ASYMPTOTIC BAYESIAN DISCRIMINATION AND REGRESSION

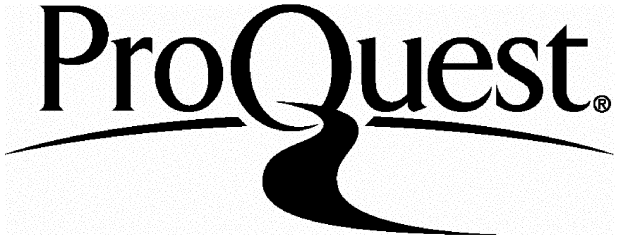
Thesis submitted to the University of London for the degree of Doctor of Philosophy in the Faculty of Science

by

Fang Biqi

University College London

June, 1995



Abstract

This thesis investigates the problems of discimination and regression using Bayesian methods with emphasis on their asymptotic properties when p, the number of variables that can be observed, is unlimited. For the problem of discriminating between two multivariate normal populations the conjugate prior is found to lead to asymptotically perfect discrimination, under certain conditions on the parameters. Similarly, in a problem of discrimination between two populations with binary variables, using a Dirichlet process prior, necessary and sufficient conditions for asymptotically perfect discrimination are found. To investigate this determinism a comparison is made between the Bayesian discriminant function and a sample-based discriminant function which fits the data exactly when p is large. It is shown that their performances are asymptotically equivalent. Similarly, for the regression of normal variables with a conjugate prior the Bayes predictor, which implies asymptotic deterministic predictability, is asymptotically equivalent to a classical least squares predictor which exacly fits the sample data for large p. Thus the conjugate Bayesian approach neglects the problem of bias due to overfitting. In contrast, it is shown that a certain nonconjugate prior does not imply asymptotic determinism for the Bayes predictor, and renders the behaviours of Bayes and least squares predictors different. This reveals the importance of the choice of prior distribution for Bayesian analysis.

Acknowledgments

This work has been carried out in the Department of Statistical Science, University College London.

I am most grateful to my supervisor Professor A. P. Dawid for his excellent guidance, enormous stimulation and continuous encouragement.

I would like to truly give my gratitude to University College London and the University of London for providing this invaluable opportunity of research.

I would like to express my sincere thanks to all my colleagues in the Department for their cordial friendship and constant encouragement and all those who thoughtfully have helped me in many ways.

Financial support from SBFSS (Sino–British Friendship Scholarship Scheme) of the Sir Y K Pao foundation Hong Kong, Chinese State Education Commission Beijing and Overseas Development Administration U.K., ORSAS (Overseas Research Student Award Scheme) of Committee of Vice–Chancellors and Principals, Fund from UCL Graduate School and University College London, Fund from The Great Britain China Educational Trust are greatly acknowledged.

Contents

Chapter 1

INTRODUCTION

1.1 The Problems of Discrimination and Regression

In this thesis two major problems in multivariate analysis, discrimination and regression, are investigated using Bayesian methods.

The basic problem of discrimination is to assign an observation to one of two or more populations on the basis of its value. Statistical decision theory gives solutions minimizing the probability or expected cost of misclassification. A criterion maximizing a function of the distance between the mean value of two samples leads to Fisher's linear discriminant function (Fisher 1936). Under the assumption of normal distributions, Anderson (1958) suggested a likelihood ratio criterion. The problem of discrimination was studied from the Bayesian point of view by Geisser and Cornfield (1963), Geisser (1964), among others. They obtained the posterior probabilities that an observation with a finite number of variables belongs to one of k multivariate normal distributions, under the assumption of prior ignorance.

In the regression problem parameters of a model are estimated from a set of

data. The least squares criterion is widely used. In normal linear models the least squares estimator is the maximum likelihood estimator (Anderson 1958). It has other optimal properties such as MRE (mimimum risk equivariance) and UMVU (uniformly minimum variance unbiased) (Lehmann, 1983, pp. 156,187,77). The regression problem was studied from the Bayesian point of view by Geisser (1965), Tiao and Zellner (1964), again using noninformative priors (Berger 1980 p.88).

Usually it is assumed that the number of observations is greater than the number of observable variables. In practice it is possible that more and more variables of the population are observed to enable the investigator to use more information on each population. Brown (1980) considered the problem of discrimination between two multinomial populations with the number of cells of each multinomial being unlimited. This is the case when more symptoms are introduced in medical diagnosis. The prior expectation of the probability of correct classification, p_n, is given by

$$p_n = \frac{1}{2}\sum_{i=1}^{n} E\max(\theta_i, \phi_i),$$

where $\{\theta_i\}$, $\{\phi_i\}$ are the probabilities for the first and second populations respectively ($1 \leq i \leq n$, $\sum \theta_i = \sum \phi_i = 1$). The asymptotic behaviour of p_n as $n \to \infty$ was studied. In particular, it was shown that $p_n \to 1$ under the assumption that the prior distributions for $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ are independent identically distributed Dirichlet $D(a_1, \cdots, a_n)$ for equal a_i. The problem of regression on an unlimited number of explanatory variables was studied by Dawid (1988). Under the assumption that the sampling distribution is normal, it was shown that the conjugate inverse Wishart prior implies degenerate prediction under certain conditions, that is, the response variable can be predicted arbitrarily closely by using a sufficiently large number of predictors.

This thesis extends these investigations to more general case, concentrating on the assymptotic properties of the Bayes approach to discrimination and regression problems when the number of observable variables tends to infinity. In Chapter 2 the problems of discrimination between two multivariate normal populations with common dispersion matrix are considered. Under the natural congugate normal

inverted Wishart prior, necessary and sufficient conditions exist for asymptotically perfect discrimination in the sense that the ratio of the posterior probability of the second population to that of the first population tends to 0 or ∞ according as the observation arises from the first or second population. Chapter 3 studies the asymptotic properties of discrimination between two populations with binary variables, using a Dirichlet process prior. Necessary and sufficient condition for asymptotically perfect discrimination between the two populations are found. To understand the phenomenon of determinism, Chapter 4 compares a Bayes discriminant function with Fisher's discriminant function in discrimination between two normal populations. The performance of a linear discriminant function is defined as the squared difference between its expectation in the two populations, normalized by its variance. The Bayes discriminant function maximizes the performance conditional on the training data, with the maximum tending to infinity under a condition on the hyper-parameters of the prior, allowing perfect discrimination also in the sense indicated in Chapter 2. When the number of observable variables is large, Fisher's discriminant function fits the data exactly. It is shown that the performances of these two discriminant functions are asymptotically equivalent. Chapter 5 and 6 are devoted to regression of normal variables, comparing the Bayes estimator with the classical least squares estimator, which exactly fits the data for a sufficiently large number of variables. Chapter 5 shows that, under a conjugate prior, the Bayes estimator, which implies asymptotic determinism under a certain condition on the prior hyper-parameters, is asymptotically equivalent to this least squares estimator. This conclusion in conjunction with Chapter 3 shows that the conjugate Bayes approach neglects the problem of bias in these problems. Chapter 6 shows that, in contrast to the conjugate prior, a certain non-conjugate prior does not imply asymptotic determinism of the Bayes predictor, and makes the Bayes and the above least squares estimator different.

1.2 Selection Bias

Our investigation shows that the usual assumptions of conjugate priors imply asymptotic determinism in the discrimination and regression problems. This may be reasonable in certain areas such as pattern recognition. However, in many statistical problems it is unreasonable to believe that inference will be deterministic if only sufficiently many variables can be observed. Hence the choice of prior must be made according to the problems considered. This undesirable determinism property induced by the use of conjugate prior was related to paradoxes of inference under selection and optimisation in Dawid, 1994, as follows.

In some optimisation problems, bias is introduced because selection is related to the variables to be studied. Suppose the distribution for the data X has a parameter θ, and we wish to make inference on ϕ_λ, a function of θ, for $\lambda \in \mathcal{L}$, a set of indexes. Each ϕ_λ is estimated by X_λ. Suppose our interest is the optimised parameter $\phi^{**} = \sup\{\phi_\lambda, \lambda \in \mathcal{L}\}$, which is achieved at λ^{**}. It is not possible to identify this value λ^{**} without fully knowing the parameters. A two–stage approach is as follows. At the first stage the data are used to select Λ^*, at which X_λ achieves its maximum, and a parameter $\phi^* = \phi_{\Lambda^*}$ is thus identified. At the second stage inference is made about the selected parameter ϕ^*. Since Λ^* is random, ϕ^* is a "data–dependent parameter". Let $X^* = X_{\Lambda^*}$. It can be shown that $E_\theta X^* \geq \phi^{**} \geq \phi^*$ (typically the strict inequalities hold). Thus X^* is positively biased for ϕ^{**} and ϕ^*. The classical approach then demands that allowance be made for the bias either by explicit modelling of the whole two–stage process or by some general de–biasing technique. In contrast, the Bayesian approach requires no adjustment for selection, since the posterior distribution of any quantity is unchanged by selection using data. Let $Y_\lambda = E(\phi_\lambda \mid X)$, $Y^{**} = E(\phi^{**} \mid X)$, $Y^* = E(\phi^* \mid X)$, $Y^+ = \sup\{Y_\lambda\} = Y_{\Lambda+}$. Y^{**}, Y^+ and Y^* can be used as the Bayesian estimates of ϕ^{**}, $\phi^+ \stackrel{\text{def}}{=} \phi_{\Lambda+}$ and ϕ^* respectively. In particular, when a proper prior is used, in the joint distribution of (X, θ), $E(Y^* - \phi^*) = 0$ so that $E_\theta Y^* > \phi^*$ does not hold for all θ. Thus the Bayes estimate Y^* is not positively biased for ϕ^*, at least for some values of θ. The same analysis applies to Y^{**} and

Y^+.

If the number of variables that can be observed is allowed to tend to infinity, the investigation of this thesis reveals that in some important multivariate statistical problems such as discrimination and regression, the use of the usual conjugate priors leads to asymptotic equivalence of the Bayesian inference and the unadjusted classical one, which suggests determinism in these problems implied by the conjugate priors may be inadequate. However, if such priors are taken seriously, the biasing effects of the determinism can be ignored. Moreover, the Bayesian inference by using a nonconjugate prior in a regression problem does not imply asymptotic determinism and is different from the unajusted classical inference, providing a possible solution of the conflict between determinism and selection bias, if we do not believe determinism in the problem considered.

Since the result of a Bayesian analysis depends on the prior assumption made, the choice of a suitable prior is a very important issue for Bayesian analysis. If a result appears unreasonable, the prior assumption is inappropriate and must be reconsidered.

1.3 Matrix Distributions

The models considered in this thesis involve certain spherical and rotatable distributions, including the matrix-variate normal, t, F, beta, the Wishart and inverse Wishart distributions. We shall use the notation and properties for these distributions developed in Dawid (1981). The notation may differ from other common conventions. It has the property of consistency under marginalization of the distribution, and hence is convenient when dealing with the distributions of infinite matrices. In this section we shall give a brief review of their definitions and properties and develop some properties which will be used in subsequent chapters.

Left–spherical, Right–spherical, Spherical and Rotatable distributions. Let Y be a random $n \times p$ matrix. Y is called left–spherical if, for any

fixed $n \times n$ orthogonal matrix P, PY has the same distribution as Y. Y is called right–spherical if YQ has the same distribution as Y, for any fixed $p \times p$ orthogonal Q. Y is called spherical if it is simultaneously left– and right–spherical. A random $p \times p$ nonnegative-definite symmetric matrix S is called rotatable if $Q'SQ$ has the same distribution as S, for any fixed $p \times p$ orthogonal Q (Dawid, 1977, 1978, 1981).

Matrix normal. The $n \times p$ random matrix Z with independent standard normal elements is denoted by $Z \sim \mathcal{N}(I_n, I_p)$. For nonrandom A, B, M such that $AA' = \Lambda$, $B'B = \Sigma$, the distribution of $M + AZB$ is denoted by $M + \mathcal{N}(\Lambda, \Sigma)$. This is denoted by $N(M, \Lambda \otimes \Sigma)$ in Muirhead (1982).

Wishart. The $p \times p$ random matrix Ψ having a Wishart distribution with ν degrees of freedom and scale matrix Σ is denoted by $\Psi \sim W(\nu; \Sigma)$, ($\Sigma \geq 0$ is $p \times p$). For $p = 1$, $W(\nu; 1)$ is equal to χ^2_ν. If $Z \sim \mathcal{N}(I_n, \Sigma)$, then $Z'Z \sim W(n; \Sigma)$.

Inverse Wishart. The $p \times p$ random matrix Φ having a standard inverse Wishart distribution with parameter δ is denoted by $\Phi \sim IW(\delta;\ I_p)$, ($\delta > 0$), for which $\Phi^{-1} \sim W(\nu; I_p)$ with $\nu = \delta + p - 1$. The parameter δ is chosen so that it does not change for any leading submatrix of Φ. The distribution $IW(\delta; I_p)$ is denoted by $W^{-1}(\delta + 2p; I_p)$ in Muirhead (1982).

Matrix-t. The $n \times p$ random matrix T having a standard matrix-t distribution with parameter δ is denoted by $T \sim T(\delta; I_n, I_p)$. It is denoted by $T(I_n, I_p,\ 0,\ \delta + n + p - 1)$ in Dickey (1967). For $p = 1$, $n > 1$, $T(\delta; I_n, 1) = \delta^{-1/2}\mathbf{t}_\delta$, where $\mathbf{t}_\delta$ is multivariate t distribution with δ degrees of freedom (Cornish 1954). The matrix-t distribution has a stochastic representation as $T \mid \Phi \sim \mathcal{N}(I_n, \Phi)$ with $\Phi \sim IW(\delta; I_p)$ or $T \mid \Lambda \sim \mathcal{N}(\Lambda, I_p)$ with $\Lambda \sim IW(\delta, I_n)$.

Matrix-variate F. The $p \times p$ random matrix U having a standard matrix-variate F distribution with parameters ν, δ is denoted by $U \sim F(\nu, \delta; I_p)$, ($\delta > 0, \nu > p - 1$ or ν integral). It is denoted by $B_{II}\{p; \frac{1}{2}\nu, \frac{1}{2}(p + \delta - 1)\}$ in Tan (1969), $G(\nu, p + \delta - 1; I_p)$ in Dempster (1969), (see also Olkin & Rubin, 1964). For $p = 1$, $F(\nu, \delta; I_p) = (\nu/\delta)F_{\nu,\delta}$. It has a stochastic representation as $F \mid \Phi \sim W(\nu; \Phi)$ with

$\Phi \sim IW(\delta; I_p)$ or $U \mid \Lambda \sim IW(\delta; \Lambda)$ with $\Lambda \sim W(\nu; I_p)$. If $T \sim T(\delta; I_n, I_p)$, then $T'T \sim F(n, \delta; I_p)$. If $U \sim F(\nu, \delta; I_p)$, then $U^{-1} \sim F(\delta + p - 1, \nu - p + 1; I_p), (\nu > p - 1)$.

Matrix-variate beta. The $p \times p$ random matrix V having a matrix-variate beta distribution with ν_1, ν_2 degrees of freedom and scale matrix Σ is denoted by $V \sim B(\nu_1, \nu_2; \Sigma)$. The standard $B(\nu_1, \nu_2; I_p)$, $(\nu_1 + \nu_2 > p - 1)$ is denoted by $B_I(p; \frac{1}{2}\nu_1, \frac{1}{2}\nu_2)$ in Tan (1969), $B_p(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2)$ in Mitra (1970). See also Olkin & Rubin (1964), Khatri (1970). For $p = 1$, $B(\nu_1, \nu_2; 1) = \beta(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2)$, the beta distribution. If $S_i \sim W(\nu_i; \Phi)$, $i = 1, 2$, independently, with Φ positive–definite, and $S = S_1 + S_2$, then the conditional distribution of S_1 given $S = \Sigma$ is $B(\nu_1, \nu_2; \Sigma)$. It also has a stochastic represantation as $D^{-1}S_1(D^{-1})' \sim B(\nu_1, \nu_2; I_p)$, where $S_i \sim W(\nu_i, \Sigma)$, $i = 1, 2$, independently with $\nu_1 + \nu_2 > p - 1$, D is such that $S \stackrel{\text{def}}{=} S_1 + S_2 = DD'$, and D is independent of S_1 given S. If $U \sim F(\nu, \delta; I_p)$, then $(I_p + U)^{-1} \sim B(\delta + p - 1, \nu; I_p)$. If $V \sim B(\nu_1, \nu_2; I_p)$, then $I_p - V \sim B(\nu_2, \nu_1; I_p)$.

The parameters of the random matrix distributions considered above have a consistency property that the leading submatrices have the same degrees of freedom. Hence we can introduce corresponding distributions for infinite arrays $Y_{\infty,\infty} \sim \Delta$ (for $\Delta = \mathcal{N}$, T; or W, IW, F, if $n = p$) such that all leading submatrices of order $n \times p$, $Y_{n,p} \sim \Delta_{n,p}$, (Dawid 1981, Section 3). Furthermore, for an $\infty \times p$ matrix $Y (p \leq \infty)$ having a left-spherical distribution Δ or, equivalently, a consistent family $\{\Delta_n\}$ of left-spherical distribution for Y_n, the first n rows of Y, we have a scale-modified left-spherical distribution $\Delta(H)$ defined as that of $\{AY_n : AA' = H, Y_n \sim \Delta_n\}$. Similarly we have scale-modified right-spherical distributions. For an infinite spherical distribution Δ for $Y_{\infty,\infty}$ corresponding to the consistent family $\{\Delta_{n,p} : n, p \geq 1\}$ we have a doubly scaled distribution $\Delta(H, K)$ defined as that of $\{AY_{n,p}B; AA' = H, B'B = K\}$. For a rotatable distribution Π for $S_{\infty,\infty}$ we have $\Pi(K)$ as the distribution of $B'S_{p,p}B$. Thus the standard matrix distribution discussed above have scale-modified infinite versions $\mathcal{N}(H, K)$, $W(\nu; K)$, $IW(\delta; K)$, $T(\delta; H, K)$, $F(\nu, \delta; K)$ and $B(\nu_1, \nu_2; K)$, (see Dawid 1981, Section 6).

The distribution of the submatrices of the matrix distribution can be found in the literature, e.g. Dawid 1988, Lemma 1 for normal distribution (or Muirhead 1982 p.12, Theorem 1.2.11), Lemma 2 for inverse Wishart distribution (or Dempster 1969, Theorem 13.4.2), Lemma 4 for matrix-t distribution (or Dickey 1967), Muirhead (1982, Theorem 3.2.10) for Wishart distribution, Mitra (1970) and Tan (1969) for matrix-variate F and beta distribution. Some properties on the moments, asymptotics and mixtures of the above distributions are summarized in the following lemmas.

Lemma 1.1.

(a) For $\Psi \sim W(\nu, \Sigma)$, the Wishart distribution, $E\Psi = \nu\Sigma$.

(b) For $\Phi \sim IW(\delta; G)$, the inverse Wishart distribution, $E\Phi = G/(\delta - 2), (\delta > 2)$.

(c) For $F \sim F(\nu, \delta; K)$, the matrix-variate F, $EF = \frac{\nu}{\delta-2}K, (\delta > 2)$.

(d) For $B \sim B(\nu_1, \nu_2; I_p)$, the matrix-variate Beta distribution, $EB = \frac{\nu_1}{\nu_1+\nu_2}I_p$.

Proof. For the proof of (a) and (b), see e.g. Muirhead (1982, p.90, p.113). Then (c) follows by the definition of $F(\nu, \delta; K)$ as a mixture of Wishart $W(\nu; \Psi)$ and inverse Wishart $IW(\delta; K)$. For the proof of (d), note that $\forall \alpha \in R^p$, $\alpha' B \alpha \sim B(\nu_1, \nu_2; \alpha'\alpha)$ which is $\beta(\nu_1/2, \nu_2/2) \cdot \alpha'\alpha$ in the conventional notation of the univariate beta distribution with parameters $\nu_1/2$, $\nu_2/2$. Thus $\alpha'(EB)\alpha = E(\alpha' B \alpha) = \frac{\nu_1}{\nu_1+\nu_2} \cdot \alpha' I_p \alpha$, $\forall \alpha \in R^p$, which establishes (d). □

Lemma 1.2. Suppose $T : (n \times p)$ has matrix-t distribution $T \sim T(\delta; K, G)$, where $\delta > 0$, $K > 0$, $G > 0$. Then

$$ET = 0, \qquad \mathrm{Var}(\mathrm{Vec}(T)) = \frac{1}{\delta - 2} G \otimes K, \ (\delta > 2).$$

Moreover, if $G = I_p$, then

$$ET'T = I_p \mathrm{tr}(K/(\delta - 2)).$$

Proof. Since $T \stackrel{d}{=} AT(\delta; I_n, I_p)B$, where $AA' = K$, $B'B = G$, and $\mathrm{Vec}(T) = (B' \otimes A)\mathrm{Vec}(T(\delta; I_n, I_p))$, we only need to prove the case in which $K = I_n$, $G = I_p$. Then T has a stochastic representation

$$T \mid \Phi \sim \mathcal{N}(\Phi, I_p), \quad \Phi \sim IW(\delta; I_n).$$

Hence

$$ET = E[E(T \mid \Phi)] = 0$$

$$\mathrm{Var}(\mathrm{Vec}(T)) = E[\mathrm{Var}(\mathrm{Vec}(T)) \mid \Phi)] = E(I_p \otimes \Phi) = I_p \otimes I_n/(\delta - 2), \ (\delta > 2).$$

If $G = I_p$, let T be represented as $T \mid \Lambda \sim \mathcal{N}(\Lambda, I_p)$, with $\Lambda \sim IW(\delta, K)$, and let $\mathbf{t}_i$ be the ith column of T. Then $E(\mathbf{t}_i'\mathbf{t}_j \mid \Lambda) = 0$, if $i \neq j$, $\mathrm{tr}\Lambda$, if $i = j$. Hence

$$ET'T = E[E(T'T \mid \Lambda)] = E[\mathrm{diag}\{\mathrm{tr}\Lambda\}] = \mathrm{diag}\{\mathrm{tr}K/(\delta - 2)\},$$

completing the proof. □

Lemma 1.3.

(a) For $\Psi \sim W(\nu; \Sigma)$, the Wishart distribution, $\Psi/\nu \xrightarrow{\mathrm{P}} \Sigma$ as $\nu \to \infty$.

(b) For the $p \times p$ random matrix $\Phi \sim IW(\delta; G)$, the inverse Wishart distribution, $\delta\Phi \xrightarrow{\mathrm{P}} G$ as $\delta \to \infty$.

(c) For $T \sim T(\delta; L, M)$, the matrix-t distribution, $T \xrightarrow{\mathcal{L}} \mathcal{N}(L, A)$ as $\delta \to \infty$, with L fixed and $M/\delta \to A$.

(d) For $F \sim F(\nu, \delta; K)$, the matrix-variate F, $F \xrightarrow{P} 0$ as $\delta \to \infty$, $F/\nu \xrightarrow{\mathcal{L}} IW(\delta; K)$ as $\nu \to \infty$.

(e) For $B \sim B(\nu_1, \nu_2; I_p)$, the matrix-variate beta distribution, $B \xrightarrow{P} I_p$ as $\nu_1 \to \infty$, $B \xrightarrow{P} 0$ as $\nu_2 \to \infty$.

Proof. Assertion (a) can be established by SLLN (Strong law of large numbers, see, e.g. Rao, 1987 p.14) and the additive property of the Wishart distribution. Assertion (b) follows from (a) and the fact that $\Phi^{-1} \stackrel{\mathrm{d}}{=} W_p(\delta + p - 1; G^{-1})$. Then assertion (c) follows from (b) and the representation of T as a mixture of a normal distribution with an inverse Wishart distribution (or Dickey 1967, p.513). The

representation of F as a mixture of a Wishart distribution and an inverse Wishart distribution, and (a) , (b), establish (d). Finally, the first part of (e) follows from (d) and the relation $B \stackrel{d}{=} [I + F(\nu_2, \nu_1 - p + 1; I_p)]^{-1}$ (Dawid 1981, Thorem 4), and the second part follows from the relation that $B(\nu_1, \nu_2; I_p) = I_p - B(\nu_2, \nu_1; I_p)$. □

Lemma 1.4. Let Δ_n denote the left-spherical distribution of an order $n \times r$ matrix. Suppose Z_p, A_p, A are random matrices with

$$Z_p \mid A_p \sim \Delta_n(A_p A_p'), \ p = 1, 2, \cdots$$

Suppose $A_p \xrightarrow{P} A$ as $p \to \infty$. Then $Z_p \xrightarrow{P}$ a random matrix Z, with distribution given by $Z \mid A \sim \Delta_n(AA')$, as $p \to \infty$ (Similar results hold for right-spherical, spherical and rotatable distributions.)

Proof. Let A, $\{A_p, p = 1, 2 \cdots\}$, and Y be independent matrices with $Y \sim \Delta_n(I)$. The (i, j) element of $A_pY - AY$ is a sum of finite products of the corresponding elements of $A_p - A$ and Y, $\sum_{k=1}^{n}(A_p(i,k) - A(i,k))Y(k,j)$. Hence $(A_p - A)Y \xrightarrow{P} 0$, i.e. $A_pY \xrightarrow{P} AY$. Since $Z_p \mid A_p \stackrel{\mathrm{d}}{=} A_p \cdot Y \mid A_p$ where Y is independent of A_p, we have $Z_p \stackrel{\mathrm{d}}{=} A_pY$ with $A_p \perp\!\!\!\perp Y$ (i.e. A_p and Y are independent). Let $Z = AY$. The same argument shows that $Z \mid A \sim \Delta_n(AA')$. This completes the proof. □

Lemma 1.5. Suppose $U \sim T(\delta; I_n, I_p)$. Then the following hold:

$$UU' \sim F(p, \delta; I_n),$$
$$(I_n + UU')^{-1} \sim B(\delta + n - 1, p; I_n),$$
$$(UU')^{-1} \sim F(\delta + n - 1, p - n + 1; I_n), \ (p \geq n),$$
$$U'U \sim F(n, \delta, I_p),$$
$$(I_p + U'U)^{-1} \sim B(\delta + p - 1, n; I_p).$$

Proof. The results follow from the definitions and properties of the matrix distributions given in this Chapter (cf. Dawid 1981, p.266, p.272, p.271). □

1.4 Matrix Algebra

In this section we shall give some formulae of matrix theory needed to investigate the distributions of statistics derived from random matrices.

Lemma 1.6. Suppose $A : p \times p$, $B : q \times q$ and all the other matrices below are of suitable orders, and the inverse matrices concerned exist. Then the following hold.

$$(A+CBD)^{-1} = A^{-1} - A^{-1}CB(B+BDA^{-1}CB)^{-1}BDA^{-1}, \tag{1.1}$$

$$BD(A+CBD)^{-1} = B(B+BDA^{-1}CB)^{-1}BDA^{-1} \tag{1.2}$$

$$(A+F)^{-1} = A^{-1} - A^{-1}F(I+A^{-1}F)^{-1}A^{-1}, \tag{1.3}$$

$$(I+F)^{-1} = I - F(I+F)^{-1} = I - (I+F)^{-1}F, \tag{1.4}$$

$$(I_p + D'D)^{-1} = I_p - D'(I_q + DD')^{-1}D, \tag{1.5}$$

$$D(I_p + D'D)^{-1} = (I_q + DD')^{-1}D, \tag{1.6}$$

Proof. The first identity is from Theorem A 5.1 of Muirhead, 1982. It follows that

$$\begin{aligned} & BD(A+CBD)^{-1} \\ = \; & BDA^{-1} - BDA^{-1}CB(B+BDA^{-1}CB)^{-1}BDA^{-1} \\ = \; & [(B+BDA^{-1}CB) - BDA^{-1}CB](B+BDA^{-1}CB)^{-1}BDA^{-1} \\ = \; & B(B+BDA^{-1}CB)^{-1}BDA^{-1}, \end{aligned}$$

yielding (1.2). The rest can be derived from (1.1) and (1.2). □

Suppose A is $m \times m$ positive–definite. Then there exists an $m \times m$ matrix B, such that $A = B'B$. We can define B as the square root of A, written as $A^{\frac{1}{2}}$, satisfying $(A^{\frac{1}{2}})'A^{\frac{1}{2}} = A$. A symmetric square root can be taken as follows: find an orthogonal matrix P such that

$$A = P\text{diag}(a_1, \cdots, a_m)P',$$

where the a_i's are the latent roots of A. Then let $A^{\frac{1}{2}} = P\mathrm{diag}(a_1^{\frac{1}{2}}, \cdots, a_m^{\frac{1}{2}})P'$. If $A > 0$, $A^{-\frac{1}{2}} = (A^{\frac{1}{2}})^{-1}$ can be defined. Then $(A^{-\frac{1}{2}})'A^{-\frac{1}{2}} = A^{-1}$. We have the following Lemma on the property of the symmetric square root of the nonnegative–definite matrix.

Lemma 1.7. If A, B are nonnegative–definite matrices of order $m \times m$ and $AB = BA$, then

$$(AB)^{\frac{1}{2}} = A^{\frac{1}{2}}B^{\frac{1}{2}} = B^{\frac{1}{2}}A^{\frac{1}{2}},$$

where $C^{\frac{1}{2}}$ denotes the symmetric square root of a matrix C.

Proof. By the assumption, we can obtain simultaneous orthogonal diagonalization (c.f. Press, 1982, p.40), i.e. there is an orthogonal matrix P: $m \times m$, such that

$$P'AP = \mathrm{diag}(a_1, \cdots, a_m),\ P'BP = \mathrm{diag}(b_1, \cdots, b_m),$$

where a_i, b_i are the latent roots of A, B respectively. Hence we can define

$$\begin{aligned} A^{\frac{1}{2}} &= P\mathrm{diag}(a_1^{\frac{1}{2}}, \cdots, a_m^{\frac{1}{2}})P', \\ B^{\frac{1}{2}} &= P\mathrm{diag}(b_1^{\frac{1}{2}}, \cdots, b_m^{\frac{1}{2}})P', \\ (AB)^{\frac{1}{2}} &= P\mathrm{diag}((ab)_1^{\frac{1}{2}}, \cdots, (ab)_m^{\frac{1}{2}})P', \end{aligned}$$

which satisfy the Lemma. □

Chapter 2

CONJUGATE BAYES CONTINUOUS DISCRIMINATION

2.1 Introduction

In this chapter we shall investigate the problem of discrimination between two multivariate normal populations with common dispersion matrix. Fuller details are given in Dawid and Fang (1992), which is submitted as part of this thesis and is attached as Appendix A. For this problem Geisser (1964) obtained posterior probabilities that an observation belongs to one of k multivariate normal distributions under the assumption of the prior reflecting ignorance, the number of variables being finite. We consider the case when the number of variables is unlimited. Along the lines of Dawid (1988), we assume a natural conjugate normal inverted Wishart distribution and study the asymptotic properties of the ratio of the posterior probabilities of the populations. In particular, we derive necessary and sufficient conditions for asymptotically degenerate discrimination, whether or not the parameters are known. Thus a conjugate prior implies asymptotic deter-

minism in discrimination under certain condition, a phenomenon similar to that in regression (Dawid, 1988). However, in many contexts this belief is unreasonable. Hence the assumption of a conjugate prior in Bayesian analysis must be considered according to the problems being investigated.

2.2 Assumptions

Let Y be a binary indicator variable with $Y = i$ denoting membership of Π_i. Suppose that associated with each individual is a countable collection of variables $\mathbf{X} = (X_1, X_2, \cdots,)'$ with normal distribution,

$$\mathbf{X} \mid Y = i \ \sim \ \boldsymbol{\mu}_i + \mathcal{N}(1, \Sigma), \ i = 1, 2, \tag{2.1}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \cdots,)'$ is $\infty \times 1$, $\Sigma = (\sigma_{ij})_{i,j=1,2,\cdots}$ is $\infty \times \infty$. Denote by $\mathbf{X}_p$, $\boldsymbol{\mu}_{ip}$, Σ_p the submatrices of $\mathbf{X}$, $\boldsymbol{\mu}_i$, Σ restricted to the first p variables, and similarly for other matrices.

Suppose the prior for the parameters (μ, Σ) is the conjugate normal inverted Wishart distribution $\mathcal{NIW}(m, H; \delta, K)$,

$$\mu \mid \Sigma \sim m + \mathcal{N}(H, \Sigma), \ \Sigma \sim IW(\delta; K), \tag{2.2}$$

where $\mu = \begin{pmatrix} \boldsymbol{\mu}_1' \\ \boldsymbol{\mu}_2' \end{pmatrix}$ and $m = \begin{pmatrix} \mathbf{m}_1' \\ \mathbf{m}_2' \end{pmatrix}$ are $2 \times \infty$, $H = \begin{pmatrix} h_1^{-1} & 0 \\ 0 & h_2^{-1} \end{pmatrix}$, $\delta > 0$, and K is $\infty \times \infty$. (cf. eqs. (3.1), (3.2) in Dawid and Fang, 1992).

2.3 Main Results

We consider in several cases the asymptotic behaviour of the ratio of the predictive probabilities.

2.3.1 Known Parameters

Suppose all parameters are known. Then the ratio of the conditional probabilities of $Y = i$ given $\mathbf{X}_p$ is

$$\frac{P(Y=2 \mid \mathbf{X}_p; \mu, \Sigma)}{P(Y=1 \mid \mathbf{X}_p; \mu, \Sigma)} = \frac{P(Y=2) f(\mathbf{X}_p \mid Y=2; \mu, \Sigma)}{P(Y=1) f(\mathbf{X}_p \mid Y=1; \mu, \Sigma)}, \tag{2.3}$$

where $f(\mathbf{X}_p \mid Y = i; \mu, \Sigma)$ is the density of $N(\boldsymbol{\mu}_{ip}, \Sigma_p)$, (cf. eqs. (2.2), (2.1) in Dawid and Fang, 1992). Let

$$\lambda_p = (\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p})' \Sigma_p^{-1} (\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p}),$$

the Mahalanobis distance between the populations Π_1 and Π_2 based on $\mathbf{X}_p$. The limit of λ_p as $p \to \infty$ exists and is denoted by λ_∞. If $\mathbf{X}$ arises from Π_1, (2.3) converges a.s. to a random variable distributed as $(\pi_2/\pi_1)\exp\{N(-\lambda_\infty/2, \lambda_\infty)\}$ if $\lambda_\infty < \infty$, 0 if $\lambda_\infty = \infty$, (cf. eq. (2.3) in Dawid and Fang, 1992). A parallel result holds if $\mathbf{X}$ arises from Π_2, (with $N(-\lambda_\infty/2, \lambda_\infty)$, 0 replaced by $N(\lambda_\infty/2, \lambda_\infty)$, ∞ respectively, cf. eq.(2.4) in Dawid and Fang, 1992). Hence the condition $\lambda_\infty = \infty$ is necessary and sufficient for asymptotically degenerate discrimination between the two populations.

2.3.2 With Prior Distribution

Suppose that the parameters are assigned the prior distribution (2.2). Then $\lambda_p \xrightarrow{\text{a.s.}} \infty$ as $p \to \infty$. Hence by result 2.3.1 we expect the parameters to be such as to permit asymptotically degenerate discrimination, were their values to be known.

2.3.3 Unknown Parameters

Suppose that the parameters are unknown, but are assigned the conjugate prior distribution (2.2). Then the ratio of the conditional probabilities of $Y = i$ given $\mathbf{X}_p$ (with the parameters μ, Σ integrated out) is

$$\frac{P(Y=2 \mid \mathbf{X}_p)}{P(Y=1 \mid \mathbf{X}_p)} = \frac{P(Y=2) f(\mathbf{X}_p \mid Y=2)}{P(Y=1) f(\mathbf{X}_p \mid Y=1)}, \tag{2.4}$$

where $f(\mathbf{X}_p \mid Y = i)$ is the density of $\mathbf{m}_i + T(\delta; K, k_i)$, $k_i = 1 + h_i^{-1}$, (cf. eqs. (4.3), (4.1) in Dawid and Fang, 1992). Let

$$\gamma_p = (\mathbf{m}_{1p} - \mathbf{m}_{2p})' K_p^{-1} (\mathbf{m}_{1p} - \mathbf{m}_{2p})$$

be the analogue of λ_p as the function of the hyper-parameters in the prior distribution for $(\mu,\ \Sigma)$. The limit of γ_p as $p \to \infty$ exists and is denoted by γ_∞. If $\mathbf{X}$ arises from Π_1, (2.4) converges a.s. to a random variable distributed as $(\pi_2/\pi_1)(k_2/k_1)^{\frac{\delta}{2}} \exp \Omega$, where the distribution of Ω is the mixture, over the distribution $(\chi^2_\delta)^{-1}$ for Λ, of $N(-(k_1\Lambda)^{-1}\gamma_\infty/2, (k_1\Lambda)^{-1}\gamma_\infty)$, if $\gamma_\infty < \infty$, 0, if $\gamma_\infty = \infty$, (cf. eq. (4.8) in Dawid and Fang, 1992). A parallel result holds if $\mathbf{X}$ arises from Π_2. Thus a necessary and sufficient condition for asymptotically degenerate discrimination between the two populations in the absence of knowledge of the parameters is that $\gamma_\infty = \infty$.

2.3.4 Using Training Data

Suppose the parameters are unknown and assigned the prior (2.2) . Suppose also we have training data of Y and $\mathbf{X}$ for a random sample of n individuals

$$\begin{pmatrix} \mathbf{y}^n & x^n \\ 1 & \infty \end{pmatrix} \begin{matrix} n \\ \\ \end{matrix},$$

where, without loss of generality, we suppose the first n_1 components of $\mathbf{y}^n$ are 1, the next n_2 components are 2, and partition x^n as

$$x^n = \begin{pmatrix} x^n_{(1)} \\ x^n_{(2)} \end{pmatrix} \begin{matrix} n_1 \\ n_2 \end{matrix},$$

$n_1 + n_2 = n$. Now on a further individual we observe the values $\mathbf{x}^0_p$ of the first p variables and wish to predict Y^0. By result 2.3.2, in the posterior distribution of the parameters given $(\mathbf{Y}^n, X^n)$, $\lambda_\infty = \infty$ with probability 1 and thus with probability 1 the parameters will be such as to support asymptotically degenerate discrimination, were they to be known. Moreover, if $\gamma_\infty = \infty$, by result 2.3.3, with probability 1, the predictive distribution of $(Y^0, \mathbf{X}^0_p)$ given $(\mathbf{Y}^n, X^n)$ allows

asymptotically almost sure identification of Y^0 on the basis of $\mathbf{X}_p^0$, even when the parameters are unknown. Let

$$\gamma_p^*(x^n) = (\mathbf{m}_{1p}^*(x^n) - \mathbf{m}_{2p}^*(x^n))' K_p^*(x^n)^{-1} (\mathbf{m}_{1p}^*(x^n) - \mathbf{m}_{2p}^*(x^n)),$$

where

$$\begin{aligned}
\mathbf{m}_i^*(x^n) &= (n_i \bar{\mathbf{x}}_i^n + h_i \mathbf{m}_i)/(h_i + n_i), \\
K_p^*(x^n) &= K_p + (x_p^n - \Gamma_n m_p)' Q_{nn}^{-1} (x_p^n - \Gamma_n m_p), \\
\bar{x}_i^n &= (x_{(i)}^n)' \mathbf{1}_{n_i}/n_i, \ \mathbf{1}_{n_i}' = (1, \cdots, 1) \text{ is } 1 \times n_i, \ i = 1, 2, \\
\Gamma_n &= \begin{pmatrix} \mathbf{1}_{n_1} & 0 \\ 0 & \mathbf{1}_{n_2} \end{pmatrix}, \ Q_{nn} = I_n + \Gamma_n H \Gamma_n'.
\end{aligned}$$

Then the predictive odds are

$$\frac{P(Y^0 = 2 \mid \mathbf{X}_p^0, x^n, \mathbf{y}^n)}{P(Y^0 = 1 \mid \mathbf{X}_p^0, x^n, \mathbf{y}^n)} = \frac{P(Y^0 = 2) f(\mathbf{X}_p^0 \mid Y = 2; x^n, \mathbf{y}^n)}{P(Y^0 = 1) f(\mathbf{X}_p^0 \mid Y = 1; x^n, \mathbf{y}^n)}, \tag{2.5}$$

where $f(\mathbf{X}_p^0 \mid Y = i; x^n, \mathbf{y}^n)$ is the density of $m_{ip}^*(x^n) + T(\delta^*; K_p^*(x^n), k_i^*)$, with $\delta^* = \delta + n$, $k_i^* = 1 + (n_i + h_i)^{-1}$, (cf. eqs.(5.14), (5.12) in Dawid and Fang, 1992).

If $\gamma_\infty < \infty$, then $\gamma_p^*(X^n)$ converges a.s. to a limit

$$\gamma_\infty^*(X^n) = \gamma_\infty + n_1 h_1^{-1} (n_1 + h_1)^{-1} + n_2 h_2^{-1} (n_2 + h_2)^{-1} \stackrel{\text{def}}{=} \gamma_\infty^*,$$

(cf. eq.(5.17) in Dawid and Fang, 1992). Hence by result 2.3.3, if $\mathbf{X}_p^0$ arises from Π_1

$$\frac{P(Y^0 = 2 \mid \mathbf{X}_p^0, X^n, \mathbf{y}^n)}{P(Y^0 = 1 \mid \mathbf{X}_p^0, X^n, \mathbf{y}^n)}$$

(cf. eq.(5.18) in Dawid and Fang, 1992) converges a.s. (under the distribution of $(\mathbf{X}^0, X^n)$ given $\mathbf{y}^n$ and $Y^0 = 1$), as $p \to \infty$, to a limit whose distribution is the mixture of

$$\frac{\pi_2}{\pi_1} \left(\frac{k_2^*}{k_1^*} \right)^{\delta^*/2} \exp \left\{ N \left(-\frac{1}{2} (k_1^* \Lambda^*)^{-1} \gamma_\infty^*, (k_1^* \Lambda^*)^{-1} \gamma_\infty^* \right) \right\}$$

over the distribution $(\chi_{\delta^*}^2)^{-1}$ for Λ^*, (cf. eq.(5.19) in Dawid and Fang, 1992). A parallel result holds if $\mathbf{X}_p^0$ arises from Π_2. Furthermore, the predictive odds

$$\frac{P(Y^0 = 2 \mid \mathbf{X}_p^0, X^n, \mathbf{Y}^n)}{P(Y^0 = 1 \mid \mathbf{X}_p^0, X^n, \mathbf{Y}^n)} \tag{2.6}$$

(cf. eq.(5.20) in Dawid and Fang, 1992). is finite a.s. (under $Y^0 = 1$ or $Y^0 = 2$) as $p \to \infty$ ($\gamma_\infty < \infty$).

