original model is by replacing z by $(r_1,...,r_k)$ which are the realisations of $R_{1|x},...,R_{k|x}$. Note that for a sufficiently small dispersion matrix $n^{-1}\Sigma_{\cdot|x}$ the rule approaches the rule

$$\min_{a} \sum_{t=1}^{k} L(t,a)r_t.$$

3.5. FULLY BAYESIAN APPROACH

In this section we shall describe an approach in which the k densities $f_{h,\theta}$, $\ell=1,...,k$ are unknown, a prior T has been put on the numbers $1,...,k$ and a prior P on the parameter set Θ. More precisely, prior T is given by $P(T=t)=\rho_t$, $t=1,...,k$, prior P has density $p(\theta)$ in $\theta\in\Theta$, and T and P are independent. In other words, a prior has been put on the parameter set $\{1,...,k\}\times\Theta$ with value $\rho_t p(\theta)$ in the parameter point (t,θ) and with above-mentioned marginals and such that independence holds. Let the observations $x,x_{11},...,x_{kn_k}$ be an element of the sample space $\mathscr{X}$, which consists of all outcomes which are generated by the independent random variables $X,X_{11},...,X_{kn_k}$. The probability density on $\mathscr{X}$ is given by

$$f_{t,\theta}(x)\prod_{h=1}^{k}\prod_{i=1}^{n_h} f_{h,\theta}(x_{hi}).$$

Further, there is the action set $\mathscr{A}=\{a_1,...,a_m\}$ and the loss function $L(t,a)$ where $a\in\mathscr{A}$ and $t\in\{1,...,k\}$. The risk of decision rule d in the parameter point (t,θ) is given by

$$R(t,\theta,d) = E_{t,\theta}L(t,d(X,X_{11},...,X_{kn_k}))$$

$$= \int_{\mathcal{X}} L(t,d(x,x_{11},\ldots,x_{kn_k}))f_{t,\theta}(x)\prod_{h=1}^{k}\prod_{i=1}^{n_k} f_{h,\theta}(x_{hi})dx dx_{11}\cdots dx_{kn_k}.$$

The Bayes risk of decision rule d with respect to prior (T,P) becomes

$$R(d) = EL(T,d(X,X_{11},\ldots,X_{kn_k}))$$
$$= \sum_{t=1}^{k}\rho_t \int_{\Theta} E_{t,\theta}L(t,d(X,X_{11},\ldots,X_{kn_k}))p(\theta)d\theta.$$

The Bayes rule is obtained by minimizing the conditional expected loss:

$$\min_a \sum_{t=1}^{k}\int_{\Theta} L(t,a)p(t,\theta|x,x_{11},\ldots,x_{kn_k})d\theta$$

which can be written as

$$\min_a \sum_{t=1}^{k} L(t,a)P_{pred}(t)$$

where $P_{pred}(t)$ is the predictive posterior probability

$$P_{pred}(t) = \frac{\rho_t f_{t,pred}(x)}{\sum_{u=1}^{k}\rho_u f_{u,pred}(x)}, \quad t=1,\ldots,k$$

with $f_{t,pred}(x)$ the predictive density of the distribution of the vector of scores in population t, evaluated at the observation x:

$$f_{t,pred}(x) = \int_{\Theta} f_{t,\theta}(x)p(\theta|x_{11},\ldots,x_{kn_k})d\theta$$

with

$$p(\theta|x_{11},\ldots,x_{kn_k}) = \frac{p(\theta)\prod_{s=1}^{k}\prod_{i=1}^{n_s} f_{s,\theta}(x_{si})}{\int_{\Theta} p(\theta)\prod_{s=1}^{k}\prod_{i=1}^{n_s} f_{s,\theta}(x_{si})d\theta}.$$

A special case is that in which the assumptions are made that $\theta=(\theta_1,\ldots,\theta_k)$, that the prior probability $p(\theta)=p_1(\theta_1)\cdots p_k(\theta_k)$, and that $f_{t,\theta}=f_{t,\theta_t}$, $t=1,\ldots,k$. Writing $\Theta=\Theta_1\times\cdots\times\Theta_k$ we obtain

$$p(\theta|x_{11},\ldots,x_{kn_k}) = p_1(\theta_1|x_{11},\ldots,x_{1n_1})\cdots p_k(\theta_k|x_{k1},\ldots,x_{kn_k})$$

and the predictive density is given by

$$f_{t,pred}(x) = \int_{\Theta_t} f_{\theta_t}(x)p_t(\theta_t|x_{t1},\ldots,x_{tn_t})d\theta_t.$$

The approach of this section is called the predictive or fully Bayesian approach. A series of articles has been published on the subject, especially by Geisser (1964, 1965, 1966, 1967, 1970, 1977, 1980, 1982a,b). The series was preceded by the articles of Geisser and Cornfield (1963) and Cornish

(1961), in which the fiducial argument of FISCHER (1935, 1954) was elaborated upon. Studies about comparisons between predictive and estimative approaches can be found in, for example, GEISSER (1982a), AITCHISON and KAY (1975), MCLACHLAN (1979), HERMANS and HABBEMA (1975), AITCHISON, HABBEMA and KAY (1977), HABBEMA and HERMANS (1978). One of their results is that the predictive approach gives less extreme estimates for the posterior probabilities than the estimative approach. However, MORAN and MURPHY (1979) showed that if adjustments for the bias of the odds in the estimative method are made, both methods are comparable.

Chapter 4

Miscellaneous results, normal densities

4.1. INTRODUCTION

Throughout this chapter we shall assume that there are $k(\geqslant 2)$ populations and p continuous variables. The variables follow a p-variate normal distribution. In some cases we shall assume that the covariance matrices are equal and in other cases that they are not, but it will always be clear which case we are dealing with. We shall present unbiased estimators and their variances for various statistics. These statistics appear in a natural way if one tries to estimate the posterior probabilities. In section 4.2 a proof is given that unbiased estimators for posterior probabilistics themselves do not exist if normality is assumed. Now, recall that the posterior probabilities can be written as

$$\rho_{t|x} = \rho_t f_t(x) / \sum_{h=1}^{k} \rho_h f_h(x), \quad t = 1,...,k \tag{4.1.1}$$

with $f_t(x)$ the density of the t-th population at vector x. Unbiased estimators with minimum variance of the k densities are given in section 4.3. If we define the log-odds

$$\zeta_{x;ht} = \log(f_h(x)/f_t(x)) = \tfrac{1}{2}(\Delta^2_{x;t} - \Delta^2_{x;h})$$

$$= (\mu_h - \mu_t)^T \Sigma^{-1} \{x - \tfrac{1}{2}(\mu_h + \mu_t)\}$$

in case of equal covariance matrices, then

$$\rho_{t|x} = \{\sum_{h=1}^{k} \rho_h \rho_t^{-1} \exp(\zeta_{x;ht})\}^{-1}, \quad t = 1,...,k.$$

In section 4.4 unbiased estimators for these $\zeta_{x;ht}$ are given together with their variances and covariances. The posterior probabilities can also be written as functions of logarithms of the population densities. The unbiased estimators of the log-densities and their variances are derived in section 4.5. Section 4.6 presents the results of a simulation study of the quality of confidence intervals for posterior probabilities constructed on the basis of asymptotic distributions.

4.2. THE NON-EXISTENCE OF UNBIASED ESTIMATORS FOR POSTERIOR PROBABILITIES

In this section it is proved that unbiased estimators for the posterior probabilities do not exist under assumptions of normality. We shall give a proof for the case that two populations are involved. The more general case with more than two populations can simply be given by extending the proof of this section.

Let the two populations correspond with the p-dimensional multivariate normal densities

$$f_t(x) = |2\pi\Sigma_t|^{-1/2} \exp\{-\tfrac{1}{2}(x-\mu_t)^T\Sigma_t^{-1}(x-\mu_t)\} \quad t=1,2.$$

The training samples are generated by the independent random variables

$$X_{j1},\ldots,X_{jn_j} \text{ with } X_{ji}\sim N_p(\mu_{ji}, \Sigma_j), \quad i=1,\ldots,n_j;\ j=1,2$$

Define

$$Y_j = n_j^{-1}\sum_{i=1}^{n_j} X_{ji} \quad j=1,2$$

and

$$S_j = \sum_{i=1}^{n_j}(X_{ji}-Y_j)(X_{ji}-Y_j)^T \quad j=1,2.$$

Then

$$Y_j\sim N_p(\mu_j, n_j^{-1}\Sigma_j) \text{ and } S_j\sim W_p(n_j-1, \Sigma_j), \quad j=1,2.$$

If we want to obtain an unbiased estimator for the posterior probability of population one, then we are searching for a function $h:\mathbb{R}^{p+p+p\times p+p\times p}\to\mathbb{R}$ which satisfies

$$E_{\mu_1,\mu_2,\Sigma_1,\Sigma_2} h(Y_1, Y_2, S_1, S_2) = \rho_{1|x} \tag{4.2.1}$$

for every point $(\mu_1,\mu_2,\Sigma_1,\Sigma_2)$ of the parameter set. Note that h depends on the vector of scores x, the prior probability ρ_1, the sizes n_1 and n_2 of the samples, and the dimension p.

However, we shall show that such a function h does not exist. Unbiasedness requires that (4.2.1) is true for every point in the parameter set. For the proof of the non-existence of an unbiased estimator it is sufficient that a subset of

the parameter set can be found on which (4.2.1) is not fulfilled.

So, let us suppose that a function h exists which satisfies (4.2.1). After taking the expectation with respect to Y_2, S_1 and S_2 a function $g: \mathbb{R}^p \to \mathbb{R}$ is obtained which satisfies

$$E_{\mu_1,\Sigma_1}\, g(Y_1) = \rho_{1|x}. \tag{4.2.2}$$

Let $x=(x_1,...,x_p)^T$ and $\mu_j=(\mu_{j1},...,\mu_{jp})^T, j=1,2$, and let us take $\mu_{1i}=x_i$, $i=2,...,p$; $\mu_2=x$ and $\Sigma_1=\Sigma_2=I$. Then the function g satisfies

$$\int \cdots \int g(y_1,...,y_p)\left(\frac{2\pi}{n_1}\right)^{-p/2} \exp\{-\tfrac{1}{2}n_1 \sum_{i=1}^{p}(y_i-\mu_{1i})^2\}dy_1 \cdots dy_p$$

$$= [1+\rho_2\rho_1^{-1} \exp\{\tfrac{1}{2}(x_1-\mu_{11})^2\}]^{-1}.$$

After integrating out $y_2,...,y_p$ and replacing y_1 by y, μ_{11} by μ, $\rho_2\rho_1^{-1}$ by a, x_1 by b, n_1 by n, $\sqrt{n}\mu$ by ν, and the transformation $z=\sqrt{n}y$, we get the following result. A function $f: \mathbb{R} \to \mathbb{R}$ for which

$$\int_{-\infty}^{+\infty} f(z)\frac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}(z-\nu)^2\}dz = [1+a \exp\{\tfrac{1}{2}(b-n^{-1/2}\nu)^2\}]^{-1} \tag{4.2.3}$$

for all $\nu \in \mathbb{R}$, where $a>0$ and $b \in \mathbb{R}$. Another formulation for the left-hand side of (4.2.3) is

$$E_\nu f(Z) = \int_{-\infty}^{+\infty} f(z)dF_\nu(z) \tag{4.2.4}$$

where $Z \sim N(\nu,1)$ and F_ν is the distribution function of Z. The left-hand side of (4.2.3) is transformed by the Gauss- or Weierstrass transformation

$$t=-z, \quad G(t) = f(-t)\exp(-t^2/2)$$

into a two-side Laplace transform (see MAGNUS et al. (1966), p. 398). The consequence is that (4.2.3.) becomes

$$\int_{-\infty}^{+\infty} G(t) \exp(-t\nu)dt = (2\pi)^{1/2} \exp(\tfrac{1}{2}\nu^2)[1+a \exp\{\tfrac{1}{2}(b-n^{-1/2}\nu)^2\}]^{-1}. \tag{4.2.5}$$

Now, by definition of Lebesgue integrability and the fact that the right-hand side of (4.2.5) is finite for every $\nu \in \mathbb{R}$, we have

$$\int_{-\infty}^{+\infty} |G(t) \exp(-t\nu)|dt < +\infty \quad \forall \nu \in \mathbb{R}. \tag{4.2.6}$$

Let us define for the complex number $z \in \mathbb{C}$ the following two functions

$$g_\ell(z) = \int_{-\infty}^{+\infty} G(t) \exp(-tz)dt$$

and

$$g_r(z) = (2\pi)^{1/2} \exp(\tfrac{1}{2}z^2)[1+a \exp\{\tfrac{1}{2}(b-n^{-1/2}z)^2\}]^{-1}.$$

Note that the function $g_\ell(z)$ can be defined for every $z \in \mathbb{C}$, because $|G(t)\exp(-tz)|$ is integrable by (4.2.6). But this means that also $G(t)\exp(-tz)$ is integrable (see RUDIN (1964), p. 250), and

$$|g_\ell(z)| \leqslant \int |G(t) \exp(-tz)| dt < +\infty, \quad \forall z \in \mathbb{C}.$$

It can be proved that $g_\ell(z)$ is analytic on $\mathbb{C}$ (see PAPOULIS (1962), p. 170). Further, the function $g_r(z)$ is analytic, with poles at the points $z = u + iv$ where

$$u = n^{-1/2}b - n^{-1/2}w, \quad v = -n^{1/2}w^{-1}(\pi + 2k\pi)$$

and

$$w = \pm\{\ln(\frac{1}{a}) \pm \{\ln^2(\frac{1}{a}) + (\pi + 2k\pi)^2\}^{1/2}\}^{1/2}$$

and $k \in \mathbb{Z}$. So, we have on $\mathbb{C}$ two analytic functions of which one has poles and the other none. This means that they are different on $\mathbb{C}$ and hence, by a result from analytic function theory, they cannot be equal on $\mathbb{R}$ (see CONWAY (1973), theorem 3.7). But this is a contradiction with (4.2.5). The conclusion is that a function G satisfying (4.2.5) does not exist. So, it has been proved that unbiased estimators for posterior probabilities do not exist under assumptions of normality.

The non-existence of unbiased estimators of posterior probabilities has also been mentioned in SCHAAFSMA (1985b).

4.3. UNBIASED ESTIMATORS FOR NORMAL DENSITIES

In section 4.2 we proved that unbiased estimators for the posterior probabilities themselves do not exist. However, the posterior probabilities are functions of other parameters, see for instance formula (4.1.1). An approach is to use the unbiased estimators of the basic parameters. This was done in the preceding chapters. However, other parameters like $f_h(x)$, $\log f_h(x)$, etc. can be used as well. Plugging unbiased estimators for these parameters into the expression of the posterior probabilities provides a modification of the estimators for the posterior probabilities studied earlier. From formula (4.1.1) we see that $f_h(x)$, $\log f_h(x)$, and $\zeta_{x;ht} = \log(f_h(t)/f_t(x))$ can be used as parameters. In this and the following sections we shall therefore concentrate upon the unbiased estimation of these parameters.

In this section we shall present the unbiased estimator with minimum variance, MVUE, for the density $f_h(x)$ of the multivariate normal distribution. Section 4.4 gives the unbiased estimator of $\zeta_{x;ht}$, and section 4.5 that of $\log f_h(x)$.

Let $X_{t1}, \ldots, X_{tn_t}$ be independently and identically distributed, $X_{ti} \sim N_p(\mu_t, \Sigma_t)$. Try to find the minimum variance unbiased estimator, based on these X_{ti}'s, for

$$f_t(x) = |2\pi\Sigma_t|^{-1/2} \exp\{-\tfrac{1}{2}(x-\mu_t)^T\Sigma_t^{-1}(x-\mu_t)\}.$$

For $p=1$, KOLMOGOROV (1950) solved this problem by deriving the MVUE for the probability $P(X_{t1}>u)$, using the Rao-Blackwell theorem. For that purpose he evaluated $P(X_{t1}\geqslant u|\bar{X}_t, S_t)$. Note that taking derivatives is a linear operation which preserves unbiasedness. So, the conditional density of X_{t1} given $\bar{X}_t$ and S_t is the best unbiased estimator for the density of the normal variable X_{t1}. Many authors have presented solutions to the problem for the case $p\geqslant 2$. For a complete proof see for instance EATON and MORRIS (1970) or GHURYE and OLKIN (1969). Recently SCHAAFSMA (1985b) also obtained the estimator. The MVUE estimator for $f_t(x)$ is equal to the conditional density of X_{t1} given $\bar{X}_{t1}$ and S_t, this is

$$\frac{\Gamma(\frac{1}{2}f_t)n_t^{p/2}}{\Gamma(\frac{1}{2}(f_t-p))(n_t-1)^{p/2}\pi^{p/2}|S_t|^{1/2}}\{1-\frac{n_t}{n_t-1}(x-\bar{X}_t)^TS_t^{-1}(x-\bar{X}_t)\}^{\frac{1}{2}(f_t-p-2)}$$

$$\cdot I_{(0,1)}(\frac{n_t}{n_t-1}(x-\bar{X}_t)^TS_t^{-1}(x-X_t)),$$

where $I_{(0,1)}(a)=1$ if $a\in(0,1)$ and 0 otherwise. Further recall to mind that $\bar{X}_t\sim N_p(\mu_t, n_t^{-1}\Sigma_t)$ and $S_t\sim W_p(f_t,\Sigma_t)$ where $f_t=n_t-1$. This f_t should not be confused with $f_t(x)$, which is the density at point x.

4.4. UNBIASED ESTIMATORS FOR LOG-ODDS AND THEIR VARIANCES

In this section it is assumed that the densities of the $k(\geqslant 2)$ populations are multivariate normal with equal covariance matrices. In section 4.1 we saw that in that case the k posterior probabilities can be written as

$$\rho_{t|x} = \{\sum_{h=1}^{k}\rho_h\rho_t^{-1}\exp(\zeta_{x;ht})\}^{-1}, \quad t=1,...,k \tag{4.4.1}$$

where the log-odds $\zeta_{x;ht}=\log(f_h(x)/f_t(x))$ is specified by

$$\zeta_{x;ht} = (\mu_h-\mu_t)^T\Sigma^{-1}\{x-\tfrac{1}{2}(\mu_h+\mu_t)\}, \quad h,\ t=1,...,k. \tag{4.4.2}$$

Unbiased estimators of the parameters $\zeta_{x;ht}$ with their variances and covariances will be derived. Subsequently, they will be used in a theorem about the asymptotic distribution of the corresponding estimator of the vector of posterior probabilities. From the independent random variables $X_{hi}, i=1,...,n_h$; $h=1,...,k$, in which X_{hi} is distributed as $N_p(\mu_h,\Sigma)$ we form the independent statistics $X_{1\cdot},...,X_{k\cdot}, S$ where

$$X_{h\cdot} = n_h^{-1}\sum_{i=1}^{n_h}X_{hi} \quad , \quad h=1,...,k$$

and

$$S = \sum_{i=1}^{k} \sum_{i=1}^{n_h} (X_{hi} - X_{h\cdot})(X_{hi} - X_{h\cdot})^T.$$

Then

$$X_{h\cdot} \sim N_p(\mu_h, n_h^{-1}\Sigma)$$

and

$$S \sim W_p(\sum_{h=1}^{k} (n_h - 1), \Sigma).$$

To deal with situations where additional information is available for estimating Σ, we shall present results for the more general case where S has the $W_p(f, \Sigma)$ distribution. As estimator for $\zeta_{x;ht}$ we use

$$U_{x;ht} = (f - p - 1)(X_{h\cdot} - X_{t\cdot})^T S^{-1}\{x - \tfrac{1}{2}(X_{h\cdot} + X_{t\cdot})\}$$

$$+ \tfrac{1}{2}p(n_h^{-1} - n_t^{-1}) \tag{4.4.3}$$

which has the property that

$$EU_{x;ht} = \zeta_{x;ht}. \tag{4.4.4}$$

This result follows immediately from the independence of $X_{h\cdot}$, $X_{t\cdot}$ and S, from $EX_{h\cdot} = \mu_h$, $EX_{t\cdot} = \mu_t$, and

$$ES^{-1} = (f - p - 1)^{-1}\Sigma^{-1}.$$

The statistic S^{-1} is said to have an Inverse-Wishart distribution if S has a Wishart distribution. For more about Inverse-Wishart distributions see for example MUIRHEAD (1982), p. 97 and EATON (1983), p. 330.

The next theorem lies at the basis of many of the computations which will follow.

