 Vol 26: 161–207, 1999

Multivariate Calibration — Direct and Indirect Regression Methodology*

ROLF SUNDBERG

Stockholm University

ABSTRACT. This paper tries first to introduce and motivate the methodology of multivariate calibration. Next a review is given, mostly avoiding technicalities, of the somewhat messy theory of the subject. Two approaches are distinguished: the estimation approach (controlled calibration) and the prediction approach (natural calibration). Among problems discussed are the choice of estimator, the choice of confidence region, methodology for handling situations with more variables than observations, near-collinearities (with countermeasures like ridge type regression, principal components regression, partial least squares regression and continuum regression), pretreatment of data, and cross-validation vs true prediction. Examples discussed in detail concern estimation of the age of a rhinoceros from its horn lengths (low-dimensional), and nitrate prediction in waste-water from high-dimensional spectroscopic measurements.

Key words: bilinear regression, collinearity, continuum regression, cross-validation, generalized least squares, least squares ridge regression, PCR, PLS, prediction, spectroscopic data

1. Introduction and background

1.1. Introduction

We see an ever increasing need to extract quantitative information in all areas of science and technology, that goes hand in hand with the developments in measurement instrument technology. Data analysis must face larger data sets, with more variables and more observations, and the computer revolution makes it possible to handle them numerically. Statistics is the science of transforming data into information and knowledge, so statistical methodology must not be left behind. Fifteen years ago instrument calibration was a univariate small-sample technique, to many statisticians known just for its history of methodological controversy and for theoretically peculiar confidence regions. Since then calibration has become a multivariate field of research and applications, for analysis of high-dimensional data and of great practical relevance.

By the term calibration of the title we should think of situations where we typically have:

1. Two types of measurements or observations for each item or individual:
   (i) A characteristic t that is laborious or expensive to measure, or even impossible to measure for some individuals,
   (ii) A quick or cheap but on the other hand perhaps less accurate measurement u. (Pure instrumental measurement noise is often quite low, actually, but there are typically more sources of error to be allowed for).

We want to use u as an indirect measurement of t, that is we want to determine (estimate or predict) the corresponding unknown t when u has been measured.

2. A model for the relationship between t and u; we will usually assume a linear regression of u on t, or of t on u, in the latter case typically in combination with a joint distribution for (t, u).

*This paper was presented as an Invited Lecture at the 16th Nordic Conference on Mathematical Statistics, Lahti, Finland August 1996.

3. A calibration (or training) sample of complete pairs (t, u).

When u is multivariate, we talk of multivariate calibration. We may also allow t to be a vector, if we only require $q \geqslant p$, where $q = \dim u$, $p = \dim t$, since otherwise t cannot be identified even from an error-free u. Multivariate calibration is used either because we want to deal with several t-variables jointly ($q \geqslant p > 1$), or only because we want to utilize (combine) the information about t in several u-values jointly ($q > p \geqslant 1$).

In today's analytical chemistry multivariate calibration examples are abundant. In the typical chemical situation the concentration t of one or several substances jointly should be determined. The true concentrations (or near so) are known for specially prepared reference samples (standards), or from a wet chemistry reference method. A physical measurement method yields the u-values. Often this is a measurement of absorption, reflection or transmission of light, that is easily and quickly performed over a sequence of wavelengths by a more or less automatized instrument. The calibration problem concerns the estimation of the relationship between u and t and the use of this relationship in inference about unknown t from observed u. The multivariate character appears when the absorption (reflection, transmission) of light is measured at several different wavelengths jointly. The number q of wavelengths can be very large and count in hundreds or even thousands, whereas the training sample size n is usually moderate (a two-digit number). An example will be presented in detail in section 3.

Many chemical calibration problems are biological in their scientific origin, concerning the chemical composition of for example meat, fat, dough or cereals, or constituents in environmental samples. More purely biometric applications have also appeared in the statistical literature (and there should be a potential for many more such applications). Wood (1982) discussed the determination of the age t of young female water rats from $q = 2$ different body measurements u, based on a non-linear regression model. Oman & Wax (1984) estimated gestational age from two fetal bone measurements made by ultra-sound. Clarke (1992) estimated age of rhinoceros from the lengths of their two horns, see also de Plessis & van der Merwe (1996). These applications differ from the typical applications of analytical chemistry in that $\dim t$ and $\dim u$ are both small ($p = 1$, $q = 2$ here) but calibration samples can be large (sizes were $n = 139$, $n > 1000$, and $n = 12$, respectively), and that they require fit of non-linear relationships.

1.2. Direct or indirect regression—natural or controlled calibration

When calibration is termed natural we think of each unit as randomly sampled from one and the same "natural" population, so the pair (t, u) of measurements has a joint distribution. When component t is missing, we are led to predict t. The best unbiased predictor is the regression of t on u, linear under normality. Thus a natural calibration situation motivates prediction of t in a direct regression of t on u. Fitting this regression to the calibration sample helps to estimate the parameters needed.

On the other hand we may prefer to assume less structure, and regard the unknown t as a fixed quantity to be estimated, and the calibration sample t-values as more or less deliberately selected. This is controlled calibration. Only a regression of u on t is then motivated, not the converse. The statistical task is to estimate the regression relationship u on t from the calibration sample and to infer about the unknown t by inverting this relationship. We have here chosen to denote this situation as indirect regression, since the regression of u on t is not the solution to the estimation problem but only a start to the more complex inversion problem, in contrast to the direct regression above where the regression was all that was needed to get the t-value.

Summarizing, we have reason to contemplate both direct regression, of t on u, and indirect

regression, of u on t, for the determination of an unknown t. Since use of conventional notations is likely to simplify for the reader we will discuss both situations in terms of a regression of a response vector y on an explanatory vector x, where inference then concerns an unknown $t = y$ in the former case (direct regression), and an unknown $t = x$ in the latter situation.

1.3. Is multivariate calibration motivated?

Why use multivariate calibration when univariate is much simpler? If we want to infer about a p-dimensional t, $p > 1$, we must measure at least p u-variables. Under ideal circumstances we could imagine making p univariate calibrations, by selecting for each component of t a specific u-variable, that is a u that reacts to changes in that component of t exclusively. However, in reality they would not be quite specific. If we were so lucky that the constituents varied randomly and independently, the consequence of this non-specificity would only be that we lost precision in the estimated t-values, but worse things could happen.

Furthermore, multivariate calibration with $q = p$ is hardly more complicated than univariate calibration. Hence there is no reason to avoid multivariate calibration when it is needed ($p > 1$).

So if we are only interested in a one-dimensional t, then univariate calibration is enough? By measuring several u-variables jointly we can gather more information about t than from one only. We must only learn how to extract this information. This is not to say that many variables necessarily put us in a better position than few—if we add variables which mostly contribute noise we will not be better off, on the contrary we might be misled by spurious correlations.

Another important aspect is the redundancy in $q > p$ correlated u-variables. By using suitable diagnostics to check their internal consistency we have good opportunities to detect a gross error in one of the u-variables, or an interference from an unknown, uncontrolled constituent, instead of risking that such errors go from a single u straight into the estimated/predicted t. With two-way response data, section 4.2, we are in an even better position, in that we may even correct for such interference in the specimen analysed.

Borrowing from Harald Martens, we may draw an analogy with music played on one string or several. The latter is richer, and we might hear from disharmony if one string has been falsely tuned (calibration error!) or the wrong note struck (outlier).

But my knowledge of multivariate statistics is not too good; will I understand the procedures? Indeed, with calibration regarded from a predictive point of view we do quite well without multivariate statistics even for theory; the essential tool then is multiple regression methodology. More generally, however, statistical understanding of multi-dimensional data structures must become wide-spread. In particular this concerns methods of estimation/prediction and of dimension reduction.

Should we estimate or predict (regress u on t or t on u)? This question can be posed in the univariate case as well as in the multivariate. Some special multivariate aspects will be given later. In a pure natural calibration situation we may of course argue that we should predict t by regression of t on u. But many situations are not so clear-cut.

In practice the training sample is typically more or less systematically selected ("controlled calibration"). However, this does not speak against the t on u prediction, if only as usual the training sample is more wide-spread than the natural population.

An argument for the u on t point of view is that it is less sensitive against future outliers in t, since the estimator (but not the predictor) is approximately unbiased irrespective of t. However, that is not a strong argument, since we should avoid estimating future t-values outside the region of the training data.

Having precise measurements in chemistry in mind I am personally inclined to demand that

both types of procedure should be able to yield similar results, else we have reason for scepticism. This point of view is not too demanding in the univariate case, where typically the estimator and the predictor differ little in the region of the training data, but is less clear-cut in the multivariate case, where there is neither one single estimator, nor one predictor, and different ones may agree in more or less important subspaces. One exception that must be allowed in the multivariate case ($q > p$) is when t can be fitted by a linear function in u, but not vice versa, cf. the example of section 3.

1.4. A brief recapitulation of some univariate calibration

The classical protocol was set by Eisenhart (1939). We think of a controlled calibration experiment and a linear regression of the measured u on the true t. In conventional regression notations of y and x we assume a simple linear regression of y on x, and the unknown x is regarded as a parameter. We estimate the regression parameters by least squares on the training data. The unknown x is estimated by solving for x in the fitted regression, with the result

$$\hat{x} = \frac{y - \hat{\alpha}}{\hat{\beta}}. \tag{1.1}$$

This is the MLE under normality. Due to the random $\hat{\beta}$ in the denominator, $\hat{x}$ does not even have an expected value under normality, but it has an approximating normal distribution with the right mean, and variance

$$\frac{\sigma^2}{\beta^2}\left\{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right\}, \tag{1.2}$$

where $\overline{x}$ and S_{xx} denote the calibration sample mean and centered sum of squares in x, respectively.

Krutchkoff (1967) started a famous controversy by proposing the "inverse" procedure: regress t on u and then simply insert the observed u in this fitted regression to obtain an estimate of t directly. His motivation was based on empirical MSE, but his procedure is strongly justified as a matter of principle by regarding all t (training and unknown) as randomly sampled from the same (normal) population of t-values, that is a natural calibration situation. As a matter of fact, if t is random and u has a linear regression on t, (t, u) has a joint (normal) distribution, and the best linear predictor of t is given by the linear regression of t on u, which we estimate from the calibration sample.

2. Multivariate calibration reviewed: the estimation approach (indirect regression)

2.1. The classical and other estimators

The basic statistical model is a multivariate regression model, that is we have q multiple regressions whose error terms are allowed to be correlated. In vector form we write

$$y = \alpha + B^{\mathrm{T}}x + \varepsilon, \tag{2.1}$$

where $\alpha(q \times 1)$, B (full rank $p \times q$) and Γ (non-singular $q \times q$) are unknown parameters, and the error vector ε is $\mathrm{N}(0, \Gamma)$.

The special multivariate case of $\dim y = \dim x$ ($q = p$) is a quite straightforward extension of the univariate procedure. When we have fitted the p multiple regressions of y on x (one for each component of y) and got (in vector form)

$$y = \hat{\alpha} + \hat{B}^{\mathrm{T}} x,$$

we solve for x to obtain

$$\hat{x} = (\hat{B}^{\mathrm{T}})^{-1}(y - \hat{\alpha}). \tag{2.2}$$

The many theoretical problems in multivariate calibration appear when $\dim y > \dim x$ ($q > p$). In the estimation approach the model is then a curved exponential family, and to some extent this explains the problems. When $q > p$ there is no proper inverse of $\hat{B}^{\mathrm{T}}$. We may use ordinary least squares with the estimated $\hat{B}$ and $\hat{\alpha}$ to derive

$$\hat{x}_{\mathrm{LS}} = (\hat{B}\hat{B}^{\mathrm{T}})^{-1}\hat{B}(y - \hat{\alpha}). \tag{2.3}$$

More efficient is generalized least squares:

$$\hat{x}_{\mathrm{GLS}} = (\hat{B}\Gamma^{-1}\hat{B}^{\mathrm{T}})^{-1}\hat{B}\Gamma^{-1}(y - \hat{\alpha}), \tag{2.4}$$

where Γ stands for $\mathrm{var}(y)$ but is also proportional to $\mathrm{var}(y - \hat{\alpha}) = \mathrm{var}(y) + \mathrm{var}(\hat{\alpha})$.

It may be noted that while both estimators LS and GLS are weighted averages of the corresponding univariate estimators of type (1.1), the weights of GLS need not all be positive. This means that the multivariate estimator can fall outside the range spanned by the univariate ones. An illustration with real data is given in the rhinoceros example below. In the different context of multivariate tests for multiple end points in clinical trials, a corresponding GLS test statistic has the same property, and for this reason has been declared unappealing (Follman, 1995).

Γ will typically be unknown, but we may insert an estimate of it, e.g. the residual covariance matrix $\hat{\Gamma}$ from the training data, to obtain what we will call the estimated generalized least squares estimator,

$$\hat{x}_{\mathrm{EGLS}} = (\hat{B}\hat{\Gamma}^{-1}\hat{B}^{\mathrm{T}})^{-1}\hat{B}\hat{\Gamma}^{-1}(y - \hat{\alpha}). \tag{2.5}$$

This is the classical estimator proposed and used already by Williams (1959, ch. 9). Note that the $q \times q$ matrix $\hat{\Gamma}^{-1}$ will exist with probability one if only the sample size n is not too small ($n > p + q$). Under this condition problems with a singular or near-singular $\hat{B}\hat{\Gamma}^{-1}\hat{B}^{\mathrm{T}}$ are not likely in practice, because they should not occur unless either the design x selection failed, or the response vector y lacks information about part of x.

The classical estimator is not ML (unless $q = p$). Heuristically this is because only a p-vector part of the q-vector y is used for estimation of the unknown x, and the remaining $q - p$ components contain some (little) information about Γ. The MLE is not only quite complicated, it also has properties such as expanding the classical estimator, which appear to make any possible gain by ML doubtful (Brown & Sundberg, 1987).

Mean and variance of $\hat{x}_{\mathrm{EGLS}}$ exist as soon as n is large enough and q is large enough ($q > p + 1$ for the mean and $q > p + 2$ for the variance; Nishii & Krishnaiah, 1988). However, as in the univariate case this existence is not crucial, if we only discuss bias and variance as quantities in an approximating normal distribution. Consider the case $q = p$, for simplicity. The reason behind the non-existence of moments is that the normally distributed $\hat{B}$ may be arbitrarily close to singular. If data happen to yield an almost singular $\hat{B}$, however, which essentially means that some linear form in x cannot be estimated, we would not simply use the estimator (2.2) uncritically, but try to find out what went wrong in the calibration.

Corresponding to the factor within braces in (1.2), let

$$c_n^2(x) = \left\{1 + \frac{1}{n} + (x - \bar{x})^{\mathrm{T}} S_{xx}^{-1}(x - \bar{x})\right\}. \tag{2.6}$$

Well approximating variance matrices for the LS and EGLS estimators (2.3) and (2.5) and for the idealized GLS estimator (2.4) are given as follows (Sundberg, 1996):

$$\mathrm{var}(\hat{x}_{\mathrm{LS}}) \approx c_n^2(x)(BB^{\mathrm{T}})^{-1}B\Gamma B^{\mathrm{T}}(BB^{\mathrm{T}})^{-1}, \tag{2.7}$$

$$\mathrm{var}(\hat{x}_{\mathrm{GLS}}) \approx c_n^2(x)(B\Gamma^{-1}B^{\mathrm{T}})^{-1}, \tag{2.8}$$

$$\mathrm{var}(\hat{x}_{\mathrm{EGLS}}) \approx \frac{n-p-2}{n-q-2}\, c_n^2(x)(B\Gamma^{-1}B^{\mathrm{T}})^{-1}. \tag{2.9}$$

The factor $(n-p-2)/(n-q-2) \geqslant 1$ is the cost we have to pay for estimating a completely unknown weights matrix Γ^{-1}. This cost can be high, if q is relatively large, and for instance it may be higher than the cost paid in $\hat{x}_{\mathrm{LS}}$ for not weighting at all, the effect of which is seen to go into the matrix factor of the variances. It should also be noted that when the variance (2.9) is to be estimated, one more correction factor of the same magnitude, $(n-p-1)/(n-q-1) \geqslant 1$, must be inserted when Γ^{-1} is replaced by $\hat{\Gamma}^{-1}$ in the matrix part of (2.9).

Thus, if we want to collect many different measurements y to form a combined (weighted) estimator $\hat{x}$ of x, we must find an alternative way to construct their weights. Since problems arise because of high dimension of y it is natural to try dimensional reduction methods like factor analysis and principal components analysis. Næs (1986) used factor analysis and fitted a latent structure model of form

$$\Gamma = UU^{\mathrm{T}} + D, \tag{2.10}$$

where U $(q \times r)$ contains only a small number r of columns and D is a diagonal (pure noise) variance matrix. In chemical applications the latent factor may have a natural interpretation as an interfering but unmeasured and uncontrolled constituent of the chemical sample. P. J. Brown (1992, personal communication) made a PCA on $\hat{\Gamma}$ and retained the larger eigenvalues only, i.e. he fitted a structure of type $\Gamma = P\Lambda P^{\mathrm{T}}$ for a P with r orthogonal columns and a diagonal Λ. Then, like in PCR, he used $P\Lambda^{-1}P^{\mathrm{T}}$ for Γ^{-1} in $\hat{x}_{\mathrm{GLS}}$, (2.4). Another possibility is to form a ridge regression type estimator by adding a diagonal matrix δI_q to the $\hat{\Gamma}$ in $\hat{x}_{\mathrm{EGLS}}$, (2.5). The dimension r and the ridge constant δ, respectively, could for instance be chosen by leave-one-out cross-validation. Note that these estimation methods have the advantage of being applicable even if there are more variables than observations so that the residual variance matrix $\hat{\Gamma}$ is singular.

Finally we must discuss the possibility of using as estimator a statistic constructed as a predictor. The estimated best linear predictor, to be denoted $\tilde{x}_{\mathrm{EBLP}}$, is obtained as the fitted multivariate regression of x on y (i.e. for each component of x a multiple regression on the q components of y; see also (3.1) below). Assuming all regression parameters precisely estimated, so that the method need not be called "estimated", Sundberg (1985) compared the efficiency of $\tilde{x}_{\mathrm{EBLP}}$ and $\hat{x}_{\mathrm{GLS}}$ in terms of MSE. Regarded as an estimator, $\tilde{x}_{\mathrm{EBLP}}$ is biased, but of smaller variance than $\hat{x}_{\mathrm{GLS}}$, in such a way that $\tilde{x}_{\mathrm{EBLP}}$ has the smaller MSE within a remarkably large region in x, covering most of the calibration sample points. The difference is typically quite small, however, and the important message is that we dare use $\tilde{x}_{\mathrm{EBLP}}$ (and other predictors) as estimator of x. For $p = 1$, Oman & Srivastava (1996) extended this comparison to incorporate uncertainty in the regression parameters. They found that for finite samples $\tilde{x}_{\mathrm{EBLP}}$ had a smaller MSE than $\hat{x}_{\mathrm{EGLS}}$ slightly more often than the asymptotic result indicated, so we are even more on the safe side when using $\tilde{x}_{\mathrm{EBLP}}$.

When there are more variables than observations, $q \geqslant n$, neither the EGLS nor the EBLP is unique. Interestingly however, in this case they are identical in the sense that they are restricted to the same $(q+1-n)$-dimensional hyperplane (Sundberg & Brown, 1989). One way to define

a (quasi-)unique estimator in this case is to choose the minimum length version of EGLS = EBLP in some selected metric. This estimator is tried in an example of the paper mentioned, but see also section 3 below.

Another way to cope with too many y-variables is to select a subset of them. For example, Brown (1982) suggests using a "test of additional information" for such selection. Brown *et al*. (1991), see also Brown (1993, ch. 7), propose a simple method for response selection in multivariate linear models for spectral data, i.e. for wavelength selection. The criterion used is the signal-to-noise ratio for the different wavelengths. In their examples they successfully select a few per cent of some thousand wavelengths. Brown (1992) allowed AR type correlation between neighbouring wavelengths. In section 3 we consider a different example of spectral data for which their method totally failed, because of lack of linearity for individual wavelengths.