In the case when n is also large, the double limit $\lim_{(n,p)\to(\infty,\infty)}$ and the repeated limits $\lim_{p\to\infty}\lim_{n\to\infty}$, $\lim_{n\to\infty}\lim_{p\to\infty}$, of (2.6) are the same and equal ∞ if $Y^0 = 2$, 0 if $Y^0 = 1$. Thus perfect discrimination is possible using extensive training data. More analysis will be made in Chapter 4 for this normal discrimination problem.

Chapter 3

CONJUGATE BAYES DISCRETE DISCRIMINATION

3.1 Introduction

In this chapter, we investigate a problem of discrimination between two populations with binary variables. Fuller details are given in Fang and Dawid (1993), which is submitted as part of this thesis as Appendix B. As in Chapter 2, we study the asymptotic property of the ratio of the probabilities of the populations conditioned on an observation as the number of variables tends to infinity. We show that the conjugate Dirichlet process prior (Ferguson, 1973) implies asymptotically perfect discrimination under certain conditions on the parameters. This extends the result obtained by Brown (1980) that, under the assumption of "uniform refinement", the prior expectation of the probability of correct classification tends to 1 as the number of predictors tends to infinity.

3.2 Assumptions

Let $\pi_i = P(\Pi_i)$, the probability of population Π_i, $i = 1, 2$. Suppose we observe $X_1, X_2, \cdots$, where X_i takes value 0 or 1. Let

$$\mathbf{X} = (X_1, X_2 \cdots), \quad \mathbf{X}^n = (X_1, \cdots, X_n).$$

Let θ [resp. ϕ] denote the joint distribution of X in Π_1 [resp. Π_2]: these are measures over the Borel σ–field $\mathcal{B}^\infty$ of $\{0,1\}^\infty$. By Kolmogorov's consistency theorem, θ is determined by its restriction, θ^n say, to each $\mathcal{B}^n$: we write $\theta^n(\mathbf{x}^n) = P(\mathbf{X}^n = \mathbf{x}^n \mid \Pi_1)$, etc.

Suppose the parameters θ are assigned Dirichlet process prior $D(\alpha)$, i.e. for each n, θ^n has the Dirichlet distribution $D(\alpha^n)$, with parameter α^n, the restriction of α to $\mathcal{B}^n$, where α is a finite measure over $\mathcal{B}^\infty$, with total mass $|\alpha|$, (cf. Ferguson, 1973). We similarly take $\Phi \sim D(\beta)$, and $\theta \perp\!\!\!\perp \phi$, thus specifying the joint prior distribution of (θ, ϕ).

Let $\lambda(\mathbf{x}^n) = \theta^n(\mathbf{x}^n)/\phi^n(\mathbf{x}^n)$, the likelihood ratio in favour of Π_1 as against Π_2, based on data $\mathbf{X}^n = \mathbf{x}^n$, when θ and ϕ are given; and let $\Lambda_n = \lambda_n(\mathbf{X}^n)$, a function of θ, ϕ and $\mathbf{X}$. We shall study the asymptotic behaviour of Λ_n as $n \to \infty$.

3.3 Main Results

We consider in several cases the asymptotic behaviour of the likelihood ratio as $n \to \infty$.

3.3.1 Known Parameters

Suppose the probability of Π_i, π_i, $i = 1, 2$ are known. Theorem 1 and Theorem 3 below give the conditions for asymptotically perfect discrimination between Π_1 and Π_2. Theorem 2 provides the basis for Theorem 3.

Theorem 1. Suppose that α and β are both non–atomic measures. Then, as $n \to \infty$,

$$\Lambda_n \xrightarrow{\text{a.s.}} \infty \quad \text{if } \Pi = \Pi_1,$$
$$\Lambda_n \xrightarrow{\text{a.s.}} 0 \quad \text{if } \Pi = \Pi_2. \tag{3.1}$$

Theorem 2. Suppose that α has decomposition $\alpha = \lambda + \mu$, where λ is continuous and μ is discrete. Arrange the atoms $\{\mathbf{x}_j\}$ of μ in descending order of $m_j = \mu(\mathbf{x}_j)$. If $\Pi = \Pi_1$, then as $n \to \infty$, the asymptotic distribution of $\theta_n^* \stackrel{\text{def}}{=} \theta^n(\mathbf{X}^n)$ is nondegenerate with p.d.f.

$$|\lambda|(1-y)^{|\alpha|-1} + \sum_{k=1}^{\infty} y^{m_k}(1-y)^{|\alpha|-m_k-1}/B(m_k, |\alpha| - m_k),$$
$$0 < y < 1.$$

Theorem 3. Perfect discrimination property (3.1) holds if and only if

$$\alpha \text{ and } \beta \text{ do not have common atoms.} \tag{3.2}$$

Now suppose we get training data from Π_1 and Π_2 and a new observation $\mathbf{X}$ which is to be classified as belonging to one of the Π_i. If α and β do not have common atoms, the parameters of θ and ϕ conditioned on the training data also do not (with probability 1) have common atoms. Hence (3.2) is necessary and sufficient for asymptotically perfect discrimination between the two populations using training data.

3.3.2 Unknown Parameters.

If θ and ϕ are unknown, then the likelihood ratio relevant for classification is

$$\gamma_n(\mathbf{x}^n) = \alpha_0^n(\mathbf{x}^n)/\beta_0^n(\mathbf{x}^n),$$

where $\alpha_0 = \alpha/|\alpha|$ is the marginal distribution for $\mathbf{X}$ in Π_1, when $\theta \sim D(\alpha)$; and similarly fo β_0. Let $\Gamma_n = \gamma_n(\mathbf{X}^n)$. Theorem 4 below gives the asymptotic property of Γ_n as $n \to \infty$.

Theorem 4.

$$\Gamma_n \xrightarrow{\text{a.s.}} \begin{cases} \infty & \text{if } \mathbf{X} \sim \alpha_0 \\ 0 & \text{if } \mathbf{X} \sim \beta_0 \end{cases}$$

if and only if α and β are mutually singular; while Γ_n is almost surely bounded away from 0 and ∞, under both α_0 and β_0, if and only if α and β are mutually absolutely continuous.

In particular, perfect discrimination in the absence of knowledge of the parameters will not be almost certain unless α and β are mutually singular—a much stronger condition than that of Theorem 3.

If we have N_i training cases from Π_i $(i = 1, 2)$. Then asymptotically perfect discrimination $(n \to \infty)$ is still possible if α and β are mutually singular or α and β are mutually absolutely continuous but the number of traning data tends to infinity $(N_i \to \infty, i = 1, 2)$.

3.3.3 Unknown Prior Probabilities

Suppose the prior probilities π_i are unknown. Let Y be a variable, taking values 1 and 2, indicating the correct population, and jointly distributed with $\mathbf{X}$, with $P(Y = i) = \pi_i$. The parameter for the distribution of $(Y, \mathbf{X})$ is now (θ, ϕ, π_1), with prior distribution given by

$$\theta \sim D(\alpha), \quad \phi \sim D(\beta),$$
$$\pi_1 \sim Beta(|\alpha|, |\beta|),$$

all independently.

In this case the asymptotic discrimination behaviour will be the same as for the case of known π_1.

Chapter 4

COMPARISON IN DISCRIMINATION

4.1 Introduction

In this Chapter we investigate the connection between the Bayesian approach and the classical approach in the problem of discrimination between two homoscedastic multivariate normal populations when the number of variables that can be observed is allowed to tend to infinity. Geisser (1964) considered the problem of discrimination between k multivariate populations using Bayesian methods. In Chapter 2 (cf. Dawid and Fang, 1992) we consider the case of discrimination between two normal populations with infinitely many variables, and showed that under certain conditions the conjugate prior will imply asymptotically degenerate discrimination, i.e. the ratio of the probability of the second population to that of the first population conditioned on the data will tend to zero or infinity according as the observation comes from the first or second population. In this Chapter we shall investigate this problem more deeply. We shall study a Bayes linear discriminant, the coefficient of which is obtained by maximizing the performance of the linear function of the observation to be classified conditioned on the training

data. We shall show that under certain conditions the performance of this Bayes discriminant tends to infinity and the probability of its correct classification tends to one as the number of the variables tends to infinity. Thus this Bayes criterion implies degenerate discrimination. In the classical approach, if the number of variables is unlimited, we can always find a best linear discriminant function maximizing the ratio of between-class sample variance to the within-class sample variance by taking the denominator as zero, which means it will classify the data exactly. If for uniqueness we choose one which maximizes the performance conditioned on the data, then we obtain a mixture solution. In another word, we define the Bayes criterion a maximizing the performance conditioned on the data and consider two estimators, of which one is unrestricted optimal, while the other is restricted optimal, subject to the condition that its within-class sample variance be zero. We shall show that the performance of this solution is asymptotically equivalent to that of the full Bayes solution using a conjugate prior. Thus the Bayes discriminant is very close to the sample-based discriminant, and so neglects the problem of selection bias (Dawid, 1994).

4.2 Assumptions

Suppose that, for each population Π_i, $\mathbf{X}_p = (X_1, \cdots, X_p)'$ has a multivariate normal distribution $N(\boldsymbol{\mu}_{ip}, \Sigma_p)$, $i = 1, 2$, where $\boldsymbol{\mu}_{ip} = (\mu_{i1}, \cdots, \mu_{ip})'$ is a p–dimensional vector, $i = 1, 2$, $\Sigma_p = (\sigma_{ij}), i, j = 1, \cdots, p$, is a $p \times p$ matrix. Discrimination may be based on a linear function $Y_{\mathbf{a}_p} \stackrel{\text{def}}{=} \mathbf{a}_p'\mathbf{X}_p$. The performance of $Y_{\mathbf{a}_p}$ is defined as the squared difference of its expectations in the two populations normalized by its variance, i.e.

$$\phi_{\mathbf{a}_p} = [E_1(Y_{\mathbf{a}_p}) - E_2(Y_{\mathbf{a}_p})]^2/\mathrm{Var}(Y_{\mathbf{a}_p}) = [\mathbf{a}_p'(\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p})]^2/\mathbf{a}_p'\Sigma_p\mathbf{a}_p. \tag{4.1}$$

Suppose we have training data on n_i cases from Π_i, and on each we observe potentially infinitely many variables $\mathbf{X} = (X_1, X_2, \cdots)'$, thus obtaining a $n \times \infty$

matrix X^n $(n_1 + n_2 = n)$. Without loss of generality we suppose its first n_1 rows arise from population Π_1, the next n_2 rows from Π_2. We then have

$$X^n \sim \Gamma\mu + \mathcal{N}(I_n, \Sigma), \tag{4.2}$$

where

$$\begin{aligned} \mu &= \begin{pmatrix} \boldsymbol{\mu}_1' \\ \boldsymbol{\mu}_2' \end{pmatrix}, \ \boldsymbol{\mu}_i' = (\mu_{i1}, \mu_{i2}, \cdots), \ i = 1, 2, \\ \Sigma &= (\sigma_{ij}), \ i, j = 1, 2, \cdots, \\ \Gamma &= \begin{pmatrix} \mathbf{1}_{n_1} & 0 \\ 0 & \mathbf{1}_{n_2} \end{pmatrix}, \ n = n_1 + n_2, \end{aligned} \tag{4.3}$$

and $\mathbf{1}_{n_i}$ is a n_i-dimensional vector with all components being 1.

Suppose the prior distribution of the parameters (μ, Σ) is the conjugate inverse Wishart distribution $\mathcal{NIW}(m, H; \delta, Q)$:

$$\mu \mid \Sigma \ \sim \ m + \mathcal{N}(H, \Sigma), \ \Sigma \sim IW(\delta; Q), \tag{4.4}$$

where H (2×2), m $(2 \times \infty), \delta > 0$, Q $(\infty \times \infty) > 0$ are given,

$$m = \begin{pmatrix} \mathbf{m}_1' \\ \mathbf{m}_2' \end{pmatrix} \ , \ H = \begin{pmatrix} h_1^{-1} & 0 \\ 0 & h_2^{-1} \end{pmatrix} \ . \tag{4.5}$$

Thus the marginal distribution of X^n is

$$X^n \sim \Gamma m + T(\delta; G, Q), \tag{4.6}$$

where

$$G = I_n + \Gamma H \Gamma'. \tag{4.7}$$

In what follows we write $\mathbf{X}_p$, X_p^n, Σ_p for the submatrices of $\mathbf{X}$, X^n, Σ restricted to the first p variables, and similarly for other matrices.

4.3 Posterior Distribution

To make Bayesian inference we need the posterior distribution of the parameters $(\mu,\ \Sigma)$ given the data X^n. Under the assumptions (4.2)–(4.5), the result is well known (see, for example, Press, 1982). The following lemma gives a somewhat more general form for r populations.

Lemma 4.1. Suppose the distribution of X^n and the prior distribution of the parameters are given by (4.2) and (4.4), where X^n is $n \times \infty$, Γ is $n \times r$, μ is $r \times \infty$, $\Sigma > 0$ is $\infty \times \infty$, m is $r \times \infty$, H is $r \times r$, $Q > 0$ is $\infty \times \infty$, and $\delta > 0$. Then the posterior distribution of $(\mu_p,\ \Sigma_p)$ conditioned on the data X^n is given by

$$\begin{aligned} \mu_p \mid (\Sigma_p,\ X^n) &\sim m_p^* + \mathcal{N}(H^*, \Sigma_p), \\ \Sigma_p \mid X^n &\sim IW(\delta^*; Q_p^*) \end{aligned} \tag{4.8}$$

where

$$\begin{aligned} H^* &= (\Gamma'\Gamma + H^{-1})^{-1}, \\ m_p^* &= H^*\Gamma'(X_p^n - \Gamma m_p) + m_p, \\ \delta^* &= \delta + n, \\ Q_p^* &= Q_p + (X_p^n - \Gamma m_p)'G^{-1}(X_p^n - \Gamma m_p), \end{aligned}$$

with G given by (4.7).

Proof. We first note, by Lemma 1 of Dawid and Fang (1992), that the conditional distribution of μ_p, Σ_p given the full data matrix X^n is the same as that given X_p^n, the data on the first p X's only. The density of $(\mu_p, \Sigma_p) \mid X_p^n$ is proportional to that of $(\mu_p,\ \Sigma_p,\ X_p^n)$, the product of the three densities of $X_p^n \mid (\mu_p, \Sigma_p)$, $\mu_p \mid \Sigma_p$ and Σ_p. It can be checked that for m_p^*, Q_p^* defined in the Lemma, the following equations hold:

$$m_p^* = (\Gamma'\Gamma + H^{-1})^{-1}(\Gamma' X_p^n + H^{-1} m_p),$$

$$
\begin{aligned}
Q_p^* - Q_p &= (X_p^n)'X_p^n + m_p' H^{-1} m_p - (\Gamma' X_p^n + H^{-1} m_p)' \\
&\quad \times (\Gamma'\Gamma + H^{-1})^{-1} (\Gamma' X_p^n + H^{-1} m_p).
\end{aligned}
$$

Using the above equations to rearrange the terms in etr$(\cdot)$ of the density of $(\mu_p, \Sigma_p,\ X_p^n)$, we establish the Lemma. □

Now suppose (4.3) and (4.5) hold. Let X^n be partitioned as $X^n = \begin{pmatrix} X_{(1)}^n \\ X_{(2)}^n \end{pmatrix}$ with $X_{(i)}^n$ being $(n_i \times \infty)$ and

$$
\begin{aligned}
\overline{X}_{ip} &= (X_{(i)p}^n)' \mathbf{1}_{n_i} / n_i, \\
\overline{\overline{X}}_p &= (\Gamma'\Gamma)^{-1}\Gamma' X_p^n = \begin{pmatrix} (\overline{X}_{1p}^n)' \\ (\overline{X}_{2p}^n)' \end{pmatrix} = (\Gamma'\Gamma)^{-1}\Gamma'(X_p^n - \Gamma m_p) + m_p, \\
S_{ip} &= (X_{(i)p}^n)'(I_{n_i} - \mathbf{1}_{n_i}\mathbf{1}_{n_i}'/n_i) X_{(i)p}^n, \\
S_p &= (X_p^n)'(I_n - \Gamma(\Gamma'\Gamma)^{-1}\Gamma') X_p^n \\
&= (X_p^n - \Gamma m_p)'(I_n - \Gamma(\Gamma'\Gamma)^{-1}\Gamma')(X_p^n - \Gamma m_p) \\
&= \sum_{i=1}^{2} S_{ip}.
\end{aligned}
$$

Then (4.8) holds with

$$
\begin{aligned}
H^* &= \begin{pmatrix} h_1^{*-1} & 0 \\ 0 & h_2^{*-1} \end{pmatrix}, \ h_i^* = n_i + h_i, \\
m_p^* &= H^*\Gamma'(X_p^n - \Gamma m_p) + m_p = (\mathbf{m}_{1p}^* \ \mathbf{m}_{2p}^*)', \\
\mathbf{m}_{ip}^* &= h_i^{*-1}(n_i \overline{X}_{ip} + h_i \mathbf{m}_{ip}), \\
\delta^* &= \delta + n_1 + n_2, \\
Q_p^* &= Q_p + (X_p^n - \Gamma m_p)'(I_n + \Gamma H \Gamma')^{-1}(X_p^n - \Gamma m_p) \\
&= Q_p + S_p + \sum_{i=1}^{2} (n_i^{-1} + h_i^{-1})^{-1} (\overline{X}_{ip} - \mathbf{m}_{ip})(\overline{X}_{ip} - \mathbf{m}_{ip})'.
\end{aligned}
$$

4.4 Discrimination

We shall be interested in finding a linear function $\mathbf{a}_p'\mathbf{x}_p$ which can be used to discriminate best between the two populations according to some criterion, on the basis of the training data X^n. When a new observation $\mathbf{X}_p$, the value of the first p X's of a new individual, is obtained, we then use $\mathbf{a}_p'\mathbf{X}_p$ to decide which population it comes from.

4.4.1 Bayesian Approach

One Bayesian approach is to choose $\mathbf{a}_p$ such that $E(\phi_{\mathbf{a}_p} \mid X^n)$ attains its maximum, with $\phi_{\mathbf{a}_p}$ given by (4.1).

Proposition 4.1. Under the assumptions (4.2)–(4.5), the posterior expectation of $\phi_{\mathbf{a}_p}$ given data X^n is

$$\begin{aligned} E(\phi_{\mathbf{a}_p} \mid X^n) &= \boldsymbol{\alpha}' H^* \boldsymbol{\alpha} + \delta^* \frac{(\boldsymbol{\alpha}' m_p^* \mathbf{a}_p)^2}{\mathbf{a}_p' Q_p^* \mathbf{a}_p} \\ &= h_1^{*^{-1}} + h_2^{*^{-1}} + \delta^* \frac{[\mathbf{a}_p'(\mathbf{m}_{1p}^* - \mathbf{m}_{2p}^*)]^2}{\mathbf{a}_p' Q_p^* \mathbf{a}_p}, \end{aligned}$$

where H^*, m_p^*, δ^*, Q_p^* are given by the equations following Lemma 4.1, $\boldsymbol{\alpha} = (1, -1)'$.

Proof. By Lemma 4.1, $\boldsymbol{\alpha}'\mu_p\mathbf{a}_p \mid \Sigma_p, X^n \sim \boldsymbol{\alpha}' m_p^* \mathbf{a}_p + \mathcal{N}(\boldsymbol{\alpha}' H^* \boldsymbol{\alpha}, \mathbf{a}_p' \Sigma_p \mathbf{a}_p)$. Hence

$$E[(\boldsymbol{\alpha}'\mu_p\mathbf{a}_p)^2 \mid \Sigma_p, X^n] = (\boldsymbol{\alpha}' H^* \boldsymbol{\alpha})(\mathbf{a}_p' \Sigma_p \mathbf{a}_p) + (\boldsymbol{\alpha}' m_p^* \mathbf{a}_p)^2,$$

and so

$$E(\phi_{\mathbf{a}_p} \mid \Sigma_p, X^n) = \boldsymbol{\alpha}' H^* \boldsymbol{\alpha} + (\mathbf{a}_p' \Sigma_p \mathbf{a}_p)^{-1} (\boldsymbol{\alpha}' m_p^* \mathbf{a}_p)^2.$$

Since $\Sigma_p \mid X^n \sim IW(\delta^*; Q_p^*)$, $(\mathbf{a}_p'\Sigma_p\mathbf{a}_p)^{-1} \mid X^n \sim W(\delta^*; (\mathbf{a}_p' Q^* \mathbf{a}_p)^{-1})$. Hence

$$E[(\mathbf{a}_p'\Sigma_p\mathbf{a}_p)^{-1} \mid X^n] = \delta^* (\mathbf{a}_p' Q^* \mathbf{a}_p)^{-1}.$$

Substituting the above two expectations into

$$E(\phi_{\mathbf{a}_p} \mid X^n) = E[E(\phi_{\mathbf{a}_p} \mid \Sigma_p, X^n) \mid X^n],$$

we get the desired result. □

Corollary 4.1. The Bayesian solution, achieving $\max_{\mathbf{a}_p} E(\phi_{\mathbf{a}_p} \mid X^n)$, is

$$\mathbf{a}_p \propto \mathbf{a}_p^B \stackrel{\text{def}}{=} Q_p^{*-1} m_p^{*\prime} \boldsymbol{\alpha} / \gamma_p(X^n)^{\frac{1}{2}} = Q_p^{*-1}(\mathbf{m}_{1p}^* - \mathbf{m}_{2p}^*)/\gamma_p(X^n)^{\frac{1}{2}}$$

and the corresponding maximum is

$$\max_{\mathbf{a}_p} E(\phi_{\mathbf{a}_p} \mid X^n) = \boldsymbol{\alpha}' H^* \boldsymbol{\alpha} + \delta^* \gamma_p(X^n),$$

where H^*, m^*, δ^*, Q_p^* are given by the equations following Lemma 4.1, $\boldsymbol{\alpha} = (1, -1)'$, and

$$\gamma_p(X^n) = \boldsymbol{\alpha}' m_p^* Q_p^{*-1} m_p^{*\prime} \boldsymbol{\alpha} = (\mathbf{m}_{1p}^* - \mathbf{m}_{2p}^*)' Q_p^{*-1}(\mathbf{m}_{1p}^* - \mathbf{m}_{2p}^*).$$

Corollary 4.1 shows that the coefficient of the Bayes discriminant function is proportional to the difference of the posterior means (transformed by Q_p^{*-1}) between the two populations.

4.4.2 Classical Approach

A classical approach is to choose $\mathbf{a}_p$ to maximise the sample analogue of $\phi_{\mathbf{a}_p}$:

$$Z_{\mathbf{a}_p} \stackrel{\text{def}}{=} [\mathbf{a}_p'(\overline{X}_{1p} - \overline{X}_{2p})]^2 / \mathbf{a}_p' S_p \mathbf{a}_p,$$

where S_p, $\overline{X}_{ip}$ are given in Section 3. If $p \leq n-2$, $S_p > 0$ with probability one, and the solution to $\max_{\mathbf{a}_p} Z_{\mathbf{a}_p}$ is $\mathbf{a}_p \propto S_p^{-1}(\overline{X}_{1p} - \overline{X}_{2p})$, i.e. the coefficient of the discriminant function is proportional to the difference of the sample means

between the two populations (transformed by S_p^{-1}). However, if $p > n-2$, S_p is singular. Then $Z_{\mathbf{a}_p}$ attains its maximum value, viz. ∞, at any $\mathbf{a}_p$ such that $S_p\mathbf{a}_p = 0$ and $\mathbf{a}_p'(\overline{X}_{1p} - \overline{X}_{2p}) \neq 0$. For definiteness we shall choose $\mathbf{a}_p$ such that it also maximizes $E(\phi_{\mathbf{a}_p} \mid X^n)$ subject to these conditions, and make comparison with the full Bayes solution $\mathbf{a}_p^B$.

Lemma 4.2. Suppose $\xi \in R^p$, A and S are symmetric, $p \times p$, A is positive-definite, $\text{rank}(S) = r \leq p$. Then the solution of

$$\max_{\mathbf{a}_p} \frac{(\mathbf{a}_p'\xi)^2}{\mathbf{a}_p' A \mathbf{a}_p} \quad \text{subject to} \quad S\mathbf{a}_p = 0$$

is

$$\mathbf{a}_p \propto A^{-\frac{1}{2}} P(\mathcal{N}(A^{-\frac{1}{2}} S A^{-\frac{1}{2}})) A^{-\frac{1}{2}} \xi$$

and the corresponding maximum is

$$\max_{\mathbf{a}_p : S\mathbf{a}_p = 0} \frac{(\mathbf{a}_p'\xi)^2}{\mathbf{a}_p' A \mathbf{a}_p} = \xi' A^{-\frac{1}{2}} P(\mathcal{N}(A^{-\frac{1}{2}} S A^{-\frac{1}{2}})) A^{-\frac{1}{2}} \xi,$$

where $A^{\frac{1}{2}}$ is the symmetric square root of A, $P(\mathcal{V})$ denotes the orthogonal projection onto a space $\mathcal{V}$, and $\mathcal{N}(U)$ denotes the null space of a matrix U.

Proof. Let $A^{\frac{1}{2}}$ be the symmetric square root of A, $\xi_0 = A^{-\frac{1}{2}}\xi$, $S_0 = A^{-\frac{1}{2}} S A^{-\frac{1}{2}}$, $\mathbf{b} = A^{\frac{1}{2}}\mathbf{a}_p$. Then maximizing $(\mathbf{a}_p'\xi)^2/\mathbf{a}_p' A \mathbf{a}_p$ subject to $S\mathbf{a}_p = 0$ is equivalent to maximizing $(\mathbf{b}'\xi_0)^2/\mathbf{b}'\mathbf{b}$ subject to $S_0\mathbf{b} = 0$. Let $\hat{\xi}_0 = P(\mathcal{N}(S_0))\xi_0$. If $\mathbf{b} \in \mathcal{N}(S_0)$, $\xi_0'\mathbf{b} = \hat{\xi}_0'\mathbf{b}$. Hence

$$(\xi_0'\mathbf{b})^2 = (\hat{\xi}_0'\mathbf{b})^2 \leq \hat{\xi}_0'\hat{\xi}_0 \cdot \mathbf{b}'\mathbf{b}$$

with equality holding iff $\mathbf{b} \propto \hat{\xi}_0$, completing the proof (cf. (1c.6.3) in Rao, 1973 p.50, for alternative proof). □

A direct consequence of Lemma 4.2 is the "mixture" solution, denoted by $\mathbf{a}_p^L$ in the following Corollary.

Corollary 4.2. The mixture solution, achieving $\max_{\mathbf{a}_p} E(\phi_{\mathbf{a}_p} \mid X^n)$ subject to $S_p\mathbf{a}_p = 0,\ p > n-2$, is

$$\begin{aligned}\mathbf{a}_p \propto \mathbf{a}_p^L &\stackrel{\text{def}}{=} P_0^* m_p^{*\prime}\boldsymbol{\alpha}/\gamma_p^L(X^n)^{\frac{1}{2}} \\ &= P_0^*(\mathbf{m}_{1p}^* - \mathbf{m}_{2p}^*)/\gamma_p^L(X^n)^{\frac{1}{2}}\end{aligned}$$

and the corresponding maximum is

$$\max_{\mathbf{a}_p : S_p\mathbf{a}_p = 0} E(\phi_{\mathbf{a}_p} \mid X^n) = \boldsymbol{\alpha}' H^* \boldsymbol{\alpha} + \delta^* \gamma_p^L(X^n),$$

where H^*, m_p^*, δ^*, Q_p^*, are given by the equations following Lemma 4.1, $\boldsymbol{\alpha} = (1,-1)'$, and

$$\begin{aligned}P_0^* &= Q_p^{*-\frac{1}{2}} P(\mathcal{N}(Q_p^{*-\frac{1}{2}} S_p Q_p^{*-\frac{1}{2}})) Q_p^{*-\frac{1}{2}}, \\ \gamma_p^L(X^n) &= \boldsymbol{\alpha}' m_p^* P_0^* m_p^{*\prime}\boldsymbol{\alpha} \\ &= (\mathbf{m}_{1p}^* - \mathbf{m}_{2p}^*)' P_0^* (\mathbf{m}_{1p}^* - \mathbf{m}_{2p}^*).\end{aligned}$$

Recall that the coefficient of the full Bayes discriminant function (transformed by $Q_p^{*\frac{1}{2}}$) is proportional to the difference of the posterior means (transformed by $Q_p^{*-\frac{1}{2}}$) between the two populations. Adding the restriction of least squares ($S_p\mathbf{a}_p = 0$) requires the transformed coefficient to be such that it is proportional to the projection of the above difference onto the null space of $Q_p^{*-\frac{1}{2}} S_p Q_p^{*-\frac{1}{2}}$.

Note that we have normalized so that

$$\|\mathbf{a}_p^B\|_{Q_p^*} = \|\mathbf{a}_p^L\|_{Q_p^*} = 1,$$

where $\|\cdot\|_{Q_p^*}$ is the norm associated with the inner product $(\mathbf{a}_p, \mathbf{b}_p)_{Q_p^*} = \mathbf{a}_p' Q_p^* \mathbf{b}_p$.

4.5 Comparison

The performances of the full Bayes discriminant and the mixed discriminant obtained by combination of Bayesian and classical criteria can be compared through investigation of the corresponding values of $E(\phi_{\mathbf{a}_p} \mid X^n)$ achieved, from the Bayes point of view, and for $Z_{\mathbf{a}_p}$, the sample analogue of $\phi_{\mathbf{a}_p}$, from the classical point of view, and also through their performance on a future observation $\mathbf{X}^0$. We first give two Lemmas needed in the investigation of the asymptotic behaviour of the statistics concerned in this section.

Lemma 4.3. Let $P_L = P(\mathcal{L}(Q_p^{*-\frac{1}{2}} S_p Q_p^{*-\frac{1}{2}}))$ be the orthogonal projection onto $\mathcal{L}(Q_p^{*-\frac{1}{2}} S_p Q_p^{*-\frac{1}{2}})$, the range of $Q_p^{*-\frac{1}{2}} S_p Q_p^{*-\frac{1}{2}}$, and $P_L^* = Q_p^{*-\frac{1}{2}} P_L Q_p^{*-\frac{1}{2}}$. The following hold:

$$P_L^* + P_0^* = Q_p^{*-1},$$
$$P_0^* Q_p^* P_0^* = P_0^*,\ P_0^* Q_p^* P_L^* = 0,\ P_L^* Q_p^* P_L^* = P_L^*.$$

Proof. The results follow from the definitions of P_L^*, P_0^* (cf. Corollary 4.2) and the property of the orthogonal projection (cf. 1c.4 in Rao, 1973, p.46). □

Lemma 4.4. Let

$$P_1 = P(\mathcal{N}(\Gamma)) = I_n - \Gamma(\Gamma'\Gamma)^{-1}\Gamma',$$
$$\gamma_p = \boldsymbol{\alpha}' m_p Q_p^{-1} m_p' \boldsymbol{\alpha},$$

and G, P_0^*, P_L^* be defined in (4.7), Corollary 4.2 and Lemma 4.3 respectively. The following hold as $p \to \infty$:

(a) $(X_p^n - \Gamma m_p) Q_p^{*-1} (X_p^n - \Gamma m_p)' \xrightarrow{P} G.$

(b) $G^{-\frac{1}{2}} (X_p^n - \Gamma m_p) Q_p^{*-1} m_p' \boldsymbol{\alpha} / \gamma_p^{\frac{1}{2}} = O_P(p^{-1}).$

(c) $\boldsymbol{\alpha}' m_p Q_p^{*-1} m_p' \boldsymbol{\alpha}/\gamma_p \xrightarrow{P} 1.$

(d) $(X_p^n - \Gamma m_p) P_L^* (X_p^n - \Gamma m_p)' \xrightarrow{P} P_1.$

(e) $(X_p^n - \Gamma m_p) P_L^* m_p' \boldsymbol{\alpha}/\gamma_p^{\frac{1}{2}} = P_1 G^{\frac{1}{2}} O_P(p^{-1}) + o_P(p^{-1}).$

(f) $\boldsymbol{\alpha}' m_p P_L^* m_p' \boldsymbol{\alpha}/\gamma_p = O_P(p^{-2}).$

Proof. Define a random matrix U by

$$U = G^{-\frac{1}{2}}(X_p^n - \Gamma m_p) Q_p^{-\frac{1}{2}}.$$

Then by (4.6), $U \sim T(\delta; I_n, I_p)$. The following hold by Lemmas 1.1, 1.3, 1.5, 1.6:

$$(I_n + UU')^{-1} \sim B(\delta + n - 1, p; I_n), \tag{4.9}$$

$$E(I_n + UU')^{-1} = \frac{\delta + n - 1}{\delta + n + p - 1} I_n = O(p^{-1}),\ (p \to \infty), \tag{4.10}$$

$$U(I_p + U'U)^{-1}U' = I_n - (I_n + UU')^{-1}$$

$$\sim B(p, \delta + n - 1; I_n) \xrightarrow{P} I_n,\ (p \to \infty), \tag{4.11}$$

$$(I_p + U'U)^{-1} \sim B(\delta + p - 1, n; I_p), \tag{4.12}$$

$$U(I_p + U'U)^{-1} = (I_n + UU')^{-1}U \tag{4.13}$$

By Lemma 4.1,

$$Q_p^* = Q_p^{\frac{1}{2}}(I_p + U'U)Q_p^{\frac{1}{2}}, \tag{4.14}$$

Hence by (4.14), (4.11),

$$\begin{aligned} & (X_p^n - \Gamma m_p) Q_p^{*-1} (X_p^n - \Gamma m_p)' \\ = \ & G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} \xrightarrow{P} G, \text{ as } p \to \infty, \end{aligned}$$

yielding (a). Since $UQ_p^{-\frac{1}{2}} m_p' \boldsymbol{\alpha}/\gamma_p^{\frac{1}{2}} \sim T(\delta; I_n, 1)$ is bounded in probability, by (4.14), (4.13), (4.10),

$$\begin{aligned} & G^{-\frac{1}{2}}(X_p^n - \Gamma m_p) Q_p^{*-1} m_p' \boldsymbol{\alpha}/\gamma_p^{\frac{1}{2}} \\ = \ & U(I_p + U'U)^{-1} Q_p^{-\frac{1}{2}} m_p' \boldsymbol{\alpha}/\gamma_p^{\frac{1}{2}} \\ = \ & (I_n + UU')^{-1} \cdot U Q_p^{-\frac{1}{2}} m_p' \boldsymbol{\alpha}/\gamma_p^{\frac{1}{2}} = O_P(p^{-1}), \end{aligned} \tag{4.15}$$

yielding (b). By (4.14), (4.12) and Lemma 1.3 (e),

$$\begin{aligned}
& \boldsymbol{\alpha}' m_p Q_p^{*-1} m_p' \boldsymbol{\alpha}/\gamma_p \\
= \; & \boldsymbol{\alpha}' m_p Q_p^{-\frac{1}{2}} (I_p + U'U)^{-1} Q_p^{-\frac{1}{2}} m_p' \boldsymbol{\alpha}/\gamma_p \\
\sim \; & B(\delta + p - 1, n; \gamma_p)/\gamma_p = B(\delta + p - 1, n; 1) \xrightarrow{P} 1.
\end{aligned}$$

This establishes (c). By Lemma 4.1,

$$S_p = Q_p^{\frac{1}{2}} U' G^{\frac{1}{2}} P_1 G^{\frac{1}{2}} U Q_p^{\frac{1}{2}}.$$

Since $\mathcal{L}(Q_p^{*-\frac{1}{2}} S_p Q_p^{*-\frac{1}{2}}) = \mathcal{L}(Q_p^{*-\frac{1}{2}} Q_p^{\frac{1}{2}} U' G^{\frac{1}{2}} P_1)$ (cf. 1b.6 in Rao, 1973 p.27),

$$\begin{aligned}
P_L^* \; = \; & Q_p^{*-1} Q_p^{\frac{1}{2}} U' G^{\frac{1}{2}} P_1 \\
& \times \; (P_1 G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} P_1)^+ P_1 G^{\frac{1}{2}} U Q_p^{\frac{1}{2}} Q_p^{*-1}. \qquad (4.16)
\end{aligned}$$

where A^+ denotes the Moore–Penrose inverse of a matrix A (cf. 1b.5 in Rao, 1973 p.26). Also, by (4.11),

$$P_1 (P_1 G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} P_1)^+ P_1 \xrightarrow{P} P_1. \qquad (4.17)$$

Hence by (4.14), (4.16), (4.11), (4.17),

$$\begin{aligned}
& (X_p^n - \Gamma m_p) P_L^* (X_p^n - \Gamma m_p)' \\
= \; & G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} P_1 (P_1 G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} P_1)^+ P_1 \\
& \times G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} \\
\xrightarrow{P} \; & G P_1 G = P_1,
\end{aligned}$$

yielding (d). By (4.14), (4.16), (4.13), (4.11), (4.17), (4.15),

$$\begin{aligned}
& (X_p^n - \Gamma m_p) P_L^* m_p' \boldsymbol{\alpha}/\gamma_p^{\frac{1}{2}} \\
= \; & G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} P_1 (P_1 G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} P_1)^+ P_1 \\
& \times G^{\frac{1}{2}} U (I_p + U'U)^{-1} Q_p^{-\frac{1}{2}} m_p' \boldsymbol{\alpha}/\gamma_p^{\frac{1}{2}} \\
= \; & G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} P_1 (P_1 G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} P_1)^+ P_1 \\
& \times G^{\frac{1}{2}} (I_n + UU')^{-1} U Q_p^{-\frac{1}{2}} m_p' \boldsymbol{\alpha}/\gamma_p^{\frac{1}{2}} \\
= \; & G^{\frac{1}{2}} (I_n + o_P(1)) G^{\frac{1}{2}} (P_1 + o_P(1)) G^{\frac{1}{2}} O_P(p^{-1}) \\
= \; & (G P_1 + o_P(1)) G^{\frac{1}{2}} O_P(p^{-1}) \\
= \; & P_1 G^{\frac{1}{2}} O_P(p^{-1}) + o_P(p^{-1}),
\end{aligned}$$

yielding (e). By (4.14), (4.16), (4.13), (4.15), (4.17),

$$\begin{aligned}
& \boldsymbol{\alpha}' m_p P_L^* m_p' \boldsymbol{\alpha}/\gamma_p \\
= \; & \boldsymbol{\alpha}' m_p Q_p^{-\frac{1}{2}} (I_p + U'U)^{-1} U' G^{\frac{1}{2}} P_1 (P_1 G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} P_1)^+ P_1 \\
& \times G^{\frac{1}{2}} U (I_p + U'U)^{-1} Q_p^{-\frac{1}{2}} m_p' \boldsymbol{\alpha}/\gamma_p \\
= \; & \boldsymbol{\alpha}' m_p Q_p^{-\frac{1}{2}} U' (I_n + UU')^{-1} G^{\frac{1}{2}} P_1 (P_1 G^{\frac{1}{2}} U (I_p + U'U)^{-1} U' G^{\frac{1}{2}} P_1)^+ P_1 \\
& \times G^{\frac{1}{2}} (I_n + UU')^{-1} U Q_p^{-\frac{1}{2}} m_p' \boldsymbol{\alpha}/\gamma_p \\
= \; & O_P(p^{-2}),
\end{aligned}$$

yielding (f). □

4.5.1 Posterior Performance

Theorem 4.1. Under the assumptions (4.2)-(4.5)

$$\frac{\max_{\mathbf{a}_p : S_p \mathbf{a}_p = 0} E(\phi_{\mathbf{a}_p} \mid X^n)}{\max_{\mathbf{a}_p} E(\phi_{\mathbf{a}_p} \mid X^n)} \xrightarrow{P} 1, \text{ as } p \to \infty. \tag{4.18}$$

Proof. Let $\Delta_p(X^n) = \gamma_p(X^n) - \gamma_p^L(X^n)$. The left side of (4.18) is, by Corollaries 4.1 and 4.2,

$$1 - \frac{\delta^* \Delta_p(X^n)}{\boldsymbol{\alpha}' H^* \boldsymbol{\alpha} + \delta^* \gamma_p(X^n)} = 1 - \frac{\delta^* \Delta_p(X^n)/\gamma_p}{(\boldsymbol{\alpha}' H^* \boldsymbol{\alpha} + \delta^* \gamma_p(X^n))/\gamma_p}.$$

Since γ_p is nondecreasing, it tends to a limit $\gamma_\infty \in [0, \infty]$. It was shown in Dawid and Fang (1992) that

$$\gamma_p(X^n) \xrightarrow{a.s.} c + \gamma_\infty, \text{ as } p \to \infty, \text{ if } \gamma_\infty < \infty. \tag{4.19}$$

where

$$c = n_1 h_1^{-1} (n_1 + h_1)^{-1} + n_2 h_2^{-1} (n_2 + h_2)^{-1}, \tag{4.20}$$

(also proved below for convergence in probability). Now suppose $\mathbf{m}_1 \neq \mathbf{m}_2$, then $\gamma_\infty > 0$. We shall show that

$$\frac{\gamma_p(X^n)}{\gamma_p} \xrightarrow{P} 1, \text{ as } p \to \infty, \text{ if } \gamma_\infty = \infty, \tag{4.21}$$

and

$$\frac{\Delta_p(X_p^n)}{\gamma_p} \xrightarrow{P} 0. \tag{4.22}$$

Then (4.19), (4.21), (4.22) lead to the desired conclusion. By Corollary 4.1 and Lemma 4.1,

$$\begin{aligned}
\gamma_p(X^n) &= \boldsymbol{\alpha}' m_p^* Q_p^{*^{-1}} (m_p^*)' \boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}' [H^* \Gamma' (X_p^n - \Gamma m_p) + m_p] Q_p^{*^{-1}} [(X_p^n - \Gamma m_p)' \Gamma H^* + m_p'] \boldsymbol{\alpha} \\
&= J_1 + J_2 + J_3, \text{ say,}
\end{aligned}$$

where

$$\begin{aligned}
J_1 &= \boldsymbol{\alpha}' H^* \Gamma' (X_p^n - \Gamma m_p) Q_p^{*^{-1}} (X_p^n - \Gamma m_p)' \Gamma H^* \boldsymbol{\alpha}, \\
J_2 &= 2\boldsymbol{\alpha}' H^* \Gamma' (X_p^n - \Gamma m_p) Q_p^{*^{-1}} m_p' \boldsymbol{\alpha}, \\
J_3 &= \boldsymbol{\alpha}' m_p Q_p^{*^{-1}} m_p' \boldsymbol{\alpha}.
\end{aligned}$$

By the definitions of G and H^* (cf. (4.7) and Lemma 4.1),

$$\begin{aligned}
& H^* \Gamma' G \Gamma H^* = H^* \Gamma' (I_n + \Gamma H \Gamma') \Gamma H^* \\
= \; & H^* (H^{-1} + \Gamma' \Gamma) H \Gamma' \Gamma H^* = H \Gamma' \Gamma H^* \\
= \; & H (\Gamma' \Gamma + H^{-1} - H^{-1}) H^* = H (I - H^{-1} H^*) = H - H^*.
\end{aligned}$$

Hence by Lemma 4.4 (a), (b), (c), as $p \to \infty$,

$$\begin{aligned}
J_1 &\xrightarrow{P} \boldsymbol{\alpha}' H^* \Gamma' G \Gamma H^* \boldsymbol{\alpha} = \boldsymbol{\alpha}' (H - H^*) \boldsymbol{\alpha} = c, \\
J_2 &= 2\boldsymbol{\alpha}' H^* \Gamma' G^{\frac{1}{2}} O_P(\gamma_p^{\frac{1}{2}} p^{-1}), \\
\frac{J_3}{\gamma_p} &\xrightarrow{P} 1.
\end{aligned}$$

Since

$$\gamma_p(X_p^n) = J_1 + J_2 + J_3,$$

(4.21) follows, as does the "in probability" version of (4.19).

