Biometrika (1989), 76, 2, *pp.* 349-61
Printed in Great Britain

Prediction diagnostics and updating in multivariate calibration

BY PHILIP J. BROWN
Department of Statistics & Computational Mathematics, University of Liverpool,
Liverpool L69 3BX, U.K.

AND ROLF SUNDBERG
Institute of Actuarial Mathematics and Mathematical Statistics, University of Stockholm,
S-11385 Stockholm, Sweden

SUMMARY

In multivariate calibration the relationship between a q-variate response vector Y and p explanatory variables X are estimated from training data in order to predict an unknown X, denoted by ξ, from further observed responses. When $q > p$ both the profile likelihood for ξ and Bayesian inference for ξ depend on a prediction inconsistency diagnostic which highlights those response vectors used for prediction which are internally inconsistent in the prediction of ξ. When several further response vectors together display systematic anomalies one is led to questioning the estimated model. The information in prediction data about changes in parameters is investigated under various assumptions. The results indicate systematic anomalies in prediction data may be detected in a variety of ways, but corrected only under strong assumptions so that recalibration might be needed.

Some key words: Likelihood ratio test; Profile likelihood; Recalibration; Unsupervized learning.

1. INTRODUCTION

Multivariate calibration uses an observational training data set X_i, Y_i $(i = 1, \ldots, n)$ to construct a relationship between the $q \times 1$ vector Y and the $p \times 1$ vector X. These training data may have X-values fixed in advance as in a controlled designed experiment, the 'controlled' calibration case, or the X-values may be as generated by a random sample of specimens, the so-called 'random' or 'natural' calibration case. Inference methods for a future X-value, denoted by ξ, corresponding to an observed response vector Y, denoted by Z to avoid confusion with the calibration data, should depend on whether the training data are controlled or natural. Controlled calibration lacks useful information on the likely distribution of ξ whereas random calibration may not estimate the relationship between Y and X in the most precise way. Informally, an amalgamation of the two methods is to be preferred and the profile likelihood for this is discussed in § 2.

The relationship between Y and X may be linear or nonlinear and the appropriate choice should be guided by prior knowledge and graphical analysis of the calibration data. For simplicity we treat the multivariate linear model under which

$$Y = 1\alpha' + XB + E, \quad Z = \alpha + B'\xi + \varepsilon, \tag{1·1}$$

where Y, X, Z are observed, α, B and ξ are $q \times 1$, $p \times q$ and $p \times 1$ unknown parameters, 1 an $n \times 1$ vector of ones and with $E = (\varepsilon_1, \ldots, \varepsilon_n)'$ with $\varepsilon_1, \ldots, \varepsilon_n$ mutually independent

$q\times 1$ vectors,

$$E(\varepsilon_i)=E(\varepsilon)=0,$$

and a common unknown covariance matrix Γ for each ε_i and for ε. In addition, multivariate normality is assumed.

Brown (1982) developed both Bayes and sampling theory inference procedures for both controlled and random calibration. For controlled calibration, Brown & Sundberg (1987) derived the profile likelihood of ξ, that is the likelihood maximized over all parameters except ξ, as proportional to

$$[\sigma_l^2(\xi)/\{\sigma_l^2(\xi)+(\bar{Z}-\hat{\alpha}-\hat{B}'\xi)'S_+^{-1}(\bar{Z}-\hat{\alpha}-\hat{B}'\xi)\}]^{\frac{1}{2}(n+1)}. \tag{1·2}$$

A formula analogous to (1·2) also holds when (1·1) is linear in the parameters but nonlinear in certain X variables, with ξ correspondingly constrained. Here $l\geqslant 1$ corresponds to an extension of (1·1) to l replicates of Z each at the same value of ξ, and having $q\times 1$ mean $\bar{Z}$,

$$\sigma_l^2(\xi)=1/l+1/n+(\xi-\bar{X})'G(\xi-\bar{X}),$$

where

$$G=\{(X-1\bar{X}')'(X-1\bar{X}')\}^{-1},$$

and S_+ is the pooled residual sum of products from both calibration and, when $l>1$, prediction experiments. The estimates $\hat{\alpha}$, $\hat{B}$ are standard maximum likelihood estimates from (1·1) without prediction data Z, but the natural generalized least-squares estimator

$$\hat{\xi}=(\hat{B}\hat{\Gamma}^{-1}\hat{B}')^{-1}\hat{B}\hat{\Gamma}^{-1}(\bar{Z}-\hat{\alpha}),$$

with $\hat{\Gamma}=S_+/(n+l-1)$, does not in general maximize (1·2) and is thus not the maximum likelihood estimator of ξ except when $p=q$ or as $n\to\infty$ so that $\hat{B}\to B$, $\hat{\Gamma}\to\Gamma$. The maximum likelihood estimator, a nonlinear estimator given explicitly by Brown & Sundberg (1987), is usually however close to $\hat{\xi}$ even when n is small, except when the prediction inconsistency diagnostic

$$\bar{R}=(\bar{Z}-\hat{\alpha}-\hat{B}'\hat{\xi})'(\hat{\Gamma}/l)^{-1}(\bar{Z}-\hat{\alpha}-\hat{\beta}'\hat{\xi}) \tag{1·3}$$

is too large. When $l=1$ notationally $\bar{Z}$ and $\bar{R}$ revert to Z and R. Such a large value of $\bar{R}$ attests to the internal inconsistency of the q components of $\bar{Z}$ in inference about the p components of ξ $(q>p)$. It widens likelihood-based confidence intervals (Brown & Sundberg, 1987) in harmony with the Bayes approach (Brown, 1982) but in natural discord with the anomalous behaviour of the sampling theory intervals elucidated by Brown (1982) which perversely narrow with increasing R. Oman (1988) adopts an approach to confidence regions which removes the influence of R altogether in order to correct the behaviour of the sampling theory approach to controlled calibration (Oman & Wax, 1984).