2.2. A non-linear example

We will illustrate some new and some previous points when we now present a biological application, of rhinoceros age estimation from horn lengths (Clarke, 1992). As in many other calibrations for age estimation, relationships are clearly non-linear. Calibration data come from $n = 12$ animals of known age x, most of them possessing two horns. Clarke used the bivariate model

$$y_1 = \alpha - \beta_1\gamma_1^x + \varepsilon_1 \quad y_2 = \alpha - \beta_2\gamma_2^x + \varepsilon_2$$

(with positive parameter values) for anterior and posterior log-transformed horn lengths, respectively. Note that the two functions have the parameter α in common, the asymptote as $x \to \infty$. In this case a GLS procedure may also be used in the calibration phase, with an iteratively estimated Γ, or the closely related ML method under normality, see e.g. Seber & Wild (1989, ch. 11) for an account of estimation in multiresponse non-linear models.

Figure 1 shows calibration data and fitted curves, but also horn lengths for two animals of unknown age, one young and one old individual. Univariate and multivariate (GLS) age estimates are indicated, all based on the fitted curves. We can note the following features.

The specification of the model is often much more crucial in non-linear than in linear calibration, in this example the assumption of a common asymptote in particular. Long horns provide rather little information about age, and horn lengths above the fitted asymptote do not even yield finite age estimates. In this case we are extremely lucky with both animals, that the univariate estimates agree so well.

The older animal provides an illustration of the feature mentioned in the previous section, see below formula (2.4), that the GLS estimate, regarded as a weighted combination of the univariate estimates, need not have weights between 0 and 1. We see that the GLS estimate falls to the right of both univariate estimates. The actual difference is slight, but could have been drastic in this non-linear case. If we imagine the anterior horn length reduced, so that the corresponding univariate age estimate is reduced, the GLS estimate would increase. For example, if the anterior horn had been of the same length as the posterior, $y_1 = y_2 = 5.90$, which seems not to be an extreme event, the univariate estimates would have been 6 and 19 years. With such discrepant age estimates, the GLS estimate falls far to the right of both of them, at 27 years. Or like one of the animals actually observed, with both y-values being 6.00 the univariate estimates are 7 and 22 years, whereas GLS yields 30. This may seem bizarre, but originates from the positive correlation in the Γ matrix, that makes GLS prefer an age estimate yielding residuals of equal sign.

Another type of non-linearity that requires special attention in calibration is when y is linear

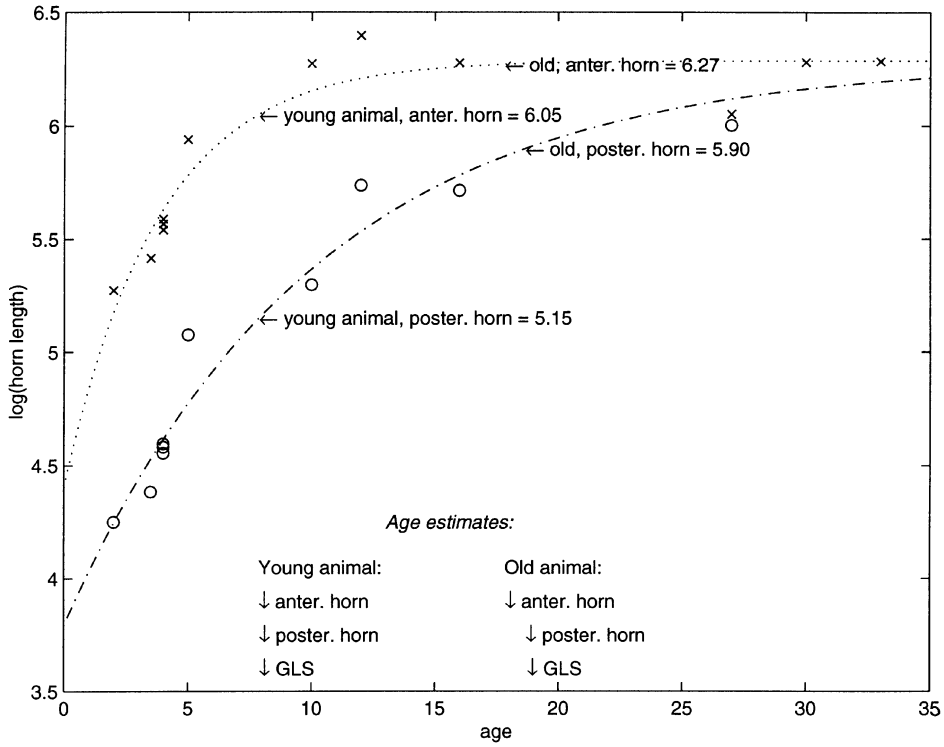


Fig. 1. Age estimation for rhinoceros, an example of bivariate calibration, from Clarke (1992). Fitted curves $y_1 = 6.287 - 1.885 \cdot 0.768^x$ and $y_2 = 6.287 - 2.496 \cdot 0.905^x$. Estimated residual variances 0.0185 and 0.0271, with covariance 0.0111.

in the parameters but non-linear in x, as exemplified by a polynomial in x. Regarded as a multivariate linear model the polynomial induces non-linear constraints on the x-vector. This situation has been studied by Brown & Oman (1991), see also Brown (1993, ch. 7). The complication is that the criterion function to be minimized in the GLS procedure may have several local minima. In particular this may lead to confidence regions formed by disjoint intervals. However, confidence regions is the topic of the next section.

2.3. Confidence regions

As we have seen above, several estimators of x are possible, and each of them has its merits. If we want to go one step further, from an estimator and its standard error to a confidence region for x, there is unfortunately no agreement about what region to use even if we restrict ourselves to regions related to the estimated generalized least squares estimator.

Following Williams (1959), Brown (1982) suggested an exact confidence region of form

$$\|y - \hat{\alpha} - \hat{B}^{\mathrm{T}} x\|^2 \leqslant K c_n^2(x), \tag{2.11}$$

where the squared distances $\|\cdot\|^2$ here and for Q and R below are in $\hat{\Gamma}^{-1}$-norm (i.e. $\|y\|^2 = y^{\mathrm{T}} \hat{\Gamma}^{-1} y$), and where K is proportional to an $F(q, n-p-q)$ quantile. This region was found to have undesirable properties, however. It turns out that the left hand side of (2.11) may be written

$$\|y - \hat{\alpha} - \hat{B}^{\mathrm{T}}x\|^2 = \|\hat{B}^{\mathrm{T}}(\hat{x}_{\mathrm{EGLS}} - x)\|^2 + \|y - \hat{y}_{\mathrm{EGLS}}\|^2 = Q + R, \tag{2.12}$$

where $\hat{y}_{\mathrm{EGLS}} = \hat{\alpha} + \hat{B}^{\mathrm{T}}\hat{x}_{\mathrm{EGLS}}$. Here R is an asymptotically ancillary statistic, asymptotically $\chi^2(q-p)$ distributed (as $n \to \infty$), that may be called an inconsistency diagnostic since it can be used to check whether y is so close to $\hat{y}$ in all q components jointly as should be expected according to $\hat{\Gamma}$ (and R was proposed for model checking already by Williams (1959)). Note that for $q = p$, R vanishes.

Now, if $q > p$ and R has a relatively high value, Q must be small in order to satisfy (2.11), that is the region will be narrow, and if R is high enough the region will be empty. In principle it is OK that a confidence region is empty when data do not fit the model, but here the shrinkage of the region with increasing R is misleading when we think of the size of the region as reflecting the precision of the estimation procedure. A number of alternatives without this annoying feature have been proposed.

1. Wood (1982) deliberately neglected R and used Q alone to form an asymptotic confidence region. The finite sample distributions of Q and R have been more closely investigated by Fujikoshi & Nishii (1984) and Davis & Hayakawa (1987).
2. Brown & Sundberg (1987) derived asymptotic profile likelihood based regions, which they found to be expanding with R.
3. Oman (1988) developed a uniformly most accurate translation (but not scale) invariant conservative region, that was constructed also with such linear models in mind which are non-linear in x (e.g. polynomials).
4. Clarke (1992), in his non-linear calibration situation, is primarily using a least-squares-criterion-based interval specified by a weakly motivated F-distribution.
5. Mathew & Kasala (1994) constructed a pivot yielding a fairly complicated but both translation and scale invariant region.
6. Mathew & Zha (1996), inspired by Oman (1988), have derived a (slightly) conservative invariant region, easier to deal with than that of Mathew & Kasala.

The time does not yet, if ever, appear ripe for declaring one region superior to the others.

2.4. Diagnostics

In the previous section we pointed out the use of $R = \|y - \hat{y}_{\mathrm{EGLS}}\|^2$ as an inconsistency diagnostic when $q > p$. A single high value of R ($\gg q - p$) primarily indicates that something is wrong with the corresponding y-vector. In the rhinoceros example it might be asked if this diagnostic would detect anomalies in those y-vectors which gave so extreme discrepancies between the two univariate estimates and between them and the GLS, and the answer is "not necessarily". If $y = (5.90, 5.90)$, then $R = 8.0$, which is the 98% quantile in a $\chi^2(2)$, but with $y = (6.00, 6.00)$ the residuals are much smaller, and $R = 4.4$ only.

Repeatedly appearing high R-values indicate that the parameters of the calibration should not be relied upon. If $q > p$ we obtain some additional residual information from y each time we predict an x, but negative results by Brown & Sundberg (1989) show that there is not much relevant information in these residuals. Rather, a recalibration might be needed.

The norm of R may in fact equivalently be taken to be with respect to the inverse of the estimated variance of the prediction residuals, $\hat{\Gamma}^{-1}(\hat{\Gamma} + \hat{B}^{\mathrm{T}}S_{xx}\hat{B})\hat{\Gamma}^{-1}$. Næs & Martens (1987) showed that with respect to the latter norm,

$$\|y - \tilde{y}_{\mathrm{EBLP}}\|^2 = \|y - \hat{y}_{\mathrm{EGLS}}\|^2 + \|\hat{y}_{\mathrm{EGLS}} - \tilde{y}_{\mathrm{EBLP}}\|^2.$$

where in analogy with $\hat{y}_{\mathrm{EGLS}}$, $\tilde{y}_{\mathrm{EBLP}} = \hat{\alpha} + \hat{B}^{\mathrm{T}}\tilde{x}_{\mathrm{EBLP}}$. The second statistic on the right hand

side is also a useful diagnostic. Under normality it is approximately $\chi^2(2)$ distributed. A high value signals an outlier in x-space, that requires caution.

More discussion about residuals and outlier detection is found in Martens & Næs (1989, ch. 5) and in Brown (1993, ch. 5).

3. Multivariate calibration reviewed: the prediction approach (direct regression)

3.1. Introduction

In "natural calibration" we think of each item or individual as randomly sampled from one and the same natural population, so both the calibration t and the future unknown t are regarded as random. In combination with the regression model for u given t we obtain a joint normal distribution for (t, u), in which we should predict t. The best linear predictor of t is given by the theoretical regression of t on u, one regression for each component of t. Note that even if we do not have a calibration sample from a natural population we might choose to regress t on u, for convenience or efficiency, as mentioned in section 2.1. Thus we discuss prediction in regression models, and we do this in the conventional notations of multiple regression of y on x. Mostly y will be scalar but we will not totally forget about multi-dimensional y, see section 3.8.

Since the theoretical regression of y on x is unknown it is natural to try the empirical regression from the calibration sample,

$$\tilde{y}_{\mathrm{EBLP}} = \overline{y} + S_{xy}S_{xx}^{-1}(x - \overline{x}), \tag{3.1}$$

where S_{xy} and S_{xx} are the sample covariance and variance matrices. When $n > \dim x$, the regression is typically unique. We do not have the problems of the estimation approach of section 2. There is no residual covariance matrix Γ that should be estimated. Ellipsoidal prediction regions, derived in the same way as the confidence regions (2.11), seem less controversial than the confidence regions. Problems do exist, however. We will concentrate on some problems which have a clear bearing on multivariate calibration situations:

(1) collinearity and near-collinearity of regressors; regularization methods;
(2) pretreatment of data: scaling, differentiation, etc.;
(3) smoothness requirements;
(4) cross-validation versus true prediction;
(5) multidimensional response.

3.2. A real data example

We will base the discussion on a calibration example of real data with a large number of potential regressor variables. This is an investigation by Karlsson *et al*. (1995) that aims at in-line monitoring of nitrate content in municipal waste water by spectroscopic methods, without need for filtering and other traditional specimen preparation. Data were collected at several different combinations of treatment plant and basin type during a period of 6 months, giving a total of $n = 125$ specimens of water. Each sample comprises a mixture of daily aliquots taken during the week. For each specimen a spectrum was collected in the UV–visible region, one absorbance value at each second nanometer in the range 190–820 nm, in total 316 values for each specimen. Additionally, the concentrations of nitrate (and some other constituents) were determined by an accredited laboratory using official (traditional) methods.

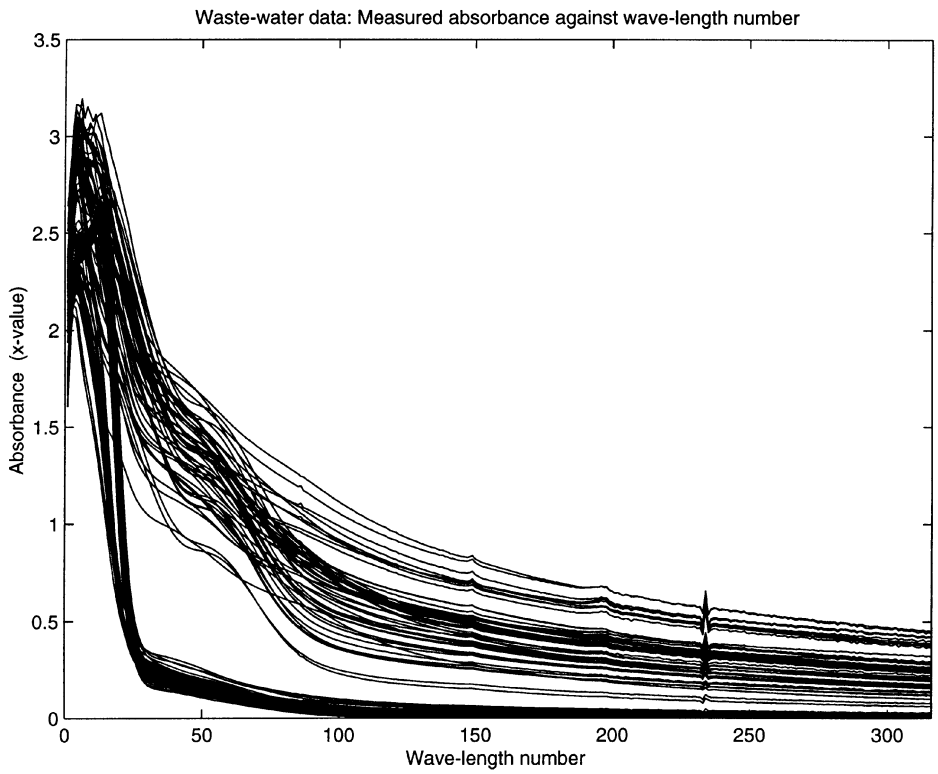


Fig. 2. Waste water data: Absorbance spectra plotted against wavelength number, all 125 specimens, 316 wavelengths.

Figure 2 shows all the 125 absorption spectra plotted as continuous curves (by joining consecutive points). We notice that after a short rise all curves go down more or less steeply and form at least two distinct groups in this respect. This corresponds to a grouping in nitrate values. Figure 3 plots absorbance against nitrate concentration for a representative selection of wavelengths. We see that this dependence is highly non-linear, typically not even monotone, and that there are in fact three groups of data. It is clear that a multivariate linear model for regressing absorbance on concentration would not have much chance of success. The wavelength subset selection method of Brown *et al.* (1991) mentioned in section 2.2, that also assumes individual linearity, is of no value; in fact it selects the uninformative right hand part of the spectrum. However, this does not exclude the possibility that some linear combination of absorbance values could form a good description of and a good predictor for nitrate. In fact, we will see that this works to some extent, when we look for a regression of nitrate y on the absorbance spectrum x.

3.3. Collinearity and near-collinearity problems

In the example we have 316 explanatory x-variables but only $n = 125$ observations. Thus we have a large number of exact collinearities in x, just because the sample size is not big enough. The observed x-vectors span only a 125-dimensional Euclidean subspace U_{125} of the

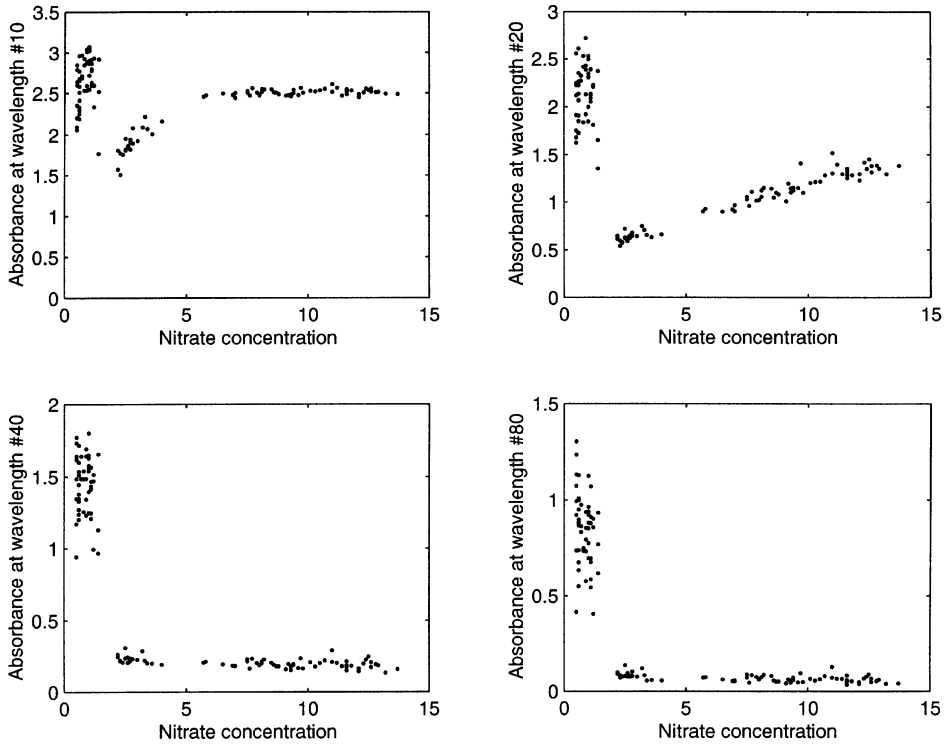


Fig. 3. Waste water data: Absorbance at four different wave-lengths plotted against nitrate concentration.

potential U_{316}. Above this we have near-collinearities, meaning that the 125 x-vectors nearly fall in a smaller-dimensional hyperplane of U_{125}. We return to near-collinearities shortly.

In traditional regression theory we would think of a unique true regression function $\beta^{\mathrm{T}}x$ that we wanted to determine (estimate). This task would not be possible here since our x-vectors are confined to U_{125}, implying that an infinity of linear functions fit data perfectly. Our aim is a different one, however, being "only" to find a good predictor for y. The idea is now that the x-variables do not vary independently, but are very much correlated, and that all essential sources of variation for x should show up in U_{125}, including those connected with the explainable part of the variation in y. In the formation of a predictor for y we thus simply assume that directions of variation in x orthogonal to those represented in the data are uncorrelated with y, and we immediately obtain a unique least squares predictor. This corresponds to the minimum length (or minimum norm) least squares estimator of β, calculated by using the Moore–Penrose generalized inverse of S_{xx} in (3.1).

