Mathematical Statistics
Stockholm University

Small sample and selection bias effects in calibration under latent factor regression models

Rolf Sundberg

Research Report 2006:13

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2006:13**,
http://www.math.su.se/matstat

Small sample and selection bias effects in calibration under latent factor regression models

Rolf Sundberg*

December 2006

Abstract

We study bias of predictors when a multivariate calibration procedure has been applied to relate a scalar y (concentration of an analyte, say) to a vector x (spectral intensities, say). The model for data is assumed to be of latent factor regression type, with multiple regression models and errors-in-variables models as special cases. The calibration procedures explicitly studied are OLSR, PLSR and PCR. When y has been more or less systematically selected in the calibration in order to achieve increased variation (overdispersion), a practical device to increase precision, this leads to biased coefficients in the predictor, possible to see when observed y is regressed on predicted $\widehat{y}(x)$ for a separate validation set. Another bias effect is a sample size effect, increasing with reduced calibration sample size and with increasing dimension of x (absent when x is univariate). Formulae are given for these bias effects, both separately and in combination, and the formulae are illustrated and compared with simulation results. As a qualitative example, PLSR and PCR are less sensitive than OLSR to small samples, but equally sensitive to selection.

Key words: Bias, chemometrics, cross-validation, errors-in-variables, multivariate calibration, OLS, overdispersed training set, PCR, PLSR, prediction,

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: rolfs@math.su.se. Website: www.math.su.se/∼rolfs

1 Introduction

In multivariate calibration, practice is dominated by explicit or implicit multiple regression models and latent factor models relating, say, concentration y of a substance linearly to a measurement vector x of absorbances measured at specified wavelengths. Typical statistical methods used are ordinary least squares (OLS) multiple regression for low-dimensional x, and latent variable methods such as PCR and PLSR for higher-dimensional x. Such methods are justified in the so-called *natural calibration* setting, when we regard unknown concentrations as quantities to be predicted, corresponding to chemical specimens randomly sampled from a natural infinite population of possible such samples. This assumption implies a joint randomness of x and y, and hence that prediction of unknown y from measured x is justified. The best linear predictor of y would be the population linear regression of y on x. In a latent factor model we can extend this result, by explicitly allowing additional measurement errors and similar uncertainty in both x and y.

If the underlying relationship between x and y were known, nothing would be controversial with such a procedure. However, complications appear as soon as the relationship between x and y must be estimated from training (calibration) data. Admittedly, the classical univariate results by Lindley (1947) indicate that if the calibration specimens are sampled from the same population as the future specimens, whose concentrations are to be predicted, we need not worry. However, it will be seen below that this is not true for OLS in a multi-dimensional setting. The result to be shown is strongly related to those of Copas (1983, 1987), who demonstrated that the OLS predictor in a multiple regression model should be shrunk, and who proposed estimators of the shrinkage factor. The main aim of the present paper is different, namely to investigate the shrinkage appearing when the calibration specimens do not represent the natural prediction population. In the context of a structural linear model and with applications to different seismological measurements in mind, this bias was discussed by Ganse et al. (1983). In a multiple regression model setting, Jones & Copas (1986) investigate the robustness of Copas' shrinkage estimator to differences between the calibration population and the validation population. More generally, however, the situation appears to have been largely neglected in the literature.

In chemometrics, we may be tempted to select calibration specimens with a larger spread in concentrations than natural, because the more widely

spread the data points are, the more precisely can a linear relationship be estimated. It is particularly likely to occur in calibrations where the concentrations of the training samples can be completely controlled, but even otherwise it is quite common with some kind of "pseudo-selection" aimed at increasing variation and precision. However, the bias caused by selection can be substantial, as we will see, and it will not be detected in leave-out cross-validation. Therefore, the reader is warned against uncritical use of selection in calibration, and against sloppy statements such as "don't bother, it's just to run a PLS", which have been heard in the context of multivariate calibration.

We start the theory parts with the univariate case, which allows us more easily than in the multivariate case to express and to understand the basic relationships. After that, we turn to OLS and latent factor regression methods in the multivariate case. The bias formulae derived will be compared with the results of simulations from a model representing a multivariate data-set from fat-grading in pig slaughteries. The same pig grading data were used more directly in a companion paper (Sundberg, 2006), where the effects were illustrated by sampling from this dataset, but no formulae discussed.

2 Univariate linear calibration, without and with selection

2.1 Natural calibration

In a *natural calibration* situation, calibration specimens are randomly selected from the target population for the calibration. This makes both x and y inherently random, through an underlying latent linear relationship. We additionally allow random measurement type errors in both x and y in the statistical model for data:

$$\begin{array}{rcll} x_i & = & \xi_i + \delta_i\,, & \delta_i \sim N(0, \sigma_\delta^2), \\ y_i & = & \eta_i + \epsilon_i\,, & \epsilon_i \sim N(0, \sigma_\epsilon^2), \\ \eta_i & = & \alpha + \beta\xi_i\,, & \xi_i \sim N(\mu, \sigma^2). \\ i & = & 1, ..., n. & \end{array} \tag{1}$$

We think of x_i as the observed instrument response for specimen i, and y_i as a concentration (or some other property) determined by a reference method

for the calibration specimens. Symbols δ and ϵ represent unobservable contributions of measurement error type, mutually uncorrelated, with variances σ_δ^2 and σ_ϵ^2, respectively. In classical calibration, ϵ is assumed to vanish, i.e. $\sigma_\epsilon^2 = 0$. Symbols ξ and η represent linearly related latent random variables, or equivalently here, in the univariate case, interpreted as being the same latent variable on different scales. It means that ξ has an unknown expected value μ and a variance σ^2, and $\sigma_\eta^2 = \beta^2\sigma^2$. The model terms in (1) are also assumed normally distributed, partially for notational convenience. This is not a crucial assumption, see below.

We should have two possible interpretations of model (1) in mind. In practice we could have a mixture between these two pure situations:

- η is the true concentration of the specimen, y is a determination of η with measurement error ϵ, and ξ is the ideal, error-free instrumental response according to the Lambert–Beer law. Without the assumption of a randomly sampled specimen, this is the classical errors-in-variables regression (EIVR) model, see for example Gleser (1991).

- Alternatively, y is the true concentration and η (or equivalently ξ) is some other intrinsic property of the specimens, which relates concentration imperfectly to instrument response.

If all model parameters are known, the theoretically best predictor of y or η from x for a new specimen from the same population is the conditional expected value, given x, $E(y|x)$. 'Best' means unbiased with least variance. This is the same as the theoretical regression of y or η on x. A sufficient condition for this theoretical regression to be automatically linear in x is that ξ and δ be normally distributed, so that x and ξ are jointly normal. When the theoretical regression is linear, the best predictor ($\widehat{y}$ or $\widehat{\eta}$) can be expressed as

$$\widehat{y}(x) = \widehat{\eta}(x) = E(y) + b\,(x - E(x)), \tag{2}$$

where

$$E(y) = \alpha + \beta\,E(x) = \alpha + \beta\,\mu \tag{3}$$

and

$$b = \mathrm{cov}(y, x)\,\mathrm{var}(x)^{-1}. \tag{4}$$

Here, expressed in terms of (1)

$$\mathrm{cov}(y, x) = \mathrm{cov}(\eta, \xi) = \beta\,\sigma^2,$$

$$\mathrm{var}(x) = \sigma^2 + \sigma_\delta^2,$$

so

$$b = \beta/(1 + \sigma_\delta^2/\sigma^2). \tag{5}$$

The deflation of β seen in (5) precisely compensates for the random "error" in x. Note the important interpretation of this regression: $\widehat{y}$ (and $\widehat{\eta}$) is a conditional expected value over the joint randomness in ξ, δ and ϵ, and the regression coefficient b depends explicitly on the relationship between the variances of ξ and δ.

