CREDIBILITY INTERVALS

IN

BAYESIAN DISCRIMINANT ANALYSIS

ROBERT ANTHONY RIGBY

PHD

IMPERIAL COLLEGE LONDON

1987

ACKNOWLEDGEMENTS

I would like to acknowledge the many helpful discussions with my supervisor Professor Sir David Cox and the support provided by a Science Research Council grant.

ABSTRACT

A method of approximating the posterior distribution of P, the probability that a new observation y belongs to one of two populations with no parameters in common is developed and applied generally to Exponential family populations and specifically to multivariate Normal linear model populations with unequal covariance matrices, Gamma, Binomial and Poisson populations. The posterior moment generating function of the log likelihood function at y is derived and the posterior moments of the likelihood, log likelihood and L, the log likelihood ratio function are found. The method is extended to populations with parameters in common and applied to multivariate Normal linear model populations with equal covariance matrices. The posterior distribution and credibility interval for L are approximated using Pearson curves. The posterior distribution and credibility interval for P are obtained by transformation from L.

The method is generalised to the situation where the new observation y belongs to one of more than two populations (with or without parameters in common). The method is also generalised to the situation where the mixing parameters are unknown and the random samples were obtained by mixture sampling from the combined population distribution, and their population of membership subsequently identified. A method of combining subsets of variables modelled seperately is shown, including the location linear discriminant model and the situation of independent subsets of variables as special cases.

Finally the stability of the predictive density function is considered. A method of obtaining credibility interval bands for the likelihood and log likelihood functions is given. For the likelihood function, this gives a credibility interval band around the predictive density function. The joint posterior moment generating function of the log likelihood functions at multiple new values is also obtained.

CONTENTS

APPENDICES

PAGES

FIGURES

CHAPTER 1

Estimates, predictive distribution and credibility interval for P, the probability that a new observation belongs to one of two populations

1.1 Introduction

Suppose that observations belong to one of two distinct populations Π_1 or Π_2, and that in population Π_i the d-dimensional random variable X has probability density function $p_i(x|\Theta)$ of known, but not neccessarily the same functional form, for i=1,2, dependent on some unknown parameter vector Θ which combines the parameters from both populations. This allows for the possibility of common parameters between the populations. Suppose that independent random samples, $\underline{x}_1 = (x_{1j}, j=1,n_1)$ and $\underline{x}_2 = (x_{2j}, j=1,n_2)$, from Π_1 and Π_2 respectively, are to be used to gain information concerning which population, Π_1 or Π_2, a new observation y belongs to. Prior to being observed y is assumed to have been drawn from Π_1 or Π_2 with probabilities q_1 and q_2 respectively,i.e. y is assumed to have been drawn from the combined population distribution, $q_1p_1(y|\Theta)+q_2p_2(y|\Theta)$. The probability P that the new observation y belongs to Π_1 given the value y, the combined population parameter vector Θ and the prior probabilities is,

$$P = \frac{q_1p_1(y|\Theta)}{q_1p_1(y|\Theta) + q_2p_2(y|\Theta)} \quad . \qquad (1.1.1)$$

The purpose of this research is to develop a method for approximating the posterior distribution and finding a credibility interval for P. The method is outlined later in chapter 1 and is applied to multivariate Normal populations both with equal (chapter 2) and unequal (chapter 3) covariance matrices, and to Exponential family populations, including Gamma, Binomial and Poisson populations (chapter 4).

In chapter 5 the problem is generalised to allow the previously fixed known prior probabilities of population membership, q_1 (and $q_2=1-q_1$), to be unknown. Assuming a known or vague prior distribution for q_1, then random samples $\underline{x}_1$ and $\underline{x}_2$, obtained by mixture sampling, i.e. sampled from the combined population distribution $q_1p_1(x|\Theta) + q_2p_2(x|\Theta)$, and then identified as belonging to Π_1 and Π_2 respectively, can be used to update the prior distribution for q_1. The number of observations in the data set from each of the two populations gives information concerning the distribution of q_1. If the mixing parameter q_1 was unknown and the random samples were obtained by seperate sampling, then the problem is insoluble since no information is available concerning the mixing parameters, unless a non vague prior distribution is assumed for q_1.

The problem of more than two populations is considered in chapter 6. In chapter 7 the general situation of multiple new observations from one or more populations, where the mixing parameters may be unknown, is considered and applied to multivariate Normal linear model populations, both with equal and unequal population error precision matrices. The stability of the predictive density function is investigated in chapter 8. The problem of combining subsets of variables modelled seperately is considered in chapter 9, including as a special case the situation of independent subsets of variables. In chapter 10 the problem of using the extra information concerning the unknown population parameters Θ provided by the new observations of unknown population membership sampled from the mixture distribution is considered. Conclusions are given in chapter 11.

The notation used for population probability density functions used throughout this thesis is that of Aitchison and Dunsmore (1975), except where stated otherwise.

1.2 Estimates of P

Estimates of P using independent random samples from Π_1 and Π_2 have been either 'estimative' or 'predictive' (Aitchison, Habbema and Kay, 1977).

An 'estimative' estimate of P, denoted by P_e, is obtained by substituting into $p_i(y|\Theta)$ point estimates $\hat{\Theta}$ of the population parameters Θ, for i=1,2. The resulting estimates of $p_i(y|\Theta)$, i=1,2 , are then substituted into equation (1.1.1), giving

$$P_e = \frac{q_i p_i(y|\hat{\Theta})}{q_1 p_1(y|\hat{\Theta}) + q_2 p_2(y|\hat{\Theta})} \quad . \qquad (1.2.1)$$

A 'predictive' estimate of P, denoted by P_p, is obtained by substituting the predictive densities $p_i(y|\underline{x}_1,\underline{x}_2)$ for the population densities $p_i(y|\Theta)$, i=1,2 , into equation (1.1.1). The predictive density is defined by

$$p_i(y|\underline{x}_1,\underline{x}_2) = \int p_i(y|\Theta)\, p(\Theta|\underline{x}_1,\underline{x}_2)\, d\Theta \quad ,$$

(Aitchison and Dunsmore, 1975, Ch.2), where $p(\Theta|\underline{x}_1,\underline{x}_2)$ is the posterior distribution of Θ given the independent random samples $\underline{x}_1$ and $\underline{x}_2$, and assuming either a prior distribution for Θ or the limiting case of vague prior knowledge about Θ.

Hence,

$$P_p = \frac{q_1 p_1(y|\underline{x}_1,\underline{x}_2)}{q_1 p_1(y|\underline{x}_1,\underline{x}_2) + q_2 p_2(y|\underline{x}_1,\underline{x}_2)} \quad . \qquad (1.2.2)$$

The 'predictive' estimate of P can be shown to be equal to $p(y\in\Pi_1|y,\underline{x}_1,\underline{x}_2)$ from a fully Bayes approach, i.e.

$$p(y\in\Pi_1|y,\underline{x}_1,\underline{x}_2) = \frac{q_1 p_1(y|\underline{x}_1,\underline{x}_2)}{q_1 p_1(y|\underline{x}_1,\underline{x}_2) + q_2 p_2(y|\underline{x}_1,\underline{x}_2)}$$

(Aitchison and Dunsmore, 1975, Ch.11).

The 'estimative' and 'predictive' estimates of P have been compared for a particular data set concerning Conn's syndrome by Aitchison and Dunsmore(1975, Ch11). They found that the 'estimative' and 'predictive' estimates of P can differ dramatically, particularly in their practical implications. The 'estimative' and 'predictive' estimates of P have also been compared in a simulation study by Aitchison, Habbema and Kay (1977). They also found that the 'estimative' and 'predictive' estimates of P can differ dramatically for the multivariate Normal populations they considered, and that the 'predictive' estimates of P generally gave more accurate estimates of P as judged by the absolute error in the estimate of the log likelihood ratio.

An alternative Classical estimate of P to the usual 'estimative' estimate P_e has been proposed and considered by Moran and Murphy (1979). They propose using bias-corrected estimates of the log likelihood functions $\log p_i(y|\Theta)$ for i=1,2, and hence obtaining estimate P_u of P by transformation from (1.1.1). This implies using a bias-corrected estimate, L_u, of the log likelihood ratio function, $L = \log p_1(y|\Theta) - \log p_2(y|\Theta)$. They found that the Classical estimate P_u of P performed comparably with the 'predictive' estimate P_p of P, as judged by the absolute error in the log likelihood ratio, but also that all three estimates of P, P_u, P_e and P_p can differ dramatically.

1.3 The posterior distribution and credibility interval for P

Since the 'estimative' and 'predictive' estimates of P can differ considerably as shown by Aichison and Dunsmore (1975, Ch11), Aitchison, Habbema and Kay, (1977) and Moran and Murphy (1979), it would be useful to obtain a Bayesian credibility interval for P. In particular this will throw some light on the relationship between the 'estimative' and 'predictive' estimates of P.

If, given y, P is treated as a randon variable,(a function of the random variables Θ), then the posterior distribution of P given $\underline{x}_1$ and $\underline{x}_2$ can be obtained.

In order to achieve this, the posterior moments of the logarithm of the likelihood ratio, L, are obtained, where $L=l_1-l_2$, and $l_i = \log(p_i(y|\Theta))$ for i=1,2. The posterior distribution of L can then be approximated, either using Pearson curves (Elderton and Johnson, 1969, Chapter 5) or, if appropriate, using a Normal distribution. A Normal approximation may be justified in a particular case provided calculation of the posterior skewness and kurtosis of L shows them to be close to the values for a Normal distribution, i.e. 0 and 3 respectively. An approximate posterior credibility interval for L can easily be obtained using percentage points for Pearson curves, tabled in full by Johnson, Nixon and Amos (1963), and in summary by Biometrika tables, No. 42.

The posterior distribution and credibility interval for P are obtained by transformation from the posterior distribution and credibility interval for L since,

$$P = \frac{1}{1 + \frac{q_2}{q_1}.e^{-L}} \quad . \qquad (1.3.1)$$

Hence the posterior probability density fuctions p(P) and p(L) are related by,

$$p(P) = p(L)|dL/dP| = p(L)e^{L}.(q_1/q_2).(1 + e^{-L}q_2/q_1)^2 \quad \text{,for } 0<P<1. \qquad (1.3.2)$$

The problem of obtaining an interval for P has been considered extensively from a Classical viewpoint by Critchley and Ford. They propose two main methods, the first being based on finding the variance and higher moments of the bias corrected estimate of the log likelihood ratio, L_u, and using these to obtain an interval for P (Critchley and Ford, 1984, 1985). The problems with this approach are, firstly, that the exact formulae for the variance and higher moments of L_u have so far appeared intractable, at least for multivariate Normal populations so far considered, and secondly, even when found, the formulae for the variance and higher moments will be functions of the unknown parameters Θ and hence will have to be estimated by substituting some estimate of Θ into the formulae. The second Classical method they propose is to obtain an interval for the log likelihood ratio L from its profile, (Critchley, Ford and Rijal, 1986a) and (Critchley, Ford and Hirst, 1986).

A review of several approaches to obtaining an interval for P is given by Critchley, Ford and Rijal (1986b), where they compare the intervals obtained for Conn's syndrome data from Aitchison and Dunsmore (1975, Ch11). They found that their profile likelihood approach and the Bayesian approach, derived here, gave very similar results for this particular data set. They are currently carrying out a simulation study to compare the performance of the different approaches to obtaining an interval for P.

1.4 The posterior moments of L, the log likelihood function at the new value y, when the two population distributions have no parameters in common

If the set of parameters involved in $p_1(y|\Theta)$ is different from the set in $p_2(y|\Theta)$, and the two sets of parameters have independent prior distributions, then l_1 and l_2 have independent posterior distributions and the posterior cumulant generating function, $K_L(t)$, of L can be obtained by finding the posterior cumulant generating functions of l_1 and l_2 separately since,

$L = l_1 + l_2$ and hence

$$\Phi_L(t) = E\left[e^{tL}\right] = E\left[e^{t(l_1-l_2)}\right] = E\left[e^{tl_1}\right].E\left[e^{-tl_2}\right] = \Phi_{l_1}(t).\Phi_{l_2}(-t),$$

since l_1 and l_2 have independent posterior distributions.

Hence,

$$K_L(t) = \log \Phi_L(t) = \log \Phi_{l_1}(t) + \log \Phi_{l_2}(-t) = K_{l_1}(t) + K_{l_2}(-t) ,$$

$$\text{i.e. } K_L(t) = K_{l_1}(t) + K_{l_2}(-t) . \qquad (1.4.1)$$

Hence the r^{th} cumulant of L, $(K_r)_L$, is related to the r^{th} cumulants of l_1 and l_2, i.e. $(K_r)_{l_1}$ and $(K_r)_{l_2}$ respectively, for r=1,2,..., by,

$$(K_r)_L = (K_r)_{l_1} + (-1)^r.(K_r)_{l_2} . \qquad (1.4.2)$$

The posterior cumulant generating function and cumulants of l_i are found , for i=1,2 , from which the posterior cumulant generating function and cumulants of L are obtained using (1.4.1) and (1.4.2). The posterior central moments of L, i.e. moments of L about its mean, can then be found from its posterior cumulants using a special case of the recurrence relationship derived in Appendix 2,

$$\mu'_r = \sum_{j=0}^{r-2} \binom{r-1}{j} \mu'_j \, K_{r-j} \qquad \text{for } r \geq 1, \qquad (1.4.3)$$

where μ'_r is the r^{th} moment about the mean and K_r is the r^{th} cumulant.

Hence
$$\mu'_2 = K_2 ,$$
$$\mu'_3 = K_3 , \qquad (1.4.4)$$
$$\mu'_4 = K_4 + 3\mu'_2 K_2 = K_4 + 3K_2^2 .$$

The posterior moments of L are thus found and are used to approximate the posterior distribution of P as explained in section 1.3.

Chapter 2 considers the case of multivariate Normal populations with unequal means and unequal covariance matrices. This case has been considered from a Classical rather than a Bayesian viewpoint by Crithley, Ford and Rijal (1986a, 1986b). Chapter 4 considers the case of Exponential family populations having no parameters in common, including as examples, pairs of Gamma, Binomial and Poisson populations, each pair having no parameters in common. The case of multivariate Normal linear model populations with unequal covariance matrices is considered in chapter 7.

1.5 Posterior moments of L when the two population distributions have parameters in common

If the populations have common parameters, then l_1 and l_2 will not have independent posterior distributions and hence the moments of L cannot be found from the moments of l_1 and l_2 separately, but can be found either from the joint posterior moment generating function of l_1 and l_2 or directly from the posterior moment generating function of L itself. The posterior cumulant generating function of L is used to obtain the posterior mean and cumulants of L, from which the posterior moments of L about its mean are obtained using (1.4.4) above. The posterior distribution of P can then be approximated as explained in section 1.3.

Chapter 3 considers the case of two multivariate Normal populations with unequal means but equal covariance matrices. Critchley and Ford (1985) and Crichley, Ford and Hirst (1986) have considered this case using a Classical rather than a Bayesian approach. The case of multivariate Normal linear model populations with equal covariance matrices is considered in chapter 7.

1.6 Motivation for obtaining a credibility interval for P

The dramatic differences in the estimates of P, discussed in section 1.2, indicate that the posterior distibution and, in particular, the posterior credibility interval for P would be desirable. It is hoped that this will throw light on the relationship between the different estimates of P.

Different new observations y could lead to the same estimate of P but have dramatically different posterior distributions and posterior credibility intervals and so should not be treated identically.

From a practical point of view, given a particular utility loss-gain structure, relating possible actions a to the true population Π_i of the new observation y, i.e. U(a,i), a point value for P would give a point optimum action a (i.e. with maximum expected utility or gain over all possible actions a), while a distribution for P would give a distribution of optimum actions a or a distribution for the expected utility or gain for a particular action a. If the utility structure, U(a,i,y), also depends on the value of the new observation y, as suggested by Aitchison and Dunsmore (1975, Ch12), then since y is known the utility structure given the value y returns to being a function of a and i. For example, suppose that all values of P in a credibility interval for P lead to the same optimum action a, one would feel confident in choosing this action. If, however, the values of P in a credibility interval for P lead to different optimum actions a, more consideration should perhaps be given to the choice of action, for example, by collecting further data before choosing an action.

CHAPTER 2

Populations with no parameters in common: Multivariate Normal populations with unequal means and unequal covariance matrices

2.1 Introduction

The method described in chapter 1 of approximating the posterior distribution and credibility interval for P, the probability that a new observation y belongs to one of two populations Π_1 or Π_2, where y is assumed to have been drawn from the combined population distribution $q_1p_1(y|\Theta) + q_2p_2(y|\Theta)$ is applied to d-dimensional multivariate Normal populations, Π_1 and Π_2, with unknown and unequal population means μ_1 and μ_2 and unknown and unequal population covariance matrices Σ_1 and Σ_2 respectively. Let $T_i=\Sigma_i^{-1}$ be the population precision matrix for i=1,2, then $\Theta = (\mu_1,T_1,\mu_2,T_2)$ and the i^{th} population density, for i=1,2, is given by

$$p_i(y|\Theta) = p_i(y|\mu_i,T_i) = (2\pi)^{-d/2}|T_i|^{1/2}\exp\left[-\tfrac{1}{2}(y-\mu_i)'T_i(y-\mu_i)\right] . \qquad (2.1.1)$$

Estimates of P

The true probability, P, that a new observation y belongs to population Π_1 is given by substituting $p_i(y|\Theta)$ from equation (2.1.1), for i=1,2 into equation (1.1.1).

The 'estimative' estimate of P, P_e, is given by substituting point estimates for (μ_i,T_i) into equation (2.1.1) and then substituting the resulting $p_i(y|\Theta)$ for i=1,2 into equation (1.2.1). The maximum likelihood estimates of (μ_i,T_i) are given by $(\bar{x}_i,n_iS_i^{-1})$ where $\bar{x}_i$ is the mean and S_i is the corrected sum of squares and products matrix for the i^{th} random sample, $\underline{x}_i$, for i=1,2.

The 'predictive' estimate of P, P_p, is given by substituting the predictive density $p_i(y|\underline{x}_1,\underline{x}_2)$ for i=1,2 into equation (1.2.2), where $p_i(y|\underline{x}_1,\underline{x}_2)$ is given by equation (2.2.2) in section 2.2, and (b,c,v,h) are the parameters of the posterior distribution of (μ_i,T_i) for i=1,2. Assuming vague prior information about the parameters (μ_i,T_i) then $p_i(y|\underline{x}_1,\underline{x}_2)$ is given, for i=1,2, by substituting $(\bar{x}_i,n_i,n_i-1,S_i)$ for (b,c,v,h) in equation (2.2.2), i.e.

$$p_i(y|\underline{x}_1,\underline{x}_2) = \frac{\Gamma[n_i/2]}{\pi^{d/2}.\,\Gamma[(n_i-d)/2]} \cdot \left|\left[\frac{1+n_i}{n_i}\right]S_i\right|^{-1/2}\left[1+\left[\frac{n_i}{n_i+1}\right]D_i\right]^{-n_i/2} \qquad (2.1.2)$$

where $D_i = (y-\bar{x}_i)'S_i^{-1}(y-\bar{x}_i)$.

Credibility interval for P

In section 2.2 the moments of the posterior distribution of the likelihood $z=p(y|\mu,T)$ considered as a function of the random variables (μ,T) for a given y are obtained, by integrating z^t over the posterior distribution of (μ,T) given a random sample $\underline{x}=(x_1,x_2,\ldots,x_n)$ from a multivariate Normal population with mean μ and precision matrix T, where $\bar{x}$ is the sample mean and S is the sample corrected sum of squares and products matrix.

Since $E(z^t) = E(e^{tlogz})$, the formula obtained in section 2.2 for the posterior moments of z is also the formula for $\Phi_l(t)$, the posterior moment generating function of the logarithm of the likelihood $l=logz$. In section 2.3 the posterior cumulants of l are obtained from $\log \Phi_l(t)$. Hence the posterior cumulants of l_i [$= logz_i = \log p_i(y|\mu_i,T_i)$], for i=1,2, are then obtained by appropriate substitution for (b,c,v,h) in the results of section 2.3. If vague prior information was assumed for (μ_i,T_i) then

$$p(\mu_i,T_i|\underline{x}_i) = p(\mu_i,T_i|\bar{x}_i,S_i) = NoWi_d(\bar{x}_i,n_i,n_i-1,S_i) \quad ,$$

in the notation of Aitchison and Dunsmore (1975, Ch2), and hence $(\bar{x}_i,n_i,n_i-1,S_i)$ is substituted for (b,c,v,h) in the results of section 2.3 to give the posterior cumulant generating function and posterior cumulants of l_i for i=1,2.

The cumulant generating function and cumulants of $L=l_1-l_2$ are then found using equations (1.4.1) and (1.4.2) since the posterior distributions of l_1 and l_2 will be independent. They will be independent because the population distributions have different parameter sets (μ_1,T_1) and (μ_2,T_2) respectively and the random samples from the two populations are independent. Hence the posterior distributions of the two sets of parameters will be independent, provided it is assumed that the prior distributions of the two sets of parameters are independent.

For the example in section 2.4 the cumulants of l_1,l_2 and hence L are found assuming independent vague prior information for the two sets of parameters (μ_1,T_1) and (μ_2,T_2). Pearson curves were then used to approximate p(L), the posterior probability density function of L using the first four cumulants of L. The transformation (1.3.1) gives the relationship between P and L and hence p(P), the posterior probability density function of P is found using equation (1.3.2). In the example in section 2.4 equal prior probabilities, q_1 and q_2, of population membership were assumed. The conclusions are given in section 2.5.

2.2 The posterior moments of z, the likelihood function at y

Let $z = p(y|\mu,T)$

$$= (2\pi)^{-d/2}|T|^{1/2}\exp\left[-\tfrac{1}{2}(y-\mu)'T(y-\mu)\right] ,$$

where d is the dimension of the vector y, y is a new observation of known value, and (μ,T) has a Normal-Wishart posterior distribution $p(\mu,T|\bar{x},S)$ given the sample mean vector $\bar{x}$, the sample corrected sum of squares and products matrix S and the prior distribution $p(\mu,T)$.

If $p(\mu,T) = \text{NoWi}_d(\mu_0,n_0,v_0,h_0)$, then for $n_0 \rangle 0$,

$$p(\mu,T|\bar{x},S) = \text{NoWi}_d\left[\frac{n_0\mu_0+n\bar{x}}{n_0+n},n_0+n,v_0+n,h_1\right] ,$$

where $h_1 = h_0 + S + \frac{n_0 n}{n_0+n}(\bar{x}-\mu_0)(\bar{x}-\mu_0)'$, and n is the sample size,

(Aitchison and Dunsmore, 1975, Ch2).

If vague prior knowledge is assumed for (μ,T), i.e. $p(\mu,T) \propto |T|^{-\frac{1}{2}(d+1)}$, then $p(\mu,T|\bar{x},S) = \text{NoWi}_d(\bar{x},n,n-1,S)$, (Aitchison and Dunsmore, 1975, Ch2).

For a general result let $p(\mu,T|\bar{x},S) = \text{NoWi}_d(b,c,v,h)$, then,

$$E(z^t) = \int_{T\rangle 0}\int_{\mu\in R^d} z^t.p(\mu,T|\bar{x},S)\, d\mu dT$$

$$= \int_{T\rangle 0}\int_{\mu\in R^d} (2\pi)^{-\frac{1}{2}dt}\,|T|^{\frac{1}{2}t}\exp[-\tfrac{1}{2}t(y-\mu)'T(y-\mu)].$$

$$(2\pi)^{\frac{1}{2}d}\,|cT|^{\frac{1}{2}}\exp[-\tfrac{1}{2}(\mu-b)'cT(\mu-b)].$$

$$\frac{|\tfrac{1}{2}h|^{\frac{1}{2}v}\,|T|^{\frac{1}{2}(v-d-1)}\exp[-\tfrac{1}{2}\text{tr}hT]}{\Gamma_d(\tfrac{1}{2}v)}\, d\mu\, dT ,$$

where $\Gamma_d(\tfrac{1}{2}v) = \pi^{d(d-1)/4}.\prod_{j=1}^{d}\Gamma[\tfrac{1}{2}(v-d+j)]$,

and $T\rangle 0$ means T is positive definite.

Completing the square and integrating out μ, gives,

$$E(z^t) = \frac{K}{(c+t)^{\frac{1}{2}d}} \cdot \int |T|^{\frac{1}{2}(v-d+t-1)} \cdot \exp\left[-\tfrac{1}{2}\mathrm{tr}\left[h + \frac{ct}{c+t}(y-b)(y-b)'\right]T\right] dT,$$

where $K = \dfrac{|\frac{1}{2}h|^{\frac{1}{2}v}\, c^{\frac{1}{2}d}}{(2\pi)^{\frac{1}{2}d(t+1)}\, \Gamma_d(\frac{1}{2}v)}$.

By noting that the integrand above is in the form of the Wishart probability density function, T can be integrated out to the constant of the Wishart distribution giving,

$$E(z^t) = \frac{\Gamma_d[\frac{1}{2}(v+t)]}{\pi^{\frac{1}{2}td} \cdot \Gamma_d(\frac{1}{2}v)} \cdot \left[\frac{c}{c+t}\right]^{\frac{1}{2}d} \cdot |h|^{-\frac{1}{2}t} \cdot \left[\, 1+\left[\frac{ct}{c+t}\right]D \,\right]^{-\frac{1}{2}(v+t)} , \qquad (2.2.1)$$

where $D = (y-b)'h^{-1}(y-b)$.

Setting t=1 gives the multivariate Student t distribution evaluated at y, i.e.,

$$p(y|\underline{x}) = St_d\left[v, b, \left[1+\frac{1}{c}\right].\frac{h}{v}\right]$$

$$= \frac{\Gamma[\frac{1}{2}(v+1)]}{\pi^{\frac{1}{2}d} \cdot \Gamma[\frac{1}{2}(v-d+1)]} \cdot \left[\frac{c}{c+1}\right]^{\frac{1}{2}d} \cdot |h|^{-\frac{1}{2}} \cdot \left[\, 1+\left[\frac{c}{c+1}\right]D \,\right]^{-\frac{1}{2}(v+1)} . \qquad (2.2.2)$$

This is the predictive density function for y discussed in section 2.1.
The predictive density at y can be considered to be either the posterior expected value of the likelihood function at y, or the estimate of the likelihood function at y which minimises the posterior squared errror loss.

2.3 The posterior moment generating function of l=logz, and moments of l

The posterior moment generating function of l=logz is given by,
$\Phi_l(t) = E[e^{t.logz}] = E[z^t]$, which was obtained in equation (2.2.1).
On taking logs and differentiating (2.2.1) with respect to t,

$$\frac{d}{dt}\log \Phi_l(t) = -\tfrac{1}{2}d.\log\pi - \tfrac{1}{2}\log|h| + \sum_{i=1}^{d} \tfrac{1}{2}.\Psi[\tfrac{1}{2}(v+t-i+1)]$$

$$- \frac{d}{2(c+t)} - \tfrac{1}{2}.\log\left[1 + \frac{ctD}{c+t}\right] - \frac{(v+t).c^2.D}{2(c+t)^2\left[1 + \frac{ctD}{c+t}\right]} \qquad (2.3.1)$$

where $\Psi(s) = \Gamma'(s)/\Gamma(s)$ is the psi (digamma) function.

Now let $Q = - \dfrac{c^2.D}{2(c+t).[c+t+ctD]}$.

Using partial fractions,

$$\frac{d^{(r)}}{dt^{(r)}} Q = \frac{(-1)^r.r!}{2(c+t)^{r+1}} + \frac{(-1)^{r+1}.r!.(1+cD)^{r+1}}{2[c+t+ctD]^{r+1}} . \qquad (2.3.2)$$

Expressing the higher derivatives of (2.3.1) in terms of derivatives of Q and then substitution from (2.3.2) gives the r^{th} kumulant of l=logz, $(K_r)_l$,

$$(K_r)_l = \left[\frac{d^{(r)}}{dt^{(r)}} \log \Phi(t)\right]_{t=0} = (\tfrac{1}{2})^r \sum_{i=1}^{d} \Psi^{(r-1)}[\tfrac{1}{2}(v-i+1)]$$

$$+ \frac{(r-2)!(-1)^r}{2c^r}.[cr+(d-v)(r-1)]$$

$$+ \frac{(r-2)!(-1)^r}{2c^r}.[-cr+v(r-1)(1+cD)].(1+cD)^{r-1} \qquad (2.3.3)$$

for $r \geq 2$.

The r^{th} moment of $l=\log z$ about its mean, i.e. $(\mu_r')_l$, can now be obtained from the kumulants of l, given by equation (2.3.3), using the recurrence relationship (1.4.3).

Hence,

$$E(l) = (K_1)_l = -\frac{d}{2}.\log \pi - \frac{d}{2c} - \frac{1}{2}.\log|h| - \frac{vD}{2} + \frac{1}{2}.\sum_{i=1}^{d} \Psi[\tfrac{1}{2}(v-i+1)] ,$$

$$V(l) = (\mu_2')_l = (K_2)_l = \frac{d}{2c^2} + \left[\frac{v-c}{c}\right].D + \frac{1}{2}.v.D^2 + \frac{1}{4}.\sum_{i=1}^{d} \Psi'[\tfrac{1}{2}(v-i+1)] ,$$

$$(\mu_3')_l = (K_3)_l = -\frac{d}{c^3} - \frac{3D}{c}.\left[\frac{v}{c} - 1\right] - \frac{3D^2}{2}.\left[\frac{2v}{c} - 1\right] - v.D^3 + \frac{1}{8}.\sum_{i=1}^{d} \Psi^{(2)}[\tfrac{1}{2}(v-i+1)],$$

$(\mu_4')_l = (K_4)_l + (K_2)_l^2$, where,

$$(K_4)_l = \frac{3d}{c^4} + \frac{12(v-c)D}{c^3} + \frac{(3v-2c)6D^2}{c^2} + \frac{(3v-c)4D^3}{c} + 3vD^4 + \frac{1}{16}.\sum_{i=1}^{d} \Psi^{(3)}[\tfrac{1}{2}(v-i+1)].$$

(2.3.4)

If v is an integer then the value of the Psi function Ψ and its derivatives in $E(l)$, $V(l)$, $(\mu_3')_l$ and $(\mu_4')_l$ can be found using the formulae given in Appendix 3. If v is not an integer then an efficient numerical method of approximation which is particularly useful when the Psi function and its derivatives are required at the same value is given in Appendix 4.

For the examples in section 2.4 vague prior knowledge about (μ_i,T_i), for $i=1,2$, was assumed, so that $p(\mu_i,T_i|\bar{x}_i,S_i) = \text{NoWi}_d(\bar{x}_i,n_i,n_i-1,S_i)$ and hence $(\bar{x}_i,n_i,n_i-1,S_i)$ was substituted for (b,c,v,h) in the above formulae to give the posterior cumulants of $l_i=\log z_i$ for $i=1,2$. The posterior cumulants of the log likelihood ratio $L=\log(z_1/z_2)=l_1-l_2$ about its mean were then found using equations (1.4.2). Pearson curves were then used to approximate the posterior distribution of L (Elderton and Johnson, 1969, Ch5) using the first four cumulants of L. The posterior distribution of P was then obtained by transformation using equation (1.3.1), assuming equal prior probabilities i.e. $q_1=q_2$.

2.4 Examples of the posterior distribution of P

A sample of size 10 was generated from each of two populations with multivariate Normal density functions $p(y|\mu_i,T_i)$ with mean μ_i and precision matrix T_i $(=\Sigma_i^{-1})$ giving sample statistics $\bar{x}_i$ and S_i for i=1,2, where,

$$\mu_1' = (0.0 \quad 0.0) \quad , \quad \Sigma_1 = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix} ,$$

$$\mu_2' = (1.0 \quad 1.0) \quad , \quad \Sigma_2 = \begin{bmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{bmatrix} ,$$

$$\bar{x}_1' = (0.169 \; 0.407) \quad , \quad S_1 = \begin{bmatrix} 8.694 & 0.393 \\ 0.393 & 5.838 \end{bmatrix} ,$$

$$\bar{x}_2' = (0.981 \; 0.959) \quad , \quad S_2 = \begin{bmatrix} 4.188 & 1.454 \\ 1.454 & 4.049 \end{bmatrix} .$$

A new set of observations was independently sampled from the combined population distribution $\frac{1}{2}[p(y|\mu_1,T_1) + p(y|\mu_2,T_2)]$. Four observations y_1 to y_4 were chosen from the new set having distinctly different posterior distributions p(P) as shown by figures 1 to 4, where,

y_1 = (0.023 0.208) ,

y_2 = (-1.697 -0.4480) ,

y_3 = (-1.299 -2.818) and,

y_4 = (0.006 1.962).

On each graph the following are marked,

(P_1,P_2) , a cental 90% credibility interval for P, i.e. equal tail areas.

T , the true value of P obtained using the true population parameters, given by equation (1.1.1).

Q , the 'estimative' estimate of P assuming unequal covariance matrices for the populations, and using maximum likelihood estimates $(\bar{x}_i,S_i/n_i)$ for (μ_i,Σ_i) for i=1,2, i.e. a quadratic discriminant estimate.

E , the 'estimative' estimate of P assuming an equal covariance matrix, Σ, for the two populations and using estimates $(\bar{x}_i,(S_1+S_2)/(n_1+n_2-2))$ for (μ_i,Σ) for i=1,2, i.e. the usual linear discriminant estimate.

B , the 'predictive' estimate of P assuming unequal covariance matrices and vague prior knowledge about the parameters, given by substituting equation (2.1.2) for i=1,2 into equation (1.2.2).

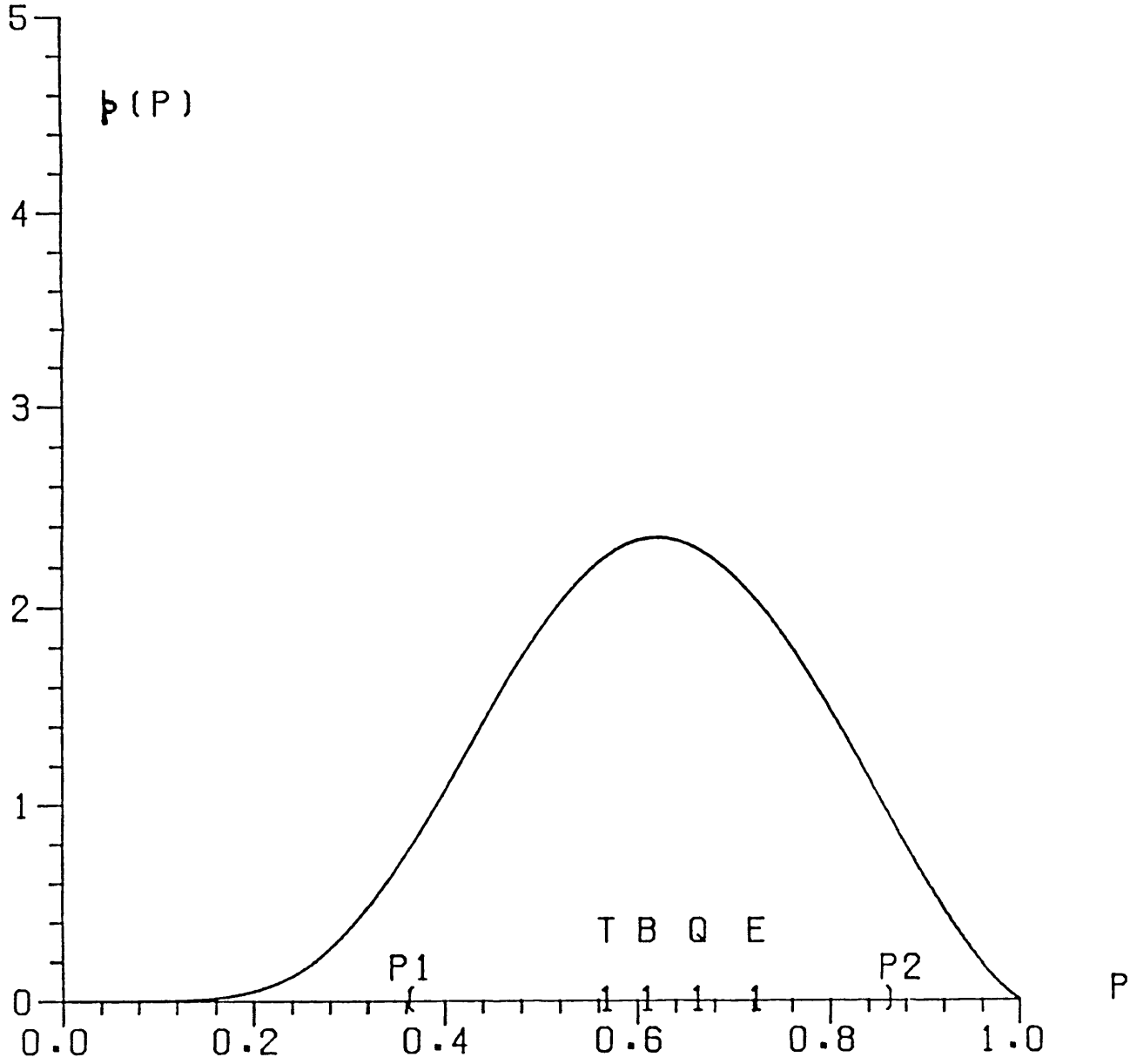


Figure 1. The posterior density function p(P) for new observation y_1.

