Bayesian Discrimination with Longitudinal data

P. J. Brown[*], M.G. Kenward[†] and E. E. Bassett[‡]

21st October 2000

Abstract

We give a new method of Bayesian discrimination for multivariate longitudinal data. This involves a Kronecker product covariance structure for the time by measurements (markers) data on each individual. This structure is estimated by an empirical Bayes approach, using an EM algorithm, within a Bayesian Gaussian discrimination model. In future one may have markers for an individual at one or more time point.

The application concerns a double-blind clinical trial designed to estimate the effect of regular injection of growth hormone, with the purpose of identifying growth hormone abusers in sport. The data formed part of a multicentre investigation jointly sponsored by the European Union and the International Olympic Committee. The data is such that for each individual there is a matrix of marker variables by time point (nominally 8 markers at each of 7 time points). Data arises out of a double-blind trial in which individuals are given growth hormone at one of two dose levels or placebo daily for 28 days. Monitoring by means of blood samples is at 0, 21, 28, 30, 33, 42 and 84 days.

The method gives probabilities that an individual is on placebo or one of the two dose regimes.

KEYWORDS: Discrimination; Bayesian Methods; Type II Likelihood; Smoothing; Repeated Measures; Kronecker Covariance structures; Double blind clinical trial.

1 INTRODUCTION

Multivariate data on individuals are often collected repeatedly over time so that an individual's observation consists of a data matrix for which covariability may also

[*]University of Kent at Canterbury, Institute of Mathematics and Statistics, Cornwallis Building, Canterbury, CT2 7NF, UK

[†]London School of Hygiene and Tropical Medicine

[‡]University of Kent at Canterbury

have two dimensions, across time and variables. Each individual may belong to one of several different groups. In future an individual may present some or all of the variables at a subset of the times and one wishes to calculate the probability that the individual belongs to each group, and perhaps even where in the time trajectory the individual fits.

Earlier comprehensive references to the general area of multivariate discrimination are provided by McLachlan (1992), Krzanowski and Marriott (1995). There is relatively little work on discrimination involving repeated measures. In the two population case, a components of variance model is developed by Logan and Gupta (1993). This extends Geisser (1964), as does Brown [chapter 8](1993) in the direction of informative prior structures, albeit natural conjugate ones.

Two-way covariance structures have been the subject of some interest in repeated measures and spatial statistics. In the latter field see Brown *et al* (1994).

Many aspects of modelling joint covariances structures has general applicability. Our emphasis on discrimination arises out of our work on the detection of growth hormone abuse in sport (GH-2000 EU/IOC 1999 BMH14 CT950678, Co-ordinator P. H. Sönksen), a multi-centre project analysing the effects of growth hormone in a range of trials, as described in more detail in section 4. In this paper we are concerned with the methodology for dealing with multivariate discrimination when repeated measures are involved, using the growth hormone data as a motivating example.

2 THE MODEL

First let us establish some notation. The number of groups is $g+1$, which will be 3 in the example considered corresponding to doses $d=0$ (Placebo), $d=1$ or $d=2$. The training data, for which we know both variables and population group, comprise n_d independent observations forming a $n_d \times qp$ matrix where the qp observations on an individual are ordered by time and marker, all the p biochemical markers measured at q time points. There may well be other covariates such as age, sex, weight which may need to be incorporated. These variables are logically different from the biochemical measurements in that they are constant or near constant within an individual and known beforehand and may be conditioned on. For the dth population we assume multivariate normality, perhaps after a suitable transformation. Conditional on parameters α_d, B, Ω the standard multivariate normal regression model assumed is (in the matrix variate notation discussed in Appendix A):

$$Y_d - 1\alpha_d' - X_d B \sim \mathcal{N}(I_{n_d}, \Omega), \tag{1}$$

with $n_d \times qp$ random matrix Y_d, 1 an $n_d \times 1$ vector of ones, $n_d \times c$ model matrix X_d regarded as fixed and corresponding to c covariates, and B the $c \times qp$ matrix of regression coefficients for these covariates (eg. age, body-mass index, etc). The matrix Y_d could have been structured as a three-way array of dimensions $n_d \times q \times p$, and although we will avoid such notational considerations at this point our model and analysis will separately structure the times and variables.

Here by virtue of the identity matrix first argument of $\mathcal{N}(.,.)$ rows of Y_d are independently distributed. We have chosen a relatively simple covariance structure where the covariance matrix Ω does not depend on dose level. Straightforward model elaboration would have Ω_d and possibly also B_d. The former implicitly allows quadratic discrimination, the latter allows differing dependence on the covariates by groups. Without loss of generality we assume that columns of the concatenated matrix $X, (\sum_0^g n_d \times c)$ where $X' = (X_0', X_1', \ldots, X_g')$, have been centered by subtracting their column means. Similarly the matrix $Y, (\sum_0^g n_d \times qp)$ is post hoc centered by subtracting off means of columns.

The special forms of prior distributions for parameters α_d, B, Ω are given as follows. First, given Ω,

$$\alpha_d' - \alpha_{d0}' \sim \mathcal{N}(h, \Omega); \tag{2}$$

secondly and independently, given Ω,

$$B - B_0 \sim \mathcal{N}(H, \Omega). \tag{3}$$

Notice that from our matrix variate characterization, both priors (2) and (3) have covariances dependent on Ω in a way that directly extends the univariate regression natural conjugate prior distributions. We could have combined α and B but we felt it helpful to emphasise the different status of the covariates to which they relate. Often a different prior distribution would be appropriate and h would differ from H, although here we assume uninformative prior distributions in both cases. Thus the scale hyperparameter h of the prior distributions for α_d given in (2) will be taken to be a large value, tending to infinity, when the value ascribed to the prior mean, α_{d0}, becomes irrelevant. We will also not be specially interested in imposing anything other than vague prior structure on the regression coefficient matrix B, achieved by setting H to be very large. This has the added simplifying advantage of decoupling the estimation of these parameters from that of the covariance structure Ω.

Now the prior marginal distribution of Ω is taken to be Inverse Wishart,

$$\Omega \sim \mathcal{IW}(\delta; Q), \tag{4}$$

in the notation of Appendix A. Here Q is a $qp \times qp$ scale matrix of the Inverse Wishart, and Ω has expectation $Q/(\delta - 2)$ for $\delta > 2$. The shape parameter δ is such that $\delta > 0$ for a proper prior density. Improper prior densities would have $\delta \leq 0$, and the Jeffreys invariant improper density corresponds to $\delta = 1 - qp$.