This does not only imply that formula (2) will be the best predictor, but also that it is an unbiased predictor for each given (observed) x,

$$E(\widehat{y}(x) - y|x) = E(\widehat{\eta}(x) - \eta|x) = 0. \tag{6}$$

Note that the predictors $\widehat{\eta}$ and $\widehat{y}$ are the same, so the predictor of the true concentration is the same, whether this concentration is represented by η or y. On the other hand, $\widehat{\eta}$ and $\widehat{y}$ will have different predictor variances, but precision is not the primary topic of the present study.

In practice the parameters of the predictor (2) are unknown, of course, so they must be estimated from training data in a calibration, typically by forming the linear OLS regression of y-data on x-data. This is the same as saying that we estimate the unknown population quantities $E(y) = E(\eta)$, $E(x) = \mu$ and b by the calibration sample means $\bar{y}$ and $\bar{x}$, and the sample regression coefficient $\widehat{b}$ (via the sample covariance and x-variance), respectively. In particular, the calibration sample yields

$$\hat{y}(x) = \hat{\eta}(x) = \bar{y} + \hat{b}(x - \bar{x}) \tag{7}$$

Note the crucial implicit appearance of the population quantities in the predictor above. If the calibration sample is taken from the same natural population, then $E(\widehat{b}) = b$, but otherwise we must expect problems. In later sections, we will examine bias effects caused by selection in the calibration.

The calibration can be checked by use of a test set intended to represent the population in question. We may compare RMSEP values (or similar measures of predictive ability) for the test set with the corresponding leave-one-out cross-validation values from the calibration, and we may regress y on $\widehat{y}$ for the test set, to check if this simple regression is reasonably close to the ideal regression line $y = \widehat{y}$, so it can be taken to represent the relation

$$E(y|\widehat{y}) = \widehat{y}. \tag{8}$$

As is well-known, calibration data satisfy the ideal regression relationship $y = \widehat{y}$, when $\widehat{y}$ are the OLS-fitted values from the calibration, because regression residuals are uncorrelated with predicted values (and this holds irrespective of any population or model). Under cross-validation (leave-one-out or leave-k-out), approximately the same relationship will be seen. Hence, from the calibration data alone it will not be possible to see if the predictor construction is biased, even if we use cross-validation or some similar method for predictor testing. In order to have a chance to see a bias when plotting y against $\widehat{y}$, a separate validation sample is required, taken from the same population as the one in which the predictor will be used.

For an infinitely large validation set, the (population) regression of y on $\widehat{y}$ has the coefficient (cf. (4))

$$\frac{\mathrm{cov}(y,\widehat{y})}{\mathrm{var}(\widehat{y})} = \frac{\widehat{b}\,\mathrm{cov}(y,x)}{(\widehat{b})^2\,\mathrm{var}(x)} = \frac{b}{\widehat{b}}\,, \tag{9}$$

for a fixed calibration sample having the regression slope $\widehat{b}$. Coefficient (9) will of course differ from 1, because no calibration is perfect. For saying that the predictor lacks systematic error, it should be enough to require that the distribution of (9) over the randomness in the calibration is centered at 1. If the calibration sample is taken from the same population as the validation data, $\widehat{b}$ has median (and mean) b, so the median of (9) is 1, as desired. Since $\widehat{b}$ appears in the denominator of (9), we should be a bit careful when talking about the mean value of (9) itself, because this mean value does not exist under a strict normality assumption. We will return to the mean value in Section 3, when for higher-dimensional x it will be seen that the quantity corresponding to (9) can suffer from a serious shrinkage effect even though the calibration sample represents the population desired.

Conclusion: In univariate natural calibration, the OLS predictor $\widehat{y}(x)$ is median-unbiased for each x.

2.2 Selective calibration in the univariate case

Now suppose the calibration sample is aimed at yielding a higher spread in y-values or in the underlying η, and in this way to achieve less randomness in the estimate of the b-value. Such selection could be carried out in different ways, and a model interpretation of the selection depends on what y and η represent in model (1). Selection could be in terms of an intrinsic property

of the specimens, strongly related to the true concentration. We take it to be represented by η, but in principle this intrinsic property could be just correlated with η. However, unless it is also correlated with δ or ϵ, it is enough to consider its influence on η in terms of selection.

Alternatively a selection could be in terms of the true concentration itself (represented by either y or η). Typical in practice would be when there are standards to choose from, or when the calibration specimens can be prepared in the laboratory to have specified concentrations. Another relation to model (1) occurs when y represents a determination subject to substantial measurement error, and we imagine that selection is in terms of the y-values. This latter type of selection is perhaps less natural.

Selection can either be deterministic, typically represented by use of standards with predetermined concentrations, or be of a more or less random or haphazard character. The latter is almost necessarily the case if selection is in terms of an intrinsic property. In any case the important feature will be the spread of concentration values in the calibration, and the origin of that spread will be less important. Therefore we will assume that calibration can be regarded as random also under selection, but with an increased variance (overdispersion) in y or η, quantified by a *variance inflation factor* θ^2. We will focus on the effects of an increased variance, but temporarily we will also allow the mean value of y or η to differ from that of the natural population. We will not unnecessarily make assumptions about the form of the distribution of the calibration concentration values. In particular, we need not assume that they are normally distributed.

2.2.1 Selection in η

The model for the calibration data is assumed the same as (1), except for the distribution of ξ and η. We use suffix c to denote calibration mean and variance for ξ, that is μ_c and $\sigma_c^2 = \theta^2\sigma^2$, respectively. The linear regression that represents the special population, from which the calibration sample is (assumed) selected, is obtained by simply substituting μ_c for μ in (1) and b_c for b in (2), where in complete analogy with (4) and (5),

$$b_c = \mathrm{cov}_c(y, x)\mathrm{var}_c(x)^{-1} = \beta/(1 + \sigma_\delta^2/\theta^2\sigma^2). \tag{10}$$

If any of these population characteristics, μ_c or b_c, differs from the natural population, the predictor will be systematically misleading, and we will say

that it is biased. More precisely, given an observed instrument response x for a new specimen, the predictor (2) formed with parameters from the calibration will have a bias that may be expressed as

$$\begin{aligned} \text{bias}(\widehat{y}(x)|x) &= E(\widehat{y}(x) - y|x) \\ &= \alpha + \beta\mu_c + b_c(x - \mu_c) - \{\alpha + \beta\mu + b(x - \mu)\} \\ &= (\beta - b_c)(\mu_c - \mu) + (b_c - b)(x - \mu). \end{aligned} \tag{11}$$

