J. R. Statist. Soc. B (1982),
44, *No.* 3, *pp.* 287–321

Multivariate Calibration

By P. J. Brown

Imperial College, London

[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, March 3rd, 1982, Professor R. N. Curnow in the Chair]

Summary

A set of q responses $\mathbf{Y} = (Y_1, ..., Y_q)^{\mathrm{T}}$ are determined by a set of p explanatory variables $\mathbf{X} = (X_1, ..., X_p)^{\mathrm{T}}$. A set of l observed vectors $\mathbf{Y}$ are available at a single unknown $\mathbf{X}$ and it is desired to draw inferences about this unknown vector $\mathbf{X}$. In order to do this calibrating data is available jointly on $(\mathbf{Y}, \mathbf{X})$ where two situations are distinguished (i), $\mathbf{X}$ is controlled, (ii) $\mathbf{X}$ is random.

Using orthodox sampling theory on the one hand and Bayesian methods on the other hand point estimators and confidence regions are derived and contrasted. A procedure for selection of a subset of responses is given. Finally a comparison is made of the methods on data from (a) a random calibration experiment of wheat quality using an infrared spectrometer and (b) a controlled experiment of point finish.

Keywords: CALIBRATION; MULTIVARIATE REGRESSION; PREDICTION; CONTROLLED AND RANDOM EXPERIMENTATION; BAYES

1. Introduction

1.1. *Controlled and Random Calibration*

The well-known problem of calibration involves, in the simplest case, making inference about an unknown $p \times 1$ vector $\mathbf{X}'$ from a single random observed $q \times 1$ response vector $\mathbf{Y}'$. To this end, the relationship between $\mathbf{Y}$ and $\mathbf{X}$ is calibrated with experimental data $(\mathbf{Y}_i, \mathbf{X}_i)$, $i = 1, ..., n$, where $\mathbf{Y}_i, \mathbf{X}_i$ are $q \times 1$ and $p \times 1$ vectors respectively. This situation, the inverse of the more usual desire to predict $\mathbf{Y}'$ from $\mathbf{X}'$, is asymmetric in $\mathbf{X}$ and $\mathbf{Y}$ in that (i) $\mathbf{X}$ and $\mathbf{X}'$ are usually accurately determined and (ii) $\mathbf{X}$ in the calibration experiment may be at fixed prechosen levels, that is the calibration is *controlled*. When $\mathbf{X}$ as well as $\mathbf{Y}$ are random, calibration is said to be *random* and then only (i) is liable to distinguish the problem from that of predicting $\mathbf{Y}'$ from given $\mathbf{X}'$. With the notable exception of Williams (1959), the existing literature is concerned with $p = q = 1$, although Draper and Smith (1981, 2nd edn, p. 125) touch on $q = 1$, $p > 1$.

When there is no standard measurement $\mathbf{X}$, comparison is of two (or possibly more) instruments in a symmetric way. For a discussion of such comparative calibration see, for example, Williams (1969) and Theobald and Mallinson (1978). We will not explicity be concerned with such comparative problems.

An example of data from a controlled calibration experiment is given in Aitchison and Dunsmore (1975, p. 185). Enzyme concentration in blood plasma can be determined by a long and costly laboratory method whereas an autoanalyser method is quick and cheap. Here to calibrate the autoanalyser nine plasma samples selected to cover the range of enzyme concentrations have each been divided into four aliquots, one aliquot being assigned to the laboratory method and the other three to separate analyser determinations.

Ideally one would like the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ but of course this cannot be obtained from the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ without data from the marginal distribution of $\mathbf{X}$, or at least data from the marginal distribution of $\mathbf{Y}$ corresponding to random $\mathbf{X}$. No problems arise in the Bayesian approach since a prior for $\mathbf{X}'$ can always be given as demonstrated by Hoadley (1970) or Aitchison and Dunsmore (1975, Chapter 10). Otherwise

0035-9246/82/145287$2.00

the simple normal theory linear calibration model has been treated from a sampling theory approach, for example Brownlee (1960, Section 11.5), and via "marginal likelihood", Minder and Whitney (1975). More general regression functions of Y on X were considered by Clark (1979), concerning exponential decay and carbon-14 dating, and by Scheffé (1973), allowing for example polynomials in X. Asymptotic confidence intervals for this are given by Lundberg and De Maré (1980). Here monotonicity over a relevant range of X ensures uniqueness of estimated X' for given Y'.

Monotonicity, it should be emphasized, should predicate any effective univariate calibration. However, we shall see in Section 5, that in multivariate calibration simple notions such as pairwise monotonicity are not essential.

If, in the calibration experiment on the other hand, $\mathbf{Y}$ and $\mathbf{X}$ are random then no difficulties arise in specification of the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$. Hence if $(\mathbf{X}', \mathbf{Y}')$ may be assumed to derive from the same joint distribution inference about $\mathbf{X}'$ for given $\mathbf{Y}'$ is immediate. Often however $\mathbf{X}$ will correspond to random true values accurately determined. Whilst the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ may be reasonably assumed to be normal, involving measurement errors superimposed by the cheap and quick measuring instrument, the distribution of $\mathbf{X}$, and hence that of $\mathbf{X}$ given $\mathbf{Y}$, may not be normal. This in particular will deserve careful checking from the data, perhaps along the lines of Healy (1968) or Cox and Small (1978). It may thus be advantageous to derive the distribution of $\mathbf{X}$ given $\mathbf{Y}$ separately from $\mathbf{Y}$ given $\mathbf{X}$ and the marginal distribution of $\mathbf{X}$. Lwin and Maritz (1980) follow this course, basing estimation of the marginal distribution of $\mathbf{X}$ on the sample distribution function. They analyse random calibration data on water content of soil using as an accurate laboratory method (X) and a quick on-site method (Y).

Briefly, for point estimation, if $f(y_i\, x_i, \beta)$ is the probability density function of Y_i conditional on $X_i = x_i, i = 1, \ldots, n$, then the predictor of X' when $Y' = y'$ is observed is

$$\sum w_i x_i$$

where

$$w_i = f(y' \mid x_i, \beta) / \sum f(y' \mid x_j, \beta).$$

This is a weighted average of the x_i in the calibration experiment with weights proportional to their probabilistic distance from y'. When as is usual, the adjustable parameter β is unknown, it may be replaced by a good estimator, perhaps the maximum likelihood estimator. No attempt however is made to allow for the increased uncertainty due to estimation of β. A multivariate extension of their method is described in Section 4.2 and applied to the data example of that Section.

Aside from such considerations, for random calibration nothing new is incurred, from going to p, q greater than unity, over and above standard multivariate regression theory. Shrinkage methods may be advantageously used as for example in Brown and Zidek (1980). The subsequent emphasis will therefore be on controlled calibration. An interesting intermediate possibility not considered is where some X variables are controlled and some random as in Brooks (1974).

Discrimination, or medical diagnosis described for example in Titterington *et al.* (1981), or pattern recognition as it is referred to in the electrical engineering literature (Kanal, 1974 provides a good review), shares similar features with the calibration problem. It differs in that typically X is univariate and discrete taking a set of nominal values which identify from which of several populations the $\mathbf{Y}$ derives. In the case of discrimination, Geisser (1964) clearly distinguishes between the random and controlled experiment, whereas the distinction is not often made within the engineering literature where the emphasis is on nonparametric procedures as opposed to probabilistic manipulation. In the context of random X and medical diagnosis Dawid (1976) and both P. J. Brown and D. R. Cox in the discussion of Titterington *et al.* (1981) emphasize the advantages of modelling X given $\mathbf{Y}$ over that of $\mathbf{Y}$ given X. This may be countered by the fact that in many parametric models estimation from X given $\mathbf{Y}$ may be improved by modelling via $\mathbf{Y}$ given X and marginal X, if $\mathbf{Y}$ is not S-ancillary for the parameters of X given $\mathbf{Y}$, as for example in Efron (1975).

The importance of the distinction between controlled and random calibration should be emphasized. It has often been ignored in the past. Of course the use of the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ in random calibration need not be confined to jointly normal $\mathbf{X}$ and $\mathbf{Y}$, as in much of our subsequent development. For example, $\mathbf{X}$ and $\mathbf{Y}$ might be considered independent Poisson entries of a $p \times p$ contingency table. Then conditional on $\mathbf{Y}$ the rows are multinomial and conditional on $\mathbf{X}$ the columns are also multinomial but with different probabilities. Fortunately also each margin is S-ancillary for these conditional probabilities.

Specifically a random batch of apples might be sorted mechanically into p sizes, $\mathbf{Y}$, whilst a careful true sorting gives random p-dimensional vector $\mathbf{X}$. Grassia and Sundberg (1982) examine this, working essentially with the multinomial distributions of $\mathbf{Y}$ given $\mathbf{X}$. However to predict the true size category X from similar random batches, we would simply advocate the use of the multinomial distributions of $\mathbf{X}$ given $\mathbf{Y}$. When $\mathbf{X}$ is controlled, as for example in fish stock calibration (Pella and Robertson, 1979), the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ is fundamental, although one message of this paper is that under some strategies for controlling $\mathbf{X}$, it is better to behave as if $\mathbf{X}$ were random and formally derive the distribution of $\mathbf{X}$ given $\mathbf{Y}$.

1.2. *Univariate Controlled Calibration Reviewed*

Adherents of various approaches to controlled calibration have generally concentrated on comparisons between point estimates of X' for given Y'. Whereas Williams (1969) and others insist on using the fitted regression of Y on X and hence deriving the estimate $\hat{X}'$ of X' from

$$Y' - \bar{y} = (S_{xy}/S_{xx})(\hat{X}' - \bar{x}),$$

Krutchkoff (1967) suggested regressing X on Y and obtaining $\hat{\hat{X}}'$ from

$$\hat{\hat{X}}' - \bar{x} = (S_{xy}/S_{yy})(Y' - \bar{y}),$$

where $\bar{x} = \Sigma x_i/n$, $\bar{y} = \Sigma y_i/n$, $S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y})$, $S_{yy} = \Sigma(y_i - \bar{y})^2$, $S_{xx} = \Sigma(x_i - \bar{x})^2$. The two estimates coincide only when X relates perfectly to Y in linear fashion. The suggestion of Krutchkoff (1967) runs counter to established protocols stemming at least from Eisenhart (1939) since the n X-values are fixed. However from a Bayesian point of view Hoadley (1970) showed that this latter approach would be justified if the X-values were chosen in a manner to reflect the prior beliefs about the future X'. Broadly, regressing Y on X corresponds to diffuse prior information about X'. In sampling theory terms Hoadley's result implies that the procedure of regressing X on Y will do well if the unknown X' happens to be in a part of X space reasonably central to the set of X-values prechosen in the controlled calibration experiment. The approach would not fare so well if X' were outside the prechosen range. The point at which Y on X does better than X on Y is not clear from the selection of simulations in the literature. At any rate for a proper comparison of point estimates a bounded loss function would be necessary since if the estimated slope of Y on X happens by chance to be near zero, $\hat{X}'$ becomes very large. The expected mean square error of $\hat{X}'$ is indeed infinite and mean square error as a criterion has been criticized by Williams (1969).

There are at least four different ways to obtain a confidence region for X': fiducial as described by Fieller (1954), essentially a *joint sampling* approach; *conditional sampling*, as in Wilks' tolerance regions and as developed by Scheffé (1973); *marginal likelihood* (Minder and Whitney, 1974); *Bayes* as in Hoadley (1970).

Perhaps the simplest approach is that of joint sampling . It is easy to see that given α, β, σ^2, X, X' the joint sampling distribution of Y', $\hat{\alpha} = \bar{Y}$, $\hat{\beta} = S_{xy}/S_{xx}$ is such that

$$(Y' - \hat{\alpha} - \hat{\beta}X')/[\sigma^2\{1 + 1/n + (X' - \bar{x})^2/S_{xx}\}]^{\frac{1}{2}} \tag{1.1}$$

is standard normal. Note that this standard normal does not involve any of the conditioning parameters α, β, σ^2 or X, X' so that probability statements are also true unconditionally and in particular over repetitions of (Y, X) where both Y *and* X are allowed to vary. Replacing σ^2 by $\hat{\sigma}^2$, the usual unbiased estimator derived from the residual sum of squares after regressing Y on X, leads to a $100(1-\gamma)$ per cent confidence region as those values of X' satisfying the quadratic inequality

$$(Y'-\hat{\alpha}-\hat{\beta}X')^2 \leq \hat{\sigma}^2 t_{n-2}^2(\gamma)\{1+1/n+(X'-\bar{x})^2/S_{xx}\}, \tag{1.2}$$

where $t_{n-2}(\gamma)$ is the two-sided 100γ per cent point of the student t-distribution on $(n-2)$ degrees of freedom. This region is a respectable interval provided the t-test of the hypothesis $\beta = 0$ is rejected. Otherwise it is either the whole real line or even two disjoint semi-infinite lines! This has been the source of some consternation, see for example Neyman's discussion of Fieller's paper or Hoadley (1970). The practical man's answer that one should not attempt calibration when one is not confident that $\beta \neq 0$ might be countered by the argument that if the procedure is obviously suspect in some circumstances then the solutions may be far from ideal in the other cases where there is no obvious flaw.

The conditional sampling approach requires the specification of two probabilities or significance levels and produces operationally dense statements like "I am 80 per cent certain that 95 per cent of the statements I make are true". In univariate (polynomial) calibration it has been extensively developed by Scheffé (1973). He gives an excellent discussion of necessary assumptions and provides tables for the implementation of his method. This approach via tolerance regions is strongly criticized by Lindley (1972, p. 56).

Both a marginal likelihood approach and a Bayesian approach to the derivation of confidence regions are described in the two papers already cited. Minder and Whitney use various approximations to their "marginal likelihood" to derive confidence regions. Hoadley emphasised the need for a proper prior distribution for X' in order that the posterior distribution be integrable. With one specially chosen but not unnatural prior, posterior intervals for X' can be obtained simply from regression of X' on Y.

1.3. *Examples of Multivariate Calibration*

Some examples of multivariate calibration are given in Williams (1959, Chapter 9). We were motivated to the present paper by two different examples. The first involved a random calibration experiment and related $p = 2$ accurately determined measurements of moisture and protein content of wheat samples to six infrared reflectance measurements at six different wavelengths. Four responses ($q = 4$), derived as differences of a subset of these reflectance measurements, (1)–(2), (4)–(3), (4)–(5) and (1)–(5), are tabulated with the accurate laboratory determinations in Table 1.

The second example concerned a controlled calibration experiment where $p = 2$ factors, pigmentation and viscosity of paint, were controlled each at three levels in a three-by-three completely balanced experiment. Again $q = 6$ responses involving optical properties and measuring appearance were obtained. The aim in future was to be able to match the paint by taking optical measurements. The complete data for this is given in Table 2.

These examples are analysed in Sections 4 and 5, the analysis being no more than comparative and illustrative of techniques developed in Sections 2 and 3.

1.4. *A Formulation of the Multivariate Controlled Calibration Problem*

The controlled calibration model assumed is

$$\mathbf{Y}_i = \mathbf{m}(\mathbf{X}_i, \Theta) + \mathbf{e}_i, \quad i = 1, \ldots, n, \tag{1.3}$$

where $\mathbf{Y}_i$ is a $q \times 1$ vector of responses, $\mathbf{X}_i$ a $p \times 1$ vector of fixed explanatory variables, $\mathbf{m}(\mathbf{X}_i, \Theta)$

is a $q \times 1$ vector function of $\mathbf{X}_i$ of known form and $\boldsymbol{\Theta}$ a matrix of unknown parameters, $\mathbf{e}_i$ is a $q \times 1$ error typically satisfying

$$E(\mathbf{e}_i) = \mathbf{0}, \quad E(\mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}) = \boldsymbol{\Gamma}, \tag{1.4}$$

so that with normality additionally assumed, given $\boldsymbol{\Gamma}$

$$\mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Gamma}), \tag{1.5}$$

where "$\sim$" denotes "distributed as". Errors are assumed independent from observation to observation. Other error assumptions may be desirable and may be incorporated at the expense of a more complicated analysis.

The observations of the prediction experiment are assumed to follow the same assumptions as those of the calibration experiment. In particular

$$\mathbf{Y}_j' = \mathbf{m}(\mathbf{X}', \boldsymbol{\Theta}) + \mathbf{e}_j', \quad j = 1, \ldots, l, \tag{1.6}$$

where $\mathbf{e}_j'$ satisfy (1.5) and are independent of $\mathbf{e}_i$, $i = 1, \ldots, n$ for $j = 1, \ldots, l$. Each observation Y_j' is observed at the same unknown $\mathbf{X}'$, $j = 1, \ldots, l$.

The full problem defined by (1.3), (1.5) and (1.6) has three distinct sets of unknowns or parameters, (i) $\boldsymbol{\Theta}$ determining the regression of $\mathbf{Y}$ on $\mathbf{X}$, (ii) $\boldsymbol{\Gamma}$ determining the distributions of deviations about the regression relation and (iii) $\mathbf{X}'$ which in conjunction with $\boldsymbol{\Theta}$ determines the mean of $\mathbf{Y}_j'$, $j = 1, \ldots, l$. For ease of analysis $\mathbf{m}(\mathbf{X}, \boldsymbol{\Theta})$ in (1.3), (1.6) may be categorized by whether it is or is not linear in $\mathbf{X}$ and $\boldsymbol{\Theta}$. Being linear in both corresponds to standard multivariate regression. When $\mathbf{m}(\mathbf{X}, \boldsymbol{\Theta})$ is only linear in $\boldsymbol{\Theta}$ as is often the case when derived variables are incorporated, for example polynomials in $\mathbf{X}$, then (1.3) is still easy to analyse but (1.6) is non-linear in the unknown $\mathbf{X}'$. Non-linearity in $\boldsymbol{\Theta}$ poses problems for both (1.3) and (1.6). Note the logical structure of (1.3) and (1.6). The calibration experiment (1.3) provides information on $\boldsymbol{\Theta}$. Even if n is infinite and $\boldsymbol{\Theta}$ is then known, estimators of $\mathbf{X}'$ from (1.6) will depend for their accuracy on the structure of $\mathbf{m}(\mathbf{X}', \boldsymbol{\Theta})$, $\boldsymbol{\Gamma}$ and the number of replicates l. Thus for example even when $\boldsymbol{\Theta}$ is known, an estimator of $\mathbf{X}'$ can only be consistent as $l \to \infty$. Finally it may be observed that if q is less than the number of independent variables, p, if there are no derived variables, $\mathbf{X}'$ cannot be completely determined. We will therefore avoid such ill-specified problems.