The scale matrix hyperparameter Q of the prior distribution for Ω from equation (4) is given as proportional to a Kronecker form $\Gamma \otimes \Sigma$. Here Γ is a $(q \times q)$ matrix of covariation across time, whereas Σ is $(p \times p)$ and measures covariability across markers. Other forms are of course possible, but here this Kronecker product structure is a key simplifying form, born out by empirical evidence in section 4.1. It can be argued that imposing a Kronecker product structure at the hyperparameter level of the prior mean of this covariance matrix, results in robust inference. A small value of δ will not strongly impose the structure whereas a large value will effectively impose the structure at the first level of covariance Ω. This Kronecker product model applies when we have a multiplicative random- effects model:$Y_{tr} - \mu_{tr} = \theta_t \times \gamma_r$, where Y_{tr} denotes the rth response at the tth time point, $r = 1, \ldots, p, t = 1, \ldots, q$. Brown, Le and Zidek (1994) in modelling spatial variation across areas rather than time used a full $pq \times pq$ covariance structure, with a similar Kronecker product form at the hyperparameter level. Like them we will adopt an empirical Bayes approach to estimation of the covariance structure, following Chen (1979). We prefer however to fix the shape parameter δ at a selection of values and examine the 'likelihood' of these so exploring sensitivity to its choice rather than estimating it. This simplifies and speeds up the algorithms and offers more control. It does mean that we will need to retain terms solely in δ, or equivalently the degrees of freedom $\nu = \delta + qp - 1$, in our relevant likelihoods.

3 BAYESIAN INFERENCE

3.1 Posterior and Marginal Distributions

The probability density function of $Y_d, d = 0, \ldots, g$ from model (1) is

$$f_Y(Y_d|\alpha_d, B, \Omega) \propto |\Omega|^{-n_d/2} exp\{-\frac{1}{2}\text{trace}(Y_d - 1\alpha_d' - X_d B)\Omega^{-1}(Y_d - 1\alpha_d' - X_d B)'\}, \quad (5)$$

for $d = 0, \ldots\ g$. Explicit forms for the constant of proportionality and other matrix-variate densities needed below, can be found in Brown (1993), Appendix A.

The likelihood marginalised over the vague priors $\alpha_d, d = 0, \ldots, g, B$ is shown in Appendix B for given Ω as proportional to

$$|\Omega|^{-N/2} exp\{-\frac{1}{2}\text{trace}(\Omega^{-1}S)\}. \quad (6)$$

with centered $Y_d^* = Y_d - 1\bar{Y}_d$ and sum of products of residuals

$$S = \sum_d (Y_d^* - X_d^*\hat{B})'(Y_d^* - X_d^*\hat{B}), \quad (7)$$

where $X_d^* = X_d - 1\bar{X}_d$ and

$$\hat{B} = [\sum X_d^{*\prime} X_d^*]^{-1}(\sum X_d^{*\prime} Y_d).$$

In the process *a posteriori* given Ω, Y it is shown that

$$B - \hat{B} \sim \mathcal{N}([\sum X_d^{*\prime} X_d^*]^{-1}, \Omega), \quad (8)$$

The final stage injects prior knowledge from the Inverse Wishart prior for Ω in (4). We take the scale matrix $Q = \nu Q^*$ so that the inverse of Ω given as $\Omega^{-1} = \Lambda$ is Wishart$(\nu, [\nu Q^*]^{-1})$ and has expectation $(Q^*)^{-1}$, independent of the degrees of freedom. The form of the prior density of Ω is

$$\pi(\Omega) = c(qp, \delta)|Q|^{\nu/2}|\Omega|^{-(\nu+qp+1)/2} exp - \frac{1}{2}\mathrm{trace}(\Omega^{-1}Q) \tag{9}$$

with $c(l, \delta) = 2^{-l\nu/2}/\Gamma_l(\nu/2)$ and $\Gamma_l(\frac{\nu}{2}) \propto \prod_{i=1}^{l} \Gamma[\frac{\nu}{2} - (i-1)/2]$. Here we have used the alternative parameters, δ, ν, interchangeably where $\delta = \nu - l + 1$ with dimension $l = qp$. Multiplying the likelihood (6) by the prior density (9) and integrating out Ω gives the integrated or Type II likelihood (Good, 1965) of Q as

$$L(Q; Y) = \frac{c(qp, \delta)}{c(qp, \delta^*)}|Q|^{\nu/2}|Q + S|^{-(\nu+N)/2} \tag{10}$$

with $\delta^* = \delta + N$. In the limit as $\nu \to \infty$ this becomes the standard likelihood for Q as if Q were the level I parameter. Explicitly with $Q = \nu Q^*$

$$L(Q^*; Y) \propto |Q^*|^{-N/2} exp - \frac{1}{2}\mathrm{trace}[(Q^*)^{-1}S]. \tag{11}$$

Where this normal likelihood has the constant of proportionality $(2\pi)^{-qpN/2}$. This is straightforward to maximise when $Q^* = \Gamma \otimes \Sigma$ using a flip-flop algorithm: alternatively maximising over Γ for fixed Σ and then Σ for fixed Γ, as shown in Brown *et al* (1999b). The same algorithm will be developed below in a slightly different context. The non-asymptotic likelihood (10) does not seem straightforward to maximise under the Kronecker product structure: we rather take an EM algorithm route suggested by Dempster *et al* (1977) and developed by Chen (1979). The use of a conditional maximisation within the M-step renders it an ECM algorithm (Meng and Rubin, 1993).

3.2 The ECM algorithm

Chen (1979) considers parameter ν as well as Ω, whereas we concentrate on Ω alone. Taking $\Lambda = \Omega^{-1}$, the complete data is (S, Λ) and S alone the incomplete data and Λ is the complete data sufficient statistic for Q^*. Now the posterior distribution of Ω for given Q^* is inverse Wishart, $\mathcal{IW}(\delta + N, S + \nu Q^*)$ and Λ is Wishart, $\Lambda \sim \mathcal{W}[\nu + N, (S + \nu Q^*)^{-1}]$. Thus for the E-step we have,

$$E(\Lambda) = (\nu + N)(S + \nu Q^*)^{-1}. \tag{12}$$

For the M-step a new value of Q^* is found by maximising the density of Λ given Q^*, that is the Wishart prior density for Λ with the current value of Λ taken at the E-step (12). This means that we are required to maximise over Q^*

$$H(Q^*) = log|Q^*| - \mathrm{trace}(\Lambda Q^*). \tag{13}$$

This can be achieved by the same type of flip-flop used in the maximisation of (11), although further simplifications are evident in the case of the normal likelihood. Notice that by virtue of the parametrization $Q = \nu Q^*$, as used by Chen (1979), the maximization of (13) is decoupled from the prior degrees of freedom ν.