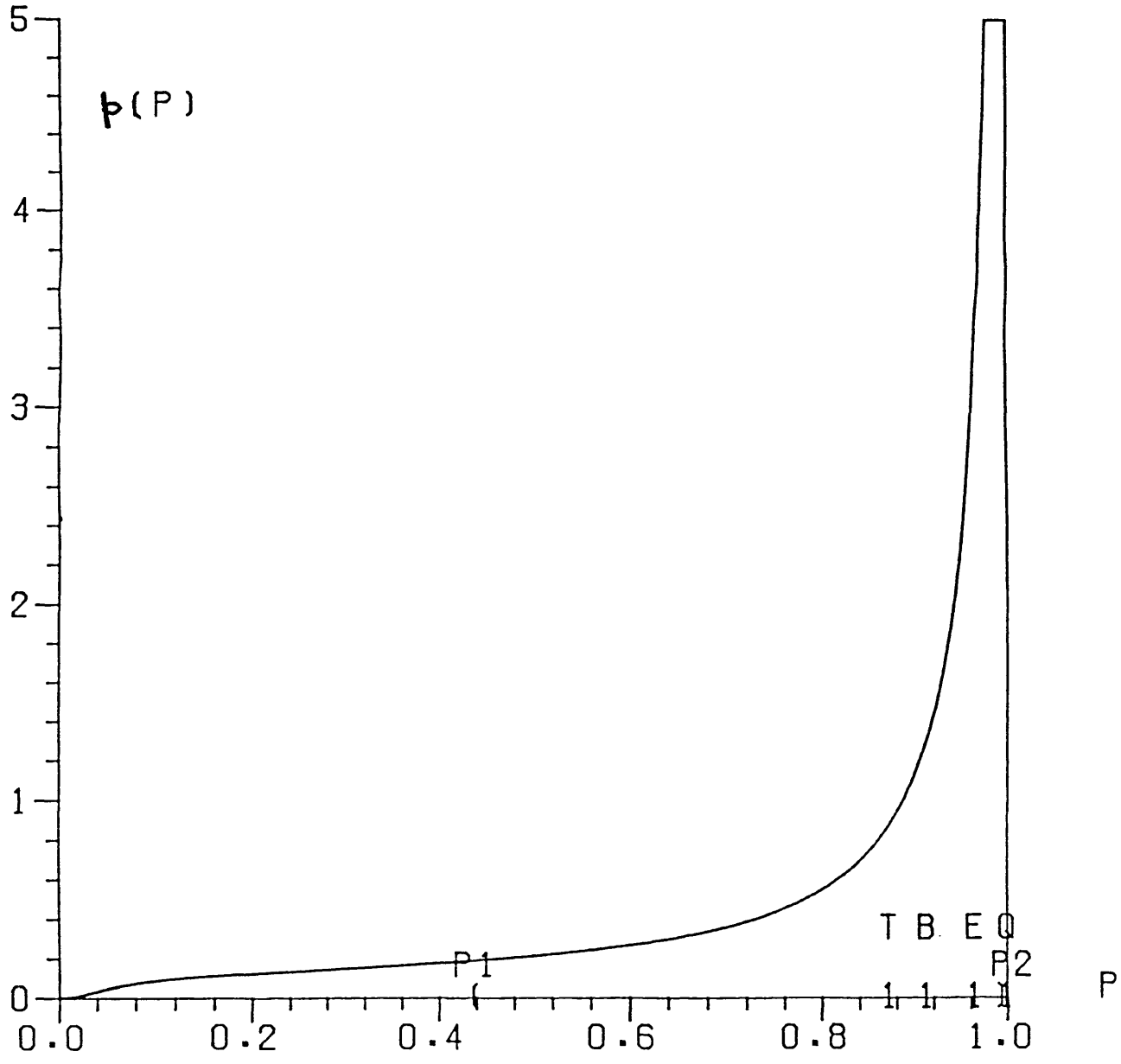


Figure 2. The posterior density function p(P) for new observation y_2.

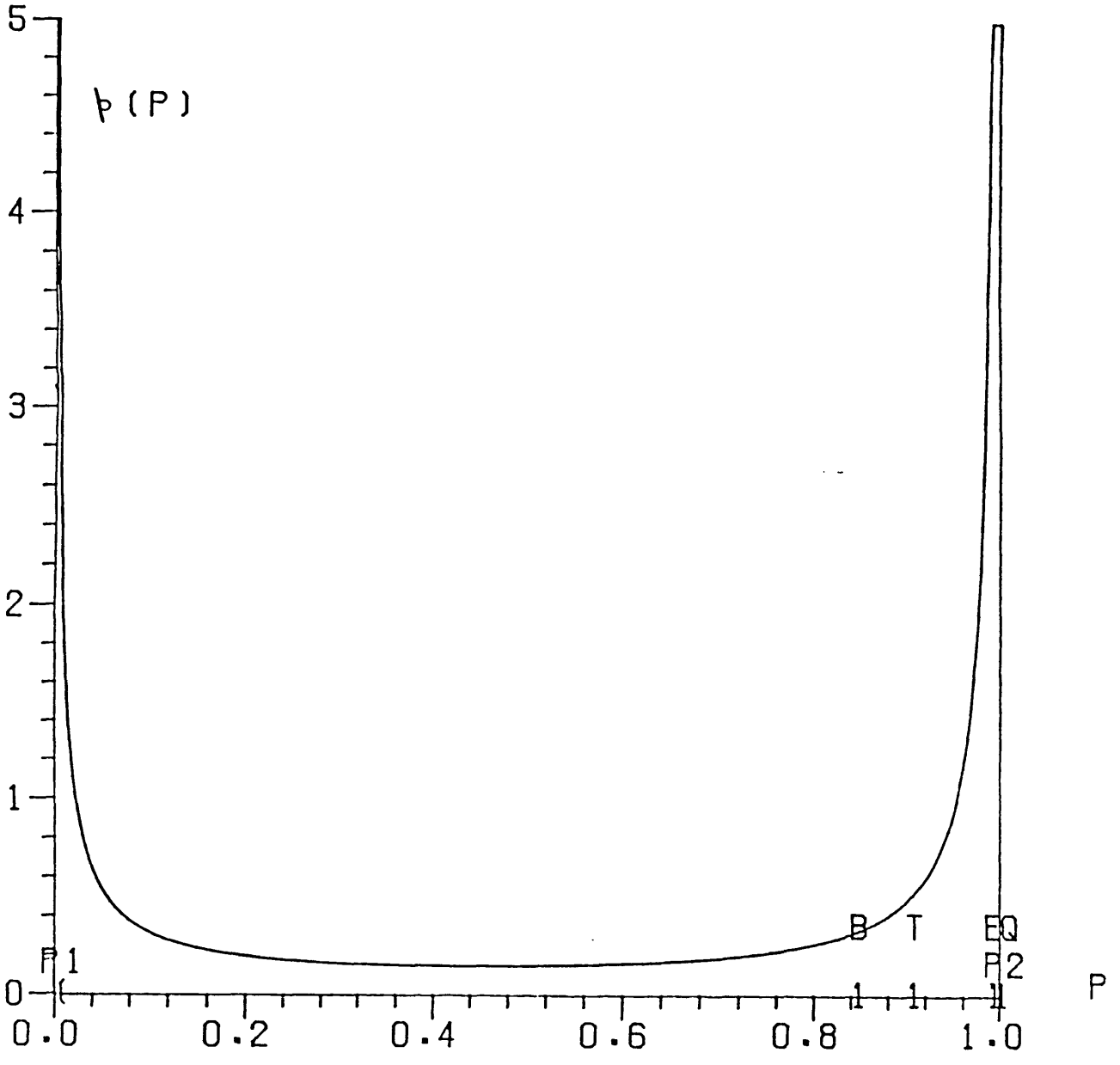


Figure 3. The posterior density function p(P) for new observation y_3.

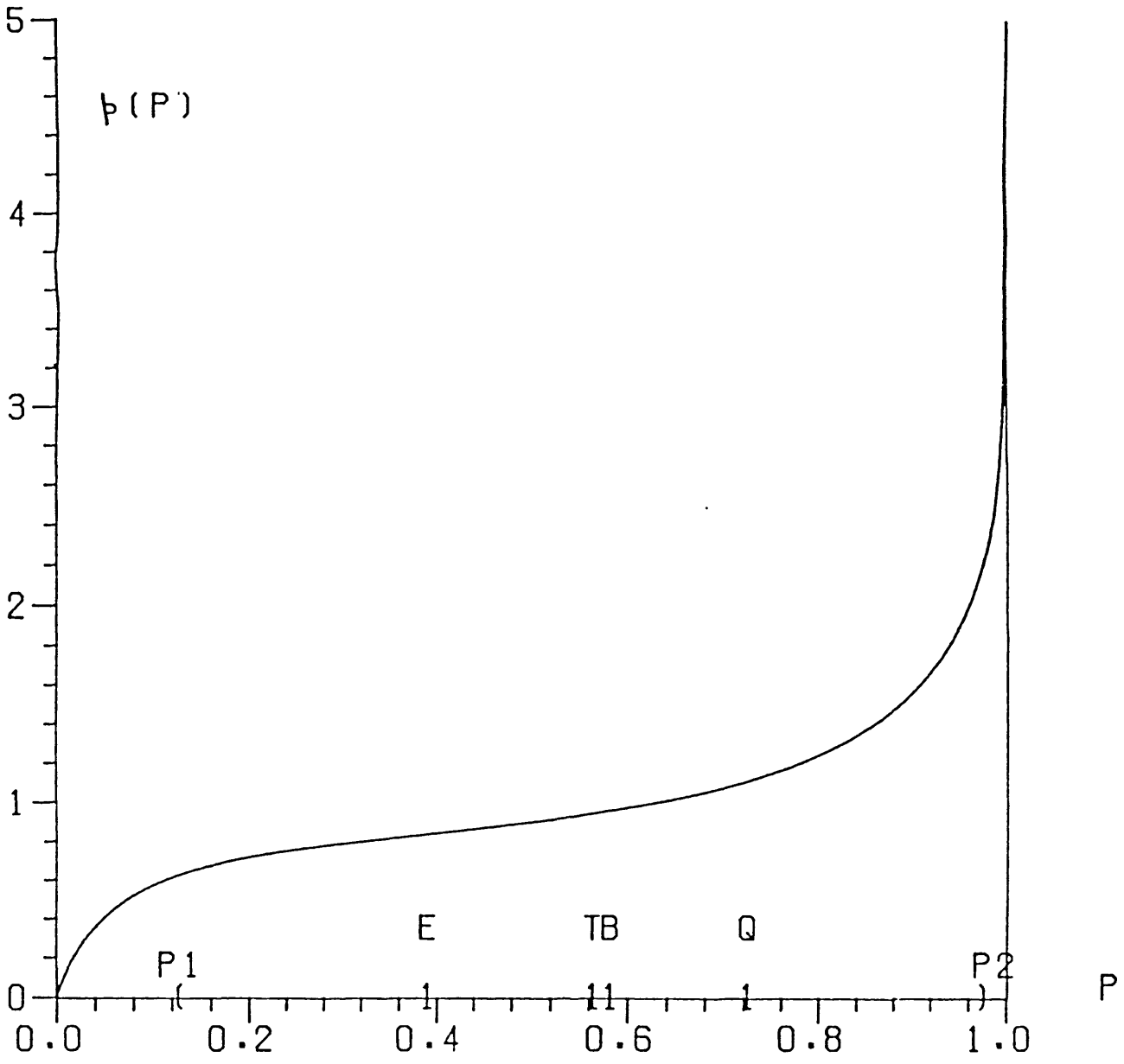


Figure 4. The posterior density function p(P) for new observation y_4.

The points y_1 to y_4 are plotted in figure 5 together with the true chi-squared and estimated T-squared 90% confidence ellipses for a new observation y for each of the two populations, respectively defined, for i=1,2, by the set of y such that,

$$(y-\mu_i)'\Sigma_i^{-1}(y-\mu_i) \leq X^2_{2,0.90} \quad \text{and,}$$

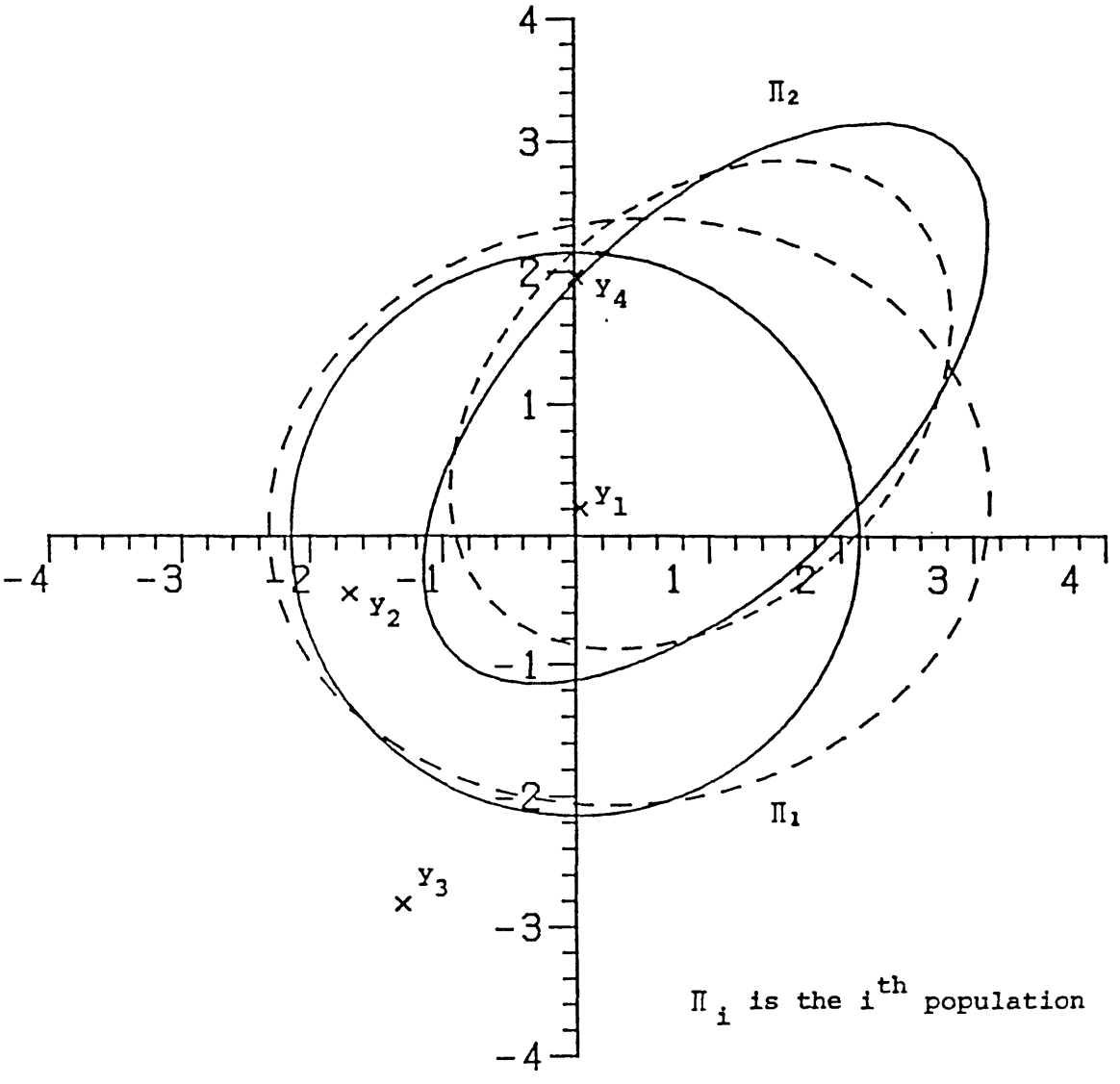
$$(n_i-1).\left[\frac{n_i}{n_i+1}\right].(y-\bar{x}_i)'S_i^{-1}(y-\bar{x}_i) \leq T^2_{n_i-1,d,0.90} = \frac{(n_i-1)d}{(n_i-d)}.F_{d,n_i-d,0.90} \qquad (2.4.1)$$

where n_i=10 is the sample size and d=2 is the dimension of the populations. The suffix 0.90 refers to the upper 90% point of the distribution. Equation (2.4.1) can be interpreted as meaning that the expected or average % of new observations captured within R_j, the region defined by (2.4.1) over a sequence of independent random samples with sample statistics $(\bar{x}_i,S_i)_j$ for j=1,2,..., each with the same number of independently sampled new observations, is 90%.

As several of the estimates of P are close to zero or one, the following table compares the corresponding estimates of the log likelihood ratio, L, given by equation (1.3.1), where $q_1=q_2=0.5$. The corresponding central 90% credibility interval, $(L(P_1),L(P_2))$, is also given. Hence the interval and the different esimates of L (and by transformation P) can be compared,

	L(T)	L(Q)	L(E)	L(B)	$L(P_1)$	$L(P_2)$
y_1	0.265	0.687	0.950	0.440	-0.577	1.903
y_2	1.896	5.759	3.237	2.313	-0.271	12.341
y_3	2.251	8.359	4.761	1.693	-5.815	23.287
y_4	0.243	0.948	-0.461	0.320	-1.970	4.004

From the above table it is clear that the different estimates of the log likelihood function L can differ quite dramatically but still lie within the credibility interval for L. This is particularly striking for y_3. Also despite the estimates of L being highly positive for y_3, the credibility interval still includes the value L=0, which corresponds to P=½, indicating that the true population of membership of y_3 is not clear cut.



_____ True chi-squared 90% confidence ellipse for y.

_ _ _ _ _ Estimated T-squared 90% confidence ellipse for y.

Figure 5. The true and estimated 90% confidence ellipses for a new observation y plotted together with the new observations y_1-y_4

Denote the posterior probabiliy density function of P based on y_j by C_j. It is interesting to compare C_1 with C_3. Curve C_3 may be desribed as 'unstable' in the sence that the value of P appears likely to be at one extreme of the interval (0,1), but which extreme is unclear. Curve C_1 may be described as 'stable' since P appears very likely to be moderately greater than 0.5. Clearly therefore further sampling would be of considerable value for deciding at which extreme P lies for y_3, while it would only serve to provide a more accurate estimate of P for y_1.

The following question therefore arises. From which population is further sampling the more helpful. In order to decide this consider the mean and variance(M and V) of the log likelihoods $l_{ij} = \log p(y_j|\mu_i,T_i)$ for j=1,2,3,4 and population i=1,2,

	i=1		i=2	
	M	V	M	V
y_1	-1.924	0.143	-2.481	0.441
y_2	-4.144	1.228	-9.266	14.170
y_3	-10.714	17.271	-18.175	63.266
y_4	-3.783	0.903	-4.576	2.497

From this table it can be seen that the 'stability' of P in C_1 is due to the low variances of l_{11} and l_{21}, while the 'instability' of P in C_3 is due to the large variances of l_{13} and l_{23}, particularly l_{23}. Hence it would be better to continue sampling from the second population if firmer knowledge about the true population to which y_3 belongs is required. This is even more apparent for y_2.

2.5 Conclusions

From section 2.4 it can be seen that even though estimates of P are close to 0 or 1 it is still possible for the 90% credibility interval for P to be very wide, as for example for y_3. Furthermore the 90% credibility interval for P can include both Bayesian and Classical estimates (B and Q) even when they differ greatly, e.g.,

	B	Q	P_1	P_2
y_3	0.845	0.99977	0.00297	$1 - 0.77\times10^{-10}$
y_2	0.910	0.9969	0.433	$1 - 0.44\times10^{-5}$

Hence, though estimates B and Q may be both well within permissible statistical limits, they may differ immensely in their practical implications.

CHAPTER 3

Populations with parameters in common: Multivariate Normal Populations with unequal means but equal covariance matrices

3.1 Introduction

The method descibed in chapter 1 of approximating the posterior distribution and credibility interval for P, the probability that a new observation y belongs to one of two populations Π_1 or Π_2, where y is assumed to have been drawn from the combined population distribution $q_1p_1(y|\Theta) + q_2p_2(y|\Theta)$, is applied to d-dimensional multivariate Normal populations Π_1 and Π_2 with unknown and unequal population means μ_1 and μ_2 and unknown but equal population covariance matrix Σ. Let $T=\Sigma^{-1}$ be the common population precision matrix, then $\Theta=(\mu_1,\mu_2,T)$ and the i^{th} population density , for i=1,2, is given by

$$p_i(y|\Theta) = p_i(y|\mu_i,T) = (2\pi)^{-d/2}|T|^{1/2}\exp\left[-\tfrac{1}{2}(y-\mu_i)'T(y-\mu_i)\right] . \qquad (3.1.1)$$

Estimates of P

The true probability, P, that the new observation y belongs to population Π_1 is given by substituting $p_i(y|\Theta)$ from equation (3.1.1), for i=1,2, into equation (1.1.1).

The 'estimative' estimate of P, P_e, is given by substituting point estimates for (μ_i,T) into equation (3.1.1) and then substituting the resulting $p_i(y|\Theta)$ for i=1,2 into equation (1.2.1). The maximum likelihood estimates of $[\mu_1,\mu_2,T]$ are given by $[\bar{x}_1,\bar{x}_2,(n_1+n_2)(S_1+S_2)^{-1}]$ where $\bar{x}_i$ is the mean and S_i is the corrected sum of squares and products matrix for the i^{th} random sample, $\underline{x}_i$, for i=1,2.

The 'predictive' estimate of P, P_p, is given by substituting the predictive density $p_i(y|\underline{x}_1,\underline{x}_2)$ for i=1,2 into equation (1.2.2), where $p_i(y|\underline{x}_1,\underline{x}_2)$ is given by equation (2.2.2) in section 2.2 and (b,c,v,h) are the parameters of the posterior distribution of (μ_i,T) given random samples $\underline{x}_1$ and $\underline{x}_2$, for i=1,2. Assuming vague prior information about the parameters (μ_1,μ_2,T) then $p_i(y|\underline{x}_1,\underline{x}_2)$ is given for i=1,2 by substituting $(\bar{x}_i,n_i,n_1+n_2-2,S)$ for (b,c,v,h) in equation (2.2.2) where $S=S_1+S_2$, i.e.

$$p_i(y|\underline{x}_1,\underline{x}_2) = \frac{\Gamma[(n_1+n_2-1)/2]}{\pi^{d/2}.\,\Gamma[(n_1+n_2-d-1)/2]}\cdot\left|\left[\frac{1+n_i}{n_i}\right]S\right|^{-1/2}\left[1+\left[\frac{n_i}{n_i+1}\right]D_i\right]^{-(n_1+n_2-1)/2} \qquad (3.1.2)$$

where $D_i = (y-\bar{x}_i)'S^{-1}(y-\bar{x}_i)$.

Credibility interval for P

In section 3.2, $\Phi_L(t)=E(e^{tL})$, the posterior moment generating function of the logarithm of the likelihood ratio, $L=l_1-l_2$, where $l_i=\log(p_i(y|\Theta))$ for i=1,2, considered as a function of random variables (μ_1,μ_2,T) for a given y is obtained by integrating e^{tL} over the posterior distribution of (μ_1,μ_2,T) given random samples $\underline{x}_1=(x_{11},x_{12},..,x_{1n1})$ and $\underline{x}_2=(x_{21},x_{22},..,x_{2n2})$ from population Π_1 and Π_2 respectively with sample means $\bar{x}_1$ and $\bar{x}_2$ and sample corrected sums of sqares and products matrices S_1 and S_2 respectively.

In section 3.3 general formulae for the posterior mean and cumulants of L are obtained. In any particular example the posterior mean and moments of L about its mean can be obtained by appropriate substitution for (b_1,c_1,b_2,c_2,v,h) in the results of section 3.3. If vague prior information is assumed for (μ_1,μ_2,T) then

$$p(\mu_1,\mu_2,T|\underline{x}_1,\underline{x}_2) = p(\mu_1,\mu_2,T|\bar{x}_1,\bar{x}_2,S) = \text{NoNoWi}(\bar{x}_1,n_1,\bar{x}_2,n_2,n_1+n_2-2,S)$$

where $S=S_1+S_2$, and hence $(\bar{x}_1,n_1,\bar{x}_2,n_2,n_1+n_2-2,S)$ are substituted for (b_1,c_1,b_2,c_2,v,h) in the results in section 3.3.

In section 3.4 the posterior distribution of L is approximated both using a Normal distribution and using Pearson curves. The posterior distribution of P is then obtained by using transformation (1.3.1). A numerical example is given in section 3.5.

3.2 The posterior moment generating function of L, the logarithm of the likelihood ratio at y.

The posterior moment generating function of L, the log of the likelihood ratio at y, given the independent random samples $\underline{x}_1$ and $\underline{x}_2$ from the populations Π_1 and Π_2 respectively, is given by

$$\Phi_L(t) = E[e^{tL}] = E\left[\left[\frac{p_1(y|\Theta)}{p_2(y|\Theta)}\right]^t\right] = \int\left[\frac{p_1(y|\Theta)}{p_2(y|\Theta)}\right]^t p(\Theta|\underline{x}_1,\underline{x}_2)d\Theta \quad , \qquad (3.2.1)$$

where $p(\Theta|\underline{x}_1,\underline{x}_2)$ is the posterior distribution of Θ given $\underline{x}_1$ and $\underline{x}_2$.

Let Π_i be a d-dimensional multivariate Normal population, $No_d(\mu_i,T)$, with mean μ_i and precision matrix T,i.e.

$$p_i(y|\Theta) = p_i(y|\mu_i,T) = \frac{|T|^{1/2}}{(2\pi)^{d/2}}\exp[-\tfrac{1}{2}(y-\mu_i)'T(y-\mu_i)] \quad \text{for } i=1,2. \quad (3.2.2)$$

Let $\Theta = (\mu_1,\mu_2,T)$ have a Normal-Normal-Wishart prior distribution, $p(\mu_1,\mu_2,T)$,prior that is to observing the values of $\underline{x}_1$ and $\underline{x}_2$, i.e.

$p(\mu_1,\mu_2,T) = NoNoWi_d(\mu_{10},n_{10},\mu_{20},n_{20},v_0,h_0)$, defined by $p(\mu_1,\mu_2,T)=p(\mu_1|T)p(\mu_2|T)p(T)$ where $p(\mu_i|T)=No_d(\mu_{i0},n_{i0}T)$ for i=1,2, and $p(T)=Wi_d(v_0,h_0)$ (Aitchison and Dunsmore,1975,Chapter 2).
Then, for $n_{10}>0$ and $n_{20}>0$, the posterior distribution of Θ is

$$p(\Theta|\underline{x}_1,\underline{x}_2) = p(\mu_1,\mu_2,T|\bar{x}_1,\bar{x}_2,S)$$
$$= NoNoWi_d\left[\frac{n_{10}\mu_{10}+n_1\bar{x}_1}{n_{10}+n_1},n_{10}+n_1,\frac{n_{20}\mu_{20}+n_2\bar{x}_2}{n_{20}+n_2},n_{20}+n_2,v_0+n_1+n_2-2,h_1\right]$$

where $h_1 = h_0+S+\sum_{i=1}^{2}\left[\frac{n_{i0}n_i}{n_{i0}+n_i}(\bar{x}_i-\mu_{i0})(\bar{x}_i-\mu_{i0})'\right]$.

If vague prior information is assumed, i.e. $p(\mu_1,\mu_2,T) \propto |T|^{-(d+1)/2}$ then
$p(\Theta|\underline{x}_1,\underline{x}_2) = p(\mu_1,\mu_2,T|\bar{x}_1,\bar{x}_2,S) = NoNoWi_d(\bar{x}_1,n_1,\bar{x}_2,n_2,n_1+n_2-2,S)$,
where $\bar{x}_1$ and $\bar{x}_2$ are the sample mean vectors, n_1 and n_2 are the sample sizes and S_1 and S_2 are the sample corrected sum of squares and products matrices for the random samples $\underline{x}_1$ and $\underline{x}_2$ respectively, and $S=S_1+S_2$.

For a general result let

$$p(\Theta|\underline{x}_1,\underline{x}_2)=p(\mu_1,\mu_2,T|\bar{x}_1,\bar{x}_2,S)=NoNoWi_d(b_1,c_1,b_2,c_2,v,h) \quad . \qquad (3.2.3)$$

Hence substituting $p_1(y|\Theta)$ and $p_2(y|\Theta)$ from (3.2.2) and $p(\Theta|\underline{x}_1,\underline{x}_2)$ from (3.2.3) into (3.2.1) gives,

$$\Phi_L(t) = \int_{T>0} \int_{\mu_1 \in R^d} \int_{\mu_2 \in R^d} \left[\frac{p_1(y|\mu_1,T)}{p_2(y|\mu_2,T)}\right]^t p(\mu_1,\mu_2,T|\underline{x}_1,\underline{x}_2,S)\, d\mu_1\, d\mu_2\, dT ,$$

$$= \iiint |T|^{t/2} (2\pi)^{-dt/2} \exp[-\tfrac{1}{2}t(y-\mu_1)'T(y-\mu_1)]$$
$$. |T|^{-t/2} (2\pi)^{dt/2} \exp[\tfrac{1}{2}t(y-\mu_2)'T(y-\mu_2)]$$
$$. |c_1T|^{1/2} (2\pi)^{-d/2} \exp[-\tfrac{1}{2}(\mu_1-b_1)'c_1T(\mu_1-b_1)]$$
$$. |c_2T|^{1/2} (2\pi)^{-d/2} \exp[-\tfrac{1}{2}(\mu_2-b_2)'c_2T(\mu_2-b_2)]$$
$$. \frac{|\frac{1}{2}h|^{v/2}}{\Gamma_d(v/2)} |T|^{(v-d-1)/2} \exp[-\tfrac{1}{2}\mathrm{tr}hT]\, d\mu_1\, d\mu_2\, dT,$$

where $T>0$ means T is positive definate, and $\Gamma_d(v/2) = \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma[(v-d+j)/2]$.

Completing the squares in each of μ_1 and μ_2 and integrating out the result to the constants of Normal distributions,

$No[(c_1b_1+ty)/(c_1+t),(c_1+t)T]$ and $No[(c_2b_2-ty)/(c_2-t),(c_2-t)T]$ gives,

$$\Phi_L(t) = \left[\frac{c_1}{c_1+t}\right]^{d/2} \left[\frac{c_2}{c_2-t}\right]^{d/2} \frac{|\frac{1}{2}h|^{v/2}}{\Gamma_d(v/2)} \int |T|^{(v-d-1)/2} \exp[-\tfrac{1}{2}\mathrm{tr}HT]\, dT,$$

where $H = h + \frac{c_1t}{c_1+t}(b_1-y)(b_1-y)' - \frac{c_2t}{c_2-t}(b_2-y)(b_2-y)'$.

By noting that the integrand is in the form of the Wishart probability density function, T can be integrated out to the normalising constant of the Wishart probabillity density function giving,

$$\Phi_L(t) = \left[\frac{c_1}{c_1+t}\right]^{d/2} \left[\frac{c_2}{c_2-t}\right]^{d/2} \frac{|h|^{v/2}}{|H|^{v/2}} \quad . \qquad (3.2.4)$$

Now by considering the determinant of a partitioned matrix of size (d+2) by (d+2) with elements h, $[\alpha_1(b_1-y),\alpha_2(b_2-y)]$, $[-(b_1-y),-(b_2-y)]'$ and I_2 the two dimensional identity matrix,

$$|H| = |h|.[1 + \alpha_1D_1 + \alpha_2D_2 + \alpha_1\alpha_2(D_1D_2-D_{12}^2)] , \qquad (3.2.5)$$

where $\alpha_1 = \frac{c_1t}{c_1+t}$, $\alpha_2 = -\frac{c_2t}{c_2-t}$, $D_i=(b_i-y)'h^{-1}(b_i-y)$ for $i=1,2$

and $D_{12}=(b_1-y)'h^{-1}(b_2-y)$.

3.3 The Posterior Moments of L

On taking logs and differentiating $\Phi_L(t)$ r times with respect to t and setting t=0 gives the r^{th} cumulant of L, $(K_r)_L$, i.e.,

$$(K_r)_L = \left[\frac{d^{(r)}}{dt^{(r)}} \log \Phi_L(t)\right]_{t=0} = \frac{(-1)^r(r-1)!d}{2c_1^r} + \frac{(r-1)!d}{2c_2^r} - \frac{v}{2}\left[\frac{d^{(r)}}{dt^{(r)}} \log u\right]_{t=0}, \quad (3.3.1)$$

where $r \geq 1$ and $u = [1 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_1\alpha_2(D_1D_2 - D_{12}^2)]$.

Differentiating the equation, $\frac{du}{dt} = u \frac{d}{dt} \log u$, (r-1) times with respect to t and setting t=0 gives the recurrence relationship,

$$\left[\frac{d^{(r)}}{dt^{(r)}} \log u\right]_{t=0} = \frac{1}{u}\left[\left[\frac{d^{(r)}}{dt^{(r)}} u\right]_{t=0} - \sum_{i=1}^{r-1}\binom{r-1}{i}\left[\frac{d^{(i)}}{dt^{(i)}} u\right]_{t=0}\left[\frac{d^{(r-i)}}{dt^{(r-i)}} \log u\right]_{t=0}\right] \quad (r \geq 1)$$

where $\left[\frac{d^{(r)}}{dt^{(r)}} u\right]_{t=0} = \frac{(-1)^{r-1}r!D_1}{c_1^{r-1}} + \frac{(-1)r!D_2}{c_2^{r-1}} - \left[\frac{c_1c_2}{c_1+c_2}\right]r!\left[\frac{(-1)^r}{c_1^{r-1}} + \frac{1}{c_2^{r-1}}\right](D_1D_2 - D_{12}^2)$

for $r \geq 1$.

The rth moment of L about its mean, $(\mu_r')_L$, is now given by,

$$(\mu_r')_L = E([L-E(L)]^r) = \sum_{j=0}^{r-2}\binom{r-1}{j}(\mu_j')_L.(K_{r-j})_L \quad , \text{ for } r \geq 1.$$

Hence $E(L) = (K_1)_L = -\frac{1}{2}\left[v(D_1 - D_2) + d\left[\frac{1}{c_1} - \frac{1}{c_2}\right]\right]$,

$$V(L) = (\mu_2')_L = (K_2)_L = \frac{1}{2}\left[v(D_1^2 + D_2^2) + 2v\left[\frac{D_1}{c_1} + \frac{D_2}{c_2}\right] + d\left[\frac{1}{c_1^2} + \frac{1}{c_2^2}\right]\right] - vD_{12}^2 ,$$

$$(\mu_3')_L = (K_3)_L = -\left[v(D_1^3 - D_2^3) + 3v\left[\frac{D_1^2}{c_1} - \frac{D_2^2}{c_2}\right] + 3v\left[\frac{D_1}{c_1^2} - \frac{D_2}{c_2^2}\right] + d\left[\frac{1}{c_1^3} - \frac{1}{c_2^3}\right]\right]$$
$$+3vD_{12}^2\left[(D_1 - D_2) + \left[\frac{1}{c_1} - \frac{1}{c_2}\right]\right] , \quad (3.3.2)$$

and $(\mu_4')_L = (K_4)_L + 3(K_2)_L^2$ where

$$(K_4)_L = 3\left[v(D_1^4 + D_2^4) + 4v\left[\frac{D_1^3}{c_1} + \frac{D_2^3}{c_2}\right] + 6v\left[\frac{D_1^2}{c_1^2} + \frac{D_2^2}{c_2^2}\right] + 4v\left[\frac{D_1}{c_1^3} + \frac{D_2}{c_2^3}\right] + d\left[\frac{1}{c_1^4} + \frac{1}{c_2^4}\right]\right]$$

$$+6vD_{12}^4 + 12vD_{12}^2\left[(D_1 - D_2) + \left[\frac{1}{c_1} - \frac{1}{c_2}\right]\right]^2 - 6vD_{12}^2\left[D_1 + \frac{1}{c_1}\right]\left[D_2 + \frac{1}{c_2}\right] .$$

3.4. Approximating the Posterior Distributions of L and P

The posterior distribution of L can now be approximated using its posterior moments either by Pearson curves (Elderton and Johnson,1969,Chapter 5), or, if appropriate, by a Normal distribution,i.e.

$L \sim N(\mu,\sigma^2)$ where $\mu = E(L)$ and $\sigma^2 = V(L)$.

A Normal approximation to the posterior distribution of L may be justified in an individual case by checking that the posterior skewness and kurtosis of L are sufficiently close to the values 0 and 3 respectively.

Using the transformation $P = [1 + (q_2/q_1)e^{-L}]^{-1}$, the posterior distribution of P can be approximated by

$$p(P) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma P(1-P)} \exp\left[-\frac{1}{2\sigma^2}(L-\mu)^2\right] \text{ where } L = \log\left[\frac{q_2}{q_1}\left[\frac{1-P}{P}\right]\right] , \qquad (3.4.1)$$

for $0<P<1$.

Hence, using a Normal approximation to the posterior distribution of L, a $(1-\alpha)$ credibility interval for L is given by,
$(L_1,L_2) = (\mu-z_{\alpha/2}\sigma,\mu+z_{\alpha/2}\sigma)$, where $z_{\alpha/2}$ is the $\alpha/2$ point of a Normal distribution.

Alternatively, using a Pearson curve approximation to the posterior distribution of L, a more accurate credibility interval (L_1,L_2) for L based on its first four cumulants can be obtained. Percentage points for Pearson curves are given by Johnson, Nixon and Amos (1963) or by Biometrika Tables, No.42.

A $(1-\alpha)$ credibility interval (P_1,P_2) for P can be obtained from the corresponding credibility interval (L_1,L_2) for L by using transformation (1.3.1), giving,

$$(P_1,P_2) = ([1 + (q_2/q_1)e^{-L_1}]^{-1} , [1 + (q_2/q_1)e^{-L_2}]^{-1}) . \qquad (3.4.2)$$

3.5. A numerical example of the calculations

As an illustration of the calculations, suppose random samples each of size 10 from Π_1 and Π_2 give the following sample statistics,

$$\underline{\bar{x}}_1 = \begin{bmatrix}0.169\\0.407\end{bmatrix}, \quad \underline{\bar{x}} = \begin{bmatrix}0.981\\0.959\end{bmatrix}, \text{ and } S = S_1+S_1 = \begin{bmatrix}12.88 & 1.85\\1.85 & 9.89\end{bmatrix}.$$

where $(\underline{\bar{x}}_i, S_i)$ are the sample mean and corrected sums of squares and products matrix for the sample from the i^{th} population, for i=1,2.

Assume that observations from populations Π_1 and Π_2 have bivariate Normal distributions with unequal means μ_1 and μ_2 but equal precision matrix T.
If equal prior probabilities q_1 and q_2, and vague prior information about $\Theta = (\mu_1,\mu_2,T)$ are assumed, then in the equations for posterior mean and cumulants of L given in section 3.3 substitute,

$$[b_1,c_1,b_2,c_2,v,h] = [\bar{x}_1,n_1,\bar{x}_2,n_2,n_1+n_2-2,S]$$

$$= \left[\begin{bmatrix}0.169\\0.407\end{bmatrix},10,\begin{bmatrix}0.981\\0.959\end{bmatrix},10,18,\begin{bmatrix}12.88 & 1.85\\1.85 & 9.89\end{bmatrix}\right]$$

For new observation, $y = \begin{bmatrix}0.023\\0.208\end{bmatrix}$, $D_1 = (b_1-y)'h^{-1}(b_1-y) = 0.00495$,

$D_2 = (b_2-y)'h^{-1}(b_2-y) = 0.11035$, and $D_{12} = (b_1-y)'h^{-1}(b_2-y) = 0.0222$.
Hence for y, E(L) = 0.9486, V(L) = 0.3285 and an approximate 95% credibility interval for L is (-0.175,2.072), using a Normal approximation to the posterior distribution of L as in section 3.4. Hence the corresponding approximate credibility interval for P is (0.456,0.888), using equation (3.4.2).
The posterior skewness and kurtosis of L are given by,

$(\mu_3')_L = (K_3)_L = 0.1495$

$(K_4)_L = 0.1021$ and so

$(\mu_4')_L = 0.1021 + (3\times0.3285^2) = 0.4258$

Hence the posterior skewness of L is $(\sqrt{\beta_1})_L = \dfrac{(\mu_3')_L}{(\mu_2')_L^{3/2}} = 0.7942$

and the posterior kurtosis of L is $(\beta_2)_L = \dfrac{(\mu_4')_L}{(\mu_2')_L^{2}} = 3.9462.$

A more accurate credibility interval for L can be obtained by approximating the posterior distribution of L using Pearson curves. Using percentage points for Pearson curves, an approximate 95% credibility interval for L is given by,
$(\mu_L - 1.588\sigma_L, \mu_L + 2.308\sigma_L) = (0.03844, 2.27143)$,
and hence an approximate 95% credibility interval for P is given by (0.5096,0.9065).