THEOREM 4.4.1.
Suppose $S \sim W_p(f, \Sigma)$. *Let* S *and* Σ *be partitioned as*

$$S = \begin{bmatrix} S_{(1,1)} & S_{(1,2)} \\ S_{(2,1)} & S_{(2,2)} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{(1,1)} & \Sigma_{(1,2)} \\ \Sigma_{(2,1)} & \Sigma_{(2,2)} \end{bmatrix}$$

where $S_{(1,1)}$ *and* $\Sigma_{(1,1)}$ *are* $q \times q$ *submatrices. Define*

$$S_{11.2} = S_{(1,1)} - S_{(1,2)}S_{(2,2)}^{-1}S_{(2,1)}$$

and

$$\Sigma_{11.2} = \Sigma_{(1,1)} - \Sigma_{(1,2)}\Sigma_{(2,2)}^{-1}\Sigma_{(2,1)},$$

then

(1) $S_{11.2} \sim W_p(f-p+q,\ \Sigma_{11.2})$

(2) $S_{11.2}$ *is independent of* $(S_{(1,2)}, S_{(2,2)})$

(3) $S_{(1,2)}|S_{(2,2)} \sim N(\Sigma_{(1,2)}\Sigma_{(2,2)}^{-1}S_{(2,2)},\ \Sigma_{11.2} \otimes S_{(2,2)})$

(4) $S_{(2,2)} \sim W_{p-q}(f,\ \Sigma_{(2,2)})$.

References for this theorem are MUIRHEAD (1982) p. 93, SRIVASTAVA and KHATRI (1979) p. 79, and for the first two statements RAO (1965) p. 539.

The special case of theorem 4.4.1 when $\Sigma = I$ and $q = 1$, is given in the next corollary (S is replaced by V and $S_{(1,1)}$ by V_{11}).

COROLLARY 4.4.2.
If $V \sim W_p(f, I)$ *and* V_{11} *is the (1,1) element of V, then*

(1) $V_{11} - V_{(1,2)}V_{(2,2)}^{-1}V_{(2,1)} \sim W_1(f-p+1, 1) = \chi^2_{f-p+1}$

(2) $V_{11} - V_{(1,2)}V_{(2,2)}^{-1}V_{(2,1)}$ *is independent of* $(V_{(1,2)},\ V_{(2,2)})$

(3) $(V_{(1,2)}V_{(2,2)}^{-1/2})^T | V_{(2,2)} \sim N_{p-1}(0,\ I_{p-1})$

(4) $V_{(2,2)} \sim W_{p-1}(f,\ I_{p-1})$.

The following lemma about the inverse of a partitioned matrix is very useful (see RAO (1965), p. 33).

LEMMA 4.4.3.
Let A *be a* $p \times p$ *square matrix with*

$$A = \begin{bmatrix} A_{(1,1)} & A_{(1,2)} \\ A_{(2,1)} & A_{(2,2)} \end{bmatrix} \quad \text{and} \quad A^{-1} = \begin{bmatrix} A^{(1,1)} & A^{(1,2)} \\ A^{(2,1)} & A^{(2,2)} \end{bmatrix}$$

where $A_{(1,1)}$ *and* $A^{(1,1)}$ *are* $q \times q$ *submatrices, then*

(1) $A^{(1,1)} = (A_{(1,1)} - A_{(1,2)}A_{(2,2)}^{-1}A_{(2,1)})^{-1}$

(2) $A^{(1,2)} = -(A_{(1,1)} - A_{(1,2)}A_{(2,2)}^{-1}A_{(2,1)})^{-1}A_{(1,2)}A_{(2,2)}^{-1}$

(3) $A^{(2,1)} = -A_{(2,2)}^{-1}A_{(2,1)}(A_{(1,1)} - A_{(1,2)}A_{(2,2)}^{-1}A_{(2,1)})^{-1}$

(4) $A^{(2,2)} = A_{(2,2)}^{-1} + A_{(2,2)}^{-1}A_{(2,1)}(A_{(1,1)} - A_{(1,2)}A_{(2,2)}^{-1}A_{(2,1)})^{-1}A_{(1,2)}A_{(2,2)}^{-1}$.

The next lemma can be found in RAO (1965), p. 538.

LEMMA 4.4.4.
If $S \sim W_p(f, \Sigma)$ *and* B *a* $q \times p$ *matrix, then*

$$BSB^T \sim W_q(f,\ B\Sigma B^T).$$

A direct application of this lemma is the following corollary.

COROLLARY 4.4.5.
If $V \sim W_p(f, I)$ *and* Ψ *an orthogonal matrix then*

$$\mathscr{L}V = \mathscr{L}\Psi V \Psi^T$$

and

$$\mathcal{L}V^{-1} = \mathcal{L}\Psi V^{-1}\Psi^T.$$

The following lemma gives results about first and second moments of the Inverse-Wishart distribution.

LEMMA 4.4.6.
Let $V \sim W_p(f, I)$ and let V and V^{-1} be partitioned as the A and A^{-1} of lemma 4.4.3 with $q=1$, $V_{(1,1)}$ and $V^{(1,1)}$ replaced by V_{11} and V^{11}, respectively, then

(1) $E\ V^{11} = (f-p-1)^{-1}$
(2) $E\ V^{-1} = (f-p-1)^{-1} I_p$
(3) $E(V^{11})^2 = (f-p-1)^{-1}(f-p-3)^{-1}$
(4) $E\ V^{(1,2)}V^{(2,1)} = (f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}(p-1)$
(5) $E\ V^{-2} = (f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}(f-1)\ I_p$
(6) $EV^{(2,1)}V^{(1,2)} = (f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1} I_{p-1}$
(7) $E\ V^{11}V^{(1,2)} = 0$
(8) $E\ V^{11}V^{(2,1)} = 0$
(9) $E\ V^{(1,2)}V^{(2,2)} = 0$
(10) $E\ V^{11}V^{(2,2)} = (f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}(f-p-2)\ I_{p-1}$.

PROOF of (1): $E\ V^{11} = E(V_{11} - V_{(1,2)}V_{(2,2)}^{-1}V_{(2,1)})^{-1}$ which is the expectation of a $1/\chi^2_{f-p+1}$ distribution.

PROOF of (2): A proof of this can be found in many places, for example in, DAS GUPTA (1968), LACHENBRUCH (1975), MUIRHEAD (1982), EATON (1983), etc. We shall give the proof of Eaton. Corollary 4.4.5 gives that $\mathcal{L}V^{-1}$ is invariant under orthogonal transformations: if $\Psi \in O(p)$, the set of orthogonal $p \times p$ matrices, then $\mathcal{L}V^{-1} = \mathcal{L}\Psi V^{-1}\Psi^T$. Hence $EV^{-1} = E\Psi V^{-1}\Psi^T = \Psi E V^{-1}\Psi^T$ for all $\Psi \in O(p)$, which implies that $E\ V^{-1} = cI$, where c is a constant. According to (1) we get $c = E(V^{-1})_{11} = EV^{11} = (f-p-1)^{-1}$.

PROOF of (3): Lemma 4.4.3 (1) gives $E(V^{11})^2 = E((V_{11} - V_{(1,2)}V_{(2,2)}^{-1}V_{(2,1)})^{-1})^2$. From corollary 4.4.2 it follows that this is the second moment of a $1/\chi^2_{f-p+1}$ distribution.

PROOF of (4): With corollary 4.4.2 and lemma 4.4.3 we get

$$\begin{aligned} E\ V^{(1,2)}V^{(2,1)} &= E\ V^{11}V_{(1,2)}V_{(2,2)}^{-1}V_{(2,2)}^{-1}V_{(2,1)}V^{11} \\ &= E(V^{11})^2 E \text{ trace } \{V_{(2,2)}^{-1}(V_{(1,2)}V_{(2,2)}^{-1/2})^T(V_{(1,2)}V_{(2,2)}^{-1/2})\} \\ &= E(V^{11})^2 \text{ trace } EV_{(2,2)}^{-1} \end{aligned}$$

Now apply (3) and note that trace $EV_{(2,2)}^{-1} =$ trace $\{(f-(p-1)-1)^{-1} I_{p-1}\} = (f-p)^{-1}(p-1)$, where (2) is used.

PROOF of (5): The same invariance considerations as in (2) give $E\ V^{-2} = dI$,

where d is a constant. Now, $d=(EV^{-2})_{11}= E(V^{11})^2+EV^{(1,2)}V^{(2,1)}$ and apply (3) and (4) of this lemma. See DAS GUPTA (1968) and EATON (1983).

PROOF of (6): With corollary 4.4.2 and lemma 4.4.3 we get

$$\begin{aligned} E\ V^{(2,1)}V^{(1,2)} &= E\ V_{(2,2)}^{-1}V_{(2,1)}V^{11}V^{11}V_{(1,2)}V_{(2,2)}^{-1} \\ &= E(V^{11})^2E\ V_{(2,2)}^{-1/2}(V_{(1,2)}V_{(2,2)}^{-1/2})^T(V_{(1,2)}V_{(2,2)}^{-1/2})V_{(2,2)}^{-1/2} \\ &= E(V^{11})^2EV_{(2,2)}^{-1} \end{aligned}$$

Now, apply (3), lemma 4.4.4 and (2).

PROOF of (7):

$$\begin{aligned} E\ V^{11}V^{(1,2)} &= -E\ V^{11}V^{11}V_{(1,2)}V_{(2,2)}^{-1} \\ &= -E(V^{11})^2E\ V_{(1,2)}V_{(2,2)}^{-1/2}(V_{(2,2)}^{-1/2}=0 \end{aligned}$$

where lemma 4.4.3 and corollary 4.4.2 has been used.

PROOF of (8):

$$\begin{aligned} E\ V^{11}V^{(1,2)} &= -EV^{11}V_{(2,2)}^{-1}V_{(2,1)}V^{11} \\ &= -E(V^{11})^2E\ V_{(2,2)}^{-1/2}V_{(2,2)}^{-1/2}V_{(2,1)} \\ &= E(V^{11})^2E\ V_{(2,2)}^{-1/2}(V_{(1,2)}V_{(2,2)}^{-1/2})^T=0 \end{aligned}$$

PROOF of (9):

$$\begin{aligned} E\ V^{(1,2)}V^{(2,2)} &= -E\ V^{11}V_{(1,2)}V_{(2,2)}^{-1}(V_{(2,2)}^{-1}+V_{(2,2)}^{-1}V_{(2,1)}V_{11.2}^{-1}V_{(1,2)}V_{(2,2)}^{-1}) \\ &= -E\ V^{11}V_{(1,2)}V_{(2,2)}^{-1/2}V_{(2,2)}^{-3/2} \\ &\quad +E(V^{11})^2E\ V_{(1,2)}V_{(2,2)}^{-1/2}V_{(2,2)}^{-1}V_{(2,2)}^{-1/2}V_{(2,1)}V_{(1,2)}V_{(2,2)}^{-1/2}V_{(2,2)}^{-1/2} \\ &= 0 \end{aligned}$$

where $V_{11.2}=V_{11}-V_{(1,2)}V_{(2,2)}^{-1}V_{(2,1)}$ and corollary 4.4.2 (3) has been used.

PROOF of (10):

$$\begin{aligned} E\ V^{11}V^{(2,2)} &= EV^{11}V_{(2,2)}^{-1}+E\ V^{11}V_{(2,2)}^{-1}V_{(2,1)}V^{11}V_{(1,2)}V_{(2,2)}^{-1} \\ &= E\ V^{11}V_{(2,2)}^{-1}+E(V^{11})^2E\ V_{(2,2)}^{-1/2}(V_{(1,2)}V_{(2,2)}^{-1/2})^T(V_{(1,2)}V_{(2,2)}^{-1/2})V_{(2,2)}^{-1/2} \\ &= E\ V^{11}E\ V_{(2,2)}^{-1}+E(V^{11})^2E\ V_{(2,2)}^{-1} \\ &= \{(f-p-1)^{-1}+(f-p-1)^{-1}(f-p-3)^{-1}\}(f-p)^{-1}I_{p-1} \\ &= (f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}(f-p-2)I_{p-1}. \end{aligned}$$

For the derivations which follow it is good to have a complete survey of all possible second moments of an Inverse-Wishart distribution.

LEMMA 4.4.7.
If $V \sim W_p(f, I)$ and i,j,k,ℓ represent four positive integers, all different from each other, then

(1) $E\ V^{ii}V^{ii} = (f-p-1)^{-1}(f-p-3)^{-1}$
(2) $E\ V^{ii}V^{jj} = (f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}(f-p-2)$
(3) $E\ V^{ij}V^{ij} = (f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}$
(4) $E\ V^{ii}V^{ij} = 0$
(5) $E\ V^{ii}V^{jk} = 0$
(6) $E\ V^{ij}V^{ik} = 0$
(7) $E\ V^{ij}V^{kl} = 0.$

PROOF. The first six moments can be obtained directly from lemma 4.4.6 using (3), (10), (6), (8), (10) and (6) respectively. However, the moments of (4), (5) and (6) can also be obtained in the following way. Take in corollary 4.4.5 for Ψ a diagonal matrix with $\Psi_{11}=1$, $\Psi_{22}=-1$, $\Psi_{33}=1$, $\Psi_{44}=-1$ and $\Psi_{ii}=1$ for $i=5,\ldots,p$. Let B be a $4\times p$ matrix specified by $B=(I_4:O)$. Hence $\mathscr{L}BV^{-1}B^T = \mathscr{L}B\Psi V^{-1}\Psi^T B^T$ which is

$$\mathscr{L}\begin{bmatrix} V^{11} & V^{12} & V^{13} & V^{14} \\ V^{21} & V^{22} & V^{23} & V^{24} \\ V^{31} & V^{32} & V^{33} & V^{34} \\ V^{41} & V^{42} & V^{43} & V^{44} \end{bmatrix} = \mathscr{L}\begin{bmatrix} +V^{11} & -V^{12} & +V^{13} & -V^{14} \\ -V^{21} & +V^{22} & -V^{23} & +V^{24} \\ +V^{31} & -V^{32} & +V^{33} & -V^{34} \\ -V^{41} & +V^{42} & -V^{43} & +V^{44} \end{bmatrix}.$$

After integrating out variables we get: $\mathscr{L}V^{11}V^{12} = -\mathscr{L}V^{11}V^{12}$ hence $E\ V^{11}V^{12} = -E\ V^{11}V^{12} = 0$, further $\mathscr{L}V^{11}V^{23} = -\mathscr{L}V^{11}V^{23}$ which implies $E\ V^{11}V^{23} = -E\ V^{11}V^{23} = 0$ and $\mathscr{L}V^{12}V^{13} = -\mathscr{L}V^{12}V^{13}$, hence $E\ V^{12}V^{13} = -E\ V^{12}V^{13} = 0$. In order to prove (7) we apply the same technics but now with $\Psi_{11}=-1$, $\Psi_{ii}=1$, $i=2,\ldots,p$ hence

$$\mathscr{L}\begin{bmatrix} V^{11} & V^{12} & V^{13} & V^{14} \\ V^{21} & V^{22} & V^{23} & V^{24} \\ V^{31} & V^{32} & V^{33} & V^{34} \\ V^{41} & V^{42} & V^{43} & V^{44} \end{bmatrix} = \mathscr{L}\begin{bmatrix} +V^{11} & -V^{12} & -V^{13} & -V^{14} \\ -V^{21} & +V^{22} & +V^{23} & +V^{24} \\ -V^{31} & +V^{32} & +V^{33} & +V^{34} \\ -V^{41} & +V^{42} & +V^{43} & +V^{44} \end{bmatrix}.$$

Hence $\mathscr{L}V^{12}V^{34} = -\mathscr{L}V^{12}V^{34}$ from which $E\ V^{12}V^{34}=0$ follows.

LEMMA 4.4.8
If $S \sim W_p(f, \Sigma)$, $X_i \sim N_p(\mu_i, n_i^{-1}\Sigma)$, $X_j \sim N_p(\mu_j, n_j^{-1}\Sigma)$, all independent of each other, and $c,d \in \mathbb{R}^p$, then
(1) $ES^{-1}cd^TS^{-1} = (f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}$

$$\cdot \{(f-p-2)\Sigma^{-1}cd^T\Sigma^{-1}+\Sigma^{-1}dc^T\Sigma^{-1} + c^T\Sigma^{-1}d\Sigma^{-1}\},$$

(2) $$EX_i^TS^{-1}X_iX_j^TS^{-1}X_j = (f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}$$
$$\cdot \{(f-p-2)\mu_i^T\Sigma^{-1}\mu_i\mu_j^T\Sigma^{-1}\mu_j + 2\mu_i^T\Sigma^{-1}\mu_j\mu_i^T\Sigma^{-1}\mu_j$$
$$+ (2+p(f-p-2))(pn_i^{-1}n_j^{-1} + n_j^{-1}\mu_i^T\Sigma^{-1}\mu_i+n_i^{-1}\mu_j^T\Sigma^{-1}\mu_j)\},$$

(3) $$EX_i^TS^{-1}X_iX_i^TS^{-1}X_i = (f-p-1)^{-1}(f-p-3)^{-1}$$
$$\cdot \{\mu_i^T\Sigma^{-1}\mu_i\mu_i^T\Sigma^{-1}\mu_i +2(p+2)n_i^{-1}\mu_i^T\Sigma^{-1}\mu_i + p(p+2)n_i^{-2}\}.$$

PROOF. The first formula is a generalisation of the result of DAS GUPTA (1968), lemma 2.4 (ii), in which $c=d$ was given. The second and third formula can be obtained from $COV(X^TS^{-1}X)$, where

$$S\sim W_p(f,\Sigma),\quad \mathrm{vec}(X)\sim N_{r\times p}(\mathrm{vec}(\mu), D\otimes\Sigma)$$

with X and μ of size $p\times r$, $D=diag\{n_1^{-1},...,b_r^{-1}\}$, X_i and $\mu_i\,(i=1,...,r)$ the i-th column of X and μ, respectively. Formula (3) is the second moment of a non-central F-distribution. However, we shall give our own proof in which (1) is derived from lemma 4.4.7 and in which (2) and (3) are derived from (1).

PROOF of (1): Let $V=\Sigma^{-1/2}S\Sigma^{-1/2}$, $a=\Sigma^{-1/2}c$ and $b=\Sigma^{-1/2}d$ then

$$\begin{aligned} ES^{-1}cd^TS^{-1} &= E\Sigma^{-1/2}V^{-1}ab^TV^{-1}\Sigma^{-1/2} \\ &= \Sigma^{-1/2}\sum_{i=1}^{p}\sum_{j=1}^{p}a_ib_j(EV^{-1}\epsilon_i\epsilon_j^TV^{-1})\Sigma^{-1/2} \\ &= \Sigma^{-1/2}\sum_{i=1}^{p}\sum_{j=1}^{p}a_ib_j\begin{bmatrix} V^{1i}V^{j1} & \cdots & V^{1i}V^{jp} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ V^{pi}V^{j1} & \cdots & V^{pi}V^{jp}\end{bmatrix}\Sigma^{-1/2} \end{aligned}$$

where a_i and b_j are the i-th and j-th element of a and b respectively. ϵ_i is the vector with 1 at the i-th and 0 at the other positions. Now, apply lemma 4.4.7 then the expression becomes

$$\Sigma^{-1/2}\sum_{\substack{i=1\\ i\neq j}}^{p}\sum_{j=1}^{p}a_ib_j\{\epsilon_i\epsilon_j^T(f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}(f-p-2)$$
$$+ \epsilon_j\epsilon_i^T(f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}\}\Sigma^{-1/2}$$
$$+ \Sigma^{-1/2}\sum_{i=1}^{p}a_ib_i\{\epsilon_i\epsilon_i^T(f-p-1)^{-1}(f-p-3)^{-1}$$
$$+ (I-\epsilon_i\epsilon_i^T)(f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}\}\Sigma^{-1/2},$$

and the desired result can be obtained directly.

PROOF of (2): The independence of the three statistics enables an easy use of conditional expectations. We obtain

$$\begin{aligned} EX_i^T S^{-1} X_i X_j^T S^{-1} X_j &= E_{X_i, X_j} E\{X_i^T S^{-1} X_i X_j^T S^{-1} X_j | X_i, X_j\} \\ &= E_{X_i, X_j}(X_i^T E\{S^{-1} X_i X_j^T S^{-1} | X_i, X_j\} X_j). \end{aligned}$$

Now, with (1) and using the notation $a=(f-p)^{-1}(f-p-1)^{-1}(f-p-3)^{-1}$, we obtain

$$a(f-p-2)EX_i^T\Sigma^{-1}X_iX_j^T\Sigma^{-1}X_j + 2aEX_i^T\Sigma^{-1}X_jX_i^T\Sigma^{-1}X_j$$

which, using the independence between X_i and X_j, can be written as

$$\begin{aligned} a(f-p-2)\ &\text{trace}(\Sigma^{-1}E(X_iX_i^T))\ \text{trace}(\Sigma^{-1}E(X_jX_j^T)) \\ &+2a\ \text{trace}(\Sigma^{-1}E(X_jX_j^T)\Sigma^{-1}E(X_iX_i^T)). \end{aligned}$$

Now, after the substitution $EX_iX_i^T = VAR(X_i) + EX_iEX_i^T = n_i^{-1}\Sigma + \mu_i\mu_i^T$ and $EX_jX_j^T = n_j^{-1}\Sigma + \mu_j\mu_j^T$, the result is obtained by straightforward computation.