When we try minimum length LS in our example we find that the regression coefficients oscillate wildly with the wavelength at an extremely high amplitude (except in the very left, most informative part of the spectrum), see Fig. 4. The corresponding predictor is a terribly bad one as judged from cross-validation. The main reason behind this is the additional presence of near-collinearities.

A discussion of near-collinearity is somewhat more intricate. If a linear function $z = c^{\mathrm{T}}x$ in x is nearly constant over the observations, the influence of z on y is not impossible to estimate, as it was for exact collinearities, but can only be very imprecisely estimated, since z varies so little. Thus we could argue either, as above, that it would be better to let z have no influence on our

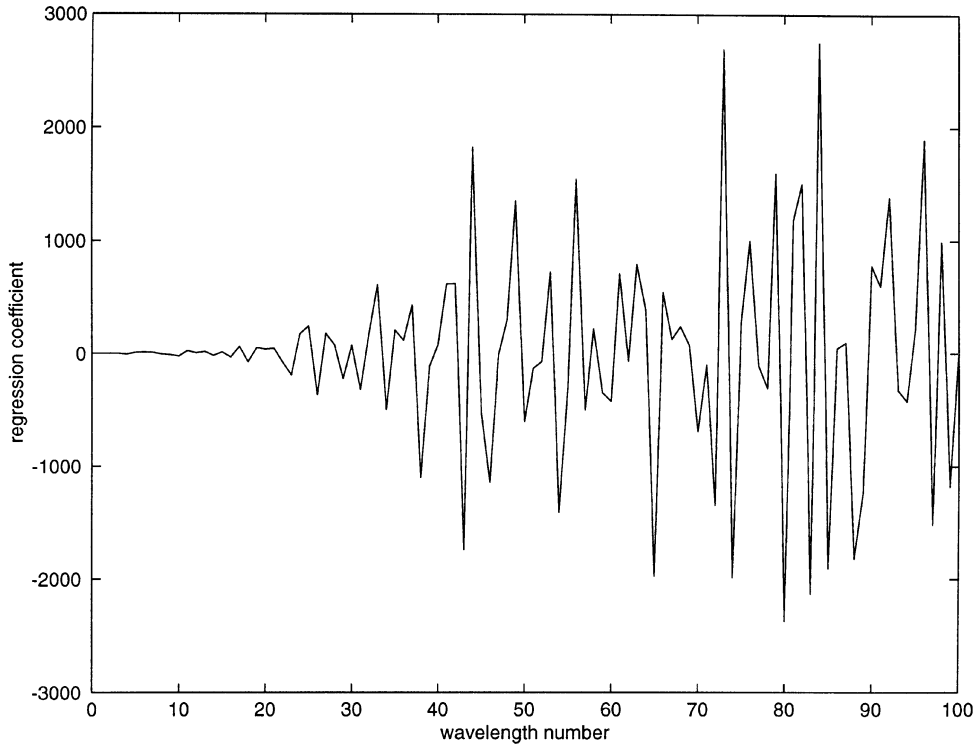


Fig. 4. Minimum length least squares (MLLS) regression of nitrate on absorbances at first 100 wavelengths, regression coefficients plotted against wavelength number. Note scale on vertical axis.

predictor (cf. PCR), or less extremely) that we should down-weight or "shrink" the LS estimate with respect to z, that is in the c direction, little or much depending on how near we are an exact collinearity, and perhaps also on how little correlated z is with y. In this way we trade bias for variance. Methods based on this idea also go under the name of regularization methods or shrinkage methods, and there are many variants, see the next section.

A complication passed over until now is the problem of lack of scaling invariance. If we apply a non-orthogonal transformation on x, for example individual scaling of the components of x, in the presence of exact or near-collinearities, a predictor formed as minimum length LS or by a regularization method will not remain invariant. This is because what is "minimum length" depends on how we measure distances, and likewise "near" in near-collinear is metric dependent. This is an intrinsic ambiguity of the regularization methods to be discussed below. It sometimes appears in disguised form, for example when Garthwaite (1994) suggests an "alternative weighting" in PLS, which in fact simply corresponds to using the metric of variance-standardized x-variables.

When there is not a single natural metric to choose, we could surpass the problem to some degree by trying several metrics and comparing prediction efficiencies by, for example, cross-validation. However, going to the extreme and letting the metric be completely free does not seem a good idea, at least not when there are 316 x-components but only 125 observations. We return to the scaling problem in section 3.5.

3.4. Counter-measures to near-collinearity

Ridge regression (RR), principal components regression (PCR) and partial least squares or projection to latent structures (PLS) regression have been popular regularization methods that have usually worked well, for a long time. For example, the book by Martens & Næs (1989) is dominated by the PLS and PCR methods.

In PCR, regressors $z = c^{\mathrm{T}}x$ (with $|c| = 1$) are not selected for their correlation $R(z, y)$ with y, but only for their own variances (small variance = near-collinearity). This can work because we are after linear forms that are not only correlated with y but also have a significant variation of their own. PLS is a compromise between OLS and PCR that brings in the correlation with y to a certain degree. This is done by successively selecting mutually orthogonal regressors (factors) to maximize covariance with y, instead of variance, as for PCR, or correlation, as for OLS. In all procedures we apply simple OLS on the set of orthogonal regressors constructed.

In its early days PLS was only algorithmically defined, based on the intuition of H. & S. Wold, but regarded with scepticism by most statisticians. By clarifying works of Frank (1987), Helland (1988, 1990), Höskuldsson (1988), Stone & Brooks (1990), Garthwaite (1994) and others it is now much less mysterious. In particular, Helland (1988) derived a non-recursive characterization of PLS, (of higher interest from a theoretical than computational point of view), saying that the first r PLS-regressors can equivalently be chosen as $z_i = c_i^{\mathrm{T}}x$ with

$$c_i = S_{xx}^{i-1}s_{xy},$$

$i = 1, \ldots, r$. In comparison with PCR, PLS does not only stress the relationship between regressor and response, but it is also computationally much faster than conventional PCR, and this can be essential in large cross-validatory studies. Faster techniques for PCR have recently been proposed, however; see Mertens *et al*. (1995).

Stone & Brooks (1990) introduced continuum regression (CR), where OLS, PLS and PCR all naturally appeared as special cases, corresponding to the different maximization criteria: correlation, covariance, and variance, respectively. In continuum regression a regressor z is selected to maximize the criterion function

$$R^2(z, y)V(z)^{\gamma} \tag{3.2}$$

for a suitable value of $\gamma \geqslant 0$, followed by OLS of y on the CR regressor(s). This yields the first factor; subsequent factors z are constrained to be orthogonal to all previous factors. In CR as a statistical method the value of γ and the number of factors are chosen by optimizing a cross-validation criterion, or similarly. For $\gamma = 0$ we obtain OLS, $\gamma = 1$ represents PLS, and PCR is approached as $\gamma \to \infty$. Furthermore it was shown by Sundberg (1993) that first factor CR can be interpreted as a scaled RR, more precisely using the RR type estimator

$$\hat{\beta}_{\mathrm{RR}} = (S_{xx} + \delta I)^{-1}S_{xy} \tag{3.3}$$

for some ridge constant δ, to form a regressor $z = \hat{\beta}_{\mathrm{RR}}^{\mathrm{T}}x$, followed by simple least squares of y on z. We call this "least squares ridge regression", LSRR. The LS scaling of the ridge estimator will compensate for such shrinkage in RR that is not motivated by near-collinearity. For example, in an orthonormal design first-factor CR does not shrink at all, but is identical to OLS for any value of γ. The scaled ridge predictor is generally less sensitive to the choice of ridge constant than conventional RR. This can be explained by the overall shrinkage of RR, increasing with δ. In particular, RR shrinks to zero as $\delta \to \infty$, whereas LSRR approaches first-factor PLS.

It was pointed out by de Jong & Farebrother (1994) that negative δ-values could be allowed in

(3.3), and that the interval $\delta < -\lambda_{\max}$, $\lambda_{\max}$ being the maximal eigenvalue of S_{xx}, represents CR between PLS and PCR, that is $\gamma > 1$. Thus all of CR(1) is covered by LSRR. Actually the opposite is not always true, as demonstrated by Björkström & Sundberg (1996), so CR is more restrictive. This is connected with the special form of the criterion (3.2) of Stone & Brooks (1990). Björkström & Sundberg (1999) show that the optimal regressor is of RR type not only for (3.2) but for *any* reasonable criterion function involving only $R^2(z, y)$ and $V(z)$, and they therefore advocate LSRR as being more fundamental than the CR as based on criterion (3.2).

Figures 5a–d show the estimated regression coefficients plotted against wavelength for a selection of PCR, PLS and LSRR predictors. Note how oscillation amplitudes increase with increasing number of factors or decreasing ridge constant, and how similar the results look for the three procedures. In particular, Fig. 5d shows PCR, PLS and LSRR jointly when the number of factors or the ridge constant has been chosen such that the highest peak should be of the same height for all three procedures.

Hence, all the methods mentioned are tied together. Is there a best method? The present author's opinion and experience from data analysis is that typically all these methods have about the same predictive ability, which is often also close to the limiting true error variance σ^2 of y. This conjecture however does not exclude the existence of data sets for which one method or the other comes out unfavourably. Further support for the approximate equivalence is found for example in the extensive simulation study by Frank & Friedman (1993), who conclude that "RR, PCR and PLS have similar properties and give similar performance", and that "the actual solutions given by the three methods on the same data are usually quite similar".

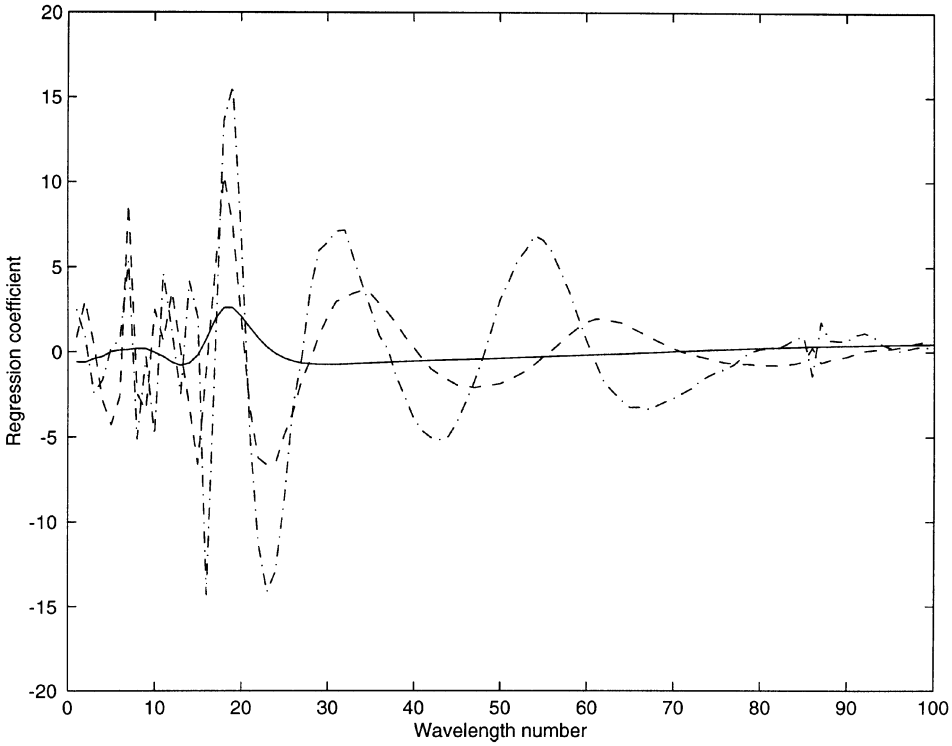
Leave-one-out cross-validation is a natural procedure for studying the predictive ability. In order to judge goodness of prediction in cross-validation leave-one-out comparisons there are several equivalent measures:

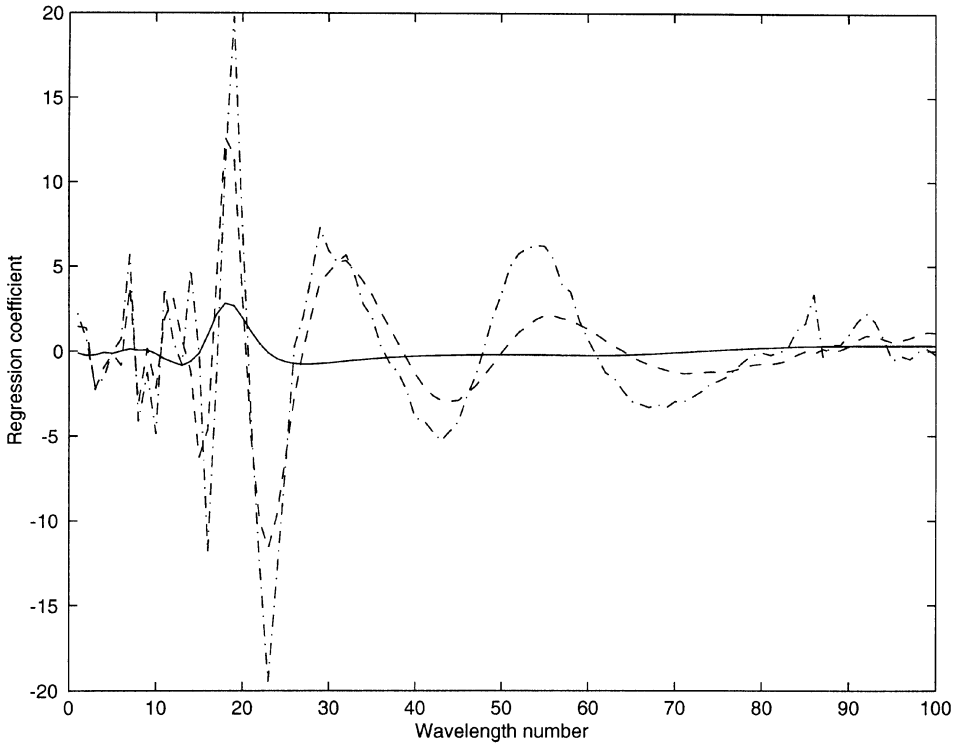
$\text{PRESS} = \sum_1^n (y_i - \hat{y}_{\setminus i})^2$	(Prediction Sum of Squares)
$\text{MSEP} = \text{PRESS}/n$	(Mean Squared Error of Prediction)
$\text{RMSEP} = \text{MSEP}^{1/2}$	(Root Mean Squared Error of Prediction)
$I = 1 - \text{PRESS}/\sum_1^n (y_i - \overline{y}_{\setminus i})^2$	(CV-index).

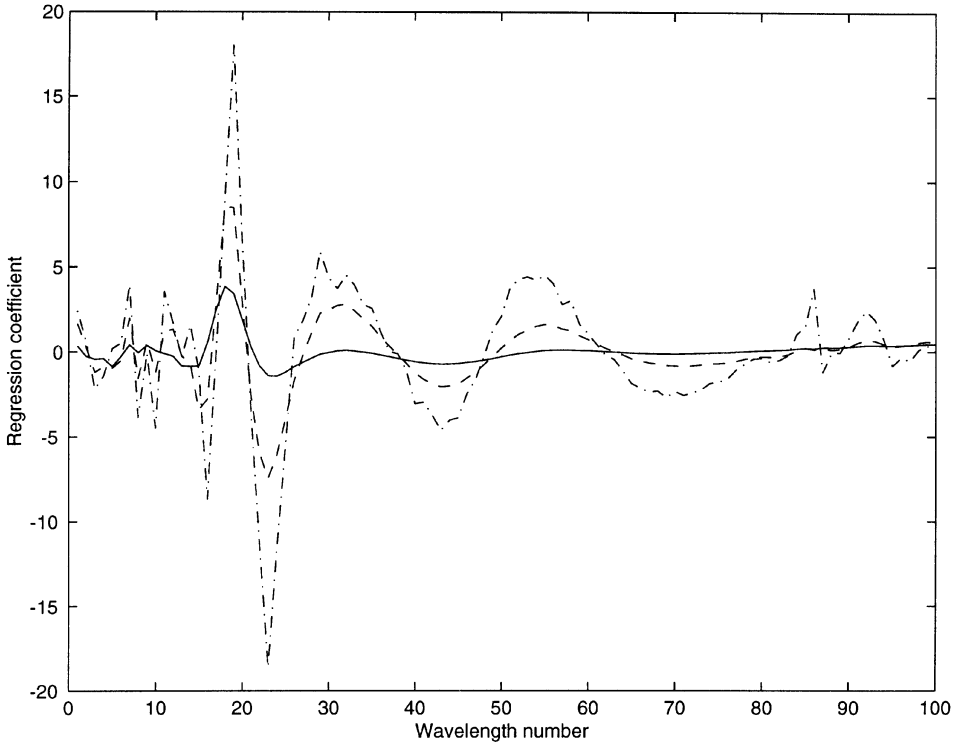
Here $\hat{y}_{\setminus i}$ denotes the prediction of y_i when observation i is excluded from the regression, and $\overline{y}_{\setminus i}$ the mean value without observation i. With separate calibration and validation samples, analogous prediction measures are formed by summing and averaging over the validation sample.

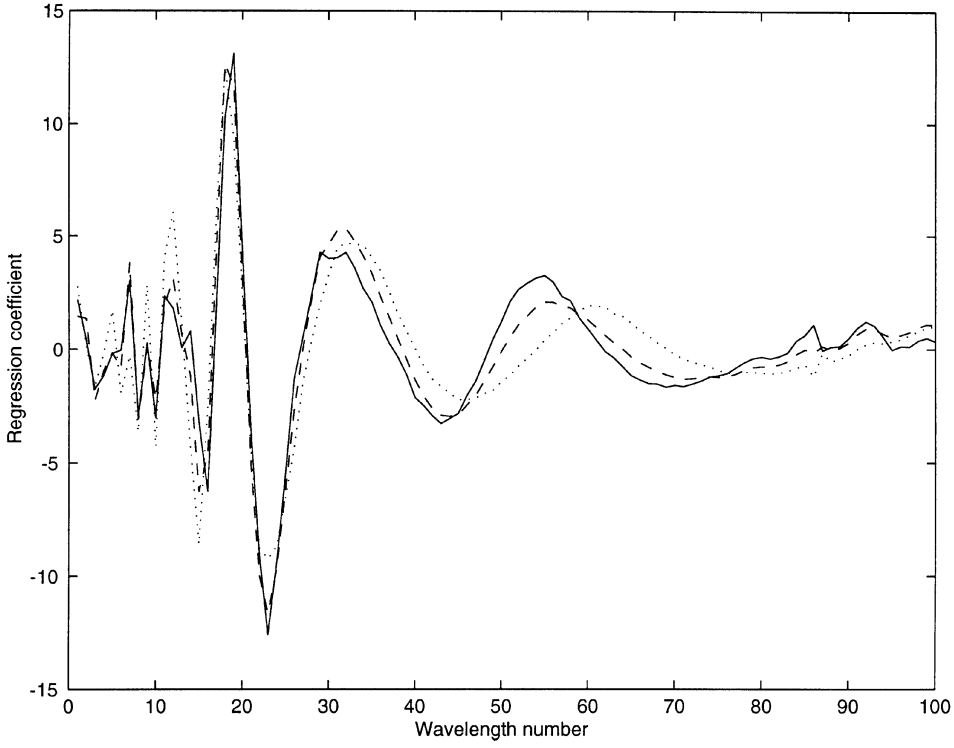
For examples, Stone & Brooks (1990) used the CV-index for their comparisons, whereas Martens & Næs (1989) and Brown (1993) prefer the MSEP or RMSEP criteria. We will here use the MSEP measure, or its logarithm. Figures 6a–b show the MSEP values for PLS and PCR with varying number of factors and for LSRR and RR with varying ridge constant. As usual PCR comes down slower than PLS but stays down longer, with increasing number of factors. This is natural since PCR would also include factors totally unrelated with y if they only have large enough variance, and only the small variance factors are dangerous. The difference between LSRR and

Fig. 5a–d. (Overleaf) Regressions of nitrate on absorbances at first 100 wavelengths, regression coefficients plotted against wavelength number, number of factors or ridge constant varied. (a) PCR for three different numbers of principal components. PCR(8): solid line; PCR(16): dashed line; PCR(24): dash-dot line. (b) PLS for three different numbers of PLS-factors. PLS(5): solid line; PLS(10): dashed line; PLS(15): dash-dot line. (c) (LS)RR for three different ridge constants. (LS)RR(0.1): solid line; (LS)RR(0.01): dashed line; (LS)RR(0.001): dash-dot line. (d) PCR(20), PLS(10) and (LS)RR(0.003), selected to have about the same amplitude of their highest peaks. PCR: dotted line; PLS; dashed line; (LS)RR: solid line.