By Corollary 4.1, Corollary 4.2, Lemma 4.3, Lemma 4.1,

$$\begin{aligned}
\Delta_p(X^n) &= \boldsymbol{\alpha}' m_p^* P_L^* m_p^{*'} \boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}' [H^* \Gamma' (X_p^n - \Gamma m_p) + m_p] P_L^* [(X_p^n - \Gamma m_p)' \Gamma H^* + m_p'] \boldsymbol{\alpha} \\
&= Z_1 + Z_2 + Z_3, \text{ say,}
\end{aligned}$$

where

$$
\begin{aligned}
Z_1 &= \boldsymbol{\alpha}' H^* \Gamma' (X_p^n - \Gamma m_p) P_L^* (X_p^n - \Gamma m_p)' \Gamma H^* \boldsymbol{\alpha}, \\
Z_2 &= 2\boldsymbol{\alpha}' H^* \Gamma' (X_p^n - \Gamma m_p) P_L^* m_p' \boldsymbol{\alpha}, \\
Z_3 &= \boldsymbol{\alpha}' m_p P_L^* m_p' \boldsymbol{\alpha}.
\end{aligned}
$$

By (d), (e), (f) of Lemma 4.4, as $p \to \infty$,

$$
\begin{aligned}
Z_1 &\xrightarrow{P} \boldsymbol{\alpha}' H^* \Gamma' P_1 \Gamma H^* \boldsymbol{\alpha} = 0, \\
Z_2 &= 2\boldsymbol{\alpha}' H^* \Gamma' (P_1 G^{\frac{1}{2}} O_P(p^{-1}) + o_P(p^{-1})) \gamma_p^{\frac{1}{2}}, \\
Z_3 &= O_P(\gamma_p p^{-2}).
\end{aligned}
$$

Thus

$$
\frac{\Delta_p(X_p^n)}{\gamma_p} = \frac{Z_1 + Z_2 + Z_3}{\gamma_p} \xrightarrow{P} 0, \text{ as } p \to \infty,
$$

establishing (4.22). If $\mathbf{m}_1 = \mathbf{m}_2$, from the proof of Lemma 4.4 and the proof above $\Delta_p(X_p^n) \xrightarrow{P} 0$, (4.19) remains true. Hence (4.18) holds. This completes the proof. □

Note by (4.19), (4.21), (4.22),

$$
\frac{\Delta_p(X^n)}{\gamma_p(X^n)} \xrightarrow{P} 0, \ \frac{\gamma_p^L(X_p^n)}{\gamma_p(X_p^n)} \xrightarrow{P} 1, \text{ as } p \to \infty. \tag{4.23}
$$

Since $|Z_2/\gamma_p| \leq (Z_1 + Z_3)/\gamma_p$, we have that (4.18) is $1 - o_P(\gamma_p^{-1})$, if $\gamma_p = o(p^2)$ (in particular, if $\gamma_\infty < \infty$), $1 - O_P(p^{-2})$, otherwise. In the former case, moreover, $Z_3 \xrightarrow{P} 0$, so that $\max_{\mathbf{a}_p} E(\phi_{\mathbf{a}_p} \mid X^n) - \max_{\mathbf{a}_p : S_p \mathbf{a}_p = 0} E(\phi_{\mathbf{a}_p} \mid X^n) \xrightarrow{P} 0$.

4.5.2 Discriminant Functions

Now suppose a new observation $\mathbf{X}_p^0$ is obtained and we wish to classify from which population it arises. We may use a discriminant function with coefficient $\mathbf{a}_p$:

$$W_{\mathbf{a}_p}(\mathbf{X}_p^0; m_p^*, Q_p^*) \stackrel{\text{def}}{=} 2\mathbf{a}_p'(\mathbf{X}_p^0 - \frac{\mathbf{m}_{1p}^* + \mathbf{m}_{2p}^*}{2})/\gamma_p(X^n)^{\frac{1}{2}}. \tag{4.24}$$

Since $W_{\mathbf{a}_p}(\mathbf{m}_{1p}^*, m_p^*, Q_p^*) > 0$, $W_{\mathbf{a}_p}(\mathbf{m}_{2p}^*, m_p^*, Q_p^*) < 0$ if $\mathbf{a}_p = \mathbf{a}_p^B$ or $\mathbf{a}_p^L$, we use the corresponding sign to classify $\mathbf{X}_p^0$. Note that the value of $\phi_{\mathbf{a}_p}$, $Z_{\mathbf{a}_p}$ will not be affected by any scalar by which $\mathbf{a}_p$ is multiplied. We shall show that the two coefficients $\mathbf{a}_p^B$ and $\mathbf{a}_p^L$ lead to the same asymptotic behaviour for (4.24).

Theorem 4.2. Under the assumptions (4.2)–(4.5), as $p \to \infty$,

(a) If $\mathbf{X}_p^0$ arises from Π_1, then

$$W_{\mathbf{a}_p^B}(\mathbf{X}_p^0; m_p^*, Q_p^*)\Big\{ \begin{array}{ll} \xrightarrow{P} 1 & \text{if } \gamma_\infty = \infty, \\ \xrightarrow{\mathcal{L}} 1 + 2(c+\gamma_\infty)^{-\frac{1}{2}} \cdot T(\delta^*; 1, k_1^*) & \text{if } \gamma_\infty < \infty. \end{array}$$

(b) If $\mathbf{X}_p^0$ arises from Π_2, then

$$W_{\mathbf{a}_p^B}(\mathbf{X}_p^0; m_p^*, Q_p^*)\Big\{ \begin{array}{ll} \xrightarrow{P} -1 & \text{if } \gamma_\infty = \infty, \\ \xrightarrow{P} -1 + 2(c+\gamma_\infty)^{-\frac{1}{2}} \cdot T(\delta^*; 1, k_2^*) & \text{if } \gamma_\infty < \infty. \end{array}$$

In the above, $k_i^* = 1 + h_i^{*-1}$, and c is given by (4.20).

(c)

$$W_{\mathbf{a}_p^B}(\mathbf{X}_p^0; m_p^*, Q_p^*) - W_{\mathbf{a}_p^L}(\mathbf{X}_p^0; m_p^*, Q_p^*) \xrightarrow{P} 0. \tag{4.25}$$

Proof. If $\mathbf{X}_p^0$ arises from Π_1, then by (5.12) of Dawid and Fang (1992)

$$\mathbf{X}_p^0 \mid X^n \sim \mathbf{m}_{1p}^* + T(\delta^*; Q_p^*, k_1^*).$$

Hence

$$W_{\mathbf{a}_p^B}(\mathbf{X}_p^0; m_p^*, Q_p^*) \mid X^n \sim 1 + T(\delta^*; 1, k_1^*) \cdot 2\gamma_p(X^n)^{-\frac{1}{2}},$$

which combined with (4.19), (4.21) establishes (a). Then (b) can be proved similarly.

For the proof of (c), without loss of generality we suppose $\mathbf{X}_p^0 \sim \Pi_1$. By Lemma 4.3,

$$\frac{Q_p^{*-1}}{\gamma_p(X_p^n)^{\frac{1}{2}}} - \frac{P_0^*}{\gamma_p^L(X_p^n)^{\frac{1}{2}}} = \frac{P_L^* + P_0^*(1-\gamma_p(X^n)^{\frac{1}{2}}/\gamma_p^L(X^n)^{\frac{1}{2}})}{\gamma_p(X^n)^{\frac{1}{2}}}. \tag{4.26}$$

Hence by Corollary 4.1, Corollary 4.2,

$$\begin{aligned} & (\mathbf{a}_p^B - \mathbf{a}_p^L)' m_p^{*'} \boldsymbol{\alpha} \\ = \; & \boldsymbol{\alpha}' m_p^* \left(\frac{Q_p^{*-1}}{\gamma_p(X_p^n)^{\frac{1}{2}}} - \frac{P_0^*}{\gamma_p^L(X_p^n)^{\frac{1}{2}}} \right) m_p^{*'} \boldsymbol{\alpha} \\ = \; & \frac{\Delta_p(X^n) + \gamma_p^L(X^n)(1-\gamma_p(X^n)^{\frac{1}{2}}/\gamma_p^L(X^n)^{\frac{1}{2}})}{\gamma_p(X^n)^{\frac{1}{2}}}, \end{aligned} \tag{4.27}$$

which divided by $\gamma_p(X^n)^{\frac{1}{2}}$ converges to 0 in probability by (4.23). Similarly, by (4.26),

$$\begin{aligned} & (\mathbf{a}_p^B - \mathbf{a}_p^L)' Q_p^* (\mathbf{a}_p^B - \mathbf{a}_p^L) \\ = \; & \boldsymbol{\alpha}' m_p^* \left(\frac{Q_p^{*-1}}{\gamma_p(X^n)^{\frac{1}{2}}} - \frac{P_0^*}{\gamma_p^L(X^n)^{\frac{1}{2}}} \right) Q_p^* \left(\frac{Q_p^{*-1}}{\gamma_p(X^n)^{\frac{1}{2}}} - \frac{P_0^*}{\gamma_p^L(X^n)^{\frac{1}{2}}} \right) (m_p^*)' \boldsymbol{\alpha} \\ = \; & \frac{\boldsymbol{\alpha}' m_p^* P_L^* m_p^{*'} \boldsymbol{\alpha}}{\gamma_p(X_p^n)} + \frac{\boldsymbol{\alpha}' m_p^* P_0^* m_p^{*'} \boldsymbol{\alpha}}{\gamma_p(X_p^n)} \cdot \left(1 - \frac{\gamma_p(X^n)^{\frac{1}{2}}}{\gamma_p^L(X^n)^{\frac{1}{2}}} \right)^2 \\ = \; & \frac{\Delta_p(X^n)}{\gamma_p(X^n)} + \frac{\gamma_p^L(X^n)}{\gamma_p(X^n)} \left(1 - \frac{\gamma_p(X^n)^{\frac{1}{2}}}{\gamma_p^L(X^n)^{\frac{1}{2}}} \right)^2 \xrightarrow{P} 0. \end{aligned} \tag{4.28}$$

Then the left hand side of (4.25), given X^n, is distributed as

$$\frac{(\mathbf{a}_p^B - \mathbf{a}_p^L)' m_p^{*'} \boldsymbol{\alpha}}{\gamma_p(X^n)^{\frac{1}{2}}} + T(\delta^*; (\mathbf{a}_p^B - \mathbf{a}_p^L)' Q_p^* (\mathbf{a}_p^B - \mathbf{a}_p^L), k_1^*) \cdot 2\gamma_p(X^n)^{-\frac{1}{2}}$$

and thus tends to zero in probability, both conditionally on X^n and unconditionally. □

From (4.28) we know that

$$(\mathbf{a}_p^B)' Q_p^* \mathbf{a}_p^L \xrightarrow{P} \frac{1}{2}((\mathbf{a}_p^B)' Q_p^* \mathbf{a}_p^B + (\mathbf{a}_p^L)' Q_p^* \mathbf{a}_p^L) = 1.$$

Thus the angle between the unit vectors $Q_p^{*\frac{1}{2}} \mathbf{a}_p^B$ and $Q_p^{*\frac{1}{2}} \mathbf{a}_p^L$ tends to 0, so that the two coefficients are asymptotically identical.

In Section 4 in order to maximize $Z_{\mathbf{a}_p}$ we let the denominator $\mathbf{a}_p' S_p \mathbf{a}_p$ be zero and obtained the mixture solution $\mathbf{a}_p^L$. But we also require that the numerator of $Z_{\mathbf{a}_p}$ not be zero. For any finite p, it is positive with probability one. We shall show that its limit is also positive.

Theorem 4.3. Under the assumptions (4.2)–(4.5),

$$[(\mathbf{a}_p^B)'(\overline{X}_{1p} - \overline{X}_{2p})]^2 \xrightarrow{P} \begin{cases} \frac{(\boldsymbol{\alpha}' H \boldsymbol{\alpha} + \gamma_\infty)^2}{c + \gamma_\infty} & \text{if } \gamma_\infty < \infty, \\ \infty & \text{if } \gamma_\infty = \infty, \end{cases} \quad p \to \infty, \tag{4.29}$$

$$[(\mathbf{a}_p^L)'(\overline{X}_{1p} - \overline{X}_{2p})]^2 \xrightarrow{P} \begin{cases} \frac{(\boldsymbol{\alpha}' H \boldsymbol{\alpha} + \gamma_\infty)^2}{c + \gamma_\infty} & \text{if } \gamma_\infty < \infty, \\ \infty & \text{if } \gamma_\infty = \infty, \end{cases} \quad p \to \infty, \tag{4.30}$$

where c is given by (4.20).

Proof. By Corollary 4.1 and Lemma 4.1,

$$(\mathbf{a}_p^B)'(\overline{X}_{1p} - \overline{X}_{2p}) = \boldsymbol{\alpha}' m_p^* Q_p^{*-1} (\overline{\overline{X}}_p)' \boldsymbol{\alpha} / \gamma_p(X^n)^{\frac{1}{2}}. \tag{4.31}$$

To prove (4.29), write

$$\begin{aligned} & \boldsymbol{\alpha}' m_p^* Q_p^{*-1} (\overline{\overline{X}}_p)' \boldsymbol{\alpha} \\ = \; & \boldsymbol{\alpha}' [H^* \Gamma' (X_p^n - \Gamma m_p) + m_p] Q_p^{*-1} [(X_p^n - \Gamma m_p)' \Gamma (\Gamma' \Gamma)^{-1} + m_p'] \boldsymbol{\alpha} \\ = \; & J_1 + J_2 + J_3, \text{ say}, \end{aligned} \tag{4.32}$$

where

$$\begin{aligned} J_1 &= \boldsymbol{\alpha}' H^* \Gamma' (X_p^n - \Gamma m_p) Q_p^{*-1} (X_p^n - \Gamma m_p)' \Gamma (\Gamma' \Gamma)^{-1} \boldsymbol{\alpha}, \\ J_2 &= \boldsymbol{\alpha}' [H^* + (\Gamma' \Gamma)^{-1}] \Gamma' (X_p^n - \Gamma m_p) Q_p^{*-1} m_p' \boldsymbol{\alpha}, \\ J_3 &= \boldsymbol{\alpha}' m_p Q_p^{*-1} m_p' \boldsymbol{\alpha}. \end{aligned}$$

By Lemma 4.4 (a), (b) and (c), as $p \to \infty$,

$$\begin{aligned} J_1 &\xrightarrow{P} \boldsymbol{\alpha}' H^* \Gamma' G \Gamma (\Gamma' \Gamma)^{-1} \boldsymbol{\alpha} = \boldsymbol{\alpha}' H \boldsymbol{\alpha}, \\ \frac{J_2}{\gamma_p^{\frac{1}{2}}} &= \boldsymbol{\alpha}' [H^* + (\Gamma' \Gamma)^{-1}] \Gamma' G^{\frac{1}{2}} O_P(p^{-1}) \xrightarrow{P} 0, \\ \frac{J_3}{\gamma_p} &\xrightarrow{P} 1. \end{aligned}$$

If $\gamma_\infty < \infty$, by (4.19), $J_2/\gamma_p^{\frac{1}{2}} \cdot (\gamma_p/\gamma_p(X^n))^{\frac{1}{2}} \xrightarrow{P} 0$. Hence (4.31) is equal to

$$\frac{J_1 + J_3}{\gamma_p(X^n)^{\frac{1}{2}}} + \frac{J_2}{\gamma_p^{\frac{1}{2}}} \cdot \left(\frac{\gamma_p}{\gamma(X^n)}\right)^{\frac{1}{2}} \xrightarrow{P} \frac{\boldsymbol{\alpha}' H \boldsymbol{\alpha} + \gamma_\infty}{(c + \gamma_\infty)^{\frac{1}{2}}}, \text{ as } p \to \infty.$$

If $\gamma_\infty = \infty$, $(J_1 + J_2)/\gamma_p \xrightarrow{P} 0$. Hence (4.31) is equal to

$$\left(\frac{J_3}{\gamma_p} + \frac{J_1 + J_2}{\gamma_p}\right) \cdot \left(\frac{\gamma_p}{\gamma_p(X^n)}\right)^{\frac{1}{2}} \cdot \gamma_p^{\frac{1}{2}} \xrightarrow{P} \infty, \text{ as } p \to \infty.$$

This completes the proof of (4.29).

By Corollary 4.2, Lemma 4.1,

$$(\mathbf{a}_p^L)'(\overline{X}_{1p} - \overline{X}_{2p}) = \boldsymbol{\alpha}' m_p^* P_0^* (\overline{\overline{X}}_p)' \boldsymbol{\alpha} / \gamma_p^L (X^n)^{\frac{1}{2}} \tag{4.33}$$

To prove (4.30), by Lemma 4.3, write

$$\boldsymbol{\alpha}' m_p^* P_0^* (\overline{\overline{X}}_p)' \boldsymbol{\alpha} = \boldsymbol{\alpha}' m_p^* Q^{*^{-1}} (\overline{\overline{X}}_p)' \boldsymbol{\alpha} - \boldsymbol{\alpha}' m_p^* P_L^* (\overline{\overline{X}}_p)' \boldsymbol{\alpha}. \tag{4.34}$$

Now

$$\boldsymbol{\alpha}' m_p^* P_L^* (\overline{\overline{X}}_p)' \boldsymbol{\alpha} = Z_1 + Z_2 + Z_3, \tag{4.35}$$

where

$$\begin{aligned} Z_1 &= \boldsymbol{\alpha}' H^* \Gamma' (X_p^n - \Gamma m_p) P_L^* (X_p^n - \Gamma m_p)' \Gamma (\Gamma'\Gamma)^{-1} \boldsymbol{\alpha}, \\ Z_2 &= \boldsymbol{\alpha}' [H^* + (\Gamma'\Gamma)^{-1}] \Gamma' (X_p^n - \Gamma m_p) P_L^* m_p' \boldsymbol{\alpha}, \\ Z_3 &= \boldsymbol{\alpha}' m_p P_L^* m_p' \boldsymbol{\alpha}. \end{aligned}$$

By Lemma 4.4 (d), (e), (f), as $p \to \infty$,

$$\begin{aligned} Z_1 &\xrightarrow{P} \boldsymbol{\alpha}' H^* \Gamma' P_1 \Gamma (\Gamma'\Gamma)^{-1} \boldsymbol{\alpha} = 0, \\ Z_2 &= \boldsymbol{\alpha}' [H^* + (\Gamma'\Gamma)^{-1}] \Gamma' (P_1 G^{\frac{1}{2}} O_P(\gamma_p^{\frac{1}{2}} p^{-1}) + o_P(\gamma_p^{\frac{1}{2}} p^{-1})) \\ &= o_p(\gamma_p^{\frac{1}{2}} p^{-1}), \text{ since } \Gamma' P_1 = 0, \\ Z_3 &= O_p(\gamma_p p^{-2}). \end{aligned}$$

If $\gamma_\infty < \infty$, by (4.31), (4.34), (4.35), we can write (4.33) as

$$(\mathbf{a}_p^B)'(\overline{X}_{1p} - \overline{X}_{2p}) \cdot \frac{\gamma_p(X^n)^{\frac{1}{2}}}{\gamma_p^L(X^n)^{\frac{1}{2}}} - \frac{Z_1 + Z_2 + Z_3}{\gamma_p^L(X^n)^{\frac{1}{2}}},$$

which combined with (4.29) establishes the first part of (4.30). If $\gamma_\infty = \infty$, since

$$\frac{J_1 + J_2 - (Z_1 + Z_2 + Z_3)}{\gamma_p} \xrightarrow{P} 0,$$

by (4.32), (4.34), (4.35), we can write (4.33) as

$$\left(\frac{J_3}{\gamma_p} + \frac{J_1 + J_2 - (Z_1 + Z_2 + Z_3)}{\gamma_p}\right) \cdot \left(\frac{\gamma_p}{\gamma_p^L(X^n)}\right)^{\frac{1}{2}} \cdot \gamma_p^{\frac{1}{2}},$$

yielding the second part of (4.30). □

We are also interested to know the performance of $\mathbf{a}_p^B$ from the classical point of view, i.e. its $Z_{\mathbf{a}_p^B}$. From Theorem 4.3 we know that its numerator (4.29) has the same limit as that of $Z_{\mathbf{a}_p^L}$, (4.30). We shall show that its denominator tends to zero in probability, so that $Z_{\mathbf{a}_p^B} \to \infty$.

Theorem 4.4. Under the assumptions (4.2)–(4.5),

$$(\mathbf{a}_p^B)' S_p \mathbf{a}_p^B \xrightarrow{P} 0, \text{ as } p \to \infty.$$

Proof. By Lemma 4.1 we have

$$(\mathbf{a}_p^B)' S_p \mathbf{a}_p^B = \boldsymbol{\alpha}' m_p^* Q_p^{*^{-1}} (X_p^n - \Gamma m_p)' P_1 (X^n - \Gamma m_p) Q_p^{*^{-1}} m_p^{*'} \boldsymbol{\alpha} / \gamma_p(X^n).$$

Now

$$\begin{aligned} & \boldsymbol{\alpha}' m_p^* Q_p^{*^{-1}} (X_p^n - \Gamma m_p)' P_1 \\ = \; & \boldsymbol{\alpha}' H^* \Gamma' (X_p^n - \Gamma m_p) Q_p^{*^{-1}} (X_p^n - \Gamma m_p)' P_1 + \boldsymbol{\alpha}' m_p Q_p^{*^{-1}} (X_p^n - \Gamma m_p)' P_1. \end{aligned}$$

By Lemma 4.4 (a), (b), the first term tends to

$$\boldsymbol{\alpha}' H^* \Gamma' G P_1 = \boldsymbol{\alpha}' H^* \Gamma' P_1 = 0$$

in probability. The second term is

$$\gamma_p^{\frac{1}{2}} G^{\frac{1}{2}} P_1 O_p(p^{-1}).$$

Hence by (4.19), (4.21),

$$\begin{aligned} & \boldsymbol{\alpha}' m_p^* Q_p^{*^{-1}} (X_p^n - \Gamma m_p)' P_1 / \gamma_p(X^n)^{\frac{1}{2}} \\ = \; & \frac{\boldsymbol{\alpha}' m_p^* Q_p^{*^{-1}} (X_p^n - \Gamma m_p)' P_1}{\gamma_p^{\frac{1}{2}}} \cdot \left(\frac{\gamma_p}{\gamma_p(X^n)} \right)^{\frac{1}{2}} \\ \xrightarrow{P} \; & 0, \text{ as } p \to \infty, \end{aligned}$$

completing the proof. □

4.6 Discussion

We have shown that, the conjugate prior implies a determinism that, if $\gamma_\infty = \infty$, the predictive odds of Π_1 to Π_2 are asymptotically degenerate (cf. Chapter 2); the posterior performance of the Bayes estimator $\mathbf{a}_p^B$, $E(\phi_{\mathbf{a}_p^B} \mid X^n)$, tends to infinity (cf. Corollary 4.1 and (4.21) in the proof of Theorem 4.1); and for a future observation, the probability of correct discrimination of $\mathbf{a}_p^B$ tends to one (cf. Theorem 4.2). To consider the possible selection bias in estimating the optimised value of the parameter $\phi_{\mathbf{a}_p}$, defined in (4.1), let

$$\phi^{**} = \sup_{\mathbf{a}_p} \{\phi_{\mathbf{a}_p}\} = (\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p})' \Sigma_p^{-1} (\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p}) = \phi_{a^{**}},$$

with

$$\mathbf{a}_p^{**} = \Sigma_p^{-1} (\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p}).$$

$Z_{\mathbf{a}_p}$ (defined in Section 4.4.2) can be used as the estimate for $\phi_{\mathbf{a}_p}$. By (4.2) and Lemma 4.1,

$$\overline{X}_{1p} - \overline{X}_{2p} = (X_p^n)' \Gamma (\Gamma' \Gamma)^{-1} \boldsymbol{\alpha} \sim \mu_p' \boldsymbol{\alpha} + \mathcal{N}(\Sigma_p, \boldsymbol{\alpha}' (\Gamma' \Gamma)^{-1} \boldsymbol{\alpha}) \tag{4.36}$$

and

$$S_p = (X_p^n)' P_1 X_p^n \tag{4.37}$$

is a noncentral Wishart distribution with covariance matrix Σ_p, $\text{rank}(P_1) = \text{tr}(P_1) = n-2$ degrees of freedom, and a matrix of noncentality parameters $(\Gamma\mu)'P_1(\Gamma\mu) = 0$, i.e.

$$S_p \sim W_p(n-2; \Sigma_p). \tag{4.38}$$

Now suppose $n - 2 \geq p$. By the independence of $\overline{\overline{X}}_p$ and S_p, we have

$$\begin{aligned} E(Z_{\mathbf{a}_p}) &= E\{[\mathbf{a}_p'(\overline{X}_{1p} - \overline{X}_{2p})]^2\}E[(\mathbf{a}_p'S_p\mathbf{a}_p)^{-1}] \\ &= [(\boldsymbol{\alpha}'\mu_p\mathbf{a}_p)^2 + (\boldsymbol{\alpha}'(\Gamma'\Gamma)^{-1}\boldsymbol{\alpha})(\mathbf{a}_p'\Sigma_p\mathbf{a}_p)](\mathbf{a}_p'\Sigma_p\mathbf{a}_p)^{-1}/(n-4) \\ &= (\phi_{\mathbf{a}_p} + n_1^{-1} + n_2^{-1})/(n-4). \end{aligned}$$

Let

$$X_{\mathbf{a}_p} = (n-4)Z_{\mathbf{a}_p} - (n_1^{-1} + n_2^{-1}).$$

Then $X_{\mathbf{a}_p}$ is an unbiased estimate of $\phi_{\mathbf{a}_p}$.

$$Z^* = \sup_{\mathbf{a}_p}\{Z_{\mathbf{a}_p}\} = (\overline{X}_{1p} - \overline{X}_{2p})'S_p^{-1}(\overline{X}_{1p} - \overline{X}_{2p})$$

and

$$X^* = \sup_{\mathbf{a}_p}\{X_{\mathbf{a}_p}\} = (n-4)Z^* - (n_1^{-1} + n_2^{-1}) = X_{\mathbf{a}_p^*},$$

where $\mathbf{a}_p^* = S_p^{-1}(\overline{X}_{1p} - \overline{X}_{2p})$. By (4.36) and (4.37),

$$(n_1^{-1} + n_2^{-1})^{-1}(\overline{X}_{1p} - \overline{X}_{2p})'\Sigma_p^{-1}(\overline{X}_{1p} - \overline{X}_{2p}) \sim \chi_p^2(b),$$

the noncentral χ^2 distribution with noncentrality parameter

$$b = (\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p})'\Sigma_p^{-1}(\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p})/(n_1^{-1} + n_2^{-1}) = \phi^{**}/(n_1^{-1} + n_2^{-1}),$$

$$\frac{(\overline{X}_{1p} - \overline{X}_{2p})'\Sigma_p^{-1}(\overline{X}_{1p} - \overline{X}_{2p})}{(\overline{X}_{1p} - \overline{X}_{2p})'S_p^{-1}(\overline{X}_{1p} - \overline{X}_{2p})} \sim \chi^2_{n-2-p+1},$$

independently. Thus

$$\begin{aligned} & (\overline{X}_{1p} - \overline{X}_{2p})'S_p^{-1}(\overline{X}_{1p} - \overline{X}_{2p}) \\ = \; & (\overline{X}_{1p} - \overline{X}_{2p})'\Sigma_p^{-1}(\overline{X}_{1p} - \overline{X}_{2p})\left(\frac{(\overline{X}_{1p} - \overline{X}_{2p})'\Sigma_p^{-1}(\overline{X}_{1p} - \overline{X}_{2p})}{(\overline{X}_{1p} - \overline{X}_{2p})'S_p^{-1}(\overline{X}_{1p} - \overline{X}_{2p})}\right)^{-1} \\ \sim \; & \frac{(n_1^{-1} + n_2^{-1})p}{n-p-1}F_{p,n-p-1}(b), \end{aligned}$$

the univariate noncentral F distribution (cf.Muirhead 1982 p.216). We have

$$\begin{aligned} E(Z^*) &= \frac{(n_1^{-1}+n_2^{-1})(p+b)}{n-p-1-2} = \frac{(n_1^{-1}+n_2^{-1})p+\phi^{**}}{n-p-3}, \\ E(X^*) &= (n-4)E(Z^*) - (n_1^{-1}+n_2^{-1}) \\ &= \frac{(n-4)\phi^{**} + (n_1^{-1}+n_2^{-1})(p-1)(n-3)}{n-p-3}. \end{aligned}$$

Let $\phi^+ = \phi_{\Lambda^+}$, $\phi^* = \phi_{\Lambda^*}$ with $\Lambda^+ = \mathbf{a}_p^B$, $\Lambda^* = \mathbf{a}_p^L$. If $p > 1$, $E(X^*) > \phi^{**} (\geq \phi^*)$, X^* is positively biased.

If $n - 2 < p$, we can find $\mathbf{a}_p^*$, (e.g. $\mathbf{a}_p^* = \mathbf{a}_p^L$), such that $X^* = \infty$, which fits the data exactly. In the Bayesian approach, we have

$$\begin{aligned} Y_{\mathbf{a}_p} &= E(\phi_{\mathbf{a}_p} \mid X_p^n) \text{ (cf. Proposition 4.1)}, \\ Y^+ &= \sup\{Y_{\mathbf{a}_p}\} = \boldsymbol{\alpha}' H^* \boldsymbol{\alpha} + \delta^* \gamma_p(X^n) \text{ with } \Lambda^+ = \mathbf{a}_p^B \text{(cf. Corollary 4.1)}, \\ Y^* &= E(\phi_{\mathbf{a}_p^L} \mid X^n) = \boldsymbol{\alpha}' H^* \boldsymbol{\alpha} + \delta^* \gamma_p^L(X^n) \text{ (cf. Corollary 4.2)}. \end{aligned}$$

By Lemma 4.1,

$$(\boldsymbol{\alpha}' H^* \boldsymbol{\alpha} \Sigma_p)^{-\frac{1}{2}} \mu_p' \boldsymbol{\alpha} \mid X_p^n, \Sigma_p \sim (\boldsymbol{\alpha}' H^* \boldsymbol{\alpha} \Sigma_p)^{-\frac{1}{2}} (m_p^*)' \boldsymbol{\alpha} + \mathcal{N}(I_p, 1).$$

Hence

$$(\boldsymbol{\alpha}' H^* \boldsymbol{\alpha})^{-1} \boldsymbol{\alpha}' \mu_p \Sigma_p^{-1} \mu_p' \boldsymbol{\alpha} \mid X_p^n, \Sigma_p \sim \chi_p^2(c),$$

the univariate noncentral χ^2 distibution, with the noncentrality parameter

$$c = (\boldsymbol{\alpha}' H^* \boldsymbol{\alpha})^{-1} \boldsymbol{\alpha}' m_p^* \Sigma_p^{-1} (m_p^*)' \boldsymbol{\alpha},$$

Also

$$\Sigma_p^{-1} \mid X_p^n \sim W(\delta^* + p - 1; (Q_p^*)^{-1}).$$

Hence

$$\begin{aligned} & E(\boldsymbol{\alpha}' \mu_p \Sigma_p^{-1} \mu_p' \boldsymbol{\alpha} \mid X_p^n) \\ = \; & E(E(\boldsymbol{\alpha}' \mu_p \Sigma_p \mu_p' \boldsymbol{\alpha} \mid X_p^n, \Sigma_p) \mid X_p^n) \\ = \; & E(\boldsymbol{\alpha}' H^* \boldsymbol{\alpha}(p+c) \mid X_p^n) \\ = \; & p \boldsymbol{\alpha}' H^* \boldsymbol{\alpha} + \boldsymbol{\alpha}' m_p^* E(\Sigma_p^{-1} \mid X_p^n)(m_p^*)' \boldsymbol{\alpha} \\ = \; & p((h_1^*)^{-1} + (h_2^*)^{-1}) + (\delta^* + p - 1)(\mathbf{m}_{1p}^* - \mathbf{m}_{2p}^*)' Q_p^{*-1} (\mathbf{m}_{1p}^* - \mathbf{m}_{2p}^*). \end{aligned}$$

This shows that

$$Y^{**} = E(\phi^{**} \mid X_p^n) = p((h_1^*)^{-1}+(h_2^*)^{-1})+(\delta^*+p-1)(\mathbf{m}_{1p}^*-\mathbf{m}_{2p}^*)'Q_p^{*^{-1}}(\mathbf{m}_{1p}^*-\mathbf{m}_{2p}^*).$$

If $p > 1$, $Y^{**} > Y^+ \geq Y^*$. Y^{**}, Y^+, Y^* can be used as the Bayes estimates for ϕ^{**}, ϕ^+, ϕ^*. For any finite p, these estimates are finite and thus adjust $X^* = X_{\mathbf{a}_p^L} = \infty$ in the right direction. However, if $p \to \infty$, we have shown that Y^+ converges a.s. to a limit, which is finite if $\gamma_\infty < \infty$ or ∞ if $\gamma_\infty = \infty$ (cf. eqs. (4.19), (4.21) in the proof of Theorem 4.1). Also by Theorem 4.1, $Y^*/Y^+ \xrightarrow{P} 1$. Hence under the condition $\gamma_\infty = \infty$, bias can not be corrected by using Y^*, Y^+ or Y^{**} for large p. In fact we have shown that the Bayes estimator $\mathbf{a}_p^B$ and the sample-based esimator $\mathbf{a}_p^L$ have similar asymptotic properties. From the Bayes point of view, their posterior performances are asymptotically equivalent (cf. Theorem 4.1). From the classical point of view, their sample performances are asymptotically identical (cf. Theorems 4.3, 4.4). When used on a future case, they give the same asymptotic discrimination between the populations (cf. Theorem 4.2). Thus the determinism in the discrimination problem considered in this chapter implied by the use of the conjugate prior might be misleading and should be considered according to context.

Chapter 5

REGRESSION WITH CONJUGATE PRIOR

5.1 Introduction

Dawid (1988) considered a response variable X_0, and a potentially infinite sequence $(X_1, X_2, \cdots)$ of explanatory variables. The joint distribution for $(X_0, X_1, \cdots)$ is supposed to be multivariate normal, $N(0^T, \Sigma)$, where Σ is a $\infty \times \infty$ dispersion matrix. Let $\Gamma_p = \text{Var}(X_0 - E(X_0 \mid X_1, \cdots, X_p))$, the residual variance. Then as $p \to \infty$, Γ_p tends a limit, $\text{Var}(X_0 - E(X_0 \mid X_1, X_2, \cdots)) \stackrel{\text{def}}{=} \Gamma_\infty$. If $\Gamma_\infty = 0$, the sampling model is called deterministic, otherwise non–deterministic. In the former case it is expected that X_0 can be predicted arbitrarily closely by using sufficiently large number of predictors. If the parameter Σ is assigned a natural conjugate inverse Wishart distribution, $IW(\delta; Q)$, the resulting overall marginal distribution for the data matrix will be the matrix–t distribution. Let the data matrix be

$$Z = \begin{pmatrix} X_{f0} & X_{fp} \\ X_{t0} & X_{tp} \end{pmatrix},$$

where $(X_{t0}\ X_{tp})$ is the training set of n independent observations on $(X_0, X_1, \cdots, X_p)$, X_{fp} is a new forecast set of m observations on the explanatory variables $(X_1, \cdots, X_p)$, and X_{f0} is the set of the associated m response vectors to be predicted (now suppose $X_0 = (Y_1, Y_2, \cdots, Y_r)$ are multiple response variables). The desired conditional distribution for predicting X_{f0} on the basis of X_{fp}, X_{t0}, X_{tp} is (cf. Dawid, 1988 eq.(5.9))

$$X_{f0} \mid (X_{fp}, X_{t0}, X_{tp}) \sim A + T(\delta + n + p; L, M),$$

where the $(r+p) \times (r+p)$ leading matrix of Q, $Q_{r+p,r+p}$, is partitioned as

$$Q_{r+p} = \begin{pmatrix} Q_{00} & Q_{0p} \\ Q_{p0} & Q_{pp} \end{pmatrix} \begin{matrix} r \\ p \end{matrix}, \qquad \begin{matrix} r & p \end{matrix}$$

$$\begin{aligned} A &= X_{fp}(Q_{pp} + S_{pp})^{-1}(Q_{p0} + S_{p0}), \\ S_{pp} &= X'_{tp}X_{tp}, \quad S_{p0} = X'_{tp}X_{t0}, \end{aligned}$$

and L, M are functions of X_{fp}, X_{t0}, X_{tp}. The distributions of L and M are given by

$$\begin{aligned} L &\sim I_m + F(p, \delta + n; I_m) \text{ (matrix} - \text{variate } F \text{ distribution)}, \\ M &\sim \Lambda_p + F(n, \delta + p; \Lambda_p), \end{aligned}$$

where $\Lambda_p = Q_{00.p}$. We have $X_{f0} - A \xrightarrow{P} 0$, if $\Lambda_p \to 0$ as $p \to \infty$. Thus under this condition a perfect prediction from infinitely many explanatory variables is possible.

In this chapter we shall investigate the Bayes estimator more deeply with the emphasis on comparison with the least squares estimator. We again assume a joint normal distribution for the response and explanatory variables. We shall show that in the case of a conjugate prior, the Bayes and the least squares estimators are essentially the same as the number of explanatory variables tends to infinity.

Consider a response variable Y and a potentially infinite sequence $\mathbf{X} = (X_1, X_2, \cdots)'$ of explanatory variables. Let $\mathbf{X}_p = (X_1, \cdots, X_p)'$. Suppose for each

$p \geq 0$,

$$E(Y \mid \mathbf{X}_p) = \mathbf{X}_p' \boldsymbol{\beta}_p, \quad \boldsymbol{\beta}_p \in R^p.$$

Suppose we have observed the values of Y and all the X's for a random sample of n individuals, thus obtaining the training data $((Y_i, X_{i1}, X_{i2}, \cdots), i = 1, \cdots, n)$. Let

$$\mathbf{Y}^n = (Y_1, \cdots, Y_n)', \; X^n = (X_{ij}, i = 1, \cdots, n, j = 1, 2, \cdots),$$

and let X_p^n be the submatrix consisting of the first p columns of X^n. Suppose we now obtain the value of $\mathbf{X}_p$, denoted by $\mathbf{X}_p^0$, on a new individual. We wish to predict Y^0 on the basis of $\mathbf{X}_p^0$, conditional on the training data $\mathbf{Y}^n$, X^n. In this chapter we assume a conjugate prior. The full Bayes linear predictor is $(\mathbf{X}_p^0)' \boldsymbol{\beta}_p^B$, where the coefficient $\boldsymbol{\beta}_p^B \in R^p$, minimizes the posterior expected squared error loss $E[(Y^0 - (\mathbf{X}_p^0)'\mathbf{b}_p)^2 \mid \mathbf{Y}^n, X^n]$, $\mathbf{b}_p \in R^p$. A property of a conjugate prior for the parameters is that it implies a degenerate prediction such that the posterior expected squared error loss tends to zero under certain conditions on the parameters as p, the number of the observed variables, tends to infinity. We shall compare it with a sample–based estimator. In the classical approach a least squares estimator of $\boldsymbol{\beta}_p$, denoted by $\boldsymbol{\beta}_p^L$, is obtained by minimizing $\|\mathbf{Y}^n - X_p^n \mathbf{b}_p\|^2$, $\mathbf{b}_p \in R^p$, $\|\cdot\|$ being ℓ_2 norm in R^n. If $p > n$, the equation $\mathbf{Y}^n = X_p^n \mathbf{b}_p$ is consistent and has non-unique solution. We require $\boldsymbol{\beta}_p^L$ be such that it miminizes $E[(Y^0 - (\mathbf{X}_p^0)'\mathbf{b}_p)^2 \mid \mathbf{Y}^n, X^n]$, $\mathbf{b}_p \in R^p$, subject to $\mathbf{Y}^n = X_p^n \mathbf{b}_p$. We then show that this mixture solution is essentially the same as the full Bayes solution in that $(\mathbf{X}_p^0)' \boldsymbol{\beta}_p^B - (\mathbf{X}_p^0)' \boldsymbol{\beta}_p^L$ tends to zero in probability as p tends to infinity.