We now consider the Kronecker product structure for Q^*, setting

$$Q^* = \Gamma \otimes \Sigma \tag{14}$$

The quantity to maximise is

$$H(Q^*) = plog|\Gamma| + qlog|\Sigma| - \text{trace}[\Lambda(\Gamma \otimes \Sigma)]. \tag{15}$$

The value of Λ is set at the current E-step value (12). Keeping this fixed we need to cycle through alternate maximisations of Γ and Σ. For given current Γ we need to maximise

$$qlog|\Sigma| - \text{trace}(\Sigma A) \tag{16}$$

where $A = \sum_{i=1}^{q} \sum_{j=1}^{q} \gamma_{ij} B_{ij}$ and B_{ij} is the ijth $p \times p$ submatrix of Λ, partitioned into blocks, and $\Gamma = (\gamma_{ij})$. The maximum of (16) is achieved at $\Sigma^{-1} = A/q$ with a consequent increase in $H(Q^*)$ in (15) due to concavity (Mardia *et al.*, 1979, p??). Fixing Σ at this new value we can achieve a further increase by maximising over Γ to give $\Gamma^{-1} = A^*/p$ with $A^* = \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma_{ij} B^*_{ij}$ and B^*_{ij} is the ijth $q \times q$ submatrix of Λ, permuted to be varying fastest for time rather than variables, in the order implied by $\Sigma \otimes \Gamma$. Cycling in this way through successive maximisations of Σ and Γ will converge to a unique maximum of

$$\hat{Q}^* = \hat{\Gamma} \otimes \hat{\Sigma},$$

even though there is no uniqueness about Γ and Σ separately, just their product.

After this cycle of maximisations within the M-step we update the value of Λ through the E-step and repeat the cycle of maximisations within the M-step. We repeat the E- and M-steps until convergence. The monotonic increase in likelihood and convergence is guaranteed by the exponential family formulation and EM theory (Meng and Rubin, 1993). Hybrid algorithms which adopt fewer CM cycles per iteration are also possible, but exploration of alternatives were not we felt warranted given the fast convergence (seconds rather than minutes) of the algorithm.

At designated convergence of the overall E- and M-step process we can evaluate the type II likelihood at the given $(\nu, \hat{Q}^*)$ in (10). The smallest integer value of ν such that $\delta > 0$ would be $\nu = qp$. We may also calculate the likelihood for $\nu \to \infty$ from (11) with the appropriate constant of proportionality. A few intervening values of ν may also be desirable. The precise value of ν may influence the tail probabilities for prediction, and we will supplement its estimation by a type II maximum likelihood estimate, denoted $\hat{\nu}$, with the two estimates of ν at 2 units of log-likelihood below the maximum.

This completes the specification of the posterior distribution of the parameters $(\alpha_0, \ldots, \alpha_g, B, \Omega)$ as a normal-inverse Wishart with estimated hyperparameters $(\hat{\alpha}_0, \ldots, \hat{\alpha}_g, \hat{B}, \hat{Q}^*, \hat{\nu})$

3.3 Predictive distribution and discrimination

We are first interested in deriving the predictive distribution for a future $qp-$dimensional $Y-$vector from group d, and denoted as the $1 \times qp$ response vector Z, from the same model as (1), with c covariates, x, $(1 \times c)$, measured on the same scale as with the training data. The model, analogous to (1), and using the same matrix notation for consistency is

$$Z - \alpha_d' - xB \sim \mathcal{N}(1, \Omega), \tag{17}$$

Conditional on (B, Ω, Y), using the conditional posterior of α_d in (32) we have

$$Z - (\bar{Y}_d - \bar{X}_d B) - xB \sim \mathcal{N}(1 + 1/n_d, \Omega). \tag{18}$$

Now we average this over the posterior distribution of B for given Ω, Y. This is symbolically accomplished by multiplying (8) by $(x - \bar{X}_d)$ on the left and adding the resultant distribution to (18) to eliminate B. This gives

$$Z^* - x^*\hat{B} \sim \mathcal{N}(1 + 1/n_d + x^*[\sum X_d^{*\prime} X_d^*]^{-1} x^{*\prime}, \Omega). \tag{19}$$

with mean corrected $x^* = (x - \bar{X}_d), Z^* = Z - \bar{Y}_d$. The posterior distribution of Ω is $\mathcal{IW}(\hat{\delta} + N, \hat{Q} + S)$ where $\hat{\delta} = \hat{\nu} - qp + 1$, and $\hat{Q} = \hat{\nu}\hat{Q}^*$. Hence we finally have the predictive distribution of Z as multivariate Student$-T$. In our matrix notation (see Appendix A)

$$Z^* - x^*\hat{B} \sim \mathcal{T}(\hat{\delta} + N; 1 + 1/n_d + x^*[\sum X_d^{*\prime} X_d^*]^{-1} x^{*\prime}, \hat{Q} + S). \tag{20}$$

We will also be interested in predicting sub-vectors of Z. With this aim let η be a binary vector of zeros and 1's of length qp so that Z^η identifies (through ones) the chosen elements of the vector Z. The predictive distribution of the sub-vector identified by η is then,

$$Z^{*\eta} - x^*\hat{B}^\eta \sim \mathcal{T}(\hat{\delta} + N; 1 + 1/n_d + x^*[\sum X_d^{*\prime} X_d^*]^{-1} x^{*\prime}, \hat{Q}^\eta + S^\eta), \tag{21}$$

where η selects columns of Z^*, B and rows and columns of $\hat{Q}, S$. Typically we will be interested in the sub-vector that corresponds to all p marker measurements at a particular time point, t. Then $\hat{Q}^\eta + S^\eta = \hat{\nu}\hat{\gamma}_{tt}\hat{\Sigma} + S_{tt}$, with S_{tt} the $p \times p$ sum of products matrix for the tth time point. Imagine a single time point from the prediction data, that is an individual whose GH-dose history is unknown provides p biochemical measurements via a blood sample. We assume such an individual history is like taking $p-$markers out of the time trajectory of q sets of p measurements. Thus the $p \times 1$ vector Z for given time t and dose d is as model (21) with the time specific selection vector η, and whose probability density function is denoted $f(Z|t, d, Y, x)$. This future observation has an unknown value of t and d and these are taken *a priori* independent and given as

$$\pi(t, d) = \pi(t)\pi(d). \tag{22}$$

Here $\pi(t)$ should have zero probability on $t = 0$ which provides baseline measurements confounded with placebo.