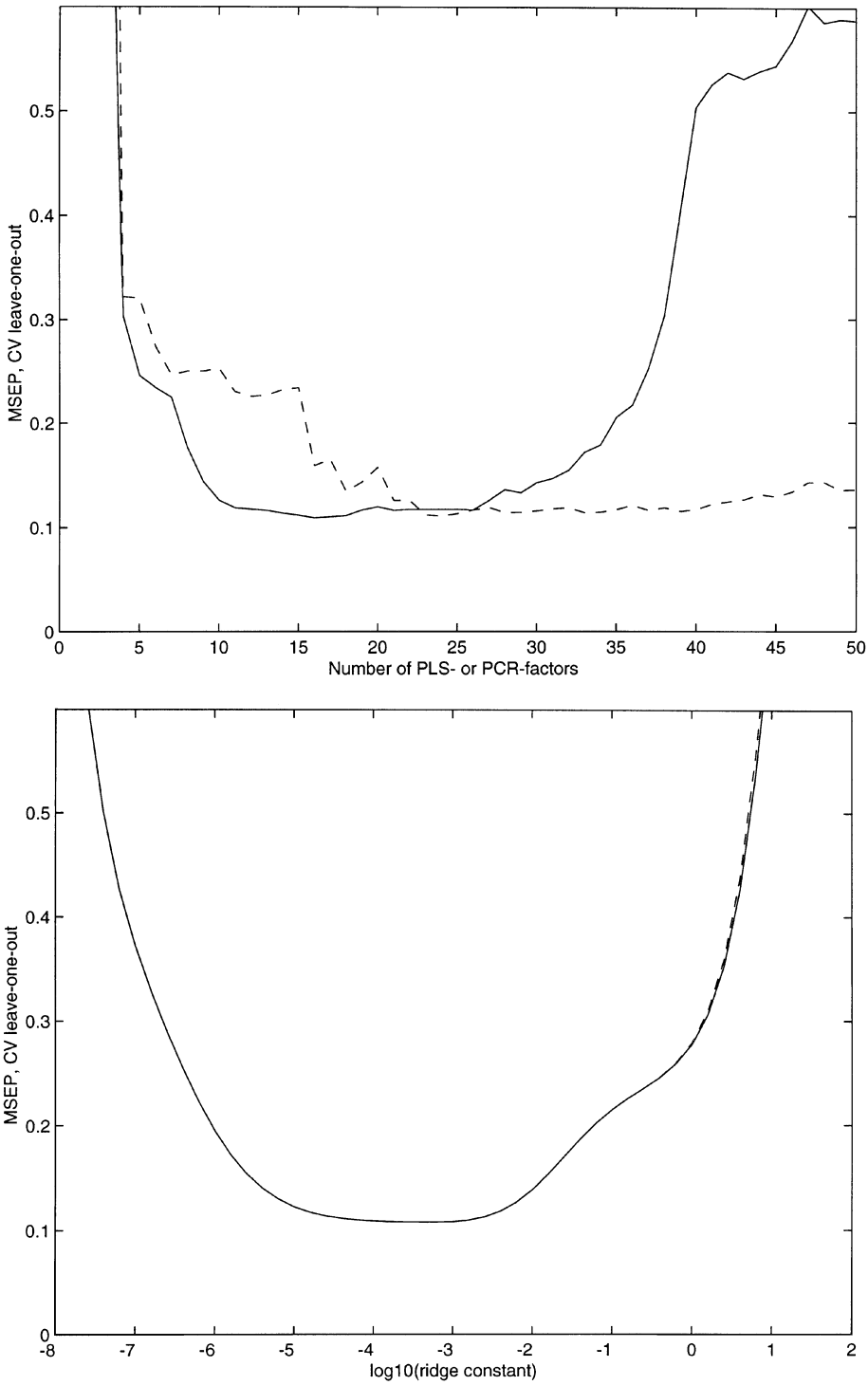


Fig. 6a–b. Regressions of nitrate on absorbances at first 100 wavelengths, CV leave-one-out MSEP values. (a) PLS and PCR for varied number of factors. PLS: solid line; PCR: dashed line; (b) LSRR and RR for varied ridge constant. LSRR: solid line; RR: dashed line.

RR is hardly noticeable within this range in this example. Note that all these MSEP curves have minimum values of 0.11 (within a range of 0.003). As an idealized reference value, $\hat{\sigma}^2 = 0.053$ when estimated from residuals in a full model with 125 observations and 100 x-variables (first 100 wavelengths). As a further reference value, OLS yields an MSEP value of 2.50.

We end this section by briefly mentioning a few other regression methods mentioned in literature. A predictor particularly intended for calibration situations was proposed by Næs (1985). The idea is the same as in his 1986 paper, described in (2.10) above, to reduce effective dimension by factor analysis (FA) modelling of Γ, the noise covariance matrix in multivariate regression of u on t. Here the FA estimate of Γ is used in a replacement of S_{xx} by $\Gamma + s_{xy}s_{yy}^{-1}s_{yx}$. A different idea is behind latent root regression (LRR), see Gunst (1983) for introduction and references. In LRR an eigenvector–eigenvalue analysis of the joint (x, y) variance–covariance matrix is undertaken, in order to sort out and eliminate "non-predictive near-collinearities". LRR seems not to have been tried much in later years, but in a recent comparison with PLS and PCR on NIR-spectra, based on a test set and a validation set, LRR was said to predict as well as the other methods and with less risk for overfitting (Vigneau *et al*., 1996). LRR is related to total least squares (orthogonal regression), see van Huffel & Vandewalle (1991), and in this way to errors-in-variables models.

3.5. Pretreatment of data

We have already stressed that regularization methods are not scale invariant. Typically the regressors are individually (auto)-scaled to unit variance, or at least that is often recommended. This is natural when all regressors are on incomparable scales, for example in QSAR regression studies, when the x-variables represent molecular properties of quite different characters. In spectroscopy, however, when the x-variables are intensities at different wavelengths, it seems more reasonable not to auto-scale. Scaling would reduce the weights of the informative wavelengths, at which the intensity varies much with concentration, and increase the weights of the uninformative wavelengths, at which the intensity varies little. Accordingly, the waste-water data of section 3.2 should not be scaled before regression. This was confirmed empirically from cross-validation performance by Karlsson *et al*. (1995), who found that auto-scaling increased the MSEP by a factor of about 10.

PLS- and PCR-users sometimes argue that they need not mean-centre the variables. We have described PLS and PCR as methods for estimating the coefficient vector β in a model with intercept, and in all methods discussed above the intercept is estimated by $\overline{y}$ at the point of gravity $\overline{x}$. If we consider x as varying around 0 instead of around $\overline{x}$, the components of x will appear to be more strongly correlated. This correlation will be taken up by the early PCs or PLS-factors, and allowing for some more factor we could expect approximately the same predictors to be formed. For support of intuition, remember that a hyperplane can always be embedded in a subspace of dimension one more. However, I cannot see any point in refraining from the mean-centring when the model should have an intercept.

In some types of applications, in particular near infrared (NIR) analysis, spectra can show a large overall influence from particle size, particle form and other irrelevant features of the specimens. Multiple scatter correction (MSC) is a preprocessing method for the spectra that has sometimes proven useful for the reduction of such variability between specimens. In MSC each spectrum is individually normalized, after subtraction of its mean (over wavelengths). The normalizing scalar is the regression coefficient for the regression of the individual absorbances on the corresponding means over all spectra, that is a measure of the relative overall intensity of the individual spectrum. Helland *et al*. (1995) provides a more detailed discussion and comparison of MSC and related methods. However, MSC will also to some extent reduce the

information provided by data about the constituents varied, so it is a risky procedure that must be used with judgement. Karlsson *et al*. (1995) tried it on the waste-water data but found a slight loss of efficiency.

Another preprocessing that might be applied in the same type of situations is second-order differencing. This will effectively filter away slow gradients in the spectra, but not so much sharp peaks. Second differencing is a linear operation, which implies that it is equivalent to a change in metric corresponding to a linear transformation. A recent paper on mathematical and statistical aspects is Goutis (1998). Karlsson *et al*. (1995) tried this preprocessing on the waste-water data and found that it could yield a slight reduction in MSEP. It has not been used in the illustrations of the present paper.

Some kind of variable selection could be natural for preprocessing, with data like in the waste-water example. Traditional variable selection techniques are hardly of much help. However, it is easily seen that the right hand part of the spectra contribute practically no (additional) information. As a crude preprocessing procedure we therefore kept only the first 100 of the 316 wave lengths. This only slightly increased the predictive ability of the techniques tried, as measured from CV leave-one-out. The reason why it did not make a larger difference is that methods like PLS and LSRR in themselves are good at putting little weight on useless variables.

3.6. Smoothness

The spectroscopic data are discretized continuous curves, and it may be argued that the corresponding regression vector $\hat{\beta}$ should reflect this by forming a reasonably smooth sequence. The standard methods discussed up to here do not regard the vector of x-variables as an ordered sequence. In latent factor procedures like PCR, PLS and CR this smoothness should also hold for the latent factors themselves, i.e. for their sequences of weights-vectors c. Automatic smoothness is in fact typically seen unless we include too many factors in latent factor procedures or choose too small ridge constants. However, there have also been some explicit attempts to favour smoothness.

Denham & Brown (1993) try a regression model where the spectrum of each component substance is modelled by a cubic spline. They conclude that "there seems little gain in applying regression splines as smoothers of the least squares coefficients when viewed against the added complication of the approach". They also fit AR type models to the noise, across wavelengths, and Brown (1993, ch. 6) in a Bayesian approach attempts priors of AR and ARMA type for the covariance structure between wavelengths. Goutis & Fearn (1996) impose a roughness penalty in the construction of PLS-like factors. All of these authors express disappointment that their procedures did not improve the predictive performance as much as they had hoped for. The reason seems to be that the continuity asked for is already inherent in the data to a large extent and will express itself automatically in the most important latent factors. In the waste-water example cross-validation tends to select many PCR or PLS factors, and the last of them are quite noisy. However, as long as they contribute little to the resulting estimator $\hat{\beta}$, their noisiness does not disturb much.

3.7. Cross-validation vs true prediction

For comparison of models intended for prediction it is highly inadequate to look just at model fit. We have used cross-validation leave-one-out above as a conceptually simple way of trying to compare prediction abilities. Alternatively we could have used cross-validations based on other splits of data into data for predictor construction and data for testing the

prediction. In chemometrics cross-validation has been standard error since Wold (1978). However, we must be aware of the limitations of simple cross-validation.

First, note that when regarding for example, PLS as a well-specified method for predictor construction, it must include a specified procedure for choice of the number of factors. If this procedure is based on cross-validation we need another test set or an "outer" cross-validation for judging the behaviour of this method, to avoid a model selection bias, see Stone (1974) or Hjorth (1994, ch. 3).

Second, there is not necessarily a single correct model for all past and future data that must be found. Even if there were one, we could not expect to find it with fewer observations than variables. Nevertheless we can expect the constructed predictor to work well if

(i) the relation between y and x remains the same; and
(ii) future x-vectors are like the calibration ones.

This is precisely when the CV procedure will be fair. However, in the waste-water example we desire a predictor that will still work next month (at least not demanding complete recali-bration), in spite of systematic (e.g. seasonal) and random variation in composition and in other conditions.

We made an effort to mimic this situation by splitting the data in two parts, of 65 and 60 observations respectively, using one part for "internal" CV and the other as a test set for "external" prediction. Only the first 100 x-variables were used in this experiment, from the informative left part of the spectrum, but many enough to have more variables than observations in the calibration.

When the splitting was done at random, leave-one-out CV and the test set gave essentially the same MSEP curves for all procedures tried (LSRR as a function of the ridge parameter, PLS and PCR as a function of the number of latent factors). Figures 7a–b show a couple of typical examples, with MSEP plotted on a logarithmic scale against number of PLS-factors or ridge constant. The message is that it is not crucial that the x-vectors of the test set be located precisely in the same subspace as the calibration data, when by construction both subsets of data represent the same population data set.

However, the waste-water data actually represent two different seasons, by purpose of the chemists to cover some variation between time periods. This corresponds to the particular split in the first 65 and the last 60 observations, respectively. We do not have a third period here for external validation, so let us imagine that we had only one of the two periods available for calibration (including internal CV), and let the other be used as a test set for external validation. When the last 60 observations were used as a test set, the MSEP functions for LSRR, PLS and PCR looked reasonably smooth for both the internal CV and the external prediction test. In particular, the results for the test set were not very sensitive to the choice of ridge constant or number of factors. The considerable loss in predictive ability from internal to external validation, demonstrating that time periods differ, is shown for PLS and LSRR on an MSEP log-scale in Figs 8a–b. The secondary minima for the calibration set illustrate that such curves need not necessarily be bowl-shaped, although they typically are. As an absolute reference, note that we still predict considerably better than we would by simply classifying each spectrum as belonging to the low, medium or high nitrate group of the calibration and using the group mean as predictor. The latter procedure yields MSEP $= 2.0$, i.e. $\log \text{MSEP} = 0.30$.

When the last 60 observations were used for calibration and the first 65 observations formed the test set, a different, bizarre MSEP behaviour was recorded. This is shown for PLS and (LS)RR in Figs 9a–b. For PCR the picture was quite similar in character to that for PLS. We see that adding or removing a couple of PLS-factors could make MSEP for the test set jump by a factor of 30(!) while the internal CV MSEP stays essentially constant. It is natural that a

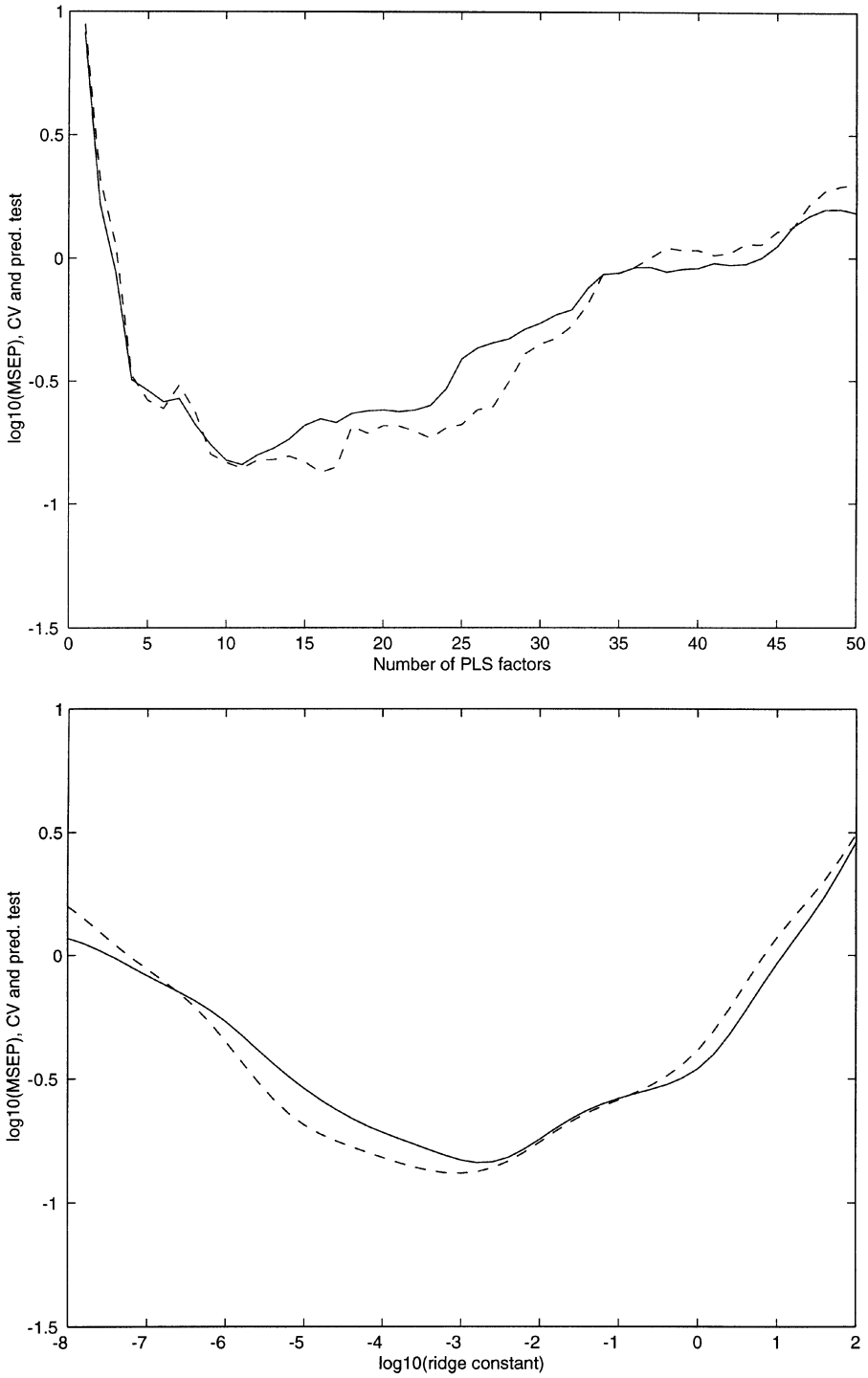


Fig. 7a–b. MSEP values, CV leave-one-out and test set validation, random split between calibration set (65 observations) and test set (60 observations). (a) PLS for varied number of PLS-factors. Solid line for CV leave-one-out, dashed line for test set; (b) LSRR for varied ridge constant. Solid line for CV leave-one-out, dashed line for test set.