The statistic R is central to diagnostic checking, whether or not it influences confidence intervals and point estimators, and is discussed further in § 3.

In routine use typically only $l=1$ replicates are available at each ξ and after a period of use $Z_1,\ldots,Z_t$ will have been observed corresponding to different and unknown $\xi_1,\ldots,\xi_t$ with each Z_j satisfying equation (1·1) and errors being independent for $j=1,\ldots,t$.

Taken individually each Z_j can be examined for consistency through (1·3). Taken collectively and assuming the distribution of ξ_j $(j=1,\ldots,t)$ is known the Z_j provide information for investigating whether the parameters of model (1·1) have changed appreciably since calibration, as discussed in § 5. If on the other hand the Z_j are consistent with the model as estimated from calibration data, they provide information about the random distribution of future ξ_j, especially lacking in controlled calibration; see also § 2.

2. Profile likelihood with some random X

Suppose (Y_i, X_i), for $i = 1, \ldots, n$, are calibration data from model (1·1) with controlled X's, but that there are also separately a random sample of X-values, $p \times 1$ vectors $X_1^*, \ldots, X_t^*$ from $N_p(m, F)$, where both the mean vector, m, and the covariance matrix F are unknown. Then if ξ, which is $p \times 1$, can be assumed to be also generated by the same mechanism, the profile likelihood for ξ solely from this marginal set of t X-values is, following Brown & Sundberg (1987), proportional to

$$\{1 + 1/t + (\xi - \bar{X}^*)' G^* (\xi - \bar{X}^*)\}^{-\frac{1}{2}(t+1)}, \tag{2·1}$$

where

$$G^* = \left\{ \sum_{j=1}^{t} (X_j^* - \bar{X}^*)(X_j^* - \bar{X}^*)' \right\}^{-1}.$$

Now (2·1) multiplied by (1·2) gives the overall profile likelihood for ξ when Y_i are observed for fixed X_i $(i = 1, \ldots, n)$ and random X_j^* $(j = 1, \ldots, t)$ are available. More interestingly, if $X_1^*, \ldots, X_t^*$ is in fact a subset of the calibration $X_i^* = X_i$, this product of (1·2) and (2·1) is still the overall profile likelihood. We are able to treat the conditional distribution of Y given X and the marginal distribution of X quite separately since X is S-ancillary for the parameters of the conditional distribution. If now $t = n$ so that we are at the other extreme of total random calibration then the product of (1·2) and (2·1), the profile likelihood, is

$$\left\{ \frac{\sigma_l^2(\xi)/\sigma_1^2(\xi)}{\sigma_l^2(\xi) + (\bar{Z} - \hat{\alpha} - \hat{B}'\xi)' S_+^{-1} (\bar{Z} - \hat{\alpha} - \hat{B}'\xi)} \right\}^{\frac{1}{2}(n+1)},$$

which when $l = 1$ specializes to

$$\{\sigma_1^2(\xi) + (Z - \hat{\alpha} - \hat{B}'\xi)' S^{-1} (Z - \hat{\alpha} - \hat{B}'\xi)\}^{-\frac{1}{2}(n+1)}. \tag{2·2}$$

A little manipulation following Brown (1982, § 2.4) shows that this is

$$(c + |\xi - \tilde{\xi}|_H^2)^{-\frac{1}{2}(n+1)},$$

where $|a|_H^2$ denotes $a'Ha$, $H = \hat{B} S^{-1} \hat{B}'$ and c is a constant. Here $\tilde{\xi}$ is the maximum likelihood estimator obtained from the regression of X on Y. Also $\tilde{\xi}$ is the best linear predictor of a random ξ from the same distribution as X neglecting estimation errors in Γ, B, α. Thus (2·2) is indeed the profile likelihood from the regression of X on Y, conforming with the natural approach to pure random calibration which provides the joint distribution of Y and X for prediction of a future X and Y. This development parallels and to some extent mimics the Bayesian analysis of § 3 of Brown (1982).

Finally, in passing, we note another source of information touched on in the introduction. In routine use of the calibrated instrument, under controlled or random calibration or a mixture of the two, $Z_1, \ldots, Z_t$ from model (1·1) are provided corresponding to unknown $\xi_1, \ldots, \xi_t$. Assuming these $\xi_1, \ldots, \xi_t$ form a random sample from $N_P(m, F)$, marginally Z_j are distributed independently as $N_q(\alpha + B'm, B'FB + \Gamma)$ for $j = 1, \ldots, t$.

Together with the data from (1·1) this provides further updating of the profile likelihood of ξ. A sampling approach to utilization of this extra information for $p=q=1$ under controlled calibration is given by Williams (1969) and could also be extended to $q \geqslant p \geqslant 1$. In controlled calibration however this utilization of future responses assumes a distribution of the future unknown ξ. This might be regarded as imposing too strong an assumption unless there is independent corroborating evidence available say from partial random calibration. Or else it is a poor substitute for directly observed random X-values, as we see later in § 6. The situation is akin to unsupervized learning as discussed, for example, by Makov (1980).