2. SAMPLING THEORY RESULTS

2.1. *Controlled Calibration Linear in* $\boldsymbol{\Theta}$

The multivariate linear regression model in which $\mathbf{m}(\mathbf{X}, \boldsymbol{\Theta})$ is linear in $\boldsymbol{\Theta}$ is assumed. Accordingly we specialize (1.3) and (1.6) to

$$\mathbf{Y} = \mathbf{1}\boldsymbol{\alpha}^{\mathrm{T}} + \mathbf{X}\mathbf{B} + \mathbf{E} \tag{2.1}$$

$$\mathbf{Y}' = \mathbf{1}\boldsymbol{\alpha}^{\mathrm{T}} + \mathbf{1}\boldsymbol{\xi}^{\mathrm{T}}\mathbf{B} + \mathbf{E}' \tag{2.2}$$

where $\mathbf{Y}(n \times q)$, $\mathbf{E}(n \times q)$, $\mathbf{Y}'(l \times q)$, $\mathbf{E}'(l \times q)$ are random matrices, $\mathbf{X}(n \times p)$ is a matrix of fixed constants as are the vectors of units, $\mathbf{1}$, which are respectively $n \times 1$ and $l \times 1$; $(\boldsymbol{\alpha}, \mathbf{B})$ replaces $\boldsymbol{\Theta}$. Here $\boldsymbol{\xi}$ is the $p \times 1$ vector of unknowns previously denoted as $\mathbf{X}'$, a Greek letter perhaps being preferable. In this formulation $\mathbf{X}$ might consist of p variables derived from a smaller set as in polynomial regression. However in this case $\boldsymbol{\xi}$ is a vector function of the same reduced number of unknowns. The use of the letter $\boldsymbol{\xi}$ rather than $\mathbf{X}'$ helps to further emphasize this possibility. The case of no derived variables will be referred to as *standard* multivariate linear regression. We reserve $\mathbf{X}'$ instead of $\boldsymbol{\xi}$ for this case.

Since explanatory variables are fixed we may without loss of generality assume

$$\sum_i x_{ij} = 0, \quad \sum_i x_{ij}^{\,2}/n = 1, \quad j = 1, \ldots, s, \tag{2.3}$$

that is columns of $\mathbf{X}$ are centred and have average sum of squares one. Using a canonical form of the model (2.1), (2.2) it is straightforward, as shown in the Appendix, to obtain the multivariate analogue of (1.1)

$$(\bar{\mathbf{Y}}'-\hat{\boldsymbol{\alpha}}-\hat{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\xi}) \sim N(\mathbf{0},\boldsymbol{\Gamma}\sigma^2(\boldsymbol{\xi})), \tag{2.4}$$

where

$$\sigma^2(\boldsymbol{\xi}) = 1/l+1/n+\boldsymbol{\xi}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\boldsymbol{\xi}. \tag{2.5}$$

Let $\mathbf{S}$ $(q\times q)$ be the residual sum of products matrix, pooled from the calibration and prediction experiments when $l>1$. Then $\mathbf{S}$ has a Wishart distribution with scale matrix $\boldsymbol{\Gamma}$ and degrees of freedom $v+q-1$ (see Appendix) where

$$v = n-p+l-q-1. \tag{2.6}$$

Hence it is shown in the Appendix that, with $\mathbf{S}^{\frac{1}{2}}$ a $q\times q$ matrix square root of $\mathbf{S}$,

$$\mathbf{S}^{-\frac{1}{2}}(\bar{\mathbf{Y}}'-\hat{\boldsymbol{\alpha}}-\hat{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\xi})/\sigma(\boldsymbol{\xi}) \sim T(v;\mathbf{I}_q), \tag{2.7}$$

where $T(v;\boldsymbol{\Sigma})$ denotes a multivariate student distribution with mean vector zero and scale matrix $\boldsymbol{\Sigma}$ such that $v^{\frac{1}{2}}T(v;\boldsymbol{\Sigma})$ is the standard multivariate student distribution defined for example in Press (1972, p. 125). From, for example, Dawid (1981), it follows that a $100(1-\gamma)$ per cent confidence region for $\boldsymbol{\xi}$ is all $\boldsymbol{\xi}$ such that

$$(\bar{\mathbf{Y}}'-\hat{\boldsymbol{\alpha}}-\hat{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\xi})^{\mathrm{T}}\mathbf{S}^{-1}(\bar{\mathbf{Y}}'-\hat{\boldsymbol{\alpha}}-\hat{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\xi})/\sigma^2(\boldsymbol{\xi})\leq(q/v)F_v^q(\gamma) \tag{2.8}$$

where $F_v^q(\gamma)$ is the upper $100(1-\gamma)$ per cent point of the standard F-distribution on q and v degrees of freedom. This reduces to (1.2) when $p = q = 1$. In standard multivariate linear regression (2.8) corresponds to the fiducial limits of Williams (1959, p. 169). In polynomial regression (2.8) may result in rather complicated regions in the reduced variable space. This is illustrated by the example of Section 5. However, standard multivariate linear regression produces natural elliptical regions under a condition which is a direct extension of that of simple univariate calibration. The results are given by Theorem 1 of the next Section.

2.2. *The Form of Confidence Region in Standard Multivariate Multiple Regression*

Here we denote $\boldsymbol{\xi} = \mathbf{X}'$ to emphasize that the standard multivariate linear regression model is adopted. Inequality (2.8) may be written as

$$\mathbf{X}'^{\mathrm{T}}(\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}-k(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})\mathbf{X}'-2(\bar{\mathbf{Y}}'-\hat{\boldsymbol{\alpha}})^{\mathrm{T}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}'+(\bar{\mathbf{Y}}'-\hat{\boldsymbol{\alpha}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{Y}'-\hat{\boldsymbol{\alpha}})-k(l^{-1}+n^{-1})\leq 0, \tag{2.9}$$

where

$$k = (q/v)F_v^q(\gamma) \tag{2.10}$$

This is a quadratic form in $\mathbf{X}'$. Let

$$\mathbf{C} = \hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}-k(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \tag{2.11}$$

This is symmetric. Suppose the following condition holds.

Condition 1

The matrix $\mathbf{C}$ defined by (2.11) is positive definite.

We have the following

Theorem 1

(i) When $p = q$, Condition 1 is sufficient to guarantee that the roots of the quadratic form (2.9) are all real and the region is a closed ellipsoid.

(ii) When $q>p$ the roots need not be real even when Condition 1 holds.

(iii) Condition 1 corresponds to a test of the null hypothesis $\mathbf{X}'^{\mathrm{T}}\mathbf{B} = 0$ for any $\mathbf{X}'$. Thus Condition 1 will typically be satisfied provided $\mathbf{X}'^{\mathrm{T}}\mathbf{B} > 0$ and the calibration experiment is of a sufficient size relative to random variation.

Proof of (i) and (ii). Represent (2.9) as

$$\mathbf{X}'^{\mathrm{T}}\mathbf{C}\mathbf{X}' - \mathbf{D}^{\mathrm{T}}\mathbf{X}' - \mathbf{X}'^{\mathrm{T}}\mathbf{D} + E - F \leq 0,$$

where

$$\mathbf{D} = \hat{\mathbf{B}}\mathbf{S}^{-1}(\bar{\mathbf{Y}}' - \hat{\boldsymbol{\alpha}}), \quad E = (\bar{\mathbf{Y}}' - \hat{\boldsymbol{\alpha}})^{\mathrm{T}}\mathbf{S}^{-1}(\bar{\mathbf{Y}}' - \hat{\boldsymbol{\alpha}}), \quad F = k(l^{-1} + n^{-1}).$$

Completing the square this becomes

$$\| \mathbf{X}' - \mathbf{C}^{-1}\mathbf{D} \|_{C}^{2} - \mathbf{D}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{D} + E - F \leq 0, \tag{2.12}$$

where $\| \mathbf{z} \|_{C}^{2} = \mathbf{z}^{\mathrm{T}}\mathbf{C}\mathbf{z}$. Now

$$E - \mathbf{D}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{D} = (\bar{\mathbf{Y}}' - \hat{\boldsymbol{\alpha}})^{\mathrm{T}}(\mathbf{S}^{-1} - \mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{C}^{-1}\hat{\mathbf{B}}\mathbf{S}^{-1})(\bar{\mathbf{Y}}' - \hat{\boldsymbol{\alpha}}),$$

and the matrix of this quadratic form is, on substitution,

$$\mathbf{S}^{-1} - \mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}(\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}} - k(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})^{-1}\hat{\mathbf{B}}\mathbf{S}^{-1}.$$

Using the binomial inverse theorem (Press, 1972, p. 23) this may be rewritten

$$(\mathbf{S} - \hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(k^{-1})\hat{\mathbf{B}})^{-1} \tag{2.13}$$

Let $\mathbf{W} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\frac{1}{2}}\hat{\mathbf{B}}\mathbf{S}^{-\frac{1}{2}}$ then $(\mathbf{W}\mathbf{W}^{\mathrm{T}} - k\mathbf{I}_p) > \mathbf{0}$ from Condition 1 so that the eigenvalues $w_1^2, \ldots, w_p^2$ are all $> k$ and then the eigenvalues of $\mathbf{W}^{\mathrm{T}}\mathbf{W}$ are $w_1^2, \ldots, w_p^2$ supplemented by $(q-p)$ zero eigenvalues. Thus (2.13) is negative definite as long as $q = p$ but, for $q > p$, $(q-p)$ eigenvalues will be positive. Thus schematically, when $q = p$, (2.9) is

$$\| \mathbf{X}' - \mathbf{C}^{-1}\mathbf{D} \|_{C}^{2} - \text{constant} \leq 0,$$

where $\mathbf{C} > \mathbf{0}$ by Condition 1. Condition 1, $\mathbf{C} > \mathbf{0}$, ensures that the quadratic form is strictly convex. The subtraction of a constant above guarantees that it cuts the $\mathbf{X}'$ axes. When $q > p$ this constant may be negative. This will occur when the q elements of $\bar{\mathbf{Y}}'$ are sufficiently contradictory in their information about the p elements of $\mathbf{X}'$ to overwhelm the constant F and the negative contributions from the quadratic form above.

(iii) Proof of this follows from standard methodology as given for example by Anderson (1958, Section 8.3). We will derive the results from the canonical form of the Appendix, namely

$$\mathbf{Z}_{1i}^{\mathrm{T}} = (n\lambda_i)^{\frac{1}{2}}\boldsymbol{\alpha}_i^{\mathrm{T}} + \mathbf{e}_i^{\mathrm{T}} \quad i = 1, \ldots, p, \tag{2.14}$$

with $\mathbf{A}^{\mathrm{T}} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p)$ and $\mathbf{A} = \mathbf{P}^{\mathrm{T}}\mathbf{B}$; the condition $\mathbf{X}'^{\mathrm{T}}\mathbf{B} = \mathbf{0}$ corresponds to $\mathbf{w}^{\mathrm{T}}\mathbf{A} = \mathbf{0}$ where $\mathbf{w} = \mathbf{p}^{\mathrm{T}}\mathbf{X}'$ is a $(p \times 1)$ vector. That is

$$w_1\boldsymbol{\alpha}_1 + w_2\boldsymbol{\alpha}_2 + \ldots + w_p\boldsymbol{\alpha}_p = 0.$$

Now choose $l_i = w_i(n\lambda_i)^{-\frac{1}{2}}/a$ for $i = 1, \ldots, p$ where $a^2 = \Sigma w_i^2/(n\lambda_i)$ so that $\mathbf{l}^{\mathrm{T}}\mathbf{l} = 1$. Next, orthogonally transform the p vectors in (2.14) by means of a $p \times p$ matrix $\mathbf{L}$ such that the first column of $\mathbf{L}$ is $\mathbf{l}$.

Multiplying (A.1.1), the stacked version of (2.14), by $\mathbf{L}^{\mathrm{T}}$, the error properties are unchanged. If $\mathbf{U}^{\mathrm{T}} = \mathbf{Z}_1^{\mathrm{T}}\mathbf{L} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)$ the null hypothesis is that the mean of $\mathbf{u}_1$ is zero and the sum of products due to this is

$$\mathbf{u}_1\mathbf{u}_1^{\mathrm{T}} = (\mathbf{l}^{\mathrm{T}}\mathbf{Z}_1)^{\mathrm{T}}(\mathbf{l}^{\mathrm{T}}\mathbf{Z}_1) = \hat{\mathbf{A}}^{\mathrm{T}}\mathbf{w}\mathbf{w}^{\mathrm{T}}\hat{\mathbf{A}}/a^2$$

or

$$\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}'\mathbf{X}'^{\mathrm{T}}\hat{\mathbf{B}}/(\mathbf{X}'^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}').$$

The residual sum of products is $\mathbf{S}$ given in Section 2.1 and hence the likelihood ratio test statistic is a function of Wilks' criterion, the ratio

$$|\mathbf{S}|/|\mathbf{S}+\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}'\mathbf{X}^{\mathrm{T}}\hat{\mathbf{B}}/a^2|,$$

which has the distribution $U_{1,q,v}$ of Anderson (1958, Section 8.4). Manipulating the ratio of determinants

$$\begin{aligned} U_{1,q,v} &= |\mathbf{I}_q+\mathbf{S}^{-\frac{1}{2}}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}'\mathbf{X}'^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{S}^{-\frac{1}{2}}/a^2|^{-1} \\ &= (1+\mathbf{X}'^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}'/a^2)^{-1} \end{aligned}$$

by Sylvester's theorem (Press, 1972, p. 20). Hence from Anderson (1958, Section 8.5.3) the statistic for testing the null hypothesis $\mathbf{X}'^{\mathrm{T}}\mathbf{B} = 0$ is

$$\mathbf{X}'^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}\mathbf{X}'/\mathbf{X}'^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}' \tag{2.15}$$

and is distributed as $(q/v)F^q_v$. The result (iii) follows.

Some remarks concerning Theorem 1 may be in order. Williams (1959) p. 169 works directly from (2.8) and does not derive our Condition 1 for real roots when $p = q$. When $q > p$ the effect on the confidence interval of the information in $\bar{\mathbf{Y}}'$ is not examined; rather he examines the consistency of the q elements of $\bar{\mathbf{Y}}'$ with the calibration experiment.

When condition 1 fails to hold confidence regions will be non-convex and possibly infinite as in the case $q = p = 1$ described for example by Hoadley (1970).

Letting $\gamma \to 1$, then $k \to 0$ and Condition 1 is sure to be satisfied if $\mathbf{B} \neq \mathbf{0}$. The resulting confidence region degenerates to the point

$$\hat{\mathbf{X}}' = (\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}})^{-1}\hat{\mathbf{B}}\mathbf{S}^{-1}(\bar{\mathbf{Y}}'-\hat{\boldsymbol{\alpha}}). \tag{2.16}$$

This is the natural estimator of $\mathbf{X}'$ which would result from maximum likelihood (or weighted least squares) estimation of $\mathbf{X}'$ in the prediction experiment when $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma})$ are replaced by $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \mathbf{S})$ from the calibration experiment. Note how $\mathbf{C}^{-1}\mathbf{D}$, the central point of the confidence region (2.9) is dependent on the confidence coefficient through k. It may be noted that $\mathbf{C}^{-1}\mathbf{D}$ might be loosely viewed in the form of an "expansion" estimator comparable to $\hat{\mathbf{X}}'$ through the term $-k(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

In the general case of functionally related components of ξ, it will be necessary to minimize numerically the left-hand side of (2.8) to estimate $\mathbf{X}'$. An example of this is given in Section 5 in the discussion of the paint data.

2.3. *Testing the Redundancy of a Subset of Responses*

In the standard multivariate regression case, the larger the eigenvalues of matrix $\mathbf{C}$ given by (2.11) the steeper the sides of the parabolic bowl given by (2.9) and hence the narrower the confidence region for $\mathbf{X}'$. If the q variables are reduced to $q_1 < q$ with $q_1 \geqslant p$, a new calculation of (2.12) gives

$$\mathbf{C}_1 = \hat{\mathbf{B}}_1\mathbf{S}_1^{-1}\hat{\mathbf{B}}_1^{\mathrm{T}} - k_1(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1},$$

where v_1 and then k_1 are recalculated from (2.6) and (2.10), respectively, on replacing q by q_1; also $\hat{\mathbf{B}}_1$ is the appropriate q_1 column-subset of $\hat{\mathbf{B}}$ and $\mathbf{S}_1$ the corresponding $(q_1 \times q_1)$ submatrix of $\mathbf{S}$. Intuitively, if the eigenvalues of $\mathbf{C}_1$ are not much less than those of $\mathbf{C}$, the $(q-q_1)$ responses deleted are redundant in specifying $\mathbf{X}'$ given the q_1 responses. The notion of "not much less" may be given more formal weight by adopting a test of additional information as given by Rao (1965, Section 8c.4). The null hypotheses of Theorem I(iii) is $\mathbf{X}'^{\mathrm{T}}\mathbf{B} = 0$, hence the hypothesis matrix is the rank one vector $\mathbf{X}'$ and Rao's distributional simplification in the rank one case may be applied.

The parallel is best seen in terms of the proof of Theorem 1(iii) given in the previous section.

Formula (2.15) is Rao's T^2/k and hence from his (8c.4.10)

$$\mathbf{X}'^{\mathrm{T}}(\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}-\hat{\mathbf{B}}_1\mathbf{S}_1^{-1}\hat{\mathbf{B}}_1^{\mathrm{T}})\mathbf{X}'/\{\mathbf{X}'^{\mathrm{T}}([\mathbf{X}^{\mathrm{T}}\mathbf{X}]^{-1}+\hat{\mathbf{B}}_1\mathbf{S}_1^{-1}\hat{\mathbf{B}}_1^{\mathrm{T}})\mathbf{X}'\}$$

has a $\{(q-q_1)/v\}\,F_v^{q-q_1}$ distribution. Thus if,

$$\mathbf{X}'^{\mathrm{T}}[\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}-k_0(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}-(1+k_0)\hat{\mathbf{B}}_1\mathbf{S}_1^{-1}\hat{\mathbf{B}}_1^{\mathrm{T}}]\mathbf{X}'>0 \tag{2.17}$$

where $k_0=\{(q-q_1)/v\}\,F_v^{q-q_1}(\gamma)$, the null hypothesis of no additional information may be rejected at the 100γ per cent level. If the matrix in the square brackets of (2.17) is positive definite the null hypothesis will be rejected for all $\mathbf{X}'$. An alternative way of utilising (2.17) would pick an *a priori* likely $\mathbf{X}'$ or even a set of likely $\mathbf{X}'$ values. One might even check with all $\mathbf{X}'$ values of the calibration experiment. A particularly appealing variation on this would be to use jack-knifed estimates of the parameters and demand that (2.17) be satisfied for all n one-at-a-time omitted values of $\mathbf{X}$. Such an approach though would be computationally expensive and anyway implicitly assumes that all $\mathbf{X}$-values of the calibration experiment are exchangeable with the unknown $\mathbf{X}'$.

In Section 5, (2.17) is used to choose a subset of the optical paint responses.