Hence, as soon as the calibration and natural population mean values differ, i.e. $\mu_c \neq \mu$, and provided $\sigma_\delta > 0$, the selection will induce a systematic error in the predictor. If $b_c = b$, i.e. $\theta^2 = 1$, the bias is independent of x. The constant part of the bias is relatively trivial, and we will devote our interest towards the x-dependent last term of (11). This term may be expressed as

$$\begin{aligned} (b_c - b)\,(x - \mu) &= \left(\frac{1}{(1 + \sigma_\delta^2/\theta^2\sigma^2} - \frac{1}{1 + \sigma_\delta^2/\sigma^2} \right) \beta(x - \mu) \\ &= \frac{(\theta^2 - 1)}{(\sigma^2 + \sigma_\delta^2)(\theta^2\sigma^2 + \sigma_\delta^2)}\,\beta(x - \mu) \end{aligned}$$

The bias can be expressed in a quite different way. Let us from now on concentrate on the effects of overdispersion in the calibration, and assume $\mu_c = \mu$. If we consider a plot of y against $\widehat{y}$ for specimens from a population represented by (x, y), when calibration has resulted in (a fixed) $\hat{b}_c$, the population linear regression of y on $\widehat{y}$ will be

$$E(y|\widehat{y}) = E(y) + \frac{\text{cov}(y, \widehat{y})}{\text{var}(\widehat{y})}\,\{\widehat{y} - E(y)\} \tag{12}$$

The population regression coefficient is

$$\frac{\text{cov}(y, \widehat{y})}{\text{var}(\widehat{y})} = \frac{\hat{b}_c\,\text{cov}(y, x)}{\hat{b}_c^2\,\text{var}(x)} = \frac{b}{\hat{b}_c} = \frac{1 + \sigma_\delta^2/\theta^2\sigma^2}{1 + \sigma_\delta^2/\sigma^2}. \tag{13}$$

If the calibration is precise, so $\hat{b}_c = E(\hat{b}_c) = b_c$, and if the (x, y) population is the same as the population used in the calibration, so that $b_c = b$, then the regression line (12) is the 45° line through the origin. On the other hand, if the calibration sample is overdispersed, the regression coefficient (13) for the precise calibration will be $b/b_c = (1 + \sigma_\delta^2/\theta^2\sigma^2)/(1 + \sigma_\delta^2/\sigma^2)$.

Conclusion: Overdispersion in η by a variance inflation factor θ^2 will yield a systematic error in the predictor $\hat{y}(x)$, quantified by the shrinkage

factor

$$\frac{1+\sigma_\delta^2/\theta^2\sigma^2}{1+\sigma_\delta^2/\sigma^2} \tag{14}$$

in the regression of y on $\widehat{y}$.

2.2.2 Selection in y, ξ or x

We now think of a selection that has taken place with respect to y. Model (1) is still assumed for the relation between x and ξ on the one hand and y and η on the other. We first assume that y is the true concentration of the specimen, so η is some concentration-related intrinsic property of the specimens. Since y and η have the same mean values, we need not further discuss change in mean value. Thus, suppose selection is used only to increase the spread in calibration y-values, and that var(y) is increased by the factor θ^2. If the true concentration is thus affected, it might be natural that also var(η) = σ^2 is increased by the same factor θ^2, or approximately so, and likewise for var(ϵ) = σ_ϵ^2. However, σ_ϵ^2 does not enter the calculations of the previous section, or the result (14), so the results derived for selection with respect to η also hold for selection with respect to y, under the assumptions formulated above. It is not even required that var(ϵ) is increased by the same factor, if only var(η) is increased by approximately the same factor as var(y).

A somewhat different and perhaps not as realistic scenario would be selection with respect to y when y is the measured concentration, with measurement error. If measurement errors are not very large (and typically in analytical chemistry they are small), then var(η) and var(y) should be approximately equal even in this case, because var(y) cannot be much increased without a corresponding increase in var(η). Hence the previous result will still hold approximately, under mild assumptions.

Since there is a linear relation between ξ and η, selections with respect to ξ and η are equivalent. Selection with respect to instrumental response x is different. Assume that an increased variance in x, by such selection, corresponds to increased variances σ^2 and σ_δ^2 by the same factor. The formulae above then show that no bias will be generated by the selection. Another, apparently quite different variation is provided by Andersen et al (2003), who investigate in a simulation study of a univariate calibration situation if it pays to modify the least squares predictor, when it is known that the error variance σ_δ^2 in x is different in calibration than in prediction (and validation). More precisely they think of averaging over different number of

replicated measurements in calibration and in prediction. In the univariate case here, however, reducing σ_δ^2 is equivalent to increasing the variance σ^2 for ξ by the same factor, so their situation actually falls within the present set-up, provided that θ^2 is given the correct interpretation.

3 Latent factor models in multivariate calibration

We now extend the univariate latent factor model (1) to multivariate calibration situations, with a multi-dimensional instrument response x. The latent structure in x is then naturally multi-dimensional, too. We will formulate two versions of a model equivalent with the special case $\dim(y) = 1$ of the *general latent variable multivariate regression (LVMR) model* of Burnham et al. (1999, 2001), but also found for example in Martens & Næs (1989). In common chemometric notation it might be written:

$$\begin{aligned} x_i - E(x) &= P' t_i + \delta_i\,, \qquad \delta_i \sim N(0, \Sigma_\delta), \\ y_i - E(y) &= \eta_i + \epsilon_i = q' t_i + \epsilon_i\,, \qquad \epsilon_i \sim N(0, \sigma_\epsilon^2), \\ t_i &\sim N(0, I_a), \qquad \dim(t) = a < \dim(x). \end{aligned} \tag{15}$$

Here x_i and y_i represent observation i, t_i is the underlying latent vector, P is a coefficients matrix, q is a coefficient vector, and δ_i is now a random error vector with covariance matrix Σ_δ (replacing the variance σ_δ^2 of the univariate model (1)). It is no restriction to assume that the model components of t are standardized and uncorrelated. Note also that the model is not unique, in the sense that a random contribution to x, uncorrelated with y, can be added to t, accompanied by an increase in $\dim(t)$, or added to δ. Even more, it is no restriction to rotate the latent vector t, so that its first component becomes proportional to $q't$. Then we may introduce the previous notation ξ for this component, and write the model as

$$\begin{aligned} x_i - E(x) &= \xi_i\,\gamma + \delta_i\,, \qquad \delta_i \sim N(0, \Sigma_\delta), \\ y_i - E(y) &= \eta_i + \epsilon_i\,, \qquad \epsilon_i \sim N(0, \sigma_\epsilon^2), \\ \eta_i &= \beta\,\xi_i\,, \qquad \xi_i \sim N(0, 1). \end{aligned} \tag{16}$$

Here γ is a theoretical regression coefficient vector, whose $\dim(x)$ components represent the relative scales of the components of x in their degree of relation

with η. The vector δ is not the same as in representation (15), but now also incorporates the complementary, residual part of t, that is uncorrelated with ξ and η. It means that Σ_δ in (16) does not only contain (small) instrument measurement errors, but also part of the latent structure in x.

In the next section we will assume that the number of observations is larger than $\dim(x)$, so the OLS method for multiple regression can be applied. In later sections we will turn our interest to regression methods more appropriate when $\dim(x)$ is large and when there are near-collinearities in x, in particular to the PLSR and PCR methods.