3. PREDICTION DIAGNOSTICS

In the introduction in (1·3) we identified

$$R = |Z - \hat{\alpha} - \hat{B}'\hat{\xi}|^2_{\hat{\Gamma}^{-1}}, \tag{3·1}$$

the residual sum of products from prediction of the q-vector Z weighted by the inverse of the least-squares error matrix $\hat{\Gamma}$, as a feature which determines the shape of the profile likelihood.

For this section, assume that the parameters α, B and Γ are precisely estimated by $\hat{\alpha}$, $\hat{B}$ and $\hat{\Gamma}$ so that they may be regarded as known. Under this assumption ξ is the unknown parameter and under normality it follows from least-squares theory that R is distributed as chi-squared with $(q-p)$ degrees of freedom.

Let $\hat{Z} = \hat{\alpha} + \hat{B}'\hat{\xi}$, $\tilde{Z} = \hat{\alpha} + \hat{B}'\tilde{\xi}$ be the fitted values of Z corresponding respectively to $\hat{\xi}$, $\tilde{\xi}$ of § 2.

Now, it may be seen that

$$E\{(Z-\tilde{Z})(Z-\tilde{Z})'\} = \hat{\Gamma}(\hat{\Gamma} + B'F\hat{B})\hat{\Gamma} = \Theta^{-1},$$

where this defines Θ and F was defined at the end of the previous section; see also Naes & Martens (1987, § 3). Defining

$$R_B = |Z - \tilde{Z}|^2_{\Theta},$$

Naes & Martens (1987) show that

$$R_B = R + R_X, \tag{3·2}$$

where $R_X = |\hat{Z} - \tilde{Z}|^2_{\Theta}$. Under normality R_X and R are independent chi-squareds with p and $(q-p)$ degrees of freedom. For further discussion, see Naes (1985, § 2; 1986). Thus in addition to the inconsistency diagnostic, R, he pinpoints R_X as a diagnostic useful in identifying outliers in X, that is ξ space. A large value of R_X would suggest that particular care should be exercized in using $\tilde{\xi}$, the natural case predictor, since it shrinks towards $\bar{x}$ the mean of the calibration experiment. In practice if R is large one would probably wish to see whether particular individual components of $\hat{\Gamma}^{-\frac{1}{2}}(Z - \hat{\alpha} - \hat{B}'\hat{\xi})$ are large. This would be especially useful if $q \gg p$.

Both diagnostics, R and R_X, assume the calibration model to be true. In practice, after calibration, many different samples are analysed corresponding to different and unknown ξ values. Consistent repeated departures from expected values lead one to question the model rather than to identifying outliers.

However whilst one is able to detect changes from the calibration model, correction of the regression model is not possible without further strong assumptions, since the ξ

are unknown. We first give a negative result to the effect that no useful inferences to correct for possible changes in the regression parameters can be made without the assumption of a distribution for ξ. We then proceed in § 5 to tests that utilize such a distribution.

4. Inference about regression change: No assumption on ξ's

Here we study whether future observations $Z_1, Z_2, \ldots$ can be used to detect and correct for deviations between the estimated matrix $\hat{B}$ of regression coefficients, regarded as known, and the true matrix B of the prediction phase. Reasons behind such deviations could be estimation 'errors' in $\hat{B}$ and a change in true relationship since calibration. For simplicity α is taken to be zero. We make no assumption about $\xi_1, \xi_2 \ldots$ here; they could be regarded as arbitrary parameters. We show a result of the following kind: without assumptions on ξ, only irrelevant deviations between $\hat{B}$ and the true B can be detected and corrected for from Z data.

Hence the best we can do from Z data is to let the detection of an inconsistency between $\hat{B}$ and B motivate a recalibration.

If $q = p$, not even irrelevant deviations can be detected. There is a one-to-one correspondence between $E(Z)$ and ξ through $E(Z) = B'\xi$, so nothing can be inferred about B from Z without assumptions on ξ. If $q > p$, however, Z has more components than needed for the estimation of ξ, so there is some information about B. Write

$$Z = KZ + (I - K)Z = Z_{(1)} + Z_{(2)}$$

for $K = \hat{B}'(\hat{B}\hat{\Gamma}^{-1}\hat{B}')^{-1}\hat{B}\hat{\Gamma}^{-1}$. Here $Z_{(1)}$ and $Z_{(2)}$ vary in subspaces $L_{(1)}$, $L_{(2)}$ of dimensions p and $q-p$ respectively. The matrix K projects Z in a q-dimensional Euclidean space onto $L_{(1)}$ in direction $L_{(2)}$, and vice versa for $I-K$, with $L_{(1)} \oplus L_{(2)} = R^q$. This may be implemented by a singular value decomposition. Also, $Z_{(1)}$ and $Z_{(2)}$ are uncorrelated. Now $Z_{(1)} = \hat{B}'\hat{\xi}$, with expected value

$$E(Z_{(1)}) = KB'\xi, \tag{4·1}$$

which is a one-to-one function of ξ, whereas $Z_{(2)}$ appears in the inconsistency diagnostic R of (3·1) and satisfies

$$E(Z_{(2)}) = (I - K)B'\xi = 0$$

when $B = \hat{B}$.