Finally it may be noted that the derivation of (2.17) allows some components of $\boldsymbol{\xi}$ to be functionally related. With $\boldsymbol{\xi}$ replacing $\mathbf{X}'$, then it is not necessary that the matrix be positive-definite to satisfy (2.17) for all $\boldsymbol{\xi}$. Only a subset of values of $\boldsymbol{\xi}$ in p-dimensional space is feasible. It is hard to know how to use this operationally. An alternative Bayesian method for polynomial and non-linear $\boldsymbol{\xi}$ is suggested in Section 3.1.

2.4. *Comparative Formulae for Regressing* X *on* Y *and* Y *on* X *in Standard Multivariate Regression*

It would be computationally simpler if regression of $\mathbf{X}$ on $\mathbf{Y}$ was performed when $\mathbf{Y}$ on $\mathbf{X}$ is linear. This as already mentioned would be the correct practice in a random calibration experiment where $(\mathbf{X},\mathbf{Y})$ are jointly multivariate normal. Armed with a formula estimating $E(\mathbf{X}\mid\mathbf{Y})$ all that is required to predict $\mathbf{X}'$ is to substitute $\bar{\mathbf{Y}}'$ for $\mathbf{Y}$. In this section the formulae are compared. Units of both $\mathbf{X}$ and $\mathbf{Y}$ are adopted so that *the variables are post hoc centred on zero.* In this case the estimated regression of $\mathbf{X}$ on $\mathbf{Y}$ predicts

$$\hat{\mathbf{X}}'^{\mathrm{T}}=\bar{\mathbf{Y}}'^{\mathrm{T}}\mathbf{D}, \tag{2.18}$$

where

$$\mathbf{D}=(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X}. \tag{2.19}$$

Lemma 2. When $\mathbf{X}$ and $\mathbf{Y}$ have been centred, $\mathbf{D}$ may be rewritten as

$$\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}+\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}]^{-1} \tag{2.20}$$

where $\hat{\mathbf{B}}$ and $\mathbf{S}$ are the usual quantities obtained from regression $\mathbf{Y}$ on $\mathbf{X}$.

Proof. Using the binomial inverse theorem (Press, 1972, p. 23), (2.20) may be rewritten

$$\begin{aligned}&\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{X}-\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{B}}(\mathbf{S}+\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}]\\&\quad=[\mathbf{S}^{-1}-\mathbf{S}^{-1}(\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{B}})(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}]\mathbf{Y}^{\mathrm{T}}\mathbf{X}\end{aligned}$$

since

$$\begin{aligned}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}&=\mathbf{S}+\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{B}}\\&=(\mathbf{S}^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}-\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{B}})(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\\&=\mathbf{S}^{-1}\mathbf{S}\mathbf{D}=\mathbf{D}.\end{aligned}$$

From (2.16), with $\mathbf{Y},\mathbf{X}$ centred and $\mathbf{Y}'$ correspondingly relocated,

$$\hat{\mathbf{X}}'=(\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}})^{-1}\hat{\mathbf{B}}\mathbf{S}^{-1}\bar{\mathbf{Y}}'$$

so that from (2.18), (2.19) and Lemma 2

$$\hat{\hat{\mathbf{X}}}' = [(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}+\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}]^{-1}(\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}})\hat{\mathbf{X}}', \tag{2.21}$$

a matrix weighted average between $\hat{\mathbf{X}}'$ and $\mathbf{0}$. It may be written

$$[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-\frac{1}{2}}\hat{\hat{\mathbf{X}}}'] = (\mathbf{I}_p+\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}\mathbf{W}^{\mathrm{T}}[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-\frac{1}{2}}\hat{\mathbf{X}}'] = [(\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}+\mathbf{I}_p]^{-1}[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-\frac{1}{2}}\hat{\mathbf{X}}']$$

using the notation introduced after (2.13). Let

$$\mathbf{U} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-\frac{1}{2}},$$

so that the positive eigenvalues of $\mathbf{U}$ are the usual canonical correlations between $\mathbf{X}$ and $\mathbf{Y}$. Then, since

$$\mathbf{S} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}-\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y},$$

$$(\mathbf{I}+\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}\mathbf{W}^{\mathrm{T}} = (\mathbf{I}+\mathbf{U}\mathbf{V}^{-1}\mathbf{U}^{\mathrm{T}})^{-1}\mathbf{U}\mathbf{V}^{-1}\mathbf{U}^{\mathrm{T}} = \mathbf{V}\mathbf{U}\mathbf{V}^{-1}\mathbf{U}^{\mathrm{T}},$$

where

$$\mathbf{V} = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-\frac{1}{2}}\mathbf{S}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-\frac{1}{2}} = \mathbf{I}-\mathbf{U}^{\mathrm{T}}\mathbf{U}.$$

Thus the eigenvalues of $(\mathbf{I}+\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}\mathbf{W}^{\mathrm{T}}$ are in fact the squared canonical correlations between $\mathbf{X}$ and $\mathbf{Y}$. This follows since $\mathbf{V}^{\frac{1}{2}}\mathbf{U}\mathbf{V}^{-\frac{1}{2}}$ is similar to $\mathbf{U}$ and

$$\mathbf{V}\mathbf{U}\mathbf{V}^{-1}\mathbf{U}^{\mathrm{T}} = \mathbf{V}^{\frac{1}{2}}(\mathbf{V}^{\frac{1}{2}}\mathbf{U}\mathbf{V}^{-\frac{1}{2}})(\mathbf{V}^{-\frac{1}{2}}\mathbf{U}^{\mathrm{T}}\mathbf{V}^{\frac{1}{2}})\mathbf{V}^{-\frac{1}{2}}$$

is similar to $\mathbf{U}\mathbf{U}^{\mathrm{T}}$.

In summary, after the same particular non-singular transformation $\mathbf{Q}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-\frac{1}{2}}$, where columns of $\mathbf{Q}$ are orthonormal latent vectors of $(\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}+\mathbf{I}_p$, components of the transformed $\hat{\hat{\mathbf{X}}}$ and $\hat{\mathbf{X}}$ are simply proportional, the p constants of proportionality being the p squared canonical correlations between $\mathbf{X}$ and $\mathbf{Y}$.

3. Bayesian Multivariate Controlled Calibration

A Bayesian formulation involving diffuse prior distributions retains many features similar to the classical sampling results of Section 2. There are interesting differences however such that the Bayesian solution is worthwhile studying in its own right. In fact the Bayesian formulation allows a development justifying regressing $\mathbf{X}$ on $\mathbf{Y}$ in some circumstances even though $\mathbf{X}$ is controlled. The approach follows that of simple linear regression given by Hoadley (1970) although we have an extra insight even in this case.

Assume the model defined by (2.1), (2.2), (2.3) and (1.5). If $\pi(\cdot)$ denotes a probability density function with a subjective status, assume in addition that

$$\pi(\mathbf{B},\boldsymbol{\alpha},\boldsymbol{\Gamma},\boldsymbol{\xi}) = \pi(\mathbf{B},\boldsymbol{\alpha},\boldsymbol{\Gamma})\,\pi(\boldsymbol{\xi}), \tag{3.1}$$

where the random $\boldsymbol{\xi}$ is *a priori* independent of the parameters (regarded as random) of the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$. Furthermore, assume

$$\pi(\boldsymbol{\xi}\mid\mathbf{X}) = \pi(\boldsymbol{\xi}), \tag{3.2}$$

i.e., the controlled $\mathbf{X}$ values provide no information on $\boldsymbol{\xi}$. Although this might at first glance seem unnatural it may be by-passed by *post hoc* assuming a prior distribution for $\boldsymbol{\xi}$ which is informative in the same fashion as the data $\mathbf{X}$. S-ancillarity allows us to do this. At this stage the prior distribution for $(\mathbf{B},\boldsymbol{\alpha},\boldsymbol{\Gamma})$ may be quite general subject to (3.1). With these assumptions it is straightforward to derive the posterior distribution of $\boldsymbol{\xi}$ on integrating out the unwanted parameters $(\mathbf{B},\boldsymbol{\alpha},\boldsymbol{\Gamma})$.

Theorem 2. With the model defined by (2.1), (2.2), (2.3) and (1.5) together with (3.1) and (3.2)

$$\pi(\boldsymbol{\xi}\mid\mathbf{Y}',\mathbf{Y},\mathbf{X}) \propto \pi(\boldsymbol{\xi})\,L(\boldsymbol{\xi}), \tag{3.3}$$

where $L(\boldsymbol{\xi})$ is the predictive distribution of $\bar{\mathbf{Y}}'$.

Proof. With the normality assumption (1.5), sufficient statistics for $\boldsymbol{\xi}$ and $\boldsymbol{\Gamma}$ are $\bar{\mathbf{Y}}'$ and $\mathbf{S}'$, the sum of products matrix on $(l-1)$ degrees of freedom.

Thus allowing proportionality signs for dropped terms not involving ξ

$$\begin{aligned}\pi(\boldsymbol{\xi} \mid \mathbf{Y}', \mathbf{Y}, \mathbf{X}) &= \pi(\boldsymbol{\xi} \mid \bar{\mathbf{Y}}', \mathbf{Y}, \mathbf{S}', \mathbf{X}) \\ &\propto \pi(\bar{\mathbf{Y}}', \mathbf{Y} \mid \boldsymbol{\xi}, \mathbf{S}', \mathbf{X})\, \pi(\boldsymbol{\xi} \mid \mathbf{S}', \mathbf{X}) \\ &= \pi(\bar{\mathbf{Y}}', \mathbf{Y} \mid \boldsymbol{\xi}, \mathbf{S}', \mathbf{X})\, \pi(\boldsymbol{\xi}),\end{aligned}$$

by (3.1), (3.2) and

$$= \pi(\bar{\mathbf{Y}}' \mid \mathbf{Y}, \boldsymbol{\xi}, \mathbf{S}', \mathbf{X})\, \pi(\mathbf{Y} \mid \boldsymbol{\xi}, \mathbf{S}', \mathbf{X})\, \pi(\boldsymbol{\xi}),$$

and by (3.1)

$$\propto \pi(\bar{\mathbf{Y}}' \mid \mathbf{Y}, \boldsymbol{\xi}, \mathbf{S}', \mathbf{X})\, \pi(\boldsymbol{\xi})$$

which is the required result.

Remark. When $l = 1$, the result applies to non-normal error distributions.

To construct $L(\boldsymbol{\xi})$, the predictive distribution of $\bar{\mathbf{Y}}'$ given $\boldsymbol{\xi}, \mathbf{Y}, \mathbf{X}, \mathbf{S}'$, is straightforward if a standard natural conjugate prior for $(\mathbf{B}, \boldsymbol{\alpha}, \boldsymbol{\Gamma})$ is assumed. The use of an invariant Jeffreys prior

$$\pi(\mathbf{B}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}) \propto |\boldsymbol{\Gamma}|^{-(q+1)} \tag{3.4}$$

leads to the predictive distribution, given, for example, when $l = 1$, as equation (14.2.3) of Press (1972), which formally coincides with the sampling theory result of (2.7). Thus as a function of $\boldsymbol{\xi}$ we have

$$L(\boldsymbol{\xi}) = \{\sigma^2(\boldsymbol{\xi})\}^{v/2} / \{\sigma^2(\boldsymbol{\xi}) + (\bar{\mathbf{Y}}' - \hat{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\xi})^{\mathrm{T}} \mathbf{S}^{-1} (\bar{\mathbf{Y}}' - \hat{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\xi})\}^{(v+q)/2}, \tag{3.5}$$

where we have *post hoc* adopted the scale of $\mathbf{Y}$ centred in the calibration experiment. This means correspondingly that we have replaced $\bar{\mathbf{Y}}' - \hat{\boldsymbol{\alpha}}$ by $\bar{\mathbf{Y}}'$. Thus in standard multivariate linear regression, replacing $\boldsymbol{\xi}$ by $\mathbf{X}'$, (3.5) is the ratio of two quadratic forms in $\mathbf{X}'$.

For large $\| \mathbf{X}' \|$ it behaves like $1/\| \mathbf{X}' \|^q$ and is integrable provided $q \geqslant 2$. When $q = 1$ as in Hoadley (1970) a proper prior $\pi(\mathbf{X}')$ is necessary in (3.4) for overall integrability. Note that (3.5) may be written in the form

$$(1/l + 1/n + \| \mathbf{X}' \|_{\mathrm{G}}^2)^{-q/2} [1 + \{R + \| \mathbf{X}' - \hat{\mathbf{X}}' \|_{\mathrm{H}}^2\} / \{1/l + 1/n + \| \mathbf{X}' \|_{\mathrm{G}}^2\}]^{-(v+q)/2}, \tag{3.6}$$

where $R = \bar{\mathbf{Y}}'^{\mathrm{T}} \mathbf{S}^{-1} \bar{\mathbf{Y}}' - \hat{\mathbf{X}}'^{\mathrm{T}} \hat{\mathbf{B}} \mathbf{S}^{-1} \hat{\mathbf{B}}^{\mathrm{T}} \hat{\mathbf{X}}'$, the residual sum of squares from the prediction experiment, and $\mathbf{G} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, $\mathbf{H} = \hat{\mathbf{B}} \mathbf{S}^{-1} \hat{\mathbf{B}}^{\mathrm{T}}$ are the crucial ingredients of Condition 1. If it were not for the first factor, (3.6) would be maximized by $\mathbf{X}' = \hat{\mathbf{X}}'$. The first factor tends to shift this maximum towards the origin but the effect will be slight as v increases relative to q. The behaviour of (3.6) could be investigated by simultaneously diagonalizing $\mathbf{G}, \mathbf{H}$; that is, it depends on the eigenvalues of $\mathbf{G}^{-1}\mathbf{H}$ or $\mathbf{W}\mathbf{W}^{\mathrm{T}}$ of Section 2.4; in other words the canonical correlations between $\mathbf{X}$ and $\mathbf{Y}$.

3.1. *A Method of Comparing Distinct Models with Regard to Prediction*

The integrated likelihood (3.5) or the posterior distribution (3.3) for $\boldsymbol{\xi}$ may be maximized under two distinct models and the ratio of the resulting maxima used as a basis for determining the preferred model in estimating $\mathbf{X}'$ from a particular $\mathbf{Y}'$. Two distinct models might for example be one involving just linear $\mathbf{X}$ and the other both linear and quadratic $\mathbf{X}$. Large ratios, say greater than five, would favour the numerator model, but see Jeffreys (1961, Appendix B) for guidelines. An average on the log-scale of those n maxima corresponding to the n values of $\mathbf{Y}$ would provide a comparison of the two models over a rather natural range of future $\mathbf{Y}'$. An approximation to this avoiding maximization would substitute the corresponding $\mathbf{X}$ in $\boldsymbol{\xi}$.

Instead of separately maximizing the posterior distributions it might be more appropriate to consider the ratio of posterior probabilities over all $\mathbf{X}'$ in a region defined as the highest posterior density region of $\mathbf{X}'$ under the larger model. See Pericchi (1981) for such an approach in a different context.

The choice of a response subset offers different problems. Since differing quantities of data are involved, the absolute values of the likelihoods are not commensurable and the posterior distributions will need different normalizing constants. One way of avoiding this difficulty is to difference the second derivatives of the log "likelihood" obtained from (3.5) or (3.3) at the maxima. In the special case of linear X throughout this would produce results similar to Section 2.2.3 with an entirely different motivation however. In this paper the method of Section 2.3 only has been tried. It has been used on the paint quality data in Section 5.

3.2. *A Special Prior for* $\mathbf{X}'$ *in Standard Multivariate Linear Calibration*

Assume that we are dealing with a prediction observation without replication, $l = 1$. Hoadley in the univariate case $q, p = 1$ noted that a particular Student distribution prior for $\mathbf{X}'$ knocks out the numerator of (3.5) and gives a Student posterior distribution for $\mathbf{X}'$. This prior may not be unreasonable depending on the design of the calibration experiment. The following theorem extends Hoadley (1970).

Theorem 3. Suppose *a priori* $\mathbf{X}'$ has a multivariate Student distribution, in our previous notation $T(\upsilon-p;\, (1+1/n)\,\mathbf{X}^{\mathrm{T}}\mathbf{X})$, then using (3.4), (3.5), *a posteriori* $(\mathbf{X}'-\hat{\mathbf{X}}')$ has a multivariate Student distribution, $T(\upsilon+q-p;\, \{1+1/n+\mathbf{Y}'^{\mathrm{T}}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}'\}(\mathbf{G}+\mathbf{H})^{-1})$, where $\mathbf{G} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, $\mathbf{H} = \hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}$ and where columns of $\mathbf{X}, \mathbf{Y}$ have been centred *post hoc*.

Proof.

$$\pi(\mathbf{X}') \propto (1+1/n+\mathbf{X}'^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}')^{-\upsilon/2}$$

and from (3.4), (3.5)

$$\begin{aligned}\pi(\mathbf{X}' \mid \mathbf{X}, \mathbf{Y}, \mathbf{Y}') &\propto [1+1/n+\mathbf{X}'^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}'+(\mathbf{Y}'-\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}')^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{Y}'-\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}')]^{-(\upsilon+q)/2}\\ &= [\mathbf{X}'^{\mathrm{T}}\{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}+\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\}\mathbf{X}'-\mathbf{Y}'^{\mathrm{T}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}'-\mathbf{X}'^{\mathrm{T}}\\ &\qquad \times \hat{\mathbf{B}}\mathbf{S}^{-1}\mathbf{Y}'+\mathbf{Y}'^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{Y}'+1+1/n]^{-(\upsilon+q)/2};\end{aligned} \tag{3.7}$$

completing the square, using (2.20), (2.21) and Lemma 2,

$$= [(\mathbf{X}'-\hat{\mathbf{X}}')^{\mathrm{T}}(\mathbf{G}+\mathbf{H})(\mathbf{X}'-\hat{\mathbf{X}}')+1+1/n+\mathbf{Y}'^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{Y}'-\hat{\mathbf{X}}'^{\mathrm{T}}(\mathbf{G}+\mathbf{H})\hat{\mathbf{X}}']^{-(\upsilon+q)/2}.$$

Now

$$\begin{aligned}\mathbf{Y}'^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{Y}'-\hat{\mathbf{X}}'^{\mathrm{T}}(\mathbf{G}+\mathbf{H})\hat{\mathbf{X}}' &= \mathbf{Y}'^{\mathrm{T}}[\mathbf{S}^{-1}-\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}+\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\}^{-1}\hat{\mathbf{B}}\mathbf{S}^{-1}]\mathbf{Y}'\\ &= \mathbf{Y}'^{\mathrm{T}}(\mathbf{S}+\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{B}})^{-1}\mathbf{Y}' = \mathbf{Y}'^{\mathrm{T}}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}',\end{aligned}$$

by the binomial inverse theorem. The result follows.