PROOF of (3): Conditioning to X_i gives

$$EX_i^T S^{-1} X_i X_i^T S^{-1} X_i = E_{X_i}(X_i^T E\{S^{-1} X_i X_i^T S^{-1} | X_i\} X_i)$$

and applying (1) gives

$$(f-p-1)^{-1}(f-p-3)^{-1}EX_i^T\Sigma^{-1}X_iX_i^T\Sigma^{-1}X_i$$

in which the second moment of a non-central χ_p^2-distribution appears. Now, $X_i = \mu_i + n_i^{-1/2}\Sigma^{1/2}U$, where $U \sim N_p(0, I)$, with $EU=0$, $EUU^TU=0$, $U^TU \sim \chi_p^2$, $EU^TU=p$ and $EU^TUU^TU = VAR(U^TU) + (EU^TU)^2 = p(p+2)$. Formula (3) can now be derived by straightforward computation.

Remember that the aim of this section is to derive expectations, variances and covariances of the statistics $U_{x;jt}$ defined in (4.4.3). The expectations have been given in (4.4.4). For the covariances we must distinguish between $COV(U_{x;jt}, U_{x;js})$ and $COV(U_{x;jt}\, U_{x;is})$ where i,j,t and s represent four different numbers. From (4.4.3) we see that $U_{x;jt}$ and hence $VAR(U_{x;jt})$, $COV(U_{x;jt}, U_{x;js})$ and $COV(U_{x;jt}, U_{x;is})$ are invariant under the transformation $y \to A(y-x)$, $\Sigma \to A\Sigma A^T$, where $A = \Sigma^{-1/2}$, $y \in \mathbb{R}^p$ and x is the vector in the sample space at which the posterior probabilities are computed. Hence the original problem with $S \sim W_p(f, \Sigma)$, $X_\ell \sim N_p(\mu_\ell, \Sigma)$ $\ell \in \{i,j,t,s\}$ is equivalent to the problem with $x=0$, $S \sim W_p(f, I)$, $X_\ell \sim N_p(\nu_\ell, I)$ where $\nu_\ell = \Sigma^{-1/2}(\mu_\ell - x)$ $\ell \in \{i,j,t,s\}$. From (4.4.3) and (4.4.4) we derive that

$$\begin{aligned} &COV(U_{x;jt}, U_{x;is}) = \\ &= \tfrac{1}{4}(f-p-1)^2 E\{(X^TS^{-1}X_t - X_j^TS^{-1}X_j)(X_S^TS^{-1}X_S - X_i^TS^{-1}X_i)\} \end{aligned}$$

$$-\{\zeta_{x;jt}-\tfrac{1}{2}p(n_j^{-1}-n_t^{-1})\}\{\zeta_{x;is}-\tfrac{1}{2}p(n_i^{-1}-n_s^{-1})\}.$$

Analogue expressions for $COV(U_{x;jt}, U_{x;js})$ and $VAR(U_{x;jt})$ are obtained by replacing the indices i by j and s by t in just-mentioned formula. By applying lemma 4.4.8 and using the notation

$$<i,j> = (x-\mu_i)^T\Sigma^{-1}(x-\mu_j) \quad i,j \in\{i,j,s,t,\}$$

we obtain

$$\begin{aligned}
VAR(U_{x;jt}) &= (f-p-1)(f-p)^{-1}(f-p-3)^{-1}.\\
&\cdot[\tfrac{1}{2}(f-p)(f-p-1)^{-1}(<j,j>-<t,t>)^2\\
&+ <t,t><j,j>-<t,j><t,j>\\
&+ (f-1)(f-p-1)^{-1}\{(f-p)n_j^{-1}-n_t^{-1}\}<j,j>\\
&+ (f-1)(f-p-1)^{-1}\{(f-p)n_t^{-1}-n_j^{-1}\}<t,t>\\
&+ p(f-1)(f-p-1)^{-1}\{\tfrac{1}{2}(f-p)(n_j^{-2}+n_t^{-2})-n_t^{-1}n_j^{-1}\}],\\
COV(U_{x;jt}, U_{x;js}) &= \tfrac{1}{2}(f-p-1)(f-p)^{-1}(f-p-3)^{-1}.\\
&\cdot[(f-p-1)^{-1}(<j,j>-<t,t>)(<j,j>-<s,s>)\\
&+ <j,j><j,j>-<j,s><j,s>\\
&+ <t,s><t,s>-<t,j><t,j>\\
&+ (f-1)(f-p-1)^{-1}(n_j^{-1}-n_s^{-1})(<j,j>-<t,t>)\\
&+ (f-1)(f-p-1)^{-1}(n_j^{-1}-n_t^{-1})(<j,j>-<s,s>)\\
&+ 2(f-1)n_j^{-1}<j,j>\\
&+ p(f-1)(f-p-1)^{-1}(n_j^{-1}-n_t^{-1})(n_j^{-1}-n_s^{-1})\\
&+ p(f-1)n_j^{-2}],
\end{aligned}$$

and

$$COV(U_{x;jt}, U_{x;is}) = \tfrac{1}{2}(f-p-1)(f-p)^{-1}(f-p-3)^{-1}.$$

$$\cdot[(f-p-1)^{-1}(<j,j>-<t,t>)(<i,i>-<s,s>)$$

$$+ <i,j><i,j>-<s,j><s,j>$$

$$+ <t,s><t,s>-<i,t><i,t>$$

$$+ (f-1)(f-p-1)^{-1}(n_i^{-1}-n_s^{-1})(<j,j>-<t,t>)$$

$$+ (f-1)(f-p-1)^{-1}(n_j^{-1}-n_t^{-1})(<i,i>-<s,s>)$$

$$+ p(f-1)(f-p-1)^{-1}(n_i^{-1}-n_s^{-1})(n_j^{-1}-n_t^{-1})].$$

These formulas have also been mentioned in AMBERGEN (1981). The formula of $VAR(U_{x;jt})$ is also given in SCHAAFSMA (1982), p. 873, and in CRITCHLEY and FORD (1984b), section 3.3.

The variances and covariances of the statistics $U_{x;jt}$ can be used for a verification of the asymptotic distribution of the estimator of the vector of posterior probabilities presented in theorem 2.4.1. We have

$$R_{t|x} = \{\sum_{h=1}^{k} \rho_h \rho_t^{-1} \exp(U_{x;ht})\}^{-1} \quad t=1,...,k,$$

where $U_{x;ht}$ is defined in (4.4.3), as estimators for the posterior probabilities

$$\rho_{t|x} = \{\sum_{h=1}^{k} \rho_h \rho_t^{-1} \exp(\zeta_{x;ht})\}^{-1} \quad t=1,...,k,$$

and where $\zeta_{x;ht}$ is defined in (4.4.2). Let us define

$$U_x = (U_{x;11},...,U_{x;k1},...,U_{x;1k},...,U_{x;kk})^T$$

and

$$\zeta_x = (\zeta_{x;11},...,\zeta_{x;k1},...,\zeta_{x;1k},...,\zeta_{x;kk})^T$$

in which the $U_{x;jj}$ and $\zeta_{x;jj}$, $j=1,...,k$ are zero.

THEOREM 4.4.9.
If $n_i/n \to b_i > 0$, $i=1,...,k$ *then*

$$\mathscr{L} n^{1/2}(U_x - \zeta_x) \to N_{k\times k}(0, \Gamma)$$

and

$$\mathscr{L} n^{1/2}(R_{\cdot|x} - \rho_{\cdot|x}) \to N_k(0, B\Gamma B^T)$$

where, with a notation in which $\Gamma_{ts,ij}$ *is the* (i,j)*-th element of the* (t,s)*-th submatrix of size* $k\times k$ *of matrix* Γ, *for* $t=1,...,k$:

$$\Gamma_{tt,jj} = \tfrac{1}{2}<j,j>^2 + \tfrac{1}{2}<t,t,>^2 - <t,j>^2$$
$$+ b_j^{-1}<j,j> + b_t^{-1}<t,t> \qquad j=1,...,k;\ j\neq t$$
$$\Gamma_{tt,ij} = \tfrac{1}{2}<t,t>^2 - \tfrac{1}{2}<t,j>^2 + \tfrac{1}{2}<i,j>^2$$
$$-\tfrac{1}{2}<i,t>^2 + b_t^{-1}<t,t,> \quad i,j = 1,...,k;\ i\neq t; j\neq t; i\neq j$$
$$\Gamma_{tt,tj} = \Gamma_{tt,jt} = 0 \qquad j=1,...,k$$

and for $t,s=1,...,k; t\neq s$:

$$\Gamma_{ts,ij} = \tfrac{1}{2}<i,j>^2 - \tfrac{1}{2}<s,j>^2 + \tfrac{1}{2}<t,s>^2 - \tfrac{1}{2}<i,t>^2$$
$$i\neq j;\ j\neq s;\ j\neq t;\ i\neq s;\ i\neq t$$
$$\Gamma_{ts,it} = -\Gamma_{tt,is} \qquad i\neq t;\ i\neq s$$
$$\Gamma_{ts,sj} = -\Gamma_{ss,tj} \qquad j\neq s;\ j\neq t$$
$$\Gamma_{ts,ii} = \Gamma_{ii,ts} \qquad i\neq t;\ i\neq s$$
$$\Gamma_{ts,tj} = \Gamma_{ts,js} = 0 \qquad j=1,...,k$$
$$\Gamma_{ts,st} = -\Gamma_{ss,tt}$$

and B *is a* $k\times k^2$ *matrix, partitioned into* k *submatrices of size* $k\times k$, *and with the notation* $B_{\ell,ts}$ *the* (t,s)*-th element of the* ℓ*-th submatrix, for* $t=1,...,k$:

$$B_{t,tj} = -\rho_{t|x}^2\rho_j\rho_t^{-1}\exp(\zeta_{x;jt}) = -\rho_{t|x}\rho_{j|x} \qquad j=1,...,k;\ j\neq t$$
$$B_{s,tj} = 0 \qquad j=1,...,k;\ s\neq t$$
$$B_{t,tt} = 0\,.$$

Note that, because of the many zeros in Γ and B, we have

$$(B\Gamma B^T)_{t,s} = \sum_{\substack{i=1\\ i\neq t}}^{k}\sum_{\substack{j=1\\ j\neq s}}^{k}\rho_{t|x}\rho_{i|x}\Gamma_{ts,ij}\rho_{s|x}\rho_{j|x}.$$

We have checked that $B\Gamma B^T=\Psi\Theta\Psi$, where the matrix Θ has been specified in theorem 2.4.1, and the matrix Ψ in formulas (2.1.3) and (2.1.4). We shall not give this computation because it is long and tedious. We confine ourselves to a short outline of the straightforward computation. Partition the set of double indices $I_k=\{(i,j); i=1,...,k; j=1,...,k\}$ into the following subsets. For $t=1,...,k$; $s=1,...,k$ and for the relations $(R_1,...,R_5)\in\times_{i=1}^{5}\{=,\neq\}$ we define

$$S(t,s,R_1,...,R_5) =$$
$$= \{(i,j); ((i,j)\in I_k)\wedge(iR_1j)\wedge(iR_2t)\wedge(jR_3t)\wedge(iR_4s)\wedge(jR_5s)\}.$$

Hence

$$I_k = \bigcup_{(R_1,\ldots,R_5)} S(t,s,R_1,\ldots,R_5)$$

for $t=1,\ldots,k$ and $s=1,\ldots,k$. Easy to verify that for $t=s$ at most 5 and for $t\neq s$ at most 10 nonempty subsets $S(t,s,R_1,\ldots,R_5)$ are defined. For any fixed pair t,s the nonempty subsets are disjunct. Subsequently

$$(B\Gamma B^T)_{t,s} = \sum_{(R_1,\ldots,R_5)} \sum_{(i,j)\in S(t,s,R_1,\ldots,R_5)} \Psi_{ti}\Theta_{ij}\Psi_{js}$$

for $t=1,\ldots,k$ and $s=1,\ldots,k$.

4.5. UNBIASED ESTIMATORS FOR LOGARITHMS OF NORMAL DENSITIES AND THEIR VARIANCES

Unbiased estimators of the logarithm of the densities of the populations are derived when normality is assumed. The variances of these estimators are obtained and used in the asymptotic distribution of the posterior probabilities under various model assumptions. Throughout this section we shall assume that the k populations are characterized by multivariate normal distributions with unequal covariance matrices: $N_p(\mu_h, \Sigma_h)$, $h=1,\ldots,k$. The k posterior probabilities in (4.1.1) can be written as

$$\rho_{t|x} = \rho_t \exp(\lambda_t)/\sum_{h=1}^{k} \rho_h \exp(\lambda_h), \quad t=1,\ldots,k \tag{4.5.1}$$

where

$$\begin{aligned}\lambda_k &= \log(f_h(x)) \\ &= -\tfrac{1}{2}p\log(2\pi) - \tfrac{1}{2}\log|\Sigma_h| - \tfrac{1}{2}\Delta^2_{x;h}\end{aligned} \tag{4.5.2}$$

with

$$\Delta^2_{x;h} = (x-\mu_h)^T\Sigma_h^{-1}(x-\mu_h).$$

The unbiased estimator of λ_h which we shall consider in this section is

$$\begin{aligned}L_h = &-\tfrac{1}{2}p\log(\pi) - \tfrac{1}{2}\log(|S_h|) \\ &-\tfrac{1}{2}(f_h - p - 1)(x - X_{h\cdot})^T S_h^{-1}(x - X_{h\cdot}) \\ &+\tfrac{1}{2}\sum_{i=1}^{p}\Psi(\tfrac{1}{2}(f_h - i + 1)) + \tfrac{1}{2}pn_h^{-1}\end{aligned} \tag{4.5.3}$$

where

$$\Psi(y) = \frac{d}{dy}\{\log(\Gamma(y))\} = \frac{\Gamma'(y)}{\Gamma(y)}$$

with

$$\Gamma(y) = \int_0^\infty x^{y-1}e^{-x}\,dx, \quad y>0,$$

and where $X_{h\cdot}$ and $f_h^{-1}S_h$ are the estimators of the mean and covariance matrix of the h-th distribution. Note that, $X_{h\cdot}\sim N_p(\mu_h, n_h^{-1}\Sigma_h)$ and $S_h\sim W_p(f_h, \Sigma_h)$. Do not confuse this f_h with $f_h(x)$, the latter being the density at point x.

Note also that $\lambda_h-\lambda_t=\zeta_{x;ht}$, defined in (4.4.2), and that L_h-L_t is an unbiased estimator for $\zeta_{x;ht}$. However, L_h-L_t is not equal to $U_{x;ht}$, defined in (4.4.3). This is because the estimators for the covariance matrices are different in the two approaches, which originate from different assumptions about these covariance matrices.

For the variance of L_h we shall derive that

$$VAR(L_h) = \frac{1}{4}\left[\sum_{j=0}^{p-1}\Psi'(\tfrac{1}{2}(f_h-j)) + n_h^{-2}(f_h-p-3)^{-1}\cdot\right.$$
$$\cdot\{2p(f_h-1) + 4n_h(f_h-1)\Delta_{x;h}^2 + 2n_h^2\Delta_{x;h}^4\}$$
$$\left.-4n_h^{-1}(f_h-p-1)^{-1}(p+n_h\Delta_{x;h}^2)\right] \qquad (4.5.4)$$

in which

$$\Psi'(y) = \frac{d^2}{dy^2}\{\log(\Gamma(y))\} = \sum_{i=0}^{\infty}\frac{1}{(y+i)^2}$$

(for this last equality see ERDÉLYI et al. (1953), p.22).

Sometimes, additional observations are available for the estimation of the covariance matrix, while these observations are not used for the estimation of the mean. This implies that the parameter f_h is different from n_h-1. Let there be the following extra observations for the estimation of the covariance matrix Σ_h of the h-th population: a number of e_{h1} extra observations from the h-th population itself, and e_{hj} extra observations from a j-th extra population, which has density $N_p(\mu_{hj}, \Sigma_h)$ with μ_{hj} unknown, for $j=2,...,a_h$. Hence $f_h=n_h+\Sigma_{j=1}^{a_h}(e_{hj}-1)$. The extra observations may not be used for the estimation of the other $k-1$ covariance matrices, and the extra populations are not selected from the other $k-1$ populations. The asymptotic distributions we are looking for will depend on the way the n_h, and e_{hj}, $j=1,...,a_h$; $h=1,...,k$ behave if n, where $n=\Sigma_{h=1}^k n_h$, tends to infinity. Let $e_{hj}/n_h\to c_{hj}$, $j=1,...,a_h$; $h=1,...,k$, and $n_h/n\to b_h>0$, $h=1,...,k$ if $n\to\infty$.

Let us define

$$R_{t|x} = \rho_t \exp(L_t)/\sum_{h=1}^{k}\rho_h \exp(L_h) \quad t=1,...,k$$

as estimators of the posterior probabilities in (4.5.1). Further, let us introduce the notations $L=(L_1,...,L_k)^T$ and $\lambda=(\lambda_1,...,\lambda_k)^T$.

THEOREM 4.5.1.
If $n \to \infty$ *then*

$$\mathscr{L} n^{1/2}(L-\lambda) \to N_k(0, D)$$

and

$$\mathscr{L} n^{1/2}(R_{\cdot|x}-\rho_{\cdot|x}) \to N_k(0, \Psi D \Psi)$$

where D *is the diagonal matrix defined by*

$$D_h = \tfrac{1}{2} b_h^{-1}(1 + \sum_{j=1}^{a_h} c_{hj})^{-1}\{p + 2\sum_{j=1}^{a_h} c_{hj}\Delta^2_{x;h} + \Delta^4_{x;h}\}$$

for $h = 1,\ldots,k$ *and where the matrix* Ψ *is defined in formulas (2.1.3) and (2.1.4).*

PROOF. The proof can be given with the same technique as used for theorem 2.4.4. However, it is interesting to see that the diagonal elements of matrix D can also be derived from (4.5.4). In this derivation one needs that

$$\lim_{n\to\infty} n\Psi'(n) = 1$$

which can be proved with

$$1 = n\int_{n-1}^{\infty} \frac{1}{(t+1)^2}dt \leqslant n\sum_{i=0}^{\infty} \frac{1}{(i+n)^2} \leqslant n\int_{n-1}^{\infty} \frac{1}{t^2}dt = \frac{n}{n-1}$$

where the term in the middle is $n\Psi'(n)$.

REMARK. A special case of theorem 4.5.1 is that in which there are no extra observations. This implies that $c_{hj}=0, j=1,\ldots,a_h; h=1,\ldots,k$. Hence the diagonal matrix D is specified by

$$D_h = \tfrac{1}{2} b_h^{-1}(p + \Delta^4_{x;h}) \quad h = 1,\ldots,k.$$

This is equal to theorem 2.4.4.

In the remaining part of this section we shall prove that L_h in (4.5.3) is an unbiased estimator for λ_h in (4.5.2) and that its variance is (4.5.4). For the sake of convenience we drop the index h and replace $X_{h\cdot}$ by Y, hence $Y \sim N_p(\mu, n^{-1}\Sigma)$ and $S \sim W_p(f, \Sigma)$.

From lemma 4.4.6 (2) and lemma 4.4.4 with $B = \Sigma^{1/2}$ we derive that $ES^{-1} = (f-p-1)^{-1}\Sigma^{-1}$, hence

$$\begin{aligned} &E(f-p-1)(x-Y)^T S^{-1}(x-Y) = \\ &= (f-p-1)\ \text{trace}\{(ES^{-1})(E(x-Y)(x-Y)^T)\} \\ &= (x-\mu)^T\Sigma^{-1}(x-\mu) + n^{-1}p. \end{aligned} \tag{4.5.5}$$

From RAO (1965), p. 540, we see that $|S|/|\Sigma|$ is distributed as the product of p independent central χ^2 variables with degrees of freedom: $f-p+1,\ldots,f-1, f$,

hence

$$\mathcal{L}\frac{|S|}{|\Sigma|} = \mathcal{L}G_1 G_2 \cdots G_p$$

where $G_i, i = 1,...,p$ has the χ^2_{f-p+i} distribution and the G_i's are mutually independent. Taking logarithms gives

$$\mathcal{L}\log\frac{|S|}{|\Sigma|} = \mathcal{L}\sum_{i=1}^{p} \log G_i$$

hence

$$E\ \log|S| = \log|\Sigma| + \sum_{i=1}^{p} E\ \log G_i\ . \tag{4.5.6}$$

Now, we need the following lemma.