Based on Z data only, $Z_1, \ldots, Z_t$ say, a test of the hypothesis $B = \hat{B}$, $\Gamma = \hat{\Gamma}$, may be constructed from the $Z_{(2)}$ values, e.g. with test statistic

$$R_. = \sum_{i=1}^{t} R_i, \tag{4·2}$$

where the R_i's are the individual values of the inconsistency diagnostic. Naes (1985) and Naes & Martens (1987) used the R_i for detection of outliers in Z-space without assumptions on ξ. This is the natural interpretation of a large value for an individual Z. A significantly large value of $R_.$ not explained by gross errors in one or a few individual Z vectors must be explained by a deviation between B and $\hat{B}$, or by an increased true Γ, as compared with $\hat{\Gamma}$.

Let us assume $\Gamma = \hat{\Gamma}$ and try to correct $\hat{B}$. Write

$$E(\hat{\xi}) = M\xi. \tag{4·3}$$

The relationship between M and K is seen from (4·1), $KB' = \hat{B}'M$. The expected value of $Z_{(2)}$ may be written

$$E(Z_{(2)}) = (I - K)B'\xi = (B' - \hat{B}'M)\xi.$$

We assume M nonsingular. In particular this condition is satisfied in the typical case when B is little different from $\hat{B}$, that is $M \simeq I$.

The linear relationship between the expected values of $Z_{(2)}$ and $\hat{\xi}$ makes it possible to estimate

$$C = (B' - \hat{B}'M)M^{-1} = B'M^{-1} - \hat{B}'.$$

Let us even assume C known. Then $\bar{B}' = \hat{B}' + C = B'M^{-1}$ would be a partially corrected alternative to $\hat{B}$, that could replace the latter in $\hat{\xi}$ to form a new estimator $\bar{\xi}$. However, it is easily seen that $E(\bar{\xi}) = M\xi = E(\hat{\xi})$, so the attempted correction does not reduce the bias. In this sense the correction is irrelevant for estimation of ξ.

5. Inference about regression change: ξ distribution assumed

In prediction, a change in the distribution of Z from that of Y in the calibration sample, as measured by the prediction sample $Z_1, \ldots, Z_t$ could be the result of (a) a change in α, B, (b) a change in Γ and (c) a change in the distribution of ξ. Section 4 showed that correction for (a) is not possible without further assumptions. Here we assume that (b) and (c) do not occur and that furthermore the distribution of ξ is known from previous experience. Note that this implies that the calibration experiment, although controlled, is augmented by random X data and prediction of ξ could take account of these random X as in § 2. However our interest here is in the Z distribution and not direct prediction of ξ.

Throughout this section we assume the parameters Γ, α, B known although in reality they will be estimated from the calibration experiment, and effectively we assume Γ is known. Even if the calibration experiment does not provide precise information about α, B repeated prediction will be in terms of such values and we are interested in deviations from them and not from some hypothesized true values. For emphasis these calibration estimated values will be denoted $\hat{\alpha}$, $\hat{B}$.

We assume that $\xi_1, \ldots, \xi_t$ form a p-variate normal sample with known mean m and known $p \times p$ variance matrix F. Thus since Γ and F are known and unchanged throughout this section, model (1·1) will be utilized after ε, Z, α, B' have been premultiplied by $\Gamma^{-\frac{1}{2}}$ and B' postmultiplied by $F^{\frac{1}{2}}$, so that the prediction part of (1·1) reads

$$Z_i = \alpha + B'\xi_i + \varepsilon_i \quad (i = 1, \ldots, t), \tag{5·1}$$

where now ε_i, ξ_i are independent q and p variate normal with identity variance matrices. To reconstruct the untransformed notation of previous sections all subsequent formulae must be transformed back by the substitutions

$$\alpha \to \Gamma^{-\frac{1}{2}}\alpha, \quad B' \to \Gamma^{-\frac{1}{2}}B'F^{\frac{1}{2}}, \quad Z_i \to \Gamma^{-\frac{1}{2}}Z_i.$$

With $p \leq q$, the relationship between Z and ξ involves p separate linear relationships. In order to avoid testing for change in α, B not relevant to prediction of ξ we immediately specialize to a canonical form of the model (5·1) estimated at $\alpha = \hat{\alpha}$, $B = \hat{B}$. For a suitable choice of orthogonal matrices Q and P, the transformed variables $Z_i^* = QZ_i$, $\xi_i^* = P\xi_i$, may be used to express estimated (5·1) in the canonical form

$$Z_i^{*(1)} = \hat{\alpha}_*^{(1)} + \hat{B}_*'\xi_i^* + \varepsilon_i^{(1)}, \tag{5·2}$$

$$Z_i^{*(2)} = \hat{\alpha}_*^{(2)} + \varepsilon_i^{(2)}, \tag{5·3}$$

where $Z_i^{*(1)}$ and $Z_i^{*(2)}$ are p and $(q-p)$ dimensional vectors respectively, $\hat{\alpha}_*$ is a new constant vector, $\hat{B}_* = \operatorname{diag}(\hat{\beta}_j)$ is a diagonal $p \times p$ matrix with singular values of $\hat{B}$ as diagonal elements, that is $\hat{\beta}_j^2$ are the eigenvalues of $\hat{B}\hat{B}'$, and $\varepsilon_i^{(1)}$, $\varepsilon_i^{(2)}$ concatenated form new independent identically distributed $N_q(0, I)$ vectors. This orthogonal linear transformation of the q components Z is similar in spirit to the linear projection adopted in § 4.