Remark 1. The prior assumed for $\mathbf{X}'$ is the same as the posterior predictive distribution that would obtain from regarding rows of $\mathbf{X}$ as independent $N(\boldsymbol{\theta}_2, \boldsymbol{\Sigma}_{22})$ with a prior

$$\pi(\boldsymbol{\theta}_2, \boldsymbol{\Sigma}_{22}) \propto |\boldsymbol{\Sigma}_{22}|^{(q-1)/2},$$

and $(1-q) = (q+1)-2q$. This prior although not the Jeffreys invariant prior is exactly the correct prior for recreating the joint posterior predictive distribution of $(\mathbf{X}', \mathbf{Y}')$ given $\mathbf{X}, \mathbf{Y}$ when that joint distribution of $p+q$ variables arises from sampling from $N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with a prior

$$\pi(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(q+1)/2}.$$

This may be seen from evaluating the Jacobian of the transformation from

$$(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \rightarrow (\boldsymbol{\alpha}, \mathbf{B}, \boldsymbol{\Gamma};\, \boldsymbol{\Sigma}_{22}, \boldsymbol{\theta}_2)$$

as given for example by Dawid, Stone and Zidek [1973, equation (A.1.2)]. That we can regard (3.3) as an actual posterior distribution hinges on $\mathbf{X}$ being S-ancillary for the parameters of the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ in normal sampling.

Remark 2. The posterior predictive distribution of Theorem 3 is the predictive conditional distribution of $\mathbf{X}'$ given $\mathbf{Y}', \mathbf{X}, \mathbf{Y}$ that obtains from sampling from a normal conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ with parameters $\boldsymbol{\alpha}^*, \mathbf{B}^*, \boldsymbol{\Gamma}^*$ and prior proportional to

$$|\boldsymbol{\Gamma}^*|^{-(q+1)/2}.$$

This is in accord with Remark 1 since the prior of the theorem exactly recreates the joint predictive distribution of $(\mathbf{X}', \mathbf{Y}')$ and we see that various routes to the conditional distribution cohere. This coherence was not noted even in the univariate case by either Hoadley (1970) or Aitchison and Dunsmore (1975).

For a Bayesian, in controlled calibration, the design implications of Theorem 3 are qualitatively self-evident. From a precision viewpoint it may well be desirable to over-sample in a less probable $\mathbf{X}$ set. It is then in principle straightforward to incorporate the prior for $\mathbf{X}'$ into (3.3), even though it no longer reflects the $\mathbf{X}$ of the controlled calibration.

Remark 3. If $l > 1$ the natural generalization of Theorem 3 is to take the posterior distribution of $\mathbf{X}'$ proportional to (3.5), defined by (2.5) with $l > 1$, divided by the same $v/2$ power of (2.5), but with $l = 1$. The mean or mode of this distribution may also be regarded as the correct generalization of the Krutchkoff estimator to $l > 1$ in the simple $p = q = 1$ case. Such a Bayes estimator is consistent as n, l both tend to infinity. For finite l but infinite n it is biased but this is hardly the drawback Berkson (1969) implies. We need only look at his simulations. The classical estimator is only superior in mean square error for $\mathbf{X}'$ extreme relative to the design $\mathbf{X}$.

4. Wheat Quality Data Analysed

4.1. *The Data and Criterion for Prediction*

The four (derived) infrared reflectance responses and accurate determinations of per cent water, X_1, and per cent protein, X_2, are given in Table 1. For a description of the theory and

Table 1

Twenty-one samples of hard wheat, four infrared reflectance measurements plus laboratory determinations of percentage water and protein

Observation number	Y_1	Y_2	Y_3	Y_4	*% Water*	*% Protein*
1	361	108	96	243	9·00	10·73
2	361	107	98	245	8·94	11·05
3	362	110	94	241	9·12	9·86
4	362	105	94	246	9·06	11·41
5	362	104	70	221	10·02	11·57
6	367	113	75	221	10·06	9·42
7	366	108	82	233	9·52	10·93
8	360	104	86	236	9·32	11·61
9	362	113	85	229	9·56	8·82
10	360	103	90	242	9·10	11·81
11	351	97	88	238	9·14	12·33
12	353	95	73	227	9·70	12·93
13	352	97	77	228	9·60	12·69
14	355	96	52	206	10·62	13·13
15	357	106	69	216	10·04	10·41
16	351	93	69	222	10·00	13·57
17	363	113	88	231	9·46	9·26
18	363	110	101	248	8·86	9·82
19	366	114	79	224	9·78	9·46
20	350	96	85	235	9·34	12·85
21	355	97	63	216	10·12	12·81

technology of infrared reflectance see Rotolo (1979). The 21 samples have been randomly ordered so that the last five, observations 17 to 21, form a random sample from the 21 observations and will be used for prediction purposes, that is the relationship between **Y** and **X** is estimated from observations 1 to 16 and then applied to the *Y*-observations 17 to 21 to obtain predictions for the corresponding five pairs of **X**-values which are then compared with their true values. The data forms an example of random calibration since both **X** and **Y** are random.

The application of the various prediction formulae are facilitated if over the 16 observations both **Y**, **X** are centred

$$\sum_{i=1}^{16} x_{ij} = 0, \quad j = 1, 2$$

and

$$\sum_{i=1}^{16} y_{ij} = 0, \quad j = 1, \dots, 4.$$

The adjustments necessary to achieve this were then also applied to observations 17 to 21. Note that this does not mean that all the 21 observations are centered. Now, since the **X**-data in the calibrating experiment have been centered and the same centring applied to $\mathbf{X}_{17}, \dots, \mathbf{X}_{21}$, in the absence of any calibratory information, zero would estimate the **X**-values and a natural criterion for prediction accuracy is

$$100 \sum_{i=17}^{21} (x_{ij} - \hat{x}_{ij})^2 \Big/ \sum_{i=17}^{21} x_{ij}^2, \tag{4.1}$$

the percentage of unexplained variation.

4.2. *The Methods of Prediction*

Three basic methods were used to predict **X**. They were as follows:

(L) From the regression of **Y** on **X** (formula 2.16).

(E) Empirical prediction. This method is a multivariate extension of Lwin and Maritz (1980). Like (L) is uses the parametric regression of **Y** and **X**, but derives from that of **X** on **Y** by means of the empirical distribution of **X**. Specifically if $\mathbf{Y}'$ (4 × 1), is a set of four responses (any one of observations $\mathbf{Y}_{17}, \dots, \mathbf{Y}_{21}$) then the prediction for the corresponding $\mathbf{X}'$ (2 × 1), is

$$\sum_{i=1}^{16} w_i (x_{1i}, x_{2i})^{\mathrm{T}},$$

where

$$w_i = f(\mathbf{Y}' \mid x_{1i}, x_{2i}) \Big/ \sum_{i=1}^{16} f(\mathbf{Y}' \mid x_{1i}, x_{2i}).$$

Here *f* was assumed to be the multivariate normal regression density with four responses and 2 regressors and parameters fixed at their least-squares values.

(LB) The regression of **X** on **Y** (formula 2.18). LB here is an abbreviation for Linear Bayes. Although the estimator can be thought of purely as arising from the regression of **X** on **Y**, it has the property of being Bayes under the special prior of section (3.2), and the property more naturally extends to the case $l > 1$, as indicated in Remark 3 following Theorem 3. Hence our preference for the Bayesian terminological link.

In addition, methods L and E were applied to the separate problems of predicting percent water ignoring all the percent protein measurements and *vice versa*. This gives methods L′, E′. Note that method LB automatically predicts separately X_1' without reference to X_2' and *vice versa*.

TABLE 3
The percentage of variation unexplained for wheat quality prediction

Method	*Water*	*Protein*
L	1·7	1·7
L′	1·7	1·7
E	6·8	9·0
E′	6·0	2·5
LB	1·5	1·7

4.3. *The Predictions Compared*

Table 3 gives the unexplained percentages of variation in predicting the five wheat samples. Overall it is evident that predictions are rather good. The best method explaining more than 98 per cent of the variation in **X**′. From the earlier discussions this implies that there will be little to choose between methods regressing **Y** on **X** and those regressing **X** on **Y**, This is indeed the case. Method LB, the preferred one for random calibration is indeed slightly better than method L, that based on regressing **Y** on **X**. The semi-non-parametric method E does rather badly in comparison. This is surprising since a plot of ordered $(\mathbf{x}_i-\bar{\mathbf{x}})^T\hat{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})$ against exponential order statistics suggests that the marginal distribution of **X** is far more flat than a normal and this is a circumstance where one might hope that method E would do better than LB. Perhaps its use of point estimators of parameters in the predictive distribution of **Y**′ is its downfall.

Finally, it is interesting to note that the methods L′, E′ which take one regression **X** at a time tend to fare better than those which consider the two regressions simultaneously.

5. PAINT FINISH DATA ANALYSED

5.1. *The Paint Data*

A single patch of paint base was tinted with a pigment at three levels (0 per cent, 0·15 per cent, 0·30 per cent), and the viscosity of these samples was adjusted before spraying to one of three levels (30, 33, 36 seconds in an efflux cup); each of the resulting paints was replicated four times, giving a total of 36 dry panels.

Each of these panels was measured for optical properties in three ways:

1. Spectrometer measurements of incident light. Measurements at two different inclinations were used to create three responses, Y_1, Y_2, Y_3 of Table 2.
2. Integrated reflectance with normal incident light, Y_4 of Table 2.
3. Peak-height and band-width on a recording goniophotometer, Y_5, Y_6 of Table 2.

In future it is desired to use a subset of the six responses to predict the pigmentation and viscosity levels used so as to match the paint.

In order to compare various methods of prediction of **X** (pigmentation and viscosity) from **Y** a random replicate, observations 2, 5, 11, 16, 18, 22, 28, 30, 35, was extracted for prediction. The nine resulting **Y**'s were to be used to predict the corresponding **X**; the 27 remaining three replicates of the 3 × 3 experiment were used to estimate the relationship between **X** and **Y** and to choose a subset of responses. Notice how there is an underlying scale to both pigmentation and viscosity. In what follows it is not therefore unreasonable to use a criterion for prediction accuracy which presupposes such a scale.

5.2. *Choice of a Subset of Responses*

It would be desirable to reduce the six responses to two, to be used jointly for subsequent prediction of pigmentation and viscosity. With this aim the test for additional information of

TABLE 2

Two factors, pigmentation (P) and viscosity (V) in a 3 × 3 experiment with four replicates and six optical responses

P	V	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
0	0	1·88	35·0	75·0	40·94	101·0	20·0
0	0	1·87	35·3	75·8	40·68	114·0	17·5
0	0	1·88	36·1	77·0	40·60	101·0	19·8
0	0	1·87	36·8	77·0	40·57	100·5	17·5
0	1	1·79	32·4	73·2	39·83	95·0	20·6
0	1	1·78	31·9	72·9	39·65	107·0	19·5
0	1	1·70	29·6	72·1	39·15	93·5	21·0
0	1	1·73	30·6	72·5	39·44	93·0	18·0
0	2	1·63	25·7	66·7	37·22	93·5	22·0
0	2	1·65	26·8	67·9	37·89	86·0	21·3
0	2	1·61	23·8	62·9	37·36	84·0	21·3
0	2	1·68	27·2	67·2	38·15	84·5	19·0
1	0	1·79	33·5	76·0	39·09	102·0	21·0
1	0	1·77	31·3	71·8	39·12	105·0	19·8
1	0	1·80	31·8	71·8	39·31	103·0	20·0
1	0	1·78	31·8	72·5	38·73	101·0	20·8
1	1	1·74	30·5	71·5	39·31	103·0	20·1
1	1	1·68	28·7	71·1	38·64	98·5	20·9
1	1	1·71	29·6	71·1	38·50	99·0	21·0
1	1	1·73	29·5	70·0	39·09	101·0	20·8
1	2	1·50	21·0	63·0	35·82	85·0	21·1
1	2	1·52	21·9	63·9	35·65	85·0	21·8
1	2	1·51	21·2	63·0	35·70	84·0	22·5
1	2	1·50	20·6	61·6	35·77	85·0	22·2
2	0	1·94	35·8	74·0	38·0	101·0	19·5
2	0	1·89	33·9	72·0	38·08	101·0	20·1
2	0	1·92	35·0	73·0	37·93	92·5	19·5
2	0	1·92	33·7	70·5	38·17	83·0	21·8
2	1	1·87	33·0	71·0	37·18	96·0	21·0
2	1	1·85	31·5	68·5	37·17	91·5	22·5
2	1	1·89	33·0	70·0	37·83	99·0	19·5
2	1	1·86	31·5	68·0	37·31	95·0	20·1
2	2	1·75	27·8	64·8	34·36	82·5	20·8
2	2	1·60	24·6	67·4	34·09	71·0	21·0
2	2	1·62	24·0	63·0	34·16	80·0	21·4
2	2	1·62	23·0	60·0	34·18	86·0	20·7

Section 2.3 was used. If it transpired that one could get away with using two responses so much the better, if not then at the price of a slightly more complicated predictor greater accuracy would obtain.

In this initial screening of the responses, it was decided to restrict attention to multivariate linear regression. This avoids difficulties over a non-monotonic relationship of **Y** on **X**. Even though additional non-linear information might be ignored by this process, it might not anyway be subsequently usable.

The matrix of the quadratic form (2.17) is 2 × 2. If its eigenvalues are both negative then, whatever $\mathbf{X}'$, the quadratic form is negative. Furthermore, both eigenvalues will be negative if and only if the trace (= sum of eigenvalues) is negative and the determinant (= product of eigenvalues) positive. Table 4 lists both trace and determinant for various subsets of the six responses when the significance level was 5 per cent. There is only one pair of responses, the first and fourth which has a negative definite quadratic form, so that discarding variables Y_2, Y_3, Y_5, Y_6 involves no loss in (linear) information. This pair is subsequently used for prediction.

TABLE 4

Test of additional information for every subset pair of six responses at the 5 per cent significance level (quadratic form (2.17))

Response subset	*Trace*	*Determinant*
110000	0·21	−0·03
101000	0·29	−0·03
100100	−0·14	0·0005
100010	0·48	−0·04
100001	0·59	−0·02
011000	0·31	−0·03
010100	−0·10	−0·01
010010	0·37	−0·07
010001	0·48	−0·04
001100	−0·11	−0·05
001010	0·18	−0·08
001001	0·38	−0·02
000110	−0·03	−0·02
000101	0·05	−0·02
000011	0·59	0·03

5.3. *The Methods of Prediction of Pigmentation and Viscosity*

Broadly two sets of methods of prediction were tried, linear methods {L′, L, LB} and quadratic methods {Q′, QB′}. As in Section 4, those methods which ignore one of the two explanatory factors, pigmentation or viscosity, in both the calibrating experiment and the prediction experiment are identified by a dash superfix. The linear methods are described in Section 4.2 and may be found explicitly by the appropriate formulae. The two quadratic methods allow a full parameterisation of the three levels of the explanatory factors taken one at a time. Orthogonal polynomials, x linear -1, 0, 1 and quadratic -1, 2, -1 were used respectively.

(Q′). The left-hand side of (2.8) was plotted as a function of pigmentation and then separately for viscosity in x steps of 0·1 with $\boldsymbol{\xi}^{\mathrm{T}} = (x, 2-3x^2)$. The matrix $\mathbf{S}$ is the residual sum of products after fitting linear and quadratic pigmentation (or viscosity). The minimum of the resulting function was taken to be the estimator (Q′). This was done in turn for each of the nine pairs $\mathbf{Y}'$.

(QB′). Here QB stands for Quadratic Bayes. By analogy with the linear Bayes case the denominator of (3.5), with $\boldsymbol{\xi}$ replacing $\mathbf{X}'$, was plotted as a function of x in steps of 0·1 where $\boldsymbol{\xi}^{\mathrm{T}} = (x, 2-3x^2)$ as in (Q′). From (3.7) it is easy to see that the denominator of (3.5) differs from the left-hand side of (2.8) in (i) the divisor $\sigma^2(\boldsymbol{\xi})$ and (ii) the matrix of the quadratic form, $\hat{\mathbf{B}}\mathbf{S}^{-1}\hat{\mathbf{B}}^{\mathrm{T}}$, is augmented by $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ in (3.5). This method is a slightly *ad hoc* alternative to using either (a) the integrated likelihood (3.5) in its entirety or (b) (3.5) modified by dividing by $(\sigma^2(x))^{v/2}$ the appropriate prior factor from the Hoadley prior.

5.4. *The Predictions Compared*

Table 5 gives the percentages of unexplained variation aggregated over the nine observations of the prediction experiment using the responses Y_1, Y_4 (cf. formula (4.1)). Inspection of the raw data emphasizes that the effect of pigmentation on Y_1 is far from linear and is not monotonic even. Although all the other quadratic effects are not significant, it is rather a surprise that Q′ does so badly in predicting both pigmentation and viscosity and quite remarkable how well the quasi-Bayes method QB′ performs. It does hardly any worse than the linear method LB with viscosity, where linearity is appropriate, and is able to utilize the quadratic relationship with pigmentation. In fact Q′ was perhaps worse than indicated. The plotted function generally had two local minima and the global minimum for the fourth of the

Table 5
Percentages of unexplained variation for the paint data

Method	*Pigmentation*	*Viscosity*
L	24	28
L′	21	28
LB	21	20
Q′	25	35
QB′	13	21

nine predictions of pigmentation was −1·4 with a second minimum at 0·3 and a 95 per cent confidence region of −1·9 to 0·7 (true value of zero). In Table 5, instead of −1·4, the mid-interval value of −0·6 was used. Otherwise the 25 per cent unexplained variation for this method would have been 52 per cent! See the next Section for more details.

As one check on the computations for Q′, QB′, the quadratic parameter estimates were set to zero and the sum of products matrices adjusted accordingly. The plots then obtained gave minima which corresponded to L′ and LB respectively. The plots were reasonable symmetric about a single maximum. The confidence interval from such a plot for L′ is adequately approximated by computing the scale factors of the univariate student distribution of Theorem 3. These range from 4·4 to 4·8 over the nine values of **Y**′ for pigmentation and also separately viscosity. Since the degrees of freedom are 25, the estimated variance may then be obtained by dividing by 23, giving a standard error of approximately 0·5.

Finally it may be noted that estimates LB are proportional to L′, a consequence shown in Section 2.4. In fact the LB estimate = 0·77 × the estimate L′, the squared multiple correlation of 0·77 being coincidently the same to two decimal places for pigmentation and viscosity.