The posterior distribution is given as

$$\pi(t, d \mid Z, Y, x) \propto f(Z \mid t, d, Y, x)\pi(t, d). \tag{23}$$

If in (21) we let

$$\begin{aligned} Z_{xd} &= Z^{*\eta} - x^*\hat{B}^\eta, \\ h_{xd} &= 1 + 1/n_d + x^*[\sum X_d^{*\prime} X_d^*]^{-1} x^{*\prime}, \\ \delta^* &= \hat{\delta} + N, \\ R_t &= \hat{\nu}\hat{\gamma}_{tt}\hat{\Sigma} + S_{tt} \end{aligned} \tag{24}$$

then this becomes,

$$Z_{xd} \sim \mathcal{T}(\delta^*; h_{xd}, R_t) \tag{25}$$

The kernel of the density $f(Z_{xd} \mid t, d, Y, x)$ is that of a multivariate Student density. This is then

$$f(Z \mid t, d, Y, x) \propto (h_{xd})^{\delta^*/2} |R_t|^{-1/2} \{h_{xd} + Z_{xd}(R_t)^{-1} Z'_{xd}\}^{-(\delta^*+p)/2}, \tag{26}$$

a simple function of the squared distance $Z_{xd}(R_t)^{-1}Z'_{xd}$ in the common metric of $(R_t)^{-1}$. When δ is large so that the Student kernel becomes that of a normal, then the quadratic dependence on Z would cancel top and bottom in (23) if γ_{tt} did not vary with time, and discrimination would be implicitly linear. With our general γ_{tt} multiplier the relative odds at a particular time loses the quadratic dependence in this asymptotic case.

The case of multiple time point data on a future individual is also subsumed within the above framework. For two time points the selection vector η would be zero except for the p variables at each of the two time points. The formulae above apply aside from R_t in (24) being specified from the the more general form given as the $2p \times 2p$ matrix

$$R_t = \hat{Q}^\eta + S^\eta. \tag{27}$$

The time index is now bivariate and would naturally consist of ${}^7C_2$ distinct ordered pairs. Since one would typically know the time lag between sampling points for the individual, there is a temptation to restrict attention to the pairs of observations with similar elapsed time within the double blind trial. This would be unwise however as elapsed time for a random competitor would be unlikely to correspond to elapsed time in the controlled doping regime of the double blind trial. It does suggest that one may wish to gather further information from a future additional double blind trial, featuring several sampled points within the doping phase, and even intermittent doping regimes.

Features of the joint posterior distribution (23) of particular interest are

1. The marginal posterior $\pi(d \mid Z, Y, x)$ formed by summing (23) over t. In particular one may require $\pi(d = 0 \mid Z)$ to be very small, say less than 1/10000 for convincing evidence against an athlete. This is the level medical investigators and lawyers connected with the project would feel justified in taking legal proceedings.
2. The marginal conditional posterior $\pi(t \mid Z, Y, x, d > 0)$, identifying the likely time phase relative to doping of the suspect.

4 GH2000 DOUBLE BLIND STUDY

This section describes the application of our methodology to one of the datasets generated through the GH2000 project mentioned in the Introduction: a double-blind multi-centre study. Pituitary GH is secreted in pulses mainly in response to exercise sleep and stress. GH acts on the liver and on bone stimulating the formation of a number of substances ('markers') which are then secreted into the blood stream producing a characteristic response. Exogenous recombinant human growth hormone (rhGH) administration, or something stimulating endogenous GH release, but nothing else, produces a characteristic pattern in these markers. Four of our chosen eight markers, IGF-1, $IGFBP_2$, $IGFBP_3$ and ALS, enter the blood stream via the Liver, whereas the other four, Osteocalcin, P-III-P, PICP, ICTP, are bone related. Whereas GH disappears very rapidly, these markers persist much longer, extinguishing themselves at different rates and producing a characteristic fingerprint. Thus we use the 'downstream' effects of growth hormone as the basis of testing for GH abuse. The trajectory of the markers for individuals over time is the starting point for developing a methodology for Bayesian discrimination on future presentation of a blood sample by an athlete.

4.1 The Data

These consisted of around 100 individual volunteers who took by injection either a single dose, a double dose or placebo daily for 28 days. The individuals and the researchers taking blood samples were both blind as to which of the three dose regimes the individual had been randomised. Blood samples at baseline, time 0, just before the first dose, at 21 and 28 days and subsequent to growth hormone intake at 30, 33, 42 and 84 days. These blood samples were split, refrigerated, and sent to two laboratories to measure the markers, each laboratory concentrating on 4 of the markers.

4.2 Data Exploration

Many differing analyses have been carried out on this and related data. Our example here is for illustrative purposes concentrating purely on males. The best known of the markers, the insulin like growth factor I, (IGF-I), shows a characteristically raised response to the recombinant human Growth hormone. This can be seen from the plot of IGF-I against time for the males of the study in Figure 1. Those individuals on Placebo have a flat response whereas those on a single or double dose of GH rise sharply over the first 28 days.

A plot of IGF-I against P-III-P at 21 days in Figure 2 shows there is complete separation at this time for this pair of markers in combination although not for either individually. A linear discriminant function optimised for day 21 is applied to the data and given as a dotplot in Figure 3. The baseline and placebo measurements have been combined, centered on zero, and scaled so as to have unit variance. Day 21 and 28 dosed individuals have discriminant function values on average four or five standard deviations above the placebo mean, and although there is some overlap of upper tail of placebos and the lower tail of dosed individuals it is easy to see the potential for setting an allocation rule with very high sensitivity, say 1 in 10,000, would correspond to 3.5 standard deviations above the placebo mean. Rather than using a single time point to calculate the discriminant function our trajectory method effectively looks to see what the best discriminator is at each time. The assumption of a Kronecker product covariance structure was motivated by the multiplicative random effects model of section 2. For selected pairs of markers it is feasible to compare the 14 by 14 ($= 2 \times 7$) full covariance matrix with that of the Kronecker structure of of the 2×2 markers covariance by the 7×7 times covariance, estimated iteratively by maximum likelihood. The agreement was good. This is not so feasible for the full $8 \times 7 = 56$ by 56 matrix due to sheer complexity and the number of variables being greater than the number of observations. As one alternative check we calculated the maximum likelihood estimates of the Kronecker product covariance structure. Assuming this as true and multivariate normality we could generate 20 new sets of data and calculate the eigenvalues of their sample covariance structures. These ordered eigenvalue profiles together with the actual profile are given in Figure 4.