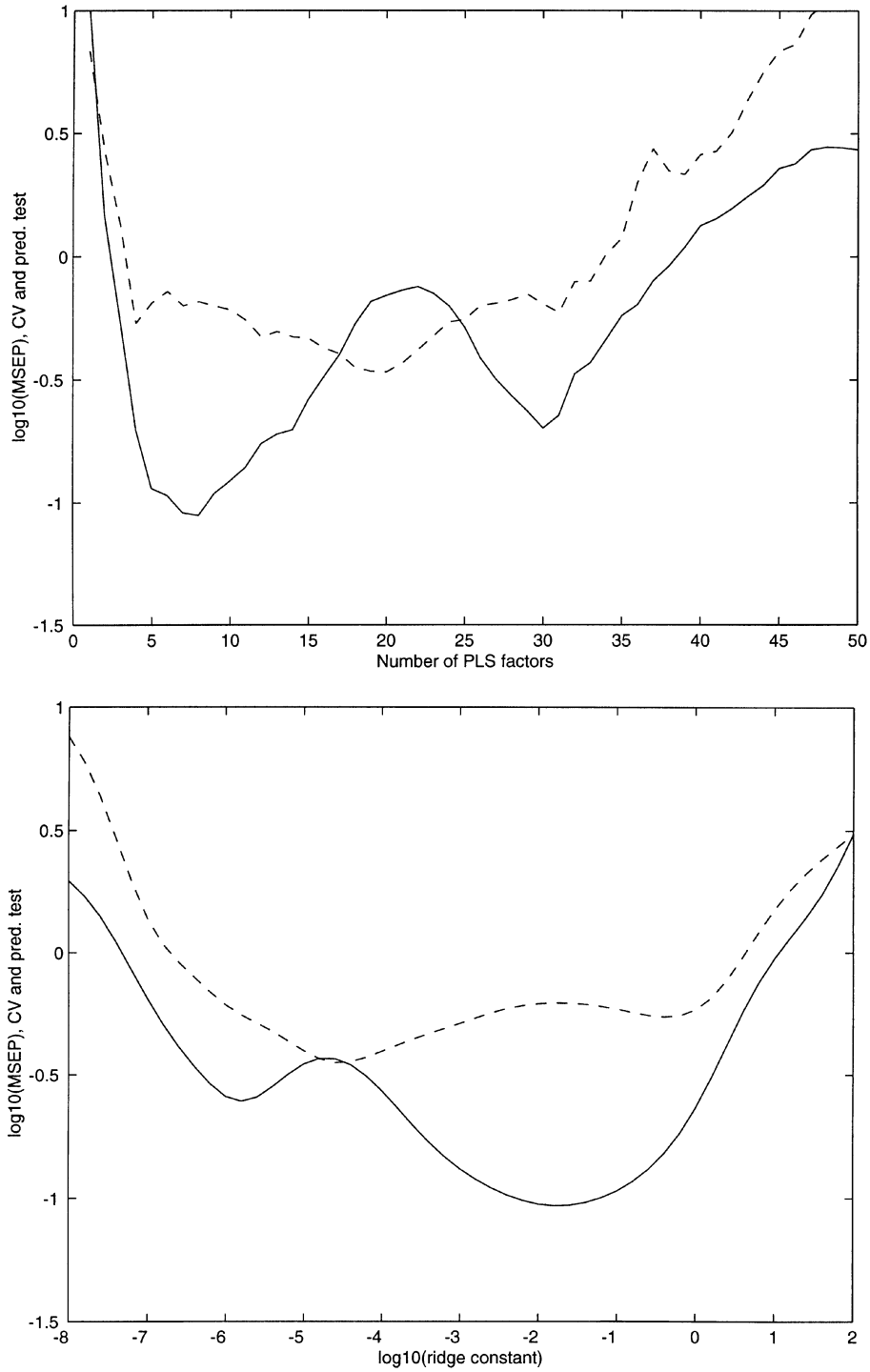


Fig. 8a–b. MSEP values, CV leave-one-out and test set validation, first 65 observations in calibration set and last 60 observations in test set. (a) PLS for varied number of PLS-factors. Solid line for CV leave-one-out, dashed line for test set; (b) LSRR for varied ridge constant. Solid line for CV leave-one-out, dashed line for test set.

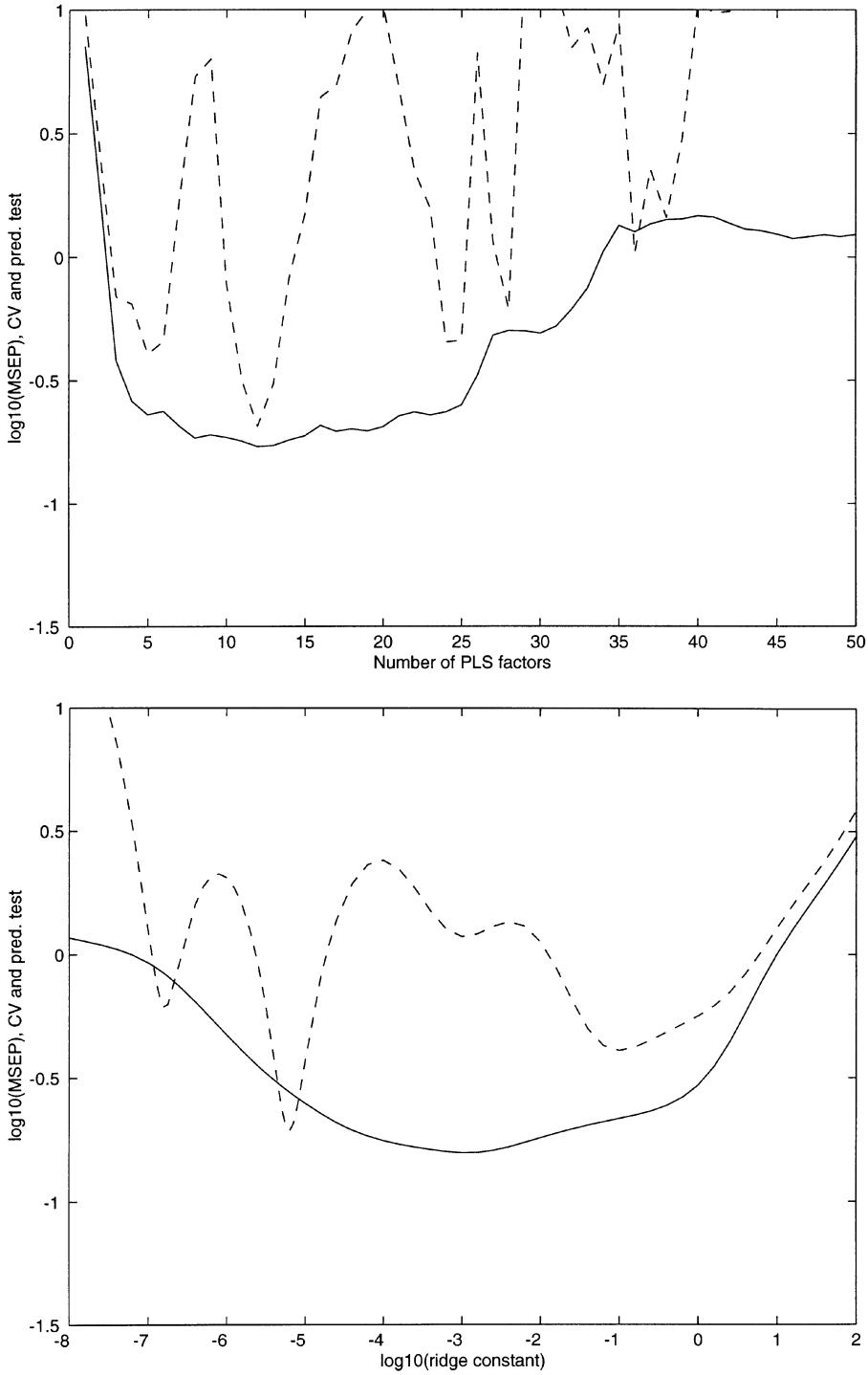


Fig. 9a–b. MSEP values, CV leave-one-out and test set validation, last 60 observations in calibration set and first 65 observations in test set. (a) PLS for varied number of PLS-factors. Solid line for CV leave-one-out, dashed line for test set; (b) LSRR for varied ridge constant. Solid line for CV leave-one-out, dashed line for test set.

difference between periods should yield worse predictions for the test set than expected from internal cross-validation, but this oscillating behaviour seems more difficult to understand.

Why can we not predict this test set? It turns out that when MSEP is found high, the low nitrate values have been badly predicted. Partial understanding was furnished by the PLS-scores of the fitted model. The first factor separates high and medium nitrate from low, the second factor nicely separates high from medium if the low values are forgotten, for calibration and test sets equally, see Fig. 10a. Subsequent score plots are more difficult to interpret, but in the fourth and many higher factors the low group from the test set stands out remarkably separated from the others scores, see Fig. 10b for examples. With the first 65 as calibration set, such a separation was not seen, but the factor scores behaved similarly in the calibration and prediction sets. It might be surprising that the homogeneous low group stands for the problems, but in any case the example drastically shows the risks connected with over-fitting, as factors are added whose behaviour in the calibration and prediction sets is different.

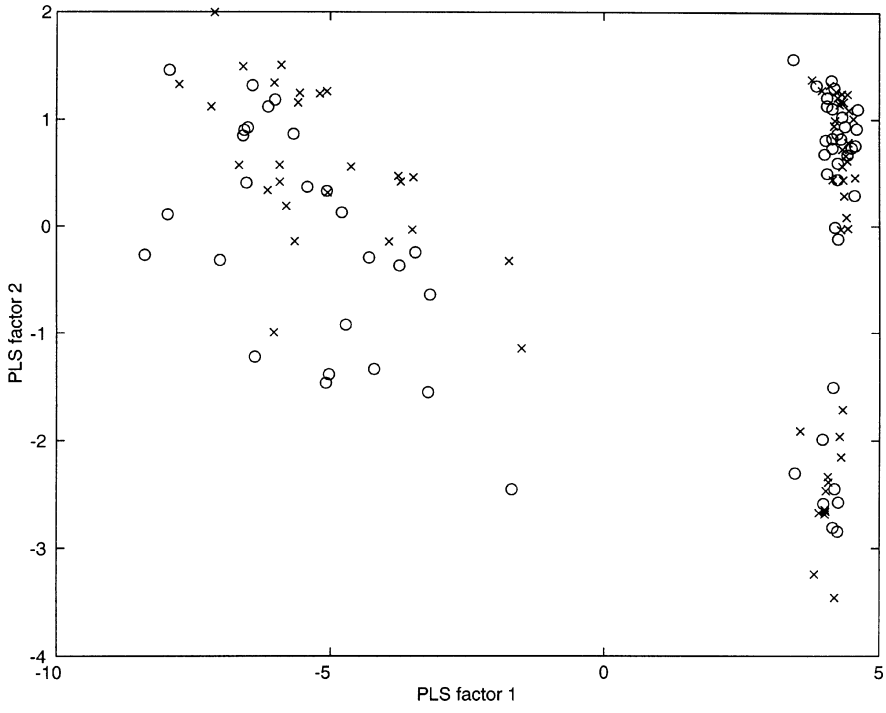
3.8. Multi-dimensional response

Suppose there are several constituents whose concentrations are determined in the calibration ($\dim y > 1$; multivariate regression). In specimens sampled from a natural population we might find that the constituents measured are highly correlated, for natural reasons likely to extend to the predictions. When we form predictors for all or for a subset of the constituents, is it possible to gain efficiency by utilizing this correlation? There are several more or less recent proposals.

Classical multivariate LS theory says that multivariate regression should be fitted as separate univariate multiple regressions, but this is not necessarily relevant for prediction and in the presence of near-collinearities. A joint PLS algorithm for multi-dimensional y, called PLS2, has been around for a long time, see e.g. Martens & Næs (1989, sect. 3.5.4) for a description. In PLS2, regressors in x-space which should be efficient for predicting some linear form in y are found by an algorithm that essentially is designed to jointly select the strongest covarying linear forms in y and x, a sort of "canonical covariance" analysis. A differing (and simpler) version of PLS2 is SIMPLS (de Jong, 1993), which is defined so as to yield orthogonal regressors $z = c^{\mathrm{T}}x$ where traditional PLS2 yields orthogonal weighting vectors c. See Burnham *et al*. (1996) for a recent discussion of these methods and some of the alternatives. SIMPLS can be fitted into the more general framework of joint continuum regression (JCR; Brooks & Stone, 1994), in the same way as univariate PLS fits into CR. Other joint methods are reduced rank regression (RRR), which also falls within the JCR framework, filtered canonical y-variate regression (FICYREG), and the recent "curds and whey" method of Breiman & Friedman (1997), who compare several methods in a controversial simulation study. There are certainly situations where it should be possible to borrow strength, for example from one precisely determined constituent to another highly correlated but imprecisely determined constituent, but the research until now is not conclusive as to when it pays and what methods should be recommended. Most experience seems to indicate that substantial gains are rare.

4. Multivariate calibration reviewed: bilinear and other less standard models

The following two situations from chemistry have in common a small number of specimens, so a prediction approach seems less natural. Consequently they ask for modelling of type multivariate regression of u on t.



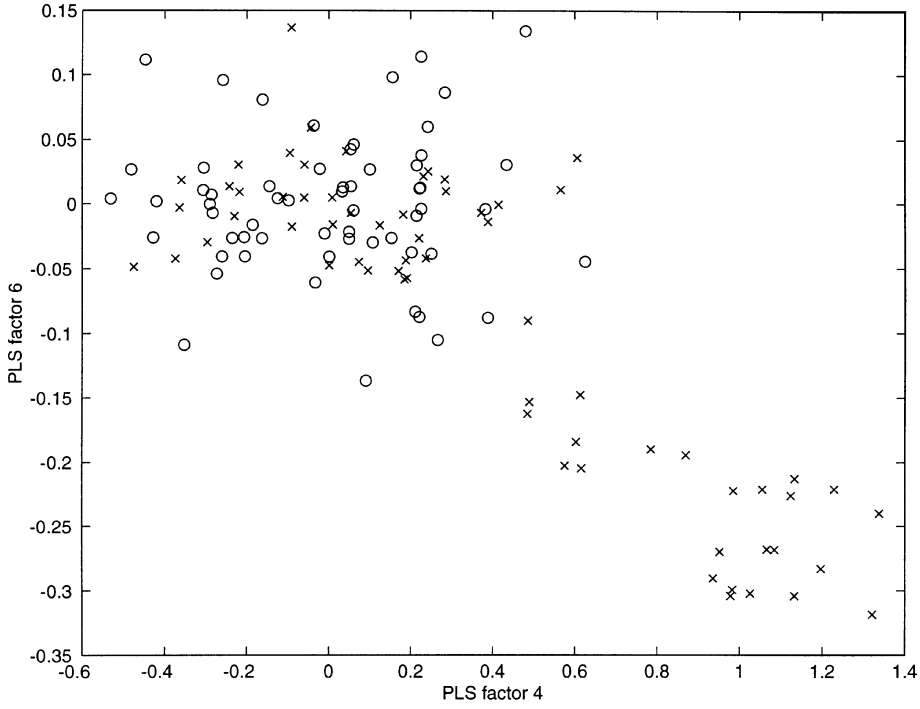


Fig. 10a–b. Two PLS score plots. PLS factors are specified by orthogonal weights (so factors are uncorrelated) and are based on the last 60 observations as calibration set. ○ scores for the calibration set (60); × scores for the test set (65); (a) PLS factors 1 and 2; (b) PLS factors 4 and 6; All of the relatively separated subset of crosses belong to the low nitrate group.

4.1. Generalized standard addition method (GSAM) of analytical chemistry

The standard addition method (SAM) is a well-known univariate spectroscopic procedure for chemical analysis in situations when so called matrix effects are suspected. By the matrix chemists think of the environment of the constituent of interest in the particular specimen to be analysed. Determination of a metal concentration in wine is an example where SAM is used. When an ordinary "external" calibration cannot mimic the specimen in all crucial aspects, SAM provides a combined calibration and estimation internal to the specimen. Disadvantages are lack of precision and sensitivity to modelling errors.

In SAM known amounts of the constituent of interest are added to the specimen. It is a crucial assumption that the instrument response is proportional to the amount present, that is we think of a linear regression model of type

$$y_i = \beta(\xi + x_i) + \varepsilon_i, \tag{4.1}$$

where the total amount $\xi + x$ behind a spectroscopic response y is the sum of the unknown amount ξ and the controlled addition $x = x_0, x_1, \ldots, x_n$, with $x_0 = 0$ for the original specimen. Making the replacement $\alpha = \beta\xi$, we see that we have an ordinary linear regression with the addition x as regressor and $\xi = \alpha/\beta$ as the parameter of particular interest. The background level of the measuring instrument must have been eliminated, because from data we cannot distinguish between intercept of type $\beta\xi$ and intercept from measurement background.

Statistical inference about ξ is straightforward. Least squares, or ML under normality, yields $\hat{\alpha}$, $\hat{\beta}$ and hence $\hat{\xi} = \hat{\alpha}/\hat{\beta}$. Error propagation formulae yield a standard error for $\hat{\xi}$. If exact confidence bounds are desired, such can be constructed by Fieller's procedure.

A generalized SAM, called GSAM, for multi-component analysis was introduced by Saxberg & Kowalski (1979), and developed by the Seattle chemometrics group. As statistical model we think of (4.1) with a matrix B replacing the scalar β and y, x, ξ and ε as vectors, the latter i.i.d. $N(0, \Gamma)$:

$$Y_i = B(\xi + x_i) + \varepsilon_i = \alpha + Bx_i + \varepsilon_i. \tag{4.2}$$

As often seen with new types of chemometrics problems, the earliest papers on GSAM neglected statistical aspects and regarded the estimation of ξ primarily as a mathematical problem of solving overdetermined equation systems. From a statistical point of view the case $q = p$ does not differ crucially from the special case of SAM. We can take the LS (= ML) estimates $\hat{\alpha}$ and $\hat{B}$ and from them form $\hat{\xi} = \hat{B}^{-1}\hat{\alpha}$. When $q > p$ there is no longer a one-to-one correspondence between parametrizations by α and by ξ, and multivariate complications appear. Various estimation procedures are suggested in Sundberg (1988). ML estimation yields a non-linear equation system for ξ, without explicit solution. However, the MLE is typically close to the estimated GLS estimator

$$\hat{\xi}_{\mathrm{EGLS}} = (\hat{B}\hat{\Gamma}^{-1}\hat{B}^{\mathrm{T}})^{-1}\hat{B}\hat{\Gamma}^{-1}\hat{\alpha}, \tag{4.3}$$

analogous with (2.5). As in section 2 the unweighted simple LS estimator, obtained by replacing $\hat{\Gamma}^{-1}$ in (4.3) by the identity, could be a better alternative. This is because approximate variance formulae corresponding to (2.7)–(2.9) hold in the GSAM case as well, by close analogy, and they show that a high price might have to be paid for using the estimated weights matrix $\hat{\Gamma}^{-1}$ in (4.3).

Recently, Booksh *et al*. (1995) have proposed and implemented an extension of the GSAM idea to the second order type of instruments to be discussed in the next section.

4.2. Bilinear models for two-way data from hyphenated instruments

In modern analytical chemistry two instruments are sometimes combined, to yield a highly informative two-way array of responses for each specimen. For example, liquid chromatography (LC) may be combined with ultra-violet (UV) spectroscopy to form the "hyphenated" instrument LC–UV. Here LC more or less efficiently separates the constituents of the specimen, which spend different time passing through the LC instrument. The UV instrument measures the spectra of the LC output at a number of different times. The resulting responses form a matrix $U = \{U_{ij}\}$, where (i, j) denotes (time, wavelength).

For a pure substance the LC time profiles at all wavelengths should be mutually proportional, and likewise the UV spectra at all times, under normal conditions. This means that the response matrix U for each specimen should follow a model of type

$$U = \sum_k t_k \alpha_k \beta_k^{\mathrm{T}} + E, \tag{4.4}$$

where t_k is the concentration or amount of substance k, $k = 1, \ldots, p$, α_k and β_k are parameter vectors (standardized LC profiles and UV spectra, respectively), and E is a random noise matrix. This is a non-linear regression model that is bilinear in its two dimensions (time, wavelength). It is also linear in the third dimension, the concentration t_k varying between specimens. As in the GSAM of the previous section, the background level of the measuring instrument (the intercept) is assumed separately eliminated.

Each matrix term $\alpha_k \beta_k^{\mathrm{T}}$ in (4.4) is rank one, and many methods devised for analysing this sort of data amount to finding low rank approximations to all the observed data matrices U simultaneously. For the pure regression problem, with all t-values known for a sample of U-matrices, Linder & Sundberg (1998) study bilinear least squares, that must be solved iteratively, and an alternative, direct estimation method based on singular value decomposition (SVD) of suitably reweighted linear combinations of the data matrices. The SVD method is not much less efficient, and it has the advantages of not only being direct but also allowing explicit standard error estimates.