5.5 *Confidence Intervals for Prediction*

Figs 1, 2(a) and 2(b) relate to giving confidence intervals for the true pigmentation values of

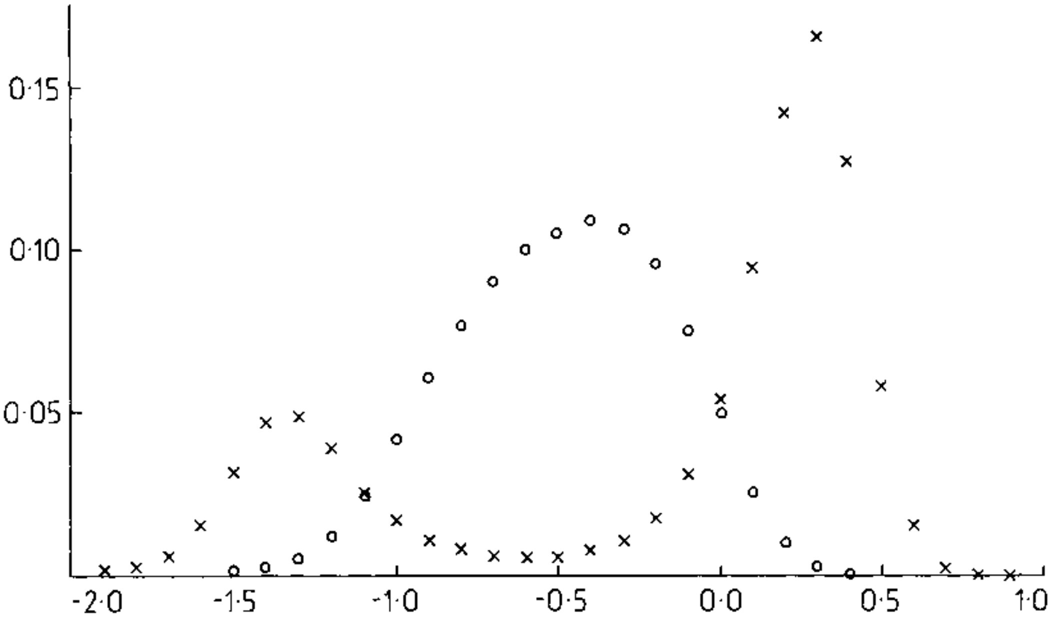

Fig. 1. Discretized posterior probabilities for the quadratic model for observation 4 (×) and 5 (○).

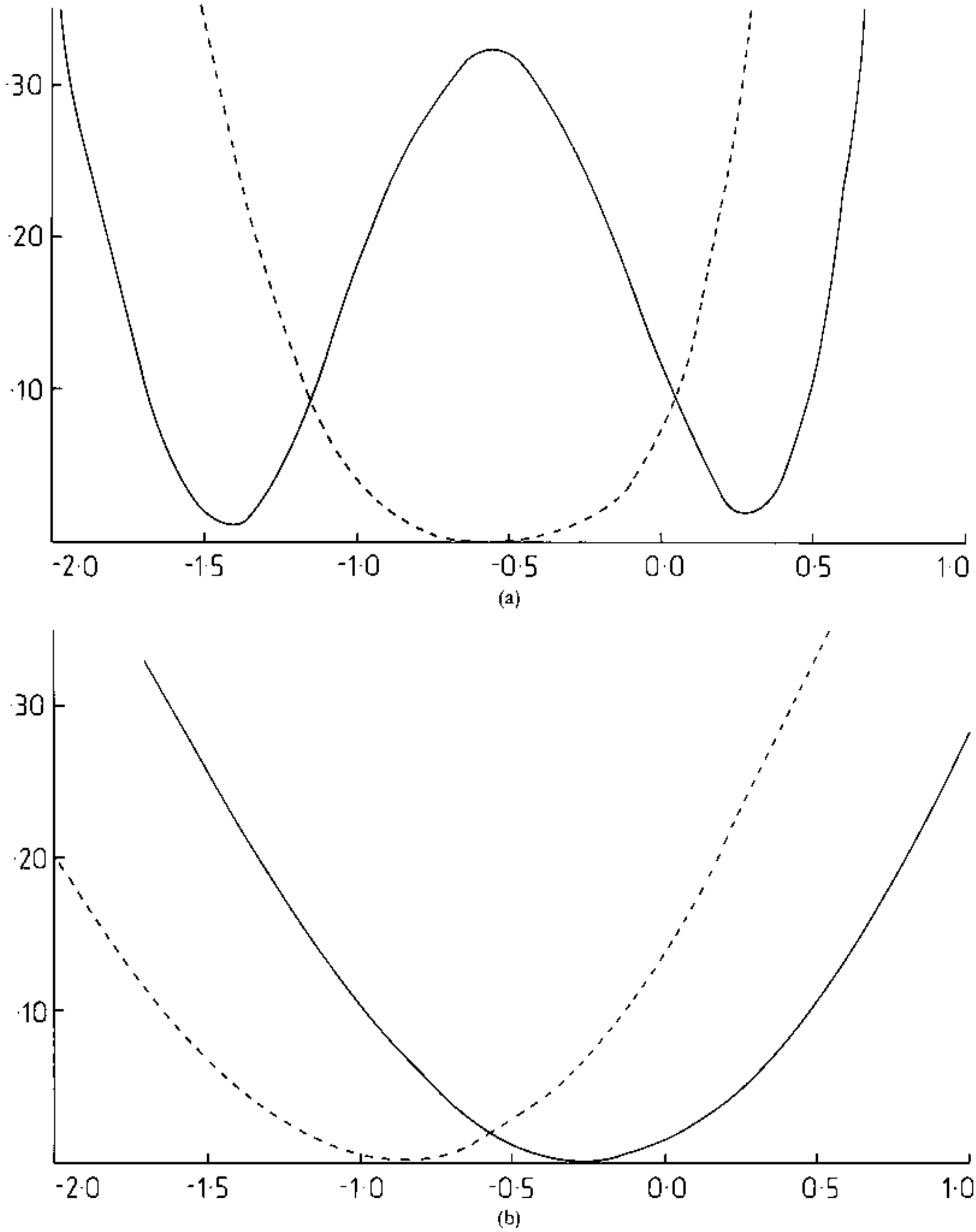

FIG. 2. Plot of the left-hand side of (2.8): continuous curve, observation 4; dotted curve, observation 5. (a) Quadratic model. (b) Linear model.

observations 4 and 5 of the nine observation prediction set (numbers 16 and 18 of the original listing). The true values are zero for each of these two observations. Pigmentation is being predicted completely ignoring viscosity measurements.

Fig. 1 gives posterior probabilities discretized in intervals of 0·1 for the quadratic model described under (b) of Section 5.3. That is, it has "linear" prior assumptions. Notice that whilst the probabilities for observation 5 constitute a smooth unimodal distribution, those for observation 4 are somewhat bimodal. Highest posterior density intervals easily obtain from the plot. Similar "confidence intervals" under LB, the regression of **X** on **Y**, have not been plotted but may be calculated from Theorem 3. The means are 0·77 times the L′ predictors, namely $-0{\cdot}20$ and $-0{\cdot}66$ from observations 4 and 5 respectively. Then from a univariate student t-distribution on 25 degrees of freedom, we have the symmetric 95 per cent "confidence intervals" $(-1{\cdot}05, 0{\cdot}65)$ and $(-1{\cdot}52, 0{\cdot}20)$.

Figs 2(a) and 2(b) are plots of the left-hand side of (2.8) for quadratic and linear models respectively. Thus confidence intervals for the two methods Q′ and L′ result from taking all **X**′ values less than $(\mathrm{q}/\mathrm{v})\,F_v^q(\gamma)$. For 95 per cent intervals $\gamma = 0{\cdot}05$ and since $q = \mathrm{q}$ then $v = 23$, 24 for Q′ and L′ respectively. Hence the threshold values are 0·30 and 0·28. Since the Q′ curve is bimodal for observation 4, two disjoint sets of points constitute the confidence "interval"! The curve even favours some points outside $(-1, 1)$ more than those inside. However sanity returns with observation 5 and this is far more typical of the behaviour for other observations not plotted here. Although observation 4 is rather a maverick it seems to be treated more reasonably by the Bayesian method: just compare Figs 1 and 2(a). Intervals from Fig. 2(b) are rather wider than those from LB′ given above.

6. CONCLUSION

It is possible to draw some tentative conclusions from the theoretical results of Sections 2 and 3 and the applications of Sections 4 and 5. In random calibration, one should regress **X** on **Y** to predict **X** in the future. The semi-nonparametric approach to this (Section 4) did not fare well. Furthermore, a controlled calibration experiment will typically have **X**-values chosen to cover the range of future values. This implies, Section 5, that the linear regression of **X** on **Y** is still superior to that of **Y** on **X**; and the generalization to polynomial dependence is best implemented using the Bayesian approach of Section 3, regression of **X** on **Y** being the linear special case, rather than the classical approach of Section 2. The classical confidence regions seem unnecessarily wide in the polynomial case. They can also be disjoint intervals.

Response variable selection, by the classical procedure of Section 2.4, presented no difficulties or evident flaws as applied in Section 5, though Section 3.1 might be worth developing further. The applications suggest that one should predict **X** one variable or factor at a time. This is necessarily so in the linear case where the regression of **X** on **Y** is superior. Finally marginal monotonicity of **Y** versus **X** is not necessary in multivariate calibration, even though it is essential for univariate calibration.

ACKNOWLEDGEMENTS

I am grateful to Dr T. Fearn of the Flour, Milling and Baking Research Association for providing the data of Table 1. A. Azzalini performed the computations for this wheat data with admirable accuracy and speed. A discussion with R. Sundberg led me to fresh thoughts on the problem of model selection.

REFERENCES

AITCHISON, J. and DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge: University Press.

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

BERKSON, J. (1969). Estimation of a linear function for a calibrating line; consideration of a recent proposal. *Technometrics*, **9**, 649–660.

BROOKS, R. J. (1974). On the choice of an experiment for prediction in linear regression. *Biometrika*, **61**, 303–11.

BROWN, P. J. and ZIDEK, J. V. (1980). Adaptive multivariate ridge regression. *Ann. Statist.*, **8**, 64–74.

BROWNLEE, K. A. (1960). *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley.

CLARK, R. M. (1979). Calibration, cross validation and carbon-14. I. *J. R. Statist. Soc.*, A **142**, 47–62.

COX, D. R. and SMALL, N. J. H. (1978). Testing multivariate normality. *Biometrika*, **65**, 263–72.

DAWID, A. P. (1976). Properties of diagnostic data distributions. *Biometrics*, **32**, 647–58.

—— (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, **68**, 265–74.

DAWID, A. P., STONE, M. and ZIDEK, J. V. (1973). Marginalisation paradoxes in Bayesian and structural inference (with Discussion). *J. R. Statist. Soc.*, B, **35**, 189–233.

DICKEY, J. M. (1967). Matricvariate generalisations of the multivariate *t* distribution and the inverted multivariate *t* distribution.

DRAPER, N. and SMITH, H. (1981). *Applied Regression Analysis*, 2nd edn. New York: Wiley.

EFRON, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Ass.*, **70**, 892–8.

EISENHART, C. (1939). The interpretation of certain regression methods and their use in biological and industrial research. *Ann. Math. Statist.*, **10**, 162–86.

FIELLER, E. C. (1954). Some problems in interval estimation. *J. R. Statist. Soc., B*, **16**, 175–85.

GEISSER, S. (1964). Posterior odds for multivariate normal classifications. *J. R. Statist. Soc.*, B, **26**, 69–76.

GRASSIA, A. and SUNDBERG, R. (1982). Statistical precision in the calibration and use of sorting machines and other classifiers. To appear in *Technometrics*.

HEALY, M. J. R. (1968). Multivariate normal plotting. *J. R. Statist. Soc.*, C, **17**, 157–61.

HOADLEY, B. (1970). A Bayesian look at inverse linear regression. *J. Amer. Statist. Ass.*, **65**, 356–69.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd Edn. Oxford: University Press.

KANAL, L. (1974). Patterns in pattern recognition: 1968–74. *IEEE Trans. Inf. Th.*, **20**, 697–722.

KRUTCHKOFF, R. G. (1967). Classical and inverse regression methods of calibration. *Technometrics*, **9**, 425–39.

LINDLEY, D. V. (1972). *Bayesian Statistics: A Review*. Philadelphia: SIAM.

LUNDBERG, E. and DE MARE, J. (1980). Interval estimates in the spectroscopy calibration problem. *Scand. J. Statist.*, **7**, 40–42.

LWIN, T. and MARITZ, J. S. (1980). A note on the problem of statistical calibration. *J. R. Statist. Soc.*, C, **29**, 135–41.

MINDER, CH. E. and WHITNEY, J. B. (1974). A likelihood analysis of the linear calibration problem. *Technometrics*, **17**, 463–71.

PELLA, J. J. and ROBERTSON, T. L. (1979). Assessment of composition of stock mixtures. *Fishery Bulletin*, **77**, 387–98.

PERICCHI, L. R. (1981). A Bayesian approach to transformations to normality. *Biometrika*, **68**, 35–44.

PRESS, S. J. (1972). *Applied Multivariate Analysis*. New York: Holt, Rinehart and Winston.

RAO, C. R. (1965). *Linear Statistical Inference and its Applications*. New York: Wiley.

ROTOLO, P. (1979). New infrared reflectance instrumentation. *Cereal Foods World*, **24**, 94–98.

SCHEFFÉ, H. (1973). A statistical theory of calibration. *Ann. Statist.* **1**, 1–37.

TALLIS, G. M. (1969). Note on a calibration problem. *Biometrika*, **56**, 505–8.

THEOBALD, C. M. and MALLINSON, J. R. (1978). Comparative calibration, linear structural relationships and congeneric measurements. *Biometrics*, **34**, 39–45.

TITTERINGTON, D. M., MURRAY, G. D., MURRAY, L. S., SPIEGELHALTER, D. J., SKENE, A. M., HABBEMA, J. D. F. and GELPKE, G. J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients (with Discussion). *J. R. Statist. Soc.* A, **144**, 145–175.

WILLIAMS, E. J. (1959). *Regression Analysis*. New York: Wiley.

— (1969). Regression methods in calibration problems. *Bull. ISI.*, **43**, 17–28.

APPENDIX

Using a canonical form of model (2.1) and (2.2) centred and scaled as in (2.3), the distributional result (2.7) is developed. Let **P** be an orthogonal $p \times p$ matrix of eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$; if we define $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ to be the $p \times p$ diagonal matrix of eigenvalues of the correlation matrix then

$$\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P} = n\Lambda,$$

where $\Sigma\, \lambda_i = p$. Limiting arguments as $n \to \infty$ are easy to incorporate since Λ may sensibly remain constant.

At times in deriving results it is simpler to work with the canonical form. Since **X** is fixed we may multiply both sides of (2.1) by a $n \times n$ orthogonal matrix **Q** so that

$$\begin{aligned} \mathbf{Z}_0^{\mathrm{T}} &= n^{\frac{1}{2}}\boldsymbol{\alpha}^{\mathrm{T}} + \mathbf{e}_0^{\mathrm{T}} \\ \mathbf{Z}_1 &= n^{\frac{1}{2}}\Lambda^{\frac{1}{2}}\mathbf{A} + \mathbf{e}_1 \qquad\qquad (\text{A}.1.1) \\ \mathbf{Z}_2 &= \mathbf{e}_2 \end{aligned}$$

where $\mathbf{Z}_0$ is $q \times 1$, $\mathbf{Z}_1$ is $p \times q$ and $\mathbf{Z}_2$ is $(n-p-1) \times q$ where $\mathbf{Y}^{\mathrm{T}}\mathbf{Q} = (\mathbf{Z}_0\, \mathbf{Z}_1^{\mathrm{T}}\, \mathbf{Z}_2^{\mathrm{T}})$ with $\mathbf{A} = \mathbf{P}^{\mathrm{T}}\mathbf{B}$. Strictly $\mathbf{E} \rightarrow \mathbf{Q}^{\mathrm{T}}\mathbf{E}$ but since $\mathbf{Q}$ is orhtogonal and $\mathbf{E}$ is normal $1 \times q$ rows of $\mathbf{e}_0^{\mathrm{T}}\mathbf{e}_1$, $\mathbf{e}_2$ are independent identically distributed $N(\mathbf{O}, \Gamma)$ as before. As each observation in the prediction experiment (2.2) has the same mean, this is particularly simple to reduce to a canonical form. The $l \times l$ orthogonal matrix $\mathbf{Q}'$ may have its first column proportional to the unit vector and the $l-1$ other columns any orthogonal set of vectors orthogonal to the unit vector. If $\mathbf{Y}'^{\mathrm{T}} = (\mathbf{Y}'_1 \ldots \mathbf{Y}'_l)$ then $\mathbf{Y}'^{\mathrm{T}}\mathbf{Q}' = (\mathbf{Z}'_1, \mathbf{Z}'^{\mathrm{T}}_2)$ and

$$\begin{aligned} \mathbf{Z}'^{\mathrm{T}}_1 &= l^{\frac{1}{2}}(\boldsymbol{\alpha}^{\mathrm{T}} + \boldsymbol{\zeta}^{\mathrm{T}}\mathbf{B}) + \mathbf{e}'^{\mathrm{T}}_1 \\ \mathbf{Z}'_2 &= \qquad\qquad\quad \mathbf{e}'_2 \end{aligned} \tag{A.1.2}$$

where $\mathbf{Z}'_1$ is $q \times 1$ and $\mathbf{Z}'_2$ is $(l-1) \times q$. Here again by virtue of orthogonality of $\mathbf{Q}'$, error structure is preserved and $\mathbf{Z}_2$ and $\mathbf{Z}'_2$ consist together of $(n-p+l-2)$ rows of independent $N(\mathbf{O}, \Gamma)$ random vectors. They provide independent information for inference about $\boldsymbol{\Gamma}$.

The first two equations of (A.1.1) may be written as

$$\begin{aligned} \mathbf{Z}_0^{\mathrm{T}} &= n^{\frac{1}{2}}\boldsymbol{\alpha}^{\mathrm{T}} + \mathbf{e}_0^{\mathrm{T}}, \\ \mathbf{Z}_{1i}^{\mathrm{T}} &= (n\lambda_i)^{\frac{1}{2}}\boldsymbol{\alpha}_i^{\mathrm{T}} + \mathbf{e}_i^{\mathrm{T}}, \quad i = 1, \ldots, p, \end{aligned}$$

since $\boldsymbol{\Lambda}$ is diagonal. The least squares (maximum likelihood) estimators of $\boldsymbol{\alpha}$, $\mathbf{A}$ are given by

$$\hat{\boldsymbol{\alpha}} = n^{-\frac{1}{2}}\mathbf{Z}_0, \quad \hat{\boldsymbol{\alpha}}_i = (n\lambda_i)^{-\frac{1}{2}}\mathbf{Z}_i, \quad i = 1, \ldots, p,$$

where $\hat{\boldsymbol{\alpha}} \sim N(\boldsymbol{\alpha}, n^{-1}\boldsymbol{\Gamma})$ and $\hat{\boldsymbol{\alpha}}_i \sim N(\boldsymbol{\alpha}_i, (n\lambda_i)^{-1}\boldsymbol{\Gamma})$ and $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_p$ are independent. Now $\hat{\mathbf{B}} = \mathbf{P}\hat{\mathbf{A}}$ and hence $\hat{\boldsymbol{\alpha}} + \hat{\mathbf{A}}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\boldsymbol{\xi} \sim N(\boldsymbol{\alpha} + \mathbf{B}^{\mathrm{T}}\boldsymbol{\xi},\ (n^{-1} + n^{-1}\Sigma_1^p\, \delta_i^2/\lambda_i))$ where $\delta_i = \mathbf{p}_i^{\mathrm{T}}\boldsymbol{\xi}$. Subtracting from (A.1.2) gives (2.4),

$$(\bar{\mathbf{Y}}' - \hat{\boldsymbol{\alpha}} - \hat{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\xi}) \sim N(\mathbf{O}, \boldsymbol{\Gamma}(l^{-1} + n^{-1} + \boldsymbol{\xi}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\boldsymbol{\xi})).$$

Finally we may utilize the independent estimator of $\boldsymbol{\Gamma}$ to obtain the predictive sampling distribution of $\bar{\mathbf{Y}}'$ consequent on joint sampling of $\mathbf{Y}$ and $\mathbf{Y}'$ for fixed $\mathbf{X}, \mathbf{X}'$. Let $\mathbf{S}$ be the $q \times q$ sum of products matrix obtained from $\mathbf{Z}_2$ and $\mathbf{Z}'_2$ where

$$\mathbf{S} = \mathbf{Z}_2^{\mathrm{T}}\mathbf{Z}_2 + \mathbf{Z}'^{\mathrm{T}}_2\mathbf{Z}'_2.$$

Thus $\mathbf{S}$ is the pooled residual sum of products matrix from the calibration and prediction experiments. This under our normal assumptions has a Wishart distribution with scale matrix $\boldsymbol{\Gamma}$ and degrees of freedom $v+q-1$ where $v = n-p+l-q-1$ which we shall denote as

$$\mathbf{S} \sim W(v+q-1; \boldsymbol{\Gamma}), \quad \text{where } v = n-p+l-q-1. \tag{A.1.3}$$

Adopting the definition of multivariate student distribution given in Section 2.1, an adaptation of Dawid (1981), the following lemma follows from Dickey (1967).