4 Multivariate calibration by OLS regression

4.1 Small sample shrinkage in natural calibration with OLS multiple regression

In this section we combine model (16) with an implicit assumption that ordinary multiple regression of y on x will be a reasonable method for constructing the predictor. A population multiple regression of y on the vector x would yield the predictor

$$\widehat{y}(x) = \widehat{\eta}(x) = E(y) + b' \{x - E(x)\}, \tag{17}$$

where now

$$b' = \mathrm{cov}(y, x)\, \mathrm{var}(x)^{-1} = \beta\, \gamma' \, (\gamma\gamma' + \Sigma_\delta)^{-1}. \tag{18}$$

Here $\gamma\gamma'$ is a rank one matrix, and b corresponds to B^0 of Burnham et al (2001).

As in Section 2, the linear regression of y on $\widehat{y}$ for data from the same population is $E(y|\widehat{y}) = \widehat{y}$, not only theoretically for the population, but also as fitted to the training data used for estimation of b.

It was argued in Section 2.1 that we should demand the regression of y on $\widehat{y}$ to have a population regression coefficient $\mathrm{cov}(y, \widehat{y})/\mathrm{var}(\widehat{y})$ not systematically deviating from 1, over the randomness of the calibration. In the case of univariate x, the median for the regression coefficient was seen to be 1, so there was really no systematic error in the regression of y on $\widehat{y}$. When $\dim(x)$ is large, this is no longer true. In Appendix A1 it is demonstrated that the mean value over the calibration randomness satisfies

$$E\left\{\frac{\mathrm{cov}(y, \widehat{y})}{\mathrm{var}(\widehat{y})}\right\} \;=\; E\left\{\frac{b'\mathrm{var}(x)\widehat{b}}{\widehat{b}'\mathrm{var}(x)\widehat{b}}\right\} \approx 1 - \frac{\dim(x) - 2}{n - \dim(x) - 2}\, \frac{\sigma_\epsilon^2}{b'\mathrm{var}(x)b}$$

$$= 1 - \frac{\dim(x) - 2}{n - \dim(x) - 2} \frac{\sigma_\epsilon^2}{\beta^2} \left\{1 + (\gamma' \Sigma_\delta^{-1} \gamma)^{-1}\right\}, \qquad (19)$$

and an illustration of this function of n is seen in Figure 1. The interpretation of the term $(\gamma' \Sigma_\delta^{-1} \gamma)^{-1}$ will be discussed in the next section. The mean value of the ratio does not exist, in a strict sense, when $\dim(x) \leq 2$. In these cases we must refer to the median or to the mean value of a linearization of the ratio.

Formula (19) shows that for $\dim(x)$ large, and in particular when the sample size n is not much larger, there can be a quite substantial systematic shrinkage effect in the regression of y on $\widehat{y}$. This was noted by Copas (1983), and an analogous approximation is found in his paper, but under the somewhat inappropriate additional assumption that the calibration sample variance matrix $\mathrm{var}(x)$ is identical with the population variance. In chemometric prediction situations it is usual that $\dim(x)$ is large, even larger than n, and formula (19) contains much of the reason why OLS does not work satisfactorily in these cases and is better replaced by for example PCR or PLSR (see Section 5 below).

Conclusion: When x is multi-dimensional, and most pronounced when $\dim(x)$ is of the same magnitude as n, the OLS predictor $\widehat{y}$ has a systematic error, quantified by the approximate mean shrinkage factor (19).

4.2 Selection under multiple regression

We concentrate here on the effect of an overdispersion in the calibration η-values. For simplicity, the mean values are taken to be the same in both populations. If they were not, the bias effect of their difference would be given by a formula quite analogous to the univariate formula (11). Suppose a selection is made in the calibration such that $\mathrm{var}(\eta)$ is increased by a variance inflation factor θ^2. This is equivalent to an increase of $\mathrm{var}(\xi)$ from 1 to θ^2 in the calibration formulae. This will yield a b-vector b_c for the calibration population given by

$$b_c' = \mathrm{cov}_c(y, x)\, \mathrm{var}_c(x)^{-1} = \beta\theta^2\, \gamma' \left(\theta^2 \gamma\gamma' + \Sigma_\delta\right)^{-1}. \qquad (20)$$

instead of (18). If b_c is used for prediction in the natural population, the bias for given x will be

$$\mathrm{bias}(\widehat{y}(x)|x) = (b_c - b)'\,(x - \mu), \qquad (21)$$

in analogy with (11). The corresponding population regression of y on $\widehat{y}$ is

$$E(y|\widehat{y}) - E(y) = \frac{\mathrm{cov}(y, \widehat{y})}{\mathrm{var}(\widehat{y})} \, (\widehat{y} - E(y)) = \frac{b' \, \mathrm{var}(x) \, b_c}{b_c' \, \mathrm{var}(x) \, b_c} \, (\widehat{y} - E(y)). \qquad (22)$$

The interpretation of (22) is as the relation seen when testing a calibration on a large natural population when the calibration used a large sample from a population characterized by the variance inflation factor θ^2. Further down we will allow smaller calibration samples.

When $\theta = 1$, making $b_c = b$, relation (22) simplifies to $E(y|\widehat{y}) = \widehat{y}$). For a general θ the expression might appear complicated, but by use of the so called binomial inverse theorem (Brown, 1993, App. D), (22) can be expressed in complete analogy with the one-dimensional version (12):

$$E(y|\widehat{y}) - E(y) = \frac{1 + (\gamma' \Sigma_\delta^{-1} \gamma)^{-1} / \theta^2}{1 + (\gamma' \Sigma_\delta^{-1} \gamma)^{-1}} \, (\widehat{y} - E(y)). \qquad (23)$$

A demonstration is given in Appendix A2.

Even for moderate values of θ, =2 or =3 say, the second term of the numerator of (24) is quite small in comparison with the corresponding term of the denominator, and a useful upper bound to the selection effect is obtained by neglecting the former term completely.

If Σ_δ contains large latent variation in x, it might appear as if the term $(\gamma' \Sigma_\delta^{-1} \gamma)^{-1}$ must also be large. This is not the case, however, as understood from the following interpretations and alternative expressions.

By analogy with $\gamma' \Sigma_\delta \gamma$ being the variance of $\gamma' \, \delta$, the expression $(\gamma' \Sigma_\delta^{-1} \gamma)^{-1}$ can also be interpreted as a variance, more precisely as the conditional variance for $\gamma' \, \delta / (\gamma' \, \gamma)$, given the $(\dim(x) - 1)$-dimensional orthogonal complement to $\gamma' \, \delta$ in δ-space. This is an extension of the well-known fact that the inverted diagonal elements of a Σ^{-1}-matrix are the conditional variances, given the other components of the random vector with the covariance matrix Σ. The extension required here can be proved by making an orthogonal transformation such that one basis vector becomes proportional to γ.