The different models $E(Y \mid \mathbf{X}_p) = \mathbf{X}_p \boldsymbol{\beta}_p$ are supposed to hold for every p simultaneously. To achieve this we make the following assumptions on the distribution of Y, $\mathbf{X}$:

Assumptions

Sampling distribution

Suppose

$$(Y, \mathbf{X}) \mid \Sigma \sim \mathcal{N}(1, \Sigma), \tag{5.1}$$

(cf. Dawid 1981, or Chapter 1 of this thesis for notation), i.e. Y, X's have a joint normal distribution with zero mean and covariances

$$\mathrm{Var(Y)} = \sigma_{00},\ \mathrm{Cov}(Y, X_i) = \sigma_{0i},\ \mathrm{Cov}(X_i, X_j) = \sigma_{ij},\ (i, j = 1, 2, \cdots),$$

σ_{ij} being the (i, j) element of the $\infty \times \infty$ symmetric matrix Σ, $i, j = 0, 1, 2, \cdots$. The assumption (5.1) is equivalent to the models (Dawid 1988):

$$Y \mid \mathbf{X}_p \sim \mathbf{X}_p'\boldsymbol{\beta}_p + \mathcal{N}(1, \Gamma_p),\ p = 1, 2, \cdots.$$

where

$$\begin{aligned} \Gamma_p &= \Sigma_{00.p} \stackrel{\text{def}}{=} \Sigma_{00} - \Sigma_{0p}\Sigma_{pp}^{-1}\Sigma_{p0}, \\ \boldsymbol{\beta}_p &= \Sigma_{pp}^{-1}\Sigma_{p0} \end{aligned}$$

with Σ_{1+p}, the leading $(1+p) \times (1+p)$ submatrix of Σ, being partitioned as

$$\Sigma_{1+p} = \begin{array}{c} \\ \\ \\ \end{array} \begin{array}{c} \begin{pmatrix} \Sigma_{00} & \Sigma_{0p} \\ \Sigma_{p0} & \Sigma_{pp} \end{pmatrix} \\ \begin{array}{cc} 1 & p \end{array} \end{array} \begin{array}{l} 1 \\ p \end{array} .$$

Prior Distribution

Suppose the parameter Σ is assigned the conjugate inverse Wishart distribution:

$$\Sigma \sim IW(\delta; Q), \tag{5.2}$$

where $\delta > 0$ is the degrees of freedom parameter, and $Q > 0$ an infinite dispersion matrix, i.e., for each q,

$$\Sigma_q \sim IW(\delta; Q_q),$$

where Σ_q, Q_q are the leading $q \times q$ submatrices of Σ, Q respectively. By Lemma 2 of Dawid 1988, or Theorem 13.4.2 of Dempster 1969, the distribution for $(\boldsymbol{\beta}_p, \Gamma_p)$ is given by

$$\begin{aligned} \Gamma_p &\sim IW(\delta + p; Q_{00.p}) \\ \boldsymbol{\beta}_p \mid \Gamma_p &\sim Q_{pp}^{-1}Q_{p0} + \mathcal{N}(Q_{pp}^{-1}, \Gamma_p), \end{aligned}$$

where Q_{1+p} is partitioned in the same fashion as Σ_{1+p}.

Let the full data matrix be

$$Z = \begin{pmatrix} Y^0 & (\mathbf{X}^0)' & 1 \\ \mathbf{Y}^n & X^n & n \\ 1 & \infty & \end{pmatrix}.$$

The overall marginal distribution of Z is the matrix-t distribution (see Dawid 1981 or Chapter 1 of this thesis):

$$Z \sim T(\delta; I_{n+1}, Q). \tag{5.3}$$

To make predictive inference of Y^0 on the basis of $\mathbf{X}^0$ and the training data $(\mathbf{Y}^n, X^n)$, we can first condition on $\mathbf{X}_p^0$ and $(\mathbf{Y}^n, X^n)$ and then let $p \to \infty$. Moreover it can be shown that if we have observed only the values of $X_1, \cdots, X_p$ for our new case, then we can use $\mathbf{Y}^n$ and the first p columns of X^n only in the training data (c.f. Lemma 4 of Dawid 1988). More precisely, the conditional distribution of $(Y^0, \mathbf{X}_p^0)$ given the full training data $(\mathbf{Y}^n, X^n)$ is the same as that of $(Y^0, \mathbf{X}_p^0)$ given $(\mathbf{Y}^n, X_p^n)$ and $Y^0 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X^n$ and $Y^0 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_p^n$ are also identically distributed. Hence in what follows we shall first constrain to the first p explanatory variables and condition on the subset $\mathbf{Y}^n$, X_p^n of the full training data $\mathbf{Y}^n$, X^n.

The following Lemma on the predictive distribution of $(Y^0, \mathbf{X}_p^0)$ given $(\mathbf{Y}^n, X_p^n)$ and the distribution of the training data $(\mathbf{Y}^n, X_p^n)$ is from Lemma 4 of Dawid 1988.

Lemma 5.1. Let

$$Q_{1+p}^* = Q_{1+p} + \begin{pmatrix} (\mathbf{Y}^n)'\mathbf{Y}^n & (\mathbf{Y}^n)'X_p^n \\ (X_p^n)'\mathbf{Y}^n & (X_p^n)'X_p^n \end{pmatrix}. \tag{5.4}$$

Partition Q_{1+p}, Q_{1+p}^* in the same fashion as Σ_{1+p}. Then under the assumptions (5.1) and (5.2),

$$(Y^0 \vdots (\mathbf{X}_p^0)') \mid \mathbf{Y}^n, X_p^n \sim T(\delta + n; 1, Q_{1+p}^*),$$

$$
\begin{aligned}
(\mathbf{X}_p^0)' \mid \mathbf{Y}^n, X_p^n &\sim T(\delta+n; 1, Q_{pp}^*), \\
(\mathbf{X}_p^0)' \mid X_p^n &\sim T(\delta+n; 1, Q_{pp}^*), \\
(\mathbf{Y}^n \vdots X_p^n) &\sim T(\delta; I_n, Q_{1+p}), \\
\mathbf{Y}^n \mid X_p^n &\sim X_p^n Q_{pp}^{-1} Q_{p0} + T(\delta+p; I_n + X_p^n Q_{pp}^{-1}(X_p^n)', Q_{00.p}), \\
X_p^n &\sim T(\delta; I_n, Q_{pp}).
\end{aligned}
$$

□

By standard analysis,

$$\Sigma \mid \mathbf{Y}^n, X^n \sim IW(\delta+n; Q^*).$$

Let

$$U = X_p^n Q_{pp}^{-\frac{1}{2}}. \tag{5.5}$$

Then $U \sim T(\delta; I_n, I_p)$. Lemma 1.5 gives the distributions of certain functions of U, which will be used in this Chapter.

5.2 Bayes Estimator

In this Section we deduce the Bayes estimator, denoted by $\boldsymbol{\beta}_p^B$, which minimizes the posterior expected squared error loss $E[(Y^0-(\mathbf{X}_p^0)'\mathbf{b}_p)^2 \mid \mathbf{Y}^n, X_p^n]$ when used as the coefficients in the predictor $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B$ and investigate its asymptotic properties. Suppose $n > 2-\delta$. By Lemma 5.1 and Lemma 1.2, $\mathrm{Var}[(Y^0 \vdots (\mathbf{X}_p^0)') \mid \mathbf{Y}^n, X_p^n] = Q_{1+p}^*/(\delta+n-2)$. Thus

$$R(\mathbf{b}_p) \stackrel{\text{def}}{=} (\delta+n-2)\cdot E[(Y^0-(\mathbf{X}_p^0)'\mathbf{b}_p)^2 \mid \mathbf{Y}^n, X_p^n] = Q_{00}^* - 2Q_{0p}^*\mathbf{b}_p + \mathbf{b}_p' Q_{pp}^* \mathbf{b}_p. \tag{5.6}$$

Proposition 5.1. The Bayes solution, denoted by $\boldsymbol{\beta}_p^B$, which minimizes the posterior expected squared error loss, or equivalently (5.6), is

$$
\begin{aligned}
&\boldsymbol{\beta}_p^B \\
=\ &E(\boldsymbol{\beta}_p \mid \mathbf{Y}^n, X_p^n)
\end{aligned} \tag{5.7}
$$

$$= Q_{pp}^{*-1} Q_{p0}^{*} \tag{5.8}$$

$$= Q_{pp}^{-1} Q_{p0} + Q_{pp}^{-1}(X_p^n)'(I_n + X_p^n Q_{pp}^{-1}(X_p^n)')^{-1}(\mathbf{Y}^n - X_p^n Q_{pp}^{-1} Q_{p0}) \tag{5.9}$$

$$= Q_{pp}^{-1} Q_{p0} + Q_{pp}^{-\frac{1}{2}} U'(I_n + UU')^{-1}(\mathbf{Y}^n - U Q_{pp}^{-\frac{1}{2}} Q_{p0}). \tag{5.10}$$

The corresponding minimum is $R(\boldsymbol{\beta}_p^B)/(\delta + n - 2)$, where

$$R(\boldsymbol{\beta}_p^B)$$

$$= Q_{00.p}^{*} \tag{5.11}$$

$$= Q_{00.p} + (\mathbf{Y}^n - U Q_{pp}^{-\frac{1}{2}} Q_{p0})'(I_n + UU')^{-1}(\mathbf{Y}^n - U Q_{pp}^{-\frac{1}{2}} Q_{p0}). \tag{5.12}$$

Proof. Write

$$R(\mathbf{b}_p) = (\mathbf{b}_p - Q_{pp}^{*-1} Q_{p0}^{*})' Q_{pp}^{*} (\mathbf{b}_p - Q_{pp}^{*-1} Q_{p0}^{*}) + Q_{00.p}^{*}. \tag{5.13}$$

Observe that $Q_{pp}^{*} > 0$, so $R(\mathbf{b}_p)$ attains its minimum $Q_{00.p}^{*}$ at $\mathbf{b}_p = Q_{pp}^{*-1} Q_{p0}^{*}$. The alternative expressions for $\boldsymbol{\beta}_p^B$ and $R(\boldsymbol{\beta}_p^B)$ can be obtained from Lemma 1.6 or (3.3), (5.10) and (5.14) of Dawid 1988.

The distribution of the Bayes estimator $\boldsymbol{\beta}_p^B$ is a mixture of the matrix–t distribution and the matrix–variate beta distribution, as stated in the following proposition.

Proposition 5.2. Under the assumptions (5.1) and (5.2), the Bayes estimator $\boldsymbol{\beta}_p^B$ has a stochastic representation

$$\boldsymbol{\beta}_p^B = Q_{pp}^{-1} Q_{p0} + T(\delta + p; V, Q_{00.p}), \quad \text{given } U,$$

where

$$V \stackrel{\text{def}}{=} Q_{pp}^{-\frac{1}{2}} U'(I_n + UU')^{-1} U Q_{pp}^{-\frac{1}{2}} \sim B(n, \delta + p - 1; Q_{pp}^{-1}).$$

Proof. Lemma 5.1 shows that

$$\mathbf{Y}^n - U Q_{pp}^{-\frac{1}{2}} Q_{p0} \sim T(\delta + p; I_n + UU', Q_{00.p}), \text{ given } U,$$

which combined with Proposition 5.1 gives the conditional distribution of $\boldsymbol{\beta}_p^B$ given U. By Lemma 1.6,

$$U'(I_n + UU')^{-1}U = I_p - (I_p + U'U)^{-1},$$

hence $Q_{pp}^{-\frac{1}{2}}U'(I_n + UU')^{-1}UQ_{pp}^{-\frac{1}{2}}$ has the stated distribution by Lemma 1.5. □

When using an estimator $\hat{\boldsymbol{\beta}}_p$ of $\boldsymbol{\beta}_p$ to predict the response Y^0 on the basis of the new observation $\mathbf{X}_p^0$, we are concerned with the predictive error $Y^0 - (\mathbf{X}_p^0)'\hat{\boldsymbol{\beta}}_p$. The distribution and asymptotic form of the predictive error for Bayes estimator are given in the following proposition.

Proposition 5.3. Under the assumptions (5.1) and (5.2), the conditional distribution of $Y^0 - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B$ given the training data is

$$Y^0 - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B \mid \mathbf{Y}^n, X_p^n \sim T(\delta + n; 1, R(\boldsymbol{\beta}_p^B)),$$

where $R(\boldsymbol{\beta}_p^B)$ is given in Proposition 5.1, and

$$\begin{aligned} R(\boldsymbol{\beta}_p^B) &\sim Q_{00.p} + F(n, \delta + p; Q_{00.p}) \\ &\xrightarrow{P} \Lambda_\infty \stackrel{\text{def}}{=} \lim_p Q_{00.p} \quad \text{as } p \to \infty. \end{aligned}$$

Moreover,

$$Y^0 - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B \xrightarrow{\mathcal{L}} T(\delta + n; 1, \Lambda_\infty), \quad \text{as } p \to \infty.$$

Proof. The distributions of $Y^0 - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B \mid (\mathbf{Y}^n, X_p^n)$ and $R(\boldsymbol{\beta}_p^B)$ are obtained from Lemma 5.1 and Lemma 1.5 or (6.11) and (6.9) of Dawid 1988. Note that $Q_{00.p} \geq 0$ is decreasing in p, so $\lim_p Q_{00.p}$ exists and is finite. The asymptotic distribution follows from Lemma 1.3 (d) and Lemma 1.4 (letting $A_p = \sqrt{R(\boldsymbol{\beta}_p^B)}$). □

From the theorem above the Bayes predictor has the property that, under the condition $Q_{00.p} \to 0$ $(p \to \infty)$, the difference between the response variable Y^0 and

the predictor $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B$ converges in $\mathcal{L}_2$ to 0, when the number p of the observed predictive variables tends to infinity.

5.3 Bayes–Least Squares Estimator

We have seen that the Bayes estimator, using a conjugate prior, implies a deterministic predictability. To investigate this property we shall compare it with the classical least squares estimator. The least squares estimator based on the training data $\mathbf{Y}^n$, X_p^n is obtained by minimizing

$$\|\mathbf{Y}^n - X_p^n \mathbf{b}_p\|^2, \ \mathbf{b}_p \in R^p.$$

If $n \geq p$, $(X_p^n)'X_p^n > 0$ with probability one so that the normal equation $(X_p^n)'X_p^n\mathbf{b}_p$ $= (X_p^n)'\mathbf{Y}^n$ has a unique solution $\hat{\mathbf{b}}_p = ((X_p^n)'X_p^n)^{-1}(X_p^n)'\mathbf{Y}^n$. If $n < p$, the equation $\mathbf{Y}^n = X_p^n\mathbf{b}_p$ is consistent and has nonunique solution. For definiteness we require that $\mathbf{b}_p$ also minimize the posterior expected error loss, or equivalently $R(\mathbf{b}_p)$ in (5.6), thus obtaining a mixture solution , denoted by $\boldsymbol{\beta}_p^L$ in the following proposition.

Proposition 5.4. Under the assumptions (5.1) and (5.2) with $p > n$, the mixture solution,which minimizes the posterior expected squared error loss, or equivalently $R(\mathbf{b}_p)$ in (5.6), subject to $\mathbf{Y}^n = X_p^n\mathbf{b}_p$, $\mathbf{b}_p \in R^p$, is

$$\begin{aligned}
\boldsymbol{\beta}_p^L &= Q_{pp}^{*^{-1}}Q_{p0}^* + Q_{pp}^{*^{-1}}(X_p^n)'(X_p^nQ_{pp}^{*^{-1}}(X_p^n)')^{-1}(\mathbf{Y}^n - X_p^nQ_{pp}^{*^{-1}}Q_{p0}^*) && (5.14)\\
&= Q_{pp}^{-1}Q_{p0} + Q_{pp}^{-1}(X_p^n)'(X_p^nQ_{pp}^{-1}(X_p^n)')^{-1}(\mathbf{Y}^n - X_p^nQ_{pp}^{-1}Q_{p0}) && (5.15)\\
&= Q_{pp}^{-1}Q_{p0} + Q_{pp}^{-\frac{1}{2}}U'(UU')^{-1}(\mathbf{Y}^n - UQ_{pp}^{-\frac{1}{2}}Q_{p0}). && (5.16)
\end{aligned}$$

The corresponding minimum is $R(\boldsymbol{\beta}_p^L)/(\delta + n - 2)$, where

$$\begin{aligned}
& R(\boldsymbol{\beta}_p^L)\\
= \ & Q_{00.p}^* + (\mathbf{Y}^n - X_p^nQ_{pp}^{*^{-1}}Q_{p0}^*)'(X_p^nQ_{pp}^{*^{-1}}(X_p^n)')^{-1}(\mathbf{Y}^n - X_p^nQ_{pp}^{*^{-1}}Q_{p0}^*) && (5.17)\\
= \ & Q_{00.p} + (\mathbf{Y}^n - X_p^nQ_{pp}^{-1}Q_{p0})'(X_p^nQ_{pp}^{-1}(X_p^n)')^{-1}(\mathbf{Y}^n - X_p^nQ_{pp}^{-1}Q_{p0}) && (5.18)
\end{aligned}$$

$$= Q_{00.p} + (\mathbf{Y}^n - UQ_{pp}^{-\frac{1}{2}}Q_{p0})'(UU')^{-1}(\mathbf{Y}^n - UQ_{pp}^{-\frac{1}{2}}Q_{p0}). \quad (5.19)$$

Proof. If $p > n$, then $\text{rank}(X_p^n) = n$ with probability one so that $(X_p^n Q_{pp}^{*-1}(X_p^n)')^{-1}$ exists and $X_p^n \mathbf{b}_p = \mathbf{Y}^n$ is consistent. Minimizing the first term in $R(\mathbf{b}_p)$ of (5.13), using (1f.1.5) in Rao (1973), p.60, we obtain (5.14) for $\boldsymbol{\beta}_p^L$ and (5.17) for $R(\boldsymbol{\beta}_p^L)$. To obtain the expression of $\boldsymbol{\beta}_p^L$, $R(\boldsymbol{\beta}_p^L)$ in terms of Q_{1+p} instead of Q_{1+p}^*, we note, by Lemma 1.6 and Lemma 5.1,

$$X_p^n Q_{pp}^{*-1}(X_p^n)' = I_n - (I_n + X_p^n Q_{pp}^{-1}(X_p^n)')^{-1},$$
$$Q_{pp}^{*-1}(X_p^n)' = Q_{pp}^{-1}(X_p^n)'(I_n + X_p^n Q_{pp}^{-1}(X_p^n)')^{-1}.$$

Hence

$$Q_{pp}^{*-1}(X_p^n)'(X_p^n Q_{pp}^{*-1}(X_p^n)')^{-1} = Q_{pp}^{-1}(X_p^n)'(X_p^n Q_{pp}^{-1}(X_p^n)')^{-1}.$$

By Proposition 5.1,

$$\begin{aligned}
& \mathbf{Y}^n - X_p^n Q_{pp}^{*-1} Q_{p0}^* \\
= \; & \mathbf{Y}^n - X_p^n [Q_{pp}^{-1}Q_{p0} + Q_{pp}^{-1}(X_p^n)'(I_n + X_p^n Q_{pp}^{-1}(X_p^n)')^{-1}(\mathbf{Y}^n - X_p^n Q_{pp}^{-1}Q_{p0})] \\
= \; & [I_n - X_p^n Q_{pp}^{-1}(X_p^n)'(I_n + X_p^n Q_{pp}^{-1}(X_p^n)')^{-1}](\mathbf{Y}^n - X_p^n Q_{pp}^{-1}Q_{p0}) \\
= \; & (I_n + X_p^n Q_{pp}^{-1}(X_p^n)')^{-1}(\mathbf{Y}^n - X_p^n Q_{pp}^{-1}Q_{p0}).
\end{aligned}$$

Substituting the above equations into (5.14) for $\boldsymbol{\beta}_p^L$, we obtain (5.15). The third expression (5.16) is from the definition of U in (5.5). By (5.6), Lemma 5.1 and the condition that $X_p^n \boldsymbol{\beta}_p^L = \mathbf{Y}^n$,

$$R(\boldsymbol{\beta}_p^L) = Q_{00}^* - 2Q_{0p}^* \boldsymbol{\beta}_p^L + (\boldsymbol{\beta}_p^L)' Q_{pp}^* \boldsymbol{\beta}_p^L \quad (5.20)$$
$$= Q_{00} - 2Q_{0p}\boldsymbol{\beta}_p^L + (\boldsymbol{\beta}_p^L)' Q_{pp} \boldsymbol{\beta}_p^L. \quad (5.21)$$

Since $R(\boldsymbol{\beta}_p^L)$ as a function of Q^* through (5.20) and (5.14) is (5.17), $R(\boldsymbol{\beta}_p^L)$ as a function of Q through (5.21) and (5.15) must be (5.17) with Q^* replaced by Q, which gives (5.18). □.

The following Proposition gives the distribution of the mixture solution $\boldsymbol{\beta}_p^L$, which is of a more complex form than that of $\boldsymbol{\beta}_p^B$.

Proposition 5.5. Under the assumptions (5.1) and (5.2) with $p > n$, the mixture solution $\boldsymbol{\beta}_p^L$ has a stochastic representation

$$\boldsymbol{\beta}_p^L = Q_{pp}^{-1} Q_{p0} + T(\delta + p; A, Q_{00.p}), \quad \text{given } U,$$

where

$$A = Q_{pp}^{-\frac{1}{2}} U'(UU')^{-1}(I_n + UU')(UU')^{-1} U Q_{pp}^{-\frac{1}{2}},$$

with $U \sim T(\delta; I_n, I_p)$.

Proof. The Proposition follows from Lemma 5.1 and Proposion 5.4. □

For the predictive error of the Bayes–least squares estimator the following Proposition holds:

Proposition 5.6. Under the assumptions (5.1) and (5.2) with $p > n$, the conditional distribution of $Y^0 - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L$ given the training data is

$$Y^0 - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L \mid \mathbf{Y}^n, X_p^n \sim T(\delta + n; 1, R(\boldsymbol{\beta}_p^L)),$$

where $R(\boldsymbol{\beta}_p^L)$ is given in Proposition 5.4, and

$$R(\boldsymbol{\beta}_p^L) \sim Q_{00.p} + F(n, p - n + 1; Q_{00.p}) \xrightarrow{P} \Lambda_\infty, \text{ as } p \to \infty.$$

Moreover,

$$Y^0 - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L \xrightarrow{\mathcal{L}} T(\delta + n; 1, \Lambda_\infty).$$

Proof. By Lemma 5.1, given $\mathbf{Y}^n, X_p^n$, the distribution of $Y^0 - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L$ is $T(\delta + n; 1, (1 \vdots (-\boldsymbol{\beta}_p^L)')Q_{1+p}^*(1 \vdots (-\boldsymbol{\beta}_p^L)')')$. The right-scale parameter is $R(\boldsymbol{\beta}_p^L)$ by Proposition 5.4. To specify the distribution of $R(\boldsymbol{\beta}_p^L)$, let

$$A = (UU')^{-\frac{1}{2}}(\mathbf{Y}^n - U Q_{pp}^{-\frac{1}{2}} Q_{p0}).$$

Then by Lemma 5.1,

$$A \sim T(\delta + p; V, Q_{00.p}) \quad \text{given } U,$$

where

$$V = (UU')^{-\frac{1}{2}}(I_n + UU')(UU')^{-\frac{1}{2}} = I_n + (UU')^{-1}.$$

By Lemma 1.5, $(UU')^{-1} \sim F(\delta + n - 1, p - n + 1; I_n)$, $(p \geq n)$. Hence $A \sim T(p-n+1, I_n; Q_{00.p})$ (cf. Lemma 5 in Dawid, 1988), so that $R(\boldsymbol{\beta}_p^L) = Q_{00.p} + A'A$ has the distribution stated in the Proposition. Its limit is obtained from Lemma 1.3 (d). The convergence of the unconditional distribution of $Y^0 - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L$ follows from Lemma 1.4. □

5.4 Comparison

We are now in a position to compare the performances of the Bayes estimator and the least squares estimator. We first investigate the difference between $\boldsymbol{\beta}_p^L$ and $\boldsymbol{\beta}_p^B$.

Theorem 5.1. Let $R = R(\boldsymbol{\beta}_p^L) - R(\boldsymbol{\beta}_p^B)$. Then under the assumptions (5.1) and (5.2) with $p > n$,

$$\begin{aligned} R &= (\boldsymbol{\beta}_p^L - \boldsymbol{\beta}_p^B)'Q_{pp}^*(\boldsymbol{\beta}_p^L - \boldsymbol{\beta}_p^B) && (5.22) \\ &= (\mathbf{Y}^n - UQ_{pp}^{-\frac{1}{2}}Q_{p0})'[(UU')^{-1} - (I_n + UU')^{-1}](\mathbf{Y}^n - UQ_{pp}^{-\frac{1}{2}}Q_{p0}) && (5.23) \\ &\stackrel{d}{=} T'FT && (5.24) \end{aligned}$$

where $T \sim T(\delta + p; I_n, Q_{00.p})$, $F \sim F(\delta + n - 1, p - n + 1; I_n)$, and $T \perp\!\!\!\perp F$.

Moreover,

$$E(R) = \frac{n(\delta + n - 1)}{(p - n - 1)(\delta + p - 2)} Q_{00.p} \to 0, \quad \text{as } p \to \infty. \tag{5.25}$$

Proof. Letting $\mathbf{b}_p = \boldsymbol{\beta}_p^L$ in the decomposition formula (5.13) for the posterior expected squared error loss in Proposition 5.1, we obtain (5.22). From (5.12) in proposition 5.1 and (5.19) in Proposition 5.4 we obtain (5.23). Let $T = (I_n + UU')^{-\frac{1}{2}}(\mathbf{Y}^n - UQ_{pp}^{-\frac{1}{2}}Q_{p0})$. We have, by Lemma 5.1, $T \mid U \sim T(\delta + p; I_n, Q_{00.p})$,

which does not depend on U, so that $T \sim T(\delta + p; I_n, Q_{00.p})$ unconditionally, and $T \perp\!\!\!\perp U$. Since

$$(I_n + UU')(UU')^{-1} = (UU')^{-1}(I_n + UU'),$$

by Lemma 1.7,

$$\begin{aligned} & (I_n + UU')^{\frac{1}{2}}(UU')^{-\frac{1}{2}} \\ = \; & [(UU')^{-1}(I_n + UU')]^{\frac{1}{2}} \\ = \; & [(UU')^{-1} + I_n]^{\frac{1}{2}}. \end{aligned}$$

Hence (5.24) holds with

$$\begin{aligned} F &= (I_n + UU')^{\frac{1}{2}}[(UU')^{-1} - (I_n + UU')^{-1}](I_n + UU')^{\frac{1}{2}} \\ &= (I_n + UU')^{\frac{1}{2}}(UU')^{-1}(I_n + UU')^{\frac{1}{2}} - I_n \\ &= I_n + (UU')^{-1} - I_n \\ &= (UU')^{-1}. \end{aligned}$$

The distribution of F follows from Lemma 1.5. The expectation of R is, by Lemma 1.2, Lemma 1.1,

$$\begin{aligned} E(R) &= E[E(R \mid F)] = E[\mathrm{tr}(F/(\delta + p - 2))]Q_{00.p} \\ &= \mathrm{tr}[E(F)] \cdot \frac{Q_{00.p}}{\delta + p - 2} = \mathrm{tr}(\frac{\delta + n - 1}{p - n + 1 - 2} \cdot I_n) \cdot \frac{Q_{00.p}}{\delta + p - 2}, \end{aligned}$$

which leads to (5.25). □

We note that apart from a constant $(\delta + n - 2)$, R is the difference of posterior expected squared error loss between the estimators $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L$ and $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B$. We have seen from Propositions 5.3 and 5.6 that $R(\boldsymbol{\beta}_p^L)$ and $R(\boldsymbol{\beta}_p^B)$ have the same limit (in probability) so that $R \xrightarrow{P} 0$ as $p \to \infty$. Theorem 5.1 gives a stronger result that $R \xrightarrow{\mathcal{L}_1} 0$ as $p \to \infty$. The above theorem also shows that, when p is large, the distance between $\boldsymbol{\beta}_p^L$ and $\boldsymbol{\beta}_p^B$ is negligible if we define the norm of $\mathbf{z} \in R^p$ by $(\mathbf{z}'Q_{pp}^*\mathbf{z})^{\frac{1}{2}}$. Next we shall give the distribution of the difference of the two predictors $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L$ and $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B$.

Theorem 5.2. Under the assumptions (5.1) and (5.2) with $p > n$, the conditional distribution of the difference of the two predictors $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L$ and $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B$ given the training data $\mathbf{Y}^n, X_p^n$ is

$$(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B \mid \mathbf{Y}^n, X_p^n \sim T(\delta + n; 1, R), \tag{5.26}$$

where R is given in Theorem 5.1 so that

$$(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B \xrightarrow{\mathcal{L}_2} 0, \quad \text{as } p \to \infty. \tag{5.27}$$

Proof. The distribution (5.26) follows from Lemma 5.1 and (5.22) in Theorem 5.1. Thus

$$E[(\mathbf{X}_p^0)'(\boldsymbol{\beta}_p^L - \boldsymbol{\beta}_p^B)]^2 = E(R)/(\delta + n - 2) \to 0, \text{as } p \to \infty,$$

by Theorem 5.1 and Lemma 1.2, which leads to (5.27). □

Theorem 5.2 shows that the two estimators $\boldsymbol{\beta}_p^L$ and $\boldsymbol{\beta}_p^B$ are asymptotically equivalent in the sense that, when they are used to make prediction of a future response Y^0, the difference of the two predictors, $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B$, tends to zero in $\mathcal{L}_2$.

Finally, we shall investigate the error sum of the squares of the Bayes estimator $\boldsymbol{\beta}_p^B$.

Theorem 5.3. Under the assumptions (5.1) and (5.2), the distribution of $\mathbf{Y}^n - X_p^n\boldsymbol{\beta}_p^B$ is given by

$$\mathbf{Y}^n - X_p^n\boldsymbol{\beta}_p^B \sim T(\delta + p; (I_n + UU')^{-1}, Q_{00.p}) \quad \text{given } U, \tag{5.28}$$

with

$$(I_n + UU')^{-1} \sim B(\delta + n - 1; p, I_n). \tag{5.29}$$

Moreover,

$$E\|\mathbf{Y}^n - X_p^n\boldsymbol{\beta}_p^B\|^2 = Q_{00.p} \cdot \frac{n(\delta + n - 1)}{(\delta + p - 2)(\delta + n + p - 1)}, \tag{5.30}$$

so that

$$\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B \xrightarrow{\mathcal{L}_2} 0, \quad p \to \infty. \tag{5.31}$$

Proof. By Lemma 1.6,

$$\begin{aligned} & -U + UU'(I_n + UU')^{-1}U \\ = \; & -[I_n - UU'(I_n + UU')^{-1}]U \\ = \; & -(I_n + UU')^{-1}U \end{aligned}$$

Hence by Proposition 5.1,

$$\begin{aligned} & \mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B \\ = \; & [I_n - UU'(I_n + UU')^{-1}]\mathbf{Y}^n + [-U + UU'(I_n + UU')^{-1}U]Q_{pp}^{-\frac{1}{2}} Q_{p0} \\ = \; & (I_n + UU')^{-1}(\mathbf{Y}^n - UQ_{pp}^{-\frac{1}{2}} Q_{p0}), \end{aligned}$$

which combined with Lemma 5.1 and Lemma 1.5 establishes (5.28) and (5.29). Thus by Lemma 1.1, Lemma 1.2,

$$\begin{aligned} & E\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B\|^2 = E[E(\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B\|^2 \mid U)] \\ = \; & E\mathrm{tr}[(I_n + UU')^{-1}] \cdot \frac{Q_{00.p}}{\delta + p - 2} = \mathrm{tr}(\frac{\delta + n - 1}{\delta + n + p - 1} \cdot I_n) \cdot \frac{Q_{00.p}}{\delta + p - 2}, \end{aligned}$$

which leads to (5.30). □

We know that when $p > n$, the least squares estimator $\boldsymbol{\beta}_p^L$ satisfies $\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^L\|^2 = 0$, the error sum of squares being zero. Theorem 5.3 shows the full Bayes estimator, using a conjugate prior, has a similar property asymptotically.

5.5 Discussion

Using the notation introduced in Section 1.2, let

$$\phi_{\mathbf{b}_p} = E(Y - \mathbf{X}_p' \mathbf{b}_p)^2 = \Sigma_{00} - 2\Sigma_{0p}\mathbf{b}_p + \mathbf{b}_p' \Sigma_{pp} \mathbf{b}_p.$$

Then the aim in regression is to minimise $\phi_{\mathbf{b}_p}$. Let

$$\phi^{**} = \phi_{\mathbf{b}_p^{**}} = \inf_{\mathbf{b}_p}\{\phi_{\mathbf{b}_p}\} = \Sigma_{00.p} = \Gamma_p, \text{ (equivalently , } -\phi^{**} = \sup_{\mathbf{b}_p}\{-\phi_{\mathbf{b}_p}\}, \text{ etc.)}$$

with

$$\mathbf{b}_p^{**} = \Sigma_{pp}^{-1}\Sigma_{p0} = \boldsymbol{\beta}_p.$$

Let

$$\begin{aligned} S &= (\mathbf{Y}^n\ X_p^n)'(\mathbf{Y}^n\ X_p^n) \\ &= \begin{pmatrix} S_{00} & S_{0p} \\ S_{p0} & S_{pp} \end{pmatrix} \begin{matrix} 1 \\ p \end{matrix} \\ &\qquad\quad \begin{matrix} 1 & p \end{matrix} \end{aligned} . \tag{5.32}$$

An unbiased estimate from the training data $(\mathbf{Y}^n, X_p^n)$ for $\phi_{\mathbf{b}_p}$ is

$$X_{\mathbf{b}_p} = \|\mathbf{Y}^n - X_p^n\mathbf{b}_p\|^2/n.$$

If $p \leq n$, S_{pp} is nonsingular, $X_{\mathbf{b}_p}$ achieves its minimum at

$$X^* = X_{\mathbf{b}_p^*} = S_{00.p}/n$$

at

$$\mathbf{b}_p^* = S_{pp}^{-1}S_{p0}.$$

Since $S \sim W_{p+1}(n; \Sigma_{p+1})$, $S_{00.p} \sim W_1(n-p; \Sigma_{00.p})$, so

$$E(X^*) = \frac{n-p}{n}\Sigma_{00.p} < \phi^{**}.$$

Thus X^* is negatively biased for ϕ^{**}. Now consider the case that $p > n$. Then a $\mathbf{b}_p^*$ can be found such that $X_{\mathbf{b}_p^*} = 0$, e.g. $\mathbf{b}_p^* = \boldsymbol{\beta}_p^L$. Let

$$\phi^* = \phi_{\mathbf{b}_p^*} = \Sigma_{00} - 2\Sigma_{0p}\boldsymbol{\beta}_p^L + (\boldsymbol{\beta}_p^L)'\Sigma_{pp}\boldsymbol{\beta}_p^L$$

be the "data-dependent parameter". For making Bayesian analysis consider

$$Y_{\mathbf{b}_p} = E(\phi_{\mathbf{b}_p} \mid \mathbf{Y}^n, X_p^n).$$

Then by (5.6) in Section 5.2,

$$Y_{\mathbf{b}_p} = (Q_{00}^* - 2Q_{0p}^*\mathbf{b}_p + \mathbf{b}_p'Q_{pp}^*\mathbf{b}_p)/(\delta + n - 2).$$

$Y_{\mathbf{b}_p}$ achieves its minimum

$$\inf_{\mathbf{b}_p}\{Y_{\mathbf{b}_p}\} = Q^*_{00.p}/(\delta+n-2) = R(\boldsymbol{\beta}^B_p)/(\delta+n-2) \stackrel{\text{def}}{=} Y^+$$

at

$$\Lambda^+ = \boldsymbol{\beta}^B_p,$$

(cf. Proposition 5.1). Since $\Sigma_{1+p} \mid \mathbf{Y}^n, X^n_p \sim IW(\delta+n, Q^*_{1+p})$, $\Sigma_{00.p} \mid \mathbf{Y}^n, X^n_p \sim IW(\delta+n+p, Q^*_{00.p})$ (cf. Dawid, 1988, Lemma 2). Thus

$$Y^{**} \stackrel{\text{def}}{=} E(\phi^{**} \mid \mathbf{Y}^n, X^n_p) = Q^*_{00.p}/(\delta+n+p-2) = R(\boldsymbol{\beta}^B_p)/(\delta+n+p-2).$$

Also

$$\begin{aligned} Y^* &\stackrel{\text{def}}{=} E(\phi^* \mid \mathbf{Y}^n, X^n_p) \\ &= E(\Sigma_{00} \mid \mathbf{Y}^n, X^n_p) - 2E(\Sigma_{0p} \mid \mathbf{Y}^n, X^n_p)\boldsymbol{\beta}^L_p + (\boldsymbol{\beta}^L_p)' E(\Sigma_{pp} \mid \mathbf{Y}^n, X^n_p)\boldsymbol{\beta}^L_p \\ &= (Q^*_{00} - 2Q^*_{0p}\boldsymbol{\beta}^L_p + (\boldsymbol{\beta}^L_p)' Q^*_{pp}\boldsymbol{\beta}^L_p)/(\delta+n-2) \\ &= R(\boldsymbol{\beta}^L_p)/(\delta+n-2), \end{aligned}$$

(cf. Proposition 5.4). Y^{**}, Y^+, Y^* can be used as Bayesian estimates for ϕ^{**}, $\phi^+ = \phi_{\Lambda^+}$, ϕ^* respectively. Since $R(\boldsymbol{\beta}^B_p) < R(\boldsymbol{\beta}^L_p)$ a.s., (cf. Theorem 5.1),

$$Y^{**} < Y^+ < Y^*.$$

For any finite p, these estimators are greater than 0, and thus adjust bias in the right direction. However, if $\Lambda_\infty = 0$, as $p \to \infty$,

$$Y^+ \xrightarrow{P} \Lambda_\infty/(\delta+n-2) = 0, \; Y^* \xrightarrow{P} \Lambda_\infty/(\delta+n-2) = 0$$

by Propositions 5.3 and 5.6. Hence asymptotically, Y^{**}, Y^+, Y^* make no correction for X^*. In another word, asymptotically perfect prediction is possible by using the Bayes estimator $\boldsymbol{\beta}^B_p$ and the Bayes–least estimator $\boldsymbol{\beta}^L_p$ if $\Lambda_\infty = 0$. This property is undesirable if we regard $\boldsymbol{\beta}^L_p$ is a sample–based estimator with $X_{\boldsymbol{\beta}^L_p} = 0$ and compare the Bayes estimator $\boldsymbol{\beta}^B_p$ with it. From the Bayes point of view, Theorem 5.1 shows that the difference of their posterior expected error losses tends to zero as $p \to \infty$. From the classical point of view, Theorem 5.3 shows that the error sum of squares of the Bayes esimator $\boldsymbol{\beta}^B_p$, equivalently $X_{\boldsymbol{\beta}^B_p}$, tends to zero

as $p \to \infty$, compatible with the condition that $\mathbf{Y}^n = X_p^n \boldsymbol{\beta}_p^L$ or $X_{\boldsymbol{\beta}_p^L} = 0$. And on a future case, the two predictors $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B$, $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L$ are asymptotically identical as shown in Theorem 5.2. The above three theorems hold without condition that $\Lambda_\infty = 0$. Hence as in the discrimination problems discussed in Chapters 2, 3 and 4, the congjugate prior neglects the problem of overfitting in regression.

Chapter 6

REGRESSION WITH NONCONJUGATE PRIOR

6.1 Introduction

Conjugate priors are frequently used in Bayesian analysis for the resulting ease in calculation and the possibility of reasonable approximation to the true prior, at least for initial analysis (Berger, 1985, Section 4.2.2). However, in the regression discussed in the last chapter, under the natural conjugate inverse Wishart prior, the Bayes predictor leads to deterministic predictivity under a condition on the hyperparameters ($\Lambda_\infty = 0$), and appears to neglect the problems of overfitting. In many cases it is unreasonable to believe this determinism holds. This suggests that we should investigate nonconjugate priors as alternatives. In this chapter we investigate regression under a certain nonconjugate prior. We shall show that, under this prior, the Bayes predictor will not lead to deterministic predictability, and is different from the sample–based Bayes–least squares predictor which fits the data exactly.