Lemma 1. Let a $q \times q$ symmetric matrix $\mathbf{V}$ have a square root $\mathbf{V}^{\frac{1}{2}}$ so that $\mathbf{V} = \mathbf{V}^{\frac{1}{2}}\mathbf{V}^{\frac{1}{2}\mathrm{T}}$. For convenience of notation we may assume a symmetric square root. If

$$\mathbf{V} \sim W(v+q-1; \mathbf{I}_q),$$

and independently

$$\mathbf{X} \sim N(\mathbf{O}, \mathbf{I}_q),$$

then

$$(\mathbf{V}^{\frac{1}{2}})^{-1}\mathbf{X} \sim T(\mathrm{v}; \mathbf{I}_q).$$

Applying this lemma to (2.4) and (A.1.3), (2.7) follows.

Discussion of Dr Brown's Paper

Mr T. C. Aitchison (University of Glasgow): Since a problem of calibration, namely estimating foetal age from an ultra-sonic scan measurement early in pregnancy, was the first taste of statistics (pure or applied) I ever had, it is really a considerable pleasure to be invited to propose the vote of thanks to the simultaneous presentation of both a sampling theory and a Bayesian approach to the multivariate extension.

Certainly this paper and that of Hunter and Lamboy (1981) in *Technometrics* has reopened and extended discussion of the calibration problem bringing with it further argument over the use of inverse regression in calibration as suggested by Krutchkoff (1967). For the random calibration context this is undoubtedly a sensible strategy but in the controlled calibration case the issue is less clear cut. In fact, even for a Bayesian, there is an open question as to the choice of prior to make since Hunter and Lamboy adopt a prior specification which results in the posterior calibration distribution $\pi(\mathbf{X}'/\mathbf{X}, \mathbf{Y}, \mathbf{Y}')$ being well approximated, under mild conditions, by a distribution where expected value/mode is the maximum likelihood estimator of $\mathbf{X}'$ and any appropriate HPD interval for $\mathbf{X}'$ is exactly that attained by the sampling theory approach using Fieller's Theorem.

Thus we have Bayesian "justifications" for both the standard calibration and inverse regression approaches but in my opinion, however, this is not really the issue—except for the Bayesian in his choice of prior. Personally I can find no sympathy for the approach of Krutchkoff to treating the controlled calibration problem as a regression of X on Y for the production of point or interval estimates of X'. In the case of $l > 1$ (i.e. replicates $Y'_1, \ldots, Y'_l$ at the unknown X') only a very convoluted argument could effect interval estimates for X' from a sampling theory viewpoint and indeed the Bayesian searching for a tractable posterior distribution for X' would require different values of l—see Remark 3 on Theorem 3. Perhaps before the computing age there was a need for tractable and simple posterior distributions but surely the dominant ingredients of any solution, Bayesian or otherwise, should be a reasonable and consistent set of assumptions.

Another issue well worth clarifying is that raised by the author in Section 1.2 when he refers to the conditional sampling approach of Scheffé (1973), and in a very clear paper by Lieberman *et al.* (1967). This really directs attention as to the use to be made of the calibration curve. Are we performing a one-off calibration for a *single* X' value *or* are we going to use this same curve for a succession of X' values (i.e. *multiple* future use of the curve)?

In my experience of calibration in estimating foetal ages from ultra-sonic measurements and in estimating true ages of organic material from radio-carbon dates it is the *multiple* use of the curve that is most common. So perhaps the conditional sampling approach is not really tackling the same problem as this paper since the methods of Scheffé *et al.* are aimed at tackling the multiple use problem. I am certainly very interested to find out what a Bayesian strategy would be to this and indeed whether it would be any different from that outlined for single use in this paper.

On this issue and the other types of problem which go under the umbrella of calibration I would like to direct attention to part of the discussion on the Hunter and Lamboy paper given by Rosenblatt and Spiegelman as an excellent example of the input of applied statisticians to the academic statistical world.

On another practical note in the paint example the author refers to the purpose of the calibration as being to match the "new batch" of paint—presumably to one of a set of standard batches. However, I feel that the presentation of either marginal intervals or marginal posterior densities for the components of $\mathbf{X}'$ may not give an adequate answer to this problem. In this case of $p = 2$ obviously the joint interval and/or posterior density will provide a better answer but for the case of $p \geqslant 3$ (if it exists!) the difficulty of presentation of a multivariate calibration cannot be overlooked. At least for the sampling theory case the use of some form of simultaneous confidence intervals for the components of $\mathbf{X}'$ may go some way to provide a middle ground between the marginal and joint intervals.

I was delighted to see Sections 4 and 5 in this paper concerned with a comparison of various calibration techniques on two interesting data sets. John Anderson in proposing the vote of thanks to a comparison of discrimination techniques on a real data set by Titterington *et al.* (1981) was surely correct when he said that this was a much braver approach than simulation but fraught with difficulty. Any example chosen may have some strange and atypical features which make generalizing conclusions rather a perilous affair. The paint example of this paper does leave something to be desired with respect to model choice (and comparison) since the design appears rather inadequate on the basis of only nine distinct points in the two-dimensional $\mathbf{X}$ space. There may also be problems with potential outliers such as the $\mathbf{Y}_1$ component of observation 33 and the influence of any outliers on multivariate calibration may indeed be rather strange. Further I think that, in this example, it is unfortunate that the author has not included the empirical approach of Lwin and Maritz in his comparison.

Lastly on Section 5 the choice of comparing "confidence" intervals for $\mathbf{X}'_1$ *only* based on the two Y components is perhaps questionable both in terms of whether only one Y component might be more effective for calibration, and in terms of a different assumption of sampling mechanism in comparing, say, methods L′ and LB. There also seems to be a question of an extra parameter in the L′ model over the LB—two simple linear regressions for L′ against one multiple regression on two variables for LB.

This latter problem is only a small glimpse into the Pandora's box of strange properties of interval estimates in the multivariate calibration context particularly when $q > p$. Even if there is "marginal" monotonicity for all pairwise components of $\mathbf{Y}$ with $\mathbf{X}$ there is no guarantee of "nice" interval estimates for $\mathbf{X}'$ if there is, in some sense, contradictory information about $\mathbf{X}'$ in the components of $\mathbf{Y}'$. Great care should certainly be taken in setting up any multivariate calibration "curve" and in the monitoring of future $\mathbf{Y}'$ for their typicality with the $\mathbf{Y}$ observations used in the construction of the calibration curve.

Finally, may I say that one of the pleasant consequences of having to propose this vote of thanks is that I read most of the papers presented to the RSS over the past few years at the one point in time and was particularly struck by two. Namely, those by Box (1980) trying to bring sampling theory and Bayesian methods closer on an applied front and by David Cox (1981) emphasizing the need for a better balance between the practical and the mathematical in the world of statistics. This paper I believe is one good example of how the ideas of Box and Cox can be brought together to, dare I say it, transform a rather theoretical piece of mathematics into a useful technique for the applied statistician.

I did very much enjoy reading and thinking about this paper and have great pleasure in proposing the vote of thanks.

DR I. R. DUNSMORE (University of Sheffield): The masterly presentation today complements the subtle balance of a paper which ranges from delicate dealings with matrix manipulations through to the less mathematical but more practical problems of real data analysis. The problem seems to be fairly topical at the moment. Hunter and Lamboy (1981) initiated a discussion at the annual meeting of the American Statistical Association. There, however, the concentration was on the Bayesian analysis in the univariate case.

I should perhaps say that it was at a rather late stage in the progress of this paper that I was asked to second the vote of thanks. It does however seem appropriate (at least for the Bayesian aspects of the calibration problem) that the proposer and seconder should be named Aitchison and Dunsmore.

Arriving late on the scene and having been away from the calibration problem for some time I found that I was rather like a mountaineer returning to the rock face after an injury. I found myself somewhat out of condition for the rigours of the foothills of the Wishart ranges, the multivariate Student mountains and the central massif of the matrix-T. In scaling these peaks therefore I took the easy route and used the ropes of previous climbers (namely the original referees) in the hope that they had ensured that the route through the technical details was valid.

I will restrict my comments on the theoretical aspects of the paper to the Bayesian approach evolved in Section 3 for controlled calibration. Acknowledgement should, I think, be given to Geisser (1965), who derived a form of $L(\xi)$, namely the basic predictive distributions $\pi(\mathbf{Y}' \mid \mathbf{Y}, \xi, \mathbf{X})$ of $\mathbf{Y}'$. However the form given here in (3.5) in terms of $\bar{\mathbf{Y}}'$ and $\mathbf{S}'$ (through $\mathbf{S}$) is more readily usable.

For the $l > 1$ case in the standard multivariate linear regression model with ξ replaced by $\mathbf{X}'$, problems can occur which are slightly glossed over in Remark 3 on p. 299. The main issue lies with the prior distribution of $\mathbf{X}'$. In the paper the same prior is used for general l as for $l = 1$. Since however that prior was chosen for mathematical convenience, for consistency of criterion we would need to specify a $T(\upsilon - p;\ (1/l + 1/n)\mathbf{X}^{\mathrm{T}}\mathbf{X})$ prior in the general case. This leads to problems similar to those noted in Aitchison and Dunsmore (1975, p. 198) in the $p = q = 1$ case. In particular the prior depends on the feature l of the future experiment. Also, for example, the prior variances of individual components of $\mathbf{X}'$ decrease with l. Use of Brown's law on p. 290, which specifies that ". . . if the procedure is obviously suspect in some circumstances then the solutions may be far from ideal in the other cases where there is no obvious flaw", presumably leads us to question the use of this prior even in the case $l = 1$.

Hunter and Lamboy (1981) proposed a different stand on the prior assumptions, especially with regard to independence of the parameters $\boldsymbol{\alpha}, \mathbf{B}, \xi$. I wonder if Dr Brown has investigated whether their procedure produces much difference from the practical point of view in the multivariate case.

Following Remarks 1 and 2 on pp. 298–299 I would like to assure Dr Brown that Aitchison and Dunsmore are coherent—the corresponding result in Remark 1 for the univariate case was effectively noted in an unpublished report of mine (1970).

We now turn to the two examples given in the paper. They illustrate the theoretical aspects admirably although they leave many questions unanswered from a data analysis point of view, especially with regard to the appropriateness of the model. For example, little mention is made of normality checks for the underlying model. Indeed in the paint example some of the y-variables look palpably un-normal, and the assumption of homogeneity of variance over the different (P, V)-groups merits consideration. My main criticism here concerns the lack of "answers" however. The purpose of calibration is to calibrate, and so provide posterior or predictive distributions, or perhaps interval estimates, or even just point estimates in the extreme. The only results presented for the wheat quality data are in the form of a criterion of prediction accuracy. For that data the results appear to be incredibly good. However the criterion is somewhat questionable, not the least in terminology, since it is quite possible to have, for example, 250 per cent of unexplained variation.

With the paint example an alternative approach would be to view the problem within the framework of classification or discrimination or diagnosis models. This would ignore the underlying scale to both pigmentation and viscosity measurements and restrict the possible predictions, but less would need to be checked in the way of linearity assumptions in the model.

I am sure that we have not heard the last of calibration, nor would Dr Brown claim his paper to be the final word. I only hope that this paper induces as much discussion and further work as the previous airings of the topic have done in the past. Returning finally to the mountaineering analogy I feel that we are like climbers who have set off up the rock face. From the first night bivouac after the initial univariate climb, this paper has enabled us to advance with great care up the long multivariate face. However, the massive amount of modelling assumptions incorporated into the analysis leaves me with worries that perhaps we have only succeeded in reaching an overhang and that the way ahead is hidden from view.

It is with great pleasure that I second the vote of thanks.

Professor A. S. C. EHRENBERG (London Business School): I have three major worries with the applied side of this paper: the use of small data sets in Sections 4 and 5, the use of split-samples there, and the general specification of the calibration problem.

Starting with the least important, I am sorry that Dr Brown has fallen for the modern tendency of making a theoretical paper look more applied by pushing some minuscule sets of data through the formulae. What is worse, he implies that one could judge his methods from samples of less than 10 readings.

Next, and of greater importance, is the mistaken use of split-samples. This is the idea (unfortunately not an uncommon one) that having fitted a model to a random sub-sample of a given set of data, one can then test the model on another sub-sample from the same data. But there can be no difference between two random samples from the same population other than for sampling error. That is what random sampling is all about! One can see this more clearly (if this is needed) by visualizing two random samples of 10 000 each from the same population: Whatever we do to Sample A, we will get the same answer when we also do it to Sample B. This does not test the fitted model. (I note that many statisticians unfortunately think of prediction in the same trivial way as being just about another sample from the same population, especially in the context of regression, as here.)

Thirdly, and most important of all, I believe that Dr Brown's basic specification of the calibration problem on p. 287 is wrong, from a practical point of view. It is all put as if one had to face only a single set of data in X and Y, with the X following some supposedly meaningful statistical frequency distribution.

Instead, what one does in any well-designed calibration study is to arrange things deliberately (as Dr Brown himself says but then ignores) to cover the relevant range of values of X, say at j different levels. In his blood plasma example, he refers to nine such different sets of data, deliberately chosen to lie at or near certain values. (In the controlled case, X would be made to lie "exactly" at certain chosen values. In the uncontrolled case the choice of the word "random" is, I believe, very unfortunate; the data may be irregular and uncontrolled, but not "random").

In general one would therefore take sets of n_j readings (X_j, Y_j) at or near each of the j different levels. (In principle the n_j should first be considered as being large, to separate sample problems from model specification.) The data then reduce down to j sets of mean values $(\bar{X}_j, \bar{Y}_j)$

High: $\bar{X}_1, \bar{Y}_1$

. , .

Medium: . , .

. , .

Low: $\bar{X}_j, \bar{Y}_j$

together with the correlated scatter of the individual X, Y readings in each of the j data sets. (If the scientist is not able to select conditions yielding "High", "Medium" and "Low" values of X and Y he is hardly ready to calibrate Y against X.)

The main calibration problem in practice then concerns the relationship between the j sets of means $\bar{X}$, $\bar{Y}$. This relationship is *symmetrical*, whether X is controlled or not. This is quite different from the asymmetrical problems considerd by Dr Brown.

Professor J. B. Copas (University of Birmingham): In many statistical situations Bayes procedures might be viewed as rather exotic alternatives to standard methods, but in the topic of tonight's paper a random distribution for ξ or $\mathbf{X}'$ is an essential and inescapable ingredient of the problem. Indeed such a distribution makes practical sense since one is not interested in calibrating just one particular value but in developing a formula for calibrating a whole range of values which are likely to occur in the future. It is reasonable to suppose that these future values will arise according to some frequency distribution.

One consequence of randomness in $\mathbf{X}'$ is that, for random calibration data, least squares is no longer admissible if the number of $\mathbf{Y}_i$'s exceeds 2 (Stein, 1960). Least squares tends to overpredict: one is better off shrinking the prediction towards the overall average. Dr Brown mentions the possibility of this in Section 1.1, but does not pursue it in his examples. For instance, a graph of percentage water for cases 17–21 in Table 1 plotted against its LB prediction shows clear evidence of overprediction, and the "percentage of unexplained variation" is reduced slightly if a Stein-type shrinkage is applied (1·50 to 1·47, cf. Table 3). The overprediction seems more than that implied by the Stein formula, which raises the question of whether the 4 Y_i's are defined in advance or selected from a larger set using a stepwise method on the same data.

The importance of the distribution of X' is illustrated in Fig. D1 for the simplest case of bivariate normal data. Although the regression line, and Y', are exactly the same in the two situations shown, the calibrations X' are quite different just because of the different distributions of X'. Thus whilst Dr Brown is right to emphasize the distinction between random and controlled data, I would go further and say that unless we are prepared to say something about the distribution of X' in future cases the problem simply has no solution.

Supposing that Y' given X' is $N(\alpha+\beta X', \sigma^2)$ but that X' has some arbitrary distribution, it is easy to show that

$$E(X' \mid Y') = \frac{Y'-\alpha}{\beta} + \sigma^2 \frac{d}{dy}\{\log p(Y')\}, \tag{1}$$

where $p(Y')$ is the marginal distribution of Y'. If σ^2 is small or $p(Y')$ is flat (very dispersed data), this is just the regression line solved backwards (method L). If $p(Y')$ is log-concave the curve of X' on Y' is flattened, perhaps to the regression of X' on Y' (method LB).

As a special case, suppose that both X (old data) and X' (new data) are normal, so that

$$E(X \mid Y) = \mu_X + d(Y-\mu_Y), \tag{2}$$

and

$$E(X' \mid Y') = \mu_X^* + d^*(Y'-\mu_Y). \tag{3}$$

If the distribution of X' is only slightly displaced from that of X, it is easy to show that

$$\mu_X^* \simeq \mu_X + dAC, \quad d^* \simeq d(1+BC),$$

where

$$A = \text{change in } E(Y), \quad B = \text{proportional change in } \operatorname{var}(Y), \quad C = (1-r^2)/r^2,$$

r being the correlation between X and Y in the calibration data.