LEMMA 4.5.2.
If $U \sim \chi^2_\nu$, *then*

$$E\ \log U = \Psi(\tfrac{1}{2}\nu) + \log 2$$

$$VAR\ \log U = \Psi'(\tfrac{1}{2}\nu).$$

PROOF. We use that $E\log U = M'(s)|_{s=0}$ and $E\log^2 U = M''(s)|_{s=0}$ where $M(s)$ is the moment-generating function of $\log U$:

$$M(s) = Ee^{s\log U} = EU^s,$$

which can be computed by using the density $p_\nu(u)$ of U:

$$p_\nu(u) = u^{\frac{1}{2}\nu-1}\exp(-\tfrac{1}{2}u)(\tfrac{1}{2})^{\frac{1}{2}\nu}\ \frac{1}{\Gamma(\frac{1}{2}\nu)}.$$

We obtain that

$$M(s) = \frac{\Gamma(\frac{1}{2}\nu+s)2^s}{\Gamma(\frac{1}{2}\nu)}$$

and with

$$\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)} \quad \text{and} \quad \Psi'(x) = \frac{\Gamma''(x)}{\Gamma(x)} - (\frac{\Gamma'(x)}{\Gamma(x)})^2$$

the lemma can be proved.

The application of lemma 4.5.2 to (4.5.6) gives

$$E\ \log|S| = \log|\Sigma| + p\log 2 + \sum_{i=1}^{p}\Psi(\tfrac{1}{2}(f-p+i)). \tag{4.5.7}$$

The fact that L_h in (4.5.3) is an unbiased estimator of λ_h in (4.5.2) follows immediately from (4.5.5) and (4.5.7).

For the variance of L_h it can be derived from (4.5.3) that

$$VAR(L_h) = \tfrac{1}{4}VAR(\log|S|) + \tfrac{1}{4}(f-p-1)^2 VAR((x-Y)^T S^{-1}(x-Y)) + \tfrac{1}{2}(f-p-1)COV(\log|S|, (x-Y)^T S^{-1}(x-Y)). \quad (4.5.8)$$

The first two terms of the right-hand side are easiest to derive. For the first term, using (4.5.6) and lemma 4.5.2, we get

$$VAR(\log|S|) = VAR\left(\log\frac{|S|}{|\Sigma|}\right) = \sum_{i=1}^{p} VAR(\log G_i) = \sum_{i=1}^{p} \Psi'\left(\tfrac{1}{2}(f-p+i)\right) \quad (4.5.9)$$

and for the second term of (4.5.8) we find with (4.5.5) and lemma 4.4.8 (3) that

$$VAR((x-Y)^T S^{-1}(x-Y)) = (f-p-1)^{-2}(f-p-3)^{-1}\{2\Delta^4 + 4n^{-1}(f-1)\Delta^2 + 2pn^{-2}(f-1)\} \quad (4.5.10)$$

in which $\Delta^2 = (x-\mu)^T\Sigma^{-1}(x-\mu)$. For the derivation of the third term in (4.5.8) we need the following two lemmas.

LEMMA 4.5.3.
If $S \sim W_p(f, \Sigma)$ *and* $d \in \mathbb{R}^p$ *then*

$$COV(\log|S|, d^T S^{-1} d) = -2(f-p-1)^{-2} d^T\Sigma^{-1} d.$$

PROOF. Let Ψ be an orthogonal matrix with $d^T\Sigma^{-1/2}/\|d^T\Sigma^{-1/2}\|$ as its first row. Let $W = \Sigma^{-1/2} S \Sigma^{-1/2}$ and $V = \Psi^T W \Psi$ then $W \sim W_p(f, I)$ and $V \sim W_p(f, I)$. Now

$$\begin{aligned} COV(\log|S|, d^T S^{-1} d) &= \\ &= COV(\log|\Sigma^{1/2} W \Sigma^{1/2}|, d^T\Sigma^{-1/2} W^{-1}\Sigma^{-1/2} d) \\ &= COV(\log|W|, d^T\Sigma^{-1/2}\Psi^T V^{-1}\Psi\Sigma^{-1/2} d) \\ &= COV(\log|\Psi V \Psi^T|, \epsilon_1^T V^{-1}\epsilon_1 d^T\Sigma^{-1} d) \\ &= d^T\Sigma^{-1} d\, COV(\log|V|, V^{11}). \end{aligned} \quad (4.5.11)$$

Define

$$C = \begin{bmatrix} 1 & -V_{(1,2)}\, V_{(2,2)}^{-1} \\ 0 & I_{p-1} \end{bmatrix}$$

hence $|C|=1$ and

$$CVC^T = \begin{bmatrix} V_{11}-V_{(1,2)}V_{(2,2)}^{-1}V_{(1,2)} & 0 \\ 0 & V_{(2,2)} \end{bmatrix}$$

and

$$\begin{aligned} |V| &= |V_{11}-V_{(1,2)}V_{(2,2)}^{-1}V_{(1,2)}| \cdot |V_{(2,2)}| \\ &= (V^{11})^{-1}|V_{(2,2)}| \end{aligned}$$

where lemma 4.4.3 (1) has been used. Now, using the independence between V^{11} and $V_{(2,2)}$ (see corollary 4.4.2 (2)), expression (4.5.11) becomes

$$-d^T\Sigma^{-1}d\ COV(\log V^{11}, V^{11}) =$$

$$d^T\Sigma^{-1}d\ \{E(V^{11}\log(V^{11})^{-1})-EV^{11}E\ \log(V^{11})^{-1}\}.$$

Now, let $U=(V^{11})^{-1}$ then, according to corollary 4.4.2 (1), $U\sim\chi^2_{f-p+1}$. Further, introduce a G with a χ^2_{f-p-1} distribution, then

$$E(V^{11}\log(V^{11})^{-1}) = E(U^{-1}\log U) =$$

$$= \int_0^\infty u^{-1}(\log u)u^{(f-p+1)/2}\exp(-\tfrac{1}{2}u)(\tfrac{1}{2})^{(f-p+1)/2}\Gamma^{-1}(\tfrac{1}{2}(f-p+1))\,du$$

$$= (f-p-1)^{-1}E\ \log G$$

$$= (f-p-1)^{-1}\{\Psi(\tfrac{1}{2}(f-p-1)) + \log 2\}$$

where lemma (4.5.2) has been used. Subsequently, with lemma 4.4.6 (1) we obtain $EV^{11}=(f-p-1)^{-1}$, and once again using lemma (4.5.2)

$$E\ \log(V^{11})^{-1} = \Psi(\tfrac{1}{2}(f-p+1)) + \log 2.$$

Now, using $\Psi(x+1)=\Psi(x)+\frac{1}{x}$ the proof of the lemma can easily be completed.

LEMMA 4.5.4.
If $S\sim W_p(f, \Sigma)$, $D\sim N_p(\mu, \Sigma)$ and S and D are independent, then

$$COV(\log|S|,\ D^TS^{-1}D) = -2(f-p-1)^{-2}(p+\mu^T\Sigma^{-1}\mu).$$

PROOF. We use that

$$\begin{aligned} &COV(g(S),\ h(D,S)) = \\ &\quad = E_D\{COV((g(S)|D),\ (h(D,S)|D))\} + \\ &\qquad COV(E\{g(S)|D\},\ E\{h(D,S)|D\}) \end{aligned}$$

where g and h are suitable real valued functions. Since $E\{g(S)|D\}$ is a constant the last term disappears. So, the left-hand side of the lemma becomes

$$E_D\{COV(\log|S|,\ (D^TS^{-1}D|D=d))\}$$

and by lemma 4.5.3 this is

$$-2(f-p-1)^{-2}ED^T\Sigma^{-1}D$$

which evaluates to the right-hand side of the lemma, by which the proof is completed.

The last term of the right-hand side of (4.5.8) can be evaluated by using lemma 4.5.4 with $D=n^{1/2}(x-Y)$. So

$$\begin{aligned} &COV(\log|S|,\ (x-Y)^TS^{-1}(x-Y))= \\ &\quad = n^{-1}COV(\log|S|,\ D^TS^{-1}D) \\ &\quad = -2n^{-1}(f-p-1)^{-2}\{p+n(x-\mu)^T\Sigma^{-1}(x-\mu)\}. \end{aligned} \tag{4.5.12}$$

The variance of L_h, given in (4.5.4), can now easily be derived from (4.5.8) and the three formulas (4.5.9), (4.5.10) and (4.5.12).

4.6. A COMPARISON OF THE ACCURACY OF FOUR METHODS OF CONSTRUCTING CONFIDENCE INTERVALS FOR POSTERIOR PROBABILITIES

In this section we shall study the quality of four different methods of constructing confidence intervals for posterior probabilities in a specific model. These methods originate from the use of different estimators. The confidence intervals are constructed on the basis of the asymptotic distributions of these estimators. The quality of the approximate confidence intervals is investigated by means of simulation experiments. The model is that in which a multidimensional observation vector originates from one of k populations, specified by multivariate normal distributions with equal covariance matrices. The estimators are based on training samples from the k populations. Observation vector and prior probabilities are given.

Let x denote the observation vector which comes from one of the populations. These populations are characterized by p-dimensional multivariate normal distributions with equal covariance matrices. Accordingly, let f_h denote the probability density function of $N_p(\mu, \Sigma)$, $h=1,...,k$. The means μ_h, $h=1,...,k$ and the covariance matrix Σ are unknown. For each of the k populations the outcomes $x_{h1},...,x_{hn_h}$ are given of the independent identically distributed random variables $X_{h1},...,X_{hn_h}$ in which X_{hi} has density f_h. Let the prior probabilities be denoted by ρ_h, $h=1,...,k$ and considered given. The posterior probabilities are

$$\rho_{t|x} = \rho_t f_t(x) / \sum_{h=1}^{k} \rho_h f_h(x), \quad t=1,...,k$$

in which

$$f_h(x) = |2\pi\Sigma|^{-1/2}\exp(-\tfrac{1}{2}\Delta^2_{x;h})$$

with

$$\Delta^2_{x;h} = (x-\mu_h)^T\Sigma^{-1}(x-\mu_h)\,.$$

By cancelling the factor $|2\pi\Sigma|^{-1/2}$ we get the simpler expression

$$\rho_{t|x} = \rho_t\exp(-\tfrac{1}{2}\Delta^2_{x;t})/\sum_{h=1}^{k}\rho_h\exp(-\tfrac{1}{2}\Delta^2_{x;h})\,,$$

see also formula (2.4.14).

Let X_h denote the mean of the sample of the h-th population and let S be the pooled matrix of corrected sums of squares and cross products

$$X_h = n_h^{-1}\sum_{i=1}^{n_h} X_{hi}$$

$$S = \sum_{h=1}^{k}\sum_{i=1}^{n_h}(X_{hi}-X_h)(X_{hi}-X_h)^T.$$

Take $n=n_1+\cdots+n_k$ and $f=n-k$ then $X_h\sim N_p(\mu_h, n_h^{-1}\Sigma)$ and $S\sim W_p(f,\Sigma)$. With the notation

$$V^2_{x;h} = (x-X_h)^TS^{-1}(x-X_h)\,,$$

the four different estimators $R^{(j)}_{t|x}$, $j=1,\ldots,4$ for $\rho_{t|x}$, $t=1,\ldots,k$ are defined by

$$R^{(j)}_{t|x} = \rho_t\exp(-\tfrac{1}{2}\hat\Delta^{2(j)}_{x;t})/\sum_{h=1}^{k}\rho_h\exp(-\tfrac{1}{2}\hat\Delta^{2(j)}_{x;h})$$

in which $\hat\Delta^{2(j)}_{x;h}$ is an estimator for $\Delta^2_{x;h}$:

$$\hat\Delta^{2(1)}_{x;h} = (f+k)\,V^2_{x;h}$$
$$\hat\Delta^{2(2)}_{x;h} = f\,V^2_{x;h}$$
$$\hat\Delta^{2(3)}_{x;h} = (f-p-1)\,V^2_{x;h}$$
$$\hat\Delta^{2(4)}_{x;h} = (f-p-1)\,V^2_{x;h} - pn_h^{-1}\,.$$

So, the estimator $R^{(1)}_{t|x}$ is obtained by plugging the maximum likelihood estimators X_h and $(f+k)^{-1}S$ for μ_h and Σ, respectively, into the formula of $\Delta^2_{x;h}$. The other three estimators are obtained by applying the principle of unbiasedness. The second estimator is based on $ES=f\Sigma$ and the third one on $ES^{-1}=(f-p-1)^{-1}\Sigma^{-1}$ (see MUIRHEAD (1982), p. 97). The fourth estimator satisfies

$$E\hat\Delta^{2(4)}_{x;h}=\Delta^2_{x;h}\,.$$

This can be verified by showing that

$$E\;V^2_{x;h} = (f-p-1)^{-1}\Delta^2_{x;h} + n_h^{-1}\,(f-p-1)^{-1}p\,.$$

We derive this as follows

$$\begin{aligned} E\ V^2_{x;h} &= E(x-X_h)^T S^{-1}(x-X_h) \\ &= E\ \text{trace}\{(x-X_h)^T S^{-1}(x-X_h)\} \\ &= E\ \text{trace}\{S^{-1}(x-X_h)(x-X_h)^T\} \\ &= \text{trace}\ E\{S^{-1}(x-X_h)(x-X_h)^T\} \\ &= \text{trace}\ \{(ES^{-1})(E(x-X_h)(x-X_h)^T)\} \\ &= \text{trace}\ \{(ES^{-1})(E(x-X_h)E(x-X_h)^T + VAR(x-X_h))\}\ , \end{aligned}$$

where we have used that S and X_h are independent, that $E\,\text{trace}E\{\cdot\}=\text{trace}E\{\cdot\}$, and that $\text{trace}\{AB\}=\text{trace}\{BA\}$. By putting $ES^{-1}=(f-p-1)^{-1}\Sigma^{-1}$, $E(x-X_h)=x-\mu_h$ and $VAR(x-X_h)=n_h^{-1}\Sigma$ into the formula, the desired result is easily obtained.

The distributions of the four estimators defined above are in first order asymptotically the same. For the estimator $R^{(1)}_{t|x}$ the asymptotic distribution was derived in chapter 2, theorem 2.4.1. With the notation $R^{(j)}_{|x}=(R^{(j)}_{1|x},\ldots,R^{(j)}_{k|x})$ this distributional result is given by

$$\mathcal{L}n^{1/2}(R^{(j)}_{|x}-\rho_{\cdot|x}) \to N_k(0,\Psi\Theta\Psi)$$

with Θ defined in theorem 2.4.1 and Ψ in (2.1.3) and (2.1.4). From these asymptotic distributions we obtain the approximate $100(1-\alpha)\%$ confidence intervals for $\rho_{t|x}$, $t=1,\ldots,k$

$$[R^{(j)}_{t|x}-\tfrac{1}{2}L^{(j)}_{t|x},\ R^{(j)}_{t|x}+\tfrac{1}{2}L^{(j)}_{t|x}],\ j=1,\ldots,4$$

where

$$L^{(j)}_{t|x} = 2u_{\frac{1}{2}\alpha}n^{-1/2}\{(\hat{\Psi}^{(j)}\hat{\Theta}^{(j)}\hat{\Psi}^{(j)})_{t,t}\}^{1/2}$$

denotes the length of the confidence interval, and in which $u_{\frac{1}{2}\alpha}$ is defined by $P(U>u_{\frac{1}{2}\alpha})=\frac{1}{2}\alpha$ if U has a standard-normal distribution. The estimators $\hat{\Theta}^{(j)}$ and $\hat{\Psi}^{(j)}$ are defined by plugging-in the estimators $R^{(j)}_{t|x}$ for $\rho_{t|x}$, $\hat{\Delta}^{2(j)}_{h;x}$ for $\hat{\Delta}^2_{x;h}$, and $b^{(j)}(x-X_h)^T S^{-1}(x-X_\ell)+c^{(j)}$ for $(x-\mu_h)^T\Sigma^{-1}(x-\mu_\ell)$ in the matrices Θ and Ψ, respectively, $(j=1,\ldots,4)$, where $b^{(1)}=f+k$, $b^{(2)}=f$, $b^{(3)}=b^{(4)}=f-p-1$, $c^{(1)}=c^{(2)}=c^{(3)}=0$, and $c^{(4)}=-pn_h^{-1}$ if $h=\ell$ and $c^{(4)}=0$ if $h\neq\ell$.

In order to investigate the reliability of the four methods of constructing approximate confidence intervals, the following simulation experiment was carried out. Note that an overall comparison of small sample performance of the estimators $R^{(j)}_{t|x}$ and the approximate confidence intervals $R^{(j)}_{t|x}\pm\frac{1}{2}L^{(j)}_{t|x}$ $(j=1,\ldots,4)$ is rather complicated because the performance depends on the very large number of parameters

$$p,\ k,\ n_1,\ldots,n_k,\ t,\ \alpha,\ x,\ \rho_1,\ldots,\rho_k,\ \mu_1,\ldots,\mu_k,\ \Sigma$$

where t indicates the number of the population from which the vector of scores x has been drawn. We selected 500 parameter points for the simulation experiment, and we did the following for each point: compute $\rho_{t|x}$, generate 1000 times a set of training samples and compute each time $R_{t|x}^{(j)}$, $L_{t|x}^{(j)}(j=1,...,4)$. Count the number of times the intervals contains the true value $\rho_{t|x}$ and divide this number by 10, so that it can be compared with the $100(1-\alpha)$. The 500 points were grouped into 25 clusters of 20 points each. Within a cluster only the x vectors differ because they were drawn independently. For the points within a cluster the same training set was used. We made the restrictions $t=1, \alpha=0.05, \mu_1=0_p, \Sigma=I_p, \rho_h=k^{-1}$ $(h=1,...,k)$ and considered only $\rho_{1|x}$ which is the most important, because often largest, posterior probability. For each cluster we averaged the results of the 20 points. These average results with their standard deviations are presented in Tabl 4.1; a cluster corresponds with a row in the table. In order to get a nice layout of the table we introduce the following notations:

$$n=(n_1,...,n_k); \quad \mu=(\mu_1;...;\mu_k);$$

$$a=(0,0,0,0)^T; \quad b=(2,0,0,0)^T; \quad c=(0,2,0,0)^T; \quad d=(1,1,1,1)^T$$

$$e=(1,1,0,0)^T; \quad f=(0,0,2,0)^T; \quad g=(0,0,0,2)^T; \quad h=(0,0,1,1)^T$$

$$1_4=(1,1,1,1); \quad 1_8=(1_4;1_4); \quad m_4=(0,1,0,1); \quad m_8=(m_4;m_4).$$

Bias, mean square error (m.s.e) and mean absolute deviation (m.a.d.) of the point estimators $R_{t|x}^{(j)}$ $(j=1,...,4, t=1)$ were also studied.

For the chosen parameter points we conclude that the m.l. estimator $R_{t|x}^{(1)}$ has a smaller bias, a smaller m.a.d. and a smaller m.s.e. than its competitors, at least on the average. Table 4.1 shows that the confidence intervals for $j=2, 3$ and 4 are slightly more reliable than those based on the m.l. estimator $(j=1)$. Sample sizes should certainly not be smaller than 50 (25) if one requires that the true confidence coefficient of the interval based on the m.l. estimator and $1-\alpha=0.95$ should not be smaller than 0.90 (0.85).

The results of this section were published earlier in the *Journal of Multivariate Analysis, 16,* 432-439, (1985). Publication in this thesis is with permission of Academic Press, Inc.