Remark: This interpretation of $(\gamma' \Sigma_\delta^{-1} \gamma)^{-1}$ as a conditional variance is important for understanding the influence of different sources of variation. The marginal variance for $\gamma' \, \delta$ is an upper bound for the conditional variance. The latter will typically be strictly smaller, because the orthogonal complement need not be uncorrelated with $\gamma' \, \delta$. The relation between the marginal and conditional variances can be expressed as follows by means of

the total correlation coefficient ρ_{tot}, more precisely here the maximum correlation between $\gamma'\delta$ and all linear combinations of variables in the orthogonal complement to $\gamma'\delta$. Also $\gamma'\gamma = \mathrm{var}(\gamma'\xi)$ is the variance of the part of x that is correlated with y (or η). This yields the alternative expression

$$(\gamma'\Sigma_\delta^{-1}\gamma)^{-1} = \frac{\mathrm{var}(\gamma'\delta/|\gamma|)}{\mathrm{var}(\gamma'\xi)}(1-\rho_{tot}^2). \tag{24}$$

Here $\gamma'\delta/|\gamma|$ is the projection of δ on the direction of the vector γ.

We now consider how the bias factor in 23 must be modified when the training sample size is small or only moderately large, in combination with overdispersion in the calibration. In analogy with formula (19) for the expected coefficient of the regression of y on $\hat{y}(x)$, we need an approximation for

$$E\left\{\frac{\mathrm{cov}(y,\widehat{y})}{\mathrm{var}(\widehat{y})}\right\} = E\left\{\frac{b'\mathrm{var}(x)\widehat{b}_c}{\widehat{b}_c'\mathrm{var}(x)\widehat{b}_c}\right\}, \tag{25}$$

where $\widehat{b}_c$ notifies a difference from $\widehat{b}$ due to the selection. The same propagation of errors type approximations as for (19) now yield a main factor identical with the large sample factor

$$\frac{1+(\gamma'\Sigma_\delta^{-1}\gamma)^{-1}/\theta^2}{1+(\gamma'\Sigma_\delta^{-1}\gamma)^{-1}}.$$

The rest of the approximation becomes more involved, and in order to get a formula simple enough for presentation we restrict to the limiting case $\theta^2 \to \infty$. Then the following approximation formula is obtained, see Appendix A3 for a derivation and Figure 2 for an illustration.

Conclusion: In multivariate calibration by OLS multiple regression under a latent factor regression model, overdispersion in η by a variance inflation factor θ^2 will yield a systematic error in the regression of y on the predictor $\widehat{y}(x)$, for large calibration samples quantified by the shrinkage factor

$$\frac{1+(\gamma'\Sigma_\delta^{-1}\gamma)^{-1}/\theta^2}{1+(\gamma'\Sigma_\delta^{-1}\gamma)^{-1}} \tag{26}$$

in the regression of y on $\widehat{y}$, and for smaller sample sizes approximated by at worst (i.e. for $\theta^2 \to \infty$)

$$\frac{1}{1+(\gamma'\Sigma_\delta^{-1}\gamma)^{-1}}\left[1-\frac{\dim(x)-1}{n-\dim(x)-2}\,\frac{\sigma_\epsilon^2}{\beta^2}\left\{1+(\gamma'\Sigma_\delta^{-1}\gamma)^{-1}\right\}\right]. \tag{27}$$

4.3 Could the bias be adjusted for?

For the pure small-sample bias, discussed in Section 4.1 above, Copas (1983) proposed that the shrinkage be estimated and adjusted for. Since we can estimate $\sigma_\epsilon^2/(b'\mathrm{var}(x)b)$ from the calibration, we can also estimate the systematic shrinkage factor. The simple straight-on estimate of $\sigma_\epsilon^2/(b'\mathrm{var}(x)b)$ can be expressed in terms of the coefficient of determination R^2 as

$$\frac{n-1}{n-\dim(x)-1}\,\frac{1-R^2}{R^2}.$$

Cross-validation, in which one or several observations at a time are left out of the model-fitting, can also indicate small-sample bias, see Sundberg (2006) for illustrations, and in principle be used to adjust for such bias. However, there is considerable randomness involved in the particular case, and an imprecisely estimated shrinkage adjustment need not be better than no adjustment at all. We could also cite Faber (2000, p. 368), who comes to the same conclusion about possible correction for the estimation bias of the regression coefficients of PLSR.

For the bias due to selection, the situation is different. Cross-validation can of course not detect the bias (cf. Sundberg, 2006), since it uses only the training data. In order to use formula (23) above in an adjustment, we would have to rely on the underlying assumptions, for example that selection was made in η itself, and not in some other latent factor, only correlated with η. We must also have relatively accurate values of the parameters involved. As seen from formulae (11) and (12), we need to know or estimate all parameters, μ, μ_c, σ^2, σ_c^2, and σ_δ^2, and these are implicit characteristics defined within a latent structure. Therefore it is doubtful if we will ever have good enough information about them to be sure that the bias adjustment improves the situation.

Another possibility that might come to mind is to use the empirical regression of y on $\widehat{y}$ for a validation set from the natural population, and to adjust the coefficients of the predictor $\widehat{y}$ such that the fitted regression becomes $y = \widehat{y}$. This would be equivalent to using only the direction of the b_c-vector from the calibration sample and the length of b from the validation sample. That cannot be an efficient way of using the information in data. When $\dim(x) = 1$, it would imply throwing away the calibration sample completely.

5 Calibration by latent factor regression methods, such as PLSR and PCR

By latent factor regression methods, we refer to methods such as PLSR (here preferably regarded as an acronym for Projection to Latent Structures Regression) and PCR (Principal Components Regression), in which a typically high-dimensional x, suffering from collinearity or near-collinearity, is replaced by a low-dimensional $\widetilde{t} = \widetilde{t}(x)$, before OLS regression is applied. Thus the predictor construction follows a two-stage procedure;

1. Estimation of the form and dimension of the latent vector t as a linear function $\widetilde{t} = \widetilde{t}(x)$ of x, or rather of the latent subspace spanned by t.

2. When the function $\widetilde{t}(x)$ has been determined, OLS multiple regression of y on $\widetilde{t}$ yields the predictor.

The statistic $\widetilde{t}$ is a method-specific estimate of the latent vector t in the underlying latent model (15). As an example, in PLSR we successively select a suitable number of mutually orthogonal normalized linear functions of x having maximum covariance with y, whereas in PCR the criterion is maximum variance instead of covariance. For the present treatment, we need not specify much more, but detailed descriptions and discussions of PLSR and PCR are found for example in the books by Martens & Næs (1989) and Brown (1993). Among other methods satisfying the procedure are various more parsimonious variations of PLSR and PCR known under the name of orthogonal signal correction, see Svensson et al. (2002) for a review and comparisons.

The result of Stage 1 is a weights matrix $\widetilde{W}$ saying how the vector $\widetilde{t}$ should be calculated from the vector x, namely linearly as $\widetilde{t} = \widetilde{W}x$. On the other hand, $\widetilde{W}$ may depend non-linearly on calibration data. The ˜ notation is used to indicate that $\widetilde{W}$ is estimated, i.e. depends on data, and does not exactly match the model characteristic P. In the following we will argue as if $\widetilde{W}$ were predetermined and did not depend on the same data as used in the regression stage (stage 2). This is a deliberate simplification, which is not believed to have important effects on the bias question. In particular it should be so if the estimation of the latent vector space is successful, so that essentially all covariance between y and x is captured by the estimated $\widetilde{t}$. Alternatively this is expressed as $\mathrm{cov}(y, x|\widetilde{t})$ being negligible or that the

vector γ is (almost) in the column space of $\widetilde{W}'$. This argument holds equally if the calibration sampling has been selective with respect to the component $\eta = q'\,t$.