Much methodology in use is based on unfolding the two-way structure into an ordinary one-way structure, that is going from U to vec(U), at the same time sacrificing the (exact) bilinearity. Wold *et al*. (1987) proposed that the roles of x and y also be exchanged, and a PCR or PLS algorithm be used on the unfolded U-data. Bro (1996) has recently introduced a truly multilinear version of PLS that will preserve the bilinear structure of (4.2). PARAFAC (parallel factor analysis), borrowed from psychometry, also preserves the bilinearity. This is a trilinear decomposition (TLD) method, to be used with specimen forming the third dimension (Sanchez & Kowalski, 1990; Burdick *et al*., 1990). This means that PARAFAC by itself does not use the known concentrations t_k, so it must be combined with a step where estimated concentrations are related to their known values, for example by least squares.

When the problem is to estimate the t-values of a new specimen, methods for simultaneous calibration and estimation have been devised. Essentially, one reference specimen of known composition is sufficient for the calibration, because of the bilinear structure. The generalized rank annihilation method (GRAM) solves the mathematical problem in this case. For a review of its successive development, see Faber *et al*. (1994). When several calibration specimens are available, this methodology is not established. Unfolding of the sampling dimension and TLD variants have been proposed for an extended GRAM.

5. Literature

Specific references have been given above in their appropriate context. Here is a more general view.

Although the paper by Brown (1982), with its lively discussion, was not the first to consider multivariate calibration it became the starting point and basic reference for statistical research in many directions. In parallel new instruments and fitting methodology were developed by chemists. The first book on multivariate calibration was written by the Norwegian pioneers Martens & Næs (1989). Their book is specifically directed towards applications in spectrophotometry for the chemist user, and situations with more variables than observations tend to dominate. A later book by Brown (1993) gives a fairly comprehensive treatment of many statistical aspects of multivariate calibration, including the Bayesian perspective that we have left out of the present paper. Both books have many examples with real data.

These books also cover univariate calibration, of course, but more references to the extensive statistical literature on this specific problem of application are found in the review paper by Osborne (1991).

The present review has touched some chemistry literature, and the statistician should be reminded that most journal papers on multivariate calibration appear in the two journals for chemometrics, started in 1986–87, and not only scattered among statistics journals.

6. Some concluding words

Even if several topics have been deliberately omitted to keep space limited (for example the Bayesian aspect, for which we refer to Brown's (1993) book) this was an effort to tell where the central theory of multivariate calibration stands today and where it is heading. However, where it is heading must depend strongly on the developments in the fields of application. In the 1970s statisticians did not foresee the urgent and extensive needs of the 80s to go from a univariate to a high-dimensional multivariate theory of calibration. Now, while statisticians (including the present paper) are still discussing multivariate calibration for vectors u, chemists are collecting large U-matrices and even higher-dimensional arrays of data in their hyphenated instruments. The statistician searching collaboration can find many stimulating data and connected statistical problems.

Acknowledgement

The financial support by the Swedish Natural Science Research Council is gratefully acknowledged, and the kindness of Mikael Karlsson, Bo Karlberg and Ralf Olsson to provide their waste water data.

References

Björkström, A. & Sundberg, R. (1996). Continuum regression is not always continuous. *J. Roy. Statist. Soc. Ser. B* **58**, 703–710.

Björkström, A. & Sundberg, R. (1999). A generalized view on continuum regression. *Scand. J. Statist.* **26** 17–30.

Booksh, K., Henshaw, J. M., Burgess, L. W. & Kowalski, B. R. (1995). A second-order standard addition method with application to calibration of a kinetics-spectroscopic sensor for quantitation of trichloroethylene. *J. Chemometrics* **9**, 263–282.

Breiman, L. & Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 3–54.

Bro, R. (1996). Multiway calibration. Multilinear PLS. *J. Chemometrics* **10**, 47–62.

Brooks, R. & Stone, M. (1994). Joint continuum regression for multiple predictands. *J. Amer. Statist. Assoc.* **89**, 1374–1377.
Brown, P. J. (1982). Multivariate calibration (with discussion). *J. Roy. Statist. Soc. Ser. B* **44**, 287–321.
Brown, P. J. (1992). Wavelength selection in multicomponent near-infrared calibration. *J. Chemometrics* **6**, 151–161.
Brown, P. J. (1993). *Measurement, regression, and calibration*. Oxford University Press, Oxford.
Brown, P. J. & Oman, S. D. (1991). Double points in nonlinear calibration. *Biometrika* **78**, 33–43.
Brown, P. J., Spiegelman, C. H. & Denham, M. C. (1991). Chemometrics and spectral frequency selection. *Philos. Trans. Roy. Soc. London Ser. A* **337**, 311–322.
Brown, P. J. & Sundberg, R. (1987). Confidence and conflict in multivariate calibration. *J. Roy. Statist. Soc. Ser. B* **49**, 46–57.
Brown, P. J. & Sundberg, R. (1989). Prediction diagnostics and updating in multivariate calibration. *Biometrika* **76**, 349–361.
Burdick, D. S., Tu, X. M., McGown, L. B. & Millican, D. W. (1990). Resolution of multi-component fluorescent mixtures by analysis of the excitation–emission–frequency array. *J. Chemometrics* **4**, 15–28.
Burnham, A. J., Viveros, R. & MacGregor, J. F. (1996). Frameworks for latent variable multivariate regression. *J. Chemometrics* **10**, 31–45.
Clarke, G. P. Y. (1992). Inverse estimates from a multiresponse model. *Biometrics* **48**, 1081–1094.
Davis, A. W. & Hayakawa, T. (1987). Some distribution theory relating to confidence regions in multivariate calibration. *Ann. Inst. Statist. Math.* **39**, 141–152.
de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics Intell. Lab. Systems* **18**, 251–263.
de Jong, S. & Farebrother, R. W. (1994). Extending the relationship between ridge regression and continuum regression. *Chemometrics Intell. Lab. Systems* **25**, 179–181.
Denham, M. C. & Brown, P. J. (1993). Calibration with many variables. *J. Roy. Statist. Soc. Ser. C* **42**, 515–528.
de Plessis, J. L. & van der Merwe, A. J. (1996). Bayesian calibration in the estimation of the age of rhinoceros. *Ann. Inst. Statist. Math.* **48**, 17–28.
Eisenhart, C. (1939). The interpretation of certain regression methods and their use in biological and industrial research. *Ann. Math. Statist.* **10**, 162–186.
Faber, N. M., Buydens, L. M. C. & Kateman, G. (1994). Generalized rank annihilation method. I: Derivation of eigenvalue problems. *J. Chemometrics* **8**, 147–154.
Follman, D. (1995). Multivariate tests for multiple endpoints in clinical trials. *Statist. Med.* **14**, 1163–1175.
Frank, I. E. (1987). Intermediate least squares regression method. *Chemometrics Intell. Lab. Systems* **1**, 233–242.
Frank, I. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–148.
Fujikoshi, Y. & Nishii, R. (1984). On the distribution of a statistic in multivariate inverse regression analysis. *Hiroshima Math. J.* **14**, 215–225.
Garthwaite, P. H. (1994). An interpretation of partial least squares. *J. Amer. Statist. Assoc.* **89**, 122–127.
Goutis, C. (1998). Second derivative functional regression with applications to NIR spectroscopy. *J. Roy. Statist. Soc. Ser. B* near infra-red **60**, 103–114.
Goutis, C. & Fearn, T. (1996). Partial least squares regression on smooth factors. *J. Amer. Statist. Assoc.* **91**, 627–632.
Gunst, R. F. (1983). Latent root regression. In *Encyclopedia of statistical sciences*, Vol 4 (eds S. Kotz, N. L. Johnson & C. B. Read), 495–497. Wiley, New York.
Helland, I. S. (1988). On the structure of partial least squares regression. *Comm. Statist. Simulation Comput.* **17**, 581–607.
Helland, I. S. (1990). Partial least squares regression and statistical models. *Scand. J. Statist.* **17**, 97–114.
Helland, I. S., Næs, T. & Isaksson, T. (1995). Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometrics Intell. Lab. Systems* **29**, 233–241.
Hjorth, J. S. U. (1994). *Computer intensive statistical methods. Validation, model selection and bootstrap*. Chapman & Hall, London.
Höskuldsson, A. (1988). PLS regression methods. *J. Chemometrics* **2**, 211–228.
Karlsson, M., Karlberg, B. & Olsson, R. J. O. (1995). Determination of nitrate in municipal waste water by UV spectroscopy. *Anal. Chim. Acta* **312**, 107–113.
Krutchkoff, R. G. (1967). Classical and inverse regression methods of calibration. *Technometrics* **9**, 425–439.

Linder, M. & Sundberg, R. (1998). Second order calibration: bilinear least squares regression and a simple alternative. *Chemometrics Intell. Lab. Systems* **42**, 159–178.
Martens, H. & Næs, T. (1989). *Multivariate calibration*. Wiley, Chichester.
Mathew, T. & Kasala, S. (1994). An exact confidence region in multivariate calibration. *Ann. Statist.* **22**, 94–105.
Mathew, T. & Zha, W. (1996). Conservative confidence regions in multivariate calibration. *Ann. Statist.* **24**, 707–725.
Mertens, B., Fearn, T. & Thompson, M. (1995). The efficient cross-validation of principal components applied to principal component regression. *Statist. Comput*. **5**, 227–235.
Næs, T. (1985). Multivariate calibration when the error covariance matrix is structured. *Technometrics* **27**, 301–311.
Næs, T. (1986). Multivariate calibration using covariance adjustment. *Biometrical J.* **28**, 99–107.
Næs, T. & Martens, H. (1987). Testing adequacy of linear random models. *Statistics* **18**, 323–331.
Nishii, R. & Krishnaiah, P. R. (1988). On the moments of the classical estimates of the explanatory variables under a multivariate calibration model. *Sankhyā Ser. A* **50**, 137–148.
Oman, S. D. (1988). Confidence regions in multivariate calibration. *Ann. Statist.* **16**, 174–187.
Oman, S. D. & Srivastava, M. S. (1996). Exact mean squared error comparisons of the inverse and classical estimators in multi-univariate linear calibration. *Scand. J. Statist.* **23**, 473–488.
Oman, S. D. & Wax, Y. (1984). Estimating fetal age by ultrasound measurements: an example of multivariate calibration. *Biometrics* **40**, 947–960; corrigendum (1985) **41**, 821–822.
Osborne, C. (1991). Statistical calibration: a review. *Internat. Statist. Rev.* **59**, 309–336.
Sanchez, E. & Kowalski, B. R. (1990). Tensorial resolution: a direct trilinear decomposition. *J. Chemometrics* **4**, 29–45.
Saxberg, B. E. H. & Kowalski, B. R. (1979). Generalized standard addition method. *Anal. Chem.* **51**, 1031–1038.
Seber, G. A. F. & Wild, C. J. (1989). *Nonlinear regression*. Wiley, New York.
Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction (with discussion). *J. Roy. Statist. Soc. Ser. B* **36**, 111–147.
Stone, M. & Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *J. Roy. Statist. Soc. Ser. B* **52**, 237–269; corrigendum (1992) **54**, 906–907.
Sundberg, R. (1985). When is the inverse regression estimator MSE-superior to the standard regression estimator in multivariate controlled calibration situations? *Statist. Probab. Lett.* **3**, 75–79.
Sundberg, R. (1988). Interplay between chemistry and statistics, with special reference to calibration and the generalized standard addition method. *Chemometrics Intell. Lab. Systems.* **4**, 299–305.
Sundberg, R. (1993). Continuum regression and ridge regression. *J. Roy. Statist. Soc. Ser. B* **55**, 653–659.
Sundberg, R. (1996). The precision of the estimated generalized least squares estimator in multivariate calibration. *Scand. J. Statist.* **23**, 257–274.
Sundberg, R. & Brown, P. J. (1989). Multivariate calibration with more variables than observations. *Technometrics* **31**, 365–371.
van Huffel, S. & Vandewalle, J. (1991). *The total least squares problem: computational aspects and analysis*. SIAM, Philadelphia.
Vigneau, E., Bertrand, D. & Qannari, E. M. (1996). Application of latent root regression for calibration in near-infrared spectroscopy. Comparison with principal component regression and partial least squares. *Chemometrics Intell. Lab. Systems* **35**, 231–238.
Williams, E. J. (1959). *Regression analysis*. Wiley, New York.
Wold, S. (1978). Cross-validatory estimation of the number of components in factor analysis and principal components models. *Technometrics* **20**, 397–406.
Wold, S., Geladi, P., Esbensen, K. & Öhman, J. (1987). Multi-way principal components- and PLS-analysis. *J. Chemometrics* **1**, 41–56.
Wood, J. T. (1982). Estimating the age of an animal: an application of multivariate calibration. In *Proc. 11th Int. Biometrics Conference*.

Received July 1997, in final form April 1998

Rolf Sundberg, Mathematical Statistics, Stockholm University, S-106 91 Stockholm, Sweden.

Discussion and Comments

PHILIP J. BROWN

University of Kent at Canterbury

There is an English saying directed at "something old, something new, something borrowed and something blue". I won't go into its origins but the "blue" here is not the acronym "Best Linear Unbiased Estimation", although if I were to level criticism at this excellent review it would be at its classical stance and its omission of much revealing Bayesian insights. Also procedures which are not Bayes with respect to some prior distribution will, at least in Bayesian eyes have demonstrable deficiencies.

Thus the EGLS estimator (2.5) is not MLE, but to interpret this as merely a quirk of some (little) information about Γ is not I think the issue. It is neither MLE for (2.4) where Γ is known (Brown, 1993, rem. 3, sec. 5.8). Both (2.4) and (2.5) treat $\hat{B}$ as known and fixed whereas this estimator is uncertain and plug-in estimators for such non-linear forms are bound to be inadequate. Much can be learnt from the Bayes form of posterior (Brown, 1982, 1993) which utilizes an average predictive likelihood. I think foibles of weights being negative in the GLS estimator are interesting only in this sub-optimal context.

If Rolf's primary practical finding is that most regularization methods (PLS, PCR, CR, RR, LSRR) do about as well as one another then I will continue to feel most comfortable in using ridge regression, because of its natural pedigree in terms of exchangeability of regression coefficients, the ridge constant being the ratio of the residual variance to the variance of these coefficients. I will continue to feel uneasy about the least squares scaled version Rolf propounds and also suggestions of negative ridge constants.

I think Rolf is right to sound warning bells about simple marginal forms of variable selection. We have been developing a number of Bayesian variable selection procedures, see Brown *et al*. (1997, 1998a–c) with software on the Web ⟨http://www.ukc.ac.uk/IMS/statistics/people/M.Vannucci.html⟩. These cannot be entirely automated, and probably should not be, and do we think provide procedures that will seek out even linear combinations of absorbances if non-linearity is present. They are also truly multivariate where prediction of several constituents is concerned. The approaches based on a latent selection vector and a good posterior fit also can gain robustness by averaging over a number of good selections. If a single model is sought where variables have genuine costs a different decision theoretic tack has been adopted, see Brown *et al*. (1997, 1998). Trying these on the nitrate pollution data in its most difficult split (last 60 observations for calibration), with a generalized cross-validation estimate of the ridge constant, and a prior expectation of 20 useful wavelengths, we find that the best selection involves five wavelengths (7, 18, 19, 21, 24) with a prediction mean squared error of 0.254 or -0.6 on the log base 10 scale, better almost everywhere than the PLS predictions depicted in Fig. 9a.

However, we would be reluctant to use this data to illustrate any simple automated procedure. The description in Karlsson *et al*. (1995) makes it clear that there are interfering turbidity problems and the water comes from three different locations and within each of these inlet, main tank, and outlet water samples are taken. A glance at fig. 2 clearly indicates three different populations and accurate predictions are more likely to come from taking the covariates explicitly into account in forming the calibrations.

I have greatly enjoyed reading Rolf's scholarly paper and particularly commend his recent work on hyphenated instruments (with Marie Linder). There are many sound insights and I'm sure I will find further highlights when I reread it.

Additional references

Brown, P. J., Vannucci, M. & Fearn, T. (1997). Multivariate Bayesian wavelength selection for NIR spectra applied to biscuit dough pieces. In *5èmes J. Agro-Industrie Méthod. Statist.*, *Versailles*, France, pp. 19.1–19.11.

Brown, P. J., Fearn, T. & Vannucci, M. (1998a). The choice of variables in multivariate regression: a Bayesian non-conjugate decision theory approach. Technical Report UKC/IMS/1998/03. University of Kent.

Brown, P. J., Vannucci, M. & Fearn, T. (1988b). Bayesian wavelength selection in multicomponent analysis. *J. Chemometrics* **12**, 173–182.

Brown, P. J., Vannucci, M. & Fearn, T. (1998c). Multivariate Bayesian variable selection and prediction. *J. Roy. Statist. Soc. Ser. B*, **60**, 627–641.

HARALD MARTENS

NTNU, Norway

Prof. Sundberg's overview article is a welcome collection of topics and experiences in this field. I shall only limit myself to addressing three points: (1) there is considerable confusion on what is meant by cross validation (two very different methods are used under the same name); (2) how to handle population heterogeneities and unobserved phenomena; and (3) the importance of using contextual background knowledge in multivariate calibration.

1. Different cross validation methods go by the same name

Cross validation is not one single technique, at least not in chemometrics. The different chemometrics software systems use at least two different cross validation principles with quite different properties. Sundberg apparently uses the standard statistical type (see e.g. Efron & Tibshirani, 1993, p. 239). I am not going to argue which of the two cross validation methods are "best". But to my knowledge the difference between the commonly used cross validation methods for multivariate calibration in chemometrics has not been properly handled before.

Local cross validation checks the predictive validity of each new component individually

Local cross validation is used in e.g. the SIMCA chemometrics program. Originally it was introduced by Svante Wold around 1980, inspired by Stone (1974), I believe. Each successive principal component or PLS component $a = 1, 2, \ldots$ is cross validated in order to see if *this* component a has the ability to improve the predictive description of the remaining, unmodelled Y (and X-) residuals after $a - 1$ factors. Then the full (not cross-validated) estimation of the model parameters for this component a is performed on *all* the objects, component a is subtracted, and one is ready for next factor, $a = a + 1$:

```
Centre X and Y, based on all objects.
For component a = 1, 2, ...
  for segment s = 1, 2, ..., S
    Estimate locally the parameters for component a from X and Y in
    the objects not belonging to s
    Using the locally estimated parameters, predict Y from X in the
    objects in s
    Accumulate prediction error in Y for component a
  end
  Check if component a adds clear predictive improvement for Y
```

```
  Estimate full-model component a from all the objects
  Subtract this full-model component from X and Y
end
```

The main purpose of local cross validation is to obtain as good an estimate as possible of the optimal rank of the model by perturbing the regression model as little as possible, so that the cross validation results are as relevant to the final regression model as possible. I believe one can say that this cross validation was developed to optimize interactive explorative multivariate analysis. There is no cross validation results for $a = 0$ components (the mean).

Full cross validation checks the predictive validity of the full model for all factors

The full cross validation (see Efron & Tibshirani, 1993) is used e.g. in the Unscrambler program. In chemometrics I believe it was introduced by this author, around 1981; the motive was to have a statistically more conservative estimator of the prediction uncertainty in Y than the one used in SIMCA. The full cross validation is not done on individual factors, but on the whole regression model, including means and every successive factor. The main purpose of the full cross validation is to obtain a good estimate of the rank of the model *and* a realistic estimate of the actual prediction error to be expected for future unknown samples from the same population. This method was developed to optimize multivariate calibration, but also to be good for explorative multivariate analysis.

I believe that the two cross validation methods differ mainly in the following two aspects.