CHAPTER 4

Exponential family populations with no parameters in common

4.1 Introduction

The method described in chapter 1 of approximating the posterior distribution and credibility interval for P, the probability that a new observation y belongs to one of two populations, where y is assumed to have been drawn from the combined population distribution $q_1p_1(y|\Theta) + q_2p_2(y|\Theta)$ is applied to Exponential family populations, Π_1 and Π_2, with unknown parameter sets Θ_1 and Θ_2 respectively, having no parameters in common. Here the parameter vector $\Theta=(\Theta_1,\Theta_2)$ and the i^{th} population density, for i=1,2, is given by

$$p_i(y|\Theta) = a(\Theta_i)b(y)e^{s(y)'r(\Theta_i)} , \qquad (4.1.1)$$

where y is a d dimensional observation,

Θ_i is a d_0 dimensional parameter vector,

s is a k dimensional vector function of y, and

r is a k dimensional vector function of Θ.

Estimates of P

The true probability P that the new observation y belongs to Π_1 is given by substituting $p_i(y|\Theta)$, given by equation (4.1.1), for i=1,2, into equation (1.1.1).

The 'estimative' estimate of P, P_e, is given by substituting point estimates of Θ_i into equation (4.1.1) and then substituting the resulting $p_i(y|\Theta)$, for i=1,2, into equation (1.2.1).

The 'predictive' estimate of P, P_p, is given by substituting the predictive density $p_i(y|\underline{x}_1,\underline{x}_2)$, for i=1,2, into equation (1.2.2), where $p_i(y|\underline{x}_1,\underline{x}_2)$ is given by equation (4.2.13) and (ν,ω) are the parameters of the posterior distribution of Θ_i given the random sample $\underline{x}_i$, for i=1,2, given by (4.2.5).

Credibility interval for P

In section 4.2 the joint posterior moment generating function of the log likelihood functions at multiple new values $(y_1,y_2,..,y_m)$ from an Exponential family population is obtained, and in particular the posterior moment generating function of the log likelihood function $l = \log p(y|\Theta)$ at a single future value y is obtained in section 4.2.4.

The posterior cumulants of l are obtained from the posterior cumulant generating function $K_l(t) = \log \Phi_l(t)$ and hence the posterior mean and central moments of l are found using equation (1.4.4). The posterior mean and central moments of $l_i = \log p_i(y|\Theta)$, for i=1,2, are then obtained by substituting for (ν,ω) the parameters of the posterior distribution of Θ_i into the results in section 4.2.5. The posterior moments of $L=l_1-l_2$ are then found using equations (1.4.1) and (1.4.2). The posterior distribution of L can then be approximated either by Pearson curves using the first four moments of L or, if appropriate, by a Normal approximation using the first two moments. The posterior distribution of P can then be found by transformation using equation (1.3.2).

4.2 The joint posterior moment generating function of the log likelihood function at multiple new values from an Exponential family population

4.2.1 Introduction

Let X be a d-dimensional random variable with probability density function, $p(x|\Theta)$, a member of the Exponential family, i.e.

$$p(x|\Theta) = a(\Theta)b(x)e^{s(x)'r(\Theta)} \quad , \qquad (4.2.1)$$

where Θ is a d_0 dimensional parameter vector,
s is a k dimensional vector function of x, and
r is a k dimensional vector function of Θ.

Let $\underline{x} = (x_1,x_2,\ldots,x_n)$ be a random sample of observations of random variable X. Interest centres on the joint posterior moment generating function, $\Phi_{\underline{l}}(\underline{t})$, of the log likelihood functions of possible future observations $(y_j; j=1,2,\ldots,m)$ of the same random variable X.

$$\Phi_{\underline{l}}(\underline{t}) = E\left[e^{\underline{t}'\underline{l}}\right] \quad ,$$

where here $\underline{t} = (t_1,t_2,\ldots,t_m)$, $\underline{l} = (l_1,l_2,\ldots,l_m)$,

and $l_j = \log p(y_j|\Theta)$ for $j=1,2,..,m$.

Hence

$$\Phi_{\underline{l}}(\underline{t}) = E\left[\prod_{j=1}^{m} [p(y_j|\Theta)]^{t_j}\right]$$

$$= \int \left[\prod_{j=1}^{m} [p(y_j|\Theta)]^{t_j}\right].p(\Theta|\underline{x})\, d\Theta \quad , \qquad (4.2.2)$$

where $p(\Theta|\underline{x})$ is the posterior distribution of the parameter vector Θ given the random sample $\underline{x}$.

4.2.2 The posterior distribution of the parameter vector Θ given the random sample x for an Exponential family population

The natural conjugate prior distribution for Θ is given by,

$$p(\Theta) = c[\nu_0,\omega_0]a(\Theta)^{\nu_0}e^{\omega_0'r(\Theta)} , \qquad (4.2.3)$$

where ν_0 is a scalar, ω_0 is a k dimensional vector and $c[\nu_0,\omega_0]$ is the normalising constant.

The posterior distribution for Θ given the random sample x is given by

$$p(\Theta|\underline{x}) \propto p(\Theta).\prod_{j=1}^{n} p(x_j|\Theta) ,$$

$$\propto a(\Theta)^{\nu}e^{\omega'r(\Theta)} ,$$

where $\nu = \nu_0 + n$ and $\omega = \omega_0 + \sum_{j=1}^{n} s(x_j)$, $\qquad (4.2.4)$

i.e. $p(\Theta|\underline{x}) = c[\nu,\omega]a(\Theta)^{\nu}e^{\omega'r(\Theta)}$. $\qquad (4.2.5)$

4.2.3 Evaluating $\Phi_{\underline{l}}(\underline{t})$

From equation (4.2.2),

$$\Phi_{\underline{l}}(\underline{t}) = \int \left[\prod_{j=1}^{m} \left[a(\theta) b(y_j) e^{s(y_j)'r(\theta)} \right]^{t_j} \right] . c[\nu,\omega] a(\theta)^{\nu} e^{\omega' r(\theta)} \, d\theta ,$$

$$= c[\nu,\omega] \left[\prod_{j=1}^{m} [b(y_j)]^{t_j} \right] . \int a(\theta)^{\nu+\Sigma t_j} \, e^{[\omega+\Sigma t_j s(y_j)]'r(\theta)} \, d\theta ,$$

where $\Sigma = \sum_{j=1}^{m}$ and $\Pi = \prod_{j=1}^{m}$ throughout section 4.2.3.

Hence,

$$\Phi_{\underline{l}}(\underline{t}) = \frac{c[\nu,\omega] . \Pi \, [b(y_j)]^{t_j}}{c[\nu+\Sigma t_j, \omega+\Sigma t_j s(y_j)]} \qquad (4.2.6)$$

Hence the cumulant generating function of $\underline{l}$ is given by,

$$K_{\underline{l}}(\underline{t}) = \log \Phi_{\underline{l}}(\underline{t}) ,$$

$$= \log c[\nu,\omega] - \log c[\nu+\Sigma t_j, \omega+\Sigma t_j s(y_j)] + \Sigma t_j \log b(y_j) . \qquad (4.2.7)$$

The predictive distribution for $\underline{y}=(y_1,y_2,..,y_m)$

Setting $t_j=1$ for $j=1,2,..,m$ in equation (4.2.2) gives from equation (4.2.6),

$$p(\underline{y}|\underline{x}) = \frac{c[\nu,\omega] . \Pi \, b(y_j)}{c[\nu+m, \omega+\Sigma s(y_j)]} \, . \qquad (4.2.8)$$

4.2.4 The posterior moment generating function and moments of the log likelihood function at a single future value y

Setting m=1 in equation (4.2.6) and omitting the subscript 1 from l_1, t_1 and y_1 and setting s=s(y) gives,

$$\Phi_1(t) = \frac{c[\nu,\omega].\ b(y)^t}{c[\nu+t,\omega+ts]} \quad , \tag{4.2.9}$$

$$K_1(t) = \log c[\nu,\omega] - \log c[\nu+t,\omega+ts] + t.\log b(y) \ , \tag{4.2.10}$$

$$E(l) = \left[\frac{d}{dt} K_1(t)\right]_{t=0} = - \frac{d}{dt} \log c[\nu+t,\omega+ts] + \log b(y) \ , \text{ and} \tag{4.2.11}$$

$$(K_r)_1 = \left[\frac{d^{(r)}}{dt^{(r)}} K_1(t)\right]_{t=0} = - \left[\frac{d^{(r)}}{dt^{(r)}} \log c[\nu+t,\omega+ts]\right]_{t=0} + \log b(y) \tag{4.2.12}$$

for $r \geq 2$, and hence the posterior central moments of l, $(\mu_r')_1$, can be obtained from the cumulants of l, $(K_r)_1$, using equation (1.4.3).

The predictive distribution for a single future value y

Setting t=1 in equation (4.2.9) gives the predictive distribution for a single future observation y given the random sample $\underline{x}$,

$$p(y|\underline{x}) = \frac{c[\nu,\omega].b(y)}{c[\nu+1,\omega+s]} \quad . \tag{4.2.13}$$

4.3 Gamma populations

4.3.1 Introduction

Suppose that in populations Π_1 and Π_2 the random variable X has Gamma probability density functions, $Ga(k_1,\theta_1)$ and $Ga(k_2,\theta_2)$, respectively, where k_1 and k_2 are known but θ_1 and θ_2 are unknown parameters. Here the parameter vector is $\theta=(\theta_1,\theta_2)$ and the i^{th} population density for i=1,2 is given by

$$p_i(y|\theta) = \frac{\theta_i^{k_i}\, y^{k_i-1}\, e^{-\theta_i y}}{\Gamma(k_i)} \quad . \tag{4.3.1}$$

Estimates of P

An 'estimative' estimate of P, P_e, is given by substituting the maximum likelihood point estimates of θ_i, $k_i/\bar{x}_i$, for i=1,2, into equation (4.3.1) and then substituting the resulting $p_i(y|\theta)$, for i=1,2, into equation (1.2.1).

The 'predictive' estimate of P, P_p, is given by substituting the predictive density $p_i(y|\underline{x}_1,\underline{x}_2)$, for i=1,2, into equation (1.2.2), where $p_i(y|\underline{x}_1,\underline{x}_2)$ is given by equation (4.3.6) and (g,h) are the parameters of the posterior distribution of θ_i given the random sample $\underline{x}_i$ for i=1,2. Assuming vague prior knowledge about parameter θ_i then $p_i(y|\underline{x}_1,\underline{x}_2)$ is given, for i=1,2, by substituting $(n_ik_i,n_i\bar{x}_i)$ for (g,h) in equation (4.3.6), i.e.

$$p_i(y|\underline{x}_1,\underline{x}_2) = \frac{\Gamma[k_i(n_i+1)].(n_i\bar{x}_i)^{k_in_i}\, y^{k_i-1}}{\Gamma(k_in_i)\,\Gamma(k_i).(y+n_i\bar{x}_i)^{k_i(n_i+1)}} \quad . \tag{4.3.2}$$

Credibility interval for P

The posterior mean and central moments of $l_i = \log p_i(y|\theta)$ about, for i=1,2, are obtained by substituting for (g,h) the parameters of the posterior distribution of θ_i into the results in section 4.3.3. If vague prior information about θ_i was assumed then $(n_ik_i,n_i\bar{x}_i)$ is substituted for (g,h) in the results in section 4.3.3. The posterior moments of $L=l_1-l_2$ are then found using equations (1.4.1) and (1.4.2). The posterior distributions of L and P are then approximated as explained in section 4.1.

4.3.2 The posterior moment generating function of the log likelihood functions at multiple new values from a Gamma population

Let $p(x|\theta) = \dfrac{\theta^k x^{k-1} e^{-\theta x}}{\Gamma(k)}$, where k is fixed and known.

Comparison with (4.1.1) gives $a(\theta) = \theta^k$, $b(x) = \dfrac{x^{k-1}}{\Gamma(k)}$, $s(x)=x$ and $r(\theta) = -\theta$.

If a natural conjugate prior distribution $p(\theta) = Ga(g_0,h_0)$ is assumed for θ, then given the random sample $\underline{x}=(x_1,x_2,..,x_n)$ from the Gamma $Ga(k,\theta)$ distribution, θ has a Gamma posterior distribution given by,

$p(\theta|\underline{x}) = Ga(g_0+nk,h_0+n\bar{x})$

(Aitchison and Dunsmore, 1975, Chapter 2), where $\bar{x}=\Sigma x_i/n$.

If vague prior information about θ is assumed then $p(\theta|\underline{x}) = Ga(nk,n\bar{x})$.

For a general result let $p(\theta|\underline{x}) = Ga(g,h)$, i.e.

$$p(\theta|\underline{x}) = \frac{h^g \theta^{g-1} e^{-h\theta}}{\Gamma(g)} .$$

Comparison with (4.2.5) gives $\nu = \dfrac{g-1}{k}$, $\omega = h$ and $c[\nu,\omega] = \dfrac{w^{k\nu+1}}{\Gamma(k\nu+1)}$.

Hence from (4.2.6),

$$\Phi_1(\underline{t}) = \frac{h^g \Gamma(g+k\Sigma t_j) \Pi y_j^{t_j(k-1)}}{[\Gamma(k)]^{\Sigma t_j} \Gamma(g) [h+\Sigma t_j y_j]^{g+k\Sigma t_j}} , \qquad (4.3.3)$$

and from (4.2.8),

$$p(\underline{y}|\underline{x}) = \frac{h^g \Gamma(g+km) \Pi y_j^{(k-1)}}{[\Gamma(k)]^m \Gamma(g) [h+\Sigma y_j]^{g+km}} . \qquad (4.3.4)$$

4.3.3 The posterior moment generating function of the log likelihood function at a single future value y

From equation (4.2.9),

$$\Phi_1(t) = \frac{y^{t(k-1)}\, h^g\, \Gamma(g+tk)}{[\Gamma(k)]^t\, \Gamma(g)\, (h+ty)^{g+tk}} \quad , \tag{4.3.5}$$

and from (4.2.13) the predictive distribution of y is the Inverse-Beta probability density function, given by

$$p(y|\underline{x}) = \frac{y^{k-1}\, h^g\, \Gamma(g+k)}{\Gamma(k)\, \Gamma(g)\, (h+y)^{g+k}} = \text{InBe}[k,g,h] \; . \tag{4.3.6}$$

Hence from (4.3.5),

$$\log \Phi_1(t) = t(k-1).\log y + g.\log h + \log[\Gamma(g+tk)] - t.\log[\Gamma(k)] - \log[\Gamma(g)] - (g+tk).\log[h+ty] \; ,$$

$$\frac{d \log \Phi_1(t)}{dt} = (k-1).\log y + k.\Psi(g+tk) - \log[\Gamma(k)] - k.\log[h+ty] - \frac{y(g+tk)}{(h+ty)}$$

$$\frac{d^{(r)} \log \Phi_1(t)}{dt^{(r)}} = k^r.\Psi^{(r-1)}(g+tk) + \frac{(-1)^{r-1} y^{r-1} (r-2)!.[(h+ty)k-(gy-hk)(r-1)]}{(h+ty)^r}$$

for $r \geq 2$.

Hence,

$$(K_r)_1 = \left[\frac{d^{(r)} \log \Phi_1(t)}{dt^{(r)}}\right]_{t=0} = k^r.\Psi^{(r-1)}(g) + (-1)^{r-1} \frac{y^{r-1}}{h^r} (r-2)![rhk-(r-1)gy] \tag{4.3.7}$$

for $r \geq 2$, where $(K_r)_1$ is the r^{th} cumulant of 1.

Equation (1.4.3) gives a recurrence relationship between the moments of a random variable about its mean and its cumulants, hence using the resulting equations (1.4.4),

$$E(1) = (K_1)_1 = (k-1).\log y + k.\Psi(g) - \log[\Gamma(k)] - k.\log h - \frac{gy}{h} \; ,$$

$$V(1) = (\mu_2')_1 = (K_2)_1 = k^2.\Psi^{(1)}(g) + (-1).\frac{y}{h^2}[2hk-gy] \; ,$$

$$(\mu_3')_1 = (K_3)_1 = k^3.\Psi^{(2)}(g) + (-1)^2.\frac{y^2}{h^3}[3hk-2gy] \; , \text{ and} \tag{4.3.8}$$

$$(\mu_4')_1 = (K_4)_1 + 3(K_2)_1^2 \; ,$$

where $(K_4)_1 = k^4.\Psi^{(3)}(g) + (-1)^3.2!.\frac{y^3}{h^4}[4hk-3gy]$.

4.3.4 Special case of Exponential populations, $Ex(\Theta_1)$ and $Ex(\Theta_2)$

Since $Ex(\Theta) = Ga(1,\Theta)$, the Exponential distribution $Ex(\Theta)$ is just a special case of the Gamma distribution $Ga(k,\Theta)$ where k=1. Hence the results in section 4.3.3 reduce to,

$$E(1) = \Psi(g) - \log h - \frac{g}{h}.y \ ,$$

$$V(1) = (\mu_2')_1 = \Psi^{(1)}(g) - \frac{y}{h^2}[2h-yg] \ ,$$

$$(\mu_3')_1 = (K_3)_1 = \Psi^{(2)}(g) + \frac{y^2}{h^3}[3h-2yg] \ , \text{ and} \qquad (4.3.9)$$

$$(\mu_4')_1 = (K_4)_1 + 3(K_2)_1^2 \ ,$$

where $(K_4)_1 = \Psi^{(3)}(g) - 2\frac{y^3}{h^4}[4h-3yg]$.

Note that V(1) is minimised when y takes the value h/g.

For vague prior information about Θ substitute $(n,n\bar{x})$ for (g,h).

Exponential populations example (special case of Gamma populations)

The following example is taken from Aitchison and Dunsmore, 1975, Ch11.

Sample 1 : 47,17,32,19 from $Ex(\theta_1)$, population Π_1

Sample 2 : 75,31 from $Ex(\theta_2)$, population Π_2

New Cases: 40,70 unknown population Π_1 or Π_2

Assume fixed chosen sample sizes i.e. the data were not collected from a 'naturally occurring experiment', and assume equal prior probabilities $q_1=q_2$.

'Estimative' estimate of P

$$P_e = \frac{q_1 p_1(y|\hat{\theta})}{q_1 p_1(y|\hat{\theta})+q_2 p_2(y|\hat{\theta})} = \frac{Ex(\hat{\theta}_1)}{Ex(\hat{\theta}_1)+Ex(\hat{\theta}_2)}$$

where for i=1,2, $Ex(\hat{\theta}_i) = \hat{\theta}_i . e^{-\hat{\theta}_i y}$

and the maximum likelihood estimate of θ_i is $\hat{\theta}_i = 1/\bar{x}_i$.

'Predictive' estimate of P

$$P_p = \frac{q_1 p_1(y|\underline{x}_1,\underline{x}_2)}{q_1 p_1(y|\underline{x}_1,\underline{x}_2)+q_2 p_2(y|\underline{x}_1,\underline{x}_2)} = \frac{InBe(1,n_1,n_1\bar{x}_1)}{InBe(1,n_1,n_1\bar{x}_1)+InBe(1,n_2,n_2\bar{x}_2)}$$

assuming vague prior knowledge about θ_1 and θ_2, where for i=1,2,

$$InBe(1,n_i,n_i\bar{x}_i) = \frac{n_i.(n_i\bar{x}_i)^{n_i}}{(n_i\bar{x}_i+y)^{n_i+1}}$$

in the notation of Aitchison and Dunsmore (1975, Ch2).

Posterior moments of the log likelihood $l_i = \log p_i(y|\Theta)$

$E(l_i) = \Psi(n_i) - \log(n_i\bar{x}_i) - y/\bar{x}_i$

$V(l_i) = \Psi^{(1)}(n_i) - y.[2\bar{x}_i - y]/(n_i\bar{x}_i^2)$

New case 1: x=40

$$P_e = \frac{0.008652}{0.008652+0.008871} = 0.49376$$

$$P_p = \frac{0.007820}{0.007820+0.007208} = 0.51991$$

Hence $E(l_1) = \Psi(4) - \log 115 - 40/(28\times75)$

$$= -C + \sum_{j=1}^{3} \frac{1}{j} - \log 115 - \frac{40}{28\times75}$$ using the results of Appendix 3,

$$= -4.88013$$

Similarly $E(l_2) = -4.995376$

$E(L) = E(l_1) - E(l_2) = 0.115246 = \mu_L$

$$V(l_1) = \Psi^{(1)}(4) - \frac{40}{4*28*75*75}.[2*28*75 - 40]$$

$$= \frac{\pi^2}{6} - \sum_{j=1}^{3} \frac{1}{j^2} - \frac{40}{4*28*75*75}.[2*28*75 - 40]$$

$$= 0.072103$$

Similarly $V(l_2) = 0.175016$

Hence $V(L) = V(l_1) + V(l_2) = 0.247119 = \sigma_L^2$

and an approximate 95% credibility interval for L is given by

$[\mu_L - 1.96*\sigma_L , \mu_L + 1.96*\sigma_L] = [-0.85909, 1.08958] = [L_1, L_2]$

and an approximate 95% credibility interval for P is given by

$[(1+e^{-L_1})^{-1}, (1+e^{-L_2})^{-1}] = [0.29753, 0.7483]$

New case 2: x=70

Similarly for new case 2,

$P_e = 0.37698$, $P_p = 0.43922$,

$E(l_1) = -5.9236$, $E(l_2) = -5.5614$ and $E(L) = -0.3622$

$V(l_1) = 0.435052$, $V(l_2) = 0.388615$ and $V(L) = 0.823667$

and an approximate 95% credibility interval for L is given by

[-2.14102,1.41662]

and an approximate 95% credibility interval for P is given by

[0.1052,0.80481].

4.4 Binonial populations

4.4.1 Introduction

Suppose that in populations Π_1 and Π_2 the random variable X has Binomial probability density functions, $Bi(k,\theta_1)$ and $Bi(k,\theta_2)$, respectively, where k is known but θ_1 and θ_2 are unknown parameters. Here the parameter vector is $\theta=(\theta_1,\theta_2)$ and the i^{th} population density for i=1,2 is given by

$$p_i(y|\theta) = \begin{bmatrix} k \\ y \end{bmatrix} \theta_i^{\,y}(1-\theta_i)^{k-y} \quad . \qquad (4.4.1)$$

where $\begin{bmatrix} k \\ y \end{bmatrix} = \frac{k!}{y!(k-y)!}$, throughout section 4.4.

Estimates of P

An 'estimative' estimate of P, P_e, is given by substituting the maximum likelihood point estimates of θ_i, $\bar{x}_i/k$, for i=1,2, into equation (4.4.1) and then substituting the resulting $p_i(y|\theta)$, for i=1,2, into equation (1.2.1).

The 'predictive' estimate of P, P_p, is given by substituting the predictive density $p_i(y|\underline{x}_1,\underline{x}_2)$, for i=1,2, into equation (1.2.2), where $p_i(y|\underline{x}_1,\underline{x}_2)$ is given by equation (4.4.6) and (g,h) are the parameters of the posterior distribution of θ_i given the random sample $\underline{x}_i$ for i=1,2. Assuming vague prior knowledge about parameter θ_i then $p_i(y|\underline{x}_1,\underline{x}_2)$ is given, for i=1,2, by substituting $(n_i\bar{x}_i, n_i(k-\bar{x}_i))$ for (g,h) in equation (4.4.6), i.e.

$$p_i(y|\underline{x}_1,\underline{x}_2) = \begin{bmatrix} k \\ y \end{bmatrix} \cdot \frac{B(n_i\bar{x}_i+y, n_i(k-\bar{x}_i)+k-y)}{B(n_i\bar{x}_i, n_i(k-\bar{x}_i))} \quad . \qquad (4.4.2)$$

Credibility interval for P

The posterior mean and central moments of $l_i = \log p_i(y|\theta)$, for i=1,2, are obtained by substituting for (g,h) the parameters of the posterior distribution of θ_i into the results in section 4.4.3. If vague prior information about θ_i was assumed then $(n_i\bar{x}, n_i(k-\bar{x}_i))$ is substituted for (g,h) in the results in section 4.4.3. The moments of $L=l_1-l_2$ are then found using equations (1.4.1) and (1.4.2). The posterior distributions of L and P are then approximated as explained in section 4.1.

4.4.2 The joint posterior moment generating function of the log likelihood function at multiple new values from a Binomial population

Let $p(x|\theta) = \binom{k}{x} \theta^x (1-\theta)^{k-x}$, where k is fixed and known.

Comparison with (4.1.1) gives $a(\theta) = (1-\theta)^k$, $b(x) = \binom{k}{x}$, $s(x) = x$, and $r(\theta) = \log[\theta/(1-\theta)]$.

If a natural conjugate prior distribution $p(\theta) = Be(g_0,h_0)$ is assumed for θ, then given the random sample $\underline{x}=(x_1,x_2,..,x_n)$ from the Binomial, $Bi(k,\theta)$, distribution, θ has a Beta posterior distribution given by

$p(\theta|\underline{x}) = Be(g_0+n\bar{x},h_0+nk-n\bar{x})$,

(Aitchison and Dunsmore, 1975, Chapter 2), where $\bar{x}=\Sigma x_j/n$.

If vague prior information about θ is assumed then $p(\theta|\underline{x}) = Be(n\bar{x},nk-n\bar{x})$.

For a general result let $p(\theta|\underline{x}) = Be(g,h)$, i.e.

$$p(\theta|\underline{x}) = \frac{\theta^{g-1}.(1-\theta)^{h-1}}{B(g,h)} .$$

Comparison with (4.2.5) gives $\nu = (g+h-2)/k$,

$\omega = g-1$ and $c[\nu,\omega] = [B(\omega+1,k\nu-\omega+1]^{-1}$.

Hence from (4.2.6),

$$\Phi_1(\underline{t}) = \frac{B(g+\Sigma t_j y_j, h+k\Sigma t_j - \Sigma t_j y_j)}{B(g,h)} . \prod_{j=1}^{m} \binom{k}{y_j}^{t_j} , \qquad (4.4.3)$$

where $\Sigma = \sum_{i=1}^{m}$ throughout section 4.4.2.

Also from (4.2.8),

$$p(\underline{y}|\underline{x}) = \frac{B(g+\Sigma y_j, h+km-\Sigma y_j)}{B(g,h)} . \prod_{j=1}^{m} \binom{k}{y_j} . \qquad (4.4.4)$$

4.4.3 The posterior moment generating function of the log likelihood function at a single future value y

From equation (4.2.9),

$$\Phi_1(t) = \binom{k}{y} t \frac{B(g+ty,h+t(k-y))}{B(g,h)} . \qquad (4.4.5)$$

and from (4.2.13) the predictive distribution of y is the Beta-Binomial probability function given by,

$$p(y|\underline{x}) = \binom{k}{y} \frac{B(g+y,h+k-y)}{B(g,h)} = BeBi(k,g,h) . \qquad (4.4.6)$$

Hence from (4.4.5),

$$\log \Phi_1(t) = t.\log\binom{k}{y} + \log[\Gamma(g+ty)] + \log[\Gamma(h+t(k-y))] - \log[\Gamma(g+h+tk)] - \log[B(g,h)]$$

$$\frac{d}{dt}\log \Phi_1(t) = \log\binom{k}{y} + y.\Psi(g+ty) + (k-y).\Psi[h+t(k-y)] - k.\Psi(g+h+tk)$$

$$\frac{d^{(r)}}{dt^{(r)}}\log \Phi_1(t) = y^r.\Psi^{(r-1)}(g+ty) + (k-y)^r\Psi^{(r-1)}[h+t(k-y)] - k^r\Psi^{(r-1)}(g+h+tk)$$

for $r \geq 2$.

Hence,

$$(K_r)_1 = \left[\frac{d^{(r)}}{dt^{(r)}} \log \Phi_1(t)\right]_{t=0} = y^r.\Psi^{(r-1)}(g) + (k-y)^r\Psi^{(r-1)}(h) - k^r\Psi^{(r-1)}(g+h) , \qquad (4.4.7)$$

for $r \geq 2$, where $(K_r)_1$ is the r^{th} cumulant of 1.

Equation (1.4.3) gives a recurrence relationship between the moments of a random variable about its mean and its cumulants, hence using the resulting equations (1.4.4),

$$E(1) = (K_1)_1 = \log\binom{k}{y} + y.\Psi(g) + (k-y).\Psi(h) - k.\Psi(g+h)$$

$$V(1) = (\mu_2')_1 = (K_2)_1 = y^2.\Psi^{(1)}(g) + (k-y)^2.\Psi^{(1)}(h) - k^2.\Psi^{(1)}(g+h)$$

$$(\mu_3')_1 = (K_3)_1 = y^3.\Psi^{(2)}(g) + (k-y)^3.\Psi^{(2)}(h) - k^3.\Psi^{(2)}(g+h) \qquad (4.4.8)$$

$$(\mu_4')_1 = (K_4)_1 + 3(K_2)_1^2$$

where $(K_4)_1 = y^4.\Psi^{(3)}(g) + (k-y)^4.\Psi^{(3)}(h) - k^4.\Psi^{(3)}(g+h)$

4.5 Poisson populations

4.5.1 Introduction

Suppose that in populations Π_1 and Π_2 the random variable X has Poisson probability density functions, $Po(\theta_1)$ and $Po(\theta_2)$, respectively, where θ_1 and θ_2 are unknown parameters. Here the parameter vector is $\theta=(\theta_1,\theta_2)$ and the i^{th} population density for i=1,2 is given by

$$p_i(y|\theta) = \frac{e^{-\theta_i}\,\theta_i^{\,y}}{y!} \quad . \qquad (4.5.1)$$

Estimates of P

An 'estimative' estimate of P, P_e, is given by substituting the maximum likelihood point estimates of θ_i, $\bar{x}_i$, for i=1,2, into equation (4.5.1) and then substituting the resulting $p_i(y|\theta)$, for i=1,2, into equation (1.2.1). The 'predictive' estimate of P, P_p, is given by substituting the predictive density $p_i(y|\underline{x}_1,\underline{x}_2)$, for i=1,2, into equation (1.2.2), where $p_i(y|\underline{x}_1,\underline{x}_2)$ is given by equation (4.5.6) and (g,h) are the parameters of the posterior distribution of θ_i given the random sample $\underline{x}_i$ for i=1,2. Assuming vague prior knowledge about parameter θ_i then $p_i(y|\underline{x}_1,\underline{x}_2)$ is given, for i=1,2, by substituting $(n_i\bar{x}_i,n_i)$ for (g,h) in equation (4.5.6), i.e.

$$p_i(y|\underline{x}_1,\underline{x}_2) = \frac{(n_i\bar{x}_i+y-1)!}{y!\,(n_i\bar{x}_i-1)!}\, n_i^{\,n_i\bar{x}_i}(n_i+1)^{-(n_i\bar{x}_i+y)} \quad . \qquad (4.5.2)$$

Credibility interval for P

The posterior mean and central moments of $l_i = \log p_i(y|\theta)$, for i=1,2, are obtained by substituting for (g,h) the parameters of the posterior distribution of θ_i into the results in section 4.5.3. If vague prior information about θ_i was assumed then $(n_i\bar{x},n_i)$ is substituted for (g,h) in the results in section 4.5.3. The moments of $L=l_1-l_2$ are then found using equations (1.4.1) and (1.4.2). The posterior distributions of L and P are then approximated as explained in section 4.1.

4.5.2 The joint posterior moment generating function of the log likelihood functions at multiple new values from a Poisson population

Let $p(x|\Theta) = \dfrac{e^{-\Theta}\,\Theta^{x}}{x!}$.

Comparison with (4.1.1) gives $a(\Theta) = e^{-\Theta}$, $b(x) = 1/x!$, $s(x) = x$ and $r(\Theta) = \log\Theta$.

If a natural conjugate prior distribution $p(\Theta) = Ga(g_0,h_0)$ is assumed for Θ then given the random sample $\underline{x}=(x_1,x_2,..,x_n)$ from the Poisson, $Po(\Theta)$, distribution, Θ has a Gamma posterior distribution given by,

$p(\Theta|\underline{x}) = Ga(g_0+n\bar{x},h_0+n)$

(Aitchison and Dunsmore, 1975, Chapter 2), where $\bar{x}=\Sigma x_j/n$.

If vague prior information about Θ is assumed then $p(\Theta|\underline{x}) = Ga(n\bar{x},n)$.

For a general result let $p(\Theta|\underline{x}) = Ga(g,h)$, i.e.

$$p(\Theta|\underline{x}) = \frac{h^{g}.\Theta^{g-1}.e^{-h\Theta}}{\Gamma(g)}$$

Comparison with (4.2.5) gives $\nu = h$, $\omega = g-1$ and $c[\nu,\omega] = \dfrac{\nu^{\omega+1}}{\Gamma(\omega+1)}$.

Hence from (4.2.6),

$$\Phi_1(\underline{t}) = \frac{\Gamma(g+\Sigma t_j y_j)\; h^{g}}{\Gamma(g)\;[h+\Sigma t_j]^{g+\Sigma t_j y_j}\;\Pi(y_j!)^{t_j}} \quad . \qquad (4.5.3)$$

where $\Sigma = \sum_{j=1}^{m}$ and $\Pi = \prod_{j=1}^{m}$ throughout section 4.5.2.

Also from (4.2.8),

$$p(\underline{y}|\underline{x}) = \frac{\Gamma(g+\Sigma y_j)\;\; h^{g}}{\Gamma(g)\;[h+m]^{g+\Sigma y_j}\;\Pi\, y_j!} \quad . \qquad (4.5.4)$$

4.5.3 The posterior moment generating function of the log likelihood function at a single future value y.

From equation (4.2.9),

$$\Phi_1(t) = \frac{\Gamma(g+ty)\, h^{g}}{\Gamma(g)\,[h+t]^{g+ty}.y!^{t}} \quad . \tag{4.5.5}$$

Also from (4.2.13) the predictive distribution for y is the Negative-Binomial probability function, given by,

$$p(y|\underline{x}) = \frac{\Gamma(g+y)\, h^{g}}{\Gamma(g)\,[h+m]^{g+y}\, y!} = \text{NeBi}\left[g, \frac{1}{1+h}\right] \quad . \tag{4.5.6}$$

Hence from (4.5.5),

$$\log \Phi_1(t) = \log[\Gamma(g+ty)] - t.\log[\Gamma(y+1)] - \log[\Gamma(g)]$$

$$+g.\log h - (g+ty).\log[h+t] \ ,$$

$$\frac{d}{dt}\log \Phi_1(t) = y.\Psi(g+ty) - \log[\Gamma(y+1)] - y.\log[h+t] - y + \frac{(hy-g)}{(h+t)} \ ,$$

$$\frac{d^{(r)}}{dt^{(r)}}\log \Phi_1(t) = y^{r}.\Psi^{(r-1)}(g+ty) - \frac{(-1)^{r-2}.y.(r-2)!}{(h+t)^{r-1}} + \frac{(-1)^{r-1}.(hy-g).(r-1)!}{(h+t)^{r}}$$

for $r \geq 2$.

Hence,

$$(K_r)_1 = \left[\frac{d^{(r)}}{dt^{(r)}} \log \Phi_1(t)\right]_{t=0} = y^{r}.\Psi^{(r-1)}(g) + \frac{(-1)^{r-1}(r-2)!.[rhy-(r-1)g]}{h^{r}} \ , \tag{4.5.7}$$

for $r \geq 2$, where $(K_r)_1$ is the r^{th} cumulant of 1.

Equation (1.4.3) gives a recurrence relationship between the moments of a random variable about its mean and its cumulants, hence using the resulting equations (1.4.4),

$$E(1) = (K_1)_1 = y.\Psi(g) - \log[\Gamma(y+1)] - y.\log h - \frac{g}{h} \ ,$$

$$V(1) = (\mu_2')_1 = (K_2)_1 = y^2.\Psi^{(1)}(g) - \frac{1}{h^2}[2hy-g] \ ,$$

$$(\mu_3')_1 = (K_3)_1 = y^3.\Psi^{(2)}(g) + \frac{1}{h^3}[3hy-2g] \ , \tag{4.5.8}$$

$$(\mu_4')_1 = (K_4)_1 + 3(K_2)_1^2 \ ,$$

where $(K_4)_1 = y^4.\Psi^{(3)}(g) + \frac{(-1)^3.2!}{h^4}.[4hy-3g]$.

CHAPTER 5

The situation where the mixing parameters are unknown and the observations are obtained by mixture sampling

5.1 Introduction

The probability P that new observation y belongs to Π_1, given the value y, the population parameters Θ and the prior probabilities q_1 and q_2 $(=1-q_1)$ is given by equation (1.1.1). Previously it was assumed that the prior probability $q_1=p(y\in\Pi_1)$ was known (and hence $q_2=1-q_1$ was also known). Suppose however that q_1 is itself unknown and the random samples $\underline{x}_1$ and $\underline{x}_2$ were obtained by mixture sampling from the combined population distribution $q_1p_1(x|\Theta)+q_2p_2(x|\Theta)$ and then identified as belonging to Π_1 and Π_2 respectively, so that the number of observations from Π_1 and Π_2 in the random samples $\underline{x}_1$ and $\underline{x}_2$ respectively gives information about q_1.

5.2 Estimates of P

The 'estimative' estimate of P, P_e, is obtained by substituting into $p_i(y|\Theta)$ point estimates $(\hat{\Theta},\hat{q}_1)$ for the population parameters (Θ,q_1), for i=1,2. The resulting estimates of $p_i(y|\Theta)$, for i=1,2, are then substituted into equation (1.1.1), giving,

$$P_e = \frac{\hat{q}_1.p_1(y|\hat{\Theta})}{\hat{q}_1.p(y|\hat{\Theta}) + \hat{q}_2.p_2(y|\hat{\Theta})} \qquad (5.2.1)$$

The maximum likelihood estimate of q_i is given by $n_i/(n_1+n_2)$, for i=1,2.