The observed eigenvalue profile lies outside the 20 simulated ones so that the Kronecker product assumption is some way from perfect, with an over representation of larger eigenvalues, and so warrants the cautious assumption of a Kronecker covariance at the secondary rather than primary level, the parameter δ of the inverse

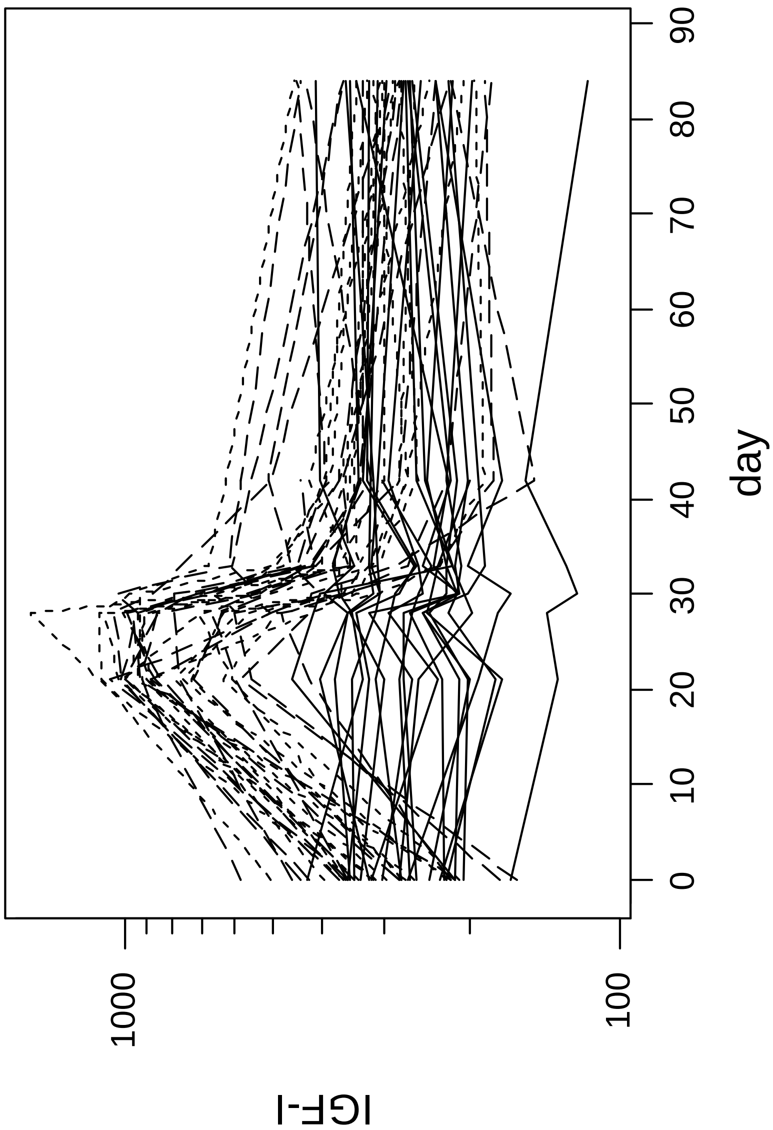

Figure 1: IGF-I against time in days, Placebo (line), GH Low (dash), GH High (dots)

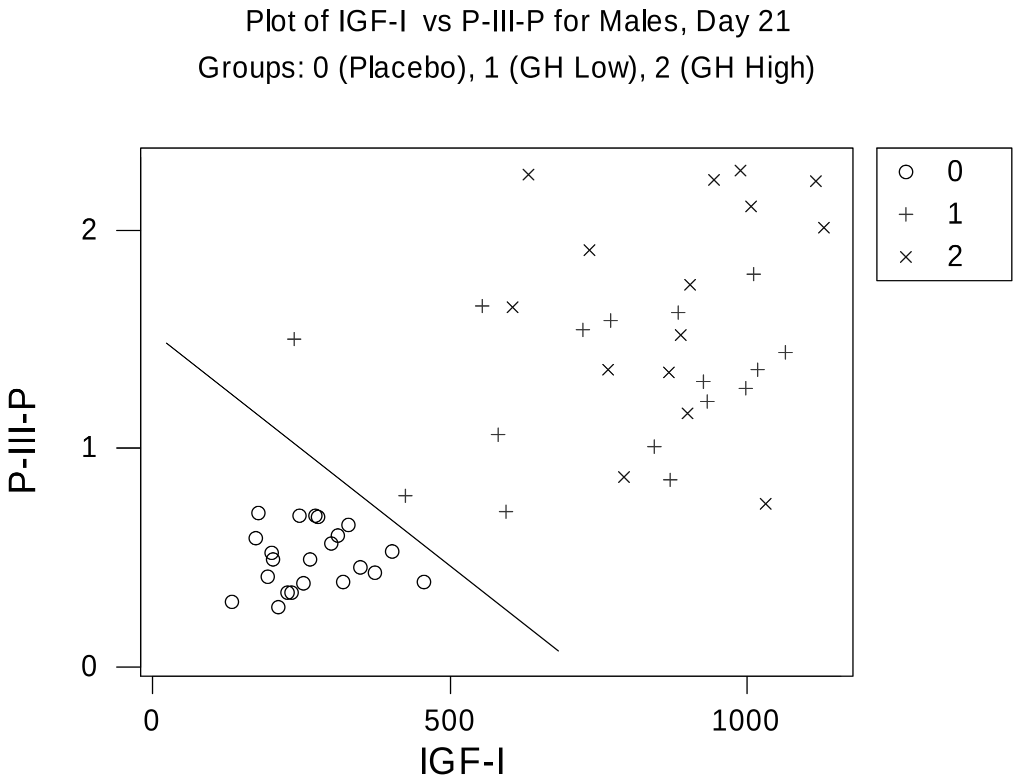

Figure 2: IGF-I v PIIIP, Day 21, Groups: 0 (Placebo), 1 (GH Low), 2 (GH High)