Prior Assumption

To construct a nonconjugate prior, we suppose the response Y is the sum of an unobservable variable η and an error α, independent of each other; the joint distribution of η and the explanatory variables X_i is normal; α is also normally distributed. We thus suppose:

$$\begin{aligned} & Y = \eta + \alpha, \\ & (\eta,\ \mathbf{X}) \sim \mathcal{N}(1, \Sigma),\ \alpha \sim \mathcal{N}(1, \Phi),\ (\eta,\ \mathbf{X}) \perp\!\!\!\perp \alpha, \end{aligned} \tag{6.1}$$

where $\mathbf{X} = (X_1, X_2, \cdots)$, $\Sigma > 0$ is $\infty \times \infty$, $\Phi > 0$. The assumption (6.1) is equivalent to the following models:

$$Y \mid \mathbf{X}_p \sim (\mathbf{X}_p)'\boldsymbol{\beta}_p + \mathcal{N}(1, \sigma^2), \tag{6.2}$$

where

$$\begin{aligned} \boldsymbol{\beta}_p &= \Sigma_{pp}^{-1}\Sigma_{p0}, \\ \Gamma_p &= \Sigma_{00.p} \stackrel{\text{def}}{=} \Sigma_{00} - \Sigma_{0p}\Sigma_{pp}^{-1}\Sigma_{p0}, \\ \sigma^2 &= \Gamma_p + \Phi. \end{aligned}$$

Suppose the prior distribution for the parameters Σ, Φ is

$$\Sigma \sim IW(\delta; Q),\ \Phi \sim IW(\nu; K),\ \Sigma \perp\!\!\!\perp \Phi, \tag{6.3}$$

where $\delta > 0$, $Q > 0$ is $\infty \times \infty$, $\nu > 0$, $K > 0$.

Now suppose we have obtained the $n \times \infty$ training data $(\mathbf{Y}^n, X^n)$ on the response variable Y and all the X's, where $\mathbf{Y}^n$ is represented by $\mathbf{Y}^n = \boldsymbol{\eta} + \boldsymbol{\alpha}$. A further case for the response and the potentially infinitely many explanatory variables is obtained and denoted by $(Y^0\ (\mathbf{X}^0)')$, where Y^0 is represented by $Y^0 = \eta_0 + \alpha_0$. We wish to predict Y^0 on the basis of the training data and $\mathbf{X}_p^0$, the observation on the first p variables, and shall investigate the asymptotic property of the predictor as $p \to \infty$.

From the asumptions the overall marginal distribution of the full data matrix is a sum of two independent matrix–t distributions,

$$Z = \begin{pmatrix} Y^0 & (\mathbf{X}^0)' \\ \mathbf{Y}^n & X^n \end{pmatrix} = Z_1 + (Z_2, 0), \tag{6.4}$$

where

$$\begin{aligned} Z_1 &= \begin{pmatrix} \eta_0 & (\mathbf{X}^0)' \\ \boldsymbol{\eta} & X^n \end{pmatrix} \sim T(\delta; I_{n+1}, Q), \\ Z_2 &= \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha} \end{pmatrix} \sim T(\nu; I_{n+1}, K), \\ & \text{and } Z_1 \perp\!\!\!\perp Z_2. \end{aligned}$$

As in Chapter 5, we use $\mathbf{X}_p$, X_p^n to indicate the corresponding submatrices of $\mathbf{X}$, X^n restricted to the first p variables, and Σ_{1+p}, Q_{1+p} the $(1+p) \times (1+p)$ leading submatrices of Σ, Q. Let Σ_{1+p} be partitioned as

$$\Sigma_{1+p} = \begin{array}{c} \\ \begin{pmatrix} \Sigma_{00} & \Sigma_{0p} \\ \Sigma_{p0} & \Sigma_{pp} \end{pmatrix} \\ \begin{matrix} 1 & p \end{matrix} \end{array} \begin{array}{c} 1 \\ p \\ \\ \end{array}$$

and let Q_{1+p} be partitioned in the same fashion. Then by assumption (6.3),

$$\begin{aligned} \boldsymbol{\beta}_p \mid \Gamma_p &\sim Q_{pp}^{-1} Q_{p0} + \mathcal{N}(Q_{pp}^{-1}, \Gamma_p), \\ \Gamma_p &\sim IW(\delta + p; Q_{00.p}), \\ \sigma^2 &= \Gamma_p + \Phi,\ \Phi \sim IW(\nu; K),\ \Phi \perp\!\!\!\perp \Gamma_p. \end{aligned}$$

Note the assumptions that $\alpha \perp\!\!\!\perp (\eta, \mathbf{X})$ given the parameters, and $\Sigma \perp\!\!\!\perp \Phi$, imply that $\alpha \perp\!\!\!\perp (\eta, \mathbf{X})$ in the marginal distribution of $(\eta, \alpha,\ \mathbf{X})$. To make predictive inference for Y^0 on the basis of $\mathbf{X}^0$ and the training data $(\mathbf{Y}^n,\ X^n)$, we shall first condition on the training data $(\mathbf{Y}^n,\ X_q^n)$ and $\mathbf{X}_p^0$, $(p \leq q)$, and then let $q \to \infty$, $p \to \infty$. A detailed discussion of the reason is in Section 6.2. We here give a lemma on the posterior variance of $(Y^0\ (\mathbf{X}_p^0)')$ given $(\mathbf{Y}^n, X_q^n)$, which is the basis of the calculation of this chapter.

Lemma 6.1. Let $Q^*_{1+p}(q) = \mathrm{Var}[(Y^0 \ (\mathbf{X}^0_p)')' \mid \mathbf{Y}^n, X^n_q] : (1+p) \times (1+p)$. $(p \le q \le \infty)$. Then under the assumptions (6.1) and (6.3), the following hold:

$$E[(Y^0 \vdots (\mathbf{X}^0_p)') \mid \mathbf{Y}^n, X^n_q] = 0, \tag{6.5}$$

$$\begin{aligned}
Q^*_{00}(q) &= E(Y^2 \mid \mathbf{Y}^n, X^n_q] = E(\Sigma_{00} + \Phi \mid \mathbf{Y}^n, X^n_q), && (6.6)\\
&= \frac{Q_{00} + E(\boldsymbol{\eta}'\boldsymbol{\eta} \mid \mathbf{Y}^n, X^n_q)}{\delta + n - 2} + \frac{K + E(\boldsymbol{\alpha}'\boldsymbol{\alpha} \mid \mathbf{Y}^n, X^n_q)}{\nu + n - 2} && (6.7)\\
Q^*_{p0}(q) &= E(\mathbf{X}^0_p Y \mid \mathbf{Y}^n, X^n_q) = E(\Sigma_{p0} \mid \mathbf{Y}^n, X^n_q) && (6.8)\\
&= \frac{Q_{p0} + (X^n_p)' E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X^n_q)}{\delta + n - 2} && (6.9)\\
&= \frac{Q_{p0} + (X^n_p)'\mathbf{Y}^n - (X^n_p)' E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X^n_q)}{\delta + n - 2} && (6.10)\\
Q^*_{pp}(q) &= E(\mathbf{X}^0_p (\mathbf{X}^0_p)' \mid \mathbf{Y}^n, X^n_q) = E(\Sigma_{pp} \mid \mathbf{Y}^n, X^n_p) && (6.11)\\
&= \frac{(Q_{pp} + (X^n_p)' X^n_p)}{\delta + n - 2} \stackrel{\text{def}}{=} Q^*_{pp} && (6.12)
\end{aligned}$$

where Q_{1+p} and $Q^*_{1+p}(q)$ is partitioned as

$$Q_{1+p} = \begin{array}{c} \\ 1 \\ p \end{array} \begin{array}{c} \begin{array}{cc} 1 & p \end{array} \\ \begin{pmatrix} Q_{00} & Q_{0p} \\ Q_{p0} & Q_{pp} \end{pmatrix} \end{array}, \quad Q^*_{1+p}(q) = \begin{array}{c} \\ 1 \\ p \end{array} \begin{array}{c} \begin{array}{cc} 1 & p \end{array} \\ \begin{pmatrix} Q^*_{00}(q) & Q^*_{0p}(q) \\ Q^*_{p0}(q) & Q^*_{pp}(q) \end{pmatrix} \end{array}.$$

Moreover, $\lim_{q\to\infty} Q^*_{1+p}(q)$ exists a.s. $(\delta > 2,\ \nu > 2)$ and is denoted by $Q^*_{1+p}(\infty)$.

Proof. By the independence property of $(Y^0, \mathbf{X}^0_p)$ and $(\mathbf{Y}^n, X^n_q)$ given (Σ, Φ), we have

$$\begin{aligned}
& E[(Y^0 \vdots (\mathbf{X}^0_p)') \mid \mathbf{Y}^n, X^n_q]\\
= \ & E\{E[(\eta_0 + \alpha_0 \vdots (\mathbf{X}^0_p)') \mid \mathbf{Y}^n, X^n_q, \Sigma, \Phi] \mid \mathbf{Y}^n, X^n_q\}\\
= \ & E\{E[(\eta_0 + \alpha_0 \vdots (\mathbf{X}^0_p)') \mid \Sigma, \Phi] \mid \mathbf{Y}^n, X^n_q\}\\
= \ & 0.
\end{aligned}$$

Similarly,

$$E(\alpha_0 \eta_0 \mid \mathbf{Y}^n, X^n_q)$$

$$
\begin{aligned}
&= E[E(\alpha_0\eta_0 \mid \Sigma, \Phi) \mid \mathbf{Y}^n, X_q^n] \\
&= E[E(\alpha_0 \mid \Phi)E(\eta_0 \mid \Sigma) \mid \mathbf{Y}^n, X_q^n] \\
&= 0, \\
&\quad E(\alpha_0(\mathbf{X}_p^0)' \mid \mathbf{Y}^n, X_q^n) \\
&\quad E[E(\alpha_0(\mathbf{X}_p^0)' \mid \Sigma, \Phi) \mid \mathbf{Y}^n, X_q^n] \\
&= 0,
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\mathrm{Var}\left(\begin{pmatrix} Y^0 \\ \mathbf{X}_p^0 \end{pmatrix} \middle| \mathbf{Y}^n, X_q^n\right) \\
&= E\left[\begin{pmatrix} Y^0 \\ \mathbf{X}_p^0 \end{pmatrix} \begin{pmatrix} Y^0 & (\mathbf{X}_p^0)' \end{pmatrix} \middle| \mathbf{Y}^n, X_q^n\right] \\
&= E\left[\begin{pmatrix} \eta_0 \\ \mathbf{X}_p^0 \end{pmatrix} \begin{pmatrix} \eta_0 & (\mathbf{X}_p^0)' \end{pmatrix} \middle| \mathbf{Y}^n, X_q^n\right] \\
&+ E\left[\begin{pmatrix} \alpha_0 \\ 0 \end{pmatrix} \begin{pmatrix} \alpha_0 & 0 \end{pmatrix} \middle| \mathbf{Y}^n, X_q^n\right].
\end{aligned}
$$

By first conditioning on the parameters Σ, Φ and the training data $\mathbf{Y}^n$, X_q^n, we have, by (6.1),

$$
\begin{aligned}
&E\left[\begin{pmatrix} \eta_0 \\ \mathbf{X}_p^0 \end{pmatrix} \begin{pmatrix} \eta_0 & (\mathbf{X}_p^0)' \end{pmatrix} \middle| \mathbf{Y}^n, X_q^n\right] \\
&= E\left\{E\left[\begin{pmatrix} \eta_0 \\ \mathbf{X}_p^0 \end{pmatrix} \begin{pmatrix} \eta_0 & (\mathbf{X}_p^0)' \end{pmatrix} \middle| \Sigma, \Phi\right] \middle| \mathbf{Y}^n, X_q^n\right\} \\
&= E(\Sigma_{1+p} \mid \mathbf{Y}^n, X_q^n), \\
&\quad E(\alpha_0^2 \mid \mathbf{Y}^n, X_q^n) \\
&= E[E(\alpha_0^2 \mid \Phi) \mid \mathbf{Y}^n, X_q^n] = E(\Phi \mid \mathbf{Y}^n, X_q^n),
\end{aligned}
$$

yielding the first expressions for $Q_{ij}^*(q)$. Now

$$
\begin{pmatrix} \eta_0 & (\mathbf{X}_q^0)' \\ \boldsymbol{\eta} & X_q^n \end{pmatrix} \perp\!\!\!\perp \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha} \end{pmatrix}
$$

implies that

$$(\eta_0 \quad (\mathbf{X}_q^0)') \perp\!\!\!\perp \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha} \end{pmatrix} \Big| (\boldsymbol{\eta} \quad X_q^n),$$

and

$$\alpha_0 \perp\!\!\!\perp \begin{pmatrix} \eta_0 & (\mathbf{X}_q^0)' \\ \boldsymbol{\eta} & X_q^n \end{pmatrix} \Big| \boldsymbol{\alpha}.$$

Hence by Lemma 5.1,

$$\begin{aligned} & E\left[\begin{pmatrix} \eta_0 \\ \mathbf{X}_p^0 \end{pmatrix} (\eta_0 \quad (\mathbf{X}_p^0)') \Big| \boldsymbol{\eta}, \boldsymbol{\alpha}, X_q^n \right] \\ = \; & E\left[\begin{pmatrix} \eta_0 \\ \mathbf{X}_p^0 \end{pmatrix} (\eta_0 \quad (\mathbf{X}_p^0)') \Big| \boldsymbol{\eta}, X_q^n \right] \\ = \; & E\left[\begin{pmatrix} \eta_0 \\ \mathbf{X}_p^0 \end{pmatrix} (\eta_0 \quad (\mathbf{X}_p^0)') \Big| \boldsymbol{\eta}, X_p^n \right] \\ = \; & \frac{Q_{1+p} + (\boldsymbol{\eta} \quad X_p^n)'(\boldsymbol{\eta} \quad X_p^n)}{\delta + n - 2} \end{aligned}$$

and

$$E(\alpha_0^2 \mid \boldsymbol{\eta}, \boldsymbol{\alpha}, X_q^n) = E(\alpha_0^2 \mid \boldsymbol{\alpha}) = \frac{K + \boldsymbol{\alpha}'\boldsymbol{\alpha}}{\nu + n - 2}.$$

Thus by first conditioning on $\boldsymbol{\eta}$, $\boldsymbol{\alpha}$, X_q^n, we have

$$\begin{aligned} & E\left[\begin{pmatrix} \eta_0 \\ \mathbf{X}_p^0 \end{pmatrix} (\eta_0 \quad (\mathbf{X}_p^0)') \Big| \mathbf{Y}^n, X_q^n \right] \\ = \; & E\left\{ E\left[\begin{pmatrix} \eta_0 \\ \mathbf{X}_p^0 \end{pmatrix} (\eta_0 \quad (\mathbf{X}_p^0)') \Big| \boldsymbol{\eta}, X_p^n \right] \Big| \mathbf{Y}^n, X_q^n \right\} \\ = \; & \frac{Q_{1+p} + \begin{pmatrix} E(\boldsymbol{\eta}'\boldsymbol{\eta} \mid \mathbf{Y}^n, X_q^n) & E(\boldsymbol{\eta}' \mid \mathbf{Y}^n, X_q^n) X_p^n \\ (X_p^n)' E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X_q^n) & (X_p^n)' X_p^n \end{pmatrix}}{\delta + n - 2}, \end{aligned}$$

and

$$E(\alpha_0^2 \mid \mathbf{Y}^n, X_q^n) = E[E(\alpha_0^2 \mid \boldsymbol{\alpha}) \mid \mathbf{Y}^n, X_q^n] = \frac{K + E(\boldsymbol{\alpha}'\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)}{\nu + n - 2},$$

yielding the second expressions for $Q_{ij}^*(q)$.

Since $\Sigma_{1+p} \sim IW(\delta; Q_{1+p})$, $\Phi \sim IW(\nu; K)$,

$$\begin{aligned} E(\Sigma_{1+p}) &= \frac{Q_{1+p}}{\delta - 2}, \\ E(\Phi) &= \frac{K}{\nu - 2}, (\delta > 2, \nu > 2), \end{aligned}$$

are finite for fixed p, by the martingale convergence theorem, $E(\Sigma_{1+p} \mid \mathbf{Y}^n, X_q^n)$ and $E(\Phi \mid \mathbf{Y}^n, X_q^n)$ tend almost surely to $E(\Sigma_{1+p} \mid \mathbf{Y}^n, X^n)$, $E(\Phi \mid \mathbf{Y}^n, X^n)$ respectively as $q \to \infty$. This completes the proof. □.

Note that in the proof of Lemma 6.1 we use the independence property of the rows of Z given the parameters and the conditional independence of certain subsets of Z in the overall marginal distribution to obtain certain conditional expectations. These methods will be used in later sections.

6.2 Bayes and Bayes–Least Squares Estimators

For predicting the new response using the training data $\mathbf{Y}^n$, X_q^n and the observation $\mathbf{X}_p^0$ on the explanatory variables ($q \geq p$), the Bayes rule $f(\mathbf{X}_p^0, \mathbf{Y}^n, X_q^n)$ will mimimize the Bayes risk

$$E(Y^0 - f(\mathbf{X}_p^0, \mathbf{Y}^n, X_q^n))^2,$$

which is equivalent to minimizing the posterior expected loss

$$E[(Y^0 - f(\mathbf{X}_p^0, \mathbf{Y}^n, X_q^n))^2 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n]. \tag{6.13}$$

If the predictor is a linear function of $\mathbf{X}_p^0$, $(\mathbf{X}_p^0)'\mathbf{b}_p(\mathbf{Y}^n, X_q^n)$, where $\mathbf{b}_p(\mathbf{Y}^n, X_q^n)$ is a coefficient depending on $\mathbf{Y}^n, X_q^n$, then the Bayes risk and the posterior expected loss conditional on the training data will be

$$E(Y^0 - (\mathbf{X}_p^0)'\mathbf{b}_p(\mathbf{Y}^n, X_q^n))^2$$

and

$$E[(Y^0 - (\mathbf{X}_p^0)'\mathbf{b}_p(\mathbf{Y}^n, X_q^n))^2 \mid \mathbf{Y}^n, X_q^n] \tag{6.14}$$

respectively. In the conjugate case case (i.e. under the assumptions (5.1), (5.2)), the two problems of minimizing (6.13) and (6.14) are equivalent since by Lemma 3 and Lemma 4 in Dawid 1988,

$$\begin{aligned} & E(Y^0 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) = E(Y^0 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_p^n) \\ = \; & (\mathbf{X}_p^0)'[Q_{pp}^{-1}Q_{p0} + Q_{pp}^{-1}(X_p^n)'(I_n + X_p^n Q_{pp}^{-1}(X_p^n)')^{-1}(\mathbf{Y}^n - X_p^n Q_{pp}^{-1}Q_{p0})]. \end{aligned}$$

However, in the nonconjugate case of (6.1) and (6.3), this equivalence is not so obvious. To investigate this equivalence, we shall first give two lemmas on conditional independence under assumptions (6.1) and (6.3).

Lemma 6.2. Under the assumptions (6.1) and (6.3), $(\mathbf{X}_q^0, \Sigma_{qq})$ and $(\boldsymbol{\beta}_q, \Sigma_{00.q}, \Phi)$ are conditionally independent given $(\mathbf{Y}^n, X_q^n)$.

Proof. Let f, π denote the relevant density functions. Consider the sampling distribution. The assumption $\boldsymbol{\alpha} \perp\!\!\!\perp (\boldsymbol{\eta}, X_q^n)$ leads to $\boldsymbol{\eta} \perp\!\!\!\perp \boldsymbol{\alpha} \mid X_q^n$ so that

$$\mathbf{Y}^n \mid X_q^n \stackrel{\mathrm{d}}{=} (\boldsymbol{\eta} + \boldsymbol{\alpha}) \mid X_q^n \stackrel{\mathrm{d}}{=} (\boldsymbol{\eta} \mid X_q^n) + \boldsymbol{\alpha}$$

with

$$\boldsymbol{\eta} \perp\!\!\!\perp \boldsymbol{\alpha} \mid X_q^n.$$

Thus, by properties of the normal distribution,

$$\begin{aligned} & f(\mathbf{Y}^n, X_q^n \mid \Sigma_{1+q}, \Phi) \\ = \; & f(X_q^n \mid \Sigma_{1+q}, \Phi) f(\mathbf{Y}^n \mid X_q^n, \Sigma_{1+q}, \Phi) \\ = \; & f(X_q^n \mid \Sigma_{qq}) f(\mathbf{Y}^n \mid X_q^n, \boldsymbol{\beta}_q, \Sigma_{00.q}, \Phi). \end{aligned}$$

Also, for the inverse Wishart distribution,

$$\pi(\Sigma_{1+q}) = \pi(\Sigma_{qq}) \cdot \pi(\boldsymbol{\beta}_q, \Sigma_{00.q}).$$

Hence

$$\begin{aligned} & f(\Sigma_{1+q}, \Phi, \mathbf{X}_q^0 \mid \mathbf{Y}^n, X_q^n) \\ = \; & f(\mathbf{X}_q^0, \mathbf{Y}^n, X_q^n \mid \Sigma_{1+q}, \Phi) \pi(\Sigma_{1+q}, \Phi) / f(\mathbf{Y}^n, X_q^n) \\ = \; & f(\mathbf{X}_q^0 \mid \Sigma_{qq}) f(X_q^n \mid \Sigma_{qq}) \pi(\Sigma_{qq}) \\ & \times f(\mathbf{Y}^n \mid X_q^n, \boldsymbol{\beta}_q, \Sigma_{00.q}, \Phi) \pi(\boldsymbol{\beta}_q, \Sigma_{00.q}, \Phi) / f(\mathbf{Y}^n, X_q^n), \end{aligned}$$

which is the product of two factors depending on $(\mathbf{X}_q^0, \Sigma_{qq}, X_q^n)$ and $(\boldsymbol{\beta}_q, \Sigma_{00.q}, \Phi,$ $\mathbf{Y}^n, X_q^n)$ respectively. Hence by (1b) of Dawid (1979), p.3, $(\mathbf{X}_q^0, \Sigma_{qq}) \perp\!\!\!\perp (\boldsymbol{\beta}_q, \Sigma_{00.q}, \Phi) \mid$ $(\mathbf{Y}^n, X_q^n)$, completing the proof. □

Lemma 6.3. Under the assumptions (6.1) and (6.3), $\mathbf{X}_q^0$ and $\boldsymbol{\eta}$ are conditionally independent given $(\mathbf{Y}^n, X_q^n)$.

Proof. The assumptions (6.1) and (6.3) imply that $(\boldsymbol{\eta}, \mathbf{X}_q^0, X_q^n) \perp\!\!\!\perp \boldsymbol{\alpha}$ so that $(\boldsymbol{\eta}, \mathbf{X}_q^0) \perp\!\!\!\perp \boldsymbol{\alpha} \mid X_q^n$. By properties of the matrix-t distribution (cf. Dawid 1988, Lemma 4), $\boldsymbol{\eta} \perp\!\!\!\perp \mathbf{X}_q^0 \mid X_q^n$. Hence $\perp\!\!\!\perp \{\boldsymbol{\eta},\ \mathbf{X}_q^0,\ \boldsymbol{\alpha}\} \mid X_q^n$, i.e. $\boldsymbol{\eta}$, $\mathbf{X}_q^0$, $\boldsymbol{\alpha}$ are independent given X_q^n. Since $\mathbf{Y}^n = \boldsymbol{\eta} + \boldsymbol{\alpha}$, $(\mathbf{Y}^n, \boldsymbol{\eta}) \perp\!\!\!\perp \mathbf{X}_q^0 \mid X_q^n$, which leads to $\boldsymbol{\eta} \perp\!\!\!\perp \mathbf{X}_q^0 \mid (\mathbf{Y}^n, X_q^n)$, (cf. Lemma 4.2, Dawid, 1979). □

The solution minimizing (6.13) is

$$\hat{f}(\mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) \stackrel{\text{def}}{=} E(Y^0 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n).$$

By using the same argument as in the proof of Lemma 6.1, we have

$$\begin{aligned} E(Y^0 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) &= E[E(Y^0 \mid \mathbf{X}_p^0, \Sigma) \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n] \\ &= (\mathbf{X}_p^0)' E(\boldsymbol{\beta}_p \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n), \end{aligned} \tag{6.15}$$

and

$$\begin{aligned} & E(Y^0 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) \\ = \ & E[E(\eta_0 \mid \mathbf{X}_p^0, \boldsymbol{\eta}, X_q^n) + E(\alpha_0 \mid \boldsymbol{\alpha}) \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n] \\ = \ & E[E(\eta_0 \mid \mathbf{X}_p^0, \boldsymbol{\eta}, X_q^n) \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n]. \end{aligned} \tag{6.16}$$

Since by assumptions (6.1) and (6.3),

$$\begin{pmatrix} \eta_0 & (\mathbf{X}_q^0)' \\ \boldsymbol{\eta} & X_q^n \end{pmatrix} \sim T(\delta; I_{n+1}, Q_{1+q}),$$

by Lemma 4 in Dawid, 1988,

$$E(\eta_0 \mid \mathbf{X}_p^0, \boldsymbol{\eta}, X_q^n) = E(\eta_0 \mid \mathbf{X}_p^0, \boldsymbol{\eta}, X_p^n)$$

$$= (\mathbf{X}_p^0)'[Q_{pp}^{-1}Q_{p0} + Q_{pp}^{-1}(X_p^n)'(I_n + X_p^n Q_{pp}^{-1}(X_p^n)')^{-1}(\boldsymbol{\eta} - X_p^n Q_{pp}^{-1}Q_{p0})]$$
$$= (\mathbf{X}_p^0)'(Q_{pp} + (X_p^n)'X_p^n)^{-1}(Q_{p0} + (X_p^n)'\boldsymbol{\eta}).$$

Hence by Lemma 6.3, Lemma 6.1, (6.16) is equal to

$$(\mathbf{X}_p^0)'[Q_{pp}^{-1}Q_{p0} + Q_{pp}^{-1}(X_p^n)'(I_n + X_p^n Q_{pp}^{-1}(X_p^n)')^{-1}$$
$$\times(E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X_q^n) - X_p^n Q_{pp}^{-1}Q_{p0})] \tag{6.17}$$
$$= (\mathbf{X}_p^0)'(Q_{pp} + (X_p^n)'X_p^n)^{-1}[Q_{p0} + (X_p^n)'E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X_q^n)]$$
$$= (\mathbf{X}_p^0)'Q_{pp}^{*-1}Q_{p0}^*(q), \tag{6.18}$$

The corresponding minimum of (6.13) is

$$\mathrm{Var}(Y^0 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) = E((Y^0)^2 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) - [(\mathbf{X}_p^0)'E(\boldsymbol{\beta}_p \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n)]^2. \tag{6.19}$$

Let

$$R(\mathbf{b}_p) = Q_{00}^*(q) - 2Q_{0p}^*(q)\mathbf{b}_p + \mathbf{b}_p'Q_{pp}^*\mathbf{b}_p, \quad \mathbf{b}_p \in R^p, \tag{6.20}$$

where Q^* is given by Lemma 6.1. If $\mathbf{b}_p$ is a function of $\mathbf{Y}^n$, X_q^n, then

$$R(\mathbf{b}_p) = E((Y^0)^2 \mid \mathbf{Y}^n, X_q^n) - 2\mathbf{b}_p'E(\mathbf{X}_p^0 Y^0 \mid \mathbf{Y}^n, X_q^n) + \mathbf{b}_p'E(\mathbf{X}_p^0(\mathbf{X}_p^0)' \mid \mathbf{Y}^n, X_q^n)\mathbf{b}_p$$
$$= E[(Y^0 - \mathbf{b}_p'\mathbf{X}_p^0)^2 \mid \mathbf{Y}^n, X_q^n],$$

i.e. (6.14). The solution of (6.14) is given in the following proposition.

Proposition 6.1. Under the assumptions (6.1), (6.3), the Bayes solution which minimizes (6.14) is

$$\boldsymbol{\beta}_p^B(q)$$
$$= Q_{pp}^{*-1}Q_{p0}^*(q) \tag{6.21}$$
$$= (Q_{pp} + (X_p^n)'X_p^n)^{-1}[Q_{p0} + (X_p^n)'E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X_q^n)] \tag{6.22}$$
$$= E(\boldsymbol{\beta}_p \mid \boldsymbol{\eta}, X_q^n) + (Q_{pp} + (X_p^n)'X_p^n)^{-1}(X_p^n)'(\boldsymbol{\alpha} - E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)), \tag{6.23}$$
$$p \le q \le \infty,\ p < \infty.$$

The corresponding minimum is

$$R(\boldsymbol{\beta}_p^B(q)) = Q_{00.p}^*(q), \tag{6.24}$$

where Q^* is given by Lemma 6.1, R is defined by (6.20). Moreover,

$$\lim_{q\to\infty} \boldsymbol{\beta}_p^B(q),\ \lim_{q\to\infty} R(\boldsymbol{\beta}_p^B(q)) \text{ exist a.s. } (\delta > 2, \nu > 2), \tag{6.25}$$

and will be denoted by $\boldsymbol{\beta}_p^B(\infty)$, $R(\boldsymbol{\beta}_p^B(\infty))$ respectively.

Proof. As in the proof of (5.8) and (5.11) in Proposition 5.1, we obtain (6.21) and (6.24) with $Q^*(q)$ given by Lemma 6.1. Then (6.22) follows by the expressions (6.9), (6.12) for $Q_{p0}^*(q)$ and Q_{pp}^*. Note that in the conjugate case (Chapter 5) we use $Q_{1+p}^* = \mathrm{Var}\left(\begin{pmatrix} Y^0 \\ \mathbf{X}_p^0 \end{pmatrix} \mid \mathbf{Y}^n, X_p^n\right)(\delta + n - 2)$, but (5.6) ((5.8), (5.11) hold for Q_{1+p}^*, $R(\mathbf{b}_p)$ multiplied by a constant $(\delta + n - 2)^{-1}$. By (6.1), (6.3) and Proposition 5.1,

$$(Q_{pp} + (X_p^n)'X_p^n)^{-1}(Q_{p0} + (X_p^n)'\boldsymbol{\eta}) = E(\boldsymbol{\beta}_p \mid \boldsymbol{\eta}, X_p^n) = E(\boldsymbol{\beta}_p \mid \boldsymbol{\eta}, X_q^n).$$

Substituting the above equation and $\mathbf{Y}^n = \boldsymbol{\eta} + \boldsymbol{\alpha}$ into (6.22), we obtain

$$\begin{aligned} & \boldsymbol{\beta}_p^B(q) \\ = \ & (Q_{pp} + (X_p^n)'X_p^n)^{-1}[Q_{p0} + (X_p^n)'\mathbf{Y}^n - (X_p^n)'E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)] \\ = \ & (Q_{pp} + (X_p^n)'X_p^n)^{-1}[Q_{p0} + (X_p^n)'(\boldsymbol{\eta} + \boldsymbol{\alpha}) - (X_p^n)'E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)], \end{aligned}$$

which is equal to (6.23). The assertion (6.25) is obtained by Lemma 6.1 and (6.21), (6.24). □.

By Proposition 6.1 and (6.18) we conclude that the two problems of minimizing (6.13) and (6.14) are equivalent, i.e. $\hat{f} = (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B(q)$. In what follows we shall be mainly concerned with the predictor linear in $\mathbf{X}_p^0$ given $\mathbf{Y}^n$, X_q^n. In the same line as the conjugate case (Chapter 5), we wish to investigate the property of the full Bayes estimator and make comparison with the Bayes–least squares estimator.

Proposition 6.2. Under the assumptions (6.1) and (6.3) with $p > n$, a least squares estimator based on the training data $\mathbf{Y}^n$, X_p^n, satisfying $\mathbf{Y}^n = X_p^n\mathbf{b}_p$, $p >$

n, which further minimizes the posterior expected squared error loss (6.14) is

$$\begin{aligned}
&\boldsymbol{\beta}_p^L(q) \\
&= Q_{pp}^{*^{-1}} Q_{p0}^*(q) + Q_{pp}^{*^{-1}} (X_p^n)' (X_p^n Q_{pp}^{*^{-1}} (X_p^n)')^{-1} (\mathbf{Y}^n - X_p^n Q_{pp}^{*^{-1}} Q_{p0}^*(q)), \quad (6.26) \\
&\qquad p \le q \le \infty,\ p < \infty.
\end{aligned}$$

The corresponding minimum is

$$\begin{aligned}
&R(\boldsymbol{\beta}_p^L(q)) = Q_{00.p}^*(q) \\
&+ (\mathbf{Y}^n - X_p^n Q_{pp}^{*^{-1}} Q_{p0}^*(q))' (X_p^n Q_{pp}^{*^{-1}} (X_p^n)')^{-1} (\mathbf{Y}^n - X_p^n Q_{pp}^{*^{-1}} Q_{p0}^*(q)), \quad (6.27) \\
&\qquad p \le q \le \infty,\ p < \infty.
\end{aligned}$$

Moreover,

$$\lim_{q\to\infty} \boldsymbol{\beta}_p^L(q) = \boldsymbol{\beta}_p^L(\infty),\ \lim_{q\to\infty} R(\boldsymbol{\beta}_p^L(q)) = R(\boldsymbol{\beta}_p^L(\infty)) \text{ exist a.s.} \qquad (6.28)$$

Proof. Replacing Q^* defined in Lemma 5.1 by $Q^*(q)$ defined in Lemma 6.1 in the proof of Proposition 5.4, we obtain the Proposition. □

From Propositions 6.1 and 6.2 we know, if we obtain a training data set $\mathbf{Y}^n$, X_q^n for sufficiently large q, then the estimators $\boldsymbol{\beta}_p^B(q)$, $\boldsymbol{\beta}_p^L(q)$ are very close to $\boldsymbol{\beta}_p^B(\infty)$, $\boldsymbol{\beta}_p^L(\infty)$, the estimators based on the full training data set $\mathbf{Y}^n$, X^n. Thus we shall first fix a $q < \infty$ and take $p \le q$. Then we study the properties of the estimators as $q \to \infty$ and then $p \to \infty$. The result will also hold as $(p, q) \to (\infty, \infty), (p \le q)$.

6.3 A Property of the Bayes Estimator

The distributions of the estimators $\boldsymbol{\beta}_p^B(q)$, $\boldsymbol{\beta}_p^L(q)$ are of more complex form in the nonconjugate case than that in the conjugate case. We shall give lower bounds for

the posterior expected squared error loss conditioned on the observed explanatory variables $\mathbf{X}_p^0$ and the training data $\mathbf{Y}^n$, X_q^n, and conditioned on the training data $\mathbf{Y}^n$, X_q^n, $(p \le q \le \infty)$.

Proposition 6.3. Under the assumptions (6.1) and (6.3), the following hold for $p \le q < \infty$ or $p \le q = \infty$, $\nu > 2$, $\delta > 2$:

$$
\begin{aligned}
& E[(Y^0 - \hat{f}(\mathbf{X}_p^0, \mathbf{Y}^n, X_q^n))^2 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n] \\
= \; & (\mathbf{X}_p^0)' \mathrm{Var}(\boldsymbol{\beta}_p \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) \mathbf{X}_p^0 + E(\Sigma_{00.p} + \Phi \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) \\
\ge \; & E(\Phi \mid \mathbf{Y}^n, X_q^n). \qquad (6.29) \\
& R(\boldsymbol{\beta}_p^L(q)) \ge R(\boldsymbol{\beta}_p^B(q)) \\
= \; & E(\Sigma \mid \mathbf{Y}^n, X_q^n)_{00.p} + E(\Phi \mid \mathbf{Y}^n, X_q^n) \\
\ge \; & E(\Phi \mid \mathbf{Y}^n, X_q^n), \qquad (p > n). \qquad (6.30) \\
& E(\Phi \mid \mathbf{Y}^n, X_q^n) \ge \frac{K}{\nu - 2 + n}. \quad (\nu + n - 2 > 0). \qquad (6.31)
\end{aligned}
$$

Proof. Since by (6.2),

$$
\begin{aligned}
E((Y^0)^2 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) &= E[E((Y^0)^2 \mid \mathbf{X}_p^0, \Sigma, \Phi) \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n] \\
&= E[(\boldsymbol{\beta}_p' \mathbf{X}_p^0)^2 + \sigma^2 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n],
\end{aligned}
$$

we have by (6.19),

$$
\begin{aligned}
& E[(Y^0 - \hat{f}(\mathbf{X}_p^0, \mathbf{Y}^n, X_q^n))^2 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n] \\
= \; & (\mathbf{X}_p^0)' E(\boldsymbol{\beta}_p \boldsymbol{\beta}_p' \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) \mathbf{X}_p^0 + E(\sigma^2 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) \\
& - (\mathbf{X}_p^0)' E(\boldsymbol{\beta}_p \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) E(\boldsymbol{\beta}_p' \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) \mathbf{X}_p^0,
\end{aligned}
$$

which leads to (6.29) by Lemma 6.2.

By Proposition 6.1 and Lemma 6.1,

$$
R(\boldsymbol{\beta}_p^B(q)) = Q_{00.p}^*(q)
$$

$$
\begin{aligned}
&= E(\Sigma_{00} + \Phi \mid \mathbf{Y}^n, X_q^n) - E(\Sigma_{0p} \mid \mathbf{Y}^n, X_q^n) E(\Sigma_{pp} \mid \mathbf{Y}^n, X_q^n)^{-1} E(\Sigma_{p0} \mid \mathbf{Y}^n, X_q^n) \\
&= E(\Sigma_{00} \mid \mathbf{Y}^n, X_q^n) - E(\Sigma_{0p} \mid \mathbf{Y}^n, X_q^n) E(\Sigma_{pp} \mid \mathbf{Y}^n, X_q^n)^{-1} E(\Sigma_{p0} \mid \mathbf{Y}^n, X_q^n) \\
&\quad + E(\Phi \mid \mathbf{Y}^n, X_q^n),
\end{aligned}
$$

yielding (6.30).

The assumptions (6.1), (6.3) imply that

$$
(\boldsymbol{\eta}, X_q^n, \Sigma_{1+q}) \perp\!\!\!\perp (\Phi, \boldsymbol{\alpha})
$$

so that

$$
\Phi \perp\!\!\!\perp (\boldsymbol{\eta}, X_q^n, \Sigma_{1+q}) \mid \boldsymbol{\alpha}.
$$

Hence

$$
\begin{aligned}
E(\Phi \mid \mathbf{Y}^n, X_q^n) &= E[E(\Phi \mid \boldsymbol{\eta}, \boldsymbol{\alpha}, X_q^n) \mid \mathbf{Y}^n, X_q^n] \\
&= E[E(\Phi \mid \boldsymbol{\alpha}) \mid \mathbf{Y}^n, X_q^n] \\
&= E(\frac{\boldsymbol{\alpha}'\boldsymbol{\alpha} + K}{\nu + n - 2} \mid \mathbf{Y}^n, X_q^n) \\
&\geq \frac{K}{\nu + n - 2}, \text{ if } (\nu + n - 2) > 0,
\end{aligned}
$$

by the conjugate property of $\Phi \sim IW(\nu; K)$.

The above results hold for any q such that $p \leq q < \infty$. Since $\alpha_0 \sim T(\nu; 1, K)$, $\eta_0 \sim T(\delta; 1, \Sigma_{00})$, $\boldsymbol{\beta}_p \sim Q_{pp}^{-1} Q_{p0} + T(\delta + p; Q_{pp}^{-1}, Q_{00.p})$, $\Sigma_{00.p} \sim IW(\delta + p; Q_{00.p})$, $\Phi \sim IW(\nu; K)$, the following hold for fixed p:

$$
\begin{aligned}
&E|Y^0| \leq E(|\alpha_0| + |\eta_0|) = \frac{K}{\nu - 2} + \frac{\Sigma_{00}}{\delta - 2} < \infty, \\
&E(\boldsymbol{\beta}_p \boldsymbol{\beta}_p') = \frac{Q_{pp}^{-1} Q_{00.p}}{\delta + p - 2} + Q_{pp}^{-1} Q_{p0} Q_{0p} Q_{pp}^{-1}, \text{ finite} \\
&E(\Sigma_{00.p}) = \frac{Q_{00.p}}{\delta + p - 2} < \infty, \\
&E(\Phi) = \frac{K}{\nu - 2} < \infty,
\end{aligned}
$$

for $\delta > 2$, $\nu > 2$ and the bounds do not depend on q. Hence, by the martingale convergence theorem, $\hat{f}(\mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) = E(Y^0 \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n)$, $\mathrm{Var}(\boldsymbol{\beta}_p \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) = E(\boldsymbol{\beta}_p \boldsymbol{\beta}_p' \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) - E(\boldsymbol{\beta}_p \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n) E(\boldsymbol{\beta}_p' \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n)$, $E(\Sigma_{00.p} \mid \mathbf{X}_p^0, \mathbf{Y}^n, X_q^n)$, $E(\Phi \mid \mathbf{Y}^n, X_q^n)$ tend almost surely to certain limits as

$q \to \infty$. Also the existence of the a.s. limits of $R(\boldsymbol{\beta}_p^B(q))$, $R(\boldsymbol{\beta}_p^L(q))$ were shown in Propositions 6.1 and 6.2, so that by the equality in (6.30) the a.s. limit of $E(\Sigma \mid \mathbf{Y}^n, X_q^n)_{00.p}$ exists. Letting $q \to \infty$ in (6.29)–(6.31), these equations also hold for $q = \infty$. Furthermore, these equations hold if taking limits $\liminf_{p\to\infty} \lim_{q\to\infty}$ on both sides. □

From (6.31) in Proposition 6.3, we have

$$\lim_q E(\Phi \mid \mathbf{Y}^n, X_q^n) > 0, \quad \text{a.s.} .$$

Hence by (6.30),

$$\liminf_p R(\boldsymbol{\beta}_p^B(\infty)) > 0,$$

which shows that the Bayes estimator $\boldsymbol{\beta}_p^B(q)$ in the nonconjugate case, unlike in the conjugate case, will not be expected to incorporate deterministic predictability of the response. The fact

$$\lim_q E(\Phi \mid \mathbf{Y}^n, X_q^n) > 0, \text{ a.s. } \nu > 2,$$

can be also deduced from the martingale argument as follows. Since $\Phi^{-1} \sim W(\nu; K^{-1})$ with $E(\Phi^{-1}) = \nu K^{-1} \in (0, \infty)$, we can define a nonnegative martingale $\{\mu_q\}$ with finite expectation by

$$\mu_q = E(\Phi^{-1} \mid \mathbf{Y}^n, X_q^n).$$

Then μ_q tends to a r.v $\mu < \infty$ a.s. as $q \to \infty$. By Jensen's inequality,

$$E(\Phi^{-1} \mid \mathbf{Y}^n, X_q^n) \geq E(\Phi \mid \mathbf{Y}^n, X_q^n)^{-1}, \ (\nu > 2).$$

Hence

$$E(\Phi \mid \mathbf{Y}^n, X_q^n) \geq \mu_q^{-1}.$$

Thus $E(\Phi \mid \mathbf{Y}^n, X_q^n)$ converges to a positive limit a.s. as $q \to \infty$.