The 'predictive' estimate of P is given by $P_p = p(y\in\Pi_1|y,n_1,n_2,\underline{x}_1,\underline{x}_2)$.
From a fully Bayes approach assuming prior independence of Θ and q_1 and assuming that $p(y|y\in\Pi_i,q_1,\Theta) = p(y|y\in\Pi_i,\Theta)$ $[=p_i(y|\Theta)]$, for i=1,2, and that $p(y\in\Pi_i|q_1,\Theta) = p(y\in\Pi_i|q_1)$, for i=1,2, then it can be shown that,

$$P_p = \frac{p(y\in\Pi_1|n_1,n_2).p(y|y\in\Pi_1,\underline{x}_1,\underline{x}_2)}{p(y\in\Pi_1|n_1,n_2).p(y|y\in\Pi_1,\underline{x}_1,\underline{x}_2) + p(y\in\Pi_2|n_1,n_2).p(y|y\in\Pi_2,\underline{x}_1,\underline{x}_2)}$$

$$= \frac{p(y\in\Pi_1|n_1,n_2).p_1(y|\underline{x}_1,\underline{x}_2)}{p(y\in\Pi_1|n_1,n_2).p_1(y|\underline{x}_1,\underline{x}_2) + p(y\in\Pi_2|n_1,n_2).p_2(y|\underline{x}_1,\underline{x}_2)} \qquad (5.2.2)$$

(Aitchison and Dunsmore, 1975, Ch11).
Assuming vague prior knowledge about q_1 then $p(y\in\Pi_i|n_1,n_2) = n_i/(n_1+n_2)$.

5.3 Posterior distribution and credibility interval for P

The true probability, P, that new observation y belongs to population Π_1, defined in equation (1.1.1), can be written as,

$$P = (1+e^{-L})^{-1} \qquad (5.3.1)$$

where $L = L + L_q$,

$L = l_1 - l_2 = \log[p_1(y|\Theta)] - \log[p_2(y|\Theta)]$,

and $L_q = l_{q_1} - l_{q_2} = \log q_1 - \log q_2 = \log[q_1/(1-q_1)]$.

If, given y, P is treated as a random variable, (a function of the random variables Θ and q_1), then the posterior distribution of P given randon samples $\underline{x}_1$ and $\underline{x}_2$, with sample sizes n_1 and n_2 respectively, can be obtained.

In order to achieve this first the posterior moments of L are obtained. The posterior distribution of L can then be approximated either using Pearson curves or a Normal approximation. The posterior distribution of P can then be obtained by transformation using equation (5.3.1).

The posterior distributions of L and q_1 will be independent provided the assumptions given in section 5.2 are made (Aitchison and Dunsmore, 1975, Ch11), and hence their cumulant generating functions are related by,

$$K_L(t) = K_{L_q}(t) + K_L(t) \quad , \qquad (5.3.2)$$

and hence, for r=1,2,..., their posterior r^{th} cumulants are related by,

$$(K_r)_L = (K_r)_{L_q} + (K_r)_L \quad . \qquad (5.3.3)$$

The cetral moments of L can be obtained from its cumulants using equations (1.4.3) and (1.4.4).

5.4 The posterior moments of L_q, the log odds for the mixing parameter q_1

Assuming a Beta $Be(g_0,h_0)$ prior distribution, $p(q_1)$, for q_1, then the posterior distribution of q_1, given random samples $\underline{x}_1$ and $\underline{x}_2$ of sizes n_1 and n_2, obtained by mixture sampling and identified as coming from populations Π_1 and Π_2 respectively, is given by,

$p(q_1|n_1,n_2) = Be(g_0+n_1,h_0+n_2)$.

If vague prior information about q_1 is assumed then,

$p(q_1|n_1,n_2) = Be(n_1,n_2)$.

For a general result let $p(q_1|n_1,n_2) = Be(g,h)$, then

$$E\left[\left[\frac{q_1}{1-q_1}\right]^t\right] = \int_0^1 \left[\frac{q_1}{1-q_1}\right]^t \frac{q_1^{g-1}(1-q_1)^{h-1}}{B(g,h)}\,dq_1$$

$$= \frac{B(g+t,h-t)}{B(g,h)}$$

$$= \frac{\Gamma(g+t).\Gamma(h-t)}{\Gamma(g).\Gamma(h)}$$

Let $L_q = \log[q_1/(1-q_1)]$ then

$$\Phi_{L_q}(t) = E\left[e^{tL_q}\right] = E\left[\left[\frac{q_1}{1-q_1}\right]^t\right]$$

$$\log \Phi_{L_q}(t) = \log[\Gamma(g+t)] + \log[\Gamma(h-t)] - \log[\Gamma(g)] - \log[\Gamma(h)]$$

$$(K_r)_{L_q} = \left[\frac{d^{(r)}}{dt^{(r)}} \log \Phi_1(t)\right]_{t=0} = \Psi^{(r-1)}(g) + (-1)^r.\Psi^{(r-1)}(h) \qquad (5.4.1)$$

for $r \geq 1$, where $(K_r)_{L_q}$ is the r^{th} cumulant of L_q.

Equation (1.4.3) gives a recurrence relationship between the moments of a random variable about its mean and its cumulants, hence using the resulting equations (1.4.4),

$$E(L_q) = (K_1)_{L_q} = \Psi(g) - \Psi(h)$$

$$V(L_q) = (\mu_2')_{L_q} = (K_2)_{L_q} = \Psi^{(1)}(g) + \Psi^{(1)}(h)$$

$$(\mu_3')_{L_q} = (K_3)_{L_q} = \Psi^{(2)}(g) - \Psi^{(2)}(h) \qquad (5.4.2)$$

$$(\mu_4')_{L_q} = (K_4)_{L_q} + 3(K_2^2)_{L_q}$$

where $(K_4)_{L_q} = \Psi^{(3)}(g) + \Psi^{(3)}(h)$

For vague prior information about q_1 substitute (n_1,n_2) for (g,h) in the above results. Hence, using the results from Appendix 3,

$$E(L_q) = \Psi(n_1) - \Psi(n_2) = \sum_{j=n_2}^{n_1-1} \frac{1}{j} \quad , \text{ for } n_1 \rangle n_2,$$

$$V(L_q) = \Psi^{(1)}(n_1) + \Psi^{(1)}(n_2) = \frac{\pi^2}{3} - \sum_{j=1}^{n_1-1} \frac{1}{j^2} - \sum_{j=1}^{n_2-1} \frac{1}{j^2} \quad ,$$

$$(\mu_3')_{L_q} = (K_3)_{L_q} = \Psi^{(2)}(n_1) - \Psi^{(2)}(n_2) = 2.\sum_{j=n_2}^{n_1-1} \frac{1}{j^3} \quad , \text{ for } n_1 \rangle n_2, \tag{5.4.3}$$

$$(\mu_4')_{L_q} = (K_4)_{L_q} + 3(K_2)_{L_q}^2 \quad ,$$

where $(K_4)_{L_q} = \Psi^{(3)}(n_1) + \Psi^{(3)}(n_2) = 3!.\left[2.\zeta(4) - \sum_{j=1}^{n_1-1} \frac{1}{j^4} - \sum_{j=1}^{n_1-1} \frac{1}{j^4}\right]$,

and $\zeta(4) = \pi^4/90 = 1.08232323$ is the Zeta function evaluated at 4.

An alternative approach is to find the posterior distribution of $Q = q_1/(1-q_1)$ and $L_q = \log[q_1/(1-q_1)]$ where, for a general result, q_1 has a $Be(g,h)$ posterior distribution.

Hence Q will then have an Inverted-Beta $InBe(g,h,1)$ posterior distribution with probability density function given by,

$$p_Q(r) = \frac{1}{B(g,h)}.r^{g-1}(1+r)^{-(g+h)} \qquad \text{for } r \rangle 0, \tag{5.4.4}$$

and L_q has the following probability density function,

$$p_{L_q}(s) = \frac{1}{B(g,h)}.e^{gs}(1+e^s)^{-(g+h)} \qquad \text{for } -\infty \langle s \langle +\infty. \tag{5.4.5}$$

The posterior moments of Q and L_q could then be directly obtained from their posterior probability density functions. For vague prior information about q_1 substitute (n_1,n_2) for (g,h) in the above results.

5.5 Example

In the example in section 4.3.4 the data were assumed to have been obtained by seperate sampling and equal prior probabilities were assumed. Assume instead that the data were obtained by mixture sampling and subsequently identified as belonging to Π_1 or Π_2, and that the mixing probability q_1 is unknown. Assume a vague prior for q_1.

New case 1 : x=40

$n_1=4$ and $n_2=2$

$$P_e = \frac{0.008652\times(4/6)}{0.008652\times(4/6) + 0.008871\times(2/6)} = 0.661089$$

$$P_p = \frac{0.007820\times(4/6)}{0.007820\times(4/6) + 0.007208\times(2/6)} = 0.684524$$

$E(L) = 0.115246$ and $V(L) = 0.247119$ where

$L=l_1-l_2=\log[p_1(y|\Theta)]-\log[p_2(y|\Theta)]$

$$E(L_q) = \sum_{j=2}^{3} \frac{1}{j} = 0.83333 \quad \text{where } L_q=\log[q_1/(1-q_1)]$$

$$V(L_q) = \frac{\pi^2}{3} - \sum_{j=1}^{3} \frac{1}{j^2} - \sum_{j=1}^{1} \frac{1}{j^2} = 0.92876$$

Hence $E(L) = 0.115246 + 0.83333 = 0.948576 = \mu_L$ and

$V(L) = 0.247119 + 0.92876 = 1.17588 = \sigma_L^2$

Hence an approximate 95% credibility interval for L is given by

$[\mu_L-1.96\sigma_L , \mu_L+1.96\sigma_L] = [-1.17681 , 3.07396] = [L_1,L_2]$

and an approximate 95% credibility interval for P is given by

$[(1+e^{-L_1})^{-1} , (1+e^{-L_2})^{-1}] = [0.23563 , 0.95581]$

New case 2 : x=70

$P_e = 0.54755$ and $P_p = 0.610358$

$E(L) = E(L)+E(L_q) = -0.3622+0.83333 = 0.47113 = \mu_L$

$V(L) = V(L)+V(L_q) = 0.823667+0.92876 = 1.75243 = \sigma_L^2$

A 95% credibility interval for L is given by [-2.12350,3.06576] and a 95% credibility interval for P is given by [0.10683,0.95546].

In conclusion, by comparison with the results of section 4.3.4, it can be seen that the uncertainty about the mixing parameters leads to a much greater posterior variance for L resulting in a much wider credibility interval for P.

CHAPTER 6

The situation where the new observation belongs to one of more than two populations

6.1 Introduction

Suppose that individuals belong to one of I distinct populations Π_i, for i=1,2,..,I. Independent random samples, $\underline{x}_i = (x_{ij}, j=1,n_i)$, from population Π_i for i=1,2,..,I are to be used to gain information concerning which population a new observation y belongs to. Let q_i be the prior probability that y belongs to Π_i and let $p_i(y|\Theta)$ be the ith population density, given the population parameters Θ, for i=1,2,..,I. The probability P_j that y belongs to Π_j given the value y, the population parameters Θ and the prior probabilities, for j=1,2,..,I, is given by,

$$P_j = \frac{q_j p_j(y|\Theta)}{\sum_{i=1}^{I} q_i p_i(y|\Theta)} . \qquad (6.1.1)$$

6.2 Estimates of P_j

An 'estimative' estimate of P_j, denoted by P_{ej}, is obtained by substituting point estimates $\hat{\Theta}$ of the population parameter(s) Θ into (6.1.1), i.e.

$$P_{ej} = \frac{q_j p_j(y|\hat{\Theta})}{\sum_{i=1}^{I} q_i p_i(y|\hat{\Theta})} . \qquad (6.2.1)$$

A 'predictive' estimate of P_j, denoted by P_{pj}, is obtained by substituting the predictive densities $p_i(y|\underline{x}_1,\underline{x}_2,..,\underline{x}_I)$ for the population densities $p_i(y|\Theta)$, i=1,2,..,I in (6.1.1), i.e.

$$P_{pj} = \frac{q_j p_j(y|\underline{x}_1,\underline{x}_2,..,\underline{x}_I)}{\sum_{i=1}^{I} q_i p_i(y|\underline{x}_1,\underline{x}_2,..,\underline{x}_I)} . \qquad (6.2.2)$$

The 'predictive' estimate of P_j can be shown to be equal to $p(y\in\Pi_j|y,\underline{x}_1,\underline{x}_2,..,\underline{x}_I)$ from a fully Bayes approach.

6.3 Credibility interval for P_j

6.3.1 Introduction

P_j can be written as,

$$P_j = \frac{q_j z_j}{\sum_{i=1}^{I} q_i z_i} = \frac{1}{1 + R_j}, \qquad (6.3.1)$$

where $R_j = \frac{1}{q_j z_j} \sum_{\substack{i=1 \\ i \neq j}}^{I} q_i z_i$ and $z_i = p_i(y|\Theta)$ [$= e^{l_i}$ where $l_i = \log[p_i(y|\Theta)]$]

Hence approximating the posterior distribution of P_j requires approximating the posterior distribution of R_j, which in turn depends on the posterior distribution of either the likelihood functions at y i.e. $\underline{z}=(z_1,z_2,..,z_I)$ or the log likelihood functions at y i.e. $\underline{l}=(l_1,l_2,..,l_I)$.

6.3.2 Population distributions with no parameters in common.

In this case, provided independent prior distributions are assumed for the different sets of population parameters, then the posterior distributions of z_i (or equivalently l_i) for i=1,2,..,I, will be independent. Hence the moments of R_j can be obtained from the moments of z_i obtained seperately for i=1,2,..,I. The moments of z_i for Normal (unequal covariance matrices), Gamma, Binomial and Poisson population distributions are obtained in chapters 2 and 4.

Since $R_j = \frac{1}{q_j z_j} \sum_{\substack{i=1 \\ i \neq j}}^{I} q_i z_i$, and hence,

$$E(R) = (1/q_j).E(z_j^{-1}) \sum_{\substack{i=1 \\ i \neq j}}^{I} q_i E(z_i)$$

$$E(R_j^2) = (1/q_j^2).E(z_j^{-2}) \sum_{\substack{i_1=1 \\ i_1 \neq j}}^{I} \sum_{\substack{i_2=1 \\ i_2 \neq j}}^{I} q_{i_1} q_{i_2} .E(z_{i_1} z_{i_2}) \qquad (6.3.2)$$

which can be simplified, since z_{i_1} and z_{i_2} are independent for $i_1 \neq i_2$.

Similarly the higher moments of R_j can be obtained.

6.3.3 Populations with parameters in common

In this case the posterior distributions of z_i (or equivalently l_i) will not be independent and hence the posterior moments of R cannot be found from the seperate posterior moments of z_i for i=1,2,..,I. To find the posterior moments of R_j the joint posterior moments of $\underline{z}=(z_1,z_2,..,z_I)$ must be obtained, since

$$E(R_j) = \sum_{\substack{i=1 \\ i\neq j}}^{I} (q_i q_j^{-1}).E(z_i z_j^{-1})$$

$$E(R_j^2) = \sum_{\substack{i_1=1 \\ i_1\neq j}}^{I} \sum_{\substack{i_2=1 \\ i_2\neq j}}^{I} (q_{i_1} q_{i_2} q_j^{-2}).E(z_{i_1} z_{i_2} z_j^{-2}) \qquad (6.3.3)$$

and similarly for higher moments of R_j.

Consider the example of multivariate Normal opulations $No_d(\mu_i,T)$, for i=1,2,..,I, with unequal means but equal covariance matrices Σ (where $T=\Sigma^{-1}$ is the common precision matrix). In section 6.4 the joint posterior moment generating function of $\underline{l}=(l_1,l_2,..,l_I)$ is obtained. In section 6.5 the joint posterior moments of $\underline{z}=(z_1,z_2,..,z_I)$ are obtained , and in section 6.6 the joint posterior moments of $\underline{l}$ are obtained.

6.3.4 Conclusions

Once the posterior moments of R_j are found, the posterior distribution of R_j can be approximated (for example by using Pearson curves) and hence the posterior distribution of P_j can be obtained by transformation using equation (6.3.1).

Furthermore, the joint posterior moments of $(R_j;j=1,2,..,I-1)$ can be obtained similarly and their joint posterior distribution approximated. Hence the joint posterior distribution of $(P_j;j=1,2,..,I-1)$ can be obtained by multivariate transformation from $(R_j;j=1,2,..,I-1)$, where each seperate univariate transformation from R_j to P_j is given by equation (6.3.1).

6.4 The joint posterior moment generating function of the population log likelihood functions at y

The joint posterior moment generating function of the log of the likelihood functions $\underline{l}=(l_1,l_2,\ldots,l_I)$, where $l_i=\log[p_i(y|\Theta)]$ for i=1,2,..,I, at y, given the independent random samples $\underline{x}_1,\underline{x}_2,..,\underline{x}_I$, from populations $\Pi_1,\Pi_2,..,\Pi_I$ respectively, is given by

$$\Phi_{\underline{l}}(\underline{t}) = E[e^{\underline{t}'\underline{l}}] = E\Big[\prod_{i=1}^{I}[p_i(y|\Theta)^{t_i}\Big]$$

$$= \int \prod_{i=1}^{I}\Big[[p_i(y|\Theta)]^{t_i}\Big]p(\Theta|\underline{x}_1,\underline{x}_2,..,\underline{x}_I)d\Theta \quad , \qquad (6.4.1)$$

where $p(\Theta|\underline{x}_1,\underline{x}_2,..,\underline{x}_I)$ is the posterior distribution of Θ given $\underline{x}_1,\underline{x}_2,..,\underline{x}_I$.

Let Π_i be a d-dimensional multivariate Normal population, $No_d(\mu_i,T)$, with mean μ_i and precision matrix T,i.e.

$$p_i(y|\Theta) = p_i(y|\mu_i,T) = \frac{|T|^{1/2}}{(2\pi)^{d/2}}\exp[-\tfrac{1}{2}(y-\mu_i)'T(y-\mu_i)] \quad , \qquad (6.4.2)$$

for i=1,2,..I.

Let $\Theta = (\underline{\mu},T)$ have a $Normal^I$-Wishart prior distribution, $p(\underline{\mu},T)$, prior that is to observing $\underline{x}_1,\underline{x}_2,..,\underline{x}_I$, i.e.

$p(\underline{\mu},T) = No^I Wi_d(\underline{\mu}_0,\underline{n}_0,v_0,h_0)$,

where $\underline{\mu}_0=(\mu_{10},\mu_{20},..,\mu_{I0})$ and $\underline{n}_0=(n_{10},n_{20},..,n_{I0})$, defined by,

$p(\underline{\mu},T)=p(\mu_1|T).p(\mu_2|T)...p(\mu_I|T).p(T)$,

where $p(\mu_i|T)=No_d(\mu_{io},n_{io}T)$ for i=1,2,..,I and $p(T)=Wi_d(v_o,h_o)$.

Then, for $n_{i0}>0$ for i=1,2,..,I, the posterior distribution of Θ is

$$p(\Theta|\underline{x}_1,\underline{x}_2,..,\underline{x}_I) = p(\underline{\mu},T|\bar{x}_1,\bar{x}_2,..,\bar{x}_I,S)$$
$$= No^I Wi_d[\underline{\mu}_p,\underline{n}_p,v_p,h_p] ,$$

where $\underline{\mu}_p=(\mu_{1p},\mu_{2p},..,\mu_{Ip})$ and $\mu_{ip} = \frac{n_{i0}\mu_{i0}+n_i\bar{x}_i}{n_{i0}+n_i}$,

$\underline{n}_p=(n_{1p},n_{2p},..,n_{Ip})$ and $n_{ip} = n_{i0}+n_i$,

$v_p = v_0 + \sum_{i=1}^{I} n_i - I$ and

$$h_p = h_0+S+\sum_{i=1}^{I}\left[\frac{n_{i0}n_i}{n_{i0}+n_i}(\bar{x}_i-\mu_{i0})(\bar{x}_i-\mu_{i0})'\right] ,$$

where $S = \sum_{i=1}^{I} S_i$, and subscript p indicates parameters of the posterior distribution of Θ and where n_i, $\bar{x}_i$ and S_i are the sample size, the mean and the corrected sum of squares and products matrix for the i^{th} random sample $\underline{x}_i$ for i=1,2,..,I.

If vague prior information is assumed, i.e. $p(\underline{\mu},T) \propto |T|^{-(d+1)/2}$ then,
$p(\Theta|\underline{x}_1,\underline{x}_2,..,\underline{x}_I) = p(\underline{\mu},T|\underline{\bar{x}},S) = No^I Wi_d(\underline{\bar{x}},\underline{n},\Sigma n_i-I,S)$
where $\underline{\bar{x}}=(\bar{x}_1,\bar{x}_2,....,\bar{x}_I)$ and $\underline{n}=(n_1,n_2,..,n_I)$.

For a general result let

$$p(\Theta|\underline{x}_1,\underline{x}_2,..,\underline{x}_I) = p(\underline{\mu},T|\underline{\bar{x}},S) = No^I Wi_d(\underline{b},\underline{c},v,h) , \qquad (6.4.3)$$

where $\underline{b}=(b_1,b_2,....,b_I)$ and $\underline{c}=(c_1,c_2,....,c_I)$.

Hence substituting $p_i(y|\Theta)$ from (6.4.2) for i=1,2,..,I, and $p(\Theta|\underline{x}_1,\underline{x}_2,..,\underline{x}_I)$ from (6.4.3) into (6.4.1) gives,

$$\Phi_1(\underline{t}) = \int_{T>0}\int_{\mu_1\in R^d}\cdots\int_{\mu_I\in R^d}\left[\prod_{i=1}^{I}[p_i(y|\mu_i,T)]^{t_i}\right].p(\underline{\mu},T|\underline{x}_1,\underline{x}_2,..,\underline{x}_I,S)\,d\underline{\mu}\,dT ,$$

$$= \int\int\cdots\int\left[\prod_{i=1}^{I}|T|^{t_i/2}(2\pi)^{-dt_i/2}\exp[-\tfrac{1}{2}t_i(y-\mu_i)'T(y-\mu_i)]\right]$$

$$.\left[\prod_{i=1}^{I}|c_iT|^{1/2}(2\pi)^{-d/2}\exp[-\tfrac{1}{2}(\mu_i-b_i)'c_iT(\mu_i-b_i)]\right]$$

$$.\frac{|\frac{1}{2}h|^{v/2}}{\Gamma_d(v/2)}|T|^{(v-d-1)/2}\exp[-\tfrac{1}{2}trhT]\,d\mu_1\,d\mu_2....d\mu_I\,dT,$$

where $T>0$ means T is positive definate and $\Gamma_d(v/2) = \pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma[(v-d+j)/2]$.

Completing the squares in each of the μ_i and integrating out the result to the constants of Normal $No[(c_ib_i+t_iy)/(c_i+t_i),(c_i+t_i)T]$ distributions for $i=1,2,\ldots,I$ gives,

$$\Phi_1(\underline{t}) = \left[\prod_{i=1}^{I}\left[\frac{c_i}{c_i+t_i}\right]^{d/2}\right] \frac{|\frac{1}{2}h|^{v/2}}{\Gamma_d[\frac{1}{2}v]} \cdot \frac{1}{(2\pi)^{d\Sigma t_i}} \int |T|^{(v-d-1)/2} \exp[-\tfrac{1}{2}\mathrm{tr}HT]\, dT,$$

where $H = h + \sum_{i=1}^{I} \frac{c_it_i}{c_i+t_i}(b_i-y)(b_i-y)'$.

By noting that the integrand is in the form of the Wishart probability density function, T can be integrated out to the normalising constant of the Wishart probability density function giving,

$$\Phi_1(\underline{t}) = \left[\prod_{i=1}^{I}\left[\frac{c_i}{c_i+t_i}\right]^{d/2}\right] \cdot \frac{\Gamma_d[\frac{1}{2}(v+\Sigma t_i)]}{\Gamma_d[\frac{1}{2}v].\pi^{d\Sigma t_i/2}} \cdot \frac{|h|^{v/2}}{|H|^{(v+\Sigma t_i)/2}} \cdot \qquad (6.4.4)$$

Now by considering the determinant of a partitioned matrix A of size (d+I) by (d+I) with elements A_1, A_2, A_3 and A_4 given by
h, $[\alpha_1(b_1-y),\alpha_2(b_2-y),..,\alpha_I(b_I-y)]$, $[-(b_1-y),-(b_2-y),\ldots,-(b_I-y)]'$ and the I dimensional identity matrix respectively, then

$$|A| = \begin{vmatrix} A_1 & A_3 \\ A_2 & A_4 \end{vmatrix} = |A_4|.|A_1-A_3A_4^{-1}A_2| = |A_1|.|A_4-A_2A_1^{-1}A_3| ,$$

and hence $|H| = |h|.|u|$ where, (6.4.5)

$$u = \begin{vmatrix} 1+\alpha_1D_{11} & \alpha_1D_{12} & \ldots. & \alpha_1D_{1I} \\ \alpha_2D_{21} & 1+\alpha_2D_{22} & \ldots. & \alpha_2D_{2I} \\ . & . & \ldots. & . \\ . & . & \ldots. & . \\ . & . & \ldots. & . \\ \alpha_ID_{I1} & \alpha_ID_{I2} & \ldots. & 1+\alpha_ID_{II} \end{vmatrix}$$

where $\alpha_i = \frac{c_it_i}{c_i+t_i}$, for $i=1,2,..,I$, and $D_{ij}=(b_i-y)'h^{-1}(b_j-y)$

for $i=1,2,..,I$ and $j=1,2,..,I$.

Hence from (6.4.4) and (6.4.5),

$$\Phi_1(\underline{t}) = \left[\prod_{i=1}^{I}\left[\frac{c_i}{c_i+t_i}\right]^{d/2}\right] \cdot \frac{\Gamma_d[\frac{1}{2}(v+\Sigma t_i)]}{\Gamma_d[\frac{1}{2}v].\pi^{d\Sigma t_i/2}|u|^{(v+\Sigma t_i)/2}|h|^{\Sigma t_i/2}} \cdot \qquad (6.4.6)$$

6.5 The joint posterior moments of the population likelihood functions at y

Let $z_i=p_i(y|\Theta)$ be the i^{th} population likelihood function at y, for i=1,2,..,I.

Since $\Phi_1(\underline{t}) = E\left[\prod_{i=1}^{I} z_i^{t_i}\right]$,

the joint posterior moments of $\underline{z}=(z_1,z_2,....,z_I)$ can be found from $\Phi_1(\underline{t})$ by appropriate substitution for $\underline{t}$ in the results of section 6.4.

For example $E(z_1z_2z_3)$ is obtained by substituting $\underline{t}=(1,1,1,0,....,0)$ in section 6.4 giving

$$E(z_1z_2z_3) = \left[\prod_{i=1}^{3}\left[\frac{c_i}{c_i+1}\right]^{d/2}\right] \cdot \frac{\Gamma_d[\frac{1}{2}(v+3)]}{\Gamma_d[\frac{1}{2}v].\pi^{3d/2}} \cdot \frac{|h|^{v/2}}{|H|^{(v+3)/2}} ,$$

where $|H| = |h|.|u|$ and

$$u = \begin{vmatrix} 1+\alpha_1D_{11} & \alpha_1D_{12} & \alpha_1D_{13} & 0 .... 0 \\ \alpha_2D_{21} & 1+\alpha_2D_{22} & \alpha_2D_{23} & 0 .... 0 \\ \alpha_3D_{31} & \alpha_3D_{32} & 1+\alpha_3D_{33} & 0 .... 0 \\ 0 & 0 & 0 & 1 .... 0 \\ . & . & . & . .... . \\ . & . & . & . .... . \\ 0 & 0 & 0 & 0 .... 1 \end{vmatrix}$$

where $\alpha_i = \frac{c_it_i}{c_i+t_i} = \frac{c_i}{c_i+1}$ for i=1,2,3.

6.6 The joint posterior moments of the population log likelihood functions at y

In order to find the joint posterior moments of $\underline{l}=(l_1,l_2,..,l_I)$ about their mean, the following are required,

1. First find $a^{(\underline{r})} = D_{\underline{t}}^{(\underline{r})}u = \frac{d^{(r_1)}}{dt_1^{(r_1)}} \frac{d^{(r_2)}}{dt_1^{(r_2)}} .... \frac{d^{(r_I)}}{dt_I^{(r_I)}} u$

   where $\underline{r}=(r_1,r_2,..,r_I)$.

2. Then find $b^{(\underline{r})} = D_{\underline{t}}^{(\underline{r})}\log u$

3. Then find $K_{\underline{r}} = \left[D_{\underline{t}}^{(\underline{r})}\log \Phi_{\underline{l}}(\underline{t})\right]_{\underline{t}=\underline{0}}$

4. Then find $\mu_{\underline{r}}'$ from $K_{\underline{r}}$ using the results of Appendix 2.

For example for I=2 :

1. $u = [1 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_1\alpha_2(D_1D_2 - D_{12}^2)]$ where $D_1=D_{11}$ and $D_2=D_{22}$

$$D_{t_2}^{(r_2)}u = [D_2 + \alpha_2(D_1D_2 - D_{12})] \cdot D_{t_2}^{(r_2)}\alpha_2 \qquad \text{for } r_2 \geq 1$$

$$D_{t_1,t_2}^{(r_1,r_2)}u = (D_1D_2 - D_{12}) \cdot D_{t_1}^{(r_1)}\alpha_1 \cdot D_{t_2}^{(r_2)}\alpha_2 \qquad \text{for } (r_1,r_2) \geq (1,1)$$

where $D_{t_1}^{(r_1)}\alpha_1 = \dfrac{(-1)^{r_1} r_1! c_1^2}{(c_1+t_1)^{r_1+1}}$ and $\left[D_{t_1}^{(r_1)}\alpha_1\right]_{\underline{t}=\underline{0}} = \dfrac{(-1)^{r_1} r_1!}{c_1^{r_1-1}}$

Let $a_{\underline{0}}^{(\underline{r})} = [a^{(\underline{r})}]_{\underline{t}=\underline{0}}$, then

$$a_{\underline{0}}^{(r_1,r_2)} = \begin{cases} 1 & \text{if } (r_1,r_2)=(0,0) \\ \dfrac{(-1)^{r_1} r_1! D_1}{c_1^{r_1-1}} & \text{if } r_2=0 \\ \dfrac{(-1)^{r_2} r_2! D_2}{c_2^{r_2-1}} & \text{if } r_1=0 \\ \dfrac{(-1)^{r_1+r_2} r_1! r_2! (D_1D_2 - D_{12}^2)}{c_1^{r_1-1} c_2^{r_2-1}} & \text{if } (r_1,r_2)=(0,0) \end{cases} \qquad (6.6.1)$$

2. $$D_{t_2}^{(r_2)}u = \sum_{j_2=0}^{r_2-1} \binom{r_2-1}{j_2} D_{t_2}^{(j_2)}u \cdot D_{t_2}^{(r_2-j_2)}\log u$$

$$D_{t_1,t_2}^{(r_1,r_2)}u = \sum_{j_2=0}^{r_2-1} \sum_{j_1=0}^{r_1} \binom{r_2-1}{j_2} \binom{r_1}{j_1} D_{t_1,t_2}^{(j_1,j_2)}u \cdot D_{t_1,t_2}^{(r_1-j_1,r_2-j_2)}\log u$$

i.e. $$a^{(r_1,r_2)} = \sum_{j_2=0}^{r_2-1} \sum_{j_1=0}^{r_1} \binom{r_2-1}{j_2} \binom{r_1}{j_1} a^{(j_1,j_2)} \; b^{(r_1-j_1,r_2-j_2)}$$

i.e. $$b^{(r_1,r_2)} = a^{(r_1,r_2)} - \sum_{j_2=0}^{r_2-1} \sum_{j_1=0}^{r_1} \binom{r_2-1}{j_2} \binom{r_1}{j_1} a'^{(j_1,j_2)} \; b^{(r_1-j_1,r_2-j_2)} \qquad (6.6.2)$$

where $a'^{(\underline{r})} = \begin{cases} 0 & \text{if } \underline{r}=\underline{0} \\ a^{(\underline{r})} & \text{if } \underline{r}=\underline{0} \end{cases}$

Setting $\underline{t}=\underline{0}$ in equation (6.6.2) and using equation (6.6.1) gives the following results,

$$b_{\underline{0}}^{(1,0)} = D_1 \quad , \quad b_{\underline{0}}^{(2,0)} = -2\frac{D_1}{c_1} - D_1^2 \quad , \quad b_{\underline{0}}^{(3,0)} = 2D_1^3 + 6\frac{D_1^2}{c_1} + 6\frac{D_1}{c_1^2} \quad ,$$

$$b_{\underline{0}}^{(4,0)} = -6D_1^4 - 24\frac{D_1^3}{c_1} - 36\frac{D_1^2}{c_1^2} - 24\frac{D_1}{c_1^3} \quad ,$$

$$b_{\underline{0}}^{(1,1)} = -D_{12}^2 \quad ,$$

$$b_{\underline{0}}^{(2,1)} = 2D_{12}^2.\left[D_1 + \frac{1}{c_1}\right] \quad ,$$

$$b_{\underline{0}}^{(3,1)} = -6D_{12}^2.\left[D_1^2 + 2\frac{D_1}{c_1} + \frac{1}{c_1^2}\right] \quad ,$$

$$b_{\underline{0}}^{(2,2)} = -2D_{12}^2.[D_{12}^2 + 2D_1D_2] - 4D_{12}^2.\left[\frac{D_1}{c_2} + \frac{D_2}{c_1}\right] - \frac{4D_{12}^2}{c_1c_2} \quad ,$$

3.

$$\frac{d}{dt_1}\log \Phi_{l_1,l_2}(t_1,t_2) = -\frac{d}{2(c_1+t_1)} + \frac{d}{dt_1}\log \Gamma_d[\tfrac{1}{2}(v+t_1+t_2)] - \tfrac{1}{2}d.\log\pi$$
$$-\tfrac{1}{2}\log[h.u] - \tfrac{1}{2}(v+t_1+t_2).\frac{d}{dt_1}\log[h.u] \quad . \qquad (6.6.3)$$

Setting $\underline{t}=\underline{0}$ in (6.6.3) gives,

$$K_{1,0} = -\tfrac{1}{2}d\log\pi - \tfrac{1}{2}\log h - d/(2c_1) + \tfrac{1}{2}\sum_{j=1}^{d}\Psi[\tfrac{1}{2}(v-j+1)] - \tfrac{1}{2}v.b_{\underline{0}}^{(1,0)} \quad .$$

Differentiating (6.6.3) r_1 times with respect to t_1 and setting $\underline{t}=\underline{0}$ gives,

$$K_{r_1,0} = \frac{(-1)^{r_1}(r_1-1)!d}{2c_1^{r_1}} + (\tfrac{1}{2})^{r_1}\sum_{j=1}^{d}\Psi^{(r_1-1)}[\tfrac{1}{2}(v-j+1)] - \tfrac{1}{2}v.b_{\underline{0}}^{(r_1,0)} - \tfrac{1}{2}r_1.b_{\underline{0}}^{(r_1-1,0)}$$

for $r_1 \geq 2$.

Differentiating (6.6.3) r_i times with respect to t_i, for i=1,2, and setting $\underline{t}=\underline{0}$ gives

$$K_{r_1,r_2} = (\tfrac{1}{2})^{r_1+r_2}.\sum_{j=1}^{d}\Psi^{(r_1+r_2-1)}[\tfrac{1}{2}(v-j+1)] - \tfrac{1}{2}v.b_{\underline{0}}^{(r_1,r_2)} - \tfrac{1}{2}r_1.b_{\underline{0}}^{(r_1-1,r_2)}$$
$$-\tfrac{1}{2}r_2.b_{\underline{0}}^{(r_1,r_2-1)} \quad \text{for } (r_1,r_2) \geq (1,1)$$

4. Appendix 2 gives a recurrence relationship relating the moments of a p-dimensional random variable about its mean to its cumulants. Hence the moments of (l_1,l_2) about its mean can now be found, from the cumulants of (l_1,l_2) obtained above, using equation (A2.2) from Appendix 2.

CHAPTER 7

The general situation : multiple new observations from each of I populations, where the mixing parameters may be unknown

7.1 Introduction

In this chapter the general situation is considered in which observations of a d-dimensional random variable X belong to one of I populations Π_i with probability density functions, $p_i(x|\Theta)$, of known but not necessarily the same functional form, for i=1,2,..,I, dependent on some unknown parameter vector Θ which comprises the parameters from all the I populations. This allows for the possibility of common parameters between populations (as for example in the case of Normal populations with different means but equal precision matrix T, so that $\Theta=(\mu_1,\mu_2,..,\mu_I,T)$). Suppose independent random samples $x_i=(x_{ij};j=1,2,..,n_i)$ are obtained from the mixture probability density function $\Sigma q_i p_i(x|\Theta)$ and identified as belonging to Π_i for i=1,2,..,I. The mixing parameters $q=(q_1,q_2,..,q_I)$ satisfy $\Sigma q_i=1$ and may be unknown.

Interest usually lies in the relative likelihoods of a single future observation y of fixed known value belonging to Π_i, i.e. on $q_i p_i(y|\Theta)$, for i=1,2,..I. Here interest lies in the joint posterior distribution of the log likelihoods which is investigated using their posterior moment generating function.

For a general result, however, consider the joint posterior moment generating function of the log likelihoods of multiple new values $(y_{i1},y_{i2},\ldots,y_{im_i})$ from each of the I populations, Π_i for i=1,2,...,I.

The special case where $m_i=1$ and $y_{i1}=y$ for i=1,2,..,I, gives the usual situation of interest of a single future observation y.

The log likelihood of new value y_{ij} is given by,

$l_{ij} = \log q_i p_i(y_{ij}|\Theta) = \log q_i + l_{ij}$

where $l_{ij} = \log p_i(y_{ij}|\Theta)$ for j=1,2,..,m_i and i=1,2,..,I.

Let $l_i=(l_{i1},l_{i2},..,l_{im_i})'$ for i=1,2,..,I and let $l=(l_1',l_2',..,l_I')'$,be a vector of length Σm_i, with l_i and l and t_i and t similarly defined.

Let $\Phi_l(t)$ be the posterior moment generating function of l.

In section 7.2, $\Phi_l(t)$ is expressed as the product of two integrals, $\Phi(t)$ and $\Phi_1(t)$, where $\Phi(t)$ involves the mixing parameters q only and $\Phi_1(t)$, the posterior moment generating function of 1, involves the combined population parameters Θ only, provided that Θ and q are assumed to have independent prior distributions.

In section 7.3, the integral $\Phi(t)$ is obtained assuming a Dirichlet prior distribution for the mixing parameters q. The joint posterior moment generating function of the logs of the mixing parameters and also the joint posterior moments of the mixing parameters q and their logs l_q are obtained. In section 7.4 the joint posterior cumulants of *l* are obtained in general from $\Phi(t)$ and $\Phi_1(t)$, and in particular for a single future observation y from $\Phi_{l_q}(t)$ and $\Phi_1(t)$.

Furthermore if the I populations have no parameters in common and have independent prior distributions for their parameters, then $\Phi_1(t)$ can be expressed as a product of the posterior moment generating functions $\Phi_{l_i}(t)$ of l_i for i=1,2,..,I. Omitting the subscript i from l_i, the posterior moment generating function of l_i is found for a multivariate Normal linear model population in section 7.5. Hence $\Phi_1(t)$ can be found for multivariate Normal linear model populations with unequal precision matrices.