(a) They are both bad for validating regressions on orthonormal X, where PLS regressions yields MLR already in the first component, but local cross validation may be the least bad. For data where X has multiple singular values (after centring), such as when $X =$ dummy variables from an orthogonal design, the PLS regression (not cross-validated) has no help from X-structure, so it delivers the MLR solution directly in the FIRST factor, at least for "PLS1" regression which applies to a single Y-variable. But any perturbation that changes the multiple singular values structure yields more than 1 component. The cross validation (both methods) changes the singular values structure. Hence the cross validation results are not fully relevant for the not-cross validated model. The local cross validation, however, detects this problem already after the first factor, and signals that there is only one component in the system. The full cross validation may sometimes continue, to indicate a couple of more factors as "significant". I believe the y-loadings for these "falsely significant" factors are usually very small, so the practical damage is limited. For "PLS2" regression with more than one Y-variable, components 2, 3 etc. apparently make sense in their own right when the Y-variables span several valid components between themselves.

(b) Full cross validation gives realistic prediction error estimates, local cross validation does not. Since the local cross validation only validates the individual factors, it is not intended to estimate the actual prediction error, and indeed does not do so; it strongly underestimates the actual prediction error in Y.

For the full cross validation, on the contrary, theoretical considerations as well as empirical studies indicate that the full cross validation gives very good prediction uncertainty estimates, when compared to the performance in very big independent control sets.

A Monte Carlo simulation study on simulated data under various circumstances was published in Martens (1997). A larger study based on empirical data (NIR determination of maize composition; 25–70 X-variables, 1 Y-variable, 10–15 latent variables, 20–120 objects), with

the same conclusions, was recently presented by Martens & Dardenne (1998). The results showed that full cross validation under the specified conditions gave quite satisfactory estimates of both model rank as well as of prediction uncertainty level.

2. How to handle population heterogeneities and unobserved phenomena

The leave-one-out cross validation worked well in Sundberg's two-season waste water samples, but gave overfitting when he split the data into the two seasons, calibrated in one season and predicted the other one. This phenomenon is well known for any one working with multivariate calibration e.g. in the process industry, or in traditional agricultural use of NIR spectrophotometry, where instrument drift or population drift is a fact of life.

If the data from one of Sundberg's seasons were affected by phenomena not present the other season, one cannot expect good prediction from one season to the other. To fine-tune a model for a given population (e.g. for one season), one amplifies minor differences between X-variables into Y-adjustments to correct automatically e.g. for minor nonlinearities etc. This amplification of minor X-differences also causes amplification of noise and other unmodelled problems in X. If future X-data do not belong to the same population and therefore have larger X-residuals, these are amplified and lead to variance inflation in Y, as discussed in e.g. Martens & Næs (1989, p. 245).

A statistical model is primarily defined for one given assumed population. Multivariate regression allows automatic detection of many types of outliers relative to the assumed population. Both for practical purposes and seen from philosophy of science it is important that this model critique is used continually. Otherwise outliers, instrument drift etc. can pass unnoticed and play havoc with any calibration model's predictive performance. It is possible that a check of Mahalanobis distances or the automatic outlier warnings and estimated uncertainties for individual Y-predictions, used e.g. in the Unscrambler, would have told Sundberg not to apply the PLSR or PCR models uncritically from one season to the next, but instead to update the models.

We often know that we must expect instrument or population drift or other unknown problems in future X-data. Then one is wise to reduce the ambition level during calibration, by using fewer factors than apparently optimal. This reduces the length of the regression vector and hence reduces the noise amplification.

But I do find it ugly to force this conservatism in blindly, by holding some of the available, already too few, calibration samples out as 'test set' during the calibration, as is very often advocated (fortunately not by Sundberg).

The above referenced Monte Carlo studies were based on random resampling within unusually large empirical data sets: repeatedly a small calibration set was drawn, each time with a very large 'secret' control set, to simulate small-sample situations, and to summarize how small-sample calibration performs in the long run. These studies support the following theoretical expectation for homogenous population: when the number of objects is limited, the full cross validation (implemented as leave-one-independent-object-out) is far superior to the splitting of the available data set into calibration set and an independent test set. This is so both with respect to finding the 'correct' rank and to giving good Y-predictions. The use of a second evaluation test set just made things worse. The variance of an estimated variance is very big, so the use of small test set(s) is like playing Russian Roulette. But the use of big test set(s) makes the remaining calibration set too small to give valid models, unless the number of available objects is very high, and then it does not matter what one does in any way, because overfitting is no longer a problem.

For small sample sets, a certain under-estimation of the true prediction error will always occur, be it in test set validation or in cross validation, because statistically, the limited number of available samples will cause some real variation phenomena in the population to be absent in

the available calibration sample set. If, in a very big park (the population), we are to calibrate for $Y =$ the number of flowers per m^2 grass from $X =$ whatever observations we can make, and if our sampled patch of grass (the available calibration samples) contains many red and blue flowers, but no yellow ones, how can we bring yellow flowers into our considerations? Can we extrapolate "flower-ness" statistics from red and blue flowers to other colours not yet seen?

3. The importance of using contextual background knowledge in multivariate calibration

As mentioned above, Sundberg might have found it useful to check the validity of one season's calibration model for the next season. There must be constant model critique in multivariate calibration.

Multivariate calibration has been proven to allow scientifically valid and industrially important quantitative analysis in real-world, "dirty" systems, even in systems where causal theory is inadequate or lacking. The reason is that the application experts use a combination of statistical sampling (learning from the world) and interactive data analysis (using their more or less "tacit", contextual domain-specific knowledge). In order for this to work, the application experts must do their own data analysis, and in order to be able to do this, they must have tools that make the problem cognitively accessible to them, have conservative validation and powerful explorative functionality.

Modern chemometric software is designed so that the relevant and reliable main structures in the data are made visually accessible in the graphs of the first few latent variables, and unexpected phenomena in higher dimensions and/or in future samples are tagged by automatic outlier warnings. Latent variables-methods like PLSR in complex cases give a more valid and safe dimension-reduction than, on one hand, traditional chemistry and physics (senselessly collapsing to parameters in known, but incomplete "laws" of nature), and on the other hand, traditional statistics (senselessly reducing the number of variables, just to avoid the collinearity "problem").

The more traditional statistical rank-reduced regression methods like ridge regression (RR) work similarly to the bilinear regressions like PLSR and PCR when the sun is up and everyone is happy. But in murky systems, where there is a need to learn more chemistry or physics from the calibration data, the output from RR, the regression coefficient vector, is useless as interpretation tool. Even the *sign* of the regression coefficients is often meaningless.

Additional references

Efron, B. & Tibshirani, R. J. (1993) *An introduction to the bootstrap*. Monographs on Statistics and Applied Probability **57**, Chapman & Hall, New York.

Martens, H. (1997) Determining rank and evaluating performance in regression: independence test sets or cross-validation/cross-evaluation? In *19th Symp. on applied statistics* (eds L. Nørgaard, and A. Høskuldsson), pp 15–44.

Martens, H. & Dardenne, P. (1998) Validation and verification of regression in small data sets. *Chemometrics Intell. Lab. Systems* in press.

TORMOD NÆS

MATFORSK, Norway

Rolf Sundberg gives a nice overview of some important aspects of multivariate calibration. He discusses a lot of important topics and methods found in the statistical literature. In

addition he also describes a number of techniques traditionally discussed mainly in the chemometric literature. Such a paper is very welcome.

The present contribution is not a point by point discussion of Sundberg's viewpoints and results. It is merely a discussion of some additional and in my viewpoint important ideas and topics which are little focused in the paper. The emphasis will be on background arguments and philosophical aspects rather than on specific statistical solutions and methodology.

Multivariate calibration methods will always be methods left in the hands of non-expert "statisticians". This is because calibrations have to be done for each particular application and it may not always be possible to have a fully trained statistical expert at hand. It is therefore of vital importance that methods which are developed and used in this area are not only accurate and based on sound statistical principles; they must also satisfy a number of additional criteria (see Martens and Næs, 1989). Some of these are: a calibration method should be

(i) easy to understand;
(ii) easy to use;
(iii) flexible (applicable in many situations); and
(iv) easy to interpret and lend itself easily to development of (outlier) diagnostic procedures.

All these aspects are very important in practice. Some of the methods discussed in the paper satisfy many of these criteria, while others do not. In particular I would like to point at the regression methods PCR and PLS (and some of their modifications, see below) as methods of special importance. In addition to being methods that give precise and reliable predictions in a lot of practical applications, they satisfy the other criteria listed as well: they are easy to visualize and interpret geometrically by users. In addition, they provide graphical tools that are very useful for interpreting the data structures at hand. These tools are called scores plots and loadings plots and can provide a lot of useful insight. It is often of vital importance that such information is made available, especially in an early stage of a calibration when little is known. In addition, tools like residuals and leverage are easy to develop and useful for detection of outliers and extreme observations (see e.g. Martens & Næs, 1989). It is my belief that, equipped with these techniques chemists and other users with special subject matter knowledge can do very good and reliable calibration work.

Methods like for instance ridge regression, on the other hand, are more questionable with respect to the listed criteria. A small advantage in prediction ability (see Frank & Friedman, 1993) can sometimes be of little interest if most of the other criteria fail to hold.

Limiting the attention to modelling and estimation aspects of calibration, the following four problems are in my opinion the most basic and important ones:

(i) lack of selectivity;
(ii) collinearity;
(iii) outliers; and
(iv) non-linearity.

The first of these, the selectivity problem, is the main reason why multivariate calibration methods are needed. This important aspect is mentioned briefly in Sundberg's paper as lack of specificity, but I think the problem is basic enough to deserve a more detailed discussion. This discussion will here be based on NIR (near infrared) spectroscopy where this problem is of particular importance. In this area, the lack of selectivity most often comes from the fact that individual spectra for the chemical constituents in the chemical compound are overlapping. This means that it is not possible to find one particular variable (wavelength) that contains all the information that we need about the chemical constituent to be calibrated for. For instance, when analysing wheat, it is known that each signal (wavelength) has a contribution from each of the constituents starch, protein and water. In

addition to this, physical effects like difference in light scatter level (due to particle size effects) preclude the situation further. An example of a serious selectivity problem is given in Martens & Næs (1989, p. 7). In this case wheat samples are measured by NIR spectroscopy and the main focus is calibration of protein%. The most dominant visual effect spotted in the spectra has nothing to do with the chemistry at all, it is just a light scatter effect. The correlation between the wavelength with the highest correlation with protein% is equal to 0.31, which is far too small for serious calibration work. Combining several wavelengths in a multivariate PCR approach, however, gave a correlation between measured and predicted protein % equal to $R = 0.97$. In other words, using a multivariate calibration method solved the selectivity problem!

The next problem is the collinearity problem. This is related to redundancy in the vector of X-variables. Another way of putting it is to say that there are too many variables compared to the number of systematic phenomena in the "spectrum". It occurs extremely often in practice (especially in chemistry) and is one of the reasons why methods like the PLS and PCR, which solve the problem elegantly, have become so popular. It is obvious that from an information point of view, more measurements of the same phenomena can never be a problem. It first starts becoming a problem when standard classical least squares (LS) methods for calibration are being used. Using the LS criterion, the problem can, however, sometimes be solved by variable selection. More frequently, however, the problem is solved by data compression (PCR, PLS) or one of the other methods mentioned by Sundberg. It should be mentioned that sometimes it can be advantageous to combine data compression with deletion of variables. An important reason for this is that certain variables may have a more complicated relationship to the reference value Y than others. Therefore, it may be advantageous to get rid of these complicated variables before the actual regression takes place (see e.g. Brown, 1992).

The outlier problem is discussed only briefly by Sundberg. There are many different reasons for outliers to occur and it is important to have a number of different criteria available to get insight into what is causing the problem. For instance, an outlier problem in chemistry can occur if

(i) the sample has extreme chemical composition (large or small concentration);
(ii) one or several of the peaks in the spectrum are shifted;
(iii) the instrument is out of order; or
(iv) the sample does not belong to the actual population.

All these different cases will have different effects on the model and/or the predictions. It is extremely important to have tools, so-called diagnostics, that can be used to give insight into which of these phenomena causes an outlier problem to occur. One will seldom be able to diagnose exactly what has happened, but a great deal of information can be provided. The diagnostics discussed by Sundberg can handle/detect some of these situations, but sometimes more outlier information can be provided. Again we will point the methods PCR and PLS as methods of special interest. For these methods a number of different criteria like residuals (X and Y residuals), leverage and influence have been developed to provide information about possible outliers. Leverage represents position within the component space, X-residual represents lack of fit of X to the component space and the Y residual represents the lack of fit of Y to the regression model. Together these tools provide a set of diagnostics that can be used to detect any of the situations mentioned above, both for individual variables and for the spectrum as a whole. For instance a large leverage only is clearly an indication that this sample belongs to the model (i.e. to the space estimated), but the sample is abnormally positioned within the model. This may be a clear indication of an abnormal chemical composition. All these tools are easy to visualize geometrically. In addition, they have more or less the same interpretation for all methods in the class of

regression methods (PCR, PLS, latent root regression (LRR), continuum regression (CR) etc.). We refer to Martens and Næs (1989, p. 272) for an application where the diagnostics mentioned above are tested.

The non-linearity problem is probably the least important of the above mentioned problems, but still it plays an important role in many chemical applications. This problem is discussed in an example in Sundberg's paper and in a discussion of a number of data pre-treatment methods. Here we will draw the reader's attention to a way that non-linearity problems can be solved by a technique which is very similar to PLS and PCR. The method has proven to give very precise and reliable results in chemistry and shares many of the same advantages as PCR and PLS (see the criteria above). First of all it should be mentioned that a number of individual non-linearities (i.e. non-linear relations between Y and each individual X) can be compensated for by adding extra factors in the PLS and PCR models. This means that it is often a projection into the large multivariate space that is reasonably linearly related to Y even though the space of variability (in X) may be highly curved. In some cases, however, a non-linear method (or alternatively a transformation) is necessary. The method to be discussed here is the so-called locally weighted regression (LWR). It is based on an idea presented in Cleveland & Devlin (1988), but is extended to handle multivariate colinear data. The algorithm can be found in Næs *et al*. (1990) and goes as follows.

1. First perform a PCA of the full set of data.
2. Use only the A first components.
3. For each new sample to be predicted, find the C samples in the calibration set which are closest within the A dimensional subspace.
4. Perform a weighted linear regression for the C samples and A components. The weights are dependent on the distance to the new sample.
5. Predict the new sample using the estimated equation.

Note that this is a natural extension of PCR obtained by adding an extra parameter C. The parameters C and A can be determined by cross-validation or prediction testing in the same way as A is always estimated in PCR. The method can be extended and modified in different ways (see Wang *et al*., 1994). Since both the distance and regression is computed only for the A first principal components, the collinearity problem is solved in an easy way. Note also that since the method is based on the same building blocks as the PCR, i.e. linear regression and PCA, the same diagnostics and interpretation tools as used for PCR carry over to this situation. It should be mentioned that it is often found that with this method, good results can be obtained even for a very small number of components, for instance 2 or 3. This can make the method more robust than many other techniques, since the most reliable variability is most often found in the first few components. Note that updating of the method is easy since this corresponds to just adding new data to the data bank. Collecting and keeping all previous data can, however, also be considered a drawback with the method.

Additional references

Cleveland, W. & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83**, 596–610.

Næs, T., Isaksson, T. & Kowalski, B. R. (1990). Locally weighted regression and scatter correction for near-infrared reflectance data. *Anal. Chem.* **62**, 664–673.

Wang, Z., Isaksson, T. & Kowalski, B. R. (1994). New approaches for distance measurement in locally weighted regression. *Anal. Chem.* **66**, 249–260.

SAMUEL D. OMAN

Hebrew University, Jerusalem

I congratulate the author on a very well-written and stimulating review. I would like to illustrate a prediction technique which Professor Sundberg did not mention in his discussion of direct regression in section 3. Suppose first that there are more observations than variables, ignore the intercept and assume normal errors with known variance. This gives the multiple regression model

$$Y_{n\times 1} = \mu + \varepsilon = X\beta_{q\times 1} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I). \tag{1}$$

Reasoning as in section 3.3 suggests that for strongly multicollinear data, predicting y should depend primarily on the first k PCs, for some $k < q$. This corresponds to the submodel

$$\mu \in \text{span}(X_k) \tag{2}$$

where $X_k = XT_k$ and the k columns of T_k are the first k eigenvectors of $X^{\mathrm{T}}X$. Consider then Sclove's (1968) modification of the James–Stein (1961) estimator,

$$\beta^*(k) = \hat{\beta}(k) + \phi(k)[\hat{\beta} - \hat{\beta}(k)] \tag{3}$$

for

$$\phi(k) = \left[1 - \frac{(q-k-2)\text{SSE}/(n-q+2)}{F_k}\right]^+. \tag{4}$$

Here, $\hat{\beta}$ denotes the LS estimator from the full model (1), $\hat{\beta}(k)$ is the PC estimator based on the first $k < q-2$ components, $F_k = \text{SSR} - \text{SSR}(k)$ where SSR and SSR(k) are the regression sums of squares from models (1) and (2) respectively and SSE is the error sum of squares from the full model.

This estimator has several desirable properties. (i) It protects against overshrinking by smoothly weighting $\hat{\beta}(k)$ and $\hat{\beta}$ according to the F-statistic we would use for making a preliminary test of model (1) vs model (2) (Sclove *et al*., 1972; Bock *et al*., 1973; Yancey & Judge, 1976). (ii) It is an empirical Bayes estimator which incorporates prior beliefs both in (2) and that the explanatory variables are intrinsically correlated as in the observed X (Raiffa & Schlaiffer, 1961; Tiao & Zellner, 1964; Efron & Morris, 1973; Zellner, 1983; Oman, 1984, 1985). (iii) Because $\beta^*(k)$ is a Stein estimator, $\text{PMSE}(\beta^*(k), \beta) < \text{PMSE}(\hat{\beta}, \beta)$ for all β, where for an estimator β^*, $\text{PMSE}(\beta^*, \beta) = E\|X\beta^* - X\beta\|^2$. Choosing a good submodel (2) can give substantial PMSE improvement (Jennrich & Oman, 1986; Oman, 1991; Oman *et al*., 1993).

Note that ridge estimators generally do not provide PMSE domination (Casella, 1980). Also, the arguments in (ii) suggest that ridge estimators (with contraction to the origin) correspond to the prior beliefs that there is no connection between y and x ($\beta = 0$), and that the components of x are "really uncorrelated"—neither of which seems reasonable for the problem at hand.

The catch with (3) is choosing a good value of k: minimal shrinkage occurs if k is too small (since then F_k in (4) is very large) or too large (since then $\hat{\beta}(k)$ is close to $\hat{\beta}$). I shall use the multiple-shrinkage approach of George (1986a, b), illustrated by George & Oman (1996). This "hedges its bets", using a weighted combination

$$\beta^*(\text{MS}) = \sum_{k=1}^{q-3} \rho(k)\beta^*(k). \tag{5}$$

The weights $\rho(k)$ emphasize those $\beta^*(k)$ which shrink more, and are chosen to give PMSE domination of $\hat{\beta}$.

I now illustrate using the waste water data of section 3.2, kindly supplied by Professor Sundberg. When $n < q$ there are no degrees of freedom for SSE, so (3) and (5) must be modified. I therefore defined a reduced q_{full}-dimensional "full" model by converting to principal components, graphing the $n-1$ regression coefficients in PC units (analogous to Sundberg's fig. 4), and choosing q_{full} to be the largest value of k before the point at which wild coefficient oscillation began. This generally gave $q_{\text{full}} = 35$. I examined the following estimators:

(a) The LS estimator using the q_{full}-dimensional model.
(b) PC(1), ..., PC(q_{full}), the PC estimators for varying subspace sizes.
(c) PC(MS), the multiple-shrinkage PC estimator suggested in George & Oman (1996).
(d) Stein(1), ..., Stein($q_{\text{full}} - 3$), the Stein estimators defined by (3).
(e) Stein(MS), the multiple shrinkage Stein estimator β^*(MS).

Observe that neither (b) nor (d) define proper estimators as discussed at the beginning of section 3.7, as they do not specify a rule for choosing k. Test set MSEP results for Sundberg's three types of 65/60 splits follow.