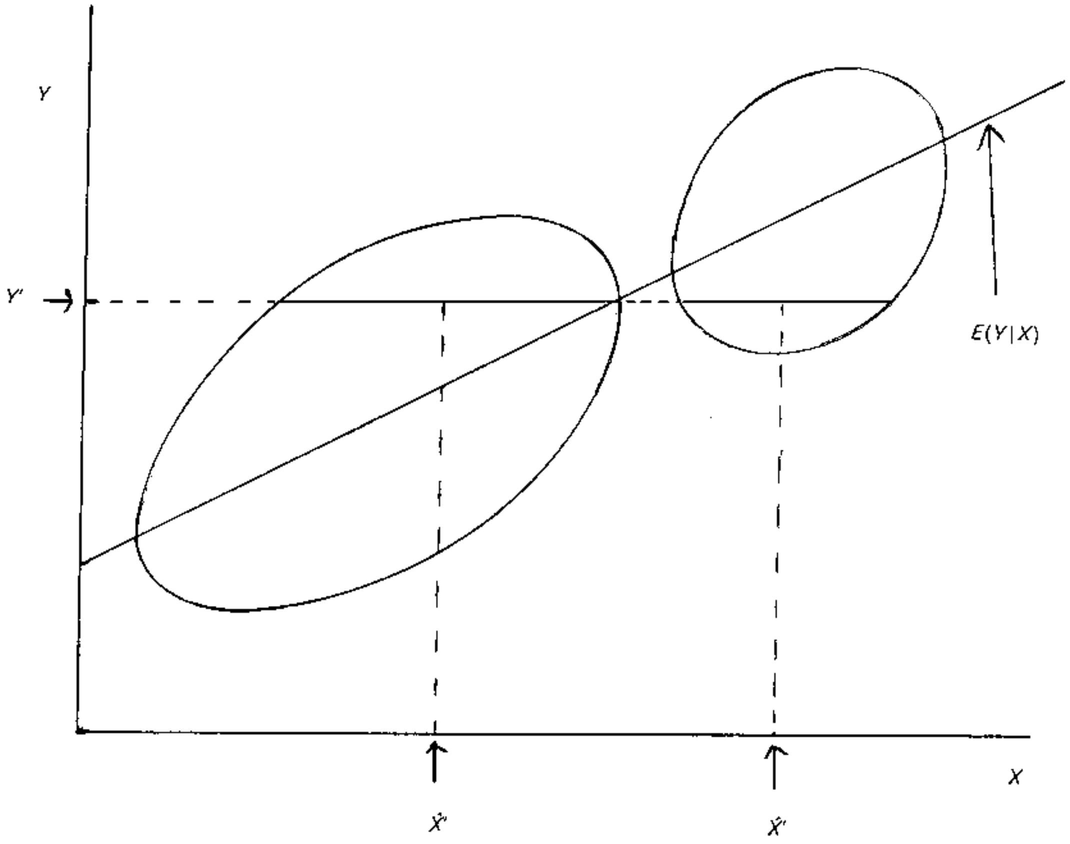

FIG. D1

This suggests a three-stage procedure. Initially, calibrate using (2), i.e. LB. After several future values of Y' have been observed, monitor the changes in the mean and variance of Y' from those of Y and update the slope and intercept giving (3). When a very large sample of Y' has accumulated, use a density estimation method to estimate the logarithmic derivative and hence the optimum calibrator in (1).

Dr H. P. WYNN (Imperial College, London): I want to make a few comments about one of the important issues raised by this paper. This is the distinction between controlled and random experiments. Think of all possible kinds of experiment that could be performed on two variables X and Y. We could just observe X, just observe Y, observe random pairs (X, Y_i), observe Y on controlled X or X on controlled Y. We could also consider two-stage experiments in which X (or Y) is observed and then, using information from this experiment, observe Y (or X). With costs attached to the different experiments the whole problem can be set up as an optimization problem. It is often assumed that the solution is to observe Y on controlled X or some other controlled experiment. However, simple examples show that a typical solution involves mixtures of controlled and random experiments.

The distinction between the two kinds of experiment has a long history in the philosophy of science. John Stuart Mill calls random experiments "spontaneous" and controlled experiments "artificial". A discussion of some of this philosophical background with a modern example will appear shortly.

Mr P. J. SCOTT (Imperial College, London): I have just a couple of comments to make: Firstly, in the wheat data analysis, the first sixteen observations were used for calibration and the remaining five for prediction purposes. I would have liked to have seen these predicted values to see how the unexplained percentages of variation given in Table 3 (p. 301) broke up; for instance, in the empirical predictions, were the large values in the unexplained percentages of variation due to a large difference in one of the predictions, or in a general under- or over-prediction of these values? Secondly, in the derivation of the empirical method one has to estimate the marginal density of the X's. In this derivation this marginal density is estimated by putting delta functions at the observed values of the X's. Might not better results be possible with some other method of density estimation, for instance one based on a kernel method?

Dr T. Fearn (Flour Milling and Baking Research Association, Rickmansworth): As the supplier of the wheat quality data I am relieved at the author's conclusion that componentwise regression of X on Y (to use the notation of the paper) is as good a method of analysis as any. This is the simplest and most natural way of tackling the problem and is universally adopted. With such a precise relationship it does not seem to matter what approach is adopted; I am only surprised that Dr Brown has managed to find a method (E) which fails.

Because the data, having once appeared in the literature, are likely to be reanalysed I would like to record the exact nature of their "random" status. The samples were not explicitly chosen on the basis of their protein or moisture contents but they would have been chosen to include a good range of wheat varieties. They would not therefore be in any sense a random sample from the throughput of the laboratory although there is some "randomness" in both variables. The flat marginal distribution of X noted in Section 4.3 is in part a consequence of this selection. Given the way the data have been used, with a randomly selected prediction set, these comments do not affect Dr Brown's analysis or conclusions.

It is a common, and desirable, practice to select the calibration samples to flatten the marginal X distribution when calibrating these instruments. Any bias introduced by such selection combined with the regression of X on Y is more than compensated for by the increase in the precision of estimation of the regression.

If the selection is modelled probabilistically in the following rather idealized way it is possible to quantify its effect. Suppose (x, y) have a bivariate normal distribution but that samples are selected on the basis of their x value with probability proportional to $1/p(x)$. If we ignore problems with norms this gives a uniform distribution for the selected x. The p.d.f. of x conditional on y for the selected samples is proportional to $p(x \mid y)/p(x)$ which in turn is proportional to $p(y \mid x)$. It follows from the form of this density (as a function of x) that the regression of selected x on y is $1/\beta$ where β is the regression of y on x. Thus the effect of flattening the marginal x distribution but still regressing x on y is roughly the same as that of doing the regression the other way round.

The following contributions were received in writing after the meeting:

Professor G. A. Barnard (University of Essex): I was glad to see Minder and Whitney's marginal likelihood approach referred to, although I was sorry it was not further discussed. It is one of the most intuitively appealing solutions. In the case where the estimated slope of the regression fails to differ significantly from zero, it tells us that large positive and negative values are quite as plausible as moderate values, and that no single point estimate is at all reasonable in this case.

Minder and Whitney had difficulty in getting their paper past one referee, who noted the possibility of not having a single point estimate and went on to say "it is the statistician's job to come up with an estimate, no matter what the data are"!

Given data, it is the statistician's primary job to express what the data say—and if this amounts to nothing, then he should say so.

Dr Brown seemed to come close to demanding that the data speak in prescribed forms when he spoke of the unpleasantness of cases where no simple confidence sets could be given for a ratio, or of the difficulties in his Section 2.2 (p. 293) when $q > p$, or when two probabilities are needed to give proper expression to the uncertainty of the conclusions. The fact that we may have been drilled in the past into expressing all uncertainties as single P values gives no excuse for continuing the practice.

In relation to the problem when $q > p$, we should remember that an assertion with 95 per cent confidence that C is true means only that, unless an event of probability 5 per cent has occurred, then C must be true. If the data make it clear that an event of probability 5 per cent really *has* occurred, there is no reason to suppose that C is true.

Professor J. M. Bernardo (Universidad de Valencia): I have truly enjoyed this important, stimulating paper. It provides a unified review of a large amount of previous work on the common problem of calibration and roughly shows, once more, that sampling theory recipes are, at best, a limiting case of a sensible Bayesian analysis. The attention to comparative studies, however, lead Dr Brown to concentrate on the estimates of $\mathbf{X}$ given $\mathbf{Y}$ and the data rather than discussing the entire predictive distribution $p(\mathbf{X} \mid \mathbf{Y}, \text{data})$ which, as he mentions, is the natural answer to the problem posed.

I certainly agree with the main conclusion in the random variables case; one should regress $\mathbf{X}$ on $\mathbf{Y}$ to predict $\mathbf{X}$ in the future. I wonder however to what extent the same type of analysis may be usefully extended to the non-linear in $\mathbf{\Theta}$ case, to include, for instance, logistic models. At a more specific level, I do not find appealing (although it is mathematically ingenious) the selection of an *ad hoc* prior for $\mathbf{X}'$. I believe one should try, either to describe personal beliefs, or to use a general systematic methodology to specify reference priors (Bernardo, 1979).

When the variables are controlled, they are usually set to a finite number of values. If, as Dr Brown rightly suggests, it is better to predict $\mathbf{X}$ one variable at a time, multivariate calibration in the controlled case becomes mathematically identical to discrimination (or medical diagnosis, or pattern recognition as the author reminds us).

A very welcome feature of this paper is the inclusion of sets of rough data which are then analysed according to the methods described. This allows the reader to compare the results with other approaches he may like to try.

I have analyzed the paint finish data of Table 2 using a Bayesian normal discrimination procedure (Bernardo, 1978), similar to that proposed by Aitchison and Dunsmore (1975, Chapter 12) but with a different reference prior specification. The results are obviously probability distributions over the possible values of $\mathbf{X}$, the goodness of which may be assessed with a proper scoring rule (Savage, 1971). If one insists in having a point estimate, on may use the mean or the mode of such distribution.

The predictive distributions of the nine observations randomly chosen by the author to compare the methods he analyses, using the other 27 observations as data, are given in Table D1.

Table D1

Observations	P	V	$\Pr(P=0)$	$\Pr(P=1)$	$\Pr(P=2)$	$E(P)$	$\Pr(V=0)$	$\Pr(V=1)$	$\Pr(V=2)$	$E(V)$
2	0	0	0·166	0·834	0·000	0·834	0·008	0·992	0·000	0·992
5	0	1	1·000	0·000	0·000	0·000	0·794	0·000	0·206	0·412
11	0	2	0·991	0·001	0·007	0·150	0·004	0·737	0·259	1·255
16	1	0	0·013	0·987	0·000	0·987	0·481	0·518	0·001	0·538
18	1	1	0·692	0·308	0·000	0·308	0·908	0·036	0·056	0·148
22	1	2	0·001	0·999	0·000	0·999	0·000	0·000	1·000	2·000
28	2	0	0·014	0·000	0·986	1·972	0·948	0·049	0·003	0·055
30	2	1	0·003	0·004	0·992	1·988	0·004	0·988	0·008	1·004
35	2	2	0·000	0·000	1·000	2·000	0·000	0·000	1·000	2·000

The results where obtained with a computer program, written to be used routinely in medical diagnosis, by a student of mine, José D. Bermudez. The unexplained percentages of variation, using as estimates the distribution means are found to be 8·0 for pigmentation and 19·4 for viscosity, clearly better than those obtained by the methods considered in the paper (cf. Table 5).

Yet an alternative analysis of the same data could have been performed using logistic regression; I do not have at hand, however, a computer program to do it.

Professor D. R. Cox (Imperial College, London): This seems to me a very impressive paper. A crucial aspect in applications is often the stability of the calibration curve in time: how often is recalibration needed, and how can checks of stability be incorporated into the routine use of the procedure? Has Dr Brown any comments?

As in other aspects of statistics, sweeping statements about what happens in practice deserve scepticism: they often do less than justice to the rich variety of the real world! Nevertheless, I am unconvinced about the wide practical appropriateness of the random model, especially that with the same distribution for future values as used in the calibration. This is partly because often sensible calibration will be done over a rather wider range likely to arise in future use. This might seem to lend support to the "regress x on y" approach, but, as Mr Aitchison stressed, a more explicit consideration of the way the results are to be used seems necessary. Here are three possibilities:

(a) is the true x outside some tolerance limits? Is a further test on the same individual needed?

(b) a number of individuals are measured and then some simple comparisons made, or more generally the resulting estimates put through some further statistical analysis, only qualitative contrasts being important;

(c) a number of individuals are measured, possible on different pieces of apparatus or even in different laboratories, and then contrasts of individuals examined, absolute values now being important.

In Case (b) the precise calibration formula will often be unimportant. In Case (c) the "classical" approach seems indicated, in the absence of very specific prior knowledge. In Case (a) the form of the calibration curve near the critical values is clearly of central importance.

Professor M. De Groot (Carnegie–Mellon University, Pittsburgh): I very much enjoyed reading this paper by P. J. Brown. Part of this enjoyment, I must confess, was due to the somewhat unwholesome self-satisfied pleasure I derived because Bill Davis and I had not long ago worked on similar problems (Davis and DeGroot, 1982). The similarity between Dr Brown's outlook and ours is substantial enough to be reassuring to us (it is always reassuring to find other good statisticians are thinking along the same lines that we are), while the overlap between his results and ours is small enough not to be embarrassing.

Davis and I, like Dr Brown, also considered calibration problems in which $\mathbf{X}_i$ and $\mathbf{Y}_i$ are vectors of arbitrary dimensions, also assumed normality, and also distinguished carefully between controlled and random calibration. However, our work is more limited than Dr Brown's in that we considered only a single observation $\mathbf{Y}'$ that was to be calibrated at an unknown $\mathbf{X}'$ and we restricted ourselves to the Bayesian approach. On the other hand, our work is more general than Dr Brown's in that we considered a 2 × 2 table of models in which $\mathbf{X}_1, \ldots, \mathbf{X}_n$ might be either all controlled or all random and, separately, $\mathbf{X}'$ might also be either controlled or random. Furthermore, we considered problems of prediction as well as problems of calibration. Finally, and this was our original motivation, we considered problems in which the experimenter is not sure which one of several multiple linear regression models is correct, i.e. in which he is not sure which components of the vector $\mathbf{X}_i$ actually appear in the regression function with non-zero coefficients.

Here we have been somewhat more formal than Dr Brown. When he briefly discusses such problems in Section 3.1, he presents a few somewhat arbitrary procedures for comparing models, such as comparing the maxima of the densities obtained under the different models or comparing the posterior probabilities obtained by integrating these densities over conveniently chosen regions. Davis and I formally develop posterior probabilities of the different models and perform prediction and calibration by using appropriate weighted averages based on these probabilities.

The examples in Sections 4 and 5 of Dr Brown's paper provide a fascinating and useful finish. Where Davis and I contented ourselves with the calculation of various posterior and predictive distributions, he has tried out his methods, looked at some numbers, and reached some tentative conclusions which appear to lend support to the Bayesian approach. I congratulate him on an interesting and stimulating paper.

Professor D. V. Lindley (Somerset): The Bayesian paradigm applied to calibration is straightforward. What is the random quantity of interest? Here it is $\mathbf{X}'$. What is known? Here the data are $\mathbf{Y}', \mathbf{X}, \mathbf{Y}$. It is therefore necessary to calculate the probability of the random quantity given the data: here $p(\mathbf{X}' \mid \mathbf{Y}', \mathbf{X}, \mathbf{Y})$. How is the calculation to proceed? By use only of the calculus of probabilities. Here, by Bayes' theorem,

$$p(\mathbf{X}' \mid \mathbf{Y}', \mathbf{X}, \mathbf{Y}) \propto p(\mathbf{Y}' \mid \mathbf{X}', \mathbf{X}, \mathbf{Y})\, p(\mathbf{X}' \mid \mathbf{X}, \mathbf{Y})$$

and, if the experiment is controlled, $(\mathbf{X}, \mathbf{Y})$ will give no information about $\mathbf{X}'$, so the last factor is simply $p(\mathbf{X}')$. The other factor on the right-hand side is $p(\mathbf{Y}', \mathbf{Y} \mid \mathbf{X}', \mathbf{X})/p(\mathbf{Y} \mid \mathbf{X}, \mathbf{X}')$, and $\mathbf{X}'$ may be omitted from the denominator so that the latter may be absorbed into the constant of proportionality. Finally

$$p(\mathbf{X}' \mid \mathbf{Y}', \mathbf{X}, \mathbf{Y}) \propto p(\mathbf{Y}', \mathbf{Y} \mid \mathbf{X}', \mathbf{X})\, p(\mathbf{X}').$$

(This is essentially Theorem 2.) With the usual judgement of exchangeability (and a thought about its applicability in the practical problem) the first probability on the right-hand side is

$$\int \prod_{i=1}^{l} p(\mathbf{Y}'_i \mid \boldsymbol{\theta}, \mathbf{X}'_i) \prod_{j=1}^{n} p(\mathbf{Y}_j \mid \boldsymbol{\theta}, \mathbf{X}_j)\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

A typical term in the products may be factored into

$$p(y_1 \mid \theta_1, \mathbf{X})\, p(y_2 \mid \theta_2, y_1, \mathbf{X}) \ldots p(y_q \mid \theta_q, y_1, \ldots, y_{q-1}, \mathbf{X})$$

with $\mathbf{Y} = (y_1, \ldots, y_q)$ and the analysis is a series of univariate regressions which I find easier to understand. In a similar view, is it necessary to consider $l > 1$? If $l = 2$, the second calibration proceeds, conditional on $\mathbf{X}'_1$; so that n is effectively increased by one.

I am sorry to see the author pursue incoherent ideas; though it was pleasant to see them do badly. In particular the claims for S-ancillarity are unfounded. Suppose X and Y have the same, unknown variance: then S-ancillarity does not obtain. If they have distinct variances, it does. But suppose the variances are known to be about equal: of what use is S-ancillarity then? Ancillarity is a will-o'-the-wisp chased by those who forgo logic in their inference.

Is it necessary to consider distinct models (Section 3.1)? Savage advocated models "as big as an elephant". Aside from technical complexity and additional computing time, they cause no logical problems. Again, ill-specified problems cause no real difficulties within the Bayesian framework and can often prove useful: see, for example, problems in econometrics.

Dr R. SUNDBERG (Royal Institute of Technology, Stockholm): I want to thank the author for his stimulating paper on statistical methods in multivariate calibration. There is no agreement among statisticians about the univariate methodology, and the author offers still more suggestions from which to choose in the multivariate case. My attention was caught by the estimation method (L′), "one x-variable at a time", used in the two examples. It first appeared paradoxical to me that (L′) worked about as well as (L), since (L′) seemed to neglect influential variables. I will discuss (L′) and (L) as defined by (2.16), which may be a reasonable estimator although the statements made about (2.16) are not quite correct when $q > p$ (apparently the author has neglected in (2.8) the dependence on ξ through $\sigma^2(\xi)$).