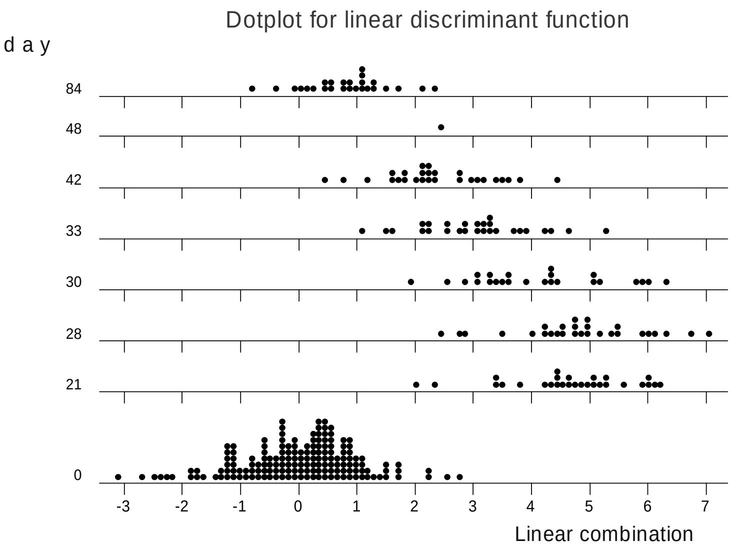

Figure 3: Discriminant scores evaluated at 21 days plotted by time for all males

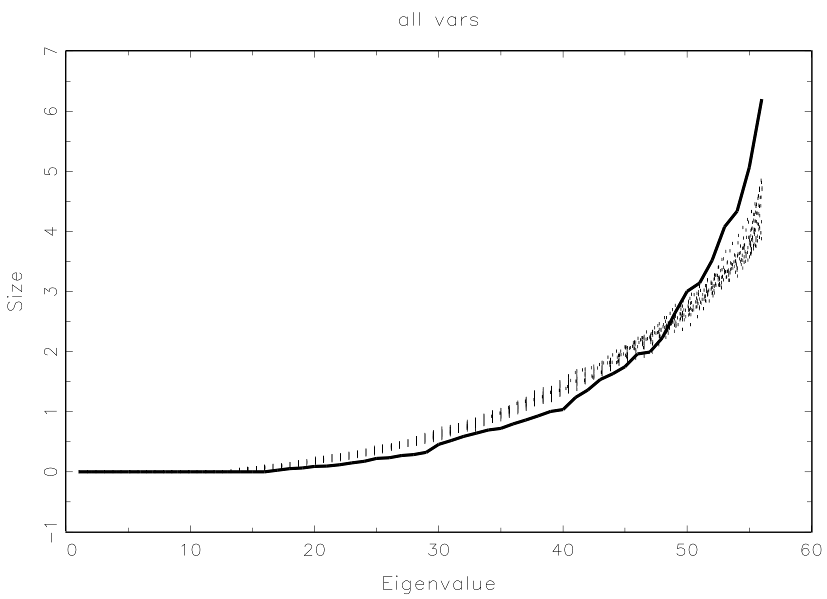

Figure 4: 20 simulated eigenvalue profiles from a true Kronecker covariance structure with the observed profile

Wishart accounting for variability about the Kronecker hyperparameter Q in (4).

4.3 Training and Validation

The data from the double blind study consisted of 43 male individuals with complete sets of blood samples taken at all seven time points. The other nine male individuals had samples at fewer than seven points. We used the complete data as training data and the individuals with incomplete records for validation. The pattern of missingness was not generally one of dropout, rather some intervening time point was missed, so we felt somewhat reassured that perhaps those with missing data were not particularly different from those without. We did check back as far as was feasible to find out whether there were particular reasons for missingness, but none emerged.

It may be noted that even though these validation records are incomplete, since we are using them at a future time point, they provide several records for validation.

Females in the study behaved quite differently in their response to growth hormones and will be reported elsewhere. A more thorough analysis will need to adjust also for age as a covariate. Adjustment for age would seem to be desirable because of strong age dependence of some of the biochemical markers (eg IGF-I). This may not be important for the actual validation of the model for the Double- Blind data alone but it will be needed to form a common ground for adjustment in the light of the cross-sectional study of elite athletes, another component of the GH2000 project. All eight markers were characteristically positively skewed and were symmetric and normal when logged and this form was used throughout.

4.4 Prior Settings

We have assumed vague prior distributions for the location constants, α_d, $d = 0, \ldots, 2$. The prior for Ω in (4) has two hyperparameters, δ and Q. The scalar δ is fixed at a grid of values so to form a posterior distribution for this. Implicitly this has been assumed to have a constant vague prior of the positive real line. The scale matrix hyperparameter Q is structured as $\Gamma \otimes \Sigma$ as discussed in section 2. The hyperparameter matrices are estimated by maximum integrated likelihood via the EM algorithm discussed in section 3.2, implicitly assuming vague third stage priors for these (Smith, 1973).

The prior probabilities for prediction are taken to be equally likely on the 3 groups and equally likely on time points 21, 28, 30, 33, 42, 84 days, that is omitting day 0 since this provides baseline measurements for all individuals.

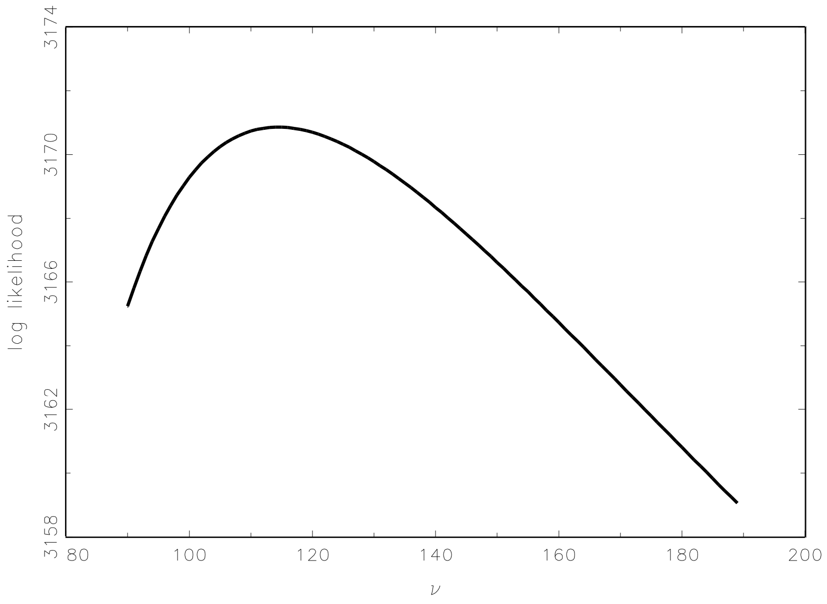

Figure 5: Profile of the Type II likelihood of ν at the maximised values of Γ, Σ

4.5 Results

The Type II profile log–likelihood of ν is plotted in Figure 5 This is formed by choosing a grid of ν values and recording the marginal likelihoods maximised over Γ and Σ by the EM algorithm. Changes in ν which result in changes in the log–likelihood of 2 units from the maximised value of $\nu = 114$ roughly correspond to a generalised likelihood ratio 95% confidence interval, that is ν in (98,137). This approximates a Bayesian 95% credibility interval for ν if the maximised values of Γ and Σ change little over this range, which appears to be the case.