Now the orthogonal matrix Q is a function of $\hat{B}$ and transforms the estimated model to canonical form. If α, B change from $\hat{\alpha}$, $\hat{B}$, so will both α_* and B_* in (5·2), (5·3), the latter from zero in (5·3). However, although (5·2) and (5·3) thus both provide testable information about α, B, only (5·2) involves ξ as used for prediction so that change in α, B manifest in (5·3) will not effect the behaviour of $\hat{\xi}$ or $\tilde{\xi}$ for prediction of ξ. Thus our null hypothesis is $(\alpha_*^{(1)}, B_*) = (\hat{\alpha}_*^{(1)}, \hat{B}_*)$ and the alternative $(\alpha_*^{(1)}, B_*) \neq (\hat{\alpha}_*^{(1)}, \hat{B}_*)$.

Marginally, averaging over the unobserved ξ_i^* distributed as $N_p(PF^{-\frac{1}{2}}m, I)$, (5·2) becomes

$$Z_i^{*(1)} = \mu_* + \delta_i^*, \tag{5·4}$$

where $\mu_* = \alpha_*^{(1)} + B_*'PF^{-\frac{1}{2}}m$ and δ_i^* are independent identically $N_p(0, \Delta_*)$ with $\Delta_* = B_*'B_* + I$, and under the null hypothesis of no change $\alpha_*^{(1)}$ and B_* are completely specified whereas under the alternative μ_* and B_* are arbitrary and B_* enters the problem through Δ_*.

If the null hypothesis had been $\alpha_*^{(1)} = \hat{\alpha}_*^{(1)}$ versus $\alpha_*^{(1)} \neq \hat{\alpha}_*^{(1)}$, with B_* constant throughout then this test could be accomplished through (5·4) merely by testing for a mean change with a known constant covariance matrix for the error. A null hypothesis concerning B_* alone is more problematic since B_* enters both μ_* and Δ_* but could be accomplished with a likelihood ratio test procedure for example. As a consequence of both $\alpha_*^{(1)}$, B_* being allowed to change as is the present case, one is allowed to regard μ_* as arbitrary under the alternative.

To conduct a likelihood ratio test we need to obtain the difference of the log likelihood under null hypothesis and log likelihood maximized over the alternative choices of $\alpha_*^{(1)}$, B_*. Under the null hypothesis the log likelihood is

$$-\tfrac{1}{2}pt\log(2\pi) - \tfrac{1}{2}t\left\{\log|\hat{\Delta}| + t^{-1}\sum_{i=1}^{t}|Z_i^{*(1)} - \hat{\mu}|^2_{\hat{\Delta}^{-1}}\right\}, \tag{5·5}$$

where $\hat{\Delta}$, $\hat{\mu}$ are values of Δ_*, μ_*, with B_*, $\alpha_*^{(1)}$ set to their hypothesized values $\hat{B}_*$, $\hat{\alpha}_*^{(1)}$.

Now if we take the general log likelihood and maximize over B_*, μ_*, first maximizing over μ_*, we have that

$$-\tfrac{1}{2}pt\log(2\pi) - \tfrac{1}{2}t\left\{\log|\Delta_*| + t^{-1}\sum_{i=1}^{t}|Z_i^{*(1)} - \bar{Z}^{*(1)}|^2_{\Delta_*^{-1}}\right\}$$

and the remaining maximization over B_* is equivalent to minimization of

$$h(\Delta_*, V) = \log|\Delta_*| + \operatorname{tr}(\Delta_*^{-1}V)$$

with

$$V=\sum_{i=1}^{t}(Z_i^{*(1)}-\bar{Z}^{(1)})(Z_i^{*(1)}-\bar{Z}^{*(1)})'/t$$

the sample covariance matrix of $Z^{*(1)}$ remembering that $\Delta_*=B'_*B_*+I$, the minimization over B_* is for $t>p$ given by the maximum likelihood result of Jöreskög used in factor analysis. Denoting $\phi_1,\ldots,\phi_p\geqslant 0$, the eigenvalues of the $p\times p$ covariance matrix V and setting

$$a_i=\max(\phi_i-1,0)\quad(i=1,\ldots,p), \tag{5·6}$$

the maximum of $h(\Delta_*, V)$ is at $\Delta_*=\tilde{\Delta}_*$ with

$$h(\tilde{\Delta}_*, V)=\sum_{i=1}^{p}\{\log(a_i+1)-a_i\phi_i/(a_i+1)+\operatorname{tr}(V)\};$$

see, for instance, Mardia, Kent & Bibby (1979, p. 265). The maximized log likelihood to be compared with (5·5) is then

$$-\tfrac{1}{2}pt\log(2\pi)-\tfrac{1}{2}th(\tilde{\Delta}_*, V). \tag{5·7}$$

The reference value of unity in (5·6) stems from the transformation to identity error variance matrix preceding (5·1). Twice {(5·7) minus (5·5)} gives a log likelihood ratio test which is asymptotically distributed as chi-squared with

$$p(p+1)-\tfrac{1}{2}p(p-1)=\tfrac{1}{2}p(p+3)$$

degrees of freedom, where this calculation of degrees of freedom takes into account the indeterminacy in $\hat{\Delta}_*$ up to a $p\times p$ orthogonal matrix.

When $p=1$ twice the log likelihood ratio specializes to

$$W=t\log(\Delta/V)+t\{(\hat{V}/\hat{\Delta})-1\}, \tag{5·8}$$

where from (5·5)

$$\hat{V}=t^{-1}\sum|Z_i^{*(1)}-\hat{\mu}|_{\hat{\Delta}^{-1}}.$$

Note that, had we not reduced the model to a canonical form as in (5·1), then the degrees of freedom of the likelihood ratio chi-squared would have been

$$q(p+1)-\tfrac{1}{2}p(p-1),$$

corresponding to the more specialized null hypothesis $\alpha=\hat{\alpha}$, $B=\hat{B}$.