Input parameter values for the clusters		Averaged confidence coefficients with standard deviations for the four methods							
p, k, μ	n	$j=1$		$j=2$		$j=3$		$j=4$	
$p=4$	50.1_4	92.0	2.0	92.8	1.8	93.3	1.6	93.0	1.6
$k=4$	50.1_4-25m_4	90.8	2.6	91.8	2.3	92.8	1.9	92.3	2.0
$\mu=(abcd)$	25.1_4+25m_4	90.1	2.9	91.2	3.0	92.1	2.7	91.6	2.4
	25.1_4	88.9	3.2	90.6	2.7	92.0	2.1	91.4	1.9
	15.1_4	84.3	4.5	87.2	4.1	89.7	3.4	88.4	3.2
$p=4$	50.1_8	92.4	2.2	93.1	1.7	93.3	1.4	92.9	1.2
$k=8$	50.1_8-25m_8	92.3	2.2	93.1	1.4	93.4	1.1	92.4	1.2
$\mu=$	25.1_8+25m_8	90.0	3.0	91.8	2.3	92.2	1.8	91.7	1.5
$(ab\cdots gh)$	25.1_8	90.5	3.3	91.6	2.3	92.2	1.6	91.3	1.3
	15.1_8	87.7	4.8	89.8	3.3	90.0	2.2	89.4	1.6
$p=8$	50.1_4	88.7	1.6	89.3	1.5	90.9	1.5	90.4	1.6
$k=4$	50.1_4-25m_4	87.0	2.4	87.9	2.2	90.0	2.1	89.4	1.9
	25.1_4+25m_4	86.3	2.4	87.5	2.2	89.6	2.0	89.1	2.0
$\mu=\begin{bmatrix} abcd \\ aaaa \end{bmatrix}$	25.1_4	83.4	2.4	85.0	2.2	88.2	2.3	87.2	2.3
	15.1_4	76.2	4.0	78.7	3.8	84.8	3.3	83.1	3.3
$p=8$	50.1_4	86.4	2.3	88.0	2.1	91.4	2.7	91.1	2.7
$k=4$	50.1_4-25m_4	83.8	2.5	86.0	2.4	90.4	2.8	90.5	3.1
	25.1_4+25m_4	84.4	2.6	86.2	2.7	90.3	3.5	89.2	3.2
$\mu=\begin{bmatrix} abcd \\ aaaa \end{bmatrix}$	25.1_4	80.2	2.7	83.3	2.8	89.0	4.4	88.6	4.6
	15.1_4	73.3	3.8	77.5	4.3	86.5	6.3	85.6	6.7
$p=8$	50.1_8	89.4	2.2	90.0	1.8	90.8	1.4	90.3	1.4
$k=8$	50.1_8-25m_8	89.1	2.1	90.0	2.1	90.9	1.6	89.4	1.7
$\mu=$	25.1_8+25m_8	85.9	3.1	86.7	2.6	87.9	1.9	88.2	1.7
$\begin{bmatrix} ab\cdots gh \\ aa\cdots aa \end{bmatrix}$	25.1_8	85.0	4.2	86.6	3.2	88.0	2.3	87.3	2.2
	15.1_8	79.2	5.2	82.0	3.6	84.8	2.8	83.5	2.6

TABLE 4.1. The Reliability of the Confidence Intervals

Chapter 5

Application and simulation study

5.1. THE BORDER CAVE CRANIUM

In this section we shall illustrate our theory with an application from physical anthropology. Group membership discussions are important in physical anthropology and the often occurring impossibility to increase the size of the training samples in this field of science leads to the need to express the statistical uncertainties in the group membership probabilities. In our theory this need is satisfied by presenting standard deviations for the posterior probabilities.

The following application was suggested by the physical anthropologist G.N. van Vark, University of Groningen, and deals with the Border Cave cranium, a famous specimen in physical anthropology.

In 1940, W.E. Horton, while digging for guano at Border Cave (near the boundary between Swaziland and Zululand, South Africa) found human remains including a partial adult cranium. More of the cranium was found during 1941-42. It is widely agreed that the cranium can be associated with Middle Stone Age industry, date 90.000-110.000 B.P.. However, the group membership of the cranium has been the subject of a controversy, described in a number of articles. The basic question of interest being whether it is of Negro (Zulu, Sotho, Venda, Teita, Dogon, Nguni, Shangana-Tonga, etc.) or of Khoisan (Bushman and Hottentot, with present-day descendants San and Khoikhoin, respectively) origin. Among the first articles we mention COOK, MALAN and WELLS (1945), WELLS (1950, 1969, 1972) and BROTHWELL (1963). Discriminant analysis techniques were used in DE VILLIERS (1973), RIGHTMIRE (1979, 1981), CAMPBELL (1980, 1984), DE VILLIERS and FATTI (1982), AMBERGEN and SCHAAFSMA (1984), and FATTI (1985). Some more interesting articles about the Border Cave cranium are those of BEAUMONT et al. (1972, 1978).

More about multivariate statistical techniques used in research to skeletal remains can be found in VAN VARK (1970, 1976) and in VAN VARK and VAN DER SMAN (1982).

Our evaluation of Border Cave is based on a comparison with crania from $k=8$ African populations, namely males and females of the Bushman, Zulu, Dogon and Teita (see Table 5.1). We used samples from van Varks's database.

Measurements	Border Cave	Bushmen		Zulu		Dogon		Teita	
		males	females	males	females	males	females	males	females
		$N_1=41$	$N_2=49$	$N_3=55$	$N_4=46$	$N_5=48$	$N_6=53$	$N_7=34$	$N_8=49$
1 SOS, Supraorbital projection	10	6.73	5.69	6.18	5.24	5.40	4.08	6.44	4.94
2 FMB, Bifrontal breadth	112	97.27	93.90	101.98	97.74	99.54	94.34	100.06	95.43
3 NAS, Nasio-frontal subtense	15	15.41	16.20	17.84	16.48	16.46	15.45	18.79	17.12
4 NFA, Nasio-frontal angle	150	143.20	143.65	141.51	142.70	143.46	143.68	138.88	140.49
5 WMH, Check height	21	20.93	19.84	20.73	20.06	21.21	19.96	22.21	20.18
6 FRC, Nasion-bregma chord	116	109.17	105.10	111.69	109.39	110.00	105.66	108.71	105.76
7 FRS, Nasion-bregma subtense	32	28.46	28.22	27.71	27.70	26.69	25.64	26.62	27.02
8 FRF, Nasion-subtense fraction	51	47.59	45.08	47.16	46.04	47.88	44.62	48.82	47.37
9 FRA, Frontal angle	122	124.29	122.73	126.33	125.33	127.58	127.28	127.41	125.43
10 OBB, Orbit breadth, left	45	39.27	37.67	40.44	39.20	39.71	38.08	39.65	37.76
11 MDH, Mastoid height	26	25.24	21.61	28.42	25.61	29.06	25.21	29.09	24.18

TABLE 5.1. Measurements of Border Cave compared with means for eight modern African populations

Unfortunately Hottentots were absent from this database. RIGHTMIRE (1979) had concluded that Border Cave is closest to the Hottentot centroid. However, RIGHTMIRE's (1979) Table 5.2 shows that Hottentot males and Bushman males are very similar. As a consequence we are not very concerned

		standard deviation	Correlation-matrix										
			1	2	3	4	5	6	7	8	9	10	11
1	SOS	1.18	1.00										
2	FMB	3.43	0.29	1.00									
3	NAS	2.21	0.25	0.33	1.00								
4	NFA	4.43	-0.18	-0.07	-0.96	1.00							
5	WMH	2.17	0.06	0.26	0.03	0.04	1.00						
6	FRC	4.63	0.04	0.22	0.09	-0.04	0.22	1.00					
7	FRS	2.62	-0.01	0.06	-0.12	0.15	-0.02	0.60	1.00				
8	FRF	3.46	-0.03	0.10	0.02	0.01	0.22	0.53	0.18	1.00			
9	FRA	3.78	0.03	0.04	0.19	-0.19	0.14	-0.18	-0.89	0.18	1.00		
10	OBB	1.65	0.08	0.63	0.25	-0.09	0.04	0.11	-0.01	0.02	0.06	1.00	
11	MDH	3.14	0.11	0.18	0.04	0.01	0.16	0.12	0.03	0.00	0.02	0.12	1.00

TABLE 5.2. Standard deviations and correlation matrix for the eleven measurements in the eight populations for the case with homogeneity of dispersion matrices

about the missing Hottentots. We used $p=11$ variables, see Table 5.1, conforming to the definitions of HOWELLS' (1973) measurement system.

We shall show that the Border Cave cranium is not typical for any of the populations investigated by us. The original question whether Border Cave is of Negro or of Khoison origin is not property solved by comparing Border Cave with samples from subpopulations. Nevertheless this may be regarded as a first step.

By making univariate comparisons (e.g. on the basis of Student's two sample test) using the scores of Border Cave, the means of the populations in Table 5.1, and the standard deviations in Table 5.2, it is clear that Border Cave is not very typical for any of the eight populations, in fact it looks rather atypical.

The same conclusions can be reached on the basis of multivariate considerations. We shall consider three methods to discuss the typicality of the specimen.

(a) First we compute estimates of the Mahalanobis distances between Border Cave and the populations. Two cases, one with homogeneity of covariance matrices, the other without this assumption are considered. The following notations are used. X_h is the mean of the h-th sample, S_h the matrix of corrected sums of squares and cross products of the measurements for the h-th sample, $S=\Sigma_{h=1}^{k} S_h$ the pooled matrix of corrected sums of squares and cross products and $n=n_1+\cdots+n_k$. In the case of homogeneity of covariance matrices, the squared Mahalanobis distances are defined as

$$\Delta^2_{x;hh} = (x-\mu_h)^T\Sigma^{-1}(x-\mu_h)$$

for $h=1,\ldots,k$ and their maximum likelihood estimators are

$$\hat{\Delta}^2_{x;hh} = n(x-X_h)^T S^{-1}(x-X_h).$$

In the case that no assumptions are made about the covariance matrices the corresponding formulas become

$$\Delta^2_{x;h} = (x-\mu_h)^T\Sigma_h^{-1}(x-\mu_h)$$

and

$$\hat{\Delta}^2_{x;h} = n_h(x-X_h)^T S_h^{-1}(x-X_h).$$

The numerical values of $\hat{\Delta}_{x;hh}=\{\hat{\Delta}^2_{x;hh}\}^{1/2}$ and $\hat{\Delta}_{x;h}=\{\hat{\Delta}^2_{x;h}\}^{1/2}$ are presented in Table 5.3 and Table 5.4. The shortest distances in the two models are $\hat{\Delta}_{x;11}=6.4822$ for the Bushman males and $\hat{\Delta}_{x;3}=6.1420$ for the Zulu males. These are so large that Border Cave can not be seen as a random drawing from any of the 8 populations, at least not if the estimates X_h, S and S_h are regarded as the true values of the corresponding population parameters.

(b) Secondly, we can perform Hotelling tests to test the null hypothesis that Border Cave is from the same population as the h-th sample. Thus the analysis is based on the assumption of normality. Tests for the homogeneity of covariance matrices can also be performed, though this has not been done.

In the case of homogeneity, the Hotelling T^2 statistic

$$n_h(n-k-p+1)n^{-1}(n_h+1)^{-1}p^{-1}\hat{\Delta}^2_{x;hh}$$

has the $F(p, n-k-p+1)$ distribution (see RAO (1965), 8.b.2. XII). In the case without the assumption of homogeneity the statistic

$$(n_h-p)\,(n_h+1)^{-1}p^{-1}\hat{\Delta}^2_{x;h}$$

has the $F(p, n_h-p)$ distribution. Table 5.3 and Table 5.4 show the results. In the column with the heading "F-prob" the level of significance is denoted. We see that all the levels are below 5%. Hence the null hypothesis that Border Cave and the h-th training sample are from the same population is rejected ($h=1,\ldots,8$).

(c) The third method deals with typicality probabilities. If x is the vector of scores of Border Cave then we define the typicality probability of Border Cave with respect to population h as

$$\alpha_h(x) = P(G>(x-\mu_h)^T\Sigma_h^{-1}(x-\mu_h))$$

where G has the χ^2_p distribution. The atypicality of vector x with respect to population h is given by $1-\alpha_h(x)$. The larger the Mahalanobis distance between x and the centre μ_h of the population, the less typical x is for this population, and the smaller the typicality probability $\alpha_h(x)$. Note that, although x is considered fixed, $\alpha_h(x)$ is an unknown parameter because $\alpha_h(x)$ depends on μ_h and Σ_h. Hence it makes sense to consider confidence intervals for the typicality probabilities. Under the assumption of normality of distributions and homogeneity of dispersion matrices, an exact confidence interval for $\Delta^2_{x;hh}$ can be obtained by using the fact that

$$n_h\,(n-k-p+1)\,n^{-1}p^{-1}\hat{\Delta}^2_{x;hh}$$

has the noncentral F distribution with p and $n-k-p+1$ degrees of freedom and noncentrality parameter $n_h\Delta^2_{x;hh}$ (see RAO (1965) 8.b.2 XII). In the present study one might content oneself with approximate results based on the unbiased estimator

$$\tilde{\tilde{\Delta}}^2_{x;hh} = (n-k-p-1)n^{-1}\Delta^2_{x;hh}-n_h^{-1}p$$

as estimator for $\Delta^2_{x;hh}$ and with corresponding standard deviation

$$\text{st.dev.}\,(\tilde{\tilde{\Delta}}^2_{x;hh}) = [(n-k-p-3)^{-1}\{2\Delta^4_{x;hh} + \\ + 4(n-k-1)n_h^{-1}\Delta^2_{x;hh} + 2p(n-k-1)n_h^{-2}\}]^{1/2}.$$

Approximate 95% confidence intervals for the $\Delta^2_{x;hh}$ can be obtained with the formula

$$\tilde{\tilde{\Delta}}^2_{x;hh} \pm 1.96\ \text{st.dev.}\,(\tilde{\tilde{\Delta}}^2_{x;hh})$$

and an approximate 95% confidence interval for the typicality probability $\alpha_h(x)$ is given by

$$(P(\chi_p^2 \geqslant U.B.\ (\Delta_{x;hh}^2)),\ P(\chi_p^2 \geqslant L.B.\ (\Delta_{x;hh}^2)))$$

where *U.B.* and *L.B.* mean upper and lower bound of the confidence interval, respectively. Numerical values are given in Table 5.3. We may conclude that Border Cave is very atypical for any of the 8 populations when normality of distributions is postulated. If we drop the assumption of homogeneity of dispersion matrices, then additional uncertainty appears because Σ_h has to be estimated on the basis of sample h only, $(h = 1,...,8)$. We shall derive approximate confidence intervals in the same way as before. Exact confidence intervals for $\Delta_{x;h}^2$ can be derived from

$$(n_h - p)\, p^{-1} \hat{\Delta}_{x;h}^2$$

which has a noncentral $F(p,\ n_h - p;\ n_h \Delta_{h,x}^2)$ distribution. The unbiased estimator

$$\tilde{\tilde{\Delta}}_{x;h}^2 = (n_h - p - 2) n_h^{-1} \hat{\Delta}_{x;h}^2 - n_h^{-1}\, p$$

for $\Delta_{x;h}^2$ has standard deviation

$$\text{st.dev.}\ (\tilde{\tilde{\Delta}}_{x;h}^2) = [(n_h - p - 4)^{-1} \{2\Delta_{x;h}^4 + 4(n_h - 2) n_h^{-1} \Delta_{x;h}^2 + 2p(n_h - 2) n_h^{-2}\}]^{1/2}.$$

See Table 5.4 for the numerical values of both quantities. Approximate 95% confidence intervals for the $\Delta_{x;h}^2$ are given by

$$\tilde{\tilde{\Delta}}_{x;h}^2 \pm 1.96\ \text{st.dev.}\ (\tilde{\tilde{\Delta}}_{x;h}^2)$$

and with

$$(P(\chi_p^2 \geqslant U.B.(\Delta_{x;h}^2)),\ P(\chi_p^2 \geqslant L.B.(\Delta_{x;h}^2)))$$

we obtain the approximate 95% confidence intervals for the typicality probabilities $\alpha_h(x), h = 1,...,k$. From Table 5.4 we see that the intervals are much larger than in Table 5.3, the case of homogeneity of dispersion matrices. The two largest intervals are those of the Bushman males and Zulu males. Interesting is that for these two populations the $\tilde{\tilde{\Delta}}_{x;h}$'s are almost equally large, but that the lengths of the approximate intervals for the typicality probabilities are very different (for Bushman males, twice as large as for Zulu males). This is caused by the difference in sample sizes (41 for the Bushman males and 55 for the Zulu males). For both populations the estimate 0.0026 for $\alpha_h(x) = P(\chi_p^2 \geqslant \Delta_{x;h}^2)$ is very small. Hence the conclusion is that Border Cave is not typical for any of the populations.

Our overall conclusion is that, whichever way we turn, Border Cave is not typical for any of the 8 populations and cannot be seen as a random drawing from any of the populations involved.

In spite of the above conclusion it is illustrative to compute the posterior probabilities and their standard deviations. They are shown in Table 5.3 and Table 5.4 for equal prior probabilities. Because Border Cave is not a random

drawing from any of the 8 populations, these posterior probabilities have not the usual meaning and should not be attached to Border Cave. However, let us assume that, in some purely hypothetical situation, a random drawing from

	N_i	$\hat{\Delta}_{x;hh}$	F-prob	$\hat{\bar{\Delta}}_{x;hh}$	st.dev$(\hat{\tilde{\Delta}}^2_{x;hh})$	$L.B.(\alpha_h(x))$	$U.B.(\alpha_h(x))$	$\hat{\rho}_{h\|x}$	st.dev$(\hat{\rho}_{h\|x})$
Bushman males	41	6.4822	0.025	39.4976	3.5915	0.0000	0.0006	0.7493	0.2827
Bushman females	49	7.5935	0.008	54.3450	4.6302	0.0000	0.0000	0.0003	0.0005
Zulu males	55	6.6615	0.021	41.7962	3.6213	0.0000	0.0003	0.2307	0.2713
Zulu females	46	7.1284	0.013	47.8503	4.1680	0.0000	0.0000	0.0092	0.0131
Dogon males	48	7.1309	0.013	47.8940	4.1496	0.0000	0.0000	0.0091	0.0131
Dogon females	53	8.3023	0.004	65.0248	5.4028	0.0000	0.0000	0.0000	0.0000
Teita males	34	7.3834	0.009	51.2680	4.6099	0.0000	0.0000	0.0015	0.0024
Teita females	49	7.9342	0.006	59.3516	5.0106	0.0000	0.0000	0.0000	0.0000

TABLE 5.3. Statistical quantities of Border Cave in the case of homogeneity of dispersion matrices.

one of 8 populations has the posterior probabilities and standard deviations presented in Table 5.3 and Table 5.4. If conclusions are based only on the point estimates of the posterior probablities, then Table 5.3 suggests that the drawing is from population 1 (probability 0.75) and Table 5.4 suggests that the drawing is from population 3 (probability 0.94). However, if standard deviations are taken into account, Table 5.3 as wel as Table 5.4 suggest that the

	N_i	$\hat{\Delta}_{x;h}$	F-prob	$\hat{\bar{\Delta}}_{x;h}$	st.dev$(\hat{\tilde{\Delta}}^2_{x;h})$	$L.B.(\alpha_h(x))$	$U.B.(\alpha_h(x))$	$\hat{\rho}_{h\|x}$	st.dev$(\hat{\rho}_{h\|x})$
Bushman males	41	6.5025	0.044	28.6076	8.1951	0.0000	0.3241	0.0558	0.3112
Bushman females	49	7.6926	0.014	43.2518	10.7208	0.0000	0.0226	0.0000	0.0002
Zulu males	55	6.1420	0.046	28.6075	6.6095	0.0000	0.1545	0.9434	0.3128
Zulu females	46	7.3389	0.020	38.3992	9.9941	0.0000	0.0646	0.0006	0.0040
Dogon males	48	7.4473	0.018	40.2121	10.1334	0.0000	0.0408	0.0002	0.0014
Dogon females	53	9.7556	0.002	71.6202	16.6504	0.0000	0.0001	0.0000	0.0000
Teita males	34	9.8733	0.004	59.8860	19.7334	0.0000	0.0313	0.0000	0.0000
Teita females	49	10.9610	0.001	88.0442	21.5855	0.0000	0.0000	0.0000	0.0000

TABLE 5.4. Statistical quantities of Border Cave in the case of heterogeneity of dispersion matrices.

drawing can be from both population 1 and population 3. So, rather than the population with the largest posterior probability, the two populations with largest posterior probablities are chosen. Fortunately, in both tables, these are the same two populations. Hence the conflict between the two allocations in the case where assignments were made to the most probable population, has disappeared.

5.2. THE COMPUTER PROGRAM POSCON

This section contains a short description of the computer program POSCON. The name POSCON was chosen by its programmer D.M. van der Sluis as a contraction of the words "posterior probability" and "confidence interval". Estimates of posterior probabilities, standard deviations, and correlations between the estimates of the posterior probabilities can be computed by the program. It is assumed that k mutually exclusive populations are involved and that the individual under investigation belongs to one of these k populations. The individual is characterized by a vector of scores. The k prior probabilities for the individual can be given, possibly with involved uncertainties. The probability densities of the populations are computed from training samples. The formula of Bayes is used for the computation of the estimates of the posterior probabilities. The standard deviations of the estimates of the posterior probabilities are computed from the asymptotic distribution of the estimates. Further output of the program are estimates of typicality probabilities with their standard deviations, and F-probabilities of related Hotelling tests.

The computer program POSCON consists of three parts CREATE, CHANGE and RUN. A so-called POSCON system file contains the relevant information, such as a database of training samples from the populations. The two parts CREATE and CHANGE are used for creating and changing this database. The real computation is carried out by the part RUN on the basis of the statistical model specified by the user.