6.4 Comparison

In this section we shall make comparison between the Bayes estimator $\boldsymbol{\beta}_p^B(q)$ and the least squares estimator $\boldsymbol{\beta}_p^L(q)$ under the nonconjugate prior for the parameters. The difference between the conjugate and the nonconjugate prior is determined by the error $\boldsymbol{\alpha}$. We shall first give a lemma on the properties of $\boldsymbol{\alpha}$.

Lemma 6.4. Under the assumptions (6.1) and (6.3), the following hold:

(a) If $\nu > 1$, then $\boldsymbol{\mu} \stackrel{\text{def}}{=} \lim_{q\to\infty} E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n) = E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X^n)$ exists a.s.

(b) Let $S_\infty = \|\boldsymbol{\mu}\|^2$. If $\nu > 2$, then

$$E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n) \xrightarrow{\mathcal{L}_2} \boldsymbol{\mu}, \quad (q \to \infty), \tag{6.32}$$

$$(\text{i.e. } \|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n) - \boldsymbol{\mu}\|^2 \to 0),$$

$$\|E(\boldsymbol{\alpha}|\mathbf{Y}^n, X_q^n)\|^2 \xrightarrow{\mathcal{L}_1} S_\infty, \quad (q \to \infty), \tag{6.33}$$

$$\lim_{q\to\infty} E(\|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)\|^2) = E(S_\infty) < \infty, \tag{6.34}$$

Proof. (a) Since $\alpha_i \sim T(\nu; 1, K)$, $(i = 1, 2, \cdots, n)$,

$$E|\alpha_i|^r \propto \int |\alpha_i|^r (K + \alpha_i^2)^{-\frac{\nu+1}{2}}\, d\alpha_i.$$

Thus $E|\alpha_i|^r$ exists iff $-1 < r < 0, \nu > r$ or $\nu > r > 0$. For $\nu > 1$, $\{E(\alpha_i|\mathbf{Y}^n, X_q^n)$, $q = 1, 2, \cdots, \}$ is a $\mathcal{L}_1$ martingale so that $\mu_i = \lim_{q\to\infty} E(\alpha_i \mid \mathbf{Y}^n, X_q^n)$ exists a.s.

(b) Since $\boldsymbol{\alpha}'\boldsymbol{\alpha} \sim F(n, \nu; K)$, $E(\boldsymbol{\alpha}'\boldsymbol{\alpha}) = \frac{nK}{\nu-2}$, $(\nu > 2)$ by Lemma 1.1. Also, by Jensen's inequality,

$$\|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)\|^2 \le E(\|\boldsymbol{\alpha}\|^2 \mid \mathbf{Y}^n, X_q^n).$$

Thus $\{\|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)\|^2, q = 1, 2, \cdots\}$ is a submartingale closed by $\|\boldsymbol{\alpha}\|^2$ and is u.i. (uniformly integrable). By Lemma 6.5 (b) below, $\{(E(\alpha_i \mid \mathbf{Y}^n, X_q^n))^2, q = 1, 2, \cdots\}$ is u.i. Hence by (a) and the Mean Convergence Theorem and its Corollary

(Theorem 4.2.3 , Corollary 4.2.5, Chow and Teicher, 1978) $E(\alpha_i \mid \mathbf{Y}^n, X_q^n) \xrightarrow{\mathcal{L}_2} \mu_i$, with $E(E(\alpha_i \mid \mathbf{Y}^n, X_q^n))^2 \to E\mu_i^2$, $q \to \infty$, $(i = 1, \cdots, n)$. Thus (6.32) holds by Lemma 6.5 (a). Since

$$E(\|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n\|^2) = \sum_{i=1}^{n} E(E(\alpha_i \mid \mathbf{Y}^n, X_q^n))^2,$$

(6.34) holds. Since $\|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)\|^2$, $\|\boldsymbol{\mu}\|^2$ are nonnegative, by Corollary 4.2.4 of Chow and Teicher (1978), we establish (6.33). □

The following lemma is on the relation of $\mathcal{L}_2$ convergence and u.i between a random vector sequence and its components.

Lemma 6.5 Let $\mathbf{X}, \mathbf{X}_n, n = 1, 2, \cdots$ be random vectors in R^m. Denote by $X_{(i)}$, $X_{n,(i)}$ their ith components.

(a) If $\mathbf{X}_n$, $n = 1, 2 \cdots$, are $\mathcal{L}_2$ random vectors ($E\|\mathbf{X}_n\|^2 < \infty$), then $\mathbf{X}_n \xrightarrow{\mathcal{L}_2} \mathbf{X}$ iff $X_{n,(i)} \xrightarrow{\mathcal{L}_2} X_i$, $(n \to \infty)$.

(b)$\{\|\mathbf{X}_n\|^2, n = 1, 2, \cdots\}$ is u.i. iff $\{X_{n,(i)}^2, n = 1, 2, \cdots\}$ is u.i., $(i = 1, \cdots, m)$.

Proof. The Lemma is established by the following two equations.

$$E\|\mathbf{X}_n - \mathbf{X}\|^2 = \sum_{i=1}^{m} E|X_{n,(i)} - X_{(i)}|^2,$$

$$\sup_n E(X_{n,(i)}^2 I_A) \le \sup_n E(\|\mathbf{X}_n\|^2 I_A) \le \sum_{i=1}^{m} \sup_n E(X_{n,(i)}^2 I_A),$$

for any measurable set A.

□

Next we shall investigate the behaviour of the error sum of squares of the Bayes estimator $\boldsymbol{\beta}_p^B(q)$.

Theorem 6.1. Under the assumptions (6.1) and (6.3) with $p \geq n$, the following hold:

(a) If $\nu > 1$, then

$$\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(\infty) \xrightarrow{P} \boldsymbol{\mu} = E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X^n), \quad \text{as } p \to \infty.$$

(b) If $\nu > 2$, then

$$\begin{aligned}
&\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(\infty) \xrightarrow{\mathcal{L}_2} \boldsymbol{\mu}, \\
&\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(\infty)\|^2 \xrightarrow{\mathcal{L}_1} S_\infty = \|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X^n)\|^2, \\
&E\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(\infty)\|^2 \to E(S_\infty), \quad \text{as } p \to \infty.
\end{aligned}$$

(c) If $\delta > 2$, $\nu > 2$, then

$$E(S_\infty) = n\left(\frac{K}{\nu - 2} - \frac{Q_{00}}{\delta - 2}\right) + E\|E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X^n)\|^2.$$

Proof. (a) Let $q \geq p$. By considering $(\eta, \mathbf{X}) \sim \mathcal{N}(1, \Sigma)$ with $\Sigma \sim IW(\delta; Q)$, we can apply the result in Section 1 of Chapter 5 to obtain $E(\boldsymbol{\beta}_p \mid \boldsymbol{\eta}, X_q^n) = E(\boldsymbol{\beta}_p \mid \boldsymbol{\eta}, X_p^n)$. By Proposition 6.1, we represent $\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q)$ as follows:

$$\begin{aligned}
& \mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q) \qquad (6.35) \\
= \;& \boldsymbol{\eta} + \boldsymbol{\alpha} - X_p^n \{E(\boldsymbol{\beta}_p \mid \boldsymbol{\eta}, X_q^n) + [Q_{pp} + (X_p^n)' X_p^n]^{-1} (X_p^n)' (\boldsymbol{\alpha} - E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n))\} \\
= \;& \boldsymbol{\eta} - X_p^n E(\boldsymbol{\beta}_p \mid \boldsymbol{\eta}, X_q^n) + \{I_n - X_p^n [Q_{pp} + (X_p^n)' X_p^n]^{-1} (X_p^n)'\} \boldsymbol{\alpha} \\
& + X_p^n [Q_{pp} + (X_p^n)' X_p^n]^{-1} (X_p^n)' E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n) \\
= \;& [\boldsymbol{\eta} - X_p^n E(\boldsymbol{\beta}_p \mid \boldsymbol{\eta}, X_p^n)] + \{I_n - X_p^n [Q_{pp} + (X_p^n)' X_p^n]^{-1} (X_p^n)'\} \boldsymbol{\alpha} \\
& + X_p^n [Q_{pp} + (X_p^n)' X_p^n]^{-1} (X_p^n)' E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n) \\
\stackrel{\text{def}}{=} \;& J_1(p) + J_2(p) + J_3(p, q). \qquad (6.36)
\end{aligned}$$

Let $q \to \infty$. By Proposition 6.1 and Lemma 6.4, (6.35) tends a.s. to

$$\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(\infty)$$

and (6.36) tends a.s. to

$$J_1(p) + J_2(p) + J_3(p, \infty),$$

where $J_3(p, \infty)$ is obtained by $J_3(p, q)$ with $E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)$ replaced by $E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X^n) = \boldsymbol{\mu}$. Now let $p \to \infty$. By Theorem 5.3, $J_1(p) \xrightarrow{\mathcal{L}_2} 0$. By Lemma 1.3, Lemma 1.5 and Lemma 1.6,

$$\begin{aligned} & I_n - X_p^n[Q_{pp} + (X_p^n)'X_p^n]^{-1}(X_p^n)' \\ = \; & [I_n + X_p^n Q_{pp}^{-1}(X_p^n)']^{-1} \\ \sim \; & B(\delta + n - 1, p; I_n) \xrightarrow{P} 0. \end{aligned}$$

Hence $J_2(p) \xrightarrow{P} 0$ and $J_3(p, \infty) \xrightarrow{P} E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X^n) = \boldsymbol{\mu}$ as $p \to \infty$. This completes the proof of (a).

(b) We have, by Theorem 5.3, $E\|J_1(p)\|^2 \to 0$, as $p \to \infty$. Since $\sup_p \|J_2(p)\|^2 \leq \|\boldsymbol{\alpha}\|^2$, $E\|\boldsymbol{\alpha}\|^2 = \frac{nK}{\nu - 2} < \infty$ $(\nu > 2)$, by Lebesgue's Dominated Convergence Theorem and the proof of (a),

$$\lim_p E\|J_2(p)\|^2 = E \lim_p \|J_2(p)\|^2 = 0.$$

Also $\|J_3(p, \infty)\|^2 \leq \|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X^n)\|^2 \leq E(\|\boldsymbol{\alpha}\|^2 \mid \mathbf{Y}^n, X^n)$. Hence $\{\|J_3(p, \infty)\|^2, p = 1, \cdots\}$ is u.i. By Mean Convergence Criterion and its Corollary (Theorem 4.2.3, Corollary 4.2.5, Chow and Theicher, 1978) and Lemma 6.5,

$$\begin{aligned} & J_3(p, \infty) \xrightarrow{\mathcal{L}_2} \boldsymbol{\mu}, \\ & \lim_p E\|J_3(p, \infty)\|^2 = E \lim_p \|J_3(p, \infty)\|^2 = ES_\infty. \end{aligned}$$

Applying Hölder's Inequality, we obtain

$$\begin{aligned} & E\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(\infty) - \boldsymbol{\mu}\|^2 \to 0 \\ & \lim_p E\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(\infty)\|^2 = E\|\boldsymbol{\mu}\|^2 = ES_\infty, \end{aligned}$$

and thus by Corollary 4.2.4 of Chow and Teicher (1978),

$$\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(\infty)\|^2 \xrightarrow{\mathcal{L}_1} S_\infty.$$

(c) By assumption (6.1),

$$E(\boldsymbol{\alpha}'\boldsymbol{\eta}) = E(\boldsymbol{\alpha}')E(\boldsymbol{\eta}) = 0,$$
$$E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X^n) = \mathbf{Y}^n - E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X^n).$$

Hence

$$\begin{aligned} ES_\infty &= E\|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X^n)\|^2 \\ &= E\|\mathbf{Y}^n\|^2 - 2E[(\mathbf{Y}^n)'E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X^n)] + E\|E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X^n)\|^2 \\ &= E\|\boldsymbol{\alpha}\|^2 + 2E\boldsymbol{\alpha}'E\boldsymbol{\eta} + E\|\boldsymbol{\eta}\|^2 - 2E[(\mathbf{Y}^n)'\boldsymbol{\eta}] + E\|E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X^n)\|^2 \\ &= E\|\boldsymbol{\alpha}\|^2 + E\|\boldsymbol{\eta}\|^2 - 2E\|\boldsymbol{\eta}\|^2 + E\|E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X^n)\|^2 \\ &= E\|\boldsymbol{\alpha}\|^2 - E\|\boldsymbol{\eta}\|^2 + E\|E(\boldsymbol{\eta} \mid \mathbf{Y}^n, X^n)\|^2. \end{aligned}$$

Since $\boldsymbol{\alpha}'\boldsymbol{\alpha} \sim F(n, \nu; K)$, $\boldsymbol{\eta}'\boldsymbol{\eta} \sim F(n, \delta; Q_{00})$, we have $E\|\boldsymbol{\alpha}\|^2 = \frac{nK}{\nu-2}$ and $E\|\boldsymbol{\eta}\|^2 = \frac{Q_{00}}{\delta-2}$, which, combined with the above equation, establish (c). □

Theorem 6.1 shows that, unlike in the conjugate case (cf. Theorem 5.3), the Bayes estimator $\boldsymbol{\beta}_p^B(q)$ is quite different from the Bayes–least squares estimator $\boldsymbol{\beta}_p^L(q)$ in the behaviour of the error sum of squares. In particular, if $\frac{K}{\nu-2} > \frac{Q_{00}}{\delta-2}$, we have $\lim_p E\|\mathbf{Y}^n - X_p^n\boldsymbol{\beta}_p^B(\infty)\|^2 > 0$ in contrast to $\mathbf{Y}^n - X_p^n\boldsymbol{\beta}_p^L(q) = 0$, $p \le q \le \infty$. The condition that $\frac{K}{\nu-2} > \frac{Q_{00}}{\delta-2}$ may be interpreted as there being a sufficiently large random error $\boldsymbol{\alpha}$, which makes our prior differ from the conjugate case considered in Chapter 5. The next theorem investigates the difference of the posterior expected squared error loss between the two estimators.

Theorem 6.2. Let $R(p,q) = R(\boldsymbol{\beta}_p^L(q)) - R(\boldsymbol{\beta}_p^B(q))$, $(p \le q \le \infty)$ with $R(\boldsymbol{\beta}_p^L(q))$, $R(\boldsymbol{\beta}_p^B(q))$ defined in Propositions 6.1 and 6.2. Then under the assumptions (6.1) and (6.3) with $p > n$,

$$R(p,q) = (\boldsymbol{\beta}_p^L(q) - \boldsymbol{\beta}_p^B(q))'Q_{pp}^*(\boldsymbol{\beta}_p^L(q) - \boldsymbol{\beta}_p^B(q)) \tag{6.37}$$
$$= (\mathbf{Y}^n - X_p^n\boldsymbol{\beta}_p^B(q))'[X_p^nQ_{pp}^{*-1}(X_p^n)']^{-1}(\mathbf{Y}^n - X_p^n\boldsymbol{\beta}_p^B(q)). \tag{6.38}$$

Moreover, as $p \to \infty$,

$$R(p,\infty) \xrightarrow{\mathcal{L}_1} S_\infty/(\delta+n-2), \tag{6.39}$$

$$E[R(p,\infty)] \to E(S_\infty)/(\delta+n-2)\ (\nu>2), \tag{6.40}$$

where S_∞ is defined in Lemma 6.4.

Proof. Write $R(\mathbf{b}_p)$ defined in (6.20) as

$$R(\mathbf{b}_p) = (\mathbf{b}_p - Q_{pp}^{*-1}Q_{p0}^*(q))'Q_{pp}^*(\mathbf{b}_p - Q_{pp}^{*-1}Q_{p0}^*(q)) + Q_{00.p}^*(q). \tag{6.41}$$

Letting $\mathbf{b}_p = \boldsymbol{\beta}_p^L(q)$, by (6.21) and (6.24) in Proposition 6.1, we establish (6.37). The next equation (6.38) follows from (6.24) and (6.27). Now consider $q=\infty$. As in the proof of (a) of Theorem 6.1, we have

$$[X_p^n Q_{pp}^{*-1}(X_p^n)']^{-1} \xrightarrow{P} I_n/(\delta+n-2),\ p\to\infty, \tag{6.42}$$

and

$$\sup_p [X_p^n Q_{pp}^{*-1}(X_p^n)']^{-1} \le I_n/(\delta+n-2). \tag{6.43}$$

By (6.38) and (a) of Theorem 6.1, we have

$$R(p,\infty) \xrightarrow{P} S_\infty/(\delta+n-2),\ (\nu>1). \tag{6.44}$$

Note that, in the proof (b) of Theorem 6.1, we can further deduce that $\{\|\mathbf{Y}^n - X_p^n\boldsymbol{\beta}_p^B(\infty)\|^2,\ p=1,2,\cdots\}$ is u.i., and thus, by (6.43), $\{R(p,\infty), p=1,2,\cdots\}$ is u.i. By Corollary 4.2.4 of Chow and Teicher, 1978, (6.39) and (6.40) hold. □

Theorem 6.3. Under the assumptions (6.1) and (6.3) with $p>n$, the conditional distribution of the difference of the two predictors $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L(q) - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B(q)$ $p\le q\le\infty$ given the training data is

$$(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L(q) - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B(q) \mid \mathbf{Y}^n, X_q^n \sim T(\delta+n; 1, R(p,q)\cdot(\delta+n-2)), \tag{6.45}$$

where $R(p,q)$ is given by Theorem 6.2. Moreover, as $p \to \infty$, $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L(\infty) - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B(\infty)$ converges in probability to a random variable, say Z, with the distribution given by

$$\begin{aligned} Z \mid S_\infty &\sim T(\delta+n;1,S_\infty), \\ S_\infty &\stackrel{\mathrm{d}}{=} \|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X^n)\|^2 \ (\nu > 1). \end{aligned} \tag{6.46}$$

Proof. The assumptions (6.1) and (6.3) imply $(\eta_0, \mathbf{X}_p^0, \boldsymbol{\eta}, X_q^n) \perp\!\!\!\perp \boldsymbol{\alpha}$, so that $\mathbf{X}_p^0 \perp\!\!\!\perp \boldsymbol{\alpha} \mid \boldsymbol{\eta}, X_q^n$ and thus

$$\begin{aligned} &\mathbf{X}_p^0 \mid \boldsymbol{\eta}, X_q^n, \boldsymbol{\alpha} = \mathbf{X}_p^0 \mid \boldsymbol{\eta}, X_q^n = \mathbf{X}_p^0 \mid \boldsymbol{\eta}, X_p^n \\ \sim\ & T(\delta+n; Q_{pp} + (X_p^n)'X_p^n, 1) = T(\delta+n; Q_{pp}^* \cdot (\delta+n-2), 1) \end{aligned}$$

by Lemma 5.1. By Propositions 6.1 and 6.2,

$$(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L(q) - (\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B(q) = (\mathbf{X}_p^0)'Q_{pp}^{*-1}(X_p^n)'[X_p^n Q_{pp}^{*-1}(X_p^n)']^{-1}(\mathbf{Y}^n - X_p^n\boldsymbol{\beta}_p^B(q)),$$

whence

$$(\mathbf{X}_p^0)'(\boldsymbol{\beta}_p^L(q) - \boldsymbol{\beta}_p^B(q)) \mid \boldsymbol{\eta}, X_q^n, \boldsymbol{\alpha} \sim T(\delta+n; 1, (\delta+n-2)R(p,q))$$

by Theorem 6.2. Since $R(p,q)$ depend on $\boldsymbol{\eta}, \boldsymbol{\alpha}$ through $\mathbf{Y}^n = \boldsymbol{\eta} + \boldsymbol{\alpha}$ only, (6.45) obtains. The asymptotic distribution is obtained by Lemma 1.4 and Theorem 6.2 for $\nu > 1$. □

We have

$$\mathrm{Var}(Z) = E[E(Z^2 \mid S_\infty)] = E(S_\infty)/(\delta+n-2).$$

If $K/(\nu-2) - Q_{00}/(\delta-2) > 0$, $\nu > 2$, $\delta > 2$, then $\mathrm{Var}(Z) > 0$ so that Z is a nondegenerate random variable. Thus, unlike in the conjugate case (c.f. Theorem 5.2, eq.(5.27)), for large p, the two predictors $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L(\infty)$ and $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B(\infty)$ will not be always identical.

We have investigated the case that $p \to \infty$, $q = \infty$. We shall show that these limits are identical to the corresponding double limits as $(p, q) \to (\infty, \infty)$, $(p \leq q)$. That is to say, the asymptotic properties discussed do not depend on the way how we increase the numbers of explanatory variables in the training data and observation to be predicted, p and q, so long as $p \leq q$. Now consider $(p, q) \to (\infty, \infty)$ with $p \leq q$. The following hold:

$$\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q) \xrightarrow{P} \boldsymbol{\mu} \ (\nu > 1), \tag{6.47}$$

$$E\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q) - \boldsymbol{\mu}\|^2 \to 0 \ (\nu > 2), \tag{6.48}$$

$$\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q)\|^2 \xrightarrow{\mathcal{L}_1} S_\infty \ (\nu > 2), \tag{6.49}$$

$$E\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q)\|^2 \to ES_\infty \ (\nu > 2), \tag{6.50}$$

$$R(p, q) \xrightarrow{P} S_\infty/(\delta + n - 2) \ (\nu > 1), \tag{6.51}$$

$$R(p, q) \xrightarrow{\mathcal{L}_1} S_\infty/(\delta + n - 2) \ (\nu > 2), \tag{6.52}$$

$$ER(p, q) \to ES_\infty/(\delta + n - 2) \ (\nu > 2), \tag{6.53}$$

$$(\mathbf{X}_p^0)'(\boldsymbol{\beta}_p^L(q) - \boldsymbol{\beta}_p^B(q)) \xrightarrow{P} Z \ (\nu > 1), \tag{6.54}$$

where Z is distributed as (6.46).

Proof. To prove (6.47), consider the equality (6.35)=(6.36) in the proof of Theorem 6.1. $J_3(p, q)$ is the product of two factors, $X_p^n[Q_{pp} + (X_p^n)'X_p^n]^{-1}(X_p^n)'$ and $E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)$. The first depends on p only and converges in probability to I_n as $p \to \infty$ and hence as $(p, q) \to \infty$. The second depends on q only and converges a.s. to $\boldsymbol{\mu}$ as $q \to \infty$ (cf. Lemma 6.4) and hence as $(p, q) \to \infty$. Thus $J_3(p, q) \xrightarrow{P} \boldsymbol{\mu}$ as $(p, q) \to \infty$. Also $J_1(p)$, $J_2(p)$ converges in probability to zero as $p \to \infty$ and hence as $(p, q) \to (\infty, \infty)$. Since the equation (6.35)=(6.36) holds for $p \leq q$, (6.47) holds.

To prove (6.48)–(6.50) , since

$$0 \leq X_p^n[Q_{pp} + (X_p^n)'X_p^n]^{-1}X_p^n \leq I_n, \ (\text{all } p),$$

we have

$$\begin{aligned}
&\sup_{p,q} E\{\|J_3(p,q)\|^2 I_A\} \\
&\leq \sup_q \sup_p E\{\|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)\|^2 I_A\} \\
&\leq \sup_q E\{\|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)\|^2 I_A\}, \text{ for any } A.
\end{aligned}$$

Thus $\{\|E(\boldsymbol{\alpha} \mid \mathbf{Y}^n, X_q^n)\|^2, q = 1, 2, \cdots\}$ is u.i. (cf. Lemma 6.4) which implies that $\{\|J_3(p,q)\|^2, p, q = 1, 2, \cdots\}$ is u.i. , (replacing the one-dimensional index of a sequence of random variables by two dimensional index in the definition of u.i.). Since $J_1(p)$, $J_2(p)$ depend on p only, their properties as $p \to \infty$ remains the same as $(p,q) \to (\infty, \infty)$. The rest of the proof is similar to the proof of Theorem 6.1 (b) by replacing $p \to \infty$, $q = \infty$ by $(p,q) \to \infty$ $(p \leq q)$ and substituting the identity of (6.35)=(6.36) $(p \leq q)$ into

$$\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q) - \boldsymbol{\mu}\|^2, \ \|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q)\|^2, \ E\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q)\|^2.$$

(Remark. Furthermore, $\{\|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q)\|^2, p, q = 1, 2, \cdots\}$ is u.i. by Theorem 4.2.3 (Chow and Teicher, 1978)).

By (6.47), $\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q) \xrightarrow{P} \boldsymbol{\mu}$ as $(p,q) \to \infty$, $(p \leq q)$. Also $[X_p^n (Q_{pp}^*)^{-1} (X_p^n)']^{-1}$ $\xrightarrow{P} I_n/(\delta + n - 2)$ as $p \to \infty$. By (6.38), $R(p,q)$ is the product of three factors which have limits in probability shown above. Hence

$$R(p,q) \xrightarrow{P} \boldsymbol{\mu}' \cdot I_n/(\delta + n - 2) \cdot \boldsymbol{\mu} = S_\infty/(\delta + n - 2), \ (p,q) \to (\infty, \infty), \ (p \leq q).$$

This establishes (6.51). By (6.43),

$$R(p,q) \leq \|\mathbf{Y}^n - X_p^n \boldsymbol{\beta}_p^B(q)\|^2,$$

where the right hand side is u.i. by the remark above. Hence $\{R(p,q), p, q = 1, 2, \cdots\}$ is u.i., which combined with (6.51) establishes (6.52) and (6.53) by Theorem 4.2.3 and Corollary 4.2.5 of Chow and Theicher (1978).

It is obvious that Lemma 1.4 holds for two dimensional index of the sequence Z_p, A_p. Hence by (6.51) and (6.45) we obtain the desired asymptotic distribution as $(p,q) \to (\infty, \infty)$, $(p \leq q)$ in (6.54) for $(\mathbf{X}_p^0)'(\boldsymbol{\beta}_p^L(q) - \boldsymbol{\beta}_p^B(q))$. □

6.5 Discussion

In the nonconjugate case, the parameter to be minimised is

$$\phi_{\mathbf{b}_p} = E(Y - \mathbf{X}_p'\mathbf{b}_p)^2 = \Phi + \Sigma_{00} - 2\Sigma_{0p}\mathbf{b}_p + \mathbf{b}_p'\Sigma_{pp}\mathbf{b}_p.$$

The minimum is then

$$\phi^{**} = \phi_{\mathbf{b}_p^{**}} = \inf_{\mathbf{b}_p}\{\phi_{\mathbf{b}_p}\} = \Phi + \Sigma_{00.p} = \Phi + \Gamma_p$$

with

$$\mathbf{b}_p^{**} = \Sigma_{pp}^{-1}\Sigma_{p0} = \boldsymbol{\beta}_p.$$

An unbiased estimate from the training data $(\mathbf{Y}^n, X_p^n)$ for $\phi_{\mathbf{b}_p}$ is

$$X_{\mathbf{b}_p} = \|\mathbf{Y}^n - X_p^n\mathbf{b}_p\|^2/n.$$

If $p \leq n$, the minimum of $X_{\mathbf{b}_p}$ is

$$X^* = X_{\mathbf{b}_p^*} = S_{00.p}/n,$$

where S is given by (5.32),

$$\mathbf{b}_p^* = S_{pp}^{-1}S_{p0}.$$

Now $\mathbf{Y}^n = \boldsymbol{\eta} + \boldsymbol{\alpha}$, hence

$$\begin{aligned} S_{00.p} &= S_{00} - S_{0p}S_{pp}^{-1}S_{p0} \\ &= [\boldsymbol{\eta}'\boldsymbol{\eta} - \boldsymbol{\eta}'X_p^n((X_p^n)'X_p^n)^{-1}(X_p^n)'\boldsymbol{\eta}] + [2\boldsymbol{\eta}'\boldsymbol{\alpha} - 2\boldsymbol{\eta}'X_p^n((X_p^n)'X_p^n)^{-1}(X_p^n)'\boldsymbol{\alpha}] \\ &\quad +[\boldsymbol{\alpha}'\boldsymbol{\alpha} - \boldsymbol{\alpha}'X_p^n((X_p^n)'X_p^n)^{-1}(X_p^n)'\boldsymbol{\alpha}] \\ &\stackrel{\text{def}}{=} J_1 + J_2 + J_3. \end{aligned}$$

Since $(\boldsymbol{\eta}\ X_p^n) \sim \mathcal{N}(I_n, \Sigma_{1+p})$,

$$J_1 \sim W_1(n-p; \Sigma_{00.p}),\ E(J_1) = (n-p)\Sigma_{00.p} = (n-p)\Gamma_p.$$

By the independence property of $\boldsymbol{\alpha}$ and $(\boldsymbol{\eta}, X^n)$,

$$E(J_2) = 0.$$

Since $\boldsymbol{\alpha} \sim \mathcal{N}(I_n, \Phi), E(\boldsymbol{\alpha}\boldsymbol{\alpha}') = \Phi I_n$. Hence

$$\begin{aligned} E(J_3) &= E\{\mathrm{tr}[\boldsymbol{\alpha}'(I_n - X_p^n((X_p^n)'X_p^n)^{-1}(X_p^n)')\boldsymbol{\alpha}]\} \\ &= \mathrm{tr}E[\boldsymbol{\alpha}\boldsymbol{\alpha}'(I_n - X_p^n((X_p^n)'X_p^n)^{-1}(X_p^n)')] \\ &= \mathrm{tr}E(\boldsymbol{\alpha}\boldsymbol{\alpha}')E(I_n - X_p^n((X_p^n)'X_p^n)^{-1}(X_p^n)') \\ &= \Phi\mathrm{tr}E(I_n - X_p^n((X_p^n)'X_p^n)^{-1}(X_p^n)') \\ &= \Phi E\mathrm{tr}(I_n - X_p^n((X_p^n)'X_p^n)^{-1}(X_p^n)') \\ &= (n-p)\Phi. \end{aligned}$$

The above calculation leads to

$$E(X^*) = \frac{n-p}{n}(\Phi + \Sigma_{00.p}) = \frac{n-p}{n}\phi^{**} < \phi^{**},$$

which shows that X^* is negatively biased for ϕ^{**}. If $p > n$, we take

$$X^* = X_{\boldsymbol{\beta}_p^L(q)} = \|\mathbf{Y}^n - X_p^n\boldsymbol{\beta}_p^L(q)\|^2/n = 0,$$

with $\mathbf{b}_p^* = \boldsymbol{\beta}_p^L(q)$, which is negatively biased for ϕ^{**}, achieving the lowest possible value. The "data-dependent parameter" ϕ^* is now

$$\phi^* = \phi_{\mathbf{b}_p^*} = \Phi + \Sigma_{00} - 2\Sigma_{0p}\boldsymbol{\beta}_p^L(q) + (\boldsymbol{\beta}_p^L(q))'\Sigma_{pp}\boldsymbol{\beta}_p^L(q).$$

To conduct Bayesian analysis, let

$$\begin{aligned} Y_{\mathbf{b}_p} &= E(\phi_{\mathbf{b}_p} \mid \mathbf{Y}^n, X_q^n) \\ &= (Q_{00}^*(q) - 2Q_{0p}^*(q)\mathbf{b}_p + \mathbf{b}_p'Q_{pp}^*\mathbf{b}_p) \\ &= R(\mathbf{b}_p), \quad (p \le q \le \infty), \end{aligned}$$

(cf. Lemma 6.1 and (6.20)). Then

$$\begin{aligned} Y^* &\stackrel{\text{def}}{=} E(\phi^* \mid \mathbf{Y}^n, X_q^n) \\ &= E(\Sigma_{00} + \Phi \mid \mathbf{Y}^n, X_q^n) - 2E(\Sigma_{0p} \mid \mathbf{Y}^n, X_q^n)\boldsymbol{\beta}_p^L(q) \\ &\quad +(\boldsymbol{\beta}_p^L(q))'E(\Sigma_{pp} \mid \mathbf{Y}^n, X_q^n)\boldsymbol{\beta}_p^L(q) \\ &= Q_{00}^*(q) - 2Q_{0p}^*(q)\boldsymbol{\beta}_p^L(q) + (\boldsymbol{\beta}_p^L(q))'Q_{pp}^*\boldsymbol{\beta}_p^L(q) \\ &= R(\boldsymbol{\beta}_p^L(q)), \end{aligned}$$

(cf. Proposition 6.2).

$$\begin{aligned} Y^+ &= \inf_{\mathbf{b}_p}\{Y_{\mathbf{b}_p}\} = Q^*_{00.p} = R(\boldsymbol{\beta}^B_p(q)) \\ &= E(\Sigma \mid \mathbf{Y}^n, X^n_q)_{00.p} + E(\Phi \mid \mathbf{Y}^n, X^n_q) \end{aligned}$$

with

$$\Lambda^+ = \boldsymbol{\beta}^B_p(q),$$

(cf. Propositions 6.1, 6.3).

$$Y^{**} \stackrel{\text{def}}{=} E(\phi^{**} \mid \mathbf{Y}^n, X^n_q) = E(\Phi + \Sigma_{00.p} \mid \mathbf{Y}^n, X^n_q).$$

Again Y^{**}, Y^+, Y^* can be used as Bayesian estimates for ϕ^{**}, $\phi^+ = \phi_{\Lambda^+}$, ϕ^* respectively. Taking expectation conditional on $\mathbf{Y}^n$, X^n_q on both sides of (6.29) in Proposition 6.3, we have

$$\begin{aligned} R(\boldsymbol{\beta}^B_p(q)) &= E\{E[(Y^0 - \hat{f}(\mathbf{X}^0_p, \mathbf{Y}^n, X^n_q))^2 \mid \mathbf{X}^0_p, \mathbf{Y}^n, X^n_q] \mid \mathbf{Y}^n, X^n_q\} \\ &= E[(\mathbf{X}^0_p)'\mathrm{Var}(\boldsymbol{\beta}_p \mid \mathbf{X}^0_p, \mathbf{Y}^n, X^n_q)\mathbf{X}^0_p \mid \mathbf{Y}^n, X^n_q] + E(\Sigma_{00.p} + \Phi \mid \mathbf{Y}^n, X^n_q), \end{aligned}$$

which compared with (6.30) gives

$$E(\Sigma \mid \mathbf{Y}^n, X^n_q)_{00.p} \geq E(\Sigma_{00.p} \mid \mathbf{Y}^n, X^n_q).$$

Hence

$$Y^{**} \leq Y^+ \leq Y^*.$$

In this nonconjugate case, Y^{**} has a positive lower bound $K/(\nu - 2 + n)$, which does not depend on p, q. Hence the Bayes estimates Y^{**}, Y^+, Y^* for ϕ^{**}, ϕ^+, ϕ^* do not share the bias associated with X^* for finite or infinite p, q. Also as the optimal estimator achieving the minimum of the posterior squared error loss, the Bayes estimator $\Lambda^+ = \boldsymbol{\beta}^B_p(q)$ does not imply deterministic predictability. Moreover the asymptotic behaviours of the Bayes estimator $\boldsymbol{\beta}^B_p(q)$ and the Bayes–least squares estimator $\boldsymbol{\beta}^L_p(q)$ act quite differently. From the Bayes point of view, Theorem 6.2 shows that the difference of their posterior expected error losses is nondegenerate as $p \to \infty$. From the classical point of view, Theorem 6.1 shows that the error sum of squares of the Bayes esimator $\boldsymbol{\beta}^B_p(q)$, equivalently $X_{\boldsymbol{\beta}^B_{p(q)}}$ is nondegenerate

as $p \to \infty$, in contrast to the condition that $\mathbf{Y}^n = X_p^n \boldsymbol{\beta}_p^L(q)$ or $X_{\boldsymbol{\beta}_p^L(q)} = 0$. Also on a future case, the difference of the two predictors $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^B(q)$, $(\mathbf{X}_p^0)'\boldsymbol{\beta}_p^L(q)$ is nondegenerate as $p \to \infty$ (cf. Theorem 6.3). The above conclusions hold on condition that $K/(\nu - 2) > Q_{00}/(\delta - 2)$ (cf. Theorem 6.1 (c)). Thus by assuming a nonconjugate prior such as studied in this chapter, one may avoid overfitting in regression when using a large number of explanatory variables.

Chapter 7

CONCLUSION

Perhaps one of the main features of the Bayesian inference is the use of the non-sample information represented by the prior distribution of the parameters. This information is obtained from the prior belief of the statistician rather than his statistical investigation. Conjugate priors are often adopted because of their perceived richness and ease in calculation. Our investigation in this thesis has shown that, in some common multivariate problems the usual conjugate priors imply determinism of the inference if the number of variables that can be observed tends to infinity, i.e. if sufficiently many variables are observed, inference can be done perfectly. We have considered the Dirichlet process prior in discrete discrimination, normal inverse Wishart prior in continuous discrimination and inverse Wishart prior in regression, which all lead to asymptotically perfect discrimination between populations, and prediction of the response variable, respectively. In many contexts such determinism is unbelievable, because of the statistical nature of the problems considered. In such a case these conjugate priors seem inappropriate. We also considered a nonconjugate prior in continuous regression, which does not imply determinism and may be suitable for certain problems in which we do not believe in determinism.

We have connected our investigation of the determinism property to selection

bias caused by the inference made for the data–dependent parameters following random noise in the data rather than signal. As the number of variables that can be observed tends to infinity, our Bayesian inference closely mimics the unadjusted classical one in the case of conjugate priors. Thus if we belive in determinism we might just ignore the biasing effects of selection. Then the selection bias causes no problem for the Bayesian. In the case of a nonconjugate prior the Bayesian inference is not deterministic and totally different from the unadjusted classical one, even asymptotically as the number of the observed variables tends to infinity.

We conclude that the choice of the prior distribution has a great impact on Byaesian inference and must be considered according to the problem investigated with great care.

Bibliography

[1] Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, Inc. New York.

[2] Berger, J. O. (1980, 1985). *Statistical Decision Theory and Bayesian Analysis.* Springer–Verlarg, New York Inc.

[3] Brown, P. J. (1980). Coherence and complexity in classification problems, *Scand. J. Statist.* **7** 95–98.

[4] Chow, Y. S. and Teicher, H. (1978). *Probability Theory.* Springer–Verlarg, New York Inc.

[5] Cornish, E. A. (1954). The multivariate t–distribution. *Australian J. Physics.* **7** 531–42.

[6] Dawid, A. P. (1977). Spherical matrix distributions and a multivariate model. *J. R. Statist. Soc. B* **39**, 254–61.

[7] Dawid, A. P. (1978). Extendibility of spherical matrix distributions. *J. Mult. Anal.* **8** 559–66.

[8] Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 1–31.

[9] Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274.

[**10**] Dawid, A. P. (1988). The infinite regress and its conjugate analysis (with discussion). In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Eds.), pp. 95–110, Oxford University Press.

[**11**] Dawid, A. P. (1994). Selection paradoxes of Bayesian inference. *Multivariate Analysis and its Applications* (T. W. Anderson, K. T. Fang and I. Olkin, Eds.), IMS Lecture Notes–Monograph Series **24**, 211–220.

[**12**] Dawid, A. P. and Fang, B. Q. (1992). Conjugate Bayes discrimination with infinitely many variables. *J. Multivariate Anal.* **41** 27–42.

[**13**] Dempster, A. P. (1969). *Elements of Continuous Multivariate Analysis.* Reading, Mass: Addison–Wesley.

[**14**] Dickey, J. M. (1967). Matricvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution. *Ann. Math. Statist.* **38** 511–518.

[**15**] Fang, B. Q. and Dawid, A. P. (1993). Asymptotic properties of conjugate Bayes discrete discrimination *J. Multivariate Anal.* **46** 83–96.

[**16**] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

[**17**] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7** 179–188.

[**18**] Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *J. Roy. Statist. Soc. Ser. B* **25** 368–376.

[**19**] Geisser, S. (1964). Posterior odds for multivariate normal classifications. *J. Roy. Statist. Soc. Ser. B* **26** 69–76.

[**20**] Geisser, S. (1965). Bayesian estimation in multivariate analysis. *Ann. Math. Stat.* **36** 150–159.

[**21**] Katri, C. G. (1970). A note on Mitra's paper "A density free approach to the matrix variate beta distribution". *Sankhyā, A* **32** 311–318.

[22] Lehmann, E. L. (1983). *Theory of Point Estimation.* John Wiley & Sons, Inc.

[23] Mitra, S. K. (1970). A density–free approach to the matrix variate beta distribution. *Sankhyā, A* **32** 81–88.

[24] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* John Wiley & Sons, Inc. New York.

[25] Olkin, I. and Rubin, H. (1964). Multivariate Beta distributions and independence properties of the Wishart distribution. *Ann. Math. Statist.* **35** 261–269.

[26] Patil, G. P. and Taillie, C. (1977). Diversity as a concept and its implications for random communities. *Bull. Internat. Statist. Inst.* **47** 497–515.

[27] Press, S. J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference.* 2nd ed., Robert E. Krieger Publishing Co., Inc.