However if the I populations have parameters in common then $\Phi_1(t)$ cannot be expressed as a product of the posterior moment generating functions of l_i for i=1,2,..,I, but may be obtained directly. In section 7.6, $\Phi_1(t)$ is found for multivariate Normal linear model populations with equal precision matrices.

7.2 The posterior moment generating function of the log likelihood functions at multiple new observations from each of I populations where the mixing parameters are unknown

The log likelihood of new value y_{ij} is given by,

$l_{ij} = \log q_i p_i(y_{ij}|\theta) = \log q_i + l_{ij}$

where $l_{ij} = \log p_i(y_{ij}|\theta)$ for $j=1,2,..,m_i$ and $i=1,2,..,I$.

Let $l_i=(l_{i1},l_{i2},..,l_{im_i})'$ for $i=1,2,..,I$ and let $l=(l_1',l_2',..,l_I')'$,be a vector of length Σm_i, with l_i and l and t_i and t similarly defined.

Let $\Phi_l(t)$ be the posterior moment generating function of l ,i.e.

$$\Phi_l(t) = E\left[e^{t'l} \right]$$

$$= E\left[\prod_{i=1}^{I} \prod_{j=1}^{m_i} [q_i p_i(y_{ij}|\theta)]^{t_{ij}} \right]$$

$$= \int\int \left[\prod_{i=1}^{I} \prod_{j=1}^{m_i} [q_i p_i(y_{ij}|\theta)]^{t_{ij}} \right] p(\theta,q|\underline{x})\, d\theta\, dq \quad , \qquad (7.2.1)$$

where $p(\theta,q|\underline{x})$ is the posterior distribution of θ and q given the data $\underline{x}$.

The likelihood of the data $\underline{x}=(x_1,x_2,..,x_I)$ given θ and q is given by,

$$\prod_{i=1}^{I} \prod_{j=1}^{n_i} q_i p_i(x_{ij}|\theta) = \left[\prod_{i=1}^{I} q_i \right] . \prod_{i=1}^{I} \prod_{j=1}^{n_i} p_i(x_{ij}|\theta) \; . \qquad (7.2.2)$$

Hence,

$$p(\theta,q|\underline{x}) \alpha \left[\prod_{i=1}^{I} \prod_{j=1}^{n_i} q_i p_i(x_{ij}|\theta) \right] . p(\theta,q) \quad , \qquad (7.2.3)$$

where $p(\theta,q)$ is the joint prior distribution of θ and q.

Assuming that the prior distributions of θ and q are independent and given by $p(\theta)$ and $p(q)$, then their posterior distributions will also be independent,

$$p(\theta|\underline{x}) \alpha \left[\prod_{i=1}^{I} \prod_{j=1}^{n_i} p_i(x_{ij}|\theta) \right] p(\theta) \quad ,\text{and} \qquad (7.2.4)$$

$$p(q|\underline{x}) \alpha \left[\prod_{i=1}^{I} q_i^{n_i} \right] p(q) \; . \qquad (7.2.5)$$

Hence $\Phi_I(t)$ can be written as the product of two integrals,

$$\Phi_I(t) = \Phi(t) \cdot \Phi_1(t) \quad , \tag{7.2.6}$$

$$\text{where} \quad \Phi(t) = \int \left[\prod_{i=1}^{I} q_i^{\Sigma t_{ij}} \right] p(q|\underline{x})\, dq \quad , \text{ where } \Sigma = \sum_{j=1}^{m_i} \quad , \tag{7.2.7}$$

$$\text{and } \Phi_1(t) = \int \left[\prod_{i=1}^{I} \prod_{j=1}^{m_i} [p_i(y_{ij}|\Theta)]^{t_{ij}} \right] p(\Theta|\underline{x})\, d\Theta \quad . \tag{7.2.8}$$

If the populations Π_i, i=1,2,..,I, have no parameters in common, then let Θ_i be the parameters of Π_i for i=1,2,..,I, so that $\Theta=(\Theta_1,\Theta_2,..,\Theta_I)$. Assuming that the distinct parameter sets Θ_i have independent prior distributions $p(\Theta_i)$ for i=1,2,..,I, then

$$p(\Theta) = \prod_{i=1}^{I} p(\Theta_i) \quad .$$

In this case the posterior distributions of Θ_i are independent :

$$p(\Theta_i|x_i) \propto \left[\prod_{j=1}^{n_i} p_i(x_{ij}|\Theta_i) \right] . p(\Theta_i) \quad , \tag{7.2.9}$$

and hence $\Phi_1(t)$ in equation (7.1.8) can be split into I seperate integrals,i.e.

$$\Phi_1(t) = \prod_{i=1}^{n_i} \int \left[\prod_{j=1}^{m_i} [p_i(y_{ij}|\Theta_i)]^{t_{ij}} \right] p(\Theta_i|x_i)\, d\Theta_i \quad , \tag{7.2.10}$$

$$= \prod_{i=1}^{I} \Phi_{1_i}(t_i) \quad . \tag{7.2.11}$$

Hence the posterior distributions of l_i for i=1,2,..,I are independent and their posterior moment generating functions can be obtained seperately. If the populations have probability density functions of the same functional form but with different parameter sets (i.e. no parameters in common) then the posterior moment generating functions of l_i for i=1,2,..,I are all of the same form, as, for example, in section 7.5 for the case of multivariate Normal linear model populations with unequal precision matrices and where for simplicity of notation the suffix i is dropped from l_i.

However when the populations have parameters in common this simplification is not possible, as,for example, in section 7.6 for the case of multivariate Normal linear model populations with equal precision matrices.

7.3 The posterior moments of the mixing parameters and their logs

Assuming a Dirichlet $Di(\underline{g}_0)$ prior distribution for the mixing parameters q,

$$p(q) = \frac{\Gamma\left[\sum_{i=1}^{I} g_{0i}\right] \prod_{i=1}^{I} q_i^{g_{0i}-1}}{\prod_{i=1}^{I} \Gamma(g_{0i})} , \qquad (7.3.1)$$

where $q_I = 1 - \sum_{i=1}^{I-1} q_i$ and $\underline{g}_0 = (g_{01}, g_{02}, .., g_{0I})$, then the posterior distribution of q given $\underline{x}$ is $Di(\underline{g}_0+\underline{n})$ where $\underline{n} = (n_1, n_2, .., n_I)$.

For a general result let $p(q|\underline{x}) = Di(\underline{g})$.

$$\text{Hence } \Phi(t) = \int \frac{\left[\prod_{i=1}^{I} q_i^{\Sigma t_{ij}}\right] . \Gamma\left[\sum_{i=1}^{I} g_i\right] \prod_{i=1}^{I} q_i^{g_i-1}}{\prod_{i=1}^{I} \Gamma(g_i)} dq ,$$

$$= \frac{\Gamma\left[\sum_{i=1}^{I} g_i\right] . \prod_{i=1}^{I} \Gamma(g_i + \Sigma t_{ij})}{\Gamma\left[\sum_{i=1}^{I} (g_i + \Sigma t_{ij})\right] . \prod_{i=1}^{I} \Gamma(g_i)} , \qquad (7.3.2)$$

where $\Sigma t_{ij} = \sum_{j=1}^{m_i} t_{ij}$.

The posterior moment generating function of $l_q = (\log q_i; i=1,2,..,I)$

$$\Phi_{l_q}(t) = E\left[e^{t' l_q}\right] \quad \text{where here } t = (t_1, t_2, .., t_I)$$

$$= E\left[\prod_{i=1}^{I} q_i^{t_i}\right]$$

$$= \frac{\Gamma\left[\sum_{i=1}^{I} g_i\right] . \prod_{i=1}^{I} \Gamma(g_i + t_i)}{\Gamma\left[\sum_{i=1}^{I} (g_i + t_i)\right] . \prod_{i=1}^{I} \Gamma(g_i)} , \qquad (7.3.3)$$

by setting $m_i=1$ and $t_{i1}=t_i$ for i=1,2,..,I in equation (7.3.2).

The posterior moments of q

Hence suitable substitution for t in (7.3.3) gives the posterior moments of q. In particular for $r \geq 1$,

$$E\left[q_i^r\right] = \frac{\Gamma(\Sigma g_i).\Gamma(g_i+r)}{\Gamma(\Sigma g_i+r).\Gamma(g_i)} \quad ,$$

where $\Sigma = \sum_{i=1}^{I}$ throughout the rest of section 7.3.

Hence $E[q_i] = \frac{g_i}{\Sigma g_i}$,

and $E\left[q_i^r\right] = \frac{(g_i+r-1)}{(\Sigma g_i+r-1)} .E\left[q_i^{r-1}\right]$, for $r \geq 1$.

The posterior moments of log q

The posterior cumulant generating function of log q , $K_{l_q}(t)$ is given by

$$K_{l_q}(t) = \log \Phi_{l_q}(t)$$

$$= \log \Gamma(\Sigma g_i) - \log \Gamma(\Sigma[g_i+t_i]) - \Sigma[\log\Gamma(g_i) - \log\Gamma(g_i+t_i)]. \quad (7.3.4)$$

Hence the r^{th} cumulant of $l_{q_i} = \log q_i$ is given by,

$$(K_r)_{l_{q_i}} = \left[\frac{d^{(r)}}{dt_i^{(r)}} K_{l_{q_i}}(t) \right]_{t=0} = \psi^{(r-1)}(g_i) - \psi^{(r-1)}(\Sigma g_i) .$$

In particular, $E(l_{q_i}) = E(\log q_i) = (K_1)_{l_{q_i}} = \Psi(g_i) - \Psi(\Sigma g_i)$.

Furthermore

$$D^{(\underline{r})}K_{l_q}(t) = \frac{d^{(\underline{r})}}{dt^{(\underline{r})}} K_{l_q}(t) = -\psi^{(\Sigma r_i-1)}[\Sigma(g_i+t_i)] \quad ,$$

for $\underline{r} \geq \underline{0}$ with at least two r_i's non-zero, where $\underline{r}=(r_1,r_2,..,r_I)$.

Hence $(K_{\underline{r}})_{l_q} = -\psi^{(\Sigma r_i-1)}[\Sigma g_i]$,

and in particular $\text{cov}(\log q_{i1}, \log q_{i2}) = -\psi^{(1)}[\Sigma g_i]$, for $q_{i1} \neq q_{i2}$.

7.4 Combining the moments of the log likelihood functions with the moments of the logs of the mixing parameters

From equation (7.2.6) the posterior cumulant generating function of l is given by,

$$K_l(t) = \log \Phi(t) + K_{l_1}(t) \quad \text{, where } \Phi_{(t)} \text{ is given by (7.3.2).} \qquad (7.4.1)$$

Hence the $\underline{r}^{th}$ posterior cumulant of l, where $\underline{r}=(r_{ij};j=1,2,..,m_i;i=1,2,..,I)$ is given by,

$$(K_{\underline{r}})_l = \left[\frac{d^{(\underline{r})}}{dt^{(\underline{r})}} \log \Phi(t)\right]_{\underline{t}=\underline{0}} + (K_{\underline{r}})_{l_1} \quad , \qquad (7.4.2)$$

and hence the posterior cumulants and moments of l can easily be found.

In particular for a single future value y

Set $m_i=1$ and y_{i1} for $i=1,2,..,I$, and hence from (7.4.1) and (7.4.2),

$$K_l(t) = K_{l_q}(t) + K_{l_1}(t) \quad ,$$

$$\text{and } (K_{\underline{r}})_l = (K_{\underline{r}})_{l_q} + (K_{\underline{r}})_{l_1} \quad .$$

The posterior cumulant generating function and cumulants of $l_q=\log q$ are given in section 7.3.

The posterior cumulant generating function and cumulants of l for multivariate Normal linear model populations with equal precision matrices are given in section 7.6.3 for the case of a single future observation y.

For populations with no parameters in common then from section 7.2,

$$\Phi_{l_1}(t) = \prod_{i=1}^{I} \Phi_{l_i}(t_i) \quad \text{and hence}$$

$$K_{l_1}(t) = \sum_{i=1}^{I} K_{l_i}(t_i) \quad \text{and}$$

$$(K_{\underline{r}})_{l_1} = \sum_{i=1}^{I} (K_{r_i})_{l_i} \quad .$$

and so the populations can be dealt with seperately and later the cumulants combined. Omitting the subscript i from l_i and r_i, the posterior cumulant generating function and cumulants of each l_i can be found for multivariate Normal linear model populations with unequal precision matrices from sections 7.5.8 and 7.5.9, and for exponential family populations from section 4.1.

7.5 Evaluating $\Phi_1(t)$ for a multivariate Normal linear model population.

7.5.1 Introduction

Let $\underline{x}=(x_1,x_2,..,x_n)$ be a random sample from a d-dimensional multivariate Normal linear model population given by

$x_j' = z_j'B + \epsilon_j'$, where $\epsilon_j \sim No_d(0,T)$ independently for j=1,2,..,n.

Hence $X = ZB + \epsilon$, where $\epsilon \sim No_{n,d}(0,I_n,T)$, a Matrix Normal distribution defined by (7.5.5), and

$X = (x_1,x_2,..,x_n)'$ is a nxd matrix of observations,

$Z = (z_1,z_2,..,z_n)'$ is a nxk known design matrix,

$\epsilon = (\epsilon_1,\epsilon_2,..,\epsilon_n)'$ is a nxd matrix of errors, and

B is a kxd matrix of unknown regression parameters.

Interest centres on the joint posterior moment generating function $\Phi_1(t)$ of the log likelihood functions of possible future observations $(y_j;j=1,2,..,m)$ defined by the same model, i.e.

$y_j' = w_j'B + e_j'$, where $e_j \sim No_d(0,T)$ independently for j=1,2,..,m.

Hence $Y = WB + e$, where $e \sim No_{m,d}(0,I_m,T)$, and

$Y = (y_1,y_2,..,y_m)'$ is a mxd matrix of future observations,

$W = (w_1,w_2,..,w_m)'$ is a mxk known future design matrix, and

$e = (e_1,e_2,..,e_m)'$ is a mxd matrix of errors.

Hence $\Phi_1(t) = E[e^{t'l}]$,

where here $t=(t_1,t_2,..,t_m)'$, $l=(l_1,l_2,..,l_m)'$ and $l_j = \log p(y_j|B,T,w_j)$ for j=1,2,..m.

$$\text{i.e. } \Phi_1(t) = \int\int \left[\prod_{j=1}^{m} [p(y_j|B,T,w_j)]^{t_j} \right] p(B,T|X,Z)\, dB\, dT \quad , \qquad (7.5.1)$$

where $p(B,T|X,Z) = p(B|T,X,Z).p(T|X,Z)$ is the joint posterior distribution of B and T given the observed data X and the known design matrix Z.

$$\text{Hence } \Phi_1(t) = \int G(T)\, p(T|X,Z)\, dT \qquad (7.5.2)$$

$$\text{where } G(T) = \int \left[\prod_{j=1}^{m} [p(y_j|B,T,w_j)]^{t_j} \right] p(B|T,X,Z)\, dB \qquad (7.5.3)$$

7.5.2.The joint posterior distribution of B and T given observed data X and known design matrix Z.

Assuming vague prior knowledge for (B,T),

$p(B,T) \propto |T|^{-\frac{1}{2}(d+1)}$ and so

$p(B,T|X,Z) = \text{NoWi}_{k,d}(\hat{B}, Z'Z, n-k, S)$,

a Matrix Normal-Wishart distribution, (an obvious generalisation of the multivariate Normal-Wishart distribution), defined below by equation (7.5.5), where $S = (X-Z\hat{B})'(X-Z\hat{B})$ and $\hat{B} = (Z'Z)^{-1}Z'X$ provided Z'Z is non-singular.

Assuming a conjugate prior distribution for (B,T),

$p(B,T) = \text{NoWi}_{k,d}(b_0, C_0, v_0, h_0)$,

where $v_0 \geq d-1$ and C_0 and h_0 are symmetric positive definite, then

$$p(B,T|X,Z) = \text{NoWi}_{k,d}(b_1, C_0+Z'Z, v_0+n, h_1) \qquad (7.5.4)$$

where $b_1 = (C_0+Z'Z)^{-1}(C_0b_0+Z'X)$

and $h_1 = h_0 + b_0'C_0b_0 + X'X - (C_0b_0+Z'X)'(C_0+Z'Z)^{-1}(C_0b_0+Z'X)$.

[Note that letting $v=-k$, $C_0 \to 0$, $h_0 \to 0$ in the conjugate prior gives the correct posterior distribution for (B,T) for vague prior knowledge.]

For a general result let the posterior distribution for (B,T) be given by,

$p(B,T|X,Z) = \text{NoWi}_{k,d}(b,C,v,h)$,

i.e. $p(B|T,X,Z) = \text{No}_{k,d}(b,C,T)$

$$= \frac{|T|^{\frac{1}{2}k}|C|^{\frac{1}{2}d}}{(2\pi)^{\frac{1}{2}kd}} \exp\left[-\tfrac{1}{2}\text{tr}[(B-b)'C(B-b)T]\right] ,$$

and $p(T|X,Z) = \text{Wi}_d(v,h)$

$$= \frac{|\frac{1}{2}h|^{\frac{1}{2}v}|T|^{\frac{1}{2}(v-d-1)}}{\Gamma_d(\frac{1}{2}v)} \exp[-\tfrac{1}{2}\text{tr}(hT)] \quad . \qquad (7.5.5)$$

7.5.3 Evaluating G(T)

From equation (7.5.3),

$$G(T) = \int \left[\prod_{j=1}^{m} \left[\frac{|T|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}d}} \exp[-\tfrac{1}{2}(y_j - \quad)'T(y_j - \quad)] \right]^{t_j} \right]$$

$$\cdot \frac{|T|^{\frac{1}{2}k}|C|^{\frac{1}{2}d}}{(2\pi)^{\frac{1}{2}kd}} \exp\left[-\tfrac{1}{2}tr[(B-b)'C(B-b)T]\right] \, dB \quad ,$$

$$= \frac{|C|^{\frac{1}{2}d} \, |T|^{\frac{1}{2}(k+\Sigma t_j)}}{(2\pi)^{\frac{1}{2}d(k+\Sigma t_j)}} \int \exp\left[-\tfrac{1}{2}tr[H(B).T]\right] dB \quad , \qquad (7.5.6)$$

where $H(B) = [\Sigma t_j(y_j - \beta'\omega_j)(y_j - \beta'\omega_j)'] + (B-b)'C(B-b)$,

$$= (Y-WB)'\Delta(Y-WB) + (B-b)'C(B-b) \quad ,$$

$$= (B-\overset{\Lambda}{B})'(W'\Delta W+C)(B-\overset{\Lambda}{B}) + V \quad ,$$

by completing the square in B, where

$\Delta = \text{diag}(t_1, t_2, .., t_m)$,

$$\overset{\Lambda}{B} = (W'\Delta W+C)^{-1}(W'\Delta Y+C'b) \quad , \qquad (7.5.7)$$

$$V = Y'\Delta Y + b'Cb - (b'C+Y'\Delta W)\overset{\Lambda}{B} \quad , \qquad (7.5.8)$$

and $\Sigma = \sum_{j=1}^{m}$ throughout sections 7.5.3 and 7.5.4.

Now integrating out B in equation (7.5.6) to the normalising constant of a Matrix Normal distribution gives normalising constant

$(2\pi)^{\frac{1}{2}kd} \, |T|^{-\frac{1}{2}k} \, |W'\Delta W+C|^{-\frac{1}{2}d}$ from $No_{k,d}[\overset{\Lambda}{B}, W'\Delta W+C, T]$ in the notation of (7.5.5).

Hence $$G(T) = |C|^{\frac{1}{2}d} \, |W'\Delta W+C|^{-\frac{1}{2}d} \, |T|^{\frac{1}{2}\Sigma t_j} \exp[-\tfrac{1}{2}tr(VT)] \quad , \qquad (7.5.9)$$

where V is defined by (7.5.8) and $\overset{\Lambda}{B}$ by (7.5.7).

7.5.4.Evaluating $\Phi_1(t)$

Hence substituting (7.5.9) into (7.5.2) and integraing out T to the normalising constant of a Wishart, $Wi_d(v+\Sigma t_j, h+V)$, distribution gives

$$\Phi_1(t) = \frac{|C|^{\frac{1}{2}d} \, |h|^{\frac{1}{2}v} \, \Gamma_d[\frac{1}{2}(v+\Sigma t_j)]}{|W'\Delta W+C|^{\frac{1}{2}d} \, |h+V|^{\frac{1}{2}(v+\Sigma t_j)} \, \Gamma_d(\frac{1}{2}v) \, \pi^{\frac{1}{2}d\Sigma t_j}} \quad , \tag{7.5.10}$$

where V, given by (7.5.8), can be rewritten by completing the square in Y, i.e.

$$V = (Y-Y^*)'A(Y-Y^*) + E \quad , \text{ where} \tag{7.5.11}$$

$$A = [\Delta - \Delta W(W'\Delta W+C)^{-1}W'\Delta] \quad , \tag{7.5.12}$$

$$Y^* = A^{-1}\Delta W(W'\Delta W+C)^{-1}C'b = Wb \quad , \text{ and}$$

$$E = -Y^{*\prime}AY^* - b'C(W'\Delta W+C)^{-1}C'b + b'Cb = 0 \quad ,$$

the results being obtained after some algebra.

Hence $V = (Y-Wb)'A(Y-Wb)$ (7.5.13)

7.5.5 The joint predictive distribution for $Y = (y_1, y_2, .., y_m)'$

Setting $t_j=1$ for $j=1,2,..,m$ gives $\Delta=I_m$ and

$$p(Y|X,Z,W) = \frac{|C|^{\frac{1}{2}d} \, |h|^{\frac{1}{2}v} \, \Gamma_d[\frac{1}{2}(v+m)]}{|W'W+C|^{\frac{1}{2}d} \, |h+(Y-Wb)'A_1(Y-Wb)|^{\frac{1}{2}(v+m)} \, \Gamma_d(\frac{1}{2}v) \, \pi^{\frac{1}{2}dm}} \quad , \qquad (7.5.14)$$

where $A_1 = [I - W(W'W+C)^{-1}W']$ (7.5.15)

i.e. given X,Z and W, then Y has a Matrix t distribution,

$Y|X,Z,W \sim t_{m,d}[Wb, A_1^{-1}, h, v-d+1]$, and (7.5.16)

$Y'|X,Z,W \sim t_{d,m}[(Wb)', h, A_1^{-1}, v-d+1]$

in the notation of Box and Tiao (1973).

The predictive distribution for Y was first obtained by Geisser (1965) for vague prior knowledge about B and T, and has been considered more recently by Broemeling (1985,Ch8) for a general conjugate prior, however, apart from some typing errors in the result and omitting the normalising constant, Broemeling apparantly fails to recognise the dramatic simplification of Y^* to Wb and E to 0 and his resultant predictive distribution for Y is unneccessarily complicated.

7.5.6 Simplification of $\Phi_1(t)$

Simplification of $\Phi_1(t)$ can be obtained by using the partitioned matrix result

$$\begin{vmatrix} M_1 & M_3 \\ M_2 & M_4 \end{vmatrix} = |M_4||M_1 - M_3M_4^{-1}M_2| = |M_1||M_4 - M_2M_1^{-1}M_3| \qquad (7.5.17)$$

Letting $(M_1, M_2, M_3, M_4) = (\Delta^{-1}, W', -W, C)$ gives $|W'\Delta W + C| = |C||I_m + \Delta WC^{-1}W'|$.

Letting $(M_1, M_2, M_3, M_4) = (\Delta, W'\Delta, \Delta W, W'\Delta W + C)$ gives
$|A| = |\Delta||C||W'\Delta W + C|^{-1} = |\Delta||I_m + \Delta WC^{-1}W'|^{-1}$, where A is given by (7.5.12).

Letting $(M_1, M_2, M_3, M_4) = (h, (Y - Wb), -(Y - Wb)', A^{-1})$ gives
$|h + (Y - Wb)'A(Y - Wb)| = |h||I_m + A(Y - Wb)h^{-1}(Y - Wb)'|$

Hence

$$\Phi_1(t) = \frac{\Gamma_d[\tfrac{1}{2}(v + \Sigma t_j)]}{|I_m + \Delta WC^{-1}W'|^{\frac{1}{2}d} \; |I_m + A(Y - Wb)h^{-1}(Y - Wb)'|^{\frac{1}{2}(v + \Sigma t_j)} \; |h|^{\frac{1}{2}\Sigma t_j} \; \pi^{\frac{1}{2}d\Sigma t_j} \; \Gamma_d(\tfrac{1}{2}v)} \qquad (7.5.18)$$

Let $D = (Y - Wb)h^{-1}(Y - Wb)'$, then the $(i,j)^{th}$ element of D is given by

$D_{ij} = (y_i - b'w_i)'h^{-1}(y_j - b'w_j)$, for i=1,2,..,m and j=1,2,..,m. (7.5.19)

7.5.7 Evaluation of the elements of matrix A

The $(j,j)^{th}$ diagonal element of matrix A, for j=1,2,..,m is given by

$$A_{jj} = t_j - t_j w_j'(C + \Sigma t_k w_k w_k')^{-1} t_j w_j$$

$$= \frac{t_j . |C - t_j w_j w_j' + \Sigma t_k w_k w_k'|}{|C + \Sigma t_k w_k w_k'|} \quad , \qquad (7.5.20)$$

where $\Sigma = \sum_{k=1}^{m}$ throughout section 7.5.7,

by using (7.5.17) with $[M_1,M_2,M_3,M_4] = [t_j, t_j w_j, t_j w_j', C+\Sigma t_k w_k w_k']$

Also using (7.5.17) again with
$[M_1,M_2,M_3,M_4] = [C,(t_1w_1,t_2w_2,..,t_mw_m)',-(w_1,w_2,..,w_m),I_m]$ gives

$$|C+\Sigma t_k w_k w_k'| = |C||I_m+\Delta e| \quad , \qquad (7.5.21)$$

where the $(i,j)^{th}$ element of matrix $e=WC^{-1}W'$ is given by $e_{ij}=w_i'C^{-1}w_j$,
and the $(i,j)^{th}$ element of matrix Δe is given by $(\Delta e)_{ij}=t_i e_{ij}$.
The numerator in equation (7.5.20) can be similarly obtained.

The $(i,j)^{th}$ off diagonal element of matrix A is given by

$$A_{ij} = - t_i w_i'(C + \Sigma t_k w_k w_k')^{-1} t_j w_j \quad ,$$

for i=1,2,..,m and j=1,2,..m with $i \neq j$.

Using equation (7.5.17) with $[M_1,M_2,M_3,M_4] = [1, t_j w_j, t_i w_i', C+\Sigma t_k w_k w_k']$ gives

$$1+A_{ij} = \frac{|C - t_i t_j w_j w_i' + \Sigma\, t_k w_k w_k'|}{|C + \Sigma\, t_k w_k w_k'|} \quad . \qquad (7.5.22)$$

The denominator in (7.5.22) is the same as in (7.5.20) and hence can be obtained using (7.5.21).

To obtain the numerator in equation (7.5.22) use (7.5.17) with

$[M_1,M_2,M_3,M_4] = [C,(t_1w_1,t_2w_2,..,t_mw_m,t_iw_i)',-(w_1,w_2,..,w_m,-t_jw_j),I_{m+1}]$ giving

$$|C-t_i t_j w_j w_i'+\Sigma t_k w_k w_k'| = |I_{m+1} + (t_1w_1,t_2w_2,..,t_mw_m,t_iw_i)'C^{-1}(w_1,w_2,..,w_m,-t_jw_j)| \qquad (7.5.23)$$

Hence both numerator and denominator in $1+A_{ij}$ can easily be obtained. Matrix A is then easily obtained from A_{ij} and A_{jj}.

In conclusion note that $\Phi_1(t)$ depends on the data x only through h,v,e and D where $e_{ij} = w_i'C^{-1}w_j$ and $D_{ij} = (y_i-b'w_i)'h^{-1}(y_j-b'w_j)$ for i=1,2,..,m and j=1,2,..,m. This is easily shown by substituting the results for the elements of A from section 7.5.7 into equation (7.5.18), together with (7.5.19) and also noting that $|I_m+\Delta WC^{-1}W'| = |I_m+\Delta e|$ and is easily obtained as for (7.5.21).

Hence to conclude the joint posterior distribution of the log likelihood functions $l=(l_1,l_2,..,l_m)$ where $l_j = \log p(y_j|\Theta)$ depends on the observed data $\underline{x}$ only through the quantities h,v,e and D.

[Note, in particular, that for vague prior knowledge $(b,C,v,h) = (\hat{B},Z'Z,n-k,S)$ where $S = (X-Z\hat{B})'(X-Z\hat{B})$ and $\hat{B} = (Z'Z)^{-1}Z'X$]

7.5.8 The posterior moment generating function of the log likelihood function $l=\log p(y|B,T,w)$ at a single future value y.

Set $t_j=0$ for $j=2,3,..,m$ to give $\Delta_{11}=t_1$ and $\Delta_{ij}=0$ for $(i,j)\neq(1,1)$, and omitting the suffix 1 from l_1,t_1,y_1 and w_1 to give from (7.5.18),

$$\Phi_l(t) = \frac{\Gamma_d[\tfrac{1}{2}(v+t)]}{[1+tw'C^{-1}w]^{\frac{1}{2}d}\,[1+A_{11}(y-b'w)'h^{-1}(y-b'w)]^{\frac{1}{2}(v+t)}\,|h|^{\frac{1}{2}t}\,\pi^{\frac{1}{2}dt}\,\Gamma_d(\tfrac{1}{2}v)} \quad , \tag{7.5.24}$$

$$\text{where } A_{11} = t - tw'(tww'+C)^{-1}wt = \frac{t}{[1+tw'C^{-1}w]} \tag{7.5.25}$$

using (7.5.17).

Predictive distribution for a single future observation y

Setting t=1 in equation (7.5.24) gives the predictive distribution for y

$$p(y|X,Z,w) = \frac{\Gamma_d[\tfrac{1}{2}(v+1)]}{[1+w'C^{-1}w]^{\frac{1}{2}d}\,[1+A_{11}(y-b'w)'h^{-1}(y-b'w)]^{\frac{1}{2}(v+1)}\,|h|^{\frac{1}{2}}\,\pi^{\frac{1}{2}d}\,\Gamma_d(\tfrac{1}{2}v)} \quad , \tag{7.5.26}$$

where here $A_{11} = [1+w'C^{-1}w]^{-1}$, and

$$\frac{\Gamma_d[\tfrac{1}{2}(v+1)]}{\Gamma_d(\tfrac{1}{2}v)} = \frac{\Gamma[\tfrac{1}{2}(v+1)]}{\Gamma[\tfrac{1}{2}(v-d+1)]} \quad ,$$

i.e. the predictive distribution of y is the multivariate Student t distribution,

$$p(y|X,Z,w) = St_d[v,b'w,v^{-1}(1+w'C^{-1}w)h] \quad ,$$

in the notation of Aitchison and Dunsmore (1975), and

$$p(y|X,Z,w) = t_d[b'w,(v-d+1)^{-1}(1+w'C^{-1}w)h,v-d+1] \quad ,$$

in the notation of Box and Tiao (1973).

7.5.9 The posterior moments of the log likelihood function at a single future value y

Equation (7.5.24) for $\Phi_1(t)$ is in the same form as equation (2.2.2) of chapter 2 for the case of a multivariate Normal population except that c in chapter 2 is here replaced by $(w'C^{-1}w)^{-1}$ and $D=(y-b)'h^{-1}(y-b)$ in chapter 2 is here replaced by $(y-b'w)'h^{-1}(y-b'w)$, and hence the resulting equations (2.3.4) for the moments of the log likelihood function are the same with the appropriate substitutions for c and D. Note also that the values for the parameters (b,C,v,h) are different here, as discussed in section 7.5.2.

Here for vague prior knowledge,

$p(B,T) \alpha |T|^{\frac{1}{2}(d+1)}$ and hence set $(b,C,v,h) = (\hat{B},Z'Z,n-k,S)$, where

$S=(X-Z\hat{B})'(X-Z\hat{B})$ and $\hat{B}=(Z'Z)^{-1}Z'X$.

7.6 Evaluating $\Phi_1(t)$ for multivariate Normal linear model populations with equal error precision matrices.

7.6.1 Introduction

Observations belong to one of I multivariate Normal linear model populations, Π_i for i=1,2,..,I, having equal error precision matrix T.

Let $(x_{i1},x_{i2},..,x_{in_i})$ be a random sample from population Π_i, a d-dimensional multivariate Normal linear model population, i.e for i=1,2,..I,

$x_{ij}' = z_{ij}'B_i + \epsilon_{ij}'$, where $\epsilon_{ij} \sim No_d(0,T)$ independently for $j=1,2,..,n_i$.

Hence $X_i = Z_iB_i + \epsilon_i$, where $\epsilon_i \sim No_{n_i,d}(0,I_{n_i},T)$ a Matrix Normal distribution defined by (7.5.5), and

$X_i = (x_{i1},x_{i2},..,x_{in_i})'$ is a $n_i \times d$ matrix of observations,

$Z_i = (z_{i1},z_{i2},..,z_{in_i})'$ is a $n_i \times k_i$ known design matrix,

$\epsilon_i = (\epsilon_{i1},\epsilon_{i2},..,\epsilon_{in_i})'$ is a $n_i \times d$ matrix of errors, and

B_i is a $k_i \times d$ matrix of unknown regression parameters.

Hence $X = ZB + \epsilon$, where $\epsilon \sim No_{n,d}(0,I_n,T)$, $n = \sum_{i=1}^{I} n_i$, $k = \sum_{i=1}^{I} k_i$ and

$X = (X_1',X_2',..,X_I')'$ is a $n \times d$ matrix of observations,

$Z = \text{mat-diag}(Z_1',Z_2',..,Z_I')'$ is a $n \times k$ matrix diagonal design matrix,

$\epsilon = (\epsilon_1',\epsilon_2',..,\epsilon_I')'$ is a $n \times d$ matrix of errors, and

$B = (B_1',B_2',..,B_I')'$ is a $k \times d$ matrix of unknown regression parameters.

Interest centres on the joint posterior moment generating function $\Phi_1(t)$ of the log likelihood functions of possible future observations $(y_{ij};j=1,2,..,m_i;i=1,2,..,I)$ defined by the same model, i.e. for i=1,2,..,I,

$y_{ij}' = w_{ij}'B_i + e_{ij}'$, where $e_{ij} \sim No_d(0,T)$ independently for j=1,2,..,m.

Hence $Y_i = W_iB_i + e_i$, where $e_i \sim No_{m_i,d}(0,I_{m_i},T)$, and

$Y_i = (y_{i1},y_{i2},..,y_{im_i})'$ is a $m_i \times d$ matrix of future observations,

$W_i = (w_{i1},w_{i2},..,w_{im_i})'$ is a $m_i \times k_i$ known future design matrix, and

$e_i = (e_{i1},e_{i2},..,e_{im_i})'$ is a $m_i \times d$ matrix of errors.

Hence $Y = WB + e$, where $e \sim No_{m,d}(0,I_m,T)$, $m = \sum_{i=1}^{I} m_i$ and

$Y = (Y_1',Y_2',..,Y_I')'$ is a $m \times d$ matrix of future observations,

$W = \text{mat-diag}(W_1',W_2',..,W_I')'$ is a $m \times k$ matrix diagonal future design matrix, and

$e = (e_1',e_2',..,e_I')'$ is a $m \times d$ matrix of errors.

The model is now in the form of section 7.5 and hence the results of section 7.5 can be used with appropriate substitutions.
In particular from (7.5.18),

$$\Phi_1(t) = \int\int \left[\prod_{i=1}^{I} \prod_{j=1}^{m_i} [p(y_{ij}|B,T,w_{ij})]^{t_{ij}} \right] p(B,T|X,Z)\, dB\, dT \qquad (7.6.1)$$

$$= \frac{\Gamma_d[\tfrac{1}{2}(v+\Sigma\Sigma t_{ij})}{|I_m+\Delta WC^{-1}W'|^{\frac{1}{2}d}\,|I_m+A(Y-Wb)h^{-1}(Y-Wb)'|^{\frac{1}{2}(v+\Sigma\Sigma t_{ij})}\,|h|^{\frac{1}{2}\Sigma\Sigma t_{ij}}\,\pi^{\frac{1}{2}d\Sigma\Sigma t_{ij}}\,\Gamma_d(\frac{1}{2}v)} \qquad (7.6.2)$$

where $\Sigma\Sigma t_{ij} = \sum_{i=1}^{I}\sum_{j=1}^{m_i} t_{ij}$, $m = \sum_{i=1}^{I} m_i$, $k = \sum_{i=1}^{I} k_i$,

$\Delta = \text{mat-diag}(\Delta_1,\Delta_2,..,\Delta_I)$ where $\Delta_i = \text{diag}(t_{i1},t_{i2},..,t_{im_i})$, and

$A = [\Delta - \Delta W(W'\Delta W+C)^{-1}W'\Delta]$.

7.6.2 Case where C is matrix diagonal

For vague prior knowledge the parameter C in the posterior distribution of (B,T) is matrix diagonal, $C = Z'Z = \text{mat-diag}[Z_1'Z_1, Z_2'Z_2, .., Z_I'Z_I]$.
Also, provided that the parameter C_0 of the conjugate prior distribution for (B,T) is matrix diagonal, i.e. $C_0 = \text{mat-diag}[C_{10}, C_{20}, .., C_{I0}]$, then C will also be matrix diagonal, i.e. $C = \text{mat-diag}[C_{10}+Z_1'Z_1, C_{20}+Z_2'Z_2, .., C_{I0}+Z_I'Z_I]$.
Note that the condition for C_0 to be matrix diagonal is equivalent to assuming independent prior distributions for $B_i; i=1,2,..I$ given T.

If C is $\text{mat-diag}[C_1, C_2, .., C_I]$ then

$$A = \text{mat-diag}[\Delta_i - \Delta_i W_i (W_i'\Delta_i W_i + C_i)^{-1} W_i'\Delta_i \; ; i=1,2,..,I] \;,$$

$$[I_m + \Delta W C^{-1} W'] = \text{mat-diag}[I_{m_i} + \Delta_i W_i C_i^{-1} W_i' \; ; i=1,2,..,I] \;, \text{ and hence}$$

$$|I_m + \Delta W C^{-1} W'| = \prod_{i=1}^{I} |I_{m_i} + \Delta_i W_i C_i^{-1} W_i'| \;.$$

7.6.3 The posterior moment generating function $\Phi_l(t)$ of the log likelihood functions $l=(l_1,l_2,..,l_I)$ where $l_i = \log p_i(y|B,T,w)$ is the log likelihood of a single future value y belonging to Π_i for i=1,2,..,I.

First set $m_i=1$ for i=1,2,..,I in (7.6.2) and assume C is mat-diag$[C_1,C_2,..,C_I]$.