POSCON, written in fortran 77, can be used interactively, but also commands from a special command file can be read. In the interactive mode the user has to type special commands or to use displayed menus. The program has a number of scratch files. One of them is a backup file of the input commands. This file can be used as a command file in a next run. Another scratch file contains the output written on the display as well as all kinds of information not automatically displayed. This scratch file can also be displayed.

In the part RUN information is asked about the individual whose posterior probabilities have to be estimated. The user has to give the vector of scores of the individual. He has to select a number of variables and populations from the database, let's say p and k, respectively, and he has to specify the k prior probabilities. Next a probabilistic context has to be chosen. This consists of (1) an eventual partitioning of the set of p variables into subsets, and (2), a specification of the probabilistic models for the subsets. The subsets are regarded as stochastically independent for each of the k populations. Furthermore they are also considered as independent of the information on the basis of which the prior probabilities are specified. For a choice of a probabilistic model for a subset of variables the following models are available.

(1) DIS, all variables are discrete.
(2) NOR, all variables are normally distributed.
(3) NEC, all variables are normally distributed with the additional assumption that the covariance matrices of the k populations are equal.

(4) MIX, some variables are discrete, the other ones are normally distributed conditional on the discrete ones.

The theory behind the above-mentioned four models has been presented in chapter two of this monograph.

The model MIX requires that some further choices have to be made. Let c be the number of possible outcomes of the vector of discrete variables. Hence a training sample can be divided into c classes by collecting all observations with the same outcome of the vector of discrete variables. With the command RECODE the user can combine various classes into one new class. Let d be the number of classes obtained in this way. A further assumption is that the continuous variables, given the discrete variables, have a multivariate normal distribution, which depends on the class to which the outcome of the discrete variables belongs. With respect to the $k \times d$ covariance matrices Σ_{hj}, $h = 1,...,k; j = 1,...,d$ the program offers the following four options.

(1) NOC, no constraints for the covariance matrices.
(2) CLC, column constraints, $\Sigma_{1j} = \cdots = \Sigma_{kj}$, $j = 1,...,d$.
(3) RWC, row constraints, $\Sigma_{h1} = \cdots = \Sigma_{hd}$, $h = 1,...,k$.
(4) MXC, maximal constraints, $\Sigma_{11} = \cdots = \Sigma_{kd}$.

The model MIX is based on theory described in section 2.5 of this monograph. More about the computer program POSCON can be found in VAN DER SLUIS et al. (1984, 1985, 1986).

5.3 A SIMULATION STUDY IN THE CASE OF BOTH CONTINUOUS AND DISCRETE VARIABLES

In this section we study the quality of point estimates, standard deviations and confidence intervals for posterior probabilities in the case of both continuous and discrete variables. The study is based on simulations for the four cases mentioned in section 2.1 under ad(3). Theoretical results can be obtained from theorems 2.5.2 and 2.5.3. The four cases correspond with the options NOC, RWC, CLC, and MXC of the model MIX in the computer program POSCON, see section 5.2.

The aim of this simulation study is certainly not to a give a comprehensive review of the goodness of the approximations based on theorems 2.5.2 and 2.5.3. Such a study would require simulation results for many inputs, because the number of parameters in the case of both continuous and discrete variables is very large. In fact, we shall present only results for one very special parameter point. It is obvious that the conclusions to be arrived at can be very misleading if they are extrapolated to other situations.

We restricted our attention to two populations ($k = 2$) and four variables, namely two discrete and two continuous ones ($p = 2$). The two discrete variables had two categories each. The probabilities for the combined categories ($d = 4$) are denoted by P_{tl}, $t = 1,2$; $l = 1,...,4$ and can be found in Table 5.5. The two continuous variables follow conditional on a combined category, a multivariate normal distribution; parameters are μ_{tl}, Σ_{tl}, $t = 1,2$; $l = 1,...,4$; μ_{tl} can be read from Table 5.5.

j	population 1		population 2	
	p_{1j}	μ_{1j}	p_{2j}	μ_{2j}
1	0.2	(0,0)	0.3	(0.25, 0.25)
2	0.2	(0,3)	0.3	(0.50, 3.50)
3	0.3	(3.0)	0.2	(3.75, 0.75)
4	0.3	(3,3)	0.2	(4.00, 4.00)

TABLE 5.5 Parameters of the subpopulations.

Various sizes of the training samples were used, namely, $n_1 = n_2 = 50$, 100, 200, 400, and 800. For each of these cases the following 4 points were done 40 times: (1) a training set was drawn from the two populations, (2) 100 vectors of scores were drawn from population 1, (3) for each of the vectors of scores the theoretical and estimated posterior probability of belonging to population 1 and the standard derivation were computed, and, whether or not the theoretical posterior probability was situated in the 95% confidence interval was registered, for each of the four models, (4) the frequency of confidence intervals with the theoretical posterior probability situated in it was computed, being some number between 0 and 100, for each of the four models. The means with their standard errors of these 40 numbers are presented in Table 5.6 for the four models considered. Note that NOC means that no constraints are imposed, CLC means that $\Sigma_{11}=\Sigma_{21}$, $\Sigma_{12}=\Sigma_{22}$, $\Sigma_{13}=\Sigma_{23}$ and $\Sigma_{14}=\Sigma_{24}$, RWC stands for $\Sigma_{11}=\cdots=\Sigma_{14}$ and $\Sigma_{12}=\cdots=\Sigma_{24}$, and MXC means $\Sigma_{11}=\cdots=\Sigma_{14}=\Sigma_{21}=\cdots=\Sigma_{24}$. Hence, for the options NOC, CLC, RWC, and MXC the observations from 1, 2, 4, and 8 subpopulations, respectively, are used for the estimation of the covariance matrices.

Note that another way of simulation would have been generating vectors of scores and then for each of them generating a large number of training samples. This would have given comparable figures for the confidence levels, but would have required much more computer time.

The figures in Table 5.6 display that the nominal confidence level of 95% is not attained in most situations. The intervals are on the average, a bit too small. This shortcoming tends to decrease if the sample sizes are increased and

$n_1=n_2$	NOC		CLC		RWC		MXC	
50	77.7	1.7	86.3	1.5	86.7	1.7	89.3	1.2
100	84.4	1.8	89.4	1.7	89.8	1.8	90.1	2.0
200	91.3	1.2	93.0	1.3	92.1	1.2	93.3	1.2
400	92.6	1.0	93.7	1.1	93.6	1.1	93.5	1.1
800	94.5	0.7	95.7	0.8	94.3	0.9	94.6	0.9

TABLE 5.6 Confidence levels with standard errors.

theoretical posterior probability	frequency	percentiles					mean	st.dev.
		0	25	50	75	100		
0.0-0.1	9	-0.050	-0.048	-0.013	0.054	0.136	0.013	0.067
0.1-0.2	53	-0.109	-0.050	-0.023	0.022	0.198	-0.003	0.066
0.2-0.3	220	-0.156	-0.064	-0.017	0.040	0.364	-0.007	0.077
0.3-0.4	582	-0.236	-0.045	-0.002	0.045	0.334	0.002	0.076
0.4-0.5	678	-0.243	-0.054	-0.001	0.054	0.298	0.001	0.078
0.5-0.6	553	-0.291	-0.044	0.007	0.068	0.305	0.007	0.088
0.6-0.7	430	-0.429	-0.033	0.016	0.068	0.288	0.013	0.086
0.7-0.8	474	-0.335	-0.049	0.007	0.053	0.235	0.002	0.079
0.8-0.9	514	-0.339	-0.036	0.008	0.040	0.124	-0.001	0.061
0.9-1.0	487	-0.686	-0.022	0.000	0.017	0.073	-0.008	0.048

TABLE 5.7 Percentiles, means and standard deviations of distributions of differences between estimated and theoretical posterior probabilities for option RWC with $n_1 = n_2 = 200$.

the number of parameters is decreased, i.e., if NOC is replaced by CLC or RWC and if CLC or RWC is replaced by MXC. That the results for NOC with $n_1 = n_2 = 50$ are rather poor should be expected because 8 different 2×2 conditional covariance matrices are estimated, each one from about 12 observations.

We may expect that goodness of approximate confidence intervals depends on the theoretical posterior probabilities themselves. Therefore, we shall study as function of the theoretical posterior probabilities, first, the differences between the estimated and theoretical posterior probabilities, secondly, the standard deviations of the estimated posterior probabilities, and, thirdly, the confidence levels of the confidence intervals.

The differences between the estimated and theoretical posterior probabilities as function of the theoretical posterior probabilities are summarized in Table 5.7 for the option RWC with $n_1 = n_2 = 200$. The $40 \times 100 = 4000$ theoretical posterior probabilities are grouped into 10 groups. The second column gives the respective frequencies. For each group the distribution of the just mentioned differences is studied. Percentiles, means and standard deviations of these distributions are given. The distributions center reasonably well around zero. The standard deviations are largest for the values near 0.5 and smallest for the values near 0 and 1 of the theoretical posterior probability. A table like Table 5.7 was obtained for each of the 4 (model options) $\times$ 5 (sample sizes) $= 20$ situations studied. The minimum and maximum of the standard deviations of the 10 distributions of differences involved are presented in Table 5.8 for each of the 20 situations. In fact, because of some outliers in the groups 0.0-0.1 and 0.1-0.2, the minimums and maximums were chosen from the groups corresponding with the interval 0.2-1.0. We see that the standard deviations become smaller if the sample sizes are increased or if the model complexity (number of unknown parameters) is decreased.

$n_1=n_2$	NOC		CLC		RWC		MXC	
	min.	max.	min.	max.	min.	max.	min.	max.
50	0.154	0.224	0.092	0.170	0.126	0.163	0.079	0.147
100	0.087	0.176	0.046	0.129	0.054	0.125	0.043	0.117
200	0.063	0.110	0.038	0.088	0.048	0.088	0.032	0.080
400	0.048	0.085	0.025	0.067	0.023	0.066	0.021	0.062
800	0.023	0.052	0.016	0.041	0.016	0.043	0.015	0.040

TABLE 5.8 Minimum and maximum of standard deviations of distributions of differences between estimated and theoretical posterior probabilities.

The standard deviation of the estimated posterior probabilities as function of the theoretical posterior probabilities are given in Table 5.9 for the option RWC with $n_1=n_2=200$. Percentiles, means and standard deviations of the 10 distributions of standard deviations of posterior probabilities are displayed.

theoretical posterior probability	percentiles					mean	st.dev.	confidence level
	0	25	50	75	100			
0.0-0.1	0.016	0.020	0.028	0.057	0.112	0.046	0.033	77.8
0.1-0.2	0.030	0.051	0.059	0.077	0.190	0.068	0.028	94.3
0.2-0.3	0.046	0.059	0.070	0.085	0.199	0.077	0.026	95.5
0.3-0.4	0.044	0.056	0.066	0.082	0.226	0.074	0.026	93.3
0.4-0.5	0.046	0.059	0.066	0.081	0.276	0.075	0.025	95.6
0.5-0.6	0.050	0.064	0.076	0.094	0.203	0.083	0.027	93.3
0.6-0.7	0.044	0.058	0.068	0.087	0.355	0.077	0.030	91.6
0.7-0.8	0.036	0.052	0.059	0.075	0.306	0.069	0.029	90.7
0.8-0.9	0.018	0.040	0.049	0.058	0.229	0.053	0.023	90.1
0.9-1.0	0.001	0.019	0.031	0.043	0.305	0.034	0.026	87.3

TABLE 5.9 Percentiles, means and standard deviations of distributions of standard deviations of posterior probabilities and confidence levels for option RWC with $n_1=n_2=200$.

The last column gives the confidence level of the confidence intervals. Note the trend in the figures of percentile 0, 25, 50, 75 and the mean. The influence of $\rho_{1|x}(1-\rho_{1|x})$ can be clearly detected, see formula 2.1.3 and 2.1.4. The means and 50-th percentiles can be compared with the standard deviations in Table 5.7. They agree reasonably well.

A comparison of the distributions of the standard deviations of the estimated posterior probabilities for the 20 situations can be made by means of Table 5.10 and Table 5.11. The minimum and maximum of the means of the 10 distributions are given in Table 5.10. We see that the larger the sample sizes

$n_1=n_2$	NOC		CLC		RWC		MXC	
	min.	max.	min.	max.	min.	max.	min.	max.
50	0.126	0.220	0.074	0.162	0.096	0.160	0.068	0.153
100	0.080	0.145	0.049	0.121	0.054	0.124	0.044	0.166
200	0.052	0.101	0.033	0.082	0.034	0.083	0.029	0.078
400	0.035	0.074	0.024	0.060	0.022	0.061	0.020	0.057
800	0.023	0.053	0.016	0.042	0.015	0.044	0.014	0.041

TABLE 5.10 Minimum and maximum of means of distributions of standard deviations of posterior probabilities.

are, the smaller the means are. Further, the means become smaller going from left to right in the table. In Table 5.11 the means of the standard deviations of the 10 distributions are displayed. Largest variation in the values for the standard deviations of the estimated posterior probabilities appear for small sample sizes. The variation becomes smaller if the model complexity decreases.

$n_1=n_2$	NOC	CLC	RWC	M×C
50	0.135	0.064	0.075	0.053
100	0.079	0.037	0.049	0.033
200	0.044	0.023	0.027	0.020
400	0.030	0.015	0.020	0.014
800	0.020	0.010	0.013	0.009

TABLE 5.11 Means of the standard deviations of distributions of standard deviations of posterior probabilities.

As said before, the last column of Table 5.9 displays the confidence level of the confidence intervals for the posterior probabilities as function of the theoretical posterior probabilities for the option RWC with $n_1=n_2=200$. The figures of the groups 0.0-0.1 and 0.1-0.2 should not be taken too seriously, because of small numbers of generated values in these groups,

$n_1=n_2$	NOC		CLC		RWC		MXC	
	min.	max.	min.	max.	min.	max.	min.	max.
50	71.3	82.4	81.8	91.7	80.4	90.6	83.9	93.9
100	74.4	90.9	85.6	93.2	83.5	94.0	83.7	94.4
200	84.2	94.2	89.9	96.4	87.3	95.6	89.1	96.8
400	86.8	95.2	91.5	95.9	91.1	95.0	90.6	96.5
800	92.5	96.3	93.2	97.7	90.2	97.0	92.5	96.9

TABLE 5.12 Minimum and maximum of confidence levels.

see Table 5.7. Maximal figures for the confidence levels are attained for theoretical posterior probabilities near 0.5, whereas for larger values of

theoretical posterior probabilities the figures become smaller. This trend could also be seen for most of the other 19 situations. We had few generated values for the first groups for all situations. Deviations from the nominal confidence level were large for these groups. In table 5.12 the minimum and maximum of the confidence levels of the groups corresponding to the interval 0.2-1.0 are displayed. The trend is the larger the sample sizes the larger the minimums and maximums. For $n_1=n_2=50$ and $n_1=n_2=100$ the levels are too small for all four models. For the other three sample sizes, for all four models, except for NOC with $n_1=n_2=200$, the nominal level of 95% is situated between the minimum and maximum. The models CLC, RWC and MXC appear to give better 95% confidence intervals than model NOC.

REFERENCES

AITCHISON, J., J.D.F. HABBEMA and J.W. KAY (1977). *A critical comparison of two methods of statistical discrimination.* Applied Statistics, 26, 15-25.

AITCHISON, J. and J.W. KAY (1975). *Principles, practice and performance in decision making in clinical medicine.* In: Proceedings of the 1973 NATO Conference on the Role and Effectiveness of Decision in Practice. D.J. White and K.C. Bowen (eds.). Hodder and Stoughton, London.

AMBERGEN, A.W. (1981). *Approximate confidence intervals for posterior probabilities.* Report TW-224, Dept. of Math., University of Groningen, The Netherlands.

AMBERGEN, A.W. (1984). *Asymptotic distributions of estimators for posterior probabilities in a classification model with both continuous and discrete variables.* Report MS-R8409, Centre for Mathematics and Computer Science, Amsterdam.

AMBERGEN, A.W. and W. SCHAAFSMA (1982). *The asymptotic variance of estimators for posterior probabilities.* Report SW 86/82, Centre for Mathematics and Computer Science, Amsterdam.

AMBERGEN, A.W. and W. SCHAAFSMA (1983). *Interval estimates for posterior probabilities in a multivariate normal classification model.* Report SW 96/83, Centre for Mathematics and Computer Science, Amsterdam.

AMBERGEN, A.W. and W. SCHAAFSMA (1984). *Interval estimates for posterior probabilities, applications to Border Cave.* In: Multivariate Statistical Methods in Physical Anthropology. G.N. van Vark and W.W. Howells (eds.), 115-134. D. Reidel, Dordrecht, The Netherlands.

AMBERGEN, A.W. and W. SCHAAFSMA (1985). *Interval estimates for posterior probabilities in a multivariate normal classification model.* Journal of Multivariate Analysis, vol. 16, no. 3, 432-439.

ANDERSON, J.A. (1972). *Separate sample logistic discrimination.* Biometrika, 59, 19-35.

ANDERSON, J.A. (1973). *Logistic discrimination with medical applications.* In: Discriminant Analysis and Applications. Proceedings of a NATO Advanced Study Institute of Discriminant Analysis and Applications, Athens, 1972. T. Cacoullos (ed.), 1-15. Academic Press, New York.

ANDERSON, T.W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

BEAN, S.J. and C.P. TSOKOS (1980). *Developments in nonparametric density estimation.* Int. Stat. Review, 48, 267-287.

BEAUMONT, P.B. and A. BOSHIER (1972). *Some comments on recent findings at Border Cave, Northern Natal.* S. Afr. J. Sci., 68, 22-24.

BEAUMONT, P.B., H. DE VILLIERS and J.C. VOGEL (1978). *Modern man in sub-Saharan Africa prior to 49000 years B.P.: A review and evaluation with particular reference to Border Cave.* S. Afr. J. Sci., 74, 409-419.

BECHHOFER, R.E., J. KIEFER and M. SOBEL (1968). *Sequential identification and ranking procedures.* The University of Chicago Press, Chicago.

BISHOP, Y.M.M., S.E. FIENBERG and P.W. HOLLAND (1975). *Discrete multivariate analysis: Theory and practice.* The MIT Press, Cambridge.

BLACKWELL, D. and M.A. GIRSHICK (1954). *Theory of games and statistical decisions.* Wiley, New York.

BROTHWELL, D.R. (1963). *Evidence of early population change in central and southern Africa: Doubts and problems.* Man, 63, 101-104.

CACOULLOS, T. (1966). *Estimation of a multivariate density.* Ann. of the Inst. of Stat. Math., 18, 178-189.

CAMPBELL, N.A. (1980). *On the study of the Border Cave remains: Statistical Comments.* Current Anthropology, vol. 21, no. 4, 532-535.

CAMPBELL, N.A. (1984). *Some aspects of allocation and discrimination.* In: Multivariate Statistical Methods in Physical Anthropology. G.N. van Vark and W.W. Howells (eds.), 177-192. D. Reidel, Dordrecht, The Netherlands.

CHUNG, K.L. (1968). *A course in probability theory.* Harcourt, Brace & World, New York.

CONWAY, J.B. (1973). *Functions of one complex variable.* Springer-Verlag, New York.

COOK, H.B.S., B.D. MALAN and L.H. WELLS (1945). *Fossil man in the Lebombo mountains, South Africa: the 'Border Cave' Ingwavuma District, Zululand.* Man, 54, 6-13.

CORNISH, E.A. (1961). Simultaneous fiducial distribution of location parameters. Tech. Paper, no. 8, CSIRO, Melbourne, Australia.

COX, D.R. (1966). *Some procedures associated with the logistic qualitative response curve.* In: Research Papers in Statistics: Festschrift for J. Neyman. F.N. David (ed.), 55-71. Wiley, New York.

COX, D.R. (1970). *The analysis of binary data.* Methuen, London.

CRAMÉR, H. (1946). *Mathematical methods of statistics.* Princeton University Press, Princeton.

CRITCHLEY, F. and I. FORD (1984a). *On the covariance of two non-central F random variables and the variance of the estimated linear discriminant function.* Biometrika, 71, 3, 637-638.

CRITCHLEY, F. and I. FORD (1984b). *Interval estimation of the log-odds ratio in discrimination: multivariate normal, equal covariances.* Report, University of Warwick, UK.

CRITCHLEY, F. and I. FORD (1985). *Interval estimation in discrimination: the multivariate normal equal covariance case.* Biometrika, 72, 109-116.

CRITCHLEY, F., I. FORD and D. HIRST (1987). *On the possible linearity of the profile log-likelihood function: strong Lagrangian theory, with applications to linear discrimination.* Submitted for publication.

CRITCHLEY, F., I. FORD and D. HIRST (1988). *An evaluation of methods of interval estimation for the odds ratio in discrimination.* To appear in the Proceedings of the Fifth International Symposium on Data Analysis and Informatics, Sept. 29 - Oct. 2, 1987, Versailles, France. North-Holland, Amsterdam.