Note finally the crucial assumption of a known ξ distribution. Actually a likelihood ratio test for detection of a change in (m, F) under the assumption of (α, B) fixed has precisely the same test statistic as one to detect a change in $(\alpha_*^{(1)}, B_*)$ under fixed (m, F).

In providing the above likelihood ratio test statistic we have also in the process provided interval and more particular point estimates of $\alpha_*^{(1)}$, B_*. Whilst B_* effects both the mean and variance of the distribution of $Z_*^{(1)}$, a change in $\alpha_*^{(1)}$ merely effects the mean of this distribution. Should a change be indicated by the likelihood ratio procedure, the maximum likelihood estimates of $\alpha_*^{(1)}$, B_* may be used to correct the calibration relationship. However in such a correction we will have to rely on the known fixed distributions of ξ and if such an assumption is deemed too strong a detected change will need to be corrected by recalibration.

Section 3 of the present paper discussed the prediction diagnostic (3·1) for a single observation. With α, B, Γ known from the calibration experiment this diagnostic for the ith prediction observation is

$$R_i = |Z_i^{*(2)} - \hat{\alpha}_*^{(2)}|^2$$

from (5·3). Thus R_i or $R_.$, the total over the t prediction observations, is not directly relevant to correcting change in the calibration parameters. An analogous insight is given in § 4 without the strong assumption of the known ξ distribution.

The above likelihood ratio procedure tests for a change in both mean and variance of the transformed and selected p components $Z_i^{*(1)}$. Tests of variance are notoriously sensitive to normality assumptions (Box, 1953), and should perhaps be made robust. Alternatively and perhaps more usefully, one may derive graphical monitoring procedures. When $p = 1$ we have a single component $Z_i^{*(1)}$ for $i = 1, \ldots, t$ whose mean and variance are known to be $\hat{\mu}$ and $\hat{\sigma}^2$ under the null hypothesis. A change in mean alone is indicative of a change in $\alpha_*^{(1)}$ from model (5·2), whereas a change in B_* will result in both a change in mean and variance and may increase or decrease either or both. A time series plot of $(Z_i^{*(1)} - \hat{\mu})/\hat{\sigma}$ will provide a plot of a random standard normal sample. Control or cusum charts could be utilized for the purpose of monitoring process progress. When $p > 1$, then $\hat{\Delta}_0^{-\frac{1}{2}}(Z_i^{*(1)} - \hat{\mu})$ is $N_p(0, I)$ under the null hypothesis and drift in the parameters $\alpha_*^{(1)}$, B_* will manifest themselves in changes of mean from zero and changes in the covariance matrix from the identity matrix I. The simplest single check is probably provided by the Mahalanobis squared distance

$$(Z_i^{*(1)} - \hat{\mu})'\Delta^{-1}(Z_i^{*(1)} - \hat{\mu}),$$

which will be a random sample from a chi-squared distribution with p degrees of freedom under the null hypothesis. Gnanadesikan (1977, p. 172) describes graphical methods involving chi-squared or gamma plots.

When the parameters are expected to change slowly through time, a Kalman filter updating procedure might be utilized. In this paper we have assumed a sharp and constant change in the parameters. If change were more progressive it would serve to lessen the power of the proposed test. Smith & Corbett (1987) have utilized the Kalman filter in the calibration of cyclists for measurement of a marathon run.

With t future Z-values, the next section looks at the amount of information available to measure change. Assuming that the unknown ξ_i $(i = 1, \ldots, t)$ form a random sample from a known distribution, as in the above, we show that there may be little information relative to the case where ξ_i are known, the calibration case. The rub then is that the proposed test of parameter change may only be able to pick up sizeable changes and regular recalibration may be unavoidable. The example of § 7 ends our investigation on a more positive note by illustrating how in practice a change in calibration parameters can be detected and corrected.

6. Comparative information, supervized versus unsupervized learning

To illustrate the comparison we assume $p = q = 1$. For so-called supervized learning, the t x-values are known, and with a simple linear regression model we have the following.

Case 1. $Z_i = \alpha + \beta x_i + \varepsilon_i$ $(i = 1, \ldots, t)$, with ε_i a random sample from $N(0, \sigma^2)$, conforming with the notation prior to the canonical reductions of the previous section.

For unsupervized learning, on the other hand, the x's are unknown but are assumed to form a random sample from a given $N(0, f^2)$ distribution.

Case 2. $Z_i = \alpha + \varepsilon_i$ $(i = 1, \ldots, t)$, with ε_i a random sample from $N(0, \delta^2)$ and $\delta^2 = \sigma^2 + \beta^2 f^2$.

Suppose α is known but β is unknown. For Case 1, the information from t observations is $I(\beta) = t\hat{f}^2/\sigma^2$, where $\hat{f}^2 = \Sigma x_i^2/t$.

For Case 2, the log likelihood for a single observation, Z, is

$$l(\beta) = c - \log(\delta) - (Z - \alpha)^2/(2\delta^2)$$

so that the information from t observations is $J(\beta) = 2t(\beta f^2/\delta^2)^2$.