Since $m_i=1$, $\Delta_i=t_i$, $W_i=w_i'$ and $Y_i=y_i'$ for i=1,2,..,I and hence using (7.5.17),

$$A = \text{diag}\,[t_i - t_i w_i'(t_i w_i w_i' + C_i)^{-1} w_i t_i \;; i=1,2,..,I]$$

$$= \text{diag}\left[\frac{t_i}{1+t_i w_i' C_i^{-1} w_i} \;; i=1,2,..,I\right] ,$$

$$|I_m + \Delta W C^{-1} W'| = \prod_{i=1}^{I} [1+t_i w_i' C_i^{-1} w_i] \quad \text{where here } m = \sum_{i=1}^{I} m_i = I \text{ , and}$$

$$[I_m + A(Y-Wb)h^{-1}(Y-Wb)']_{ij} = \delta_{ij} + \frac{t_i(y_i-b_i'w_i)'h^{-1}(y_j-b_j'w_j)}{1+t_i w_i' C_i^{-1} w_i} \quad , \quad (7.6.3)$$

where $\delta_{ij} = \begin{cases} 0 \text{ if } i \neq j \\ 1 \text{ if } i=j \end{cases}$ for i=1,2,..,I and j=1,2,..,I.

Now also set $y_i=y, t_i=t, w_i=w$ for i=1,2,..,I, assuming the same variables are measured for all I populations, then,

$$[I_m + AD]_{ij} = \delta_{ij} + \frac{t_i D_{ij}}{1+t_i w' C_i^{-1} w} \quad , \quad (7.6.4)$$

where $D = (Y-Wb)h^{-1}(Y-Wb)'$ and

$$D_{ij} = (y-b_i'w)'h^{-1}(y-b_j'w) \quad , \text{ and} \quad (7.6.5)$$

$$A_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ \dfrac{t_i}{1+t_i w' C_i^{-1} w} & \text{if } i=j \end{cases} \quad (7.6.6)$$

and hence,

$$\Phi_l(t) = \frac{\Gamma_d[\tfrac{1}{2}(v+\Sigma t_i)]}{\left[\prod_{i=1}^{I} [1+t_i w' C_i^{-1} w]\right]^{\frac{1}{2}d} |I+AD|^{\frac{1}{2}(v+\Sigma t_i)} |h|^{\frac{1}{2}\Sigma t_i} \pi^{\frac{1}{2}d\Sigma t_i} \Gamma_d(\tfrac{1}{2}v)} \quad , \quad (7.6.7)$$

where A, D and [I+AD] are I×I matrices with $(i,j)^{th}$ elements defined by (7.6.6), (7.6.5) and(7.6.4) respectively.

The equation (7.6.7) for $\Phi_1(t)$ is now in the same form as equation (6.2.6) for the case of multivariate Normal, $No_d(\mu_i,T)$, populations, except that c_i in chapter 6 is here replaced by $(w'C_i^{-1}w)^{-1}$ and $D_{ij} = (y-b_i)'h^{-1}(y-b_j)$ in chapter 6 is here replaced by $(y-b_i'w)'h^{-1}(y-b_j'w)$ and hence section 6.3 and the results of section 6.4 apply with the appropriate substitutions for $c_i;i=1,2,..,I$ and $D_{ij};i=1,2,..I;j=1,2,..,I$.

Note also that the values of the parameters (b,C,v,h) are different here as discussed in sections 7.5.2 and 7.6.2.

Here for vague prior knowledge about (B,T),

$p(B,T) \alpha |T|^{-\frac{1}{2}(d+1)}$ and hence set $(b,C,v,h) = (\hat{B},Z'Z,n-k,S)$, where

$$S = (X-Z\hat{B})'(X-Z\hat{B}) = \sum_{i=1}^{I} S_i ,$$

$$S_i = (X_i-Z_i\hat{B}_i)'(X_i-Z_i\hat{B}_i) \text{ for } i=1,2,..,I,$$

$$Z'Z = \text{mat-diag}(Z_i'Z_i ;i=1,2,..,I) ,$$

$$\hat{B}' = (\hat{B}_1',\hat{B}_2',..,\hat{B}_I') ,$$

$$\hat{B}_i = (Z_i'Z_i)^{-1}Z_i'X_i \text{ for } i=1,2,..,I,$$

$$k = \sum_{i=1}^{I} k_i = Ik, \text{ and}$$

$$n = \sum_{i=1}^{I} n_i .$$

7.6.4 Special case of I=2 populations

Let $L = l_1 - l_2$, hence

$$\Phi_L(t) = E\left[e^{tL}\right] = E\left[e^{t(l_1-l_2)}\right] = \Phi_{l_1,l_2}(t,-t) \quad .$$

Hence from (7.6.7),

$$\Phi_L(t) = \frac{[1-tw'C_2^{-1}w]^{-\frac{1}{2}d}}{[1+tw'C_1^{-1}w]^{-\frac{1}{2}d}\ |I+AD|^{\frac{1}{2}v}} \quad \text{where} \qquad (7.6.8)$$

$$|I+AD| = [1 + \alpha_1 D_{11} + \alpha_2 D_{22} + \alpha_1\alpha_2(D_{11}D_{22}-D_{12}^2)] \quad \text{and}$$

$$\alpha_i = \frac{t_i}{1+t_i w'C_i^{-1}w} \quad \text{for } i=1,2. \qquad (7.6.9)$$

The equation (7.6.8) for $\Phi_L(t)$ is now in the same form as equation (3.2.5) of chapter 3 with appropriate substitutions for $c_i;i=1,2$ and $D_{ij};i=1,2;j=1,2$, changes in the parameters (b,C,v,h) as discussed in section 7.6.3, and with $\alpha_i;i=1,2$ as defined by equation (7.6.9). Hence the results of section 3.3 for the posterior moments of L can be applied directly after these substitutions and changes.

CHAPTER 8

How stable is the predictive density function?

8.1 Introduction

Let $\underline{x}=(x_1,x_2,..,x_n)$ be a random sample of observations of a random variable X with probability density function $p(x|\theta)$, $x\epsilon R_X$. Interest centres on the likelihood function $p(y|\theta)$ and the log likelihood function $\log p(y|\theta)$ of a possible future value y of the random variable X, where y varies over the same range, i.e. R_X. In order to investigate the posterior distributions of the functions of θ, $p(y|\theta)$ and $\log p(y|\theta)$, consider values $y_1,y_2,..,y_m$ from the range of X. The posterior moment generating function of $\underline{l}=(l_1,l_2,..,l_m)$, where $l_j = \log p(y_j|\theta)$ for j=1,2,..,m is given by $\Phi_1(\underline{t})$ from chapter 7, where $\underline{t}=(t_1,t_2,..,t_m)$ and the number of populations considered is I=1, and hence the results of chapter 7 can be used.

The posterior moments of $\underline{p}=(p_1,p_2,..,p_m)$, where $p_j = p(y_j|\theta)$ for j=1,2,..,m,

$$(\mu_{\underline{r}})_{\underline{p}} = E\left[\prod_{j=1}^{m} p_j^{r_j}\right] , \qquad (8.1.1)$$

are obtained from $\Phi_1(\underline{t})$ by substituting $\underline{t}=\underline{r}$, where $\underline{r}=(r_1,r_2,..,r_m)$.

The posterior cumulants of $\underline{l}$, $(K_{\underline{r}})_{\underline{l}}$ are obtained from the cumulant generating function for $\underline{l}$, $K_1(\underline{t}) = \log \Phi_1(\underline{t})$, since,

$$(K_{\underline{r}})_{\underline{l}} = \left[\frac{d^{(\underline{r})}}{dt^{(\underline{r})}} K_1(\underline{t})\right]_{\underline{t}=\underline{0}} = K_1^{(\underline{r})}(\underline{0}) . \qquad (8.1.2)$$

In particular the posterior mean, variance and higher moments of the functions $p(y|\theta)$ and $\log p(y|\theta)$ for $y\epsilon R_X$ are obtained by setting m=1 and omitting the subscript 1 from y_1,p_1,l_1 and t_1 in equations (8.1.1) and (8.1.2),

$$\begin{aligned}
E[p(y|\theta)] &= (\mu_1)_p = \Phi_1(1) ,\\
V[p(y|\theta)] &= (\mu_2)_p - (\mu_1)_p^2 = \Phi_1(2) - \Phi_1(1)^2 ,\\
(\mu_3')_p &= (\mu_3)_p - 3(\mu_2)_p(\mu_1)_p + 2(\mu_1)_p^3 = \Phi_1(3) - 3\Phi_1(2)\Phi_1(1) + 2\Phi_1(1)^3 ,\\
(\mu_4')_p &= (\mu_4)_p - 4(\mu_3)_p(\mu_1)_p + 6(\mu_2)_p(\mu_1)_p^2 - 3(\mu_1)_p^4\\
&= \Phi_1(4) - 4\Phi_1(3)\Phi_1(1) + 6\Phi_1(2)\Phi_1(1)^2 - 3\Phi_1(1)^4 , \text{ and} \qquad (8.1.3)
\end{aligned}$$

$$\begin{aligned}
E[\log p(y|\theta)] &= (K_1)_1 = K_1^{(1)}(0) ,\\
V[\log p(y|\theta)] &= (K_2)_1 = K_1^{(2)}(0) ,\\
(\mu_3')_1 &= (K_3)_1 = K_1^{(3)}(0) ,\\
(\mu_4')_1 &= (K_4)_1 - 3(K_2)_1^2 = K_1^{(4)}(0) - 3K_1^{(2)}(0)^2 . \qquad (8.1.4)
\end{aligned}$$

Note that the posterior mean of the function $p(y|\Theta)$, for $y \epsilon R_{\underline{X}}$, is the predictive density function $p(y|\underline{x})$. The posterior moments of $p(y|\Theta)$ and log $p(y|\Theta)$ can be used to obtain credibility interval bands for each of the functions, and hence to investigate the stability of the posterior distributions of the functions. In particular, for function $p(y|\Theta)$, this will indicate the stability of the posterior distribution of the function $p(y|\Theta)$ about its posterior mean, the predictive density function $p(y|\underline{x})$.

Furthermore, the posterior correlation coefficient between the likelihood functions at two different values y_1 and y_2, i.e. $p_1 = p(y_1|\Theta)$ and $p_2 = p(y_2|\Theta)$, can easily be found since,

$$\rho(p_1,p_2) = \frac{\mathrm{Cov}(p_1,p_2)}{[V(p_1)V(p_2)]^{\frac{1}{2}}} = \frac{E(p_1p_2)-E(p_1)E(p_2)}{[V(p_1)V(p_2)]^{\frac{1}{2}}} , \qquad (8.1.5)$$

where $E(p_1p_2) = (\mu_{1,1})_{p_1,p_2} = \Phi_{l_1,l_2}(1,1)$

and $V(p_1)$ and $V(p_2)$ can be obtained using (8.1.3).

The posterior correlation coefficient between the log likelihood functions at two different values y_1 and y_2, i.e. $l_1 = \log p(y_1|\Theta)$ and $l_2 = \log p(y_2|\Theta)$ can also easily be found since,

$$\rho(l_1,l_2) = \frac{\mathrm{Cov}(l_1,l_2)}{[V(l_1)V(l_2)]^{\frac{1}{2}}} = \frac{(K_{1,1})_{l_1,l_2}}{[V(l_1)V(l_2)]^{\frac{1}{2}}} , \qquad (8.1.6)$$

where $(K_{1,1})_{l_1,l_2} = K^{(1,1)}_{l_1,l_2}(0,0)$,

and $V(l_1)$ and $V(l_2)$ can be obtained using (8.1.4).

Investigation of the credibility interval band for the likelihood (or log likelihood) function about its mean, the predictive density function, together with the posterior correlation coefficients above, will help to understand the structure or character of the predictive density function, or rather, that of the posterior distribution of the likelihood (or log likelihood) function.

8.2 Exponential family populations

For an Exponential family population, the posterior moment generating function $\Phi_1(\underline{t})$ of the log likelihood functions of m future observations $(y_1,y_2,..,y_m)$ is given by (4.2.6) and for a single future observation y by (4.2.9).

Hence from (4.2.9),

$$E[p(y|\Theta)] = \Phi_1(1) = \frac{c[\nu,\omega].b(y)}{c[\nu+1,\omega+s(y)]} ,$$

$$V[p(y|\Theta) = \Phi_1(2) - \Phi_1(1)^2 \quad \text{where,}$$

$$\Phi_1(2) = \frac{c[\nu,\omega].b(y)^2}{c[\nu+2,\omega+2s(y)]} , \qquad (8.2.1)$$

and the higher moments of $p = p(y|\Theta)$ are obtained from (4.2.9).

Also,

$$E[\log p(y|\Theta)] = (K_1)_1 = -\left[\frac{d}{dt}\log c[\nu+t,\omega+ts(y)]\right]_{t=0} + \log b(y) ,$$

$$V[\log p(y|\Theta)] = (K_2)_1 = -\left[\frac{d^{(2)}}{dt^{(2)}}\log c[\nu+t,\omega+ts(y)]\right]_{t=0} , \qquad (8.2.2)$$

and the higher cumulants of 1 are given by (4.2.12).

Also the posterior correlation coefficient between $p(y_1|\Theta)$ and $p(y|\Theta)$ is given by (8.1.5) where by setting m=2 and $(t_1,t_2) = (1,1)$ in (4.2.6),

$$E[p_1p_2] = \Phi_{1_1,1_2}(1,1) = \frac{c[\nu,\omega].b(y_1)b(y_2)}{c[\nu+2,\omega+s(y_1)+s(y_2)]} , \qquad (8.2.3)$$

and the posterior correlation coefficient between $\log p(y_1|\Theta)$ and $\log p(y_2|\Theta)$ is given by (8.1.6) where from (4.2.6),

$$(K_{1,1})_{1_1,1_2} = -\left[\frac{d}{dt_1}\frac{d}{dt_2}\log c[\nu+t_1+t_2,\omega+t_1s(y_1)+t_2s(y_2)]\right]_{\underline{t}=\underline{0}} , \qquad (8.2.4)$$

Example : Gamma population distribution for random variable X.

For a Gamma population distribution for X, $\Phi_1(\underline{t})$ is given by equation (4.3.3) for general m and by (4.3.5) for m=1.
Hence from (4.3.5),

$$E[p(y|\Theta)] = \Phi_1(1) = \frac{y^{k-1} h^g \Gamma(g+k)}{\Gamma(k)\,\Gamma(g)\,(h+y)^{g+k}} ,$$

$$V[p(y|\Theta)] = \Phi_1(2) - \Phi_1(t)^2 \quad \text{where,}$$

$$\Phi_1(2) = \frac{y^{2(k-1)} h^g \Gamma(g+2k)}{\Gamma(k)^2\,\Gamma(g)\,(h+2y)^{g+2k}} , \qquad (8.2.5)$$

and the higher moments of $p = p(y|\Theta)$ are obtained from (4.3.5).

Also,

$$E[\log p(y|\Theta)] = (K_1)_1 = (k-1).\log y + k.\psi(g) - \log[\Gamma(k)] - k.\log h - \frac{gy}{h} ,$$

$$V[\log p(y|\Theta)] = (K_2)_1 = k^2.\psi^{(1)}(g) + \frac{y}{h^2}[2hk-gy] , \qquad (8.2.6)$$

and the higher cumulants of $l = \log p(y|\Theta)$ are given by (4.3.8).

Also the posterior correlation coefficient between $p(y_1|\Theta)$ and $p(y_2|\Theta)$ is given by (8.1.5) where setting m=2 and $(t_1,t_2) = (1,1)$ in (4.3.3) gives,

$$E[p_1p_2] = \Phi_{l_1,l_2}(1,1) = \frac{y_1^{k-1} y_2^{k-1} h^g \Gamma(g+2k)}{\Gamma(k)^2\,\Gamma(g)\,[h+y_1+y_2]^{g+2k}} , \qquad (8.2.7)$$

and the posterior correlation cofficient between $l_1 = \log p(y_1|\Theta)$ and $l_2 = \log p(y_2|\Theta)$ is given by (8.1.6) where from (4.3.3),

$$(K_{1,1})_{l_1,l_2} = k^2.\psi^{(1)}(g) + \frac{g.y_1.y_2}{h^2} - \frac{k(y_1+y_2)}{h} . \qquad (8.2.8)$$

Exponemtial population distribution for X.

Setting k=1 in the results for a Gamma, Ga(k,Θ), population distribution gives the results for an Exponential, Ex(Θ), population distribution for X.

The posterior moments of $p = p(y|\Theta)$ are obtained by setting k=1 in (8.2.5). The posterior central moments of $l = \log p(y|\Theta)$ are given by (4.3.9).

Consider the following sample of size 4 from an Exponential, Ex(Θ), population taken from Aitchison and Dunsmore (1975, Ch11), $\underline{x} = (47,17,32,19)$.

The predictive density function $p(y|\underline{x})$ is given by setting k=1 in (4.3.6) and by setting $(g,h) = (nk,n\bar{x}) = (4,115)$ if vague prior knowledge is assumed for Θ. The posterior moments of either $p(y|\Theta)$ or $\log p(y|\Theta)$ can be calculated and 95% posterior credibility intervals obtained using Pearson curves. By considering a sequence of possible future values, y=0,5,10,..,100, a credibility interval band can be obtained for $p(y|\Theta)$ or $\log p(y|\Theta)$. In the diagram below the predictive density function $p(y|\underline{x}) = E[p(y|\Theta)]$ is plotted together with the credibility interval band for $p(y|\Theta)$, and from this plot the lack of stability of the predictive density function can be observed.

It is interesting to note that the posterior variance of $\log p(y|\Theta)$ has a single turning point, a minimum value when $y = h/g$, i.e. $y = \bar{x}$ assuming vague prior knowledge for Θ. For the example above $\bar{x} = 28.75$.

However the turning points of the variance of $p(y|\Theta)$ are given by the equation,

$$g.h^{g}.(h+2y)^{g+3} = (g+2).(h+y)^{2g+3} . \qquad (8.2.9)$$

Substituting $w = gy/h$ into (8.2.9) results in the elimination of h, and hence the number of turning points of the variance of $p(y|\Theta)$ depends only on g. By considering the resulting left and right functions in (8.2.9), it can be shown that for integer $g \leq 3$, equation (8.2.9) has no positive solutions for y, while for integer $g \geq 4$, equation (8.2.9) has exactly two positive solutions for w both in the range (1,2) and hence the corresponding solutions for y are in the range (h/g,2h/g). Indeed as $g \to \infty$ the solutions for w denoted by w_1 and w_2 converge to the values 1 and 2 respectively and hence the corresponding solutions for y converge to the values h/g and 2h/g. Note that for vague prior knowledge about parameter Θ, $h/g = \bar{x}$.

This is also verified by numerical calculation of the solutions w_1 and w_2 for different values of g,

g	4	5	6	7	8	9	10	50	100
w_1	1.221	1.152	1.117	1.095	1.080	1.069	1.061	1.010	1.005
w_2	1.806	1.884	1.922	1.943	1.957	1.966	1.973	1.977	1.9997

In conclusion the variance of $p(y|\Theta)$ has a minimum value at y slightly greater than h/g and a maximum value at y slighly smaller than 2h/g.

In the example above, if vague prior knowledge is assumed for Θ, then (g,h) = (4,115), and equation (8.2.9) has exactly two solutions at y=35 and y=52,each to the nearest integer. So the posterior variance of $p(y|\Theta)$ has a minimum value at y=35 and a maximum value at y=52.

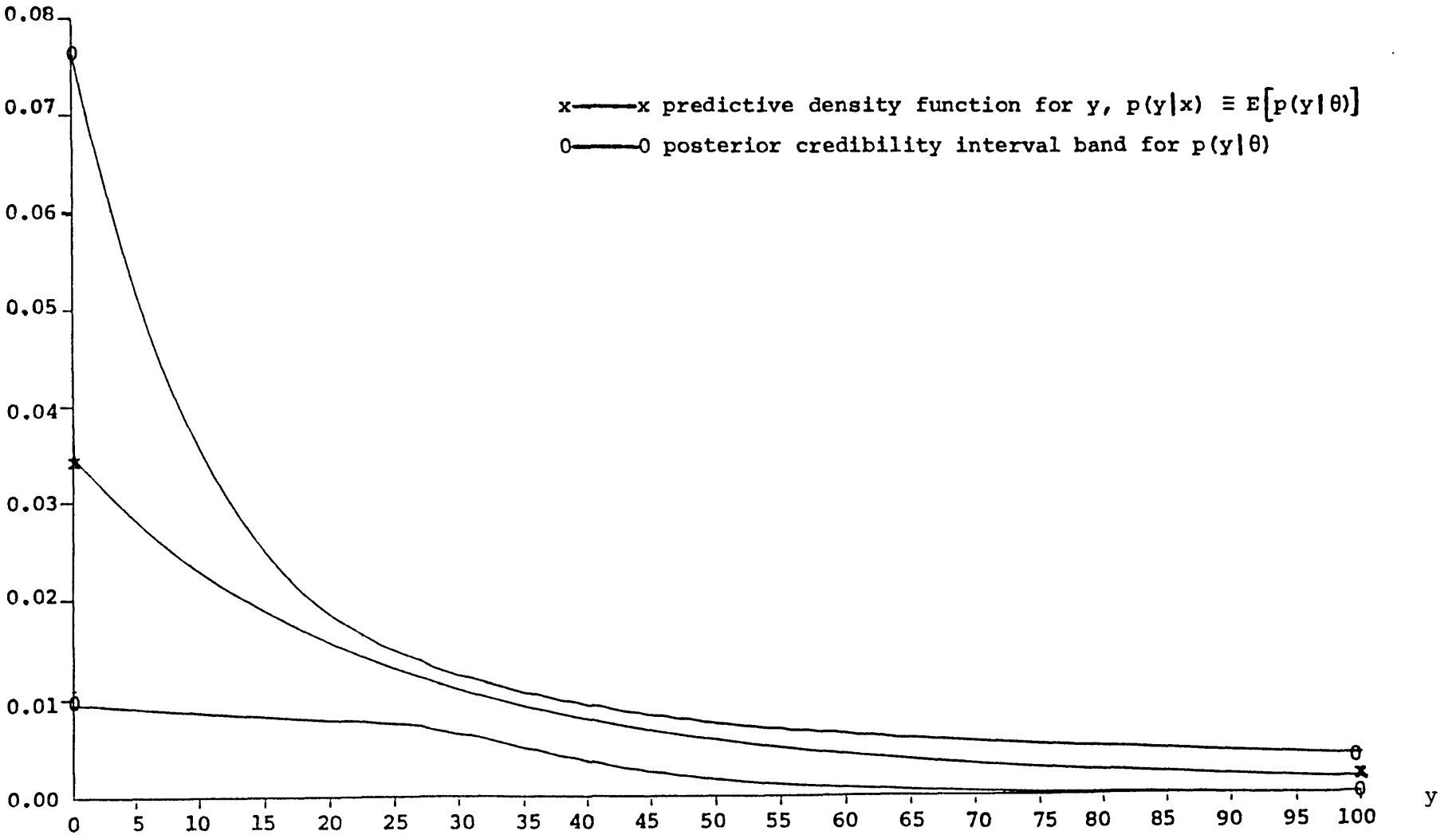


Figure 6. Posterior credibility interval band for the likelihood function around its posterior mean, the predictive density function.

8.3 Multivariate Normal linear model population

If random variable X has a multivariate Normal linear model population distribution, then $\Phi_1(t)$ is given by (7.5.18) for general m and by (7.5.24) for m=1.

Hence,

$E[p(y|\Theta)] = \Phi_1(1)$ is given by equation (7.5.26) , and

$V[p(y|\Theta)] = \Phi_1(2) - \Phi_1(1)^2$ where $\Phi_1(2)$ is given by setting t=2 in (7.5.24),

$$\Phi_1(2) = \frac{\Gamma_d[\tfrac{1}{2}(v+2)]}{[1+2w'C^{-1}w]^{\frac{1}{2}d}\,[1+A_{11}(y-b'w)'h^{-1}(y-b'w)]^{\frac{1}{2}(v+2)}\,|h|\,\pi^d\,\Gamma_d(\tfrac{1}{2}v)}$$

where $A_{11} = \dfrac{2}{1+2w'C^{-1}w}$ from (7.5.2) ,

and $\dfrac{\Gamma_d[\frac{1}{2}(v+2)]}{\Gamma_d(\frac{1}{2}v)} = \dfrac{\Gamma[\frac{1}{2}(v+1)]\;\Gamma[\frac{1}{2}(v+2)]}{\Gamma[\frac{1}{2}(v+1-d)]\;\Gamma[\frac{1}{2}(v+2-d)]}$.

Higher central moments of $p(y|\Theta)$ are obtained by setting t=3,4,.. in equation (7.5.24) and using (8.1.3).

Also the central moments of log $p(y|\Theta)$ are obtained using (8.1.4) from the cumulants of log $p(y|\Theta)$ which in turn are found from equations (2.[illegible] with appropriate substitutions as described in section 7.5.9. In particular,

$$E[\log p(y|\Theta)] = -\tfrac{1}{2}d\log\pi - \frac{d}{2c} - \tfrac{1}{2}\log|h| - \tfrac{1}{2}vD + \tfrac{1}{2}\sum_{i=1}^{d}\Psi[\tfrac{1}{2}(v-i+1)] \quad \text{and}$$

$$V[\log p(y|\Theta)] = \frac{d}{2c^2} + \left[\frac{v-c}{c}\right]D + \tfrac{1}{2}vD^2 + (\tfrac{1}{2})^2\sum_{i=1}^{d}\Psi^{(1)}[\tfrac{1}{2}(v-i+1)] \qquad (8.3.1)$$

where $c = [w'C^{-1}w]^{-1}$ and $D = (y-b'w)'h^{-1}(y-b'w)$.

The joint posterior moment generating function of $l_1 = \log p(y_1|\Theta)$ and $l_2 = \log p(y_2|\Theta)$

To obtain the posterior correlation coefficients $\rho(p_1,p_2)$ and $\rho(l_1,l_2)$, the joint moment generating function of l_1 and l_2, i.e. $\Phi_{l_1,l_2}(t_1,t_2)$, is required. Setting m=2 in (7.5.18) gives,

$$\Phi_{l_1,l_2}(t_1,t_2) = \frac{\Gamma_d[\tfrac{1}{2}(v+t_1+t_2)]}{|I_2+\Delta WC^{-1}W'|^{\frac{1}{2}d}\ |I_2+AD|^{\frac{1}{2}(v+t_1+t_2)}|h|^{\frac{1}{2}(t_1+t_2)}\pi^{\frac{1}{2}d(t_1+t_2)}\Gamma_d(\tfrac{1}{2}v)} \qquad (8.3.2)$$

where $D = (Y-Wb)h^{-1}(Y-Wb)'$, then the $(i,j)^{th}$ element of D is given by

$D_{ij} = (y_i-b'w_i)'h^{-1}(y_j-b'w_j)$, for i=1,2 and j=1,2.

Also, using the results of section 7.5.7, A is a 2×2 matrix with elements A_{ij} given by,

$$A_{11} = \frac{t_1(1+t_2e_{22})}{g(t_1,t_2)} ,$$

$$A_{22} = \frac{t_2(1+t_1e_{11})}{g(t_1,t_2)} \quad \text{and,}$$

$$A_{12} = -\frac{t_1t_2e_{12}}{g(t_1,t_2)} \quad \text{and hence,}$$

$$|I_2+AD| = 1 + \frac{1}{g}[b+\delta t_1t_2] \quad \text{and,} \qquad (8.3.3)$$

$|I_2 + \Delta WC^{-1}W'| = g$, where,

$$g = g(t_1,t_2) = (1 + t_1e_{11} + t_2e_{22} + t_1t_2[e_{11}e_{22}-e_{12}^2]) , \qquad (8.3.4)$$

$$b = b(t_1,t_2) = t_1(1+t_2e_{22})D_{11} + t_2(1+t_1e_{11})D_{22} - 2t_1t_2e_{12}D_{12} , \qquad (8.3.5)$$

$$\delta = D_{11}D_{22} - D_{12}^2 \quad \text{and} \qquad (8.3.6)$$

$$e_{ij} = w_i'C^{-1}w_j \quad \text{for i=1,2 and j=1,2.} \qquad (8.3.7)$$

The posterior correlation coefficient between $p_1 = p(y_1|\theta)$ and $p_2 = p(y_2|\theta)$

The joint posterior moments of p_1 and p_2 are now easily found by appropriate substitution for (t_1,t_2). In particular the posterior correlation coefficient between p_1 and p_2 is given by (8.1.5) where

$$E(p_1p_2) = \Phi_{l_1,l_2}(1,1)$$

$$= \frac{\Gamma[\frac{1}{2}(v+1)]\,\Gamma[\frac{1}{2}(v+2)]\,g_{11}^{\frac{1}{2}(v-d+2)}}{(g_{11}+b_{11}+\delta)^{\frac{1}{2}(v+2)}\,|h|\,\pi^d\,\Gamma[\frac{1}{2}(v+1-d)]\,\Gamma[\frac{1}{2}(v+2-d)]}\quad , \qquad (8.3.8)$$

where,

$$g_{11} = g(1,1) = 1 + e_{11} + e_{22} + e_{11}e_{22} - e_{12}^2 \quad \text{and} \qquad (8.3.9)$$

$$b_{11} = b(1,1) = (1+e_{22})D_{11} + (1+e_{11})D_{22} - 2e_{12}D_{12}\ . \qquad (8.3.10)$$

The posteior correlation coefficient between $l_1=\log p(y_1|\theta)$ and $l_2=\log p(y_2|\theta)$

The joint posterior cumulants of l_1 and l_2 are easily found since for $(r_1,r_2)\geq(1,1)$,

$$(K_{r_1,r_2})_{l_1,l_2} = \left[\frac{d^{(r_1)}}{dt_1^{(r_1)}}\,\frac{d^{(r_2)}}{dt_2^{(r_2)}}\,\log\,\Phi_{l_1,l_2}(t_1,t_2)\right]_{(t_1,t_2)=(0,0)}$$

$$= D_{(0,0)}^{(r_1,r_2)}\log\,\Phi_{l_1,l_2}(t_1,t_2)$$

$$= D_{(0,0)}^{(r_1,r_2)}\left[-\tfrac{1}{2}d.\log\,g(t_1,t_2)\right]$$

$$+ D_{(0,0)}^{(r_1,r_2)}\left[-\tfrac{1}{2}(v+t_1+t_2).\log\,|I_2+AD|\right]$$

$$+ D_{(0,0)}^{(r_1,r_2)}\left[\log\,\Gamma_d[\tfrac{1}{2}(v+t_1+t_2)]\right]\ . \qquad (8.3.11)$$

Evaluating $D_{(0,0)}^{(r_1,r_2)}[\log g(t_1,g_2)]$

$$\frac{d}{dt_1} g(t_1,t_2) = g(t_1,t_2).\frac{d}{dt_1} \log g(t_1,t_2) \qquad (8.3.12)$$

Differentiating (8.3.12) (r_1-1) times with respect to t_1 using De Moivre's theorem and setting $(t_1,t_2)=(0,0)$ gives, for $r_1 \geq 1$,

$$D_{(0,0)}^{(r_1,0)}\log g(t_1,t_2) = -(r_1-1)e_{11}D_{(0,0)}^{(r_1-1,0)}\log g(t_1,t_2) = (-1)^{r_1-1}(r_1-1)!e_{11}^{r_1}$$

Further differentiating (8.3.12) r_2 times with respect to t_2 and setting $(t_1,t_2)=(0,0)$ gives, for $(r_1,r_2) \geq (2,1)$,

$$\begin{aligned} D_{(0,0)}^{(r_1,r_2)}\log g(t_1,t_2) = & -(r_1-1)e_{11}D_{(0,0)}^{(r_1-1,r_2)}\log g(t_1,t_2) \\ & -r_2e_{22}D_{(0,0)}^{(r_1,r_2-1)}\log g(t_1,t_2) \\ & -r_2(r_1-1)(e_{11}e_{22}-e_{12}^2)D_{(0,0)}^{(r_1-1,r_2-1)}\log g(t_1,t_2) . \end{aligned} \qquad (8.3.13)$$

In particular,

$$D_{(0,0)}^{(1,0)}\log g(t_1,t_2) = e_{11} , \quad D_{(0,0)}^{(1,0)}\log g(t_1,t_2) = e_{22} ,$$

and $D_{(0,0)}^{(1,1)}\log g(t_1,t_2) = -e_{12}^2$.

Evaluating $D_{(0,0)}^{(r_1,r_2)}[(v+t_1+t_2).\log |I_2+AD|]$

$$\begin{aligned} D_{(0,0)}^{(r_1,r_2)}[(v+t_1+t_2).\log |I_2+AD|] = & \; vD_{(0,0)}^{(r_1,r_2)}\log |I_2+AD| \\ & + r_1D_{(0,0)}^{(r_1-1,r_2)}\log |I_2+AD| \\ & + r_2D_{(0,0)}^{(r_1,r_2-1)}\log |I_2+AD| \end{aligned} \qquad (8.3.14)$$

where $|I_2+AD|$ is given by equation (8.3.3).

In particular,

$$D_{(0,0)}^{(1,0)} \log |I_2+AD| = D_{(0,0)}^{(1,0)} |I_2+AD| = D_{11} ,$$

$$D_{(0,0)}^{(0,1)} \log |I_2+AD| = D_{(0,0)}^{(0,1)} |I_2+AD| = D_{22} ,$$

$$D_{(0,0)}^{(1,1)} \log |I_2+AD| = -D_{12}^2 - 2e_{12}D_{12}$$

Evaluating $D_{(0,0)}^{(r_1,r_2)}[\log \Gamma_d[\frac{1}{2}(v+t_1+t_2)]\,]$

$$D^{(r_1,r_2)}[\log \Gamma_d[\tfrac{1}{2}(v+t_1+t_2)]\,] = (\tfrac{1}{2})^{r_1+r_2} \sum_{i=1}^{d} \psi^{(r_1+r_2-1)}[\tfrac{1}{2}(v+t_1+t_2-i+1)]$$

$$D_{(0,0)}^{(r_1,r_2)}[\log \Gamma_d[\tfrac{1}{2}(v+t_1+t_2)]\,] = (\tfrac{1}{2})^{r_1+r_2} \sum_{i=1}^{d} \psi^{(r_1+r_2-1)}[\tfrac{1}{2}(v-i+1)] \qquad (8.3.15)$$

Hence the posterior correlation coefficient between l_1 and l_2

is given by equation (8.1.6) where from (8.3.11) setting $(r_1,r_2)=(1,1)$,

$$(K_{1,1})_{l_1,l_1} = \tfrac{1}{2}de_{12}^2 + \tfrac{1}{2}v(D_{12}^2 + 2e_{12}D_{12}) - \tfrac{1}{2}(D_{11}+D_{22}) + (\tfrac{1}{2})^2 \sum_{i=1}^{d} \psi^{(1)}[\tfrac{1}{2}(v-i+1)] \qquad (8.3.16)$$

Note that from equation (8.3.1),

$$V(l_j) = \tfrac{1}{2}de_{jj}^2 + \tfrac{1}{2}v(D_{jj}^2 + 2e_{jj}D_{jj}) - D_{jj} + (\tfrac{1}{2})^2 \sum_{i=1}^{d} \psi^{(1)}[\tfrac{1}{2}(v-i+1)] \qquad (8.3.17)$$

and of course $\rho(l_1,l_2)=1$ if $(y_1,w_1)=(y_2,w_2)$ since then $D_{11}=D_{12}=D_{22}$ and $e_{11}=e_{12}=e_{22}$ and so $V(l_1) = V(l_2) = Cov(l_1,l_2)$.

CHAPTER 9

Combining subsets of variables

9.1 Introduction

Suppose the d variables of the d dimensional random variable X are divided into R subsets of variables, X_r; r=1,2,..,R, which are to be modelled separately, where

$$p_X(x|\Theta) = \prod_{r=1}^{R} p_{X_r}(x_r|x_{(r-1)},\Theta_r) \qquad (9.1.1)$$

and $\Theta = (\Theta_1,\Theta_2,..,\Theta_R)$ is a parameter vector combining the parameters of the R models, $x_{(r)} = (x_r,x_{r-1},..,x_1)$ for r≥1 and $p(x_1|x_{(0)},\Theta_1) = p(x_1|\Theta_1)$, where x_r is the observed value of the variable subset X_r.

Later in section 9.1 the restricted case where the R parameter vectors Θ_r have no parameters in common is considered. In section 9.2 the special case where the first variable subset X_1 is a single randon variable indicating population membership is considered. This is the usual situation in discriminant analysis where if X_1 is assumed to have a multinomial distribution with parameter vector Θ_1, then Θ_1 is the mixing parameters q. It is shown how the joint posterior cumulant generating functions of the I population log likelihood functions for the new observation y can be obtained by combining the R separate joint posterior cumulant generating functions from the R models for X_r, for r=1,2,..,R. In section 9.3 the situation is considered where, for a particular value of r≥1, the I population models for X_r, $p(x_r|x_{r-1},x_{r-2},...,x_2,x_1=i,\Theta_r)$ for i=1,2,..,I have no parameters in common, and it is shown how the joint cumulant generating function of the I population log likelihood functions for a new observation y can be expressed as a product of the I serarate cumulant generating functions. Two examples are considered. In the first example, given the value of the categorical variable X, indicating population membership, and given the parameter sets Θ_r, for r=1,2,..,R, the subsets of variables, $X_2,X_3,..,X_R$, are independent. The second example includes as a special case the 'location linear discriminant model'. Finally in section 9.4 the special case of just two populations is considered.

Let $\underline{x}$ represent a random sample of n observations of random variable X, and $x_{j,r}$ represent the observed value of variable subset X_r for observation j for j=1,2,..,n and r=1,2,..,R.