More precisely, reduction of dimension by going from centred x-data to $\widetilde{t} = \widetilde{W}\,(x - E(x))$ in model (15) yields the following lower-dimensional model:

$$\begin{aligned} \widetilde{t}_i &= \widetilde{W}\,(x_i - E(x)) = (\widetilde{W}\,P')\,t_i + \widetilde{W}\,\delta_i\,, \qquad \delta_i \sim N(0, \Sigma_\delta), \\ y_i - E(y) &= q'\,t_i + \epsilon_i\,, \qquad \epsilon_i \sim N(0, \sigma_\epsilon^2), \\ t_i &\sim N(0, I), \qquad \dim(t) = a < \dim(x). \end{aligned} \tag{28}$$

We again reexpress this by splitting the vector $(\widetilde{W}\,P')\,t$ in one component which is a function of $\eta = q'\,t$, and a remainder term that is uncorrelated with η and is brought into the δ-term:

$$\begin{aligned} \widetilde{t}_i &= \xi_i\,\widetilde{\gamma} + \widetilde{\delta}_i\,, \qquad \widetilde{\delta}_i = \widetilde{W}\delta_i \sim N(0, \Sigma_{\widetilde{\delta}}), \\ y_i - E(y) &= \eta_i + \epsilon_i\,, \qquad \epsilon_i \sim N(0, \sigma_\epsilon^2), \\ \eta_i &= \beta\,\xi_i\,, \qquad \xi_i \sim N(0, 1). \end{aligned} \tag{29}$$

Here $\widetilde{\gamma}$ and $\widetilde{\delta}_i$ differ from the previous γ and δ_i by referring to the model with $\widetilde{t}$ instead of x. The condition for this reexpression to be possible is that the vector $\widetilde{\gamma}$ belongs to the span of $\widetilde{W}'$, so that no information about ξ is lost in the reduction to $\widetilde{t}$. This also implies that

$$\widetilde{\gamma} = \widetilde{W}\gamma. \tag{30}$$

Model (29) is seen to be of the same type as the latent variable regression model (16), except that $\widetilde{t}$ is now replacing the original explanatory vector $x - E(x)$, and that $\widetilde{\gamma}$ and $\widetilde{\delta}$ replace the previous γ and δ.

We can now see the consequences for PLSR and PCR in the small sample and selection bias effects, discussed in Sections 4.1 and 4.2 for the OLS predictor. In the bias factor formulae (19), (23) and (27) the only changes are that $\dim(x)$ is replaced by $\dim(\widetilde{t})$ and that $\gamma'\Sigma_\delta^{-1}\gamma$ is replaced by

$$\widetilde{\gamma}'\Sigma_{\widetilde{\delta}}^{-1}\widetilde{\gamma} = \gamma'\widetilde{W}'(\widetilde{W}\Sigma_\delta\widetilde{W}')^{-1}\widetilde{W}\gamma \tag{31}$$

As a first consequence, the small sample shrinkage of Section 4.1 will tend to be much smaller for the PLSR and PCR predictors than for the OLS predictor, simply because we have replaced the large $\dim(x)$ by the relatively small $\dim(\widetilde{t})$.

It remains to find the possible difference between $\widetilde{\gamma}'\Sigma_{\widetilde{\delta}}^{-1}\widetilde{\gamma}$ and $\gamma'\Sigma_{\delta}^{-1}\gamma$. We have already made the assumption that γ belongs to the subspace spanned by the rows of $\widetilde{W}$. Let $\widetilde{W}' \propto \begin{pmatrix} \gamma & \widetilde{W}'_{res} \end{pmatrix}$, where $\widetilde{W}'_{res}$ consists of column vectors spanning this space jointly with γ, and orthogonal to γ. Note that (31) is invariant to any change of $\widetilde{W}$ by a scalar factor. Inserting the expression for $\widetilde{W}'$ in formula (31) we obtain

$$\widetilde{\gamma}'\Sigma_{\widetilde{\delta}}^{-1}\widetilde{\gamma} = \begin{pmatrix} \gamma'\gamma & 0 \end{pmatrix} \begin{pmatrix} \gamma'\Sigma_\delta\gamma & \gamma'\Sigma_\delta\widetilde{W}'_{res} \\ \widetilde{W}\Sigma_\delta\gamma & \widetilde{W}\Sigma_\delta\widetilde{W}'_{res} \end{pmatrix}^{-1} \begin{pmatrix} \gamma'\gamma \\ 0 \end{pmatrix} = \frac{\gamma'\gamma}{\mathrm{var}(\gamma'\delta/\gamma'\gamma|\widetilde{W}_{res}\delta)}. \tag{32}$$

This is the same interpretation as for OLSR, see Section 4.2, except that we are now conditioning on only the part $\widetilde{W}_{res}\delta$ of the residual part δ_{res} of δ (the orthogonal complement to $\gamma'\delta$). However, both PLSR and PCR are likely to include most orthogonal variation having appreciable covariance with $\gamma'\delta$ (which is the same as the covariance with $\gamma'x$), so we have reason to expect

$$\mathrm{var}(\gamma'\delta|\widetilde{W}_{res}\delta) \approx \mathrm{var}(\gamma'\delta|\delta_{res}). \tag{33}$$

When (33) is satisfied, (32) shows that the only essential difference from OLSR is that $\dim(t)$ replaces $\dim(x)$.

For PLSR and similar methods it may happen that quite much orthogonal variation is left out of the latent structure, but that will be variation that has little covariance with $\gamma'x$, and therefore does not much influence the conditional variance. An alternative way of seeing that we should expect (33) to be satisfied is to consider the expression of the conditional variance in terms of the the total correlation coefficient ρ_{tot}, see equation (24). The population total correlation coefficient can be expected to be quite similar for OLSR, PLSR, and PCR if we include sufficiently many factors in the latter methods. (On the other hand, the corresponding sample correlation R may differ considerably if n is not large enough.)

Theoretical summary: For latent factor regression methods such as PCR or PLSR the same bias factor formulae (19), (23) and (27) as for OLSR will hold approximately, except that $\dim(t)$ replaces $\dim(x)$, reducing the small-sample bias (considerably).

6 A simulation study

For illustration and confirmation purposes a latent factor model for a data set of 344 slaughter pigs was used. The parameter values of this model

were used in the formulas, and simulations from the model were used to generate corresponding empirical quantities. The data set is briefly described in Section 6.1. These data were used also in Sundberg (2006), but differently. In that paper variability was generated by resampling from the actual data set, but here the data were utilized only to generate a model that may be realistic in applications and is adequate for at least one set of real data.

6.1 A background model for pig grading data

The simulation model, of type (15), was based on a data set from the Danish Meat Research Institute. In a study of a method called KC for grading of slaughter pig carcasses, they used a sample of 344 pigs. The quantity η of interest, to be predicted, was the lean meat percentage. A reference value y was obtained by dissection. A linear regression model for prediction was fitted, based on 11 x-variables (slaughter weight and some specific measurements of fat thickness and muscles tissue thickness).