Comparison shows that J may be much smaller than $I(\beta)$. Assuming that $\hat{f}^2 = f^2$,

$$J(\beta)/I(\beta) = 2\rho^2(1 - \rho^2),$$

where $\rho^2 = \beta^2 f^2/\delta^2$, and the ratio is always $\leq \frac{1}{2}$. The more accurate the calibration, the nearer ρ^2 to 1, and the closer the ratio to zero. However, the case of high ρ^2 corresponds to very precise determinations of β from Case 1 and, although relatively less precise, Case 2 may still provide quite precise information in absolute terms. Overall, though, knowledge of the true distribution of ξ constitutes a strong assumption and, if economically feasible and a priori warranted, regular recalibration is desirable.

7. AN EXAMPLE

Here we illustrate the use of statistics for testing and correcting for change in the calibration relationship developed in § 5, and the further diagnostics R and R_X. This will show that, despite reservations about the power of the test, real changes are quite detectable and in fact a change is signalled where one was not a priori envisaged. We have used three sets of data from the Flour Milling and Baking Research Association:

Set I: A population of protein values, X values, accurately determined for 381 samples of white flour, milled from single U.K. wheat varieties.

Set II: A calibration set of 57 of these flour samples with recorded values of protein, X, together with three logged near infrared, NIR, reflectance values at wavelengths 2180, 2100, 1680 nanometers, denoted by Y_1, Y_2, Y_3.

Set III: A further set of 39 near infrared reflectances at the six wavelengths comprising samples of 39 flours,
(a) 20 flours of similar origin to those in sets I and II;
(b) 13 flour or flour-like samples, e.g. ground wheat, wholemeal, flour from Saudi Arabia, etc.;
(c) 6 very unflour-like samples, bran, gluten, improver, etc.
True protein measurements were not available for this set of 39.

The mean protein values in sets I and II were similar although the standard deviation in set I was about twice that of sample II. It seems sample II was chosen from population I purely on grounds of ready availability and not deliberately to achieve a smaller standard deviation. Our method uses the population I standard deviation as typical of future samples and this corrects for an inbalance in the chosen samples as regards protein. The mean and variance of the 381 observations in data set I were $m = 8{\cdot}22$ and $F = (1{\cdot}22)^2$ respectively.

Sets I and II included laboratory measurements of moisture, too. The analysis has concentrated on protein. However, systematic selection on moisture in combination with protein would warrant modelling both moisture and protein in order to investigate protein calibration. We have kept analysis simple by ignoring moisture completely, but a need for recalibration might be caused by a systematic change in moisture.

By concentrating on protein alone, $p = 1$, we are able to telescope the approach through (5·1)–(5·7) by noting that there is a single canonical variate relating Y to X and that this is most easily obtained by regressing X on Y with the resultant equation from data set II, using the variables Y_1, Y_2, Y_3 for predicting protein,

$$X = 12{\cdot}7 + 0{\cdot}00494\,Y_1 - 0{\cdot}00323\,Y_2 - 0{\cdot}00243\,Y_3,$$

with 92·6% variation explained. This combination of the three NIR values, scaled to give a residual standard deviation of 1, yields $Z^{*(1)}$ from data set II and the regression $Z^{*(1)} = -81{\cdot}61 + 6{\cdot}23X$ and of course 92·6% of variation as explained above. With this residual standard deviation and mean and standard deviation of population I we are able to specify

$$\hat{\mu} = -81{\cdot}61 + 6{\cdot}23 \times 8{\cdot}22 = -30{\cdot}34,$$
$$\hat{\Delta} = 1 + (6{\cdot}23)^2 \times (1{\cdot}22)^2 = (7{\cdot}67)^2,$$

With the reduction to $p = 1$, the test statistic, W, minus twice log likelihood ratio, is given by (5·8).

Table 1 gives summary statistics for the three subsets of flour type from set III. Since $\frac{1}{2}p(p+3) = 2$, we use chi-squared tables with two degrees of freedom; all values of W in the table are greater than even the upper 0·1% point. Now it is reasonable to assume that protein values for the U.K. flours are like those of population I and that the significant difference is due to a change in the calibration relationship. This was not expected, but after further investigation turned out to be explained by a time lag and drift between data sets II and III.

Table 1. *Statistics derived from data set III*; *t, sample size*

	U.K. flours	Flour-like	Flour-unlike	Expected
				under H_0
	$t = 20$	$t = 13$	$t = 6$	
Mean	−20·4	−15·9	93·0	−30·34
$V^{\frac{1}{2}}$	5·97	12·81	167·7	7·67
$R_{.}$, (4·2)	66	8671	3720	—
R_X, (3·2)	47	125	13138	—
W, (5·8)	36·8	52·8	3898	—

Flour-like samples and even more so the flour-unlike samples could in reality manifest significant values of the test statistics due to a different distribution of protein values as well as a changed calibration relationship.

Figure 1 shows $Z^{*(1)}$ plotted for each of twelve different 'flours'. The baseline here is as given in Table 1 as 'expected under H_0', on the assumption that the ξ distribution is as in data set I. Here flours B–G constitute flour-like samples and H–L flour-unlike samples, and A represents the 20 U.K. flours.