A closer look at (L′) reveals that the choice of $\mathbf{S}$ as the estimator of $\mathbf{\Gamma}$ in the mutilated model is crucial. If we used the unbiased estimator $\hat{\mathbf{\Gamma}}$ of the full model we would get a very much worse estimator, to be called $\tilde{\mathbf{X}}$. Now, $\hat{\mathbf{X}}^{(\mathrm{L}')}$ actually is a weighted average of the full model estimator $\hat{\mathbf{X}}^{(\mathrm{L})}$ and $\tilde{\mathbf{X}}$. There is less randomness in $\hat{\mathbf{X}}^{(\mathrm{L}')}$ than in $\hat{\mathbf{X}}^{(\mathrm{L})}$, but in general a systematic error from $\tilde{\mathbf{X}}$. In the wheat quality data, the weight of $\tilde{\mathbf{X}}$ in $\hat{\mathbf{X}}^{(\mathrm{L}')}$ is only about 1 per cent, hence the observed equality of (L) and (L′) with respect to unexplained variation in Table 3. In the paint finish data, the weights are about 3 : 1, so there (L) and (L′) yield really different estimators. That Table 5 favours (L′) might be a coincidence, because no linear model fits these data, so neither (L) nor (L′) nor any other of the author's models works satisfactorily. How much unexplained variation should be tolerated in this example is indicated by using an additive quadratic model, which fits data well. Calculations gave a result of 4 per cent for P and 6 per cent for V, compare Table 5.

I suspect the paint finish data also illustrate the common (but often overlooked) type of situation where the $\mathbf{\Gamma}$ of the calibration experiment and that of the subsequent application of the instruments are not the same. The response variables used, Y_1 and Y_4, are seen to be substantially correlated within experimental points. A plausible reason for this may be a lack of precision in adjusting one (or both) of the x-variables to the pre-specified values. This source of randomness will influence the response variables only during the calibration. In univariate calibration we need only consider a proportionality factor, here we get two different covariance structures.

Finally, a few words about the random case, which I find somewhat loosely treated in the paper. A crucial assumption must be that $\mathbf{X}'$ derives from the same population as $\mathbf{X}$. It is satisfied in both of the author's data examples, by his construction, so it is not surprising that the (LB) method turns out at least as well as (L). But how often can we trust it? In the apple-sorting example cited by the author I would not trust it unless the calibration was made on the lot to be sorted, since lot = population. In most practical regression situations, I guess, each $\mathbf{X}'$ to be estimated is a unique individual in some sense or another. To consider the case when a substantial difference between methods might appear, let us say that a particular $\mathbf{Y}'$ falls in the tail of the calibration distribution of $\mathbf{Y}$, thus indicating that its $\mathbf{X}'$ might not derive from the calibration distribution of $\mathbf{X}$. Should the author then use regression of $\mathbf{X}$ on $\mathbf{Y}$, of $\mathbf{Y}$ on $\mathbf{X}$, or something else?

Professor A. ZELLNER (Alexander Research Foundation, University of Chicago): This very interesting paper provides additional support for the conclusion of Hoadley (1970, p. 369) that "... the main point in the paper is that the Bayesian approach has led to valuable insight and understanding". Brown has ingeniously extended Hoadley's analysis to the multivariate calibration problem.

On the sampling theory approach to the normal univariate calibration problem, a solution given in Section 1.2 is: $\hat{X}'-\bar{x} = (Y'-\bar{y})/\hat{\beta}$, where $\hat{\beta} = S_{xy}/S_{xx}$. For given Y', $\hat{X}'-\bar{x}$ is the ratio of two independent normal random variables and hence can have a markedly bimodal distribution. Also its moments will not in general exist and thus, as Brown notes, the MSE of $\hat{X}'-\bar{x}$ is infinite. If, however, the distribution of $\hat{X}'-\bar{x}$ is considered conditional upon the outcome of a t-test that rejects the hypothesis $\beta = 0$, that is $|\hat{\beta}|/s_{\hat{\beta}} > c > 0$, where c is a critical value for the t-test, then the moments of $\hat{X}'-\bar{x}$ exist and MSE is finite. However, this fact does not provide very strong support for general use of such an estimator. The flexible Bayesian results provided by Hoadley and Brown seem to me to be much more elegant and useful.

In Brown's analysis of the multivariate calibration problem, I would like to see some more analysis directed at characterizing the properties of the likelihood function in (3.6) as Hoadley (1970, p. 364) did in the univariate case where he found that his likelihood function can frequently possess two local maxima. Understanding the conditions giving rise to bimodal likelihood functions seems important. Brown's likelihood function (3.6),

$$(1/l+1/n+\xi^{\mathrm{T}}\mathbf{G}\xi)^{\nu/2}/[1+R+(\xi-\hat{\xi})^{\mathrm{T}}\mathbf{H}(\xi-\hat{\xi})]^{(\nu+q)/2},$$

where $\xi \equiv \mathbf{X}$ and $\hat{\xi} \equiv \hat{\mathbf{X}}$, apparently can be bimodal as well. Further, the fact that use of a special multivariate Student-t prior for ξ (or $\mathbf{X}'$) in Theorem 3 results in a unimodal posterior p.d.f. for ξ (or $\mathbf{X}'$) means that use of this prior probably has a substantial impact on the shape of the likelihood function. I believe that it would be useful to investigate the sensitivity of the form of Brown's posterior distribution to changes in the form of his prior distribution.

Overall, I congratulate Brown for his significant contribution to the analysis of the multivariate calibration problem.

The AUTHOR replied later, in writing, as follows.

I am grateful for the detailed and thoughtful contributions of the discussants. It is illuminating to watch the different methods others use to grapple with common problems.

Since I have been chastised by Professor Lindley for not being Bayesian enough I will first take issue with Mr Aitchison and Professor Barnard in their concern for error rates. Both the Scheffé method of assigning two probabilities and the approach of Lieberman *et al.* to multiple simultaneous confidence intervals emphasize the great divide between those who think statistical inference is a matter of error rates as against those who opt for notions of "support" given the actual observed data. In answer to Mr Aitchison, I take the Bayesian strategy for the multiple use of the calibration curve to be initially as in Section 3. A credibility interval for an unknown $\mathbf{X}'$ from observed $\mathbf{Y}'$ should not depend on other as yet unknown and unobserved $(\mathbf{X}', \mathbf{Y}')$. However as soon as at least two $\mathbf{Y}'$ are observed, corresponding to different but unknown $\mathbf{X}'$, there is available information which is slightly different from that considered in Section 3 and further updating by Bayes' theorem is possible. A set of observed $\mathbf{Y}'$ at different unknown $\mathbf{X}'$ provide information on the distribution of future $\mathbf{X}'$ albeit entangled with the calibration model. With assumed and tested parametric assumptions one can extend the interesting mean prediction formulae given by Professor Copas to give the full posterior predictive distribution and credibility intervals. Whether one continuously updates as $\mathbf{Y}'$ are observed whilst at the same time checking for abnormal $\mathbf{Y}'$, or whether this updating is done in stages, will perhaps depend on the practical inconvenience of continuous updating.

Many discussants refer to the distribution of future $\mathbf{X}'$. Professor Copas is right to emphasize the importance of this. I would argue that the degree of validity of approaches which eschew consideration of this depends on the validity of the implicit assumption of $\pi(\mathbf{X}')$, as judged by the $\pi(\mathbf{X}')$ which would provide a mimicking Bayes procedure, together with the proximity to unity of the canonical correlations. I have tried to stress the importance of $\pi(\mathbf{X}')$ in controlled calibration and further, I should like to dispel the dangerous notion that it be chosen for mathematical convenience.

I see the Krutchkoff solution as only valid with $l = 1$ and when the designed $\mathbf{X}$ and the future $\mathbf{X}'$ are exchangeable. By the same token I see the classical solution as really only valid if the distribution of $\mathbf{X}'$ is thought to be rather flat and wider than the designed distribution of $\mathbf{X}$. They are at two extremes. The methods of Section 3 allow one to work in the middle ground between these extremes. Here I agree with Professor Bernardo. Only if forced to choose between the two extremes alone would I generally envisage that exchangeability with $\mathbf{X}$ rather than a very flat $\mathbf{X}'$ distribution most accords with the truth. Dr Fearn's interesting design construction bears this out, as do some of Professor Cox's remarks. If I really thought the $\mathbf{X}'$ distribution to be so flat relative to the designed $\mathbf{X}$ I would want more calibrating data as the calibration curve would be rather suspect at the fringes of the **X**-range. Indeed, in reply to Dr Sundberg, if a particular **Y**′ fell well in the tail of the calibration distribution of **Y** and I had assumed **X** and **X**′ exchangeable, then after proper examination of the pedigree of the observation, I would probably still go ahead as if nothing were amiss, but I would note possible extra uncertainty due to both the weakness of the calibration curve and the questionable **X**′ distribution. It would be desirable to obtain both further designed calibratory data and further values of **Y**′.

Just as the Krutchkoff solution for $l = 1$ corresponds to a Bayes procedure with a particular prior, so the Hunter and Lamboy prior mimics the classical solution. They recommend a particular prior, in the simple linear regression case, which is vague and flat with respect to the X' distribution. They are essentially just working with the integrated likelihood (but see Hill's discussion of that paper for the exact special case of Hoadley's work which is being adopted). One undesirable feature of their prior on $E(Y')$ is that it does not naturally generalize to several future Y' at different unknown X'.

Whatever $\pi(\mathbf{X}')$ is assumed, I would re-emphasize, followed Remark 3 of Theorem 3, that the prior **X**′ should not depend on l. I agree with Mr Aitchison that dependence on l is unreasonable and here, as Dr Dunsmore notes, mathematical convenience is a very bad guide. Indeed, I am surprised that Dr Dunsmore should have engineered such an obvious pitfall. The exchangeability assumptions of **X** and **X**′ may be questioned but not by mathematical convenience. In this context of $l > 1$ I cannot understand Professor Lindley's remark concerning the equivalence to $n+1$. From his earlier formula I assume he has misunderstood the model: the l observed values of $\mathbf{Y}'_i$, $i = 1, \ldots, l$, all have the same true **X**′. Exchangeability of all $n+l$ **X**-values breaks down. We are not envisaging the situation of multiple future use, mentioned earlier.

Dr Sundberg is right in pointing out that the exchangeability assumption, that **X**′ derives from the same distribution of **X**, is satisfied in both data examples, by the construction so disliked by Professor Ehrenberg. This is indeed more favourable to (LB) than to (L). However, to suggest that two random samples from the same population are identical, as does Professor Ehrenberg, is absurd. By this token one could predict perfectly irrespective of the method. I go along with Dr Sundberg that the model assumptions of the calibrating experiment may need modifying in the prediction experiment in some applications. If this just manifests itself in a different **Γ** then $l > 1$ **Y**′ values at a single **X**′ will provide direct information on this. In the same vein, if a calibration curve estimated in one location is to be used in other locations, then calibratory data will be needed to check the validity of the relationship in each location and the model extended if such uniformity is not present. Changes in calibrating relationship over time will also require regular checking. In response to Professor Cox, I would imagine that the frequency of such recalibrations would depend on the speed of such changes. If time correlated errors feature from observation to observation then one might apply the designs of Daniel (1975) to estimate the parameters that effect such time changes.

Dr Fearn reveals a sensible design in which for estimation purposes the more extreme **X** values are proportionately over sampled. Of course this does not effect the validity of our split-sample analysis of the wheat data. In application, however, one would incline to use of a $\pi(\mathbf{X}')$ which is equal to the $p(\mathbf{X})$ given by Dr Fearn and not the marginal distribution of **X** in the calibration experiment. If one knows that a future specimen is the ith variety then one would use $p(\mathbf{X})$, the **X**-distribution for that variety. However in this example, because of the accuracy of the calibration curve, such niceties may not offer much improvement.

Would Professor Lindley kindly provide Dr Dunsmore, worried as he is by the amount of modelling assumptions and fearing that he is stranded on a mountain overhang, with some breathing apparatus to enable him to compete with his goat-like climbing agility. For my part I am happy with further model elaboration along some of the lines already indicated but I find the exhortation to build "as big as an elephant" unrealistic. Such elephants are easy to mistake for kangaroos.

Concerning such elaboration, I had hoped to find a formal solution to Section 3.1 in the paper to which Professor DeGroot refers. Unfortunately this does not seem to be the case. I addressed (a) nested models in **X**, with some retained dependence on X, and (b) models where if we partition **Y** into $(\mathbf{Y}_1, \mathbf{Y}_2)$ then the conditional distribution of $\mathbf{Y}_1$, given $\mathbf{Y}_2$ does not depend on **X**. It was heartening, however, to see corroboration of some of our Bayesian results in this same paper with Davis.

I agree with Professor Zellner that more work needs to be done to uncover the shape of the integrated likelihood (3.6) and the nature of its influence on $\pi(\mathbf{X}')$.

I was fascinated to read the opposite opinions expressed by Dr Sundberg and Professor Bernardo on the issue of marginal calibration. I stick by my original assessment but that assessment looks increasingly superficial to me and more work needs to be done. It is an important issue. The main idea was as follows.

Suppose in the wheat example we are interested in predicting protein X_2 from the four infrared measurements $Y_1, \ldots, Y_4$. We can observe jointly the values of $Y_1, \ldots, Y_4, X_2$ in the calibration experiment. In future we will only have $Y'_2 \ldots Y'_4$. Can it ever be useful to bring an irrelevant variable X_1 into consideration given that we will not know X'_1? I suspect not and Professor Bernardo agrees with me as do the practical results of Sections 4 and 5.

Dr Wynn's suggested preference for a design which chooses to observe a mixture of **Y** given **X** addresses a slightly different but related problem. The question which arises in my mind on reading his paper is under what circumstances observing **Y** alone would have been better than this mixture.

Both Dr Dunsmore and Professor Bernardo suggest in the paint example the use of discrimination, ignoring the scales of both pigmentation and viscosity. If we accept the knowledge that the true X' are also 0, 1, 2 as in our constructed prediction set then we can anticipate doing better than in the presented method. Indeed with such knowledge one could use the predicted X' of Section 5 to allocate to the nearest of the three values. This considerably reduces the unexplained variation. However such prior knowledge could not be assumed in application of the technology.

I agree with Professor Lindley that ancillarity is too fortuitous a property to become a basis for inference. Very useful, though, when it does hold, even to a Bayesian for whom it labels a type of likelihood factorization.

Dr Sundberg is right that statements made concerning (2.16) are misleading when $q>p$. In fact $\mathbf{X}'$ is the minimum of the quadratic form (2.9). This applies even when a confidence region does not exist as is sometimes the case when $q>p$.

In the examples, Mr Scott would have liked to have seen the actual predicted values. As it happens nothing is revealed by examining these. Looking at the individual results for predicting protein in the wheat example, within the limitations imposed by the paucity of five predicted values, there were no evident systematic biases and (LB) was uniformly closer than (E) and (E') over the five predictions. I did not present graphs because they did not seem to indicate anything interesting. As a general point though I agree that such exhibits are often illuminating.

Mr Scott is right that the estimate of the marginal distribution of **X** in (E) and (E') is cavalier. I doubt though whether a smoother function would improve matters. I still surmise that the neglect of the uncertainty in the parameter estimates in the conditional distribution of **Y** given **X** is the crucial element in the bad performance of these empirical methods.

Mr Aitchison alludes to the strange properties of the sampling theory confidence intervals when $q>p$ given in Section 2.2 under Theorem 1 (ii). Professor Barnard's closing remark serves to emphasize the questionable usefulness of some confidence statements. The deficiency here is one of reference set, well illustrated for example by Pierce (1973). A simple example may serve to clarify the issue. Suppose all the regression parameters are known, as follows when $n \to \infty$ in Section 2. Furthermore if $q=2$ and $p=1$, $\mathbf{\Gamma}=\mathbf{I}$ and

$$Y'_1 = x' + \varepsilon'_1,$$

$$Y'_2 = x' + \varepsilon'_2,$$

then the predictive distribution of $(Y'_1-x')^2+(Y'_2-x')^2$, under normal theory, is chi-squared on two degrees of freedom. The proposed sampling theory approach gives a 95 per cent confidence interval for x' as all x' such that

$$(Y'_1-x')^2+(Y'_2-x')^2 > 6{\cdot}0,$$

where 6·0 is the upper 5 per cent point of a chi-squared on two degrees of freedom. The confidence interval is $-\sqrt{2}<x'<\sqrt{2}$ when $\mathbf{Y}'=(1,-1)^{\mathrm{T}}$ and vanishes to the point $x'=0$ when $\mathbf{Y}'=(\sqrt{3},-\sqrt{3})^{\mathrm{T}}$.

The likelihood and Bayes approaches do not behave in this way. Even though they formally make use of the same predictive distribution of $\mathbf{Y}'$ given x', they use it in a much different way as shown by (3.5). They behave naturally under the assumption that the model is true. It is a separate but important issue to check this. It is worrying that the sampling procedure behaves far less naturally. All would be well if one conditioned above on the ancillary statistic $Y'_1-Y'_2$ but such ancillaries do not exist when one is dealing with the general regression model. Dr J. Wood of CSIRO has started on a correction to this deficiency (personal communication). I would simply recommend the use of the Bayes results of Section 3.

Finally let me once again thank the discussants for the important points they have raised.

References in the Discussion

Bernardo, J. M. (1978). Metodos Bayesianos y diagnosis clinica. *Estadistica Espanola*, **79**, 39–56.

——(1979). Reference posterior distributions for Bayesian inference (with Discussion). *J. R. Statist. Soc.* B, **41**, 113–147.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Statist. Soc.* A, **143**, 383–430.

Cox, D. R. (1981). Theory and general principle in statistics: the Address of the President (with Proceedings). *J. R. Statist. Soc.* A, **144**, 289–297.

Daniel, C. (1975). Calibration designs for machines with carry over and drift. *J. Qual. Tech.*, 7, 103–108.

Davis, W. W. and DeGroot, M. H. (1982). A new look at Bayesian prediction and calibration. In *Statistical Decision Theory and Related Topics* III, Vol. I (S. S. Gupta and J. O. Berger, eds), pp. 271–289. New York: Academic Press.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. *Ann. Math. Statist.*, **36**, 150–159.

Hunter, W. G. and Lamboy, W. F. (1981). A Bayesian analysis of the linear calibration problem. *Technometrics*, **23**, 323–328.

Lieberman, G. J., Miller, R. G. and Hamilton, M. A. (1967). Unlimited simultaneous discrimination intervals in regression. *Biometrika*, **54**, 133–145.

Pierce, D. A. (1973). On some difficulties in a frequency theory of inference. *Ann. Statist.*, **1**, 241–250.

Rosenblatt, J. R. and Spiegelman, C. H. (1981). Discussion on paper by Hunter and Lamboy. *Technometrics*, **23**, 329–333.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Ass.*, **66**, 783–801.

Stein, C. (1960). Multiple regression. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (I. Olkin, ed.), pp. 424–443. Stanford, Ca.: Stanford University Press.

Wynn, H. P. (1982). Controlled versus random experimentation. *The Statistician* (in press).