The data were found to be adequately described by a latent factor regression model with three latent variables. Gaussian distributions of the random components seemed reasonable. Variance standardization of x was not applied, but variances were already of the same magnitude. Both PLSR and PCR naturally stopped at this dimension. With the original data, OLSR with full x explained $R^2 = 77.7\%$ of the variation in y. PLSR explained 91% of the variation in x and 77.2% of the variation in y, i.e. little less than OLSR. The first three eigenvalues of the covariance matrix of x were 145, 61 and 26, and the remaining eigenvalues were between 6 and 1. In the simulation model Σ_δ was taken to be diagonal with variances 4, and the three large eigenvalues were correspondingly reduced. The data set is available at www.math.su.se/∼rolfs/Publications.html.

6.2 Simulation results

The simulations were carried out by generating a calibration sample (training set) ranging in size from $n = 14$ to $n = 100$, together with a validation sample (test set) of size 1000. The calibration sample was either a pure random sample from the model (the natural population), or else it was selected to have an increased variance in η. The θ-value actually used was extremely large, $\theta = 100$, in order to mimic the worst case ($\theta = \infty$), but it should be noted that already for moderate θ, $\theta = 3$ say, the deviation from the

worst case is small. Each such simulation of a calibration set and a test set was repeated 1000 times. In each simulation the predictor was determined from calibration data using OLSR, PCR or PLSR, the two latter methods with 3 factors, see the previous section. The predictor was tried on the validation data and for this large set the regression coefficient of y on $\hat{y}$ was determined with high precision. These coefficient values were averaged over all simulations and in this way gave a point in one of the diagrams. Figure 1 represents natural calibration samples whereas Figure 2 represents samples selected to have a very large variation in the ξ and η directions. These points, for varying calibration sample sizes, can be compared with the theoretical curves according to formulas (19) and (27) for OLSR and the corresponding formulas for PCR and PLSR.

The diagrams show first that the simulated points (with different markers) agree reasonably well with the corresponding curves, being large sample approximations of the true (expected) bias factors. Only for very small calibration sample sizes, say $n < 20$, is the lack of fit substantial. That the deviations from the curve are smaller in Fig. 2 than in Fig. 1 can mostly be explained by the fact that the precision is higher when the calibration sample is overdispersed. That PCR points in Figure 1 are clearly above the curve is mostly an effect of a quite skew distribution of the regression coefficients with PCR. If the median were plotted instead of the mean, most of this effect would disappear. See next section for more discussion about the deviations from the curve.

7 Discussion

The present paper has derived and presented formulae for the bias of predictions under a latent factor model when the predictor has been constructed by OLS regression, or PLSR, or PCR. A lot of debate can be found in the literature, largely closed by Faber (1999), concerning the biased character of PLSR and PCR regression coefficient estimation methods, as being shrinkage regression methods, in contrast to the unbiased estimation by OLSR, when the latter can be used. Most of this discussion is regarded as irrelevant here, when we consider the prediction error as function of the predictor variable x and quantify the bias likewise, and when the OLSR-based predictor is in fact found to be biased, and more seriously so than the other methods, for small traning samples. The conventional bias of PLSR and PCR is due to

the data-dependent choice of a limited number of factors to be used as regressors, but this data-dependence has been neglected here. In this respect we take as an assumption the statement by Faber (2000) about PLSR, that "in typical chemometrics applications bias is likely to be small", at the same time as we stress the restricted applicability of this statement about bias, as demonstrated by the present study.

In order to illustrate the influence of different sources of error, Martens & Næs (1989, Sec. 4.1.3) presented some simulation results for calibration by PLSR under various more or less artificial special cases of model (15), and they show plots of $\widehat{y}(x)$ against y (the converse to our diagrams), for the test set. Two of their settings yield regressions of y on $\widehat{y}(x)$ clearly differing from the ideal identity function. Their Figure 4.2 (d) shows a regression coefficient < 1, which is as expected from the theory for the small-sample bias effect, because they have a positive σ_ϵ^2 in the generation of calibration data. On the other hand, their Figure 4.2 (e) shows a coefficient > 1, as also remarked by the authors as being a typical 'least-squares effect'. In this case Martens & Næs assume $\sigma_\epsilon^2 = 0$, and $\sigma_\delta^2 = 0$ for the test set, but $\sigma_\delta^2 > 0$ for the generation of the calibration data. The formulae of the theory presented here might at first appear unable to give a coefficient > 1, but on close inspection of formula (44) of Appendix 2 below, the result follows (since $\sigma_\epsilon^2 = 0$, we can go directly to Appendix 2 and take $\theta = 1$). With $\sigma_\delta^2 = 0$ for the test set, $\mathrm{var}(x) = \gamma\gamma'$, whereas $\mathrm{var}_c(x) = \gamma\gamma' + \Sigma_\delta$. Insertion yields an expected regression coefficient of $1 + (\gamma'\Sigma_\delta^{-1}\gamma)^{-1}$, which is likely to explain the bias seen in their Figure 4.2 (e).

Geladi et al (1999) discuss bias in PLSR prediction in terms of what they call local bias, which would include the case of a bias factor. However, they are primarily interested in such bias as a diagnostic for model errors due to nonlinearities.

The influence of predictor bias on RMSEP or MSEP and similar measures of prediction errors need not be so large as we might tend to fear. A bias factor of 0.95 or even 0.90 typically does not change much, because MSEP is an average over (the whole) natural population of items, and most of them are in the centre of the population where the bias factor has little effect. More precisely, multiplying the ideal predictor $E(y|x)$ by a bias factor $k \geq 1$ increases MSEP by a factor

$$1 + (k-1)^2 \, \frac{R^2}{1-R^2}, \tag{34}$$

where the constant R^2 is the squared population correlation coefficient between y and the regression $E(y|x)$, so MSEP increases quadratically with $k-1$ (and likewise RMSEP for k near 1). However, instead of primarily considering the effect on the average measure MSEP, we should rather regard the bias formulae as saying that for not so typical items, i.e. items with y and x relatively far from their population mean values, the predictor under consideration will exaggerate their deviation from the mean.

A simple illustration of formula (34) can be found in the univariate context of Andersen et al (2003). In their simulations they observe a rapid nonlinear increase in RMSEP when their characteristic corresponding to θ becomes large (their inverse number of replicates, see Section 2.2.2) . Formulae (34) and (14) in combination give at least a qualitative explanation and understanding of the phenomenon they observed.

Even if we are given a random sample of training data from the natural population, this does of course not by itself guarantee that this sample is not variance-inflated by pure chance. If such variance inflation is suspected, there may also be reason to suspect a bias factor in the predictor, as if the variance inflation had been intended.

Finally it is worth stressing that cross-validation can help detecting the small sample bias, but is unable to detect bias generated by variance inflation in the sampling, since there is no independent test set from the population to compare with.

In the univariate case $\dim(x) = 1$ it was shown that the OLS predictor $\hat{y}(x)$ was median-unbiased for given x, whereas for larger $\dim(x)$ expressions were given for the error in mean value. It might be asked if this error possibly disappears if we go over to median instead. The answer is no. For OLSR and PLSR the distribution is relatively symmetric, so mean and median differ little. For PCR, on the other hand, the distribution can be quite skew and this fully explains the difference between empirical points for PCR (crosses) and the corresponding theoretical curve in Figure 1. Almost all difference disappears if empirical means are replaced by medians. This leaves unexplained only the difference for PLSR, which goes in the opposite direction. One explanation, cf. the last paragraph of Sec. 5, could be that with the randomness of small samples, PLSR leaves too much variation in x unexplained, and in particular more than PCR.