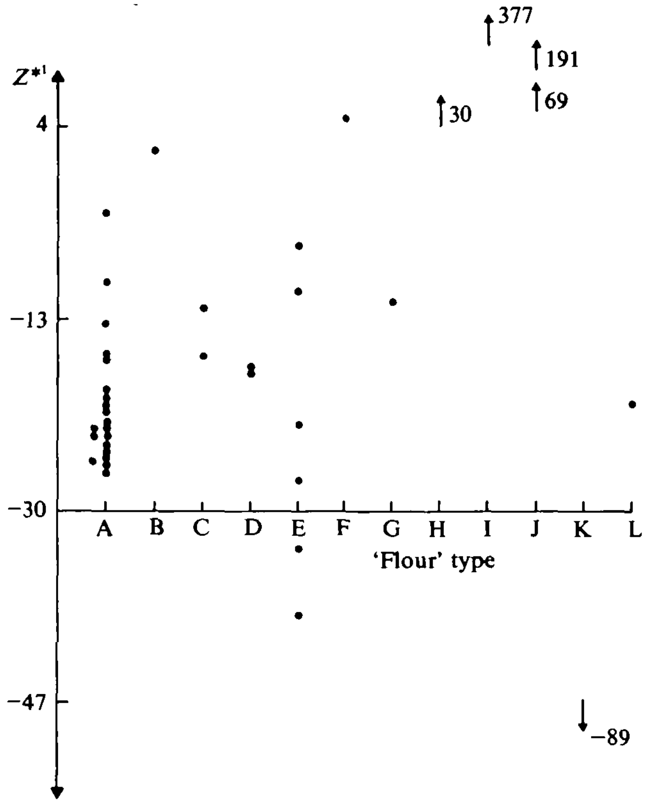

Fig. 1. Plot of $Z^{*(1)}$ against 'flour' type for 39 observations of data set III; A, U.K. flour; B, Saudi Arabian flour; C, South African flour; D, wheat flour; E, ground wheat; F, wheat feed; G, wholemeal; H, bran; I, gluten; J, improver; K, polydextrose; L, dried bread.

Looking closer at the observed means and standard deviations for the three flour types we can see that for U.K. flours the standard deviations are essentially the same whereas the change in mean from −30·34 to −20·4 is clearly significant. Thus for U.K. flours correction of the calibration relationship would seem to be adequately accomplished by adjusting $\alpha_*^{(1)}$ from the calibration value of $-81{\cdot}61$ to $-81{\cdot}61+(30{\cdot}34-20{\cdot}4)=-71{\cdot}65$.

On the other hand, for both the other flour types a chi-squared variance comparison with $(7{\cdot}67)^2$ yields significant test statistics at the 5% level, indicating that both $\alpha_*^{(1)}$ and B_* have changed. We could go on to adjust for both of these but also taking into account the large values of the prediction consistency diagnostic $R_.$ we would be more wary and led to recalibrate. The example is contrived in that we know that the flour-like samples are not entirely homogeneous and not similar to the U.K. flours, but had we thought that we were dealing with quite similar material we would be more persuaded to recalibrate than correct given such large values of $R_.$. The individual components of R, R_X are nominally independent chi-squared on $(q-p)=2$ and $p=1$ degrees of freedom. A test using statistic $R=66$ is just significant at the 1% level on $20\times2=40$ degrees of freedom for the U.K. wheat flours. All the other entries are highly significant. These statistics are more illuminating when used at an individual unit level. Figure 2 gives a plot of log R versus log R_X for the 39 observations identified by 12 symbols. While some observations are high on both measures others are high on just one of the measures. Here R is an inconsistency diagnostic whereas R_X indicates an outlier in X space.

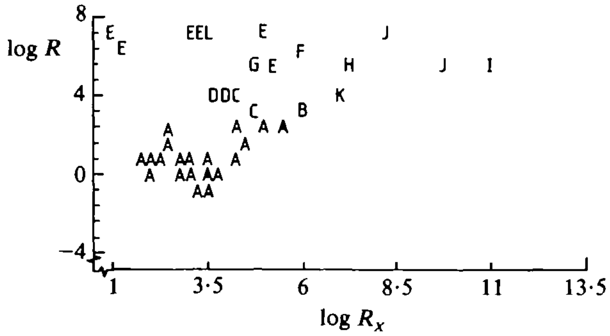

Fig. 2. Prediction diagnostics plotted for data set III; 'flour' type identified by letters given in the legend to Fig. 1.

Acknowledgement

We are grateful to the Flour Milling and Baking Research Association and in particular Dr T. Fearn for providing us with the data analysed in § 6.

References

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40**, 318-35.

Brown, P. J. (1982). Multivariate calibration (with discussion). *J. R. Statist. Soc.* B **44**, 287-321.

Brown, P. J. & Sundberg, R. (1987). Confidence and conflict in multivariate calibration. *J. R. Statist. Soc.* B **49**, 46-57.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* New York: Wiley.

Makov, U. E. (1980). Approximations of unsupervised Bayes learning procedures (with discussion). In *Bayesian Statistics*, Ed. J. M. Bernardo et al., pp. 69-82, 128-37. Valencia University Press.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis.* London: Academic Press.

Naes, T. (1985). Multivariate calibration when the error covariance matrix is structured. *Technometrics* **27**, 301-11.

Naes, T. (1986). Detection of multivariate outliers in linear mixed models. *Comm. Statist.* A **15**, 33-47.

Naes, T. & Martens, H. (1987). Testing adequacy of linear random models. *Statistics* **18**, 323-31.

Oman, S. D. (1988). Confidence regions in multivariate calibration. *Ann. Statist.* **16**, 174-87.

Oman, S. D. & Wax, Y. (1984). Estimating fetal age by ultrasound measurements: An example of multivariate calibration. *Biometrics* **40**, 947-60.

Smith, R. S. & Corbett, M. (1987). Measuring marathon courses: An application of statistical calibration theory. *Appl. Statist.* **36**, 283-95.

Williams, E. J. (1969). Regression methods in calibration problems. *Bull. Int. Statist. Inst.* **43**, 17-28.

[*Received July* 1986. *Revised August* 1988]