CRITCHLEY, F., I. FORD and O. RIJAL (1987). *Uncertainty in discrimination.* In: Proceedings of the Conference DIANA II held in Liblice, 1986,

Mathematical Institute of the Czechoslovak Academy of Science, Prague, 83-106.

CRITCHLEY, F., I. FORD and O. RIJAL (1988). *Interval estimation based on the profile likelihood: strong Lagrangian theory, with applications to discrimination.* Biometrika, 75, 21-28.

DARROCH, J.N., S.L. LAURITZEN and T.P. SPEED (1980). *Markov fields and log-linear interaction models for contingency tables.* Ann. Statist., 8, 522-539.

DAS GUPTA, S. (1968). *Some aspects of discrimination function coefficients.* Sankhya: The Indian Journal of Statistics, Series A, vol. 30, part 4, 387-400.

DAY, N.E. and D.F. KERRIDGE (1967). *A general maximum likelihood discriminant.* Biometrics, 23, 313-323.

DE VILLIERS, H. (1973). *Human skeletal remains from Border Cave.* Ingwavuma District, Kwazulu, South Africa. Ann. Transv. Mus., 28, 13, 229-256.

DE VILLIERS, H. and L.P. FATTI (1982). *The antiquity of the negro.* South African Journal of Science, 78, 321-332.

DEGROOT, M.H. (1970). *Optimal statistical decisions.* McGraw-Hill, New York.

DVORETZKY, A., A. WALD and J. WOLFOWITZ (1951). *Elimination of randomization in certain statistical decision procedures and zero-sum two-person games.* Ann. Math. Statist., 22, 1-21.

EATON, M.L. (1983). *Multivariate Statistics, a vector space approach.* Wiley, New York.

EATON, M.L. and C.N. MORRIS (1970). *The application of invariance to unbiased estimation.* The Ann. of Math. Stat., 1970, vol. 41, no. 5, 1708-1716.

EDWARDS, D. (1986). *Hierarchical mixed interaction models.* Research report 86/6. Statistical Research Unit, University of Copenhagen, Copenhagen.

EFRON, B. (1977). *The 1977 Rietz Lecture. Bootstrap methods: another look at the Jackknife.* The Annals of Statistics, 1979, vol. 7, no. 1, 1-26.

EFRON, B. (1981). *Nonparametric standard errors and confidence intervals.* The Canadian Journal of Statistics, vol. 9, no. 2, 139-172.

EFRON, B. (1982). *The Jackknife, the bootstrap and other resampling plans.* CBMS-NSF, Regional conference series in applied mathematics, 38. Society for Industrial and Applied Mathematics. Philadelphia, Pennsylvania.

ERDÉLYI, A., W. MAGNUS, F. OBERHETTINGER and F.G. TRICOMI (1953). *Higher transcendental functions.* Vol. I. McGraw-Hill, New York.

FATTI, L.P. (1985). *Discriminant analysis in prehistoric physical anthropology.* Technical Report TWISK 386, National Research Institute for Mathematical Sciences, Pretoria, South Africa.

FERGUSON, T.S. (1967). *Mathematical Statistics. A decision theoretic approach.* Academic Press, New York.

FISHER, R.A. (1935). *The fiducial argument in statistical inference.* Ann. Eugen., 6, 391-398.

FISHER, R.A. (1954). *Discussion on the symposium on interval estimation.* J.R. Statist., Soc. B, 16, 212-213.

GANESALINGAM, S. and G.L. MCLACHLAN (1979). *A case study of two clustering*

methods based on maximum likelihood. Statistica Neerlandica, Vol. 33, No. 2, 81-90.

GEISSER, S. (1964). *Posterior odds for multivariate normal classifications.* J. Roy. Statist. Soc., Ser. B, 26, 69-76.

GEISSER, S. (1965). *Bayesian estimation in multivariate analysis.* Ann. Math. Statist., 36, 150-159.

GEISSER, S. (1966). *Predictive discrimination.* In: Multivariate Analysis. P. Krishnaiah (ed.), 149-163. Academic Press, New York.

GEISSER, S. (1967). *Estimation associated with linear discriminants.* Ann. Math. Statist., 38, 807-817.

GEISSER, S. (1970). *Discriminatory practices.* In: Bayesian Statistics. D. Meyer and R.C. Collier (eds.), 57-70. Peacock, Illinois.

GEISSER, S. (1977). *Discrimination, allocatory and separatory, linear aspects.* In: Classification and Clustering. J. van Ryzin (ed.), 301-330. Academic Press, New York.

GEISSER, S. (1980). *Sample reuse selection and allocation criteria.* In: Multivariate Analysis V. P.R. Krishnaiah (ed.), 387-398. North-Holland, Amsterdam.

GEISSER, S. (1982a). *Aspects of the predictive and estimative approaches in the determination of probabilities.* Biometrics supplement; Current Topics in Biostatistics and Epidemiology, 75-85.

GEISSER, S. (1982b). *Bayesian discrimination.* In: Handbook of Statistics, Vol. 2. P.R. Krishnaiah and L.N. Kanal (eds.), 101-120. North-Holland, Amsterdam.

GEISSER, S. and J. CORNFIELD (1963). *Posterior distributions for multivariate normal parameters.* J. Roy. Statist. Soc., Ser. B, 25, 368-376.

GHURYE, S.G. and I. OLKIN, (1969). *Unbiased estimation of some multivariate probability densities and related functions.* The Ann. of Math. Stat., 1969, vol. 40, no. 4, 1261-1271.

GIBBONS, J.D., I. OLKIN and M. SOBEL (1977). *Selecting and ordering populations. A new statistical methodology.* Wiley, New York.

GILL, R.D. (1985). *Discussion of papers by S. Lauritzen and N. Wermuth on mixed interaction models for mixed continuous and discrete multivariate data.* Bull. Int. Statist. Inst., 51(4), 24.5.1-24.5.2.

GUPTA, S.S. and S. PANCHAPAKESAN (1979). *Multiple decision procedures. Theory and methodology of selecting and ranking populations.* Wiley, New York.

HABBEMA, J.D.F., and J. HERMANS (1978). *Statistical methods for clinical decision making.* Thesis, Leiden University, The Netherlands.

HABBEMA, J.D.F., J. HERMANS and K. VAN DEN BROEK (1974). *A stepwise discriminant analysis program using density estimation.* In: COMPSTAT-1974, Proceedings in Computational Statistics. G. Bruckmann, F. Ferschl, and L. Schmetterer (eds.), 101-110. Physica-Verlag, Vienna.

HABBEMA, J.D.F, J. HILDEN and B. BJERREGAARD (1978). *The measurement of performance in probabilistic diagnosis I, II, III.* Meth.. Inform. Med., 17, 217-246.

HABBEMA, J.D.F, J. HILDEN and B. BJERREGAARD (1981). *The measurement of*

performance in probabilistic diagnosis VI, V. Meth. Inform. Med., 20, 80-100.

HERMANS, J. and J.D.F. HABBEMA (1975). *Comparison of five methods to estimate posterior probabilities.* EDV in Medizin und Biologie, 6, 14-19.

HERMANS, J., B. VAN ZOMEREN, J.W. RAATGEVER, P.J. STERK and J.D.F. HABBEMA (1981). *Use of posterior probabilities to evaluate methods of discriminant analysis.* Meth. Inform. Med., 20, 207-212.

HOWELLS, W.W. (1973). *Cranial variation in man. A study by multivariate analysis of patterns of difference among recent human populations.* Papers of the Peabody Museum, 67.

KOLMOGOROV, A.N. (1950). *Unbiased estimators.* Izv. Akad. Nauk SSSR. Ser. Mat. 14, 303-326. English translation in Amer. Math. Soc. Transl., 98.

KRUSIŃSKA, E. (1981). *Logistic discrimination for two groups of data.* Report N-102, Institute of Computer Science, Wroclaw University, Worclaw, Poland.

KRUSIŃSKA, E. (1982). *Logistic discrimination for several groups of data.* Report N-110, Institute of Computer Science, Wroclaw University, Wroclaw, Poland.

KRUSIŃSKA, E. (1984). *Linguistic variables and their application to discrimination.* Report N-133, Institute of Computer Science, Wroclaw University, Wroclaw, Poland.

LACHENBRUCH, P.A. (1975). *Discriminant Analysis.* Hafner, New York.

LAURITZEN, S.L. and N. WERMUTH (1984). *Mixed interaction models.* Report R84-8, Institute of Electronic Systems, Aalborg University Centre, Denmark.

LEHMANN, E.L. (1950). *Notes on the theory of estimation.* University of California, Berkely.

LEHMANN, E.L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

MAGNUS, J.R. and H. NEUDECKER (1979). *The commutation matrix: some properties and applications.* The Annals of Statistics, Vol. 7, no. 2, 381-394.

MAGNUS, J.R. and H. NEUDECKER (1980). *The elimination matrix: some lemmas and applications.* SIAM, J. Alg. Disc. Meth., vol. 1, no. 4.

MAGNUS, J.R. and H. NEUDECKER (1985). *Matrix Differential Calculus with Applications to Simple, Hadamard, and Kronecker Products.* J. of Mathematical Psychology, vol. 29, no. 4.

MAGNUS, J.R. and H. NEUDECKER (1986). *Symmetry, 0-1 Matrices and Jacobians, A Review.* Econometric Theory, 2, 157-190.

MAGNUS, J.R. and H. NEUDECKER (1988). *Matrix differential calculus with applications in statistics and econometrics.* Wiley, New York.

MAGNUS, W., F. OBERHETTINGER and R.P. SONI (1966). *Formulas and theorems for the Special Functions of Mathematical Physics.* Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen. Band 52, third-edition. Springer-Verlag, New York.

MCLACHLAN, G.J. (1977). *The bias of sample based posterior probabilities.* Biom. J., vol. 19, no. 6, 421-426.

MCLACHLAN, G.J. (1979). *A comparison of the estimative and predictive methods of estimating posterior probabilities.* Comm. Statist., Theor. Meth., A8(9), 919-929.

MORAN, M.A. and B.J. MURPHY (1979). *A closer look at two alternative methods of statistical discrimination.* Applied Statistics, 28, 3, 223-232.

MUIRHEAD, R.J. (1982). *Aspects of Multivariate Statistical Theory.* Wiley, New York.

PAPOULIS, A. (1962). *The Fourier integral and its applications.* McGraw-Hill, New York.

PARZEN, E. (1962). *On estimation of a probability density function and mode.* Am. Math. Stat., 33, 1065-1076.

RAO, C.R. (1965). *Linear statistical inference and its applications.* Wiley, New York.

RAO, C.R. (1973). *Linear statistical inference and its applications.* Second edition. Wiley, New York.

REMME, J., J.D.F. HABBEMA and J. HERMANS (1980). *A simulative comparison of linear, quadratic and kernel discrimination.* J. Statist. Comput. Simul., vol. 11, 87-106.

RIGBY, R.A. (1982). *A credibility interval for the probability that a new observation belongs to one of two multivariate normal populations.* Journal of Royal Statistical Society, Ser. B, vol. 44, no. 2, 2122-220.

RIGHTMIRE, G.P. (1979). *Implications of Border Cave skeletal remains for Later Pleistocene human evolution.* Current Anthropology, vol. 20, no. 1, 23-35.

RIGHTMIRE, G.P. (1981). *More on the study of the Border Cave remains.* Current Anthropology, vol. 22, no. 2, 199-200.

RIJAL, O.M. (1984). *Topics in statistical discrimination.* Dissertation. Department of Statistics, University of Glasgow, UK.

ROGERS, G.S. (1980). *Matrix derivations.* Lexture notes in statistics. Vol. 2. Marcel Dekker, New York.

ROSENBLATT, M. (1956). *Remarks on some nonparametric estimates of a density function.* Ann. Math. Stat., 27, 832-837.

RUDIN, W. (1964). *Principles of mathematical analysis.* 2-th edition. McGraw-Hill, New York.

SAITTA, L. and P. TORASSO (1981). *Fuzzy characterization of coronary disease.* Fuzzy Sets and Systems 5, 245-258.

SCHAAFSMA, W. (1973). *Classifying when populations are estimated.* In: Discriminant Analysis and Applications. Proceedings of a NATO Advanced Study Institute of Discriminant Analysis and Applications, Athens, 1972. T. Cacoullos (ed.), 339-364. Academic Press, New York.

SCHAAFSMA, W. (1976). *The asymptotic distribution of some statistics from discriminant analysis.* Report TW-176, Dept. of Math., University of Groningen, The Netherlands.

SCHAAFSMA, W. (1982). *Selecting variables in discriminant analysis for improving upon classical procedures.* In: Handbook of Statistics, vol. 2. P.R. Krishnaiah and L.N. Kanal (eds.), 857-881. North-Holland, Amsterdam.

SCHAAFSMA, W. (1983). *Some aspects of discriminant analysis.* In: Proceedings Conference DIANA, Liblice, 1982, Czechoslovakia, 1-31.

SCHAAFSMA, W. (1985a). *Me and the anthropologist.* In: Proceedings of the Seventh Conference on Probability Theory. Aug. 29 - Sept. 4, 1982, Brasov,

Romania. M. Iosifescu (ed.), 333-343. Editura Academiei, Bucuresti, and VNU Science Press, Utrecht.

SCHAAFSMA, W. (1985b). *Standard errors of posterior probabilities and how to use them.* In: Multivariate Analysis VI. P.R. Krishnaiah (ed.), 527-548. North-Holland, Amsterdam.

SCHAAFSMA, W. and A.W. AMBERGEN (1987). *POSCON- a probalistic approach to medical expert systems.* In: Medical expertsystems using personal computers. M.K. CHYTIL and R. ENGELBRECHT (eds.), 175-186. Sigma Press, U.K.

SCHAAFSMA, W. and G.N. VAN VARK (1977). *Classification and discrimination problems with applications, part I.* Statistica Neerlandica, vol. 31, no. 1, 25-45.

SCHAAFSMA, W. and G.N. VAN VARK (1979). *Classification and discrimination problems with applications, part II.* Statistica Neerlandica, vol. 33, no. 2, 91-126.

SERFLING, R.J. (1980). *Approximation theorems of mathematical statistics.* Wiley, New York.

SRIVASTAVA, M.S. and C.G. KHATRI (1979). *An introduction to multivariate statistics.* North-Holland, Amsterdam.

STEERNEMAN, A.G.M. (1987). *On the choice of variables in discriminant analysis and regression analysis.* Thesis. University of Groningen, The Netherlands.

VAN DER SLUIS, D.M. and W. SCHAAFSMA (1984). *POSCON, a decision-support system in diagnosis and prognosis based on a statistical approach.* In: COMPSTAT-1984. T. Havranek et al. (eds.), 160-165. Physica-Verlag, Vienna.

VAN DER SLUIS, D.M., W. SCHAAFSMA and A.W. AMBERGEN (1985). *POSCON, user manual. A decision support system in diagnosis and prognosis.* University of Groningen, The Netherlands.

VAN DER SLUIS, D.M., W. SCHAAFSMA and A.W. AMBERGEN (1986). *POSCON, user manual. A decision support system in diagnosis and prognosis.* University of Groningen, The Netherlands.

VAN NESS, J. (1979). *On the effects of dimension in discriminant analysis for unequal covariance populations.* Technometrics, vol. 21, no. 1, 119-127.

VAN NESS, J.W. and C. SIMPSON (1976). *On the effects of dimension in discriminant analysis.* Technometrics, vol. 18, no. 2, 175-187.

VAN VARK, G.N. (1970). *Some statistical procedures for the investigation of prehistoric human skeletal material.* Thesis, University of Groningen, The Netherlands.

VAN VARK, G.N. (1976). *A critical evaluation of the application of multivariate statistical methods to the study of human populations from their skeletal remains.* Homo, 27, 2, 94-114.

VAN VARK, G.N. and P.G.M. VAN DER SMAN (1982). *New discrimination and classification techniques in anthropological practice.* Zeitschrift für Morphologie und Anthropologie, vol. 73, no. 1, 21-36.

VON NEUMANN, J. and O. MORGENSTERN (1944). *Theory of games and economic behavior.* 3rd-edition. Princeton University Press, Princeton.

WALD, A. (1950). *Statistical Decision Functions.* Wiley, New York.

WATSON, G.S. (1964). *A note on maximum likelihood.* Sankhya, 26A, 303-304.
WELLS, L.H. (1950). *The Border Cave skull.* Ingwavuma District, Natal. Am. J. Phys. Anthrop., 8, 241-243.
WELLS, L.H. (1969). *Homo sapiens afer Linn.: Content and earliest representatives.* South African Archeological Bulletin, 24, 172-173.
WELLS, L.H. (1972). *Late Stone Age and Middle Stone Age toolmakers.* South African Archeological Bulletin, 27, 5-9.
WERMUTH, N. and S.L. LAURITZEN (1987). *Conditional independence graphs, graphical chain models, and data.* In: Berichte zur Stochastik und verwandten Gebieten, 87-2, Johannes Gutenberg-Universität, Mainz, Federal Republic of Germany.
WITTING, H. and G. NÖLLE (1970). *Angewandte Mathematische Statistik.* Optimale finite und asymptotische Verfahren. B.G. Teubner, Stuttgart.
ZADEH, L.A. (1965). *Fuzzy sets.* Information and Control 8, 338-353.
ZADEH, L.A. (1975). *The concept of a linguistic variable and its application to approximate reasoning.* Information Sci., 8, 199-250, 301-358.
ZADEH, L.A. (1976). *The concept of a linguistic variable and its application to approximate reasoning.* Information Sci., 9, 43-80.

Author Index

Subject Index

MATHEMATICAL CENTRE TRACTS

1 T. van der Walt. *Fixed and almost fixed points.* 1963.

2 A.R. Bloemena. *Sampling from a graph.* 1964.

3 G. de Leve. *Generalized Markovian decision processes, part I: model and method.* 1964.

4 G. de Leve. *Generalized Markovian decision processes, part II: probabilistic background.* 1964.

5 G. de Leve, H.C. Tijms, P.J. Weeda. *Generalized Markovian decision processes, applications.* 1970.

6 M.A. Maurice. *Compact ordered spaces.* 1964.

7 W.R. van Zwet. *Convex transformations of random variables.* 1964.

8 J.A. Zonneveld. *Automatic numerical integration.* 1964.

9 P.C. Baayen. *Universal morphisms.* 1964.

10 E.M. de Jager. *Applications of distributions in mathematical physics.* 1964.

11 A.B. Paalman-de Miranda. *Topological semigroups.* 1964.

12 J.A.Th.M. van Berckel, H. Brandt Corstius, R.J. Mokken, A. van Wijngaarden. *Formal properties of newspaper Dutch.* 1965.

13 H.A. Lauwerier. *Asymptotic expansions.* 1966, out of print; replaced by MCT 54.

14 H.A. Lauwerier. *Calculus of variations in mathematical physics.* 1966.

15 R. Doornbos. *Slippage tests.* 1966.

16 J.W. de Bakker. *Formal definition of programming languages with an application to the definition of ALGOL 60.* 1967.

17 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 1.* 1968.

18 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 2.* 1968.

19 J. van der Slot. *Some properties related to compactness.* 1968.

20 P.J. van der Houwen. *Finite difference methods for solving partial differential equations.* 1968.

21 E. Wattel. *The compactness operator in set theory and topology.* 1968.

22 T.J. Dekker. *ALGOL 60 procedures in numerical algebra, part 1.* 1968.

23 T.J. Dekker, W. Hoffmann. *ALGOL 60 procedures in numerical algebra, part 2.* 1968.

24 J.W. de Bakker. *Recursive procedures.* 1971.

25 E.R. Paërl. *Representations of the Lorentz group and projective geometry.* 1969.

26 European Meeting 1968. *Selected statistical papers, part I.* 1968.

27 European Meeting 1968. *Selected statistical papers, part II.* 1968.

28 J. Oosterhoff. *Combination of one-sided statistical tests.* 1969.

29 J. Verhoeff. *Error detecting decimal codes.* 1969.

30 H. Brandt Corstius. *Exercises in computational linguistics.* 1970.

31 W. Molenaar. *Approximations to the Poisson, binomial and hypergeometric distribution functions.* 1970.

32 L. de Haan. *On regular variation and its application to the weak convergence of sample extremes.* 1970.

33 F.W. Steutel. *Preservation of infinite divisibility under mixing and related topics.* 1970.

34 I. Juhász, A. Verbeek, N.S. Kroonenberg. *Cardinal functions in topology.* 1971.

35 M.H. van Emden. *An analysis of complexity.* 1971.

36 J. Grasman. *On the birth of boundary layers.* 1971.

37 J.W. de Bakker, G.A. Blaauw, A.J.W. Duijvestijn, E.W. Dijkstra, P.J. van der Houwen, G.A.M. Kamsteeg-Kemper, F.E.J. Kruseman Aretz, W.L. van der Poel, J.P. Schaap-Kruseman, M.V. Wilkes, G. Zoutendijk. *MC-25 Informatica Symposium.* 1971.