Random 65/60 splits

Column 1 of Table 1 shows the average MSEP over 256 random 65/60 splits. The entries labelled "PC(best)" and "Stein(best)" are average MSEPs when, for each split, the value of k giving the lowest MSEP was used. These MSEPs are not realistic descriptions of predictive ability, but do indicate "potential" MSEP improvement. Also, if external and internal CV results for PC and Stein are typically similar (Sundberg's fig. 7a suggests they may be for random splits, but not for the other splits considered), then the "best" MSEP values should be close to those obtained when k is chosen by internal CV. We see that PC(MS) reduced the LS MSEP by 33%, while Stein(MS) obtained a reduction of 24%.

Table 1. *Test set MSEP values*

Estimator	Split		
	Random 65/60	First 65/Last 60	Last 60/First 65
Full model LS	0.2363	0.4786	1.4547
PC(MS)	0.1586	0.6052	1.2010
PC(best)	0.1339	0.3262	0.1119
Stein(MS)	0.1724	0.4125	1.3589
Stein(best)	0.1499	0.3579	0.1292

First 65/last 60 split

This corresponds to Sundberg's Fig. 8a. Column 2 of Table 1 shows that PC in fact increased the LS MSEP by 26%, corresponding to the difficulties discussed by Sundberg.

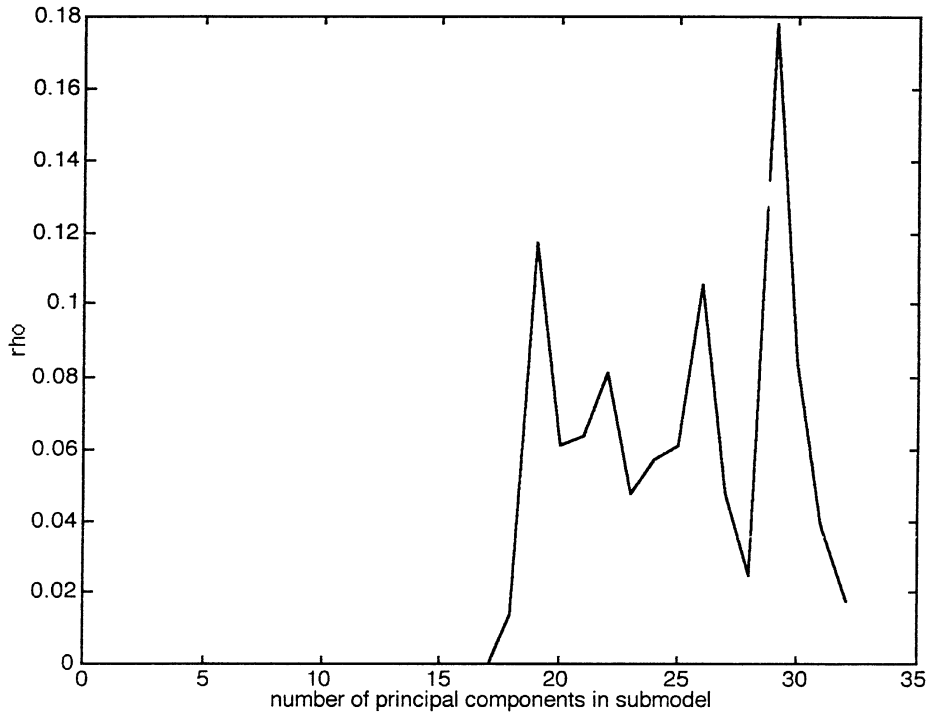


Fig. 11. Weighting function ρ for last 60/first 65 split.

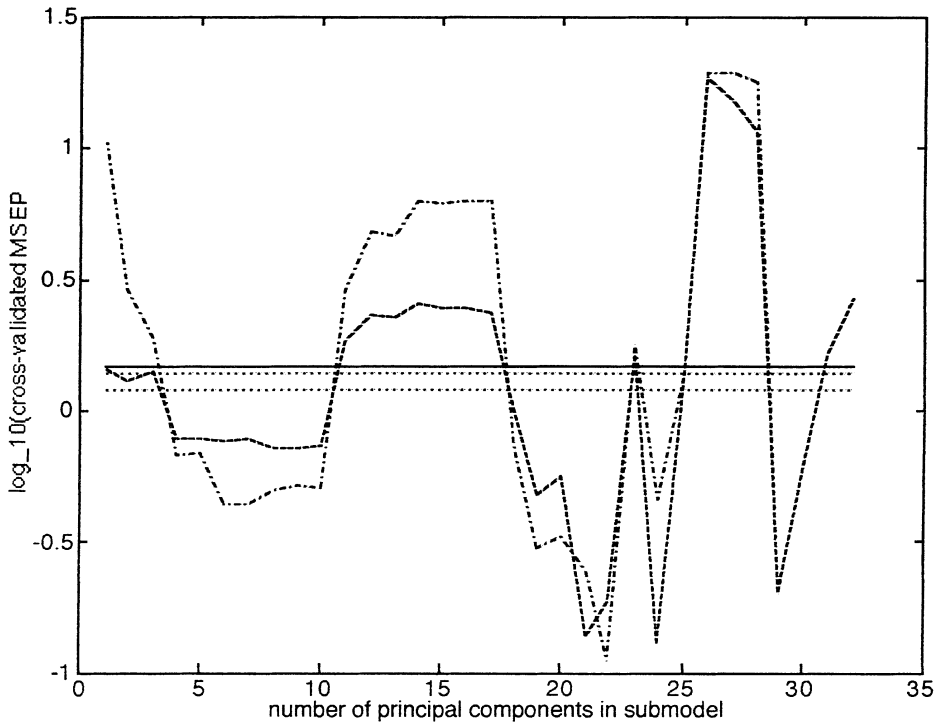


Fig. 12. Test set MSEP for last 60/first 65 split. Meaning of lines as in Fig. 13.

(Choosing k to minimize the C_p of Mallows (1973) gave the same result.) On the other hand, Stein gave a moderate MSEP reduction.

Last 60/first 65 split

The jagged form of the ρ curve (Figure 11) indicates a discontinuous dependence of y on the principal components, which is reflected in the (logged) MSEP (Figure 12). Column 3 of Table 1 shows that PC and Stein slightly reduced the MSEP of LS (Table 1, column 3). Note that ρ, which is computed entirely from the calibration data, picks up the discontinuous behaviour of the future predictions, while the internally cross-validated MSEP (for PLS; the solid curve in Sundberg's fig. 9a) does not.

These numerical results suggest the following tentative conclusions.

1. Stein contraction can usefully be extended to the case of more variables than observations. Obviously, additional work on choosing q_{full} is needed.
2. PC can dramatically improve LS prediction, provided future observations are similar to those in the calibration sample. If not, the predictions need not be better, and might even be worse.
3. Stein also gives dramatic improvements, though not quite as great as PC's. Stein also offers some protection when predicting at dissimilar future observations.
4. When predictive ability is a relatively smooth function of submodel dimension, multiple-shrinkage Stein comes close to mimicking the behaviour of Stein with optimally chosen subspace.

Additional references

Bock, M. E., Yancey, T. A. & Judge, G. G. (1973). The statistical consequences of preliminary test estimators in regression. *J. Amer. Statist. Assoc.* **68**, 109–116.

Casella, G. (1980). Minimax ridge regression estimation. *Ann. Statist.* **8**, 1036–1056.

Efron, B. & Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 17–130.

George, E. (1986a). Minimax multiple shrinkage estimation. *Ann. Statist.* **14**, 188–205.

George, E. (1986b). Combining minimax shrinkage estimators. *J. Amer. Statist. Assoc.* **81**, 437–445.

George, E. I. & Oman, S. D. (1996). Multiple-shrinkage principal component regression. *Statistician* **45**, 111–124.

James, W. & Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley symp. math. statist. probab.* **1**, 361–379.

Jennrich, R. I. & Oman, S. D. (1986). How much does Stein estimation help in multiple linear regression? *Technometrics* **28**, 113–121.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **16**, 661–675.

Oman, S. D. (1984). A different empirical Bayes interpretation of ridge and Stein estimators. *J. Roy. Statist. Soc. Ser. B* **46**, 544–557.

Oman, S. D. (1985). Specifying a prior distribution in structured regression problems. *J. Amer. Statist. Assoc.* **80**, 190–195.

Oman, S. D. (1991). Random calibration with many measurements: an application of Stein estimation. *Technometrics* **33**, 187–195.

Oman, S. D., Næs, T. & Zube, A. (1993). Detecting and adjusting for non-linearities in calibration of near-infrared data using principal components. *J. Chemometrics* **7**, 195–212.

Raiffa, H. & Schlaifer, R. (1961). *Applied statistical decision theory.* Harvard University, Graduate School of Business Administration, Boston.

Sclove, S. (1968). Improved estimators for coefficients in linear regression. *J. Amer. Statist. Assoc.* **63**, 596–606.

Sclove, S., Morris, C. & Radhakrishnan, R. (1972). Nonoptimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.* **43**, 1481–1490.

Tiao, G. C. & Zellner, A. (1964). Bayes' theorem and the use of prior knowledge in regression analysis.

Biometrika **51**, 219–230.
Yancey, T. A. & Judge, G. G. (1976). A Monte Carlo comparison of traditional and Stein-rule estimators under squared error loss. *J. Econometrics* **4**, 285–294.
Zellner, A. (1983). Applications of Bayesian analysis in econometrics. *Statistician* **32**, 23–34.

SVANTE WOLD
Umeå University

It is important both for statistics and chemistry that statisticians are becoming interested in the multivariate calibration (MC) problem. MC is today used extensively in the petroleum, pharmaceutical, food and beverage, polymer, and paper and pulp industries for process monitoring, on-line process analysis, and for rapid determination of concentrations and other properties in complicated samples from environmental and medical studies.

Rolf Sundberg is congratulated for (a) realizing the importance of the MC problems, and (b) for writing an excellent and extensive overview of the MC problem seen as a statistical regression problem.

To a chemist like myself, it is surprising that most statisticians—i.e. Sundberg, Brown, Friedman, Breiman and others—continue to stick to the regression model when it is clear that this is not an appropriate model for MC. The assumption that the signal matrix U has full rank (when the number of samples is larger than the number of variables) is, in my view, a major reason why statisticians continue to see this problem as "difficult".

The variation of the u-variables is due to the variation of a limited number, say m, of chemical compounds in the samples. Each of these m compounds has a spectrum that covers all or most of the spectral range. Under ideal circumstances, i.e. when Beer's "law" is valid for all compounds in all samples, the $(n \times q)$ U-matrix of the sample spectra can be well modelled by a latent variable model; $U = C * S' + E$ (concentrations $*$ spectra $+$ noise), where the number of components or factors equals m. Under non-ideal, more typical circumstances (interactions between the compounds, dimerization, saturation effects, etc.), the same model may still apply, but then the number of "factors" is larger than m.

Hence, the rank of U is, by definition, just m, which is usually far smaller than q. Moreover, if the latent variable model is reasonable for U, then the regression model is biased, even in the situation when $n \gg q$ and p. The approaches based on latent variables, such as principal components regression (PCR), PLS-regression, and factor analysis models, have the right form and model assumptions to be suitable for MC. This is well discussed in Martens & Næs (1989), Kvalheim (1992), as well as in the recent thesis by Burnham (1997). Hence, these "latent variable regression" approaches should not be seen just as regularization methods motivated by collinearities in many spectral variables, but being more realistic models than ordinary (multiple) regression, and hence preferable also when the number of spectral variables is smaller than the number of samples.

Sundberg is probably right in that the available methods for MC, i.e. PCR, PLS, and ridge regression (RR), do not differ much in their predictive performance with fairly linear data, at least not with typical spectral data with good precision. However, when it comes to the interpretation of the model parameters, and diagnostics of model inadequacies, outliers among samples, and inhomogeneities in the data, the methods perform differently. The latent variable methods (PCR and PLS) provide a model of the signal matrix (U), which is invaluable for the understanding of the chemistry of the situation, as well as outlier and inhomogeneity detection. This is a main reason why these methods are much more popular in chemistry than methods emerging from statistics such as RR. Also, because of the wrong model form of RR, the

coefficients (B) often are highly biased in comparison with those of PCR and PLS, and hence little use for model interpretation (Höskuldsson, 1996).

For some reason, latent variable based models are viewed with suspicion by many statisticians. This may have historical or other reasons, but at least in situations like MC where the latent variable model is derivable from first principles (and the multiple regression model is not), these models should be taken seriously. It would be interesting to see Sundberg start from the latent variable regression model, and improve our statistical understanding based on this chemically more reasonable model.

In MC, there are many problems that are still little investigated. These include data preprocessing (transformation, scaling, filtering), variable selection, and constrained modelling. Here very simple ideas can still lead to large improvements in the precision of the results and the predictability of new data. This is exemplified by the recently developed "orthogonal signal correction" (OSC), where only such structure is removed from U (the spectral matrix) that is orthogonal to t (the concentration vector(s)). With OSC the accuracy of predictions often improves by 30–50% in comparison with standard methods such as multiple signal correction (Wold *et al*., 1997).

As pointed out by Sundberg, another vast application area needing much more work is MC with 3-way data, 4-way data, and so on. Here a collaboration between chemists and statisticians will, hopefully, lead to interesting and improved methods as well as an improved understanding of both statistical and chemical issues, and I am very happy that a good statistician like Sundberg is taking an interest in these problems.

Additional references

Burnham, A. (1997). Thesis, Dept. of Mathematics and Statistics, McMaster University, Hamilton, Ontario,
Höskuldsson, A. (1996). *Prediction methods in science and technology, vol. 1, basic theory.* Thor Publishing, Denmark.
Kvalheim, O. (1992). The latent variable, an editorial. *Chemometrics Intell. Lab. Systems* **14**, 1–3.
Wold, S., Antti, H. & Öhman, J. (1997). Orthogonal signal correction of NIR spectra. *Chemometrics Intell. Lab. Syst.* submitted.

Reply to Discussion

ROLF SUNDBERG

I am grateful for the interesting discussion of my paper by this selection of distinguished calibration statisticians and chemometricians, who represent a lot of experience and insight. Nevertheless they are found to propose quite different methods. Primarily this is because they look at the world differently, but it also reflects the fact that there is no single superior method. Phil Brown advocates Bayes-motivated procedures and variable selection, while Sam Oman suggests a Stein contraction type method. Not surprisingly Martens, Næs and Wold speak for the latent variable methods, and Martens and Wold argue against RR and similar methods when the chemistry (or physics) is to be understood. I agree that PLS and PCR have much better chances to provide structural information, and in particular to tell the "chemical rank" of a set of data. But the phase of understanding should come before that of constructing a predictor for routine usage, and different phases should not necessarily use the same statistical methods.

Let me stress that I regard PLS and PCR as providing good and reasonably efficient methods of prediction in a wide range of situations with many explanatory variables. Martens, Næs and Wold deserve honour for the contributions to the popularity of these methods in chemistry and other fields. I expect that the scepticism about especially PLS found in the statistical community will gradually diminish when it is realized how PLS answers an optimality criterion and how PLS is related to other regression methods. However, these relationships and empirical evidence also indicate that someone less happy with latent factor methods can predict similarly with more direct methods.

While statisticians have been sceptical of its use, chemometricians have provoked them by the opposite tendency to view PLS as the holy grail. Personally, I remember finding PLS magical when I was a (designated) opponent to Harald Martens' doctoral thesis in 1985, but today I think there is not much magic left, due to important clarification work by a number of people. Svante Wold tries to retain a bit of the magic by stressing not only that latent variable methods represent more realistic models than the ordinary multiple regression model, but also that the latter model is biased and not derivable from first principles, etc. I must object to him on some of these points, and in doing so also briefly respond to his demand that I should start off from the latent factor model.

First, the concept of chemical rank can only be a more or less adequate approximation in practice, when pollutants and impurities will be counted or not depending on their respective concentrations (and influences), and when there is a large number of minor influential factors from non-linearities, interactions, saturation effects etc. Sometimes there is a wide gap in influence between one small group of factors and all others, and then the chemical rank is not very ambiguous and could be quite informative. But what should the chemical rank be in the waste-water example, and how should it be determined?

Second, say now there is a latent variable model generating the data. What does Svante Wold mean by saying that the regression model is biased? If both t and u are generated linearly from a latent vector s, with noise added, and the latent vectors are regarded as fixed, it is true that the regression model is inadequate, in the sense that we have an errors-in-variables type relation between t and u. Hence, ordinary regression of t on u or u on t will not yield unbiased estimators of the parameters of this model. However, it is crucial to note that our task is prediction, not estimation or structure identification. If s and the noise are (as usual) assumed normally distributed, so is t and u jointly, and the best unbiased predictor of t is its theoretical (linear) regression on u. Hence, the linear regression model *is* adequate to our prediction problem even with an underlying latent structure.

Furthermore, the regression model apparently involves fewer parameters in the linear structure than a typical latent factor model in its loading matrices. Therefore it is not surprising to me that direct regression methods can do as well as methods explicitly involving the assumed underlying latent structure. To this comes perhaps the model robustness advantage: a direct regression method clearly does not depend on the existence of a latent structure, whereas this is less clear as regards the latent model methods.

As a final point of divergence with Svante Wold, I would not condemn RR for its bias, in particular not by referring to Höskuldsson (1996). I do not know how to tell the bias of RR when the true value is missing, but I note that two examples of the book give negative values of squared bias, and the third example (NIR data) an extremely high bias. The reason for the latter is clear: Höskuldsson's bias estimation is seriously biased, bound to exaggerate. My own computations on the NIR data show that, as usual, RR, LSRR, PLS and PCR all yield very much the same minimal MSEP (in CV leave-one-out) and similar predictors. Besides, in my paper I have rather advocated LSRR, which is likely to have an even smaller bias than RR.

I do not think Phil Brown need worry about negative ridge constants, if he restricts himself to RR or LSRR (first factor CR), for which a negative value of the optimal ridge constant would be a remarkable exception. Negative ridge constants help us understand CR and the roles of special methods within CR, in particular PCR.

Phil Brown reports having been successful with variable selection in the situation of Fig. 9. This is when the fitted model fails in extrapolation. It is natural that a fit from just a few wavelengths is less vulnerable to overfitting in a situation like this, but in other cases selection of a limited number of variables could imply a substantial loss of information. As for the cost aspect, it is simpler, and therefore often also cheaper, to let the instrument automatically scan over a spectral region in a regular fashion than to seek out an irregular selection of wavelengths.

Brown and Martens comment on the important heterogeneity aspects of the waste-water data, for example pointing out that one should check the validity from one time period to the next. They are both right in their remarks, at least in principle. I must stress, however, that the aim of collecting these data was to construct, if possible, a calibration model that could be used over time and location. Therefore two time periods and several basins were covered. From that point of view the danger to remember concerns extrapolating from just a small set of different background conditions (period, basin), because by usual cross-validation we risk severe overfitting by fine-adjustment to the particular conditions present in the study.

Tormod Næs contributes a nice general discussion and an important discussion about outliers and non-linearities. My reasons for treating these problems so briefly were a wish to keep down the page number combined with a feeling that I could not write anything important not already well presented in the books by Martens & Næs and Brown. One remark to Næs, however: the fact that PCA is the natural tool for studying the distribution in x-space does not imply that PCR should be recommended for regression; again, different problems might require different tools.

Sam Oman supplements my comparisons of various regression methods by a few PCR-related variants. My scepticism against Stein shrinkage as a general principle finds support by observing that Stein shrinkage was found somewhat less efficient than PCR. The weighted multiple shrinkage approach is conservative by nature, but perhaps typically too pessimistic.

My basic feeling from the discussion, however, is of agreement. Although some opinions diverge and we might prefer different methods, I think what we have in common is more important, whether we are chemists or statisticians, Bayesians or frequentists. This could perhaps be characterized as a wish to find adequate and efficient statistical methods for extracting the available information out of multidimensional (calibration) data.