From (9.1.1) the likelihood of the data $\underline{x}$ is given by

$$L = \prod_{j=1}^{n} \prod_{r=1}^{R} p_{X_r}(x_{j,r}|x_{j,(r-1)},\Theta_r) \qquad (9.1.2)$$

Suppose the R parameter vectors Θ_r have no parameters in common and also assume that the Θ_r have independent prior distributions $p(\Theta_r)$ for r=1,2,..,R, then

$$p(\Theta) = \prod_{r=1}^{R} p(\Theta_r) \qquad (9.1.3)$$

Hence from (9.1.2) and (9.1.3), the posterior distributions of the Θ_r for r=1,2,..,R will be independent and given by

$$P(\Theta_r|\underline{x}) \propto p(\Theta_r).\prod_{j=1}^{n} p_{X_r}(x_{j,r}|x_{j,(r-1)},\Theta_r) \qquad (9.1.4)$$

The predictive density function of a possible future observation y given the data $\underline{x}$ is

$$p_X(y|\underline{x}) = \prod_{r=1}^{R} p_{X_r}(y_r|y_{(r-1)},\underline{x}) \qquad (9.1.5)$$

where $p_{X_r}(y_r|y_{(r-1)},\underline{x}) = \int p(y_r|y_{(r-1)},\Theta_r).p(\Theta_r|\underline{x})\, d\Theta_r$. (9.1.6)

9.2 Case where the first variable subset represents population membership

For the special case where the first variable subset X_1 represents a single random variable indicating population membership, i.e. $x_1=i$ when observation x belongs to population Π_i for i=1,2,..,I, then,

$$p(y_1=i|y_2,y_3,..,y_R,\Theta) = \frac{\prod_{r=1}^{R} p_{r,i}(y,\Theta_r)}{\sum_{i=1}^{I}\prod_{r=1}^{R} p_{r,i}(y,\Theta_r)} \quad \text{, and} \qquad (9.2.1)$$

$$p(y_1=i|y_2,y_3,..,y_R,\underline{x}) = \frac{\prod_{r=1}^{R} p_{r,i}(y,\underline{x})}{\sum_{i=1}^{I}\prod_{r=1}^{R} p_{r,i}(y,\underline{x})} \quad \text{, where} \qquad (9.2.2)$$

$$p_{r,i}(y,\Theta_r) = \begin{cases} p(y_r|y_{r-1},y_{r-2},..,y_2,y_1=i,\Theta_r) & \text{for } r\geq 2 \\ p(y_1=i|\Theta_1) & \text{for } r=1 \end{cases} \qquad (9.2.3)$$

$$p_{r,i}(y,\underline{x}) = \begin{cases} p(y_r|y_{r-1},y_{r-2},..,y_2,y_1=i,\underline{x}) & \text{for } r\geq 2 \\ p(y_1=i|\underline{x}) & \text{for } r=1 \end{cases} \qquad (9.2.4)$$

In discriminant analysis interest centres on which population the new observation belongs to, i.e. the value of y_1 given the values of $y_2,y_3,..,y_R$ and the past data $\underline{x}$, i.e. $p(y_1=i|y_2,y_3,..,y_R,\underline{x})$. To find this the predictive density functions $p_{r,i}(y,\underline{x})$ for r=1,2,..,R and i=1,2,..,I can be obtained seperately and later combined as in (9.2.2).

Interest here centres on the relative likelihoods of the new observation belonging to each population Π_i, i.e.

$$p(y_1=i,y_2,y_3,..,y_R|\Theta) = \prod_{r=1}^{R} p_{r,i}(y,\Theta) \quad ,$$

for i=1,2,..,I.

In particular interest here centres on the joint moment generating function of the log likelihoods of the new observation y belonging to population Π_i, for i=1,2,..,I, i.e. the joint moment generating function of $(l_i;\ i=1,2,..,I)$, where,

$$l_i = \log p(y_1=i,y_2,y_3,..,y_R|\Theta) = \sum_{r=1}^{R} \log p_{r,i}(y,\Theta) = \sum_{r=1}^{R} l_{ri}$$

where $l_{ri} = \log p_{r,i}(y,\Theta)$ for r=1,2,..,R and i=1,2,..,I.

Let $\underline{l} = (l_1,l_2,..,l_I)$, then

$$\Phi_{\underline{l}}(\underline{t}) = E\left[\exp(\underline{t}'\underline{l})\right] = E\left[\exp(\sum_{i=1}^{I} t_i l_i)\right] = E\left[\exp(\sum_{i=1}^{I}\sum_{r=1}^{R} t_i l_{ri})\right]$$

$$= \prod_{r=1}^{R} E\left[\exp(\sum_{i=1}^{I} t_i l_{ri})\right]$$

since l_{ri} depends only on the parameter set Θ_r and the Θ_r have independent posterior distributions given the data $\underline{x}$, and hence,

$$\Phi_{\underline{l}}(\underline{t}) = \prod_{r=1}^{R} \Phi_{\underline{l}_r}(\underline{t}) \quad , \qquad (9.2.5)$$

where $\underline{t} = (t_1,t_2,..,t_I)$ and $\underline{l}_r = (l_{r1},l_{r2},..,l_{rI})$ for r=1,2,..,R.

Hence the cumulant generating function of $\underline{l}$, $K_{\underline{l}}(\underline{t})$, is given by,

$$K_{\underline{l}}(\underline{t}) = \log \Phi_{\underline{l}}(\underline{t}) = \sum_{r=1}^{R} \log \Phi_{\underline{l}_r}(\underline{t}) = \sum_{r=1}^{R} K_{\underline{l}_r}(\underline{t}) \quad . \qquad (9.2.6)$$

So the posterior cumulant generating function of $\underline{l}$ can be obtained simply by summing the posterior cumulant generating functions of $\underline{l}_r$ for r=1,2,..,R. The posterior cumulants and moments of $\underline{l}$ can then easily be obtained from $K_{\underline{l}}(\underline{t})$.

Example 1 As a simple illustration of the above notation, consider the special case where random variable X comprises just two subsets of variables, the first X_1, being the categorical variable indicating the population of membership and the second, X_2, comprising all the other variables.
From (9.2.2),

$$p(y_1=i|y_2,\underline{x}) = \frac{p(y_2|y_1=i,\underline{x}).p(y_1=i|\underline{x})}{\sum_{i=1}^{I} p(y_2|y_1=i,\underline{x}).p(y_1=i|\underline{x})}$$

Here $\Theta = (\Theta_1,\Theta_2)$, where Θ_2 combines the parameters of the models for y_2, i.e. $p(y_2|y_1,\Theta_2)$ over all possible values of y_1, i.e. for $y_1=1,2,..,I$, and Θ_1 is the parameters of the model for y_1, i.e. $p(y_1|\Theta_1)$. If X_1 has a multinomial distribution given Θ_1, then Θ_1 is the multinomial parameter vector, i.e. $\Theta_1 = q$, the mixing parameters, in the notation of chapter 7 and $p(y_1=i|\Theta_1) = q_i$, for $i=1,2,..,I$.
Note that from section 7.3,

$$p(y_1=i|\underline{x}) = E[p(y_1=i|\Theta_1)] = E[q_i] = \frac{g_i}{\sum_{i=1}^{n} g_i} ,$$

where for vague prior knowledge about $\Theta_1 = q$, $g_i = n_i$, for $i=1,2,..,I$, where n_i is the number of observations in the data $\underline{x}$ with $X_1 = i$, i.e. the number of observations in $\underline{x}$ from population Π_i.

Also from (9.2.6),

$$K_{\underline{l}}(\underline{t}) = K_{\underline{l}_1}(\underline{t}) + K_{\underline{l}_2}(\underline{t}) ,$$

and from equation (7.3.4),

$$K_{\underline{l}_1}(\underline{t}) = \log \Gamma[\sum_{i=1}^{I} g_i] - \log \Gamma[\sum_{i=1}^{I} (g_i+t_i)] - \sum_{i=1}^{I} [\log\Gamma(g_i) - \log\Gamma(g_i+t_i)]$$

This situation was discussed in section 7.4.

9.3 Populations with no parameters in common

The parameter set Θ_r combines the parameters from the I sets of models for $p(x_r|x_{r-1},x_{r-2},..,x_2,x_1=i,\Theta_r)$, for i=1,2,..,I. Suppose that for some value of $r\geq 1$ these I sets of models have no parameters in common.

Then let $\Theta_r = (\Theta_{r1},\Theta_{r2},..,\Theta_{rI})$.

Let random variable X comprise r subsets of variables $X_1,X_2,..,X_R$, where as in section 9.2, X_1 indicates the population of membership.

Hence $p(x_r|x_{r-1},x_{r-2},..,x_2,x_1=i,\Theta_r) = p(x_r|x_{r-1},x_{r-2},..,x_2,x_1=i,\Theta_{ri})$, for i=1,2,..,I and some $r\geq 1$.

Suppose independent random samples $x_i = (x_{i1},x_{i2},..,x_{in_i})$ are identified as belonging to Π_i, for i=1,2,..,I.

Let $x_{ij} = (x_{ij,r}\ ;\ r=1,2,..,R)$, where $x_{ij,r}$ is the observed value of variable subset X_r for the j^{th} observation identified as belonging to population Π_i, for i=1,2,..,R, j=1,2,..,n_i and i=1,2,..,I.

Assume that the parameters Θ_{ri}, i=1,2,..,I, have independent prior distributions, $p(\Theta_{ri})$, then,

$$p(\Theta_r) = \prod_{i=1}^{I} p(\Theta_{ri}) \quad . \tag{9.3.1}$$

From (9.1.4),

$$p(\Theta_r|\underline{x}) \propto \left[\prod_{i=1}^{I} p(\Theta_{ri})\right] . \prod_{i=1}^{I}\prod_{j=1}^{n_i} p(x_{ij,r}|x_{ij,(r-1)},\Theta_{ri}) \quad ,$$

$$= \prod_{i=1}^{I} \left[p(\Theta_{ri}) . \prod_{j=1}^{n_i} p(x_{ij,r}|x_{ij,(r-1)},\Theta_{ri}) \right] \quad . \tag{9.3.2}$$

Hence the Θ_{ri}, i=1,2,..,I, have independent posterior distributions,

$$p(\Theta_{ri}|\underline{x}) \propto p(\Theta_{ri}) . \prod_{j=1}^{n_i} p(x_{ij,r}|x_{ij,(r-1)},\Theta_{ri}) \quad . \tag{9.3.3}$$

Hence for i=1,2,..,I, and r=2,3,..,R, (9.2.3) gives,

$$p_{r,i}(y,\Theta) = p(y_r|y_{r-1},y_{r-2},..,y_2,y_1=i,\Theta_{ri}) \quad , \tag{9.3.4}$$

and (9.2.4) gives,

$$p_{r,i}(y,\underline{x}) = p(y_r|y_{r-1},y_{r-2},..,y_2,y_1=i,\underline{x})$$

$$= \int p(y_r|y_{r-1},y_{r-2},..,y_2,y_1=i,\Theta_{ri}).p(\Theta_{ri}|\underline{x})\, d\Theta_{ri} \quad . \tag{9.3.5}$$

Equations (9.3.4) and (9.3.5) can be substituted directly into equations (9.2.1) and (9.2.2).

Furthermore,

$$\Phi_{\underline{l}_r}(\underline{t}) = E\left[\exp\left(\sum_{i=1}^{I} t_i l_{ri}\right)\right] ,$$

$$= \prod_{r=1}^{I} E\left[\exp(t_i l_{ri})\right] ,$$

since l_{ri} depends only on parameter θ_{ri} and the θ_{ri} have independent posterior distributions given the data $\underline{x}$, and hence,

$$\Phi_{\underline{l}_r}(\underline{t}) = \prod_{i=1}^{I} \Phi_{l_{ri}}(t_i) ,$$

$$\text{and so } K_{\underline{l}_r}(\underline{t}) = \sum_{i=1}^{I} K_{l_{ri}}(t_i) . \qquad (9.3.6)$$

Note that the mixing parameters θ_1 are common to all populations, Π_i; i=1,2,..,I, and hence result (9.3.6) does not apply for r=1.

If, for r=2,..,R, the I sets of models for $p(x_r|x_{r-1},x_{r-2},..,x_2,x_1=i,\theta_{ri})$ have no parameters in common, and the prior distributions of the θ_{ri} are all independent then,

$$K_{\underline{l}_r}(\underline{t}) = \sum_{i=1}^{I} K_{l_{ri}}(t_i) \quad \text{for } r=2,3,..,R,$$

and hence from (9.2.6),

$$K_{\underline{l}}(\underline{t}) = \sum_{r=1}^{R} K_{\underline{l}_r}(\underline{t}) = K_{\underline{l}_1}(\underline{t}) + \sum_{r=2}^{R} \sum_{i=1}^{I} K_{l_{ri}}(t_i) . \qquad (9.3.7)$$

Hence the posterior cumulants and moments of $\underline{l}$ can easily be found.

Example 2 Suppose that the R sets of models for $p(x_r|x_{(r-1)},\theta_r)$ for r=1,2,..,R have no parameters in common and also that, given the value of the categorical variable X_1 indicating the population of membership and given the parameter sets θ_r, r=1,2,..,R, the subsets of variables $X_2,X_3,..,X_R$ are independent, then for a new observation y, (9.2.3) gives,

$$p_{r,i}(y,\theta_r) = \begin{cases} p(y_r|y_1=i,\theta_r) & \text{for } r\geq 2 \\ p(y_1=i|\theta_1) & \text{for } r=1 \end{cases} \qquad (9.3.8)$$

and (9.2.4) gives,

$$p_{r,i}(y,\underline{x}) = \begin{cases} p(y_r|y_1=i,\underline{x}) & \text{for } r\geq 2 \\ p(y_1=i|\underline{x}) & \text{for } r=1 \end{cases} \qquad (9.3.9)$$

where $p(y_r|y_1=i,\underline{x}) = \int p(y_r|y_1=i,\theta_r).p(\theta_r|\underline{x})\, d\theta_r$.

From (9.2.6),

$$K_{\underline{1}}(\underline{t}) = \sum_{r=1}^{R} K_{\underline{1}_r}(\underline{t}) \quad . \qquad (9.3.10)$$

If further, for one or more values of r(>1), the I models for $p(x_r|x_1=i,\theta_{ri})$ for i=1,2,..,I, have no parameters in common, then from (9.3.6),

$$K_{\underline{1}_r}(\underline{t}) = \sum_{i=1}^{I} K_{1_{ri}}(t_i) \quad . \qquad (9.3.11)$$

Hence if the I models for $p(x_r|x_1=i,\theta_{ri})$, for i=1,2,..,I, are members of the Exponential family with no parameters in common e.g. Gamma, Binomial or Poisson, then $K_{1_{ri}}(t_i)$ can be obtained from section 4.2.4 and if the I models are multivariate Normal linear models with unequal error precision matrices, then $K_{1_{ri}}(t_i)$ can be obtained from section 7.5, for i=1,2,..,I.Equation (9.3.11) can then be applied to obtain $K_{\underline{1}_r}(\underline{t})$.

If, however, the I models for $p(x_r|x_1=i,\theta_{ri})$, for i=1,2,..,I, have parameters in common, e.g. multivariate Normal linear model populations with equal error precision matrices, then (9.3.11) cannot be applied, but $K_{\underline{1}_r}(\underline{t})$ can be obtained directly.

For the equal error precision matrix case, $K_{\underline{1}_r}(\underline{t})$ can be obtained from section 7.6.

Example 3 Suppose the random variable X comprises three subsets of variables, the first X_1 being the categorical variable indicating population of membership, the second X_2 comprising all other categorical variables and the third X_3 comprising continuous variables only.

Assuming as usual a multinomial model for $p(x_1|\Theta_1)$, then models are required for $p(x_2|x_1,\Theta_2)$ and $p(x_3|x_2,x_1,\Theta_3)$.
A possible model for $p(x_3|x_2,x_1,\Theta_3)$ is the multivariate analysis of variance (MANOVA) model. In the case of equal population error precision matrices, then in the notation of section 7.6,

$$X_3 = ZB + \epsilon$$

where $\epsilon \sim No_{n,d}(0,I_n,T)$, $\Theta_3 = (B,T)$ and Z is the design matrix for the MANOVA model chosen based on the categorical variables X_2.
The cumulant generating function for $\underline{l}_3 = (\log p(y_3|y_2,y_1=i,\Theta_3);\ i=1,2,..,I)$ was obtained in section 7.6.

If unequal error precision matrices are assumed for the populations then

$$X_{3i} = Z_iB_i + \epsilon_i$$

where $\epsilon_i \sim No_{n_i,d}(0,I_{n_i},T_i)$, independently for i=1,2,..,I,

$\Theta_{3i} = (B_i,T_i)$ and Z_i is the design matrix for the observations from population Π_i (i.e. for which $X_1=i$) for the MANOVA model chosen based on the categorical variables X_2. The cumulant generating function for $l_{3i} = \log p(y_3|y_2,y_1=i,\Theta_{3i})$ was obtained in section 7.5.

A possible model for $p(x_2|x_1,\Theta_2)$ is the simple multinomial model. If the equal error precision matrix case above is combined with a simple multinomial model for $p(x_2|x_1,\Theta_2)$ this gives the 'location linear discriminant model' discussed extensively from a Classical viewpoint by Krzanowski (1975, 1979, 1980) and also by Vlachonikolis and Mariott (1982). A general algorithm for maximum likelihood estimation in the presence of missing values which can be applied to the 'location linear discriminant model' has been given by Little and Schluchter (1985).

9.4 Special case of two populations

From equation (9.2.5), provided that the R sets of models for $p(x_r|x_{(r-1)},\theta_r)$ have no parameters in common then,

$$\Phi_{\underline{l}}(\underline{t}) = \prod_{r=1}^{R} \Phi_{\underline{l}_r}(\underline{t}) \quad . \tag{9.4.1}$$

Let $L = l_1 - l_2$, where $l_i = \sum_{r=1}^{R} l_{ri}$ for r=1,2, then,

$$\Phi_L(t) = \Phi_{l_1,l_2}(t,-t) \quad ,$$

$$= \prod_{r=1}^{R} \Phi_{l_{r1},l_{r2}}(t,-t) \quad \text{using (9.4.1), and hence,}$$

$$K_L(t) = \sum_{r=1}^{R} K_{l_{r1},l_{r2}}(t,-t) = \sum_{r=1}^{R} K_{L_r}(t) \quad \text{where } L_r = l_{r1} - l_{r2} \tag{9.4.2}$$

If the populations Π_1 and Π_2 have no parameters in common (apart from the mixing parameters θ_1), then l_{r1} and l_{r2} have independent posterior distributions for r=2,3,..,R, and hence from (9.4.2),

$$K_L(t) = K_{l_{11},l_{12}}(t,-t) + \sum_{r=2}^{R} K_{l_{r1}}(t) + \sum_{r=2}^{R} K_{l_{r2}}(-t) \quad . \tag{9.4.3}$$

In the particular case where the random variable X is seperated into just two subsets of variables, the first being X_1, the categorical variable indicating population if membership, then (9.4.2) gives,

$$K_L(t) = K_{l_{11},l_{12}}(t,-t) + K_{l_{21},l_{22}}(t,-t) = K_{L_1}(t) + K_{L_2}(t) \tag{9.4.4}$$

and (9.4.3) gives,

$$K_L(t) = K_{l_{11},l_{12}}(t,-t) + K_{l_{21}}(t) + K_{l_{22}}(-t) \quad , \tag{9.4.5}$$

where from equation (7.3.4) with I=2,

$$K_{l_{11},l_{12}}(t,-t) = -\log \Gamma(g_1) - \log \Gamma(g_2) + \log \Gamma(g_1+t) + \log \Gamma(g_2-t) \quad .$$

The cumulants and moments of L can be found from $K_L(t)$ and the posterior distribution of L approximated using Pearson curves. The posterior distribution of P can then be obtained by transformation as in section 1.3.

CHAPTER 10

Using the extra information concerning the parameters from new observations of unknown population membership

10.1 Introduction

So far the posterior distribution of P has been approximated using information about the unknown parameters Θ given by the random samples, $\underline{x}_1,\underline{x}_2,..,\underline{x}_I$ from $\Pi_1,\Pi_2,..,\Pi_I$, respectively. Any information about Θ in the new observation y has not been used. This would be appropriate if the information about Θ from y was negligible compared with the information from the random samples or if it was not certain that the new observation y was sampled (independently) from the combined population distribution.

Suppose however there are one or more new observations, of unknown population membership, which are confidently believed to have been randomly drawn from the combined population distribution. Let $\underline{x}_1,\underline{x}_2,..,\underline{x}_I$ be independent random samples obtained by mixture sampling and identified as belonging to $\Pi_1,\Pi_2,..,\Pi_I$ respectively and let $\underline{y} = (y_1,y_2,..,y_m)$ be a random sample of m new observations, of unknown population membership, obtained independently of $\underline{x}_1,\underline{x}_2,..,\underline{x}_I$ from the combined population distribution $\Sigma q_i p_i(y|\Theta)$. Let y, not neccessarily one of $(y_1,y_2,..,y_m)$, be a new observation for which an estimate, posterior distribution and credibility interval for P is required.

It is interesting to note, as pointed out in section 5.7 of Titterington, Smith and Makov (1985), that if a logistic model relates the probability of population membership to the variables measured, i.e.

$p(y,i|\beta,\gamma) = p(i|y,\beta).p(y|\gamma)$, where,

$p(y\in\Pi_i|y,\beta) = \exp(\beta_i'y).p(y\in\Pi_I|y,\beta)$, for i=1,2,..,I-1, (10.1.1)

then the uncategorised new observations provide no information at all about β, since no parametric distributional model is thereby specified concerning the population distributions and hence uncategorised observations provide no extra information to assist with the estimation of β and hence $p(y\in\Pi_i|y,\beta)$.
If, however, a discriminant analysis model is appropriate with the population distributions specified in functional form but with unknown parameters, i.e.

$p(y,i|q,\Theta) = p(y|i,\Theta).p(i|q) = q_i p_i(y|\Theta)$, (10.1.2)

then the uncategorised observations provide information about the parameters Θ and q as in equation (10.2.1) below, and hence provide extra information to assist with the estimation of Θ and q and hence $q_i p_i(y|\Theta)$.
The two alternative models (10.1.1) and (10.1.2) have been mixed by Anderson (1979).

10.2 'Estimative' estimate of P

The 'estimative' estimate of P cannot be found explicitly. The maximum likelihood estimates of Θ can be obtained, however this will, in general, require numerical maximisation of the likelihood function which may be difficult particularly when the number of parameters in Θ is large. The likelihood function is given by

$$L(\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_I,\underline{y};q,\Theta) = \left[\prod_{i=1}^{I}\prod_{j=1}^{n_i} q_i p_i(x_{ij}|\Theta)\right].\prod_{j=1}^{m}\left[\sum_{i=1}^{I} q_i p_i(y_j|\Theta)\right]. \quad (10.2.1)$$

This corresponds to the likelihood function defined by equation (1.2.3) from Titterington, Smith and Makov (1985, Ch1). In their section 4.3, they discuss use of the EM algorithm (Dempster, Laird and Rubin, 1977), or the alternative numerical algorithms, Newton-Raphson (NR), and the Method of Scoring (MS), to obtain maximum likelihood estimates of the parameters. In their section 6.4.2 they reconsider the problem by regarding the data as incomplete, the incompleteness referring to the absence of the indicator variables which would identify the true population of membership of each of the new observations. A general form of recursive algorithm for maximum likelihood estimation, together with its asymptotic properties is considered. Titterington (1984) gives regularity conditions for unique maximum likelihood estimates. Redner and Walker (1984) applied the EM algorithm to Normal mixtures and found that although convergence may sometimes be slow, only a few iterations are required to get close to the maximum value of the likelihood function. This suggests a composite algorithm, in which a few EM algorithms are followed by a few further iterations of the much faster MS or NR algorithms.

A general algorithm for maximum likelihood estimation in the presence of missing values has also been considered by Little and Schluchter (1985).

The value or efficiency of unclassified observations relative to classified observations in terms of asymptotic error rates has been investigated by O'Neill (1978) for equal covariance multivariate Normal populations. O'Neill found that the asymptotic relative efficiency increases rapidly with the separation of the populations. Ganesalingam and McLachlan (1979) have investigated the relative efficiency for small samples and found it generally higher than the asymptotic value.

10.3 'Predictive' estimate of P

The 'predictive' estimate of P can be obtained explicitly, provided that the mixing parameters are known, since from a fully Bayes approach,

$$p(y_j \in \Pi_{k_j} \text{ for } j=1,2,\ldots,m \mid \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_I, \underline{y})$$

$$= \frac{\left[\prod_{j=1}^{m} q_{k_j}\right] . p(\underline{y} \mid \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_I, y_j \in \Pi_{k_j} \text{ for } j=1,2,\ldots,m)}{\sum_{i_1=1}^{I} \sum_{i_2=1}^{I} \cdots \sum_{i_m=1}^{I} \left[\prod_{j=1}^{m} q_{k_j}\right] . p(\underline{y} \mid \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_I, y_j \in \Pi_{k_j} \text{ for } j=1,2,\ldots,m)}$$

where $k_j \in (1,2,\ldots,I)$ for $j=1,2,\ldots,m$, and

where $p(\underline{y} \mid \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_I, y_j \in \Pi_{k_j} \text{ for } j=1,2,\ldots,m)$

is the joint predictive distribution of the m new observations $\underline{y}$, i.e.

$$p(\underline{y} \mid \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_I, y_j \in \Pi_{k_j} \text{ for } j=1,2,\ldots,m)$$

$$= \int p(\underline{y} \mid \Theta, y_j \in \Pi_{k_j} \text{ for } j=1,2,\ldots,m) . p(\Theta \mid \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_I) d\Theta$$

The marginal $p(y_j \in \Pi_{k_j} \mid \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_I, \underline{y})$ for any $j \in (1,2,\ldots,m)$

can then be obtained since

$$p(y_j \in \Pi_{k_j} \mid \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_I, \underline{y})$$

$$= \sum_{k_1=1}^{I} \cdots \sum_{k_{j-1}=1}^{I} \sum_{k_{j+1}=1}^{I} \cdots \sum_{k_m=1}^{I} p(y_j \in \Pi_{k_j} \text{ for } j=1,2,\ldots,m \mid \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_I, \underline{y})$$

First consider the case where the populations have no parameters in common.

If the populations $\Pi_1, \Pi_2, \ldots, \Pi_I$ have no parameters in common then the joint predictive distribution of $\underline{y}$ factorises into the product of the joint predictive distribution of the sets of observations belonging to the separate populations since the integration over θ can be factorised into integrations over each of the population parameter sets separately, e.g.

$$p(y_1, y_2, y_3, y_4 | \underline{x}_1, \underline{x}_2, y_1 \in \Pi_1, y_2 \in \Pi_1, y_3 \in \Pi_2, y_4 \in \Pi_2)$$

$$= p(y_1, y_2 | \underline{x}_1, \underline{x}_2, y_1 \in \Pi_1, y_2 \in \Pi_1) . p(y_3, y_4 | \underline{x}_1, \underline{x}_2, y_3 \in \Pi_2, y_4 \in \Pi_2)$$

Hence essentially all that is required is the joint predictive distribution for m new observations $\underline{y}$ given that all the m new observations come from the <u>same</u> population Π, when an independent random sample $\underline{x}$ of size n from Π is available, i.e.

$$p(\underline{y} | \underline{x}, \underline{x} \in \Pi, \underline{y} \in \Pi) = \int p(\underline{y} | \theta) . p(\theta | \underline{x}) d\theta = \int \left[\prod_{j=1}^{m} p(y_j | \theta) \right] . p(\theta | \underline{x}) d\theta$$

The following results are obtained for different population distributions.

Gamma population Π

$p(y_j|\Theta) = Ga(k,\Theta)$,

$p(\Theta|\underline{x}) = Ga(g,h)$,

$$p(\underline{y}|\underline{x},\underline{x}\in\Pi,\underline{y}\in\Pi) = \frac{\left[\prod_{j=1}^{m} y_j^{k-1}\right].h^g\Gamma(mk+g)}{[\Gamma(k)]^m.\Gamma(g).(h+\sum_{j=1}^{m} y_j)^{mk+g}} \quad ,$$

where $y_j \geq 0$ for $j=1,2,\ldots,m$, from (4.3.4).

Binomial population Π

$p(y_j|\Theta) = Bi(k,\Theta)$,

$p(\Theta|\underline{x}) = Be(g,h)$,

$$p(\underline{y}|\underline{x},\underline{x}\in\Pi,\underline{y}\in\Pi) = \left[\prod_{j=1}^{m}\binom{k}{y_j}\right].\frac{B\left[g+\sum_{j=1}^{m} y_j,\ h+mk-\sum_{j=1}^{m} y_j\right]}{B(g,h)} \quad ,$$

where $y_j=0,1,2,\ldots,k$ for $j=1,2,\ldots,m$, from (4.4.4).

Poisson population Π

$p(y_j|\Theta) = Po(\Theta)$,

$p(\Theta|\underline{x}) = Ga(g,h)$,

$$p(\underline{y}|\underline{x},\underline{x}\in\Pi,\underline{y}\in\Pi) = \frac{h^g \ . \ \Gamma(g+\sum_{j=1}^{m} y_j)}{\Gamma(g).\left[\prod_{j=1}^{m} y_j!\right].(h+m)^{g+\Sigma y_j}} \quad ,$$

where $y_j=0,1,2,3,\ldots$ for $j=1,2,\ldots,m$, from (4.5.4).

Multivariate Normal population Π

$p(y_j|\Theta) = p(y_j|\mu,T) = No_d(\mu,T)$,

$p(\Theta|\underline{x}) = p(\mu,T|\underline{x}) = NoWi_d(\beta,c,g,h)$,

$$p(\underline{y}|\underline{x},\underline{x}\in\Pi,\underline{y}\in\Pi) = \left[\frac{c}{c+m}\right]^{d/2} \cdot \frac{\Gamma_d[\tfrac{1}{2}(v+m)]}{\Gamma_d[\tfrac{1}{2}v].\pi^{md/2}} \cdot \frac{|h|^{v/2}}{|H|^{(v+m)/2}} \quad , \qquad (10.3.1)$$

where $H = h + S_{\underline{y}} + \frac{mc}{m+c}.(\bar{y}-\beta)(\bar{y}-\beta)'$,

and $S_{\underline{y}} = \sum_{j=1}^{m}(y_j-\bar{y})(y_j-\bar{y})'$.

Equation (10.3.1) was obtained from (7.5.14) by the following substitutions :

$k = 1$,

$Y = [y_1, y_2, .., y_m]'$, a m×d matrix,

$X = [x_1, x_2, .., x_n]'$, a n×d matrix,

$W = [1, 1, .., 1]'$, a m×1 vector,

$Z = [1, 1, .., 1]'$, a n×1 vector,

$B = \mu'$, a 1×d vector,

$b = \beta'$, a 1×d vector,

$C = c$, a scalar, and

$A_1 = \left[I_m - \frac{1}{m+c} 1_m \right]$, where 1_m is the m×m matrix with all elements equal to 1.

Multivariate Normal linear model population Π

$p(y_j|\Theta) = p(y_j|B,T) = No_d(w_j'B, T)$,

$p(\Theta|\underline{x}) = p(B,T|\underline{x}) = NoWi_{k,d}(b,C,v,h)$ from (7.5.5), then

$p(\underline{y}|\underline{x}, \underline{x}\in\Pi, \underline{y}\in\Pi) = t_{m,d}[Wb, A_1^{-1}, h, v-d+1]$, given by equation (7.5.14).

Now consider the case where the populations have parameters in common.

Suppose the populations $\Pi_1,\Pi_2,\ldots,\Pi_I$ have parameters in common. Then the joint predictive distribution of $\underline{y}$ does not factorise as above. Instead the joint predictive distribution of the m_i observations

$$\underline{y}_i=(y_{i1},y_{i2},\ldots,y_{im_i})$$

from population Π_i for $i=1,2,\ldots,I$, is given by

$$p(\underline{y}_1,\underline{y}_2,\ldots,\underline{y}_I|\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_I,\underline{y}_1\in\Pi_1,\underline{y}_2\in\Pi_2,\ldots,\underline{y}_I\in\Pi_I)$$

$$=\int\left[\prod_{i=1}^{I}\prod_{j=1}^{m_i}p(y_{ij}|\theta,y_{ij}\in\Pi_i)\right]\cdot p(\theta|\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_I)\,d\theta$$

where $\underline{x}_i$ is an independent random sample from population Π_i for $i=1,2,\ldots,I$.

Multivariate Normal populations

Using the notation of Chapter 6 :

$$p(y_{ij}|\theta,y_{ij}\in\Pi_i)=p(y_{ij}|\mu,T,y_{ij}\in\Pi_i)=No_d(\mu_i,T)$$

$$p(\theta|\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_I)=p(\mu,T|\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_I)=No^{I}Wi_d(\beta,\underline{c},v,h)=No_{Id}Wi_{I,d}(b,C,v,h)\ ,$$

where $b=(\beta_1,\beta_2,..,\beta_I)'$ and $C=\mathrm{diag}(c_1,c_2,..,c_I)$ are $Id\times d$ and $I\times I$ matrices

$$p(\underline{y}_1,\underline{y}_2,\ldots,\underline{y}_I|\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_I,\underline{y}_1\in\Pi_1,\underline{y}_2\in\Pi_2,\ldots,\underline{y}_I\in\Pi_I)$$

$$=\left[\prod_{i=1}^{I}\left[\frac{c_i}{c_i+m_i}\right]^{d/2}\right]\cdot\frac{\Gamma_d[\tfrac{1}{2}(v+\Sigma m_i)]}{\Gamma_d[\tfrac{1}{2}v].\pi^{d\Sigma m_i/2}}\cdot\frac{|h|^{v/2}}{|H|^{(v+\Sigma m_i)/2}}\qquad(10.3.2)$$

where $H=h+\sum_{i=1}^{I}S_{\underline{y}i}+\sum_{i=1}^{I}\frac{m_ic_i}{m_i+c_i}.(\bar{y}_i-\beta_i)(\bar{y}_i-\beta_i)'$

and $S_{\underline{y}i}=\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)(y_{ij}-\bar{y}_i)'$

is the corrected sum of squares and products matrix for the m_i new observations $\underline{y}_i$ from Π_i, for $i=1,2,\ldots,I$.

Equation (10.3.2) is obtained from (7.6.2) by the following substitutions :

$t_{ij} = 1$ for $i=1,2,..,m_i$ and $i=1,2,..,I$, so that $\Delta = I_m$ where $m = \sum_{i=1}^{I} m_i$,

$k_i = 1$ for $i=1,2,..,I$, so that $k = \sum_{i=1}^{I} k_i = I$,

W = mat-diag$(W_1'W_2',..,W_I')'$, a $m \times I$ matrix diagonal design matrix, with

$W_i' = [1,1,..,1]$, a $1 \times m_i$ vector,

Z = mat-diag$(Z_1'Z_2',..,Z_I')'$, a $n \times I$ matrix diagonal design matrix, with

$Z_i' = [1,1,..,1]$, a $1 \times n_i$ vector, and $n = \sum_{i=1}^{I} n_i$,

Y and X are as defined in section 7.6.1,

$B = (\underline{\mu}_1, \underline{\mu}_2, .., \underline{\mu}_I)'$, a $I \times d$ matrix,

$b = (\beta_1, \beta_2, .., \beta_I)'$, a $I \times d$ matrix, and

$C = \text{diag}(c_1, c_2, .., c_I)$, a $I \times I$ matrix,

$W'W = \text{diag}(m_1, m_2, .., m_I)$, and

$A = \text{mat-diag}[A_1, A_2, .., A_I]$, where $A_i = \left[I_{m_i} - \frac{1}{m_i + c_i} . 1_{m_i} \right]$, for $i=1,2,..,I$.

Multivariate Normal linear model populations

$p(y_{ij}|\Theta, w_{ij}, y_{ij} \in \Pi_i) = p(y_{ij}|B, T, w_{ij}, y_{ij} \in \Pi_i) = No_d(w_{ij}'B_i, T)$,

as in section 7.6.1,

$p(\Theta|X,Z) = p(B,T|X,Z) = NoWi_{k,d}(b,C,v,h)$,

$p(\underline{y}_1, \underline{y}_2, .., \underline{y}_I | X, Z, W, \underline{y}_1 \in \Pi_1, \underline{y}_2 \in \Pi_2, .., \underline{y}_I \in \Pi_I)$ is given by substituting $\Delta = I_m$ into equation (7.6.2), where $m = \sum_{i=1}^{I} m_i$.

10.4 Credibility interval for P

For the situation of two populations, Π_1 and Π_2, the posterior distribution of P can be obtained by transformation from the posterior distribution of the log likelihood ratio L at a particular new value y (not neccessarily one of the uncategorised observations $(y_1,y_2,..,y_m)$ also obtained from the mixture distribution). The posterior distribution of L can be approximated once its posterior moments have been found, i.e.

$$E(L^r) = \int L^r.p(\theta|\underline{x}_1,\underline{x}_2,\underline{y})\, d\theta \ .$$

But $p(\theta|\underline{x}_1,\underline{x}_2,\underline{y}) = \sum_{i_1=1}^{2}\sum_{i_2=1}^{2}\cdots\sum_{i_m=1}^{2}\left[\prod_{j=1}^{m} q_{i_j}\right].p(\theta|\underline{x}_1,\underline{x}_2,\underline{y},y_j\in\Pi_{i_j} \text{ for } j=1,2,..,m)$,

provided that the mixing parameters are known and hence,

$$E(L^r) = \sum_{i_1=1}^{2}\sum_{i_2=1}^{2}\cdots\sum_{i_m=1}^{2}\left[\prod_{j=1}^{m} q_{i_j}\right].E(L^r|\underline{x}_1,\underline{x}_2,\underline{y},y_j\in\Pi_{i_j} \text{ for } j=1,2,....,m). \quad (10.4.1)$$

But $E(L^r|\underline{x}_1,\underline{x}_2,\underline{y},y_j\in\Pi_{i_j} \text{ for } j=1,2,....,m)$ can be found for specific population distributions from the results already obtained in earlier chapters since $(\underline{x}_1,\underline{x}_2,\underline{y},y_j\in\Pi_{i_j} \text{ for } j=1,2,....,m)$ reduces to just two independent random samples from populations Π_1 and Π_2 (some new observations y_j are now added to the random sample $\underline{x}_1$ while the rest are added to the random sample $\underline{x}_2$) and this is exactly the situation considered in Chapters 2,3,4 and 7. Hence essentially the posterior moments of L are a weighted average of the posterior moments of L given different allocations of the y_j to populations Π_1 and Π_2.

For two populations with no parameters in common the cumulants of L can be found using (1.4.2) from the seperate cumulants of l_i for i=1,2, (given in chapters 2 for multivariate Normal populations with unequal covariance matrices, and chapter 4 for Exponential family populations and in particular Gamma, Binomial and Poisson populations and chapter 7 for multivariate Normal linear model populations with unequal error covariance matrices, using equation (1.4.3).

Finally for I=2 populations $E(L^r)$ can then be found using equation (10.4.1) for r=1,2,3,4 and hence the posterior distribution of L given $\underline{x}_1$, $\underline{x}_2$ and $\underline{y}$ can be approximated using either Pearson curves or, if appropriate, a Normal approximation, and the posterior distribution of P obtained by transformation.