38 W.A. Verloren van Themaat. *Automatic analysis of Dutch compound words.* 1972.

39 H. Bavinck. *Jacobi series and approximation.* 1972.

40 H.C. Tijms. *Analysis of (s,S) inventory models.* 1972.

41 A. Verbeek. *Superextensions of topological spaces.* 1972.

42 W. Vervaat. *Success epochs in Bernoulli trials (with applications in number theory).* 1972.

43 F.H. Ruymgaart. *Asymptotic theory of rank tests for independence.* 1973.

44 H. Bart. *Meromorphic operator valued functions.* 1973.

45 A.A. Balkema. *Monotone transformations and limit laws.* 1973.

46 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 1: the language.* 1973.

47 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 2: the compiler.* 1973.

48 F.E.J. Kruseman Aretz, P.J.W. ten Hagen, H.L. Oudshoorn. *An ALGOL 60 compiler in ALGOL 60, text of the MC-compiler for the EL-X8.* 1973.

49 H. Kok. *Connected orderable spaces.* 1974.

50 A. van Wijngaarden, B.J. Mailloux, J.E.L. Peck, C.H.A. Koster, M. Sintzoff, C.H. Lindsey, L.G.L.T. Meertens, R.G. Fisker (eds.). *Revised report on the algorithmic language ALGOL 68.* 1976.

51 A. Hordijk. *Dynamic programming and Markov potential theory.* 1974.

52 P.C. Baayen (ed.). *Topological structures.* 1974.

53 M.J. Faber. *Metrizability in generalized ordered spaces.* 1974.

54 H.A. Lauwerier. *Asymptotic analysis, part 1.* 1974.

55 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 1: theory of designs, finite geometry and coding theory.* 1974.

56 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 2: graph theory, foundations, partitions and combinatorial geometry.* 1974.

57 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 3: combinatorial group theory.* 1974.

58 W. Albers. *Asymptotic expansions and the deficiency concept in statistics.* 1975.

59 J.L. Mijnheer. *Sample path properties of stable processes.* 1975.

60 F. Göbel. *Queueing models involving buffers.* 1975.

63 J.W. de Bakker (ed.). *Foundations of computer science.* 1975.

64 W.J. de Schipper. *Symmetric closed categories.* 1975.

65 J. de Vries. *Topological transformation groups, 1: a categorical approach.* 1975.

66 H.G.J. Pijls. *Logically convex algebras in spectral theory and eigenfunction expansions.* 1976.

68 P.P.N. de Groen. *Singularly perturbed differential operators of second order.* 1976.

69 J.K. Lenstra. *Sequencing by enumerative methods.* 1977.

70 W.P. de Roever, Jr. *Recursive program schemes: semantics and proof theory.* 1976.

71 J.A.E.E. van Nunen. *Contracting Markov decision processes.* 1976.

72 J.K.M. Jansen. *Simple periodic and non-periodic Lamé functions and their applications in the theory of conical waveguides.* 1977.

73 D.M.R. Leivant. *Absoluteness of intuitionistic logic.* 1979.

74 H.J.J. te Riele. *A theoretical and computational study of generalized aliquot sequences.* 1976.

75 A.E. Brouwer. *Treelike spaces and related connected topological spaces.* 1977.

76 M. Rem. *Associons and the closure statement.* 1976.

77 W.C.M. Kallenberg. *Asymptotic optimality of likelihood ratio tests in exponential families.* 1978.

78 E. de Jonge, A.C.M. van Rooij. *Introduction to Riesz spaces.* 1977.

79 M.C.A. van Zuijlen. *Emperical distributions and rank statistics.* 1977.

80 P.W. Hemker. *A numerical study of stiff two-point boundary problems.* 1977.

81 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 1.* 1976.

82 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 2.* 1976.

83 L.S. van Benthem Jutting. *Checking Landau's "Grundlagen" in the AUTOMATH system.* 1979.

84 H.L.L. Busard. *The translation of the elements of Euclid from the Arabic into Latin by Hermann of Carinthia (?), books vii-xii.* 1977.

85 J. van Mill. *Supercompactness and Wallman spaces.* 1977.

86 S.G. van der Meulen, M. Veldhorst. *Torrix I, a programming system for operations on vectors and matrices over arbitrary fields and of variable size.* 1978.

88 A. Schrijver. *Matroids and linking systems.* 1977.

89 J.W. de Roever. *Complex Fourier transformation and analytic functionals with unbounded carriers.* 1978.

90 L.P.J. Groenewegen. *Characterization of optimal strategies in dynamic games.* 1981.

91 J.M. Geysel. *Transcendence in fields of positive characteristic.* 1979.

92 P.J. Weeda. *Finite generalized Markov programming.* 1979.

93 H.C. Tijms, J. Wessels (eds.). *Markov decision theory.* 1977.

94 A. Bijlsma. *Simultaneous approximations in transcendental number theory.* 1978.

95 K.M. van Hee. *Bayesian control of Markov chains.* 1978.

96 P.M.B. Vitányi. *Lindenmayer systems: structure, languages, and growth functions.* 1980.

97 A. Federgruen. *Markovian control problems; functional equations and algorithms.* 1984.

98 R. Geel. *Singular perturbations of hyperbolic type.* 1978.

99 J.K. Lenstra, A.H.G. Rinnooy Kan, P. van Emde Boas (eds.). *Interfaces between computer science and operations research.* 1978.

100 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 1.* 1979.

101 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 2.* 1979.

102 D. van Dulst. *Reflexive and superreflexive Banach spaces.* 1978.

103 K. van Harn. *Classifying infinitely divisible distributions by functional equations.* 1978.

104 J.M. van Wouwe. *Go-spaces and generalizations of metrizability.* 1979.

105 R. Helmers. *Edgeworth expansions for linear combinations of order statistics.* 1982.

106 A. Schrijver (ed.). *Packing and covering in combinatorics.* 1979.

107 C. den Heijer. *The numerical solution of nonlinear operator equations by imbedding methods.* 1979.

108 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 1.* 1979.

109 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 2.* 1979.

110 J.C. van Vliet. *ALGOL 68 transput, part I: historical review and discussion of the implementation model.* 1979.

111 J.C. van Vliet. *ALGOL 68 transput, part II: an implementation model.* 1979.

112 H.C.P. Berbee. *Random walks with stationary increments and renewal theory.* 1979.

113 T.A.B. Snijders. *Asymptotic optimality theory for testing problems with restricted alternatives.* 1979.

114 A.J.E.M. Janssen. *Application of the Wigner distribution to harmonic analysis of generalized stochastic processes.* 1979.

115 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 1.* 1979.

116 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 2.* 1979.

117 P.J.M. Kallenberg. *Branching processes with continuous state space.* 1979.

118 P. Groeneboom. *Large deviations and asymptotic efficiencies.* 1980.

119 F.J. Peters. *Sparse matrices and substructures, with a novel implementation of finite element algorithms.* 1980.

120 W.P.M. de Ruyter. *On the asymptotic analysis of large-scale ocean circulation.* 1980.

121 W.H. Haemers. *Eigenvalue techniques in design and graph theory.* 1980.

122 J.C.P. Bus. *Numerical solution of systems of nonlinear equations.* 1980.

123 I. Yuhász. *Cardinal functions in topology - ten years later.* 1980.

124 R.D. Gill. *Censoring and stochastic integrals.* 1980.

125 R. Eising. *2-D systems, an algebraic approach.* 1980.

126 G. van der Hoek. *Reduction methods in nonlinear programming.* 1980.

127 J.W. Klop. *Combinatory reduction systems.* 1980.

128 A.J.J. Talman. *Variable dimension fixed point algorithms and triangulations.* 1980.

129 G. van der Laan. *Simplicial fixed point algorithms.* 1980.

130 P.J.W. ten Hagen, T. Hagen, P. Klint, H. Noot, H.J. Sint, A.H. Veen. *ILP: intermediate language for pictures.* 1980.

131 R.J.R. Back. *Correctness preserving program refinements: proof theory and applications.* 1980.

132 H.M. Mulder. *The interval function of a graph.* 1980.

133 C.A.J. Klaassen. *Statistical performance of location estimators.* 1981.

134 J.C. van Vliet, H. Wupper (eds.). *Proceedings international conference on ALGOL 68.* 1981.

135 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part I.* 1981.

136 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part II.* 1981.

137 J. Telgen. *Redundancy and linear programs.* 1981.

138 H.A. Lauwerier. *Mathematical models of epidemics.* 1981.

139 J. van der Wal. *Stochastic dynamic programming, successive approximations and nearly optimal strategies for Markov decision processes and Markov games.* 1981.

140 J.H. van Geldrop. *A mathematical theory of pure exchange economies without the no-critical-point hypothesis.* 1981.

141 G.E. Welters. *Abel-Jacobi isogenies for certain types of Fano threefolds.* 1981.

142 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 1.* 1981.

143 J.M. Schumacher. *Dynamic feedback in finite- and infinite-dimensional linear systems.* 1981.

144 P. Eijgenraam. *The solution of initial value problems using interval arithmetic; formulation and analysis of an algorithm.* 1981.

145 A.J. Brentjes. *Multi-dimensional continued fraction algorithms.* 1981.

146 C.V.M. van der Mee. *Semigroup and factorization methods in transport theory.* 1981.

147 H.H. Tigelaar. *Identification and informative sample size.* 1982.

148 L.C.M. Kallenberg. *Linear programming and finite Markovian control problems.* 1983.

149 C.B. Huijsmans, M.A. Kaashoek, W.A.J. Luxemburg, W.K. Vietsch (eds.). *From A to Z, proceedings of a symposium in honour of A.C. Zaanen.* 1982.

150 M. Veldhorst. *An analysis of sparse matrix storage schemes.* 1982.

151 R.J.M.M. Does. *Higher order asymptotics for simple linear rank statistics.* 1982.

152 G.F. van der Hoeven. *Projections of lawless sequences.* 1982.

153 J.P.C. Blanc. *Application of the theory of boundary value problems in the analysis of a queueing model with paired services.* 1982.

154 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part I.* 1982.

155 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part II.* 1982.

156 P.M.G. Apers. *Query processing and data allocation in distributed database systems.* 1983.

157 H.A.W.M. Kneppers. *The covariant classification of two-dimensional smooth commutative formal groups over an algebraically closed field of positive characteristic.* 1983.

158 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 1.* 1983.

159 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 2.* 1983.

160 A. Rezus. *Abstract AUTOMATH.* 1983.

161 G.F. Helminck. *Eisenstein series on the metaplectic group, an algebraic approach.* 1983.

162 J.J. Dik. *Tests for preference.* 1983.

163 H. Schippers. *Multiple grid methods for equations of the second kind with applications in fluid mechanics.* 1983.

164 F.A. van der Duyn Schouten. *Markov decision processes with continuous time parameter.* 1983.

165 P.C.T. van der Hoeven. *On point processes.* 1983.

166 H.B.M. Jonkers. *Abstraction, specification and implementation techniques, with an application to garbage collection.* 1983.

167 W.H.M. Zijm. *Nonnegative matrices in dynamic programming.* 1983.

168 J.H. Evertse. *Upper bounds for the numbers of solutions of diophantine equations.* 1983.

169 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 2.* 1983.

CWI TRACTS

1 D.H.J. Epema. *Surfaces with canonical hyperplane sections.* 1984.

2 J.J. Dijkstra. *Fake topological Hilbert spaces and characterizations of dimension in terms of negligibility.* 1984.

3 A.J. van der Schaft. *System theoretic descriptions of physical systems.* 1984.

4 J. Koene. *Minimal cost flow in processing networks, a primal approach.* 1984.

5 B. Hoogenboom. *Intertwining functions on compact Lie groups.* 1984.

6 A.P.W. Böhm. *Dataflow computation.* 1984.

7 A. Blokhuis. *Few-distance sets.* 1984.

8 M.H. van Hoorn. *Algorithms and approximations for queueing systems.* 1984.

9 C.P.J. Koymans. *Models of the lambda calculus.* 1984.

10 C.G. van der Laan, N.M. Temme. *Calculation of special functions: the gamma function, the exponential integrals and error-like functions.* 1984.

11 N.M. van Dijk. *Controlled Markov processes; time-discretization.* 1984.

12 W.H. Hundsdorfer. *The numerical solution of nonlinear stiff initial value problems: an analysis of one step methods.* 1985.

13 D. Grune. *On the design of ALEPH.* 1985.

14 J.G.F. Thiemann. *Analytic spaces and dynamic programming: a measure theoretic approach.* 1985.

15 F.J. van der Linden. *Euclidean rings with two infinite primes.* 1985.

16 R.J.P. Groothuizen. *Mixed elliptic-hyperbolic partial differential operators: a case-study in Fourier integral operators.* 1985.

17 H.M.M. ten Eikelder. *Symmetries for dynamical and Hamiltonian systems.* 1985.

18 A.D.M. Kester. *Some large deviation results in statistics.* 1985.

19 T.M.V. Janssen. *Foundations and applications of Montague grammar, part 1: Philosophy, framework, computer science.* 1986.

20 B.F. Schriever. *Order dependence.* 1986.

21 D.P. van der Vecht. *Inequalities for stopped Brownian motion.* 1986.

22 J.C.S.P. van der Woude. *Topological dynamix.* 1986.

23 A.F. Monna. *Methods, concepts and ideas in mathematics: aspects of an evolution.* 1986.

24 J.C.M. Baeten. *Filters and ultrafilters over definable subsets of admissible ordinals.* 1986.

25 A.W.J. Kolen. *Tree network and planar rectilinear location theory.* 1986.

26 A.H. Veen. *The misconstrued semicolon: Reconciling imperative languages and dataflow machines.* 1986.

27 A.J.M. van Engelen. *Homogeneous zero-dimensional absolute Borel sets.* 1986.

28 T.M.V. Janssen. *Foundations and applications of Montague grammar, part 2: Applications to natural language.* 1986.

29 H.L. Trentelman. *Almost invariant subspaces and high gain feedback.* 1986.

30 A.G. de Kok. *Production-inventory control models: approximations and algorithms.* 1987.

31 E.E.M. van Berkum. *Optimal paired comparison designs for factorial experiments.* 1987.

32 J.H.J. Einmahl. *Multivariate empirical processes.* 1987.

33 O.J. Vrieze. *Stochastic games with finite state and action spaces.* 1987.

34 P.H.M. Kersten. *Infinitesimal symmetries: a computational approach.* 1987.

35 M.L. Eaton. *Lectures on topics in probability inequalities.* 1987.

36 A.H.P. van der Burgh, R.M.M. Mattheij (eds.). *Proceedings of the first international conference on industrial and applied mathematics (ICIAM 87).* 1987.

37 L. Stougie. *Design and analysis of algorithms for stochastic integer programming.* 1987.

38 J.B.G. Frenk. *On Banach algebras, renewal measures and regenerative processes.* 1987.

39 H.J.M. Peters, O.J. Vrieze (eds.). *Surveys in game theory and related topics.* 1987.

40 J.L. Geluk, L. de Haan. *Regular variation, extensions and Tauberian theorems.* 1987.

41 Sape J. Mullender (ed.). *The Amoeba distributed operating system: Selected papers 1984-1987.* 1987.

42 P.R.J. Asveld, A. Nijholt (eds.). *Essays on concepts, formalisms, and tools.* 1987.

43 H.L. Bodlaender. *Distributed computing: structure and complexity.* 1987.

44 A.W. van der Vaart. *Statistical estimation in large parameter spaces.* 1988.

45 S.A. van de Geer. *Regression analysis and empirical processes.* 1988.

46 S.P. Spekreijse. *Multigrid solution of the steady Euler equations.* 1988.

47 J.B. Dijkstra. *Analysis of means in some non-standard situations.* 1988.

48 F.C. Drost. *Asymptotics for generalized chi-square goodness-of-fit tests.* 1988.

49 F.W. Wubs. *Numerical solution of the shallow-water equations.* 1988.

50 F. de Kerf. *Asymptotic analysis of a class of perturbed Korteweg-de Vries initial value problems.* 1988.

51 P.J.M. van Laarhoven. *Theoretical and computational aspects of simulated annealing.* 1988.

52 P.M. van Loon. *Continuous decoupling transformations for linear boundary value problems.* 1988.

53 K.C.P. Machielsen. *Numerical solution of optimal control problems with state constraints by sequential quadratic programming in function space.* 1988.

54 L.C.R.J. Willenborg. *Computational aspects of survey data processing.* 1988.

55 G.J. van der Steen. *A program generator for recognition, parsing and transduction with syntactic patterns.* 1988.

56 J.C. Ebergen. *Translating programs into delay-insensitive circuits.* 1989.

57 S.M. Verduyn Lunel. *Exponential type calculus for linear delay equations.* 1989.

58 M.C.M. de Gunst. *A random model for plant cell population growth.* 1989.

59 D. van Dulst. *Characterizations of Banach spaces not containing l^1.* 1989.

60 H.E. de Swart. *Vacillation and predictability properties of low-order atmospheric spectral models.* 1989.

61 P. de Jong. *Central limit theorems for generalized multilinear forms.* 1989.

62 V.J. de Jong. *A specification system for statistical software.* 1989.

63 B. Hanzon. *Identifiability, recursive identification and spaces of linear dynamical systems, part I.* 1989.

64 B. Hanzon. *Identifiability, recursive identification and spaces of linear dynamical systems, part II.* 1989.

65 B.M.M. de Weger. *Algorithms for diophantine equations.* 1989.

66 A. Jung. *Cartesian closed categories of domains.* 1989.

67 J.W. Polderman. *Adaptive control & identification: Conflict or conflux?.* 1989.

68 H.J. Woerdeman. *Matrix and operator extensions.* 1989.

69 B.G. Hansen. *Monotonicity properties of infinitely divisible distributions.* 1989.

70 J.K. Lenstra, H.C. Tijms, A. Volgenant (eds.). *Twenty-five years of operations research in the Netherlands: Papers dedicated to Gijs de Leve.* 1990.

71 P.J.C. Spreij. *Counting process systems. Identification and stochastic realization.* 1990.

72 J.F. Kaashoek. *Modeling one dimensional pattern formation by anti-diffusion.* 1990.

73 A.M.H. Gerards. *Graphs and polyhedra. Binary spaces and cutting planes.* 1990.

74 B. Koren. *Multigrid and defect correction for the steady Navier-Stokes equations. Application to aerodynamics.* 1991.

75 M.W.P. Savelsbergh. *Computer aided routing.* 1992.

76 O.E. Flippo. *Stability, duality and decomposition in general mathematical programming.* 1991.

77 A.J. van Es. *Aspects of nonparametric density estimation.* 1991.

78 G.A.P. Kindervater. *Exercises in parallel combinatorial computing.* 1992.

79 J.J. Lodder. *Towards a symmetrical theory of generalized functions.* 1991.

80 S.A. Smulders. *Control of freeway traffic flow.* 1993.

81 P.H.M. America, J.J.M.M. Rutten. *A parallel object-oriented language: design and semantic foundations.* 1992.

82 F. Thuijsman. *Optimality and equilibria in stochastic games.* 1992.

83 R.J. Kooman. *Convergence properties of recurrence sequences.* 1992.

84 A.M. Cohen (ed.). *Computational aspects of Lie group representations and related topics. Proceedings of the 1990 Computational Algebra Seminar at CWI, Amsterdam.* 1991.

85 V. de Valk. *One-dependent processes.* 1993.

86 J.A. Baars, J.A.M. de Groot. *On topological and linear equivalence of certain function spaces.* 1992.

87 A.F. Monna. *The way of mathematics and mathematicians.* 1992.

88 E.D. de Goede. *Numerical methods for the three-dimensional shallow water equations.* 1993.

89 M. Zwaan. *Moment problems in Hilbert space with applications to magnetic resonance imaging.* 1993.

90 C. Vuik. *The solution of a one-dimensional Stefan problem.* 1993.

91. E.R. Verheul. *Multimedians in metric and normed spaces.* 1993.

92. J.L.M. Maubach. *Iterative methods for non-linear partial differential equations.* 1993.

93. A.W. Ambergen. *Statistical uncertainties in posterior probabilities.* 1993.

94. P.A. Zegeling. *Moving-grid methods for time-dependent partial differential equations.* 1993.

95. M.J.C. van Pul. *Statistical analysis of software reliability models.* 1993.

96. J.K. Scholma. *A Lie algebraic study of some integrable systems associated with root systems.* 1993.

97. J.L. van den Berg. *Sojourn times in feedback and processor sharing queues.* 1993.