We summarise the results for prediction of 9 validation males from the training data of 43 males with complete records. For simplicity of presentation we amalgamate the two treated groups which would be achieved by adding their predictive probabilities or equivalently we simply give the probabilities on the Placebo group for each individual. The probabilities are predictive probabilities from section 3.3 for Placebo (versus treated) tabulated according to day of sampling, given just the information on that individual at the sampled time point as well as the training data. For each of the 9 individuals in the validation sample these Placebo probabilities are given for $\nu = 114, 98, 137$ by row in that order, the top row corresponding to the maximised ν. These are given in Table 1. The coding of the 9 subjects has as its first digit the country of the participating individual, 1=Denmark, 2=Italy, 3=Sweden, 4=UK; thus 302 stands for the second individual in Sweden, and we see from the second column that he had a double dose of GH in the Double Blind trial.

All those individuals truly on placebo have probabilities of order 1 of being drug free. Those on dose level 1 have small probabilities of being drug free when they are submitted at days 21, 28 or 30 and even at day 42 for individual 301. The one individual on double dose is detectable up to day 33 at a threshold of 1 in 10,000.

These probabilities do not vary much over the likely values of ν and so the method is robust to the maximum integrated likelihood plug-in.

5 DISCUSSION

There are various possible elaborations of the model. The use of Ω_d rather than plain Ω in the model (1) may be appealing. This could feed through to the hyperparameter Σ, the scale matrix for variation amongst markers. What is to be gained may best be judged by a pseudo-Bayes compromise such as regularised discrimination given in Friedman (1989) and underpinned by Brown, Fearn and Haque (1999a). Seeming differences in group covariances are often not useful due to the introduction of too many parameters. The Friedman approach uses a hyperparameter to compromise between $\hat{\Omega}_d$ and $\hat{\Omega}$ and has a further hyperparameter for regularisation.

We have adopted a type II likelihood approach to estimation of the level II hyperparameters Γ, Σ given by the Kronecker structure in (14). The kernel of the log-

Subject	Group	Day						
		0	21	28	30	33	42	84
102	1	0.98	2×10^{-4}	0.001	0.003	0.05	0.06	-
L	1	0.98	2×10^{-4}	0.001	0.004	0.07	0.07	-
U	1	0.97	1×10^{-4}	8×10^{-4}	0.002	0.04	0.04	-
105	0	1.0	0.97	0.99	0.87	0.66	0.40	-
L	0	1.0	0.97	0.98	0.85	0.64	0.41	-
U	0	1.0	0.97	0.99	0.88	0.68	0.40	-
106	0	1.0	1.0	0.99	-	0.96	0.99	-
L	0	1.0	1.0	0.99	-	0.95	0.98	-
U	0	1.0	1.0	0.99	-	0.96	0.99	-
115	1	0.90	6×10^{-5}	3×10^{-4}	6×10^{-4}	-	0.021	0.24
L	1	0.90	1×10^{-4}	6×10^{-4}	0.001	-	0.025	0.26
U	1	0.90	3×10^{-5}	1×10^{-4}	3×10^{-4}	-	0.016	0.22
301	1	1.0	9×10^{-4}	3×10^{-6}	-	3×10^{-5}	5×10^{-4}	0.16
L	1	1.0	0.001	4×10^{-6}	-	4×10^{-5}	6×10^{-4}	0.16
U	1	1.0	6×10^{-4}	1×10^{-6}	-	2×10^{-5}	4×10^{-4}	0.16
302	2	0.71	3×10^{-7}	2×10^{-6}	-	7×10^{-5}	0.001	0.51
L	2	0.67	4×10^{-7}	4×10^{-6}	-	9×10^{-5}	0.001	0.48
U	2	0.75	2×10^{-7}	2×10^{-6}	-	5×10^{-5}	0.001	0.55
310	1	1.0	0.029	1×10^{-4}	-	0.46	0.85	0.94
L	1	1.0	0.031	0.002	-	0.47	0.84	0.92
U	1	1.0	0.027	0.001	-	0.46	0.86	0.95
402	0	1.0	1.0	1.0	-	-	1.0	1.0
L	0	1.0	1.0	1.0	-	-	1.0	1.0
U	0	1.0	1.0	1.0	-	-	1.0	1.0
405	1	1.0	2×10^{-3}	8×10^{-5}	-	6×10^{-4}	0.014	0.58
L	1	1.0	3×10^{-3}	1×10^{-4}	-	9×10^{-4}	0.017	0.60
U	1	1.0	0.001	4×10^{-5}	-	4×10^{-4}	0.012	0.55

Table 1: Validation sample of 9 males, probabilities on placebo for $\nu = 114, 98(L), 137(U)$ by row

integrated likelihood is given by (15). Focussing on one of these hyperparameters, say Σ, the log- integrated likelihood for given Γ is given by (16). This is of the form of the log of an inverted Wishart density (Brown, 1993, Appendix A). Thus had we assumed inverted Wishart prior distributions for these then the posterior distribution would also be inverted Wishart conditional on the other matrix of the pair. This would allow a fuller inference via Gibbs sampling.

We have presented a multivariate discrimination model for discrimination using longitudinal learning data where prediction is at a single time point. A number of general issues arise to be investigated in future work.

- Selection of subsets of responses
- Incorporation of cases with missing information

Preliminary work looking at subsets of markers using a scoring rule for prediction suggests that subsets do worse than the full eight markers. An algorithm for incorporation of simple structures of missing information in the case of simple (level I) maximum likelihood, that is when $\delta \to \infty$ in our formulation, is given in Brown *et al* (1999b).

ACKNOWLEDGEMENT

This work was supported by the International Olympic Committee and the European Union under the Biomed 2 Programme, Contract BMH4 CT950678. We are also grateful to Professor Peter Sönksen the coordinator and driving force of the multicentre project, the seven partners, and researchers in Denmark, Italy, Sweden and the UK, and the volunteers who provided the data.

A Matrix-Variate Distributions

We shall follow the notation introduced by Dawid (1981) for matrix-variate distributions. This has the advantage of preserving the matrix structures without the need to string by row or column as a vector. It redefines the degrees of freedom as shape parameters for both inverse Wishart and matrix-variate T, to allow notational invariance under marginalisation and very easy symbolic Bayesian manipulations.