For the situation of more than two populations, interest centres on the posterior moment generating function of the I population log likelihood functions at a particular new value y, i.e. $\Phi_{\underline{l}}(\underline{t})$, where $\underline{t} = (t_1,t_2,..,t_I)$ and $\underline{l} = (l_1,l_2,..,l_I)$, where $l_i = \log p_i(y|\Theta)$, for i=1,2,..,I.
Hence,

$$\Phi_{\underline{l}}(\underline{t}) = E\left[e^{\underline{t}'\underline{l}}\right] = \int e^{\underline{t}'\underline{l}}.p(\Theta|\underline{x}_1,\underline{x}_2,..,\underline{x}_I,\underline{y})\, d\Theta .$$

But $p(\Theta|\underline{x}_1,\underline{x}_2,..,\underline{x}_I,\underline{y}) =$

$$\sum_{i_1=1}^{I}\sum_{i_2=1}^{I}\cdots\sum_{i_m=1}^{I}\left[\prod_{j=1}^{m} q_{i_j}\right].p(\Theta|\underline{x}_1,\underline{x}_2,..,\underline{x}_I,\underline{y},y_j\in\Pi_{i_j} \text{ for } j=1,2,..,m) , \qquad (10.4.2)$$

provided that the mixing parameters are known and hence,

$$\Phi_{\underline{l}}(\underline{t}) = \sum_{i_1=1}^{I}\sum_{i_2=1}^{I}\cdots\sum_{i_m=1}^{I}\left[\prod_{j=1}^{m} q_{i_j}\right].E\left[e^{\underline{t}'\underline{l}} \,|\underline{x}_1,\underline{x}_2,..,\underline{x}_I,\underline{y},y_j\in\Pi_j \text{ for } j=1,2,....,m\right]$$

But $E\left[e^{\underline{t}'\underline{l}}|\underline{x}_1,\underline{x}_2..,\underline{x}_I,\underline{y},y_j\in\Pi_i \text{ for } j=1,2,....,m\right]$ can be found for specific population distributions from the results already obtained in earlier chapters since,
$(\underline{x}_1,\underline{x}_2,..,\underline{x}_I,\underline{y},y_j\in\Pi_i \text{ for } j=1,2,....,m)$ reduces to just I independent random samples from populations $\Pi_1,\Pi_2,..,\Pi_I$ (for j=1,2,..,m, the new observation y_j is now allocated to population Π_{i_j} and so is added to the random sample $\underline{x}_{i_j}$).
This is exactly the situation considered in earlier chapters 4,6 and 7. Hence essentially the posterior moment generating function of $\underline{l}$ is a weighted average of the posterior moment generating functions of $\underline{l}$ given different allocations of the y_j to populations $\Pi_1,\Pi_2,..,\Pi_I$.

For I populations with no parameters in common, the moment generating function of $\underline{l}$ can be obtained from the moment generating function of l_i using equation (7.3.3).

Finally $\Phi_{\underline{l}}(\underline{t})$ can be found using (10.4.2).

CHAPTER 11

Conclusions

11.1 Summary

A method of approximating the posterior distribution of P, the probability that a new observation y belongs to one of two populations with no parameters in common has been developed and applied generally to Exponential family populations and specifically to multivariate Normal and Normal linear model populations with unequal covariance matrices, Gamma, Binomial and Poisson populations. The posterior moment generating function of the log likelihood function at y has been derived and the posterior moments of the likelihood, log likelihood and log likelihood ratio functions have been found. The method has been extended to populations with parameters in common and applied to multivariate Normal and Normal linear model populations with equal covariance matrices, for which the posterior moment generating function and moments of the log likelihood ratio L has also been found.

The posterior distribution of L has been approximated either using Pearson curves or using a Normal approximation. The posterior distribution of P has been obtained by transformation from the posterior distribution of L. A credibility interval for P can be obtained from the corresponding credibility interval for L.

The method has been generalised to the situation where the new observation y belongs to one of more than two populations (with or without parameters in common). For multivariate Normal or Normal linear model populations with equal covariance matrices this involved finding the joint posterior moment generating function and moments of the log likelihood functions at y and the joint posterior moments of the likelihood functions at y, using the recurrence relationship derived between the multivariate moments and cumulants of a random variable.

The method has also been generalised to the situation where the mixing parameters are unknown and the random samples were obtained by mixture sampling from the combined population distribution, and their population of membership subsequently identified.

A method of combining subsets of variables modelled seperately has been shown, including as a special case the situation of independent subsets of variables.

A method of using the extra information concerning the unknown population parameters, provided by new uncategorised observations sampled independently from the mixture distribution has been shown.

Finally the stability of the predictive density function has been considered. A method of obtaining credibility interval bands for the likelihood and log likelihood functions has been given. For the likelihood function, this gives a credibility interval band around the predictive density function. The joint posterior moment generating function of the log likelihood functions at multiple new values $y_1, y_2, \ldots, y_m$, has been obtained generally for an Exponential family population and specifically for the Gamma, Binomial, Poisson and multivariate Normal linear model populations. In particular the posterior correlation coefficients between the likelihood and log likelihood functions at different values y_1 and y_2 has been obtained.

The relationship between P_p the 'predictive' estimate of P and E(P) the posterior expected value of P has been investigated in Appendix 1, showing that E(P) averages P over the posterior distribution of the parameters Θ given the independent random samples $\underline{x}_1$ and $\underline{x}_2$, while P_p averages P over the posterior distribution of Θ given $\underline{x}_1$, $\underline{x}_2$ and the new observation y, specifically assuming that y has been sampled independently from the combined population distribution.

11.2 Practical uses

The posterior distribution and credibility interval for P quantify the uncertainty about estimates of P and illuminate the relationship between the different estimates. Others have found that the 'estimative' and 'predictive' estimates of P can differ dramatically in their practical implications. In chapter 2 an example was considered in which this occurred. It was found that although P_e and P_p were not statistically different (both lying within the credibility interval obtained) they were dramatically different in their practical implications with P_e extremely close to 1. The examples of posterior distributions for P obtained in chapter 2 were of two types either 'stable', in which case further sampling would only serve to provide a slightly more accurate estimate of P, or 'unstable' in which case further sampling would be of considerable value for deciding at which extreme, 0 or 1, P lies.

The extremism of P_e (paticularly when the sample sizes are small relative to the number of parameters) may well be an example of a more general problem of extremism among non-linear functions of parameters when using maximum likelihood estimation where the likelihood function is skew. It should be noted however that P_e is not in general the maximum likelihood estimate of P even when it is based on replacing Θ by its maximum likelihood estimate Θ_e, because the mapping from Θ to $p(y|\Theta)$ is not in general one to one and hence $p(y|\Theta_e)$ is not in general the maximum likelihood estimate of $p(y|\Theta)$.

Given a particular utility loss-gain structure, relating possible actions to the true population of the new observation y, a point value for P would give a point optimum action, while a distribution for P would give a distribution of optimum actions or a distribution for the expected utility or gain for a particular action.

Finally the credibility interval band around the predictive density function quantifies the uncertainty or instability about the predictive density function, and should be of assistance in understanding its character, or rather, the character of the posterior distribution of the likelihood function (or log likelihood function). This should be of some use in the general area of informed prediction.

11.3 Further work

The coverage properties of the credibility intervals, for the probability of population membership of a new observation y, derived here could be investigated by a large scale simulation study. Work is currently being undertaken by Dr. Critchley and Dr. Ford, at Warwick and Glasgow universities respectively. They are comparing the coverage properties of several Classical intervals, including one based on the profile, with the corresponding Bayesian interval derived here, and are also comparing the intervals when applied to a range of practical data sets.

The general results obtained here for Exponential family populations could be applied to specific populations other than the ones considered here. Indeed populations which are not members of the Exponential family could be considered.

More accurate approximations to the posterior distribution of the likelihood and log likelihood function than the Pearson curves used here could be considered. Indeed the exact distribution would be desirable, though generally this appears intractable.

Further consideration could have been given to the choice of prior distribution for the parameters θ. A preposterior approach was considered in Appendix 5 however this is probably only sensible in the situation involving several populations rather than just two.

Finally the credibility interval band for the likelihood function around the predictive density function, developed here, could be applied to further practical data sets and also its coverage properties could be investigated by a large scale simulation study. Indeed this would seem paticularly helpful since the likelihood (or log likelihood) functions are the 'building blocks' for the probability of population membership.

11.4 General comment

The problem of finding the posterior distribution of P when discriminating between any two populations which have no parameters in common essentially becomes one of finding the posterior distribution of the likelihood (or log likelihood) function evaluated at the known new observation y for each population separately. Assuming that the parameter sets of the two populations have independent prior distributions then the likelihood (or log likelihood) functions will have independent posterior distributions and so the posterior distribution of P can easily be found by transformation. It does not in principle matter whether the data is continuous, discete or categorical or a mixture of these as long as a suitable model can be found separately for each population and as long as the posterior distribution of the likelihood (or log likelihood) function can be found or at least approximated separately for each population. The problem is essentially one of 'density estimation' or at least approximation of the posterior distribution of the density function. In order to estimate P, 'density estimation' is required, while to obtain an interval for P, the variability of the density estimate is required. Non-parametric approaches to 'density estimation' (instead of the parametric approach considered here) have been reviewed by Seber (1984, Ch6) and Fryer (1977).

APPENDICES

Appendix 1 shows the relationship between E(P), the expected value of the posterior distribution of P, and P_p, the Bayesian 'predictive' estimate of P. It is shown that the 'predictive' estimate averages the random variable P over the posterior distibution of Θ given both the independent random samples $\underline{x}_1$ and $\underline{x}_2$ and the value of the new observation y while E(P) averages P over the posterior distribution of Θ given the independent random samples $\underline{x}_1$ and $\underline{x}_2$ only.

Hence the Bayesian 'predictive' estimate of P will be close to E(P) if the information provided by the new observation y about the population parameters is negligable compared to that provided by the samples $\underline{x}_1$ and $\underline{x}_2$. The situation of more than one new observation is also considered.

Appendix 2 derives a recurrence relationship between the multivariate moments and cumulants of a p-dimensional random variable. The multivariate moments about the mean (or about zero) can then easily be found once the multivariate kumulants have been obtained.

Appendix 3 gives some formulae for the Psi function and its derivatives at integer values and integer values plus a half, which are useful for calculating the posterior moments of L and hence for approximating the posterior distributions and credibility intervals for L and P.

Appendix 4 derives an efficient numerical method of approximation which is particularly useful when the Psi function and its derivatives are required at the same value, for use if the value is not an integer or an integer plus a half.

Appendix 5 develops a method of estimation of the parameters of the Normal-Wishart prior distribution for (μ,T) for multivariate Normal populations with unequal means and unequal covariance matrices.

APPENDIX 1

The relationship between the expected value of the posterior distribution of P and the predictive estimate of P

In this appendix the relationship between E(P), the mean of the posterior distribution of $P = P(\theta,y)$, which can be used as an estimate of P, and the more usual Bayesian 'predictive' estimate P_p of P is shown, and the situation where each is appropriate is explained. A comparison of the extension of each of the two estimates to the case where there is more than one new observation of unknown population origin is also given.

Treating P as a random variable, the expected value, E(P), of its posterior distribution, given independent random samples $\underline{x}_1$ and $\underline{x}_2$ from populations Π_1 and Π_2 respectively, can be considered as simply an estimate of P subject to a squared error loss function since

$$E(P) = \int P.p(\theta|\underline{x}_1,\underline{x}_2)\, d\theta \ , \qquad \text{(A1.1)}$$

where P is given by equation (1.1.1). Alternative estimates of P using different loss functions could be obtained, e.g. the median or mode of the posterior distribution of P.

However, assuming known prior probabilities q_1 and q_2, then the usual Bayesian 'predictive' estimate P_p of P is given by,

$$P_p = p(y\in\Pi_1|y,\underline{x}_1,\underline{x}_2) = \frac{q_1p_1(y|\underline{x}_1,\underline{x}_2)}{q_1p_1(y|\underline{x}_1,\underline{x}_2)+q_2p_2(y|\underline{x}_1,\underline{x}_2)} = \frac{q_1p_1(y|\underline{x}_1,\underline{x}_2)}{p(y|\underline{x}_1,\underline{x}_2)}$$

$$= \int \frac{q_1p_1(y|\theta)\ p(\theta|\underline{x}_1,\underline{x}_2)}{p(y|\underline{x}_1,\underline{x}_2)}\, d\theta$$

$$= \int \frac{q_1p_1(y|\theta)\ p(y,\theta|\underline{x}_1,\underline{x}_2)}{p(y|\theta,\underline{x}_1,\underline{x}_2)\ p(y|\underline{x}_1,\underline{x}_2)}\, d\theta$$

$$= \int \frac{q_1p_1(y|\theta)}{p(y|\theta)}.p(\theta|y,\underline{x}_1,\underline{x}_2)\, d\theta$$

$$= \int P.p(\theta|y,\underline{x}_1,\underline{x}_2)\, d\theta \ , \qquad \text{(A1.2)}$$

i.e. $P_p = E(P|y,\underline{x}_1,\underline{x}_2)$, and the Bayesian 'predictive' estimate of P, P_p, has been shown to be equal to the expected value of P over the predictive distribution of θ given y, $\underline{x}_1$ and $\underline{x}_2$.

The difference between the estimate E(P) in (A1.1) and the 'predictive' estimate P_p in (A1.2) can clearly be seen. The 'predictive' estimate averages the random variable P over the posterior distribution of Θ given both the data $\underline{x}_1$ and $\underline{x}_2$ and the value of the new observation y, while the estimate E(P) averages P over the posterior distribution of Θ given $\underline{x}_1$ and $\underline{x}_2$ only.

Hence, if the information about Θ from y was negligible compared with the information from the random samples or if it was not certain that the new observation was sampled (independently) from the combined population distribution then E(P) is appropriate rather than P_p.

More than one new observation

Furthermore if there are m new observations $(y_j; j=1,2,...m)$ independently sampled from the combined population distribution, which are to be allocated, then the fully Bayes approach gives, assuming q_1 and q_2 are known, for m=2,

$$p(y_1\in\Pi_i, y_2\in\Pi_j | y_1, y_2, \underline{x}_1, \underline{x}_2) = \frac{q_i q_j p(y_1, y_2 | y_1\in\Pi_i, y_2\in\Pi_j, \underline{x}_1, \underline{x}_2)}{\sum_{k=1}^{2}\sum_{l=1}^{2} q_k q_l \, p(y_1, y_2 | y_1\in\Pi_k, y_2\in\Pi_l, \underline{x}_1, \underline{x}_2)} ,$$

which is different from the product $p(y_1\in\Pi_i | y_1, \underline{x}_1, \underline{x}_2).p(y_2\in\Pi_j | y_2, \underline{x}_1, \underline{x}_2)$ for i=1,2 and j=1,2 since y_1 and y_2 are not independent given $(y_1\in\Pi_i, y_2\in\Pi_j, \underline{x}_1, \underline{x}_2)$ unless $i\neq j$.

Consequently the marginal probability given by

$$p(y_1\in\Pi_i | y_1, y_2, \underline{x}_1, \underline{x}_2) = \sum_{j=1}^{2} p(y_1\in\Pi_i, y_2\in\Pi_j | y_1, y_2, \underline{x}_1, \underline{x}_2) ,$$

is different from $p(y_1\in\Pi_i | y_1, \underline{x}_1, \underline{x}_2)$.

If however it is not certain that the new observations $(y_i; i=1,2)$ were randomly and independently obtained from the combined population distribution then as before E(P) is appropriate for $p(y_1\in\Pi_1 | y_1, y_2, \underline{x}_1, \underline{x}_2)$. This problem is discussed further in Chapter 10.

APPENDIX 2

A recurrence relationship relating the multivariate moments of a p-dimensional random variable to its multivariate cumulants

The recurrence relationships are:

$$\mu_{\underline{r}} = \sum_{i_1=0}^{r_1-1} \sum_{i_2=0}^{r_2} \cdots \sum_{i_p=0}^{r_p} \binom{r_1-1}{i_1}\binom{r_2}{i_2}\binom{r_3}{i_3}\cdots\cdots\binom{r_p}{i_p} \mu_{\underline{i}} \kappa_{\underline{r}-\underline{i}} \quad \text{for } \underline{r} \geq (1,0,0,\ldots,0)$$

and the same relationship holds with $\mu_{\underline{r}}$, $\mu_{\underline{i}}$ and $\kappa_{\underline{r}-\underline{i}}$ replaced by $\mu_{\underline{r}}'$, $\mu_{\underline{i}}'$ and $\kappa'_{\underline{r}-\underline{i}}$ respectively where, $\underline{r}=(r_1,r_2,\ldots,r_p)$, $\underline{i}=(i_1,i_2,\ldots,i_p)$ and,

$$\mu_{\underline{i}} = E\left[X_1^{i_1}X_2^{i_2}\ldots\ldots X_p^{i_p}\right] \qquad \text{for } \underline{i} \geq \underline{0},$$

$$\mu_{\underline{i}}' = E\left[\left[X_1-E(X_1)\right]^{i_1}\left[X_2-E(X_2)\right]^{i_2}\ldots\ldots\ldots\left[X_p-E(X_p)\right]^{i_p}\right] \quad \text{for } \underline{i} \geq \underline{0},$$

$$\kappa_{\underline{i}} = \left[D_{\underline{t}}^{(\underline{i})}\left[\log \Phi(\underline{t})\right]\right]_{\underline{t}=\underline{0}} = \left[\frac{d^{(i_1)}}{dt_1^{(i_1)}} \frac{d^{(i_2)}}{dt_2^{(i_2)}} \ldots\ldots \frac{d^{(i_p)}}{dt_p^{(i_p)}} \log \Phi(\underline{t})\right]_{\underline{t}=\underline{0}} \quad \text{for } \underline{i} \geq \underline{0}$$

$$\Phi(\underline{t}) = E\left[e^{\underline{t}'\underline{X}}\right], \quad \underline{t}=(t_1,t_2,\ldots,t_p) \text{ and } \underline{X}=(X_1,X_2,\ldots\ldots,X_p)$$

and $\kappa_{\underline{i}}' = \begin{cases} 0 & \text{for } \underline{1}'\underline{i}=1 \text{ i.e. } \Sigma i_j=1 \\ \kappa_{\underline{i}} & \text{for } \underline{1}'\underline{i}>1 \text{ i.e. } \Sigma i_j>1 \end{cases}$

In particular,

$$\mu_r' = \sum_{i=0}^{r-1}\binom{r-1}{i} \mu_i' \kappa_{r-i}' = \sum_{i=0}^{r-2}\binom{r-1}{i} \mu_i' \kappa_{r-i} \qquad \text{for } r \geq 1 \qquad \text{(A2.1)}$$

and

$$\mu'_{r_1,r_2} = \sum_{i_1=0}^{r_1-1} \sum_{i_2=0}^{r_2} \binom{r_1-1}{i_1} \binom{r_2}{i_2} \mu'_{i_1,i_2} \kappa'_{r_1-i_1,r_2-i_2} \qquad \text{(A2.2)}$$

for $(r_1,r_2) \geq (1,0)$ i.e. for $r_1 \geq 1$ and $r_2 \geq 0$.

Proof

$$\frac{d}{dt_1}\Phi(\underline{t}) = \Phi(\underline{t})\frac{d}{dt_1}\log\Phi(\underline{t}) \qquad \text{(A2.3)}$$

Differentiating (A2.3) , with respect to t_1, (r_1-1) times using De Moivre's Theorem gives,

$$\frac{d^{(r_1)}}{dt_1^{(r_1)}}\Phi(\underline{t}) = \sum_{i_1=0}^{r_1-1}\begin{bmatrix} r_1-1 \\ i_1 \end{bmatrix}\frac{d^{(i_1)}}{dt_1^{(i_1)}}\Phi(\underline{t})\,\frac{d^{(r_1-i_1)}}{dt_1^{(r_1-i_1)}}\log\Phi(\underline{t}) \qquad \text{for } r_1\geq 1$$

$$\text{i.e. } D_{t_1}^{(r_1)}\Phi(\underline{t}) = \sum_{i_1=0}^{r_1-1}\begin{bmatrix} r_1-1 \\ i_1 \end{bmatrix} D_{t_1}^{(i_1)}\Phi(\underline{t})\cdot D_{t_1}^{(r_1-i_1)}\log\Phi(\underline{t}) \qquad \text{for } r_1\geq 1. \text{(A2.4)}$$

Differentiating (A2.4), with respect to t_2, r_2 times using De Moivre's Theorem gives,

$$D_{t_1,t_2}^{(r_1,r_2)}\Phi(\underline{t}) = \sum_{i_1=0}^{r_1-1}\sum_{i_2=0}^{r_2}\begin{bmatrix} r_1-1 \\ i_1 \end{bmatrix}\begin{bmatrix} r_2 \\ i_2 \end{bmatrix} D_{t_1,t_2}^{(i_1,i_2)}\Phi(\underline{t})\, D_{t_1,t_2}^{(r_1-i_1,r_2-i_2)}\log\Phi(\underline{t}) \qquad \text{(A2.5)}$$

Now differentiating (A2.5), with respect to t_j, r_j times for $j=3,\ldots,p$ gives

$$D_{\underline{t}}^{(\underline{r})}\Phi(\underline{t}) = \sum_{i_1=0}^{r_1-1}\sum_{i_2=0}^{r_2}\sum_{i_3=0}^{r_3}\cdots\sum_{i_p=0}^{r_p}\begin{bmatrix} r_1-1 \\ i_1 \end{bmatrix}\begin{bmatrix} r_2 \\ i_2 \end{bmatrix}\begin{bmatrix} r_3 \\ i_3 \end{bmatrix}\cdots\begin{bmatrix} r_p \\ i_p \end{bmatrix} D_{\underline{t}}^{(\underline{i})}\Phi(\underline{t})\, D_{\underline{t}}^{(\underline{r}-\underline{i})}\log\Phi(\underline{t}) \qquad \text{(A2.6)}$$

setting $\underline{t} = \underline{0}$ gives,

$$\mu_{\underline{r}} = \sum_{i_1=0}^{r_1-1}\sum_{i_2=0}^{r_2}\sum_{i_3=0}^{r_3}\cdots\sum_{i_p=0}^{r_p}\begin{bmatrix} r_1-1 \\ i_1 \end{bmatrix}\begin{bmatrix} r_2 \\ i_2 \end{bmatrix}\begin{bmatrix} r_3 \\ i_3 \end{bmatrix}\cdots\begin{bmatrix} r_p \\ i_p \end{bmatrix}\mu_{\underline{i}}\,\kappa_{\underline{r}-\underline{i}} \qquad \text{(A2.7)}$$

Replacing $\Phi(\underline{t})$ in (A2.6) by $\Phi'(\underline{t}) = E\left[e^{\underline{t}'[X-E(X)]}\right]$ the moment generating function of $[X-E(X)]$ and setting $\underline{t}=\underline{0}$ gives

$$\mu'_{\underline{r}} = \sum_{i_1=0}^{r_1-1}\sum_{i_2=0}^{r_2}\sum_{i_3=0}^{r_3}\cdots\sum_{i_p=0}^{r_p}\begin{bmatrix} r_1-1 \\ i_1 \end{bmatrix}\begin{bmatrix} r_2 \\ i_2 \end{bmatrix}\begin{bmatrix} r_3 \\ i_3 \end{bmatrix}\cdots\begin{bmatrix} r_p \\ i_p \end{bmatrix}\mu'_{\underline{i}}\,\kappa'_{\underline{r}-\underline{i}} \qquad \text{(A2.8)}$$

where $\kappa'_{\underline{i}} = \left[D_{\underline{t}}^{(\underline{i})}\log\Phi'(\underline{t})\right]_{t=0} = \begin{cases} 0 & \text{for } \underline{1}'\underline{i}=1 \\ \kappa_{\underline{i}} & \text{for } \underline{1}'\underline{i}\neq 1 \end{cases}$

since $\log\Phi'(\underline{t}) = \log E\left[e^{-\underline{t}'E(X)}\cdot e^{\underline{t}'X}\right] = -\underline{t}'E(X) + \log\Phi(\underline{t})$.

APPENDIX 3

Formulae for evaluating the Psi function and its derivatives at integer values and integer values plus a half

For the population distributions considered here i.e. multivariate Normal and Normal linear model populations with equal and unequal covariance matrices, Binomial, Gamma and Poisson populations, if vague prior information about the parameters is assumed, then the formulae for the posterior moments of the log likelihood ratio L involve only the values of the Psi function and its derivatives at integer values or integer values plus a half. Simple formulae for these are given below for n a positive integer, (Abramovitz and Stegun, 1965),

$$\Psi(n) = -C + \sum_{k=1}^{n-1} \frac{1}{k} \quad \text{where C is Euler's constant i.e. } C = 0.577215665,$$

$$\Psi(n+\tfrac{1}{2}) = -C + 2\left[\sum_{k=1}^{n} \frac{1}{2^{k-1}} - \log_e 2 \right],$$

$$\Psi^{(1)}(n) = \frac{\pi^2}{6} - \sum_{k=1}^{n-1} \frac{1}{k^2}, \text{ and,}$$

$$\Psi^{(1)}(n+\tfrac{1}{2}) = \frac{\pi^2}{2} - 4. \sum_{k=1}^{n} \frac{1}{(2k-1)^2}.$$

In general for $r \geq 1$,

$$\Psi^{(r)}(n) = (-1)^{r+1}.r!\left[\zeta(r+1) - \sum_{k=1}^{n-1} \frac{1}{k^{r+1}} \right]$$

$$\Psi^{(r)}(n+\tfrac{1}{2}) = (-1)^{r+1}.r!\left[(2^{r+1} - 1)\,\zeta(r+1) - \sum_{k=1}^{n} \frac{1}{(k-\tfrac{1}{2})^{(r+1)}} \right]$$

where ζ is the Zeta function and $\zeta(r+1) = \sum_{k=1}^{\infty} \frac{1}{k^{r+1}}$,

and if r is an integer then $\zeta(2r) = \frac{2^{2r-1}\pi^{2r}B_r}{(2r)!}$ for $r=1,2,3,\ldots$

where B_r is the r^{th} Bernoulli number ($B_1=1/6$, $B_2=1/30$, $B_3=1/42$, $B_4=1/30$ etc.)

In particular $\zeta(2) = \frac{\pi^2}{6} = 1.64493407$, $\zeta(3) = 1.20205690$ and

$\zeta(4) = \frac{\pi^4}{90} = 1.08232323.$

APPENDIX 4

An efficient numerical method of approximation which is particularly useful when the Psi function and its derivatives are required at the same value

The following series expansions can be used to calculate the psi function and its derivatives:

$$\Psi(x) = -C + \frac{1}{x} + x \sum_{k=1}^{\infty} \frac{1}{k(x+k)} ,$$

where C is Euler's constant, C = 0.577216,

$$\psi^{(r)}(x) = (-1)^{r+1} r! \sum_{k=0}^{\infty} \frac{1}{(x+k)^{r+1}}$$

The convergence, however, is slow and the summation can be replaced by integrals for faster evaluation. In fact to evaluate

$$S_0 = \sum_{k=k_0}^{\infty} f(k) ,$$

we use a Taylor expansion for $f^{(m)}(y)$ about the nearest integer to y, integrating over the range $(k - \frac{1}{2}, k + \frac{1}{2})$ and then summing over all $k \geq k_0$ to give

$$\int_{k_0 - 1/2}^{\infty} f^{(m)}(y)\, dy = \sum_{k=k_0}^{\infty} f^{(m)}(k) + \frac{1}{3!2^2} \sum_{k=k_0}^{\infty} f^{(m+2)}(k) + \frac{1}{5!2^4} \sum_{k=k_0}^{\infty} f^{(m+4)}(k) + \ldots$$

(A4.1)

Rearranging and similarly expressing the second and third summations in (A4.1) gives

$$S_m = \sum_{j=1}^{\infty} B_j I_{m+2(j-1)} \qquad (m \geq 0)$$

where

$$S_m = \sum_{k=k_0}^{\infty} f^{(m)}(k),$$

$$B_j = -\frac{1}{2^{2(j-1)}} \sum_{l=1}^{j-1} B_l A_{2(j-l)} \qquad (j > 1)$$

$$B_1 = 1, \; A_i = \frac{1}{(i+1)!} \text{ and } I_j = \int_{k_0-1/2}^{\infty} f^{(j)}(y)dy \qquad (j \geq 0)$$

The importance of this is that once the I_j $(j \geq 0)$ and B_j $(j \geq 1)$ have been obtained, the S_m $(m \geq 0)$ are easy to obtain and hence the psi function and its derivatives can easily be obtained at the same value x by choosing $f(k) = \frac{1}{k^2}$.

Depending on the function f, if k_0 is small, the convergence of the I_j to zero may be greatly increased by writing:

$$S_m = \sum_{k=k_0}^{k_1-1} f^{(m)}(k) + \sum_{k=k_1}^{\infty} f^{(m)}(k) \qquad \text{(A4.2)}$$

The first (k_1-k_0) terms are evaluated exactly and the second summation is approximated using the first few terms in (A4.1) e.g. $f(k) = 1/k^2$, then $I_j = O\left[k_0^{-(j+1)}\right]$, while using (A4.2) $I_j = O\left[k_1^{-(j+1)}\right]$. Keeping the total number of terms n used to calculate S_m fixed, take $k_1 = k_0 + r$ in (A4.2) then, for the greatest order of accuracy, r_1 is chosen to maximise

$$2(n-r+1)\log(k_0 + r)$$

Differentiating with respect to r gives

$$-2\log(k_0 + r_1) + \frac{2(n-r_1+1)}{k_0 + r_1} = 0$$

The asymptotic relationship $n = r_1 \log r_1$ can be used to estimate the optimum value r_1 of r.

APPENDIX 5

Estimation of the parameters of the Normal-Wishart prior distribution for (μ,T) for multivariate Normal populations with unequal means and unequal covariance matrices

Consider r samples of size n_i $(i = 1,2,\ldots, r)$ from multivariate normal distributions mean μ_i and covariance Σ_i where Σ_i is a dxd matrix. Let $T_i = \Sigma_i^{-1}$ for each i. Assume each (μ_i,T_i) has a Normal-Wishart prior distribution $p(\mu_i,T_i)$ where

$$p(\mu_i,T_i) = \text{NoWi}_d(\mu_o, n_o, v_o, H_o) .$$

Let the sample mean and corrected sum of squares and products matrix be $\bar{x}_i$ and S_i, with independent normal and wishart distributions respectively:

$$p(\bar{x}_i|\mu_i,T_i) = \text{No}_d(\mu_i,n_iT_i),$$

$$p(S_i|\mu_i,T_i) = \text{Wi}_d(n_i - 1, T_i).$$

Then using the prior distributions $p(\mu_iT_i)$, the μ_i and T_i can be integrated out from the joint distribution of $\bar{x}_i$ and S_i to obtain their joint marginal distribution $M(\bar{x}_i,S_i)$ where

$$M(\bar{x}_i,S_i) = \text{StS}_{i_d}[v_o,\mu_o, (\frac{1}{n_i} + \frac{1}{n_o}) \frac{S_o}{v_o}, (n_i - 1)S_o]$$

in the notation of Aitchison and Dunsmore (1975, Ch. 2).

Maximum likelihood estimators of the prior parameters could now be obtained from $\prod_{i=1}^{r} M(\bar{x}_i,S_i)$, however the resulting equations are complicated and require non-linear minimisation in $(d^2+3d+4)/2$ variables.

Instead consider the product L of the likelihoods of (μ_iT_i) using their common Normal-Wishart prior distribution:

$$L = \prod_{i=1}^{r} [\pi(\mu_i, T_i)]$$

and replace the true $(\mu_i T_i)$ in this likelihood function by their unbiased estimates $(\bar{x}_i, (n_i-d-1)S_i^{-1})$. Estimates of the prior parameters can be obtained by maximising the resulting 'likelihood'.

Let $W_i = (n_i-d-1)S_i^{-1}$, then

$$L(n_0,\mu_0,v_0,H_0) \propto n_0^{rd/2} \exp -\frac{1}{2}[\mu_0 - (\sum_1^r W_i)^{-1}\Sigma W_i x_i]'(n_0 \sum_1^r W_i)[\quad]$$

$$\times \frac{(\prod_{i=1}^{r} |W_i|^{\frac{v_0-d-1}{2}})}{2^{\frac{rdv_0}{2}} [\Gamma_d(\frac{1}{2}v_0)]^r} |H_0|^{\frac{rv_0}{2}} \exp[-\frac{1}{2}\operatorname{tr}(H_0 \sum_1^r W_i)]$$

The maximum 'likelihood' equations give

$$\mu_0 = (\sum_1^r W_i)^{-1} \sum_1^r W_i x_i$$

$$n_0 = \frac{rd}{\sum_1^r x_i' W_i x_i - (\sum_1^r W_i x_i)'(\sum_1^r W_i)^{-1}(\sum_1^r W_i x_i)}$$

$$H_0 = rv_0 (\sum_1^r W_i)^{-1}$$

and v_0 is the solution of the equation

$$r_d \log(\frac{rv_0}{2}) - r \log|\sum_1^r W_i| + \sum_1^r \log|W_i| - r \sum_{i=1}^{d} \Psi[\frac{1}{2}(v_0+1-i)] = 0$$

which can easily be solved by the Newton-Raphson method.

The posterior distribution $p(\mu_i, T_i | \bar{x}_i, S_i)$ of (μ_i, T_i) given the sample statistics $(\bar{x}_i, S_i)$ and the prior distribution $p(\mu_i, T_i)$ is given by:

$$p(\mu_i, T_i | \bar{x}_i, S_i) = \text{NoWi}_d \left(\frac{n_o \mu_o + n_i \bar{x}_i}{n_o + n_i}, n_o + n_i, v_o + n_i, H_i \right)$$

where $H_i = H_o + S_i + \frac{n_o n_i}{n_o + n_i} (\bar{x}_i - \mu_o)(\bar{x}_i - \mu_o)'$

As an example, suppose that samples of size 10 from each of two populations give unbiased estimates $(\hat{\mu}_i, \hat{T}_i) = (\bar{x}_i, (n_i - d - 1) S_i^{-1})$ for (μ_i, T_i), $i = 1,2$; $r = 2$, $n_1 = n_2 = 10$, $\hat{\Sigma}_i = \frac{S_i}{7.0}$, $d = 2$

Suppose:-

$$\hat{\mu}_1 = (0.0 \quad 0.0), \qquad \hat{\Sigma}_1 = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix},$$

$$\hat{\mu}_2 = (1.0 \quad 1.0), \qquad \hat{\Sigma}_2 = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}.$$

The maximum 'likelihood' equations above give

$$\hat{\mu}_o = (0.399 \quad 0.536),$$

$$\hat{n}_o = 6.847, \qquad \hat{v}_o = 6.911,$$

$$\hat{H}_o = \begin{pmatrix} 9.737 & 3.562 \\ 3.562 & 6.697 \end{pmatrix},$$

The posterior distributions are therefore given by

$$p(\mu_1, T_1 | \bar{x}_1, S_1) = \text{NoWi}_d \left((0.162 \quad 0.218), 16.847, 16.911, \begin{pmatrix} 17.383 & 8.631 \\ 8.631 & 14.865 \end{pmatrix} \right)$$

$$p(\mu_1, T_1 | \bar{x}_2, S_2) = \text{NoWi}_d \left((0.756 \quad 0.811), 16.847, 16.911, \begin{pmatrix} 46.207 & 4.696 \\ 4.696 & 14.572 \end{pmatrix} \right)$$

Assuming instead vague prior knowledge

$$p(\mu_1,T_1|\bar{x}_1,S_1) = \text{NoWi}_d\left((0.0,\ 0.0),\ 10,\ 9,\ \begin{pmatrix} 7.0 & 4.2 \\ 4.2 & 7.0 \end{pmatrix} \right)$$

$$p(\mu_2,T_2|\bar{x}_2,S_2) = \text{NoWi}_d\left((1.0,\ 1.0),\ 10,\ 9,\ \begin{pmatrix} 35.0 & 0.0 \\ 0.0 & 7.0 \end{pmatrix} \right)$$

There may be some instability in $\hat{n}_0$ and $\hat{v}_0$ when using only two populations so that this procedure is only recommended where several populations are available.

References

Abramovitz, M. and Stegun, I.A. (1965). Handbook of Mathematical Functions. New York, Dover Publications.

Aitchison, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis. London, Cambridge University Press.

Aitchison, J., Habbema, J.D.F. and Kay, J.W. (1977). A critical comparison of two methods of statistical discrimination. *Appl. Statist.*, **26**, 15-25.

Anderson, J.A. (1979). Multivariate logistic compounds. *Biometrika*, **66**, 17-26.

Broemeling, L.D. (1985). Bayesian analysis of linear models. New York, Marcel Dekker.

Box, G.E.P. and Tiao, G.C. (1973). Bayesian inference in statistical analysis. Mass., Addison-Wesley.

Crichley, F. and Ford, I. (1984). On the covariance of two non-central F random variables and the variance of the estimated linear discriminant function. *Biometrika*, **71**, 637-638.

Critchley, F. and Ford, I. (1985). Interval estimation in discrimination: the multivariate normal equal covariance case. *Biometrika*, **72**, 109-116.

Critchley, F., Ford, I. and Rijal, O. (1986a). Interval estimation based on the profile likelihood: strong Lagrangian theory, with applications to discrimination. (submitted to Biometrika for publication)

Critchley, F., Ford, I. and Rijal, O. (1986b). Uncertainty in discrimination. (to appear in Proceedings of the DIANA II conference in Prague, May, 1986)

Critchley, F., Ford, I. and Hurst, D. (1986). On the possible linearity of the profile log-likelihood function: strong Lagrangian theory, with applications to linear discrimination. (Report No.96, Department of Statistics, University of Warwick)

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, **39**, 1-38.

Elderton, W.P. and Johnson, N.L. (1969). Systems of Frequency Curves. London, Cambridge University Press.

Fryer, M.J. (1977). A review of some non-parametric methods of density estimation. *J. Inst. Math. Appl.*, **20**, 335-354.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. *Ann. Math. Statist.*, **36**, 150-159.

Ganesalingam, S. and McLachlan, G.J. (1979). Small sample results for a linear discriminant function estimated from a mixture of normal populations. *J. Statist. Comput. Simul.*, **9**, 151-158.

Johnson, N.L., Nixon, E. and Amos, D.E. (1963). Tables of percentage points of Pearson curves, for given $\sqrt{\beta_1}$ and β_2 expressed in standard measure. Biometrika, **50**, 459-498.

Krzanowski, W.J. (1975). Discrimination and classification using both binary and continuous variables. *J. Amer. Statist. Soc.*, **70**, 782-790.

Krzanowski, W.J. (1979). Some linear transformations for mixtures of binary and continuous variables, with particular reference to linear discriminant analysis. *Biometrika*, **66**, 33-39.

Krzanowski, W.J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, **36**, 493-499.

Little, R.J.A. and Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, **72**, 497-512.

Moran, M.A. and Murphy, B.J. (1979). A closer look at two alternative methods of statistical discrimination. *Appl. Statist.*, **28**, 223-232.

O'Neill, T.J. (1978). Normal discrimination with unclassified observations. *J. Amer. Statist. Assoc.*, **73**, 821-826.

Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm, SIAM Rev., **26**, 195-239.

Seber, G.A.F. (1984). Multivariate Observations. New York, John Wiley and Sons.

Titterington, D.M. (1984). Recursive parameter estimation using incomplete data. *J. R. Statist. Soc. B*, **46**, 257-267.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). Statistical analysis of finite mixture distributions. New York, John Wiley and Sons.

Vlachonikolis, I.G. and Marriott, F.H.C. (1982). Discrimination with mixed binary and continuous data. *Appl. Stats.*, **31**, 23-31.