[28] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* 2nd ed., John Wiley & Sons.

[29] Rao, B. L. S. P. (1987). *Asymptotic Theory of Statistical Inference.* John Wiley & Sons, Inc.

[30] Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Ann. Math. Statist.* **18** 434–438.

[31] Tan, W. Y. (1969). Note on the multivariate and the generalized multivariate Beta distributions. *J. Amer. Staist. Assoc.* **64** 230–241.

[32] Tiao, G. C. and Zellner, A. (1964). On the Bayesian estimation of multivariate regression. *J. Roy. Statist. Soc. Ser. B* **26** 277–285.

Appendix A

Conjugate Bayes Discrimination with infinitely many variables

by Dawid, A. P. and Fang, B. Q. (1992), published in *J. Mult. Anal.* **41** 27–42,

Reprinted from JOURNAL OF MULTIVARIATE ANALYSIS
Vol. 41, No. 1, April 1992
Printed in Belgium

Conjugate Bayes Discrimination with Infinitely Many Variables

A. P. DAWID

University College London, London, England

AND

B. Q. FANG

Institute of Applied Mathematics, Academia Sinica, Beijing, China

Communicated by C. R. Rao

The problem considered is that of discrimination between two multivariate normal populations, with common dispersion structure, when the number of variables that can be observed is unlimited. We consider a Bayesian analysis, using a natural conjugate prior for the normal distribution parameters. One implication of this is that, with prior probability 1, the parameters will be such as to allow asymptotically perfect discrimination between the populations. We also find conditions under which this perfect discrimination will be possible, even in the absence of knowledge of the parameter values. © 1992 Academic Press, Inc.

1. INTRODUCTION

Dawid [3] considered the problem of predicting a continuous variable Y on the basis of a potentially infinite number of explanatory variables, assuming the joint distribution to be normal. Conditions on the parameters under which the predictive distribution would become degenerate as the number of explanatory variables increased were found, and it was shown that the usual conjugate prior distribution assigns probability 1 to this event. A necessary and sufficient condition was also given for asymptotically degenerate prediction to be possible, for the conjugate Bayesian, even in the absence of knowledge of the parameters.

Received May 16, 1991.

AMS 1980 subject classifications: 62H30, 62A15.

Key words and phrases: discrimination, Bayesian inference, conjugate prior, normal inverted Wishart distribution, predictive distribution, determinism.

0047-259X/92 $3.00

The present paper extends the above programme to the problem of discrimination between two homoscedastic multivariate normal populations. Conjugate Bayes analysis in this problem, for a finite number of variables, has been considered by Geisser [5]. We develop theory for the case of infinitely many variables. We can again specify a "natural conjugate" form for the prior distribution of the normal parameters, the *normal inverted Wishart* distribution, which renders the Bayesian analysis particularly simple. However, since we are working in an infinite-dimensional parameter space, the choice of prior will not be unimportant, even when the data are extensive. Consequently, before automatically reaching for the natural conjugate prior, it is important to be aware of its implications and to be satisfied that these are acceptable; if they are not, then a more sophisticated Bayesian analysis will be required. With this end in mind, this paper develops some of the implications of the use of the conjugate prior, with particular attention to the possibilities for "asymptotically degenerate" discrimination, whether or not the parameters are known.

Section 2 shows that a necessary and sufficient condition for asymptotically degenerate discrimination with known parameters is that the Mahalanobis distance between the two normal populations be infinite. Section 3 introduces the natural conjugate prior and shows that it assigns probability 1 to the above condition. In Section 4 we take up the problem of discrimination when the parameters are unknown and find conditions on the hyper-parameters of the conjugate prior under which asymptotically degenerate discrimination is expected in this case also. Section 5 considers the use of training data, consisting of the values of all variables for a random sample of individuals, to learn about the unknown parameters: these training data are observed before classification of a new individual is required and are to be utilized in aiding that classification. We analyse the behaviour of the probabilities, for the new individual, of belonging to either population, conditional on the training data as well as on the explanatory variables for the new individual. When the training data are sufficiently extensive, any conjugate prior will imply asymptotically degenerate discrimination in this case.

1.1. *Assumptions and Notation*

Suppose that individuals each belong to one of two populations Π_1 and Π_2: we introduce a binary indicator variable Y, with $Y=i$ denoting membership of Π_i. The probabilities $\pi_i = P(Y=i)$ $(i=1, 2)$ are supposed known. Also associated with each individual is a countable collection $\mathbf{X} = (X_1, X_2, \ldots)'$ of continuous variables. These are modelled as having a multivariate Normal distribution within either population, with

$$E(X_j \mid Y=i) = \mu_{ij} \qquad (i=1, 2; j=1, 2, \ldots),$$

and

$$\operatorname{cov}(X_j, X_k \mid Y=i) = \sigma_{jk} \qquad (i=1, 2;\ j, k=1, 2, \ldots).$$

In particular, the dispersion structure is the same within both populations. We write $\mathbf{X}_p$ for $(X_1, \ldots, X_p)'$.

Denote by $\boldsymbol{\mu}_i$ the infinite column vector $(\mu_{i1}, \mu_{i2}, \ldots)'$ and by μ the $(2 \times \infty)$ matrix $\binom{\boldsymbol{\mu}_1'}{\boldsymbol{\mu}_2'}$, whose (i, j)-entry is μ_{ij}. We write $\boldsymbol{\mu}_{ip}$ and μ_p for the sub-objects of these quantities obtained by restricting attention to the first p variables. Σ will denote the $(\infty \times \infty)$ matrix with (j, k)-entry σ_{jk} $(j, k=1, 2, \ldots)$ and Σ_p its restriction to $1 \leqslant j, k \leqslant p$.

We shall make extensive use, without further detailed description, of the notation and conventions for matrix distributions developed in Dawid [2]. The reader should be aware that these may differ from other common conventions; however, the use of this coherent notation is essential when distributions for infinite vectors and matrices are handled.

We shall also make some use, again without further detailed description, of the notation for and properties of *conditional independence* as developed in Dawid [1].

2. Known Parameters

Suppose that all parameters are known. Then for any fixed p the density of $\mathbf{X}_p$ given Y is

$$f(\mathbf{x}_p \mid Y=i;\ \mu, \Sigma) = (2\pi)^{-p/2}\, |\Sigma_p|^{-1/2} \exp\left[-\tfrac{1}{2}(\mathbf{x}_p - \boldsymbol{\mu}_{ip})'\, \Sigma_p^{-1} (\mathbf{x}_p - \boldsymbol{\mu}_{ip})\right], \tag{2.1}$$

which leads to the ratio of the conditional probabilities of $Y=i$ given $\mathbf{X}_p$ (when the parameters are known),

$$\begin{aligned} \frac{P(Y=2 \mid \mathbf{X}_p;\ \mu, \Sigma)}{P(Y=1 \mid \mathbf{X}_p;\ \mu, \Sigma)} &= \frac{P(Y=2)\, f(\mathbf{X}_p \mid Y=2;\ \mu, \Sigma)}{P(Y=1)\, f(\mathbf{X}_p \mid Y=1;\ \mu, \Sigma)} \\ &= \frac{\pi_2}{\pi_1} \exp\left[-\frac{1}{2}(S_{2p} - S_{1p})\right], \end{aligned} \tag{2.2}$$

where $S_{ip} = (\mathbf{X}_p - \boldsymbol{\mu}_{ip})'\, \Sigma_p^{-1} (\mathbf{X}_p - \boldsymbol{\mu}_{ip})$.

In order to investigate the asymptotic behaviour of (2.2) as $p \to \infty$, let

$$\begin{aligned} \mathbf{Z}_p &= \Sigma_p^{-1/2} (\mathbf{X}_p - \boldsymbol{\mu}_{1p}) \\ \mathbf{a}_p &= \Sigma_p^{-1/2} (\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p}) \\ \lambda_p &= \mathbf{a}_p' \mathbf{a}_p = (\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p})'\, \Sigma_p^{-1} (\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p}). \end{aligned}$$

Then, when $\mathbf{X}$ arises from population Π_1,

$$\mathbf{Z}_p \sim N(\mathbf{0}, I_p)$$

$$S_{2p} - S_{1p} = 2\mathbf{a}_p'\mathbf{Z}_p + \mathbf{a}_p'\mathbf{a}_p \sim N(\lambda_p, 4\lambda_p).$$

Now λ_p, being the Mahalanobis distance between the populations Π_1 and Π_2 based on $\mathbf{X}_p$, is non-decreasing as p increases. Hence there exists $\lambda_\infty \overset{\text{def}}{=} \lim_{p\to\infty} \lambda_p$, with $\lambda_\infty \leqslant \infty$. Thus, as $p \to \infty$,

$$\frac{P(Y=2\,|\,\mathbf{X}_p;\mu,\Sigma)}{P(Y=1\,|\,\mathbf{X}_p;\mu,\Sigma)} \begin{cases} \xrightarrow{L} \dfrac{\pi_2}{\pi_1}\exp\left\{N\left(-\dfrac{1}{2}\lambda_\infty, \lambda_\infty\right)\right\} & \text{if } \lambda_\infty < \infty \\ \xrightarrow{p} 0 & \text{if } \lambda_\infty = \infty. \end{cases} \tag{2.3}$$

Similarly, when $\mathbf{X}$ arises from population Π_2,

$$\frac{P(Y=2\,|\,\mathbf{X}_p;\mu,\Sigma)}{P(Y=1\,|\,\mathbf{X}_p;\mu,\Sigma)} \begin{cases} \xrightarrow{L} \dfrac{\pi_2}{\pi_1}\exp\left\{N\left(\dfrac{1}{2}\lambda_\infty, \lambda_\infty\right)\right\} & \text{if } \lambda_\infty < \infty \\ \xrightarrow{p} \infty & \text{if } \lambda_\infty = \infty. \end{cases} \tag{2.4}$$

Also, since $(P(Y=i\,|\,\mathbf{X}_p;\mu,\Sigma) : p = 1, 2, \ldots)$ is a martingale, we have

$$P(Y=i\,|\,\mathbf{X}_p;\mu,\Sigma) \xrightarrow{\text{a.s.}} P(Y=i\,|\,\mathbf{X};\mu,\Sigma) \qquad \text{as} \quad p\to\infty,$$

so that, in Eqs. (2.3) and (2.4), the left-hand side converges almost surely to a random variable with distribution given by the right-hand side. In particular, when $\lambda_\infty = \infty$, this limit is almost surely 0 or ∞, according to whether $\mathbf{X}$ arises from Π_1 or Π_2, whereas when $\lambda_\infty < \infty$, the limit is almost surely finite in both populations. Thus, when the parameters (μ, Σ) are given, the condition $\lambda_\infty = \infty$ is necessary and sufficient for there to be asymptotically degenerate discrimination between the two populations.

3. Prior Distribution

Now suppose that the parameters (μ, Σ) are assigned a conjugate normal inverted Wishart distribution $\mathcal{NIW}(m, H; \delta, K)$ (in the notation of Dawid [2]), where $\mu = \binom{\mu_1}{\mu_2}$ and $m = \binom{m_1}{m_2}$ are $(2 \times \infty)$, $H = \begin{pmatrix} h_1^{-1} & 0 \\ 0 & h_2^{-1} \end{pmatrix}$, $\delta > 0$, and K is $(\infty \times \infty)$. This means that, for any p,

$$(\mu_p, \Sigma_p) \sim \mathcal{NIW}(m_p, H; \delta, K_p),$$

where $\mu_p = \binom{\mu_{1p}'}{\mu_{2p}'}$, $m_p = \binom{\mathbf{m}_{1p}'}{\mathbf{m}_{2p}'}$, and K_p is the leading $(p \times p)$ submatrix of K (required to be non-negative definite for all p). Thus

$$\Sigma_p \sim \mathscr{I}\mathscr{W}(\delta; K_p) \tag{3.1}$$

(equivalent, if K_p is non-singular, to $\Sigma_p^{-1} \sim \mathscr{W}(\delta + p - 1; K_p^{-1})$), and, conditional on Σ_p,

$$\boldsymbol{\mu}_{ip} \mid \Sigma_p \sim N(\mathbf{m}_{ip}, h_i^{-1}\Sigma_p), \tag{3.2}$$

independently for $i = 1, 2$.

We shall show that this prior distribution assigns probability 1 to the event that the parameters (μ, Σ) are such as to yield $\lambda_\infty = \infty$ and thus are such as to permit asymptotically degenerate discrimination (were they to be known).

To see this, we write $\lambda_p = \alpha_{1p} + \alpha_{2p} + \alpha_{3p}$, with

$$\alpha_{1p} = [(\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p}) - (\mathbf{m}_{1p} - \mathbf{m}_{2p})]' \, \Sigma_p^{-1} [(\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p}) - (\mathbf{m}_{1p} - \mathbf{m}_{2p})]$$

$$\alpha_{2p} = 2(\mathbf{m}_{1p} - \mathbf{m}_{2p})' \, \Sigma_p^{-1} [(\boldsymbol{\mu}_{1p} - \boldsymbol{\mu}_{2p}) - (\mathbf{m}_{1p} - \mathbf{m}_{2p})]$$

$$\alpha_{3p} = (\mathbf{m}_{1p} - \mathbf{m}_{2p})' \, \Sigma_p^{-1} (\mathbf{m}_{1p} - \mathbf{m}_{2p}).$$

Then

$$\alpha_{3p} \sim (\mathbf{m}_{1p} - \mathbf{m}_{2p})' \, K_p^{-1} (\mathbf{m}_{1p} - \mathbf{m}_{2p}) \, \chi^2_{\delta + p - 1}$$

$$\xrightarrow{p} \infty \qquad \text{as} \quad p \to \infty$$

since $(\mathbf{m}_{1p} - \mathbf{m}_{2p})' \, K_p^{-1} (\mathbf{m}_{1p} - \mathbf{m}_{2p})$ is non-decreasing with p.

Also, given Σ,

$$\alpha_{2p} \sim N(0, 4(h_1^{-1} + h_2^{-1}) \, \alpha_{3p})$$

$$= O_p(\alpha_{3p}^{1/2}),$$

so that $\alpha_{2p} + \alpha_{3p} \xrightarrow{p} \infty$.

Finally,

$$\alpha_{1p} \sim (h_1^{-1} + h_2^{-1}) \, \chi_p^2 \xrightarrow{p} \infty.$$

We deduce that, under the given conjugate prior distribution, $\lambda_p \xrightarrow{p} \infty$ (and so $\xrightarrow{\text{a.s.}} \infty$) as $p \to \infty$, and thus we expect the parameters to be such as to permit asymptotically degenerate discrimination.

4. Discrimination with Unknown Parameters

We have seen that a normal inverted Wishart prior distribution for the parameters assigns probability 1 to the event that those parameters are such as to permit asymptotically degenerate discrimination between populations. However, this is generally of limited direct practical interest, since the values of the parameters remain unknown, and so cannot be used to perform that discrimination. An exception to this occurs, however, when the prior distribution attaches probability 1 to a set of parameter values that all lead to the same asymptotic classification rule. In this case, asymptotically degenerate discrimination is to be expected even in the absence of knowledge of the parameters. In this section we investigate the behaviour of the classification probabilities when the parameters remain unknown, but are assigned a conjugate prior distribution. In particular, we characterize those conjugate priors that imply asymptotically degenerate discrimination in this case.

4.1. *Classification Probabilities*

We continue to suppose that the prior distribution is $\mathcal{NIW}(m, H; \delta, K)$. Then the distribution of $\mathbf{X}$ given $Y=i$ and Σ only is $N(\mathbf{m}_i, k_i\Sigma)$, where $k_i = 1 + h_i^{-1}$. Thus marginalizing out further over Σ, we obtain an infinite multivariate-t distribution for $\mathbf{X}$ given only Y: in the notation of Dawid [2],

$$\mathbf{X} \mid Y=i \sim \mathbf{m}_i + T(\delta; K, k_i).$$

That is, for any p,

$$\mathbf{X}_p \mid Y=i \sim \mathbf{m}_{ip} + T(\delta; K_p, k_i), \tag{4.1}$$

with density

$$f(\mathbf{x}_p \mid Y=i) = \kappa_p^{-1} k_i^{\delta/2} \, |K_p|^{-1/2} \, [k_i + (\mathbf{x}_p - \mathbf{m}_{ip})' \, K_p^{-1} (\mathbf{x}_p - \mathbf{m}_{ip})]^{-(\delta+p)/2}, \tag{4.2}$$

where $\kappa_p = \pi^{p/2}\Gamma(\frac{1}{2}\delta)/\Gamma(\frac{1}{2}(\delta+p))$ (cf. Dickey [4, Eq. (3.2)]). Thus

$$\begin{aligned} \frac{P(Y=2 \mid \mathbf{X}_p)}{P(Y=1 \mid \mathbf{X}_p)} &= \frac{P(Y=2)\, f(\mathbf{X}_p \mid Y=2)}{P(Y=1)\, f(\mathbf{X}_p \mid Y=1)} \\ &= \frac{\pi_2 k_2^{\delta/2} (k_2 + S_{2p})^{-(\delta+p)/2}}{\pi_1 k_1^{\delta/2} (k_1 + S_{1p})^{-(\delta+p)/2}}, \end{aligned} \tag{4.3}$$

where now $S_{ip} = (\mathbf{X}_p - \mathbf{m}_{ip})' \, K_p^{-1} (\mathbf{X}_p - \mathbf{m}_{ip})$.

4.2. *Expected Behaviour of Classification Probabilities*

We shall now investigate the asymptotic behaviour of (4.3), conditional on $Y=1$, but unconditional with respect to the parameters. In this case we have $\mathbf{X} \sim \mathbf{m}_1 + T(\delta; K, k_1)$, which admits the synthetic representation

$$\mathbf{X} = \mathbf{m}_1 + k_1^{1/2} \Lambda^{1/2} D \mathbf{Z}, \tag{4.4}$$

where $\mathbf{Z} \sim N(\mathbf{0}, I_\infty)$ and $\Lambda^{-1} \sim \chi^2_\delta$, independently, and D is an infinite lower triangular matrix such that $DD' = K$.

Working now from (4.4), define

$$\mathbf{T} = k_1^{-1/2} D^{-1} (\mathbf{X} - \mathbf{m}_1) = \Lambda^{1/2} \mathbf{Z}$$

and

$$\mathbf{A} = k_1^{-1/2} D^{-1} (\mathbf{m}_1 - \mathbf{m}_2).$$

Then $S_{1p}/p = k_1 \mathbf{T}_p' \mathbf{T}_p / p = k_1 V_p \Lambda$, where $V_p = \mathbf{Z}_p' \mathbf{Z}_p / p \xrightarrow{\text{a.s.}} 1$. Hence, as $p \to \infty$, $S_{1p}/p \xrightarrow{\text{a.s.}} k_1 \Lambda$ (this relation may be regarded as identifying Λ in (4.4) as a function of $\mathbf{X}$).

Also, $S_{2p} = k_1 (\mathbf{T}_p + \mathbf{A}_p)' (\mathbf{T}_p + \mathbf{A}_p)$. So

$$\begin{aligned} \left(\frac{k_2 + S_{2p}}{k_1 + S_{1p}} \right) &= \frac{k_2 + k_1 (\mathbf{T}_p + \mathbf{A}_p)' (\mathbf{T}_p + \mathbf{A}_p)}{k_1 (1 + \mathbf{T}_p' \mathbf{T}_p)} \\ &\approx \frac{(\mathbf{T}_p + \mathbf{A}_p)' (\mathbf{T}_p + \mathbf{A}_p)}{\mathbf{T}_p' \mathbf{T}_p} \\ &= 1 + U_p / p, \end{aligned} \tag{4.5}$$

where

$$U_p = \frac{2\mathbf{A}_p' \mathbf{T}_p + \mathbf{A}_p' \mathbf{A}_p}{\mathbf{T}_p' \mathbf{T}_p / p} = (2\Lambda^{-1/2} \mathbf{A}_p' \mathbf{Z}_p + \Lambda^{-1} \mathbf{A}_p' \mathbf{A}_p) / V_p. \tag{4.6}$$

Now define $\gamma_p = (\mathbf{m}_{1p} - \mathbf{m}_{2p})' K_p^{-1} (\mathbf{m}_{1p} - \mathbf{m}_{2p}) = k_1 \mathbf{A}_p' \mathbf{A}_p$. Then γ_p is non-decreasing in p and thus tends to a limit $\gamma_\infty \leqslant \infty$. Then, conditionally on Λ,

$$\begin{aligned} &U_p \xrightarrow{L} N((k_1 \Lambda)^{-1} \gamma_\infty, 4(k_1 \Lambda)^{-1} \gamma_\infty) && \text{if} \quad \gamma_\infty < \infty; \\ &U_p \xrightarrow{p} \infty && \text{if} \quad \gamma_\infty = \infty. \end{aligned} \tag{4.7}$$

From (4.3), (4.5), and (4.7), we thus have the following representation of the asymptotic behaviour of $P(Y=2\,|\,\mathbf{X}_p)/P(Y=1\,|\,\mathbf{X}_p)$ as $p\to\infty$:

$$\frac{P(Y=2\,|\,\mathbf{X}_p)}{P(Y=1\,|\,\mathbf{X}_p)}\begin{cases}\xrightarrow{L} \dfrac{\pi_2}{\pi_1}\left(\dfrac{k_2}{k_1}\right)^{\delta/2}\exp\Omega & \text{if } \gamma_\infty<\infty\\ \xrightarrow{p} 0 & \text{if } \gamma_\infty=\infty,\end{cases} \tag{4.8}$$

where $k_i=1+h_i^{-1}$ and the distribution of Ω is the mixture, over the distribution $(\chi^2_\delta)^{-1}$ for Λ, of $N(-\frac{1}{2}(k_1\Lambda)^{-1}\gamma_\infty, (k_1\Lambda)^{-1}\gamma_\infty)$. Further, the left-hand side of (4.8) converges almost surely, to a random variable whose distribution is represented by the right-hand side.

The above analysis is all conditional on $Y=1$; a parallel result holds given $Y=2$. In particular, we see that a necessary and sufficient condition for asymptotically degenerate discrimination between the two populations in the absence of knowledge of the parameters is that $\gamma_\infty=\infty$. (More precisely, this is the condition under which, according to the prior distribution being used, such discrimination is to be expected, with probability 1. This expectation could, however, be confounded by the data, if it in fact turned out that the posterior odds (4.8) did not converge to 0 or ∞. In this case the inference must be that either the sampling model or the prior assumptions have been discredited.)

5. Training Data

Suppose that we have observed the values of Y and all the X's for a random sample of n individuals, yielding the *training data*

$$((y_i, x_{i1}, x_{i2}, \ldots), i=1, \ldots, n).$$

Let $\mathbf{y}^n$ denote the vector $(y_i,\ i=1, \ldots, n)$, and x^n the semi-infinite matrix $(x_{ij},\ i=1, \ldots, n;\ j=1, 2, \ldots)$.

We are now presented with a further individual, on which we observe the values $\mathbf{x}_p^0=(x_1^0, \ldots, x_p^0)'$ of the first p X's. We wish to make inference about the value of Y^0 for this new individual, on the basis of all the data $(\mathbf{y}^n, x^n, \mathbf{x}_p^0)$. For this we require the predictive distribution (not conditioned on the parameters) of Y^0 given $(\mathbf{y}^n, x^n, \mathbf{x}_p^0)$. In this section we shall investigate this predictive distribution.

We have already shown in Section 3 that, under the conjugate prior considered there, $\lambda_\infty=\infty$ with probability 1. From this it readily follows that the overall distribution must attach probability 1 to the event that the training data $(\mathbf{Y}^n, X^n)$ will be such that, in the posterior distribution

of the parameters given $(\mathbf{Y}^n, X^n)$, with probability 1 we shall have $\lambda_\infty = \infty$ (and thus with probability 1 the parameters will be such as to support asymptotically degenerate discrimination, were they to be known). Similarly, under the further condition $\gamma_\infty = \infty$, we shall attach probability 1 to the event that the predictive distribution of $(Y^0, \mathbf{X}^0)$ given $(\mathbf{Y}^n, X^n)$ will be such as to allow asymptotically (as $p \to \infty$) almost sure identification of Y^0 on the basis of $\mathbf{X}_p^0$, even when the parameters are unknown.

5.1. *Irrelevance of Unobserved Variables*

We now investigate more fully the nature and properties of the predictive distribution of Y^0 given $(\mathbf{y}^n, x^n, \mathbf{x}_p^0)$. First, we show that it is enough to condition on $(\mathbf{y}^n, x_p^n, \mathbf{x}_p^0)$, where $x_p^n = (x_{ij}: i = 1, ..., n;\ j = 1, ..., p)$, so that there is no useful information in the values in the training data of variables that have not been observed on the new individual for whom prediction is required.

We need the following result on conditional independence in the normal inverted Wishart distribution.

LEMMA 1. *Suppose*

$$\Sigma\ (q \times q) \sim \mathcal{IW}(\delta, K)$$

and that, conditional on Σ,

$$\mu\ (a \times q) \sim M + \mathcal{N}(H, \Sigma).$$

Partition μ, Σ, *and* M *as*

$$\mu = \underset{p}{\overset{}{(}}\underset{}{\mu_p} \quad \underset{q-p}{\mu_+}\overset{}{)}\ a$$

$$\Sigma = \begin{pmatrix} \underset{p}{\Sigma_p} & \underset{q-p}{\Sigma_{p+}} \\ \underset{p}{\Sigma_{+p}} & \underset{q-p}{\Sigma_{++}} \end{pmatrix} \begin{matrix} p \\ q-p \end{matrix}$$

$$M = \underset{p}{\overset{}{(}}\underset{}{M_p} \quad \underset{q-p}{M_+}\overset{}{)}\ a.$$

Define

$$\beta = \Sigma_p^{-1}\Sigma_{p+}$$

$$\alpha = \mu_+ - \mu_p\beta$$

$$\Sigma_{++\cdot p} = \Sigma_{++} - \Sigma_{+p}\Sigma_p^{-1}\Sigma_{p+}.$$

Then $(\alpha, \beta, \Sigma_{++\cdot p}) \perp\!\!\!\perp (\mu_p, \Sigma_p)$.

Proof. Since $\mu_p | \Sigma \sim M_p + \mathcal{N}(H, \Sigma_p)$, we have

$$\mu_p \perp\!\!\!\perp (\beta, \Sigma_{++\cdot p}) | \Sigma_p. \tag{5.1}$$

Also, by Lemma 2 of Dawid [3], we have

$$\Sigma_p \perp\!\!\!\perp (\beta, \Sigma_{++\cdot p}). \tag{5.2}$$

From (5.1) and (5.2) we obtain

$$(\mu_p, \Sigma_p) \perp\!\!\!\perp (\beta, \Sigma_{++\cdot p}). \tag{5.3}$$

Also, since

$$\mu_+ | (\mu_p, \Sigma) \sim M_+ + (\mu_p - M_p)\beta + \mathcal{N}(H, \Sigma_{++\cdot p}),$$

we have

$$\alpha | (\mu_p, \Sigma) \sim M_+ - M_p\beta + \mathcal{N}(H, \Sigma_{++\cdot p}),$$

whence we see that

$$\alpha \perp\!\!\!\perp (\mu_p, \Sigma_p) | (\beta, \Sigma_{++\cdot p}). \tag{5.4}$$

The result now follows from (5.3) and (5.4). ∎

PROPOSITION 1. *For all $q > p$,*

$$P(Y^0 = i | \mathbf{x}_p^0, \mathbf{y}^n, x_q^n) = P(Y^0 = i | \mathbf{x}_p^0, \mathbf{y}^n, x_p^n).$$

Proof. Let $x_+^n = (x_{ij}: i = 1, ..., n; j = p+1, ..., q)$. Then

$$\begin{aligned} P(Y^0 = i | \mathbf{x}_p^0, \mathbf{y}^n, x_q^n) &= P(Y^0 = i | \mathbf{x}_p^0, \mathbf{y}^n, x_p^n, x_+^n) \\ &= E[P(Y^0 = i | \mathbf{x}_p^0; \mu_p, \Sigma_p) | \mathbf{x}_p^0, \mathbf{y}^n, x_p^n, x_+^n)]. \end{aligned}$$

The result will therefore follow if we can show that

$$(\mu_p, \Sigma_p) \perp\!\!\!\perp X_+^n | (\mathbf{X}_p^0, \mathbf{Y}^n, X_p^n). \tag{5.5}$$

Now it may be seen that we can factorise the joint data density in the form

$$\begin{aligned} &f(\mathbf{x}_p^0, \mathbf{y}^n, x_p^n, x_+^n | \mu_q, \Sigma_q) \\ &\quad = f(\mathbf{x}_p^0, \mathbf{y}^n, x_p^n | \mu_p, \Sigma_p) f(x_+^n | \mathbf{y}^n, x_p^n; \alpha, \beta, \Sigma_{++\cdot p}). \end{aligned} \tag{5.6}$$

Applying Bayes' theorem, and using (5.6) and Lemma 1, we obtain the posterior density factorisation

$$\pi(\mu_p, \Sigma_p, \alpha, \beta, \Sigma_{++\cdot p} | \mathbf{x}_p^0, \mathbf{y}^n, x_p^n, x_+^n)$$
$$= \pi(\mu_p, \Sigma_p | \mathbf{x}_p^0, \mathbf{y}^n, x_p^n)\, \pi(\alpha, \beta, \Sigma_{++\cdot p} | \mathbf{x}_p^0, \mathbf{y}^n, x_p^n, x_+^n),$$

from which the result follows. ∎

Letting $q \to \infty$ in Proposition 1, we see that

$$P(Y^0 = i | \mathbf{x}_p^0, \mathbf{y}^n, x^n) = P(Y^0 = i | \mathbf{x}_p^0, \mathbf{y}^n, x_p^n).$$

Thus, if we have observed only the values $\mathbf{x}_p^0$ of a set of p predictor variables for our new case, then the only aspects of the training data that are relevant to the prediction of Y^0 for this case are the responses $\mathbf{y}^n$ and the values x_p^n of the same p predictor variables in the training data.

5.2. *Predictive Distribution*

We now investigate the distribution, for the new case, of Y^0 given $\mathbf{X}_p^0$ and the training data. As in the unconditional analysis of Section 4, we proceed by applying Bayes' theorem to the distribution of $\mathbf{X}_p^0$ given Y^0. Without loss of generality, we suppose that the cases in the training data set have been ordered so that the n_1 cases with $y_i = 1$ precede the n_2 cases with $y_i = 2$.

Given $Y^0 = 1$, $\mathbf{y}^n$, and the parameters, we have the sampling distribution

$$\begin{pmatrix} \mathbf{X}^{0\prime} \\ X^n \end{pmatrix} \sim \Gamma\mu + \mathcal{N}(I_{n+1}, \Sigma),$$

where

$$\Gamma = \begin{pmatrix} \mathbf{1}_{\nu_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{\nu_2} \end{pmatrix}, \tag{5.7}$$

with $\nu_1 = n_1 + 1$, $\nu_2 = n_2$.

If we now marginalize over the $\mathcal{NIW}(m, H; \delta, K)$ prior distribution of the parameters, we obtain the distribution conditional only on $Y^0 = 1$ and $\mathbf{y}^n$,

$$\begin{pmatrix} \mathbf{X}^{0\prime} \\ X^n \end{pmatrix} \sim \Gamma m + T(\delta; Q, K), \tag{5.8}$$

where

$$Q = I + \Gamma H \Gamma' = \begin{pmatrix} Q_1 & \mathbf{0} \\ \mathbf{0} & Q_2 \end{pmatrix} \tag{5.9}$$

with

$$Q_i\,(v_i \times v_i) = \begin{pmatrix} (1+h_i^{-1}) & \cdots & h_i^{-1} \\ \vdots & \ddots & \vdots \\ h_i^{-1} & \cdots & (1+h_i^{-1}) \end{pmatrix}. \tag{5.10}$$

By Lemma 4 of Dawid [3] we have that, conditional on $Y^0 = 1$, $\mathbf{y}^n$, and x^n,

$$\begin{aligned} \mathbf{X}^{0\prime} \sim\; & m_1' + Q_{0n} Q_{nn}^{-1}(x^n - \Gamma_n m) \\ & + T(\delta + n;\, Q_{00\cdot n},\, K + (x^n - \Gamma_n m)'\, Q_{nn}^{-1}(x^n - \Gamma_n m)), \end{aligned} \tag{5.11}$$

where we have used the partitions

$$Q = \begin{matrix} \\ 1 \\ n \end{matrix} \begin{matrix} \begin{matrix} 1 & n \end{matrix} \\ \begin{pmatrix} Q_{00} & Q_{0n} \\ Q_{n0} & Q_{nn} \end{pmatrix} \end{matrix}$$

and

$$\Gamma = \begin{matrix} 1 \\ n \end{matrix} \begin{pmatrix} \Gamma_0 \\ \Gamma_n \end{pmatrix}.$$

In particular, Γ_n and Q_{nn} are given by Eqs. (5.7), (5.9), and (5.10), with n_1 and n_2 in place of v_1 and v_2.

Simplifying (5.11) and the parallel result for $Y^0 = 2$, we obtain

$$\mathbf{X}^0 \mid (Y^0 = i,\, x^n,\, \mathbf{y}^n) \sim \mathbf{m}_i^*(x^n) + T(\delta^*;\, K^*(x^n),\, k_i^*), \tag{5.12}$$

where $\mathbf{m}_i^*(x^n) = (n_i \bar{\mathbf{x}}_i^n + h_i \mathbf{m}_i)/(h_i + n_i)$ ($\bar{\mathbf{x}}_i^n$ being the average of the $\mathbf{X}$-vectors associated with those training cases for which $Y = i$), $\delta^* = \delta + n$, $K^*(x^n) = K + (x^n - \Gamma_n m)'\, Q_{nn}^{-1}(x^n - \Gamma_n m)$, and $k_i^* = 1 + (h_i + n_i)^{-1}$. (All these quantities also depend on $\mathbf{y}^n$, but as we shall only be considering behaviour conditional on fixed $\mathbf{y}^n$ we omit this from the notation).

Restricting (5.12) to the first p predictor variables yields

$$\mathbf{X}_p^0 \mid (Y^0 = i,\, x^n,\, \mathbf{y}^n) \sim \mathbf{m}_{ip}^*(x^n) + T(\delta^*;\, K_p^*(x^n),\, k_i^*), \tag{5.13}$$

with $\mathbf{m}_{ip}^*(x^n)$ the initial p-segment of $\mathbf{m}_i^*(x^n)$ and

$$K_p^*(x^n) = K_p + (x_p^n - \Gamma_n m_p)'\, Q_{nn}^{-1}(x_p^n - \Gamma_n m_p).$$

In particular, we note the dependence on x^n through x_p^n alone. This property will thus also hold for the *predictive odds*

$$\frac{P(Y^0=2\,|\,\mathbf{X}_p^0, x^n, \mathbf{y}^n)}{P(Y^0=1\,|\,\mathbf{X}_p^0, x^n, \mathbf{y}^n)}=\frac{P(Y^0=2)\,f(\mathbf{X}_p^0|\,Y^0=2, x^n, \mathbf{y}^n)}{P(Y^0=1)\,f(\mathbf{X}_p^0|\,Y^0=1, x^n, \mathbf{y}^n)}, \tag{5.14}$$

in agreement with Section 5.1.

Comparing (5.13) with (4.1), we see from (4.8) that the limiting behaviour, as $p\to\infty$, of the predictive odds (5.14) is determined by $\gamma_\infty^*(x^n)=\lim_{p\to\infty}\gamma_p^*(x^n)$, where

$$\gamma_p^*(x^n)\overset{\text{def}}{=}(\mathbf{m}_{1p}^*(x^n)-\mathbf{m}_{2p}^*(x^n))'\,K_p^*(x^n)^{-1}\,(\mathbf{m}_{1p}^*(x^n)-\mathbf{m}_{2p}^*(x^n)).$$

The arguments at the beginning of this section demonstrate that, if $\gamma_\infty=\infty$, then $\gamma_\infty^*(X^n)=\infty$ with probability 1 (unconditionally and hence also conditionally on $\mathbf{y}^n$): this may also be verified by direct calculation. We now investigate the behaviour of $\gamma_\infty^*(X^n)$ (conditionally on $\mathbf{y}^n$) when $\gamma_\infty<\infty$.

Define $U_p=Q_{nn}^{-1/2}(X_p^n-\Gamma_n m_p)\,K_p^{-1/2}$, where we take the symmetric square roots. From (5.8) we obtain $U_p\sim T(\delta; I_n, I_p)$. Let

$$C=\begin{pmatrix}(h_1+n_1)^{-1}\,\mathbf{1}_{n_1}\\ -(h_2+n_2)^{-1}\,\mathbf{1}_{n_2}\end{pmatrix},\qquad \boldsymbol{\alpha}=\begin{pmatrix}1\\ -1\end{pmatrix}.$$

Then we can express

$$\begin{aligned}\gamma_p^*(X^n)&=(U_p'Q_{nn}^{1/2}C+K_p^{-1/2}m_p'\boldsymbol{\alpha})'\,(I_p+U_p'U_p)^{-1}\,(U_p'Q_{nn}^{1/2}C+K_p^{-1/2}m_p'\boldsymbol{\alpha})\\ &=C'Q_{nn}^{1/2}U_p(I_p+U_p'U_p)^{-1}\,U_p'Q_{nn}^{1/2}C\\ &\quad+2\boldsymbol{\alpha}'m_pK_p^{-1/2}(I_p+U_p'U_p)^{-1}\,U_p'Q_{nn}^{1/2}C\\ &\quad+\boldsymbol{\alpha}'m_pK_p^{-1/2}(I_p+U_p'U_p)^{-1}\,K_p^{-1/2}m_p'\boldsymbol{\alpha}\\ &=J_1+J_2+J_3,\text{ say.}\end{aligned} \tag{5.15}$$

Since $I_p-(I_p+U_p'U_p)^{-1}$ and $I_n-U_p(I_p+U_p'U_p)^{-1}\,U_p'$ are non-negative definite, we have $J_1\leqslant C'Q_{nn}C=n_1h_1^{-1}(n_1+h_1)^{-1}+n_2h_2^{-1}(n_2+h_2)^{-1}$, and $J_3\leqslant\gamma_\infty$. Also, by Cauchy–Schwartz, $|J_2/2|\leqslant(J_1J_3)^{1/2}$, whence

$$(J_3^{1/2}-J_1^{1/2})^2\leqslant\gamma_p^*(X^n)\leqslant(J_3^{1/2}+J_1^{1/2})^2. \tag{5.16}$$

Hence $\gamma_p^*(X^n)\leqslant((\gamma_\infty)^{1/2}+(C'Q_{nn}C)^{1/2})^2<\infty$. In particular, since $\gamma_p^*(X^n)$ is increasing with p, $\gamma_\infty^*(X^n)=\lim_{p\to\infty}\gamma_p^*(X^n)$ exists and is finite.

We now determine the behaviour of $\gamma_\infty^*(X^n)$ through a closer analysis of the behaviour of $\gamma_p^*(X^n)$ as $p\to\infty$. Consider first J_1. Applying Theorem 4

of Dawid [2] to $U_p U_p'$, which has the matrix F-distribution $F(p, \delta; I_n)$, we obtain $(I_n + U_p U_p')^{-1} \sim B(\delta + n - 1, p; I_n)$, so that $U_p(I_p + U_p' U_p)^{-1} U_p' = I_n - (I_n + U_p U_p')^{-1} \sim B(p, \delta + n - 1; I_n)$. As $p \to \infty$, this distribution becomes concentrated at I_n. Hence $J_1 \xrightarrow{p} C' Q_{nn} C$.

Similarly, we find $J_3 \sim B(\delta + p - 1, n; \gamma_p)$, so that $J_3 \xrightarrow{p} \gamma_\infty$.

Now consider

$$J_2 = 2C' Q_{nn}^{1/2} [I_n - U_p U_p'(I_n + U_p U_p')^{-1}]\, U_p K_p^{-1/2}(\mathbf{m}_1 - \mathbf{m}_2).$$

We have

$$[I_n - U_p U_p'(I_n + U_p U_p')^{-1}] \sim B(n + \delta - 1; p; I_n) \xrightarrow{p} 0.$$

Also,

$$U_p K_p^{-1/2}(\mathbf{m}_1 - \mathbf{m}_2) \sim T(\delta; I_n, \gamma_p) \xrightarrow{L} T(\delta; I_n, \gamma_\infty)$$

and is thus bounded in probability. We deduce that $J_2 \xrightarrow{p} 0$, and hence, finally, that

$$\begin{aligned} \gamma_\infty^*(X^n) &\overset{\text{a.s.}}{=} \gamma_\infty + C' Q_{nn} C \\ &= \gamma_\infty + n_1 h_1^{-1}(n_1 + h_1)^{-1} + n_2 h_2^{-1}(n_2 + h_2)^{-1}. \end{aligned} \tag{5.17}$$

Comparing with (4.8), we have thus shown that, if $\gamma_\infty < \infty$, then with probability 1 under the distribution of $(\mathbf{X}^0, X^n)$ given $\mathbf{y}^n$ and $Y^0 = 1$, the predictive odds

$$\frac{P(Y^0 = 2 \mid \mathbf{X}_p^0, X^n, \mathbf{y}^n)}{P(Y^0 = 1 \mid \mathbf{X}_p^0, X^n, \mathbf{y}^n)} \tag{5.18}$$

converges almost surely, as $p \to \infty$, to a limit whose distribution is the mixture of

$$\frac{\pi_2}{\pi_1}\left(\frac{k_2^*}{k_1^*}\right)^{\delta^*/2} \exp\left\{N\left(-\frac{1}{2}(k_1^* \Lambda^*)^{-1} \gamma_\infty^*, (k_1^* \Lambda^*)^{-1} \gamma_\infty^*\right)\right\} \tag{5.19}$$

over the distribution $(\chi^2_{\delta^*})^{-1}$ for Λ^*, where $\delta^* = \delta + n$, $k_i^* = 1 + (n_i + h_i)^{-1}$, and $\gamma_\infty^* = \gamma_\infty + h_1^{-1} + h_2^{-1} - (h_1 + n_1)^{-1} - (h_2 + n_2)^{-1}$. Again, a parallel result holds if we consider the behaviour of the predictive odds (5.18) conditional on $\mathbf{y}^n$ and $Y^0 = 2$.