With U a matrix having independent standard normal entries, $M + \mathcal{N}(\Gamma, \Sigma)$ will stand for a matrix-variate normal distribution of $V = M + A'UB$ where M, A, B are fixed matrices satisfying $A'A = \Gamma$ and $B'B = \Sigma$. Thus M is the matrix mean of V and $\gamma_{ii}\Sigma$ and $\sigma_{jj}\Gamma$ are the covariance matrices of the i-th row and j-th column, respectively, of V. There is thus a lack of uniqueness in Γ, Σ. In this Kronecker product structure they are unique up to an arbitrary scalar multiplier of Γ (which then divides Σ).

If U is of order $n \times p$ with $n \geq p$, the notation $\mathcal{IW}(\delta; \Sigma)$ with $\delta = n - p + 1$, will stand for the distribution of $B'(U'U)^{-1}B$, an inverse Wishart distribution. The shape parameter δ differs from the more conventional degrees of freedom, and may be generalised, using the density function, to take on any positive real value. The matrix-variate T distribution $M+\mathcal{T}(\delta;\Gamma,Q)$ is the distribution of T where T follows the $M+\mathcal{N}(\Gamma,\Sigma)$ distribution conditional on Σ, and $\Sigma\sim\mathcal{IW}(\delta;Q)$. Corresponding probability density functions are given in Brown (1993), Appendix A.

B Bayes Integration

The prior probability density function of $(\alpha_d, d=0,\ldots,g,B)$ given Ω is the product of

$$\pi(\alpha_d|\Omega)\propto h^{-q/2}|\Omega|^{-1/2}exp\{-\frac{1}{2h}(\alpha-\alpha_{d0})'\Omega^{-1}(\alpha-\alpha_{d0})\} \tag{28}$$

for $d=0,\ldots,g$ and

$$\pi(B|\Omega)\propto|H|^{-qp/2}|\Omega|^{-c/2}exp[-\frac{1}{2}\text{trace}\{H^{-1}(B-B_0)\Omega^{-1}(B-B_0)'\}]. \tag{29}$$

We first seek to integrate over $(\alpha_d, d=0,\ldots,g,B)$ for given Ω. In this Gaussian setting, to do this we should 'complete the square' in α_d and B within the exponentiated terms of likelihood times prior. First focusing on the likelihood for the $g+1$ sets of data comprising (5), the exponential term is

$$-\frac{1}{2}\text{trace}\Omega^{-1}\sum_{d=0}^{g}\{(Y_d-X_dB)'(Y_d-X_dB)-2(Y_d-X_dB)'1\alpha_d'+n_d\alpha_d\alpha_d'\}, \tag{30}$$

using $\text{trace}(AC)=\text{trace}(CA)$.

The α_d-term of (30) combines the prior (28) to give,

$$|\Omega|^{-1/2}exp-\frac{1}{2}\text{trace}\Omega^{-1}\sum\{-2(Y_d-X_dB)'1\alpha_d'+n_d\alpha_d\alpha_d'+(\alpha_d-\alpha_{d0})(\alpha_d-\alpha_{d0})'/h.\} \tag{31}$$

Letting $h\to\infty$ the prior is vague completing the square we see that given B,Ω,Y *a posteriori*

$$\alpha_d'-(\bar{Y}_d-\bar{X}_dB)\sim\mathcal{N}(1/n_d,\Omega) \tag{32}$$

independently for $d=0,\ldots,g$.

Turning to the integration of B for given Ω, the first term of (30) minus the quadratic in B left from completion of the square in α is

$$\sum\{(Y_d-X_dB)'(Y_d-X_dB)-n_d(\bar{Y}_d-\bar{X}_dB)'(\bar{Y}_d-\bar{X}_dB)\}.$$

For weak prior knowledge, letting H become large in (29), and completing the square in B, this gives *a posteriori* given Ω,Y

$$B-\hat{B}\sim\mathcal{N}([\sum X_d^{*\prime}X_d^*]^{-1},\Omega) \tag{33}$$

where $X_d^* = X_d - 1\bar{X}_d$ and

$$\hat{B} = [\sum X_d^{*\prime} X_d^*]^{-1}(\sum X_d^{*\prime} Y_d).$$

This completion of the square leaves the likelihood marginalised over the vague priors $\alpha_d, d = 0, \ldots, g, B$ for given Ω as proportional to

$$\{|\Omega|^{-N/2} exp\{-\frac{1}{2}\text{trace}(\Omega^{-1}S)\}. \tag{34}$$

with centered $Y_d^* = Y_d - 1\bar{Y}_d$ and sum of products of residuals

$$S = \sum_d (Y_d^* - X_d^*\hat{B})'(Y_d^* - X_d^*\hat{B}). \tag{35}$$

References

Brown, P. J. (1993). *Measurement, Regression,and Calibration.* Oxford: Clarendon Press.

Brown, P. J., Fearn, T. and Haque, M. S. (1999a). Discrimination with many variables. *Journal of the American Statistical Association*, **94,** 1320–1329.

Brown, P. J., Kenward, M. G. and Bassett, E. E. (1999b). Discrimination with longitudinal data. Technical Report UKC/IMS/99/09. University of Kent.

Brown, P. J., Le, N. D. and Zidek, J. V. (1994). Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics*, **22,** 489–509.

Chen, Chan-Fu (1979). Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. *Journal of the Royal Statistical Society,* B, **41,** 235–248.

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

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society,* B, **39,** 1–38.

Friedman, J. H. (1989). Regularised discriminant analysis. *Journal of the American Statistical Association*, **84,** 165–175.

Geisser, S. (1964). Posterior odds for multivariate normal classifications. *Journal of the Royal Statistical Society,* B, **26,** 69–76.

Good, I. J. (1965). *Estimation of probabilities: An essay on modern Bayesian methods.* MIT Press, Cambridge, MA.

Krzanowski, W. J. and Marriott, F. H. C. (1995). *Multivariate Analysis, Part 2: Classification, covariance structures and repeated measurements.* Kendall's Library of Statistics 2. Arnold, London.

Logan, T. P. and Gupta, A. K. (1993). Bayesian discrimination using multiple observations. *Communications in Statistics:- Theory and Methodology.*, **22,** 1735–1754.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate analysis.* Academic Press, London.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition.* Wiley, New York.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm. *Biometrika*, **80,** 267–278.

Smith, A. F. M. (1973). A general Bayesian linear model. *Journal of the Royal Statistical Society,* B, **35,** 67–75.