We can remove the conditioning on $\mathbf{y}^n$ by further mixing over the binomial distribution $\mathcal{B}(n; \pi_1)$ for $n_1 = n - n_2$. We thus obtain the overall asymptotic distribution, as $p \to \infty$, of the predictive odds (5.18) under either hypotheses $Y^0 = 1$ or $Y^0 = 2$. Since (for $\gamma_\infty < \infty$) this is almost surely

finite in either case, we shall *not* be in a position to perform asymptotically degenerate classification as we observe more and more variables on our new case.

5.3. *Extensive Training Data*

As the training data accumulate, so we expect to learn the parameters (μ, Σ) more and more accurately. As shown in Section 3, we believe that, were the parameters to be known exactly, then asymptotically, as the number of variables observed tends to infinity, perfect classification would be possible. To what extent does this result continue to hold when we have to learn the parameters from data?

The question concerns the behaviour of the repeated limit

$$\lim_{n\to\infty}\lim_{p\to\infty}\frac{P(Y^0=2\,|\,\mathbf{X}_p^0, X^n, \mathbf{Y}^n)}{P(Y^0=1\,|\,\mathbf{X}_p^0, X^n, \mathbf{Y}^n)}. \tag{5.20}$$

Examining the behaviour of (5.19) as $n\to\infty$, noting that $\chi^2_{\delta+n}\xrightarrow{p}\infty$, we find that, in probability, (5.20) is infinite if $Y^0=2$ and zero if $Y^0=1$.

An alternative argument, which shows these limits to be almost sure, is as follows. Since, under any distribution for $(\mathbf{X}^0, X^n, \mathbf{Y}^n)$, $P(Y^0=i\,|\,\mathbf{X}_p^0, X^n, \mathbf{Y}^n)$ forms a two-parameter martingale (with partial order $\preccurlyeq$ given by $(n,p)\preccurlyeq(n',p')$ if both $n\leqslant n'$ and $p\leqslant p'$), it follows that, with probability 1, this repeated limit is the same as the double limit as $(n,p)\to(\infty,\infty)$ or the alternative repeated limit $\lim_{p\to\infty}\lim_{n\to\infty}$.

Now, since the parameters (μ_p, Σ_p) are consistently estimable from extensive training data, as $n\to\infty$

$$\frac{P(Y^0=2\,|\,\mathbf{X}_p^0, X^n, \mathbf{Y}^n)}{P(Y^0=1\,|\,\mathbf{X}_p^0, X^n, \mathbf{Y}^n)}\xrightarrow{\text{a.s.}}\frac{P(Y^0=2\,|\,\mathbf{X}_p^0; \mu_p, \Sigma_p)}{P(Y^0=1\,|\,\mathbf{X}_p^0; \mu_p, \Sigma_p)}.$$

From the analysis of Section 3, we know that the almost sure limit of this as $p\to\infty$ will be infinity if $Y^0=2$ and 0 if $Y^0=1$. It follows that this is also the almost sure behaviour of (5.20).

6. Discussion

It is important to distinguish our opinions about the world from its behaviour, which is in no way constrained by them. Even though prior assumptions may imply almost sure asymptotically perfect discrimination, this expectation may turn out to be thwarted. Indeed, in many contexts it would, even before obtaining any data, be unreasonable to believe that knowledge of all the explanatory variables would be sufficient to determine

population membership precisely. In such a case, the above analysis should be taken as warning against the use of a conjugate prior. If such a prior is nevertheless to be used, it would seem particularly unwise to choose one for which $\gamma_\infty = \infty$, since this corresponds to a belief that population is determined in an a priori known way by the explanatory variables. It is difficult to conceive of a realistic problem where this belief would not be ridiculous.

REFERENCES

[1] DAWID, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 1–31.

[2] DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274.

[3] DAWID, A. P. (1988). The infinite regress and its conjugate analysis (with discussion). In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Eds.), pp. 95–110, Oxford Univ. Press, Oxford/New York.

[4] DICKEY, J. M. (1967). Matricvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution. *Ann. Math. Statist.* **38** 511–518.

[5] GEISSER, S. (1964). Posterior odds for multivariate normal classifications. *J. Roy. Statist. Soc. Ser. B* **25** 368–376.

Printed by Catherine Press, Ltd., Tempelhof 41, B-8000 Brugge, Belgium

Appendix B

Asymptotic Properties of Conjugate Bayes Discrete Discrimination

Reprinted from JOURNAL OF MULTIVARIATE ANALYSIS
All Rights Reserved by Academic Press, New York and London

Vol. 46, No. 1, July 1993
Printed in Belgium

Asymptotic Properties of Conjugate Bayes Discrete Discrimination

B. Q. FANG

Academia Sinica, Beijing, China

AND

A. P. DAWID

University College London, England

This paper studies a problem of discrimination between two populations with binary variables. The number of variables which can be observed is allowed to tend to infinity. Assuming the Dirichlet process prior, we find necessary and sufficient conditions for asymptotically perfect discrimination between the two populations.

1. INTRODUCTION

Dawid [3] and Dawid and Fang [4] considered, respectively, the problems of normal linear regression and of normal discrimination when the number of predictor variables is effectively infinite. In particular, it was shown how, in either case, a Bayesian approach using the usual conjugate prior distributions incorporates strong prior beliefs that the response is a deterministic function of the predictors—a property which it will often be inappropriate to assume.

In this paper we study the problem of discrimination using infinitely many binary predictors, again concentrating on the implications of the usual conjugate prior assumptions. For this problem, Brown [1] showed that, on making a suitable assumption of "uniform refinement," the prior expectation of the probability of correct classification tends to 1 as the number of predictors tends to infinity.

Here we shall find some more general conditions for asymptotically perfect discrimination. Our concern is the ratio of the probabilities of the

Received May 14, 1992.

AMS 1980 classification: 62H30, 62C10.

Key words and phrases: discrimination, Bayesian inference, Dirichlet process prior, size-biased permutation, determinism.

0047-259X/93 $5.00

populations conditioned on an observation. For a deeper study we shall make use of the Dirichlet process prior as developed by Ferguson [5]. We shall also make use of the notation for and properties of conditional independence as developed in Dawid [2].

Section 2 shows that a sufficient condition for asymptotically perfect discrimination is that the maximum of the associated Dirichlet parameters for each of the populations tends to zero. Section 3 considers a more general case and shows that a necessary and sufficient condition for asymptotically perfect discrimination is that these parameters (as measures) do not have common atoms. Section 4 considers the case that training data are available and a new observation is to be assigned to one of the populations. The same rule as above holds for asymptotically perfect discrimination based on the ratio of the population probabilities conditional on the training data and the new observation. Section 5 investigates discrimination with unknown parameters and shows that for asymptotically perfect discrimination a stronger condition, mutual singularity of α and β, is needed. And Section 6 gives an extension to the case with unknown prior probabilities.

Model Assumptions

Suppose that we observe $X_1, X_2, \ldots$, where X_i takes value 0 or 1. In each of two populations, Π_1 and Π_2, these have (different) point distributions. The probability of population Π_i, $\pi_i = P(\Pi_i)$, $i = 1, 2$, is supposed known. Let

$$\mathbf{X} = (X_1, X_2, \ldots), \qquad \mathbf{X}^n = (X_1, \ldots, X_n). \tag{1.1}$$

Let θ [resp. ϕ] denote the joint distribution of $\mathbf{X}$ in Π_1 [resp. Π_2] : these are measures over the Borel σ-field $\mathscr{B}^\infty$ of $\{0, 1\}^\infty$. By Kolmogorov's consistency theorem, θ is determined by its restriction, θ^n say, to each $\mathscr{B}^n$: we write $\theta^n(\mathbf{x}^n) = P(\mathbf{X}^n = \mathbf{x}^n \mid \Pi_1)$, etc. Then

$$\theta^{n+1}(\mathbf{x}^n, 0) + \theta^{n+1}(\mathbf{x}^n, 1) \equiv \theta^n(\mathbf{x}^n),$$

and any collection $\{\theta^n\}$ with $\theta^n \geqslant 0$, $\theta^0 = 1$, and satisfying this consistency relation is compatible with a unique θ.

If θ, ϕ are known and we observe $\mathbf{X}^n = \mathbf{x}^n$, then we obtain the ratio of the probabilities of the populations conditioned on the observation

$$\frac{P(\Pi_1 \mid \mathbf{X}^n = \mathbf{x}^n)}{P(\Pi_2 \mid \mathbf{X}^n = \mathbf{x}^n)} = \frac{\pi_1}{\pi_2} \cdot \frac{\theta^n(\mathbf{x}^n)}{\phi^n(\mathbf{x}^n)}.$$

We get asymptotically perfect discrimination if $\theta^n(\mathbf{X}^n)/\phi^n(\mathbf{X}^n)$ tends almost surely to ∞ or 0 according as $\mathbf{X}$ arises from Π_1 or Π_2, as n tends to infinity.

Now suppose θ and ϕ are unknown. We can specify a prior distribution for θ by taking a finite measure α over $\mathscr{B}^\infty$, with total mass $|\alpha|$, and requiring that, for each n, θ^n has the Dirichlet distribution $D(\alpha^n)$, with parameter α^n, the restriction of α to $\mathscr{B}^n$. Thus if $\{A_1, ..., A_k\}$ is a partition of $\{0, 1\}^n$, $(\theta^n(A_1), ..., \theta^n(A_k))$ has the ordinary Dirichlet distribution with parameter $(\alpha^n(A_1), ..., \alpha^n(A_k))$. That this specification consistently defines a unique distribution for the random probability measure θ follows by analogy with the construction of the Dirichlet process by Ferguson [5]. We write $\theta \sim D(\alpha)$ for this Dirichlet process distribution.

We similarly take $\phi \sim D(\beta)$, and $\theta \perp\!\!\!\perp \phi$, thus specifying the joint prior distribution of (θ, ϕ).

Let $\lambda_n(\mathbf{x}^n) = \theta^n(\mathbf{x}^n)/\phi^n(\mathbf{x}^n)$, the likelihood ratio in favour of Π_1 as against Π_2, based on data $\mathbf{X}^n = \mathbf{x}^n$, when θ and ϕ are given; and let $\Lambda_n = \lambda_n(\mathbf{X}^n)$, a function of θ, ϕ and $\mathbf{X}$. We shall be interested in the asymptotic behaviour of Λ_n, as $n \to \infty$, under the assumed probabilistic structure for $\mathbf{X}$, θ and ϕ and population Π: that is, when

$$\Pi = \Pi_i \text{ with probability } \pi_i \ (i = 1, 2);$$

$$\theta \sim D(\alpha);$$

$$\phi \sim D(\beta);$$

all the above being independent; and, given (Π, θ, ϕ),

$$\mathbf{X} \sim \theta \text{ if } \Pi = \Pi_1;$$

$$\mathbf{X} \sim \phi \text{ if } \Pi = \Pi_2.$$

We shall show below that, under suitable conditions, $\Lambda_n \xrightarrow{\text{a.s.}} \infty$ [resp. 0] given $\Pi = \Pi_1$ [resp. Π_2] as $n \to \infty$. That is to say, the assumed prior structure attaches probability 1 to the set of distribution pairs (θ, ϕ) allowing asymptotically perfect discrimination between Π_1 and Π_2—assuming that θ and ϕ were first to be revealed to the discriminator, thus making possible the calculation of Λ_n.

2. Smooth Parameters for the Priors

In Brown [1] uniform refinement means that for every n, $\{\theta^n(\mathbf{x}^n)\}$, $\{\phi^n(\mathbf{x}^n)\}$ have symmetric Dirichlet distributions. In this section we shall show that this condition can be replaced by a weaker condition. Our conclusion will be that if the parameters α and β for the prior distribution of θ and ϕ are non-atomic measures, then we can get asymptotically perfect

discrimination between populations Π_1 and Π_2. First we show two lemmas on the properties of the Beta distribution and gamma function.

LEMMA 1. *Suppose that Z has a Beta distribution, $Z \sim$ Beta $(b, B-b)$. Then as a functon of b, $P(Z \leqslant a)$ decreases in b, $0 < b < B$, $a > 0$.*

Proof. For any $0 < b_1 < b_2 < B$, consider three independent gamma distributed random variables $Y_1 \sim$ Gamma $(b_1, 1)$, $Y_2 \sim$ Gamma $(b_2 - b_1, 1)$, $Y_3 \sim$ Gamma $(B - b_2, 1)$. Then we have

$$Z_1 \overset{\text{def}}{=} \frac{Y_1}{Y_1 + Y_2 + Y_3} \sim \text{Beta } (b_1, B - b_1),$$

$$Z_2 \overset{\text{def}}{=} \frac{Y_1 + Y_2}{Y_1 + Y_2 + Y_3} \sim \text{Beta } (b_2, B - b_2).$$

Hence

$$P(Z_1 \leqslant a) = P\left(\frac{Y_1}{Y_1 + Y_2 + Y_3} \leqslant a\right) \geqslant P\left(\frac{Y_1 + Y_2}{Y_1 + Y_2 + Y_3} \leqslant a\right) = P(Z_2 \leqslant a),$$

completing the proof. ∎

LEMMA 2. *Suppose $\{z_i\}$ is a sequence of positive numbers (which may depend on N) satisfying*

$$\sum_{i=1}^{N} z_i \to a$$

and

$$\max_{1 \leqslant i \leqslant N} z_i \to 0, \qquad \text{as} \quad N \to \infty.$$

Then

$$\sum_{i=1}^{N} \frac{1}{\Gamma(z_i)} \to a,$$

$$\sum_{i=1}^{N} \frac{\Gamma(b)}{\Gamma(z_i)\,\Gamma(b - z_i)} \to a, \qquad \text{as} \quad N \to \infty, \ (b \geqslant a).$$

Proof. $\forall \varepsilon > 0$, by the formula

$$\lim_{z \to +0} \frac{\Gamma(z)}{1/z} = 1$$

and the continuity of $\Gamma(z)$ at b (>0), we can find $\delta>0$, such that for every $z\in(0,\delta)$,

$$\left|\frac{1}{\Gamma(z)\,z}-1\right|<\varepsilon \qquad \text{and} \qquad \left|\frac{\Gamma(b)}{\Gamma(b-z)}-1\right|<\varepsilon.$$

By the assumption, $\exists N_0$, for every $N>N_0$,

$$\max_{1\leqslant i\leqslant N} z_i<\delta \qquad \text{and} \qquad \left|\sum_{i=1}^{N} z_i-a\right|<\varepsilon.$$

Hence for every $N>N_0$,

$$\left|\sum_{i=1}^{N}\frac{1}{\Gamma(z_i)}-a\right|$$

$$\leqslant\sum_{i=1}^{N}\left|\frac{1}{\Gamma(z_i)}-z_i\right|+\left|\sum_{i=1}^{N} z_i-a\right|<\varepsilon\sum_{i=1}^{N} z_i+\varepsilon<\varepsilon(a+\varepsilon)+\varepsilon,$$

$$\text{and}\left|\sum_{i=1}^{N}\frac{\Gamma(b)}{\Gamma(z_i)\,\Gamma(b-z_i)}-\sum_{i=1}^{N}\frac{1}{\Gamma(z_i)}\right|\leqslant\varepsilon\sum_{i=1}^{N}\frac{1}{\Gamma(z_i)},$$

completing the proof. ∎

THEOREM 1. *Suppose that α and β are both non-atomic measures. Then, as $n\to\infty$,*

$$\begin{aligned} &\Lambda_n\xrightarrow{\text{a.s.}}\infty \qquad \text{if} \quad \Pi=\Pi_1,\\ &\Lambda_n\xrightarrow{\text{a.s.}}0 \qquad \text{if} \quad \Pi=\Pi_2. \end{aligned} \tag{2.1}$$

Proof. Let $\bar{\alpha}$ be the measure induced on $[0,1]$ from α by the map $(x_1,x_2,\ldots)\to\sum_{i=1}^{\infty}x_i/2^i$. Then $\bar{\alpha}$ is non-atomic, and hence has a continuous, and thus uniformly continuous, c.d.f. on $[0,1]$. Thus $\max_{0\leqslant r<2^n}\bar{\alpha}([r/2^n,(r+1)/2^n])\to 0$ as $n\to\infty$, whence

$$\max_{\mathbf{x}^n}\alpha^n(\mathbf{x}^n)\to 0 \tag{2.2}$$

and similarly

$$\max_{\mathbf{x}^n}\beta^n(\mathbf{x}^n)\to 0. \tag{2.3}$$

Since the converse is clear, we thus see that (2.2) and (2.3) are equivalent to the condition of non-atomicity.

Now suppose $\Pi = \Pi_1$. The numerator of Λ_n is $\theta_n^* = \theta^n(\mathbf{X}^n)$, where $\theta^n \sim D(\alpha^n)$ and, given θ^n, $\mathbf{X}^n \sim \theta^n$. This is thus the first component of the size-biased permutation of θ (cf. Patil and Taillie [6]). Then for $0 < y < 1$ and small dy,

$$P(y < \theta_n^* < y + dy)$$

$$= \sum_{\mathbf{x}^n} P(\mathbf{X}^n = \mathbf{x}^n \mid y < \theta^n(\mathbf{x}^n) < y + dy)\, P(y < \theta^n(\mathbf{x}^n) < y + dy)$$

$$= \sum_{\mathbf{x}^n} y \cdot y^{\alpha^n(\mathbf{x}^n)-1}(1-y)^{|\alpha|-\alpha^n(\mathbf{x}^n)-1}\, dy / B(\alpha^n(\mathbf{x}^n), |\alpha| - \alpha^n(\mathbf{x}^n)). \quad (2.4)$$

It follows from (2.2) and Lemma 2 that this expression tends to $|\alpha|(1-y)^{|\alpha|-1}\, dy$ as $n \to \infty$. Since this is a density over [0, 1], by Scheffé [7], θ_n^* has a non-degenerate limiting distribution, with p.d.f.

$$|\alpha|(1-y)^{|\alpha|-1}. \quad (2.5)$$

Now consider the denominator $\phi_n^* = \phi^n(\mathbf{X}^n)$ of Λ_n. Given $\Pi = \Pi_1$, we have $\phi \sim D(\beta)$ independently of $(\mathbf{X}, \theta)$. Thus $\phi_n^* | \mathbf{X}^n = \mathbf{x}^n \sim \text{Beta}\,(\beta^n(\mathbf{x}^n), |\beta| - \beta^n(\mathbf{x}^n))$. From (2.3) and Lemma 1 we deduce that $\phi_n^* \xrightarrow{p} 0$ $(n \to \infty)$. Taken together with the above result for θ_n^*, this implies $\Lambda_n \xrightarrow{p} \infty$, when $\Pi = \Pi_1$. Similarly $\Lambda_n \xrightarrow{p} 0$ when $\Pi = \Pi_2$. Let now $\mathscr{A}_n$ be the σ-field generated by $(\theta, \phi, \mathbf{X}^n)$. Then, when $\Pi = \Pi_1$ [resp. Π_2], (Λ_n^{-1}) [resp. (Λ_n)] is a non-negative martingale adapted to $(\mathscr{A}_n)$, and must therefore converge almost surely: hence the above probability convergence is in fact almost sure. ∎

3. A More General Case

In the last section we obtained a sufficient condition for asymptotically perfect discrimination. Now we consider a more general case in which the Dirichlet parameters α and β are not necessarily continuous. As in the proof of Theorem 1, we first derive the asymptotic distribution of the numerator θ_n^* of Λ_n.

THEOREM 2. *Suppose that α has decomposition $\alpha = \lambda + \mu$, where λ is continuous and μ is discrete. Arrange the atoms $\{\mathbf{x}_j\}$ of μ in descending order*

of $m_j = \mu(\mathbf{x}_j)$. *If* $\Pi = \Pi_1$, *then as* $n \to \infty$, *the asymptotic distribution of* θ_n^* *is nondegenerate with p.d.f.*

$$|\lambda|(1-y)^{|\alpha|-1} + \sum_{k=1}^{\infty} y^{m_k}(1-y)^{|\alpha|-m_k-1}/B(m_k, |\alpha|-m_k),$$
$$0 < y < 1. \tag{3.1}$$

Proof. The p.d.f. of θ_n^* is given by (2.4). $\forall \varepsilon > 0$, as can be seen from the proofs of Lemma 2 and Theorem 1 (2.5), we can find $\delta_1 > 0$, such that for any sequence $\{z_i > 0, i = 1, 2, \ldots\}$,

$$\sum_i z_i < \delta_1 \qquad \text{implies that}$$
$$\sum_i \frac{\Gamma(|\alpha|)}{\Gamma(z_i)\,\Gamma(|\alpha|-z_i)} < \varepsilon(1-y)/4, \tag{3.2}$$

and

$$\max_i z_i < \delta_1, \sum_i z_i \leqslant |\alpha| \qquad \text{imply that}$$
$$\left|\sum_i \frac{y^{z_i}(1-y)^{|\alpha|-z_i-1}\,\Gamma(|\alpha|)}{\Gamma(z_i)\,\Gamma(|\alpha|-z_i)} - \sum_i z_i(1-y)^{|\alpha|-1}\right| < \varepsilon/4. \tag{3.3}$$

Without losing generality we suppose $\delta_1 < \varepsilon/2(1-y)^{|\alpha|-1}$. Choose $M > 0$ such that

$$\sum_{j>M} m_j < \delta_1/2. \tag{3.4}$$

Let

$$A_1 = \{\mathbf{x}, \ldots, \mathbf{x}_M\}, \qquad A_2 = \{0, 1\}^{\infty} - A_1,$$
$$A_1^n = \{\mathbf{x}_1^n, \ldots, \mathbf{x}_M^n\}, \qquad A_2^n = \{0, 1\}^n - A_1^n.$$

We have

$$||\lambda| - \alpha(A_2)| = ||\alpha| - |\mu| - (|\alpha| - \alpha(A_1))| = \sum_{j>M} m_j < \delta_1/2. \tag{3.5}$$

Find $\delta_2 > 0$ such that

$$|z_k - m_k| < \delta_2 \qquad \text{implies that}$$
$$\left|\frac{y^{z_k}(1-y)^{|\alpha|-z_k-1}\,\Gamma(|\alpha|)}{\Gamma(z_k)\,\Gamma(|\alpha|-z_k)} - \frac{y^{m_k}(1-y)^{|\alpha|-m_k-1}\,\Gamma(|\alpha|)}{\Gamma(m_k)\,\Gamma(|\alpha|-m_k)}\right|$$
$$< \frac{\varepsilon}{4M}, \qquad k = 1, \ldots, M. \tag{3.6}$$

Now $\alpha^n(\mathbf{x}_k^n) \to \alpha(\mathbf{x}_k)$ $(k=1, ..., M)$, $\max_{\mathbf{x}^n} \lambda^n(\mathbf{x}^n) \to 0$ $(n \to \infty)$. So $\exists n_0$ such that, for $n > n_0$,

$$|\alpha^n(\mathbf{x}_k^n) - m_k| < \delta_2, \qquad k=1, ..., M, \tag{3.7}$$

and

$$\max_{\mathbf{x}^n} \lambda^n(\mathbf{x}^n) < \delta_1/2.$$

Hence

$$\begin{aligned}\max_{\mathbf{x}^n \in A_2^n} \alpha^n(\mathbf{x}^n) &\leqslant \max_{\mathbf{x}^n \in A_2^n} \lambda^n(\mathbf{x}^n) + \max_{\mathbf{x}^n \in A_2^n} \mu^n(\mathbf{x}^n) \\ &\leqslant \max_{\mathbf{x}^n \in A_2^n} \lambda^n(\mathbf{x}^n) + \sum_{j>M} m_j < \delta_1, \qquad n > n_0. \end{aligned} \tag{3.8}$$

We can write $\sum_{\mathbf{x}^n}$ in (2.4) as

$$\sum_{\mathbf{x}^n} y^{\alpha^n(\mathbf{x}^n)}(1-y)^{|\alpha| - \alpha^n(\mathbf{x}^n) - 1} / B(\alpha^n(\mathbf{x}^n), |\alpha| - \alpha^n(\mathbf{x}^n))$$

$$= \sum_{A_1^n} + \sum_{A_2^n} = J_1 + J_2, \qquad 0 < y < 1.$$

We have for $n > n_0$, by (3.6), (3.7), (3.4), and (3.2),

$$\begin{aligned}&\left| J_1 - \sum_{k=1}^{\infty} \frac{y^{m_k}(1-y)^{|\alpha| - m_k - 1}}{B(m_k, |\alpha| - m_k)} \right| \\ &\quad \leqslant \left| \sum_{A_1^n} \frac{y^{\alpha^n(\mathbf{x}^n)}(1-y)^{|\alpha| - \alpha^n(\mathbf{x}^n) - 1}\, \Gamma(|\alpha|)}{\Gamma(\alpha^n(\mathbf{x}^n))\, \Gamma(|\alpha| - \alpha^n(\mathbf{x}^n))} \right. \\ &\qquad \left. - \sum_{k=1}^{M} \frac{y^{m_k}(1-y)^{|\alpha| - m_k - 1}\, \Gamma(|\alpha|)}{\Gamma(m_k)\, \Gamma(|\alpha| - m_k)} \right| \\ &\qquad + \sum_{j>M} \frac{y^{m_j}(1-y)^{|\alpha| - m_j - 1}\, \Gamma(|\alpha|)}{\Gamma(m_j)\, \Gamma(|\alpha| - m_j)} \\ &\quad \leqslant M \cdot \frac{\varepsilon}{4M} + (1-y)^{-1} \cdot \frac{\varepsilon(1-y)}{4} = \frac{\varepsilon}{2}\end{aligned}$$

and, by (3.8), (3.3), and (3.5),

$$\begin{aligned}&|J_2 - |\lambda|(1-y)^{|\alpha|-1}| \\ &\quad \leqslant |J_2 - \alpha(A_2)(1-y)^{|\alpha|-1}| + |\alpha(A_2)(1-y)^{|\alpha|-1} - |\lambda|(1-y)^{|\alpha|-1}| \\ &\quad \leqslant \frac{\varepsilon}{4} + (1-y)^{|\alpha|-1} \cdot \frac{\delta_1}{2} < \varepsilon/2,\end{aligned}$$

which shows that the limit of the p.d.f. of θ_n^* is (3.1). Simple calculation shows that the integral of this limit from 0 to one equals one. Applying Scheffé's Theorem, we deduce the desired conclusion. ▮

Now we are ready to prove the main theorem.

THEOREM 3. *The perfect discrimination property* (2.1) *holds if and only if*

$$\alpha \text{ and } \beta \text{ do not have common atoms.} \tag{3.9}$$

Proof. We first prove sufficiency. Suppose that $\Pi = \Pi_1$. Denote a r.v. with p.d.f. (3.1) by θ^*. $\forall \varepsilon > 0$, $\forall K > 0$, find $b > 0$ such that

$$P(\theta^* \leqslant Kb) < \varepsilon/8. \tag{3.10}$$

Let $c = (\varepsilon/4)\, b\, |\beta|$ $(< |\beta|)$. Choose M such that

$$\sum_{j>M} \beta(\mathbf{x}_j) < c/2, \tag{3.11}$$

where $\mathbf{x}_j, j = 1, 2, \ldots$, are atoms of β with $\beta(\mathbf{x}_j) \geqslant \beta(\mathbf{x}_{j+1})$, $j = 1, 2, \ldots$. Let

$$A_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}, \qquad A_2 = \{0, 1\}^\infty - A_1,$$
$$A_1^n = \{\mathbf{x}_1^n, \ldots, \mathbf{x}_M^n\}, \qquad A_2^n = \{0, 1\}^n - A_1^n.$$

Since α and β do not have common atoms, $\alpha(\mathbf{x}_k) = 0$, $k = 1, \ldots, M$. Now $\alpha^n(\mathbf{x}_k^n) \to \alpha(\mathbf{x}_k)$, $k = 1, \ldots, M$, $\max_{\mathbf{x}^n \in A_2^n} \beta^n(\mathbf{x}^n) \to 0$ (cf. (3.8)) as $n \to \infty$. So $\exists n_0$, such that for $n > n_0$,

$$|\alpha^n(\mathbf{x}_k^n)| = |\alpha^n(\mathbf{x}_k^n) - \alpha(\mathbf{x}_k)| < d \stackrel{\text{def}}{=} \frac{\varepsilon}{12M}, \qquad k = 1, \ldots, M, \tag{3.12}$$

$$\max_{\mathbf{x}^n \in A_2^n} \beta^n(\mathbf{x}^n) < c, \tag{3.13}$$

and

$$P(\theta_n^* \leqslant Kb) \leqslant P(\theta^* \leqslant Kb) + \varepsilon/8. \tag{3.14}$$

Then we have, by (3.12),

$$P(\mathbf{X}^n \in A_2^n) = 1 - P(\mathbf{X}^n \in A_1^n) = 1 - \sum_{k=1}^{M} \alpha^n(\mathbf{x}_k^n) > 1 - Md. \tag{3.15}$$

Let Ψ be a r.v. such that

$$\Psi \perp\!\!\!\perp (\theta, \mathbf{X}) \qquad \text{and} \qquad \Psi \sim \text{Beta}(c, |\beta| - c).$$

Then

$$P(\Psi > b) \leqslant \frac{E\Psi}{b} = \frac{c}{|\beta|\, b} = \frac{\varepsilon}{4}. \tag{3.16}$$

By an argument similar to the proof of Theorem 1 and (3.13),

$$P(\Lambda_n > K \mid \mathbf{X}^n \in A_2^n) \geqslant P\left(\frac{\theta_n^*}{\Psi} > K \mid \mathbf{X}^n \in A_2^n\right).$$

Moreover, it is easy to prove that for any $0 < c_0 < 1/2$,

$$P(B) > 1 - c_0 \qquad \text{implies that}$$
$$|P(A \mid B) - P(A)| < 2c_0, \qquad \text{any } A, B. \tag{3.17}$$

Hence by (3.15), (3.10), (3.14), and (3.16), for $n > n_0$,

$$\begin{aligned} P(\Lambda_n > K) &\geqslant P(\Lambda_n > K \mid \mathbf{X}^n \in A_2^n) - 2Md \\ &\geqslant P\left(\frac{\theta_n^*}{\Psi} > K \mid \mathbf{X}^n \in A_2^n\right) - 2Md \\ &= 1 - P\left(\frac{\theta_n^*}{\Psi} \leqslant K,\ \Psi \leqslant b \mid \mathbf{X}^n \in A_2^n\right) \\ &\quad - P\left(\frac{\theta_n^*}{\Psi} \leqslant K,\ \Psi > b \mid \mathbf{X}^n \in A_2^n\right) - 2Md \\ &\geqslant 1 - P(\theta_n^* \leqslant Kb) - P(\Psi > b) - 6Md \\ &\geqslant 1 - \varepsilon/4 - \varepsilon/4 - \varepsilon/2 = 1 - \varepsilon. \end{aligned}$$

This shows that $\Lambda_n \to \infty$ in probability and hence almost surely as $n \to \infty$. If $\Pi = \Pi_2$, the parallel result holds.

Next we prove necessity. Without losing generality, suppose $\Pi = \Pi_1$. Denote a common atom of α and β by ξ. Then $\lim_n \alpha^n(\xi^n) = \alpha(\xi) > 0$, $\lim_n \beta^n(\xi^n) = \beta(\xi) > 0$. Let

$$c_1 = \frac{\alpha(\xi)}{2\,|\alpha|}\ (> 0), \qquad c_2 = 1 - \frac{\alpha(\xi)}{2\,|\alpha|}$$

and let Ψ_i, $i = 1, 2$, be two random variables such that

$$\Psi_1 \sim \text{Beta}(\alpha(\xi), |\alpha| - \alpha(\xi)), \qquad \Psi_2 \sim \text{Beta}(\beta(\xi), |\beta| - \beta(\xi)).$$

Find $c_3 > 0$ such that

$$P(\Psi_2 < c_3) < c_1/2. \tag{3.18}$$

Let

$$K=\frac{4}{c_1 c_3}.$$

Then

$$P(\Psi_1 > Kc_3) \leqslant \frac{E\Psi_1}{Kc_3} = \frac{\alpha(\xi)}{|\alpha|\, Kc_3} < c_1/2. \tag{3.19}$$

Since $\Gamma(z)$ is continuous at $z>0$, by Scheffé's Theorem,

$$\theta^n(\xi^n) \xrightarrow{L} \Psi_1, \qquad \phi^n(\xi^n) \xrightarrow{L} \Psi_2, \qquad \text{as} \quad n \to \infty. \tag{3.20}$$

Hence we have

$$\begin{aligned} &P(\Lambda_n > K, \mathbf{X}^n = \xi^n) \\ &\quad = P\left(\frac{\theta_n^*}{\phi_n^*} > K, \phi_n^* < c_3, \mathbf{X}^n = \xi^n\right) + P\left(\frac{\theta_n^*}{\phi_n^*} > K, \phi_n^* \geqslant c_3, \mathbf{X}^n = \xi^n\right) \\ &\quad \leqslant P(\phi^n(\xi^n) < c_3) + P(\theta^n(\xi^n) > Kc_3) \\ &\quad \to P(\Psi_2 < c_3) + P(\Psi_1 > Kc_3) \qquad (\text{as } n \to \infty) \\ &\quad \leqslant c_1/2 + c_1/2 = c_1. \end{aligned} \tag{3.21}$$

Also we have

$$\begin{aligned} P(\mathbf{X}^n \neq \xi^n) &= 1 - P(\mathbf{X}^n = \xi^n) \\ &= 1 - \frac{\alpha^n(\xi^n)}{|\alpha|} \to 1 - \frac{\alpha(\xi)}{|\alpha|}, \qquad \text{as} \quad n \to \infty. \end{aligned} \tag{3.22}$$

Thus for large n,

$$\begin{aligned} P(\Lambda_n > K) &= P(\Lambda_n > K, \mathbf{X}^n = \xi^n) + P(\Lambda_n > K, \mathbf{X}^n \neq \xi^n) \\ &\leqslant P(\Lambda_n > K, \mathbf{X}^n = \xi^n) + P(\mathbf{X}^n \neq \xi^n) \\ &\leqslant c_1 + 1 - \frac{\alpha(\xi)}{|\alpha|} = c_2 < 1. \end{aligned}$$

Recalling that K, c_2 are independent of n, we conclude that Λ_n does not tend to infinity in probability as $n \to \infty$. ∎

From the above theorem we see that under the condition that α and β do not have common atoms we shall have asymptotically perfect discrimination.

4. Discrimination Using Training Data

Suppose we get training data from Π_1 and Π_2 and a new observation $\mathbf{X}$ which is to be classified as belonging to one of the Π_i. By the conjugate property of the Dirichlet process, θ and ϕ conditional on the training data are still Dirichlet processes. Suppose $\mathbf{Z}_1, ..., \mathbf{Z}_m$ arise from Π_1, then the parameter for θ given $\mathbf{Z}_1, ..., \mathbf{Z}_m$ is $\alpha + \sum_{i=1}^{m} \delta_{\mathbf{Z}_i}$, where $\delta_{\mathbf{z}}$ denotes the measure giving mass one to the point $\mathbf{z}$. If α and β do not have common atoms, the parameters of θ and ϕ conditioned on the training data also do not (with probability 1) have common atoms. To show this we can just study the case that the training data consist of only one observation $\mathbf{Z}$ arising from Π_1. Then the parameter for the posterior of θ is $\alpha + \delta_{\mathbf{Z}}$. Let A be the set of atoms of β. Then $P(\mathbf{Z} \in A) = \alpha(A)/|\alpha| = 0$. With probability one, $\mathbf{Z}$ is not at an atom of β. Conversely, if α and β have a common atom, it is easy to see that it remains as a common atom of the parameters of the posteriors of θ and ϕ conditioned on training data. Substituting the prior distributions of θ and ϕ with their posterior distributions given training data in Theorem 3, we see that (3.9) holds if and only if the ratio of probabilities of Π_1 and Π_2 conditioned on training data and the new observation $\mathbf{X}$ tends to ∞ or 0 in probability according as $\mathbf{X}$ arises from Π_1 or Π_2, as n tends to infinity.

5. Discrimination with Unknown Parameters

Our results so far refer to prior beliefs about the asymptotic behaviour of the likelihood ratio $\Lambda_n = \theta^n(\mathbf{X}^n)/\phi^n(\mathbf{X}^n)$ relevant for classification when the parameters θ and ϕ, as well as the data $\mathbf{X}^n$, are known. In most applications θ and ϕ will remain unknown. Then the relevant likelihood ratio becomes

$$\gamma_n(\mathbf{x}^n) = \alpha_0^n(\mathbf{x}^n)/\beta_0^n(\mathbf{x}^n),$$

where $\alpha_0 = \alpha/|\alpha|$ is the marginal distribution for $\mathbf{X}$ in Π_1, when $\theta \sim D(\alpha)$; and similarly for β_0. Letting $\Gamma_n = \gamma_n(\mathbf{X}^n)$, standard martingale results for the likelihood ratio process now yields.

Theorem 4.

$$\Gamma_n \xrightarrow{\text{a.s.}} \begin{cases} \infty & \text{if } \mathbf{X} \sim \alpha_0 \\ 0 & \text{if } \mathbf{X} \sim \beta_0 \end{cases}$$

if and only if α and β are mutually singular; while Γ_n is almost surely bounded away from 0 and ∞, under both α_0 and β_0, if and only if α and β are mutually absolutely continuous.

In particular, perfect discrimination in the absence of knowledge of the parameters will not be almost certain unless α and β are mutually singular—a much stronger condition than that of Theorem 3. If, for example, both α and β are multiples of Lebesgue measure on $\{0, 1\}^\infty$, then perfect discrimination is (with prior probability 1) possible with the knowledge of θ and ϕ, but not so (again with prior probability 1) in the absence of such knowledge.

Clearly the mutual singularity property of α and β is (with probability 1) preserved if they are replaced by the new parameters, posterior to training data. This is not so for mutual absolute continuity, since the posterior parameters will now possess distinct sets of atoms. In this case, if we have N_i training cases from $\Pi_i\,(i=1,2)$, say $(\mathbf{Z}_1, ..., \mathbf{Z}_{N_1})$ and $(\mathbf{W}_1, ..., \mathbf{W}_{N_2})$, the relevant likelihood ratio Γ_n^t is now based on the measures $(\alpha+\sum_{j=1}^{N_1}\delta_{\mathbf{Z}_j})/(|\alpha|+N_1)$ and $(\beta+\sum_{j=1}^{N_2}\delta_{\mathbf{W}_j})/(|\beta|+N_2)$. It will tend to ∞ under Π_1 with posterior (and hence prior) probability $N_1/(|\alpha|+N_1)$, and to 0 under Π_2 with probability $N_2/(|\beta|+N_2)$. In particular, as N_1 and $N_2 \to \infty$, asymptotically perfect $(n \to \infty)$ discrimination becomes possible, since we effectively learn the parameters θ and ϕ.

6. Unknown Prior Probabilities

We have thus far supposed that the prior probabilities $\pi_i = P(\Pi_i)$ are known. As a minor extension, we can introduce a variable Y, taking values 1 and 2, indicating the correct population, and jointly distributed with $\mathbf{X}$. Let ψ denote this joint distribution of $(Y, \mathbf{X})$, now supposed completely unknown. Then ψ determines, and is determined by, (θ, ϕ, π_1), The conjugate prior for ψ is again Dirichlet, and may be described by the following properties: for some finite measures α and β over $\mathscr{B}^\infty$,

$$\theta \sim D(\alpha) \tag{6.1}$$

$$\phi \sim D(\beta) \tag{6.2}$$

and

$$\pi_1 \sim \text{Beta}(|\alpha|, |\beta|), \tag{6.3}$$

all independently.

Thus the only difference from our previous analysis is the distribution (6.3) for π_1, previously considered known.

However, since under (6.3) $0 < \pi_1 < 1$ with probability 1, it is clear that the asymptotic discrimination behaviour will be the same as for the case of known π_1, so that our results above will continue to apply to this extension—and indeed to any modification in which (6.3) is replaced by an arbitrary distribution for which $0 < \pi_1 < 1$ almost surely.

7. Discussion

We have shown that, in ordinary circumstances, the conjugate Bayes approach to discrete discrimination incorporates a prior belief that the classification will be essentially determined if only sufficiently many predictor variables can be observed. Whether this belief is reasonable must, of course, depend on context. In particular, such an assumption appears to express a world view prevalent among non-statistical workers in pattern recognition. However, it might seem unreasonable in many statistical problems to believe that all residual uncertainty will be eliminated by extensive observation of predictors. For such problems the conjugate Dirichlet analysis will thus be inappropriate, and more complex Bayesian approaches will need to be developed.

References

[1] Brown, P. J. (1980). Coherence and complexity in classification problems. *Scand. J. Statist.* **7** 95–98.

[2] Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. B.* **41** 1–31.

[3] Dawid, A. P. (1988). The infinite regress and its conjugate analysis (with discussion). In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, Eds.), pp. 95–110. Oxford Univ. Press, Oxford.

[4] Dawid, A. P., and Fang, B. Q., (1992). Conjugate Bayes discrimination with infinitely many variables. *J. Multivariate Anal.* **41** 27–42.

[5] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

[6] Patil, G. P., and Taillie, C. (1977). Diversity as a concept and its implications for random communities. *Bull. Internat. Statist. Inst.* **47** 497–515.

[7] Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Ann. Math. Statist.* **18** 434–438.

Printed by Catherine Press, Ltd., Tempelhof 41, B-8000 Brugge, Belgium