Appendix. The expected value of $\mathbf{cov}(y, \hat{y})/\mathbf{var}(\hat{y})$.

A1. Natural calibration, small sample shrinkage

It was found in Section 4.1 that the possible systematic error in the regression of future y on $\hat{y}(x)$ is given by the expected value over the random calibration sample of the theoretical regression coefficient $\mathrm{cov}(y, \widehat{y}(x))/\mathrm{var}(\widehat{y}(x))$.

For the investigation in Section 4.1 of the bias effect for small calibration samples on future predictions $\widehat{y}$ of y, we required the expected theoretical regression coefficient

$$E\left\{\frac{\mathrm{cov}(y,\widehat{y})}{\mathrm{var}(\widehat{y})}\right\} = E\left\{\frac{\mathrm{cov}(y,x)\widehat{b}}{\widehat{b}'\mathrm{var}(x)\widehat{b}}\right\} = E\left\{\frac{b'\mathrm{var}(x)\widehat{b}}{\widehat{b}'\mathrm{var}(x)\widehat{b}}\right\}. \tag{35}$$

The calibration sample, over which the expected value is calculated, is here assumed to be a sample from the natural population, so $E(\widehat{b}) = b$. When $\dim(x) = 1$, the ratio simplies to $b/\widehat{b}$.

The expected value (35) exists in a strict sense for $\dim(x) > 2$, but for any value of $\dim(x)$, we may Taylor expand the ratio in (35) to second order in $\widehat{b}$ around b. The constant term is 1, the first order term has expected value zero, and the second order contribution simplifies to

$$2\,\frac{\{b'\mathrm{var}(x)(\widehat{b}-b)\}^2}{\{b'\mathrm{var}(x)b\}^2} - \frac{(\widehat{b}-b)'\mathrm{var}(x)(\widehat{b}-b)}{b'\mathrm{var}(x)b}. \tag{36}$$

The numerators of these second order terms have expected values

$$E\{(b'\mathrm{var}(x)(\widehat{b}-b))^2\} = b'\mathrm{var}(x)\mathrm{var}(\widehat{b})\mathrm{var}(x)b \tag{37}$$

and

$$\begin{aligned} & E\{(\widehat{b}-b)'\mathrm{var}(x)(\widehat{b}-b)\} = E[\mathrm{tr}\{(\widehat{b}-b)'\mathrm{var}(x)(\widehat{b}-b)\}] \\ = \quad & E[\mathrm{tr}\{\mathrm{var}(x)(\widehat{b}-b)(\widehat{b}-b)'\}] = \mathrm{tr}\{\mathrm{var}(x)\mathrm{var}(\widehat{b})\} \end{aligned} \tag{38}$$

respectively, where tr denotes the trace. For both (37) and (38) we need an expression for $\mathrm{var}(\widehat{b})$. Since $\widehat{b}$ is conditionally unbiased, its variance can be calculated via conditioning on the calibration x-data x_c, as the calibration sample sum of centered squares and products matrix S_{xx} matrix of x, which is proportional to the inverse of the conditional variance of $\widehat{b}$:

$$\mathrm{var}(\widehat{b}) = E\{\mathrm{var}(\widehat{b}|x_c)\} = \sigma_\epsilon^2\, E(S_{xx}^{-1}) = \sigma_\epsilon^2\,\frac{\mathrm{var}(x)^{-1}}{n-\dim(x)-2}. \tag{39}$$

Here S_{xx} is the calibration sample sum of centered squares and products matrix, which is Wishart distributed and inversely proportional to the conditional variance of $\widehat{b}$. For the expected value of the inverse Wishart distributed matrix S_{xx}^{-1}, see for example Brown (1993, Appendix A). Insertion of (39) in (37) and (38) yields

$$\frac{\sigma_\epsilon^2\, b'\mathrm{var}(x)b}{n-\dim(x)-2}$$

for the first term numerator, and

$$\frac{\sigma_\epsilon^2}{n-\dim(x)-2}\,\mathrm{tr}(I_{\dim(x)}) = \frac{\dim(x)\,\sigma_\epsilon^2}{n-\dim(x)-2} \tag{40}$$

for the second numerator (assuming $n-\dim(x)-2>0$, of course). Together they yield the simple approximation result

$$\begin{aligned} E\left\{\frac{\mathrm{cov}(y,\widehat{y})}{\mathrm{var}(\widehat{y})}\right\} &\approx 1-\frac{\dim(x)-2}{n-\dim(x)-2}\,\frac{\sigma_\epsilon^2}{b'\mathrm{var}(x)b} \\ &= 1-\frac{\dim(x)-2}{n-\dim(x)-2}\,\frac{\sigma_\epsilon^2}{\beta^2\gamma'(\gamma\gamma'+\Sigma_\delta)^{-1}\gamma} \\ &= 1-\frac{\dim(x)-2}{n-\dim(x)-2}\,\frac{\sigma_\epsilon^2}{\beta^2}\left\{1+(\gamma'\Sigma_\delta^{-1}\gamma)^{-1}\right\}. \end{aligned} \tag{41}$$

The last identity in (41) is derived by using the so called binomial inverse theorem (Brown, 1993, Appendix D) on $\mathrm{var}(x)^{-1}=(\gamma\gamma'+\Sigma_\delta)^{-1}$, which yields

$$\begin{aligned} \gamma'(\gamma\gamma'+\Sigma_\delta)^{-1}\gamma &= \gamma'\left\{\Sigma_\delta^{-1}-\frac{\Sigma_\delta^{-1}\gamma\gamma'\Sigma_\delta^{-1}}{1+\gamma'\Sigma_\delta^{-1}\gamma}\right\}\gamma \\ &= \frac{\gamma'\Sigma_\delta^{-1}\gamma}{1+\gamma'\Sigma_\delta^{-1}\gamma} = \left\{1+(\gamma'\Sigma_\delta^{-1}\gamma)^{-1}\right\}^{-1}. \end{aligned} \tag{42}$$

A2. Selective calibration

In Section 4.2 it was stated that for an infinitely large calibration sample, overdispersed by a variance inflation factor θ^2 in ξ or η, the regression coefficient of y on $\widehat{y}(x)$ to be seen for the natural population will suffer from shrinkage by the factor $\mathrm{cov}(y,\widehat{y})/\mathrm{var}(\widehat{y})<1$. We will here prove that this shrinkage factor can be expressed as

$$\frac{1+(\gamma'\Sigma_\delta^{-1}\gamma)^{-1}/\theta^2}{1+(\gamma'\Sigma_\delta^{-1}\gamma)^{-1}}. \tag{43}$$

First we note that

$$\frac{\mathrm{cov}(y,\widehat{y})}{\mathrm{var}(\widehat{y})} = \frac{1}{\beta}\frac{\mathrm{cov}(y,x)b_c}{b_c'\mathrm{var}(x)b_c} = \frac{\gamma' b_c}{\gamma'\mathrm{var}_c(x)^{-1}\mathrm{var}(x)b_c}\,\frac{1}{\theta^2}, \tag{44}$$

where $b_c' = \beta\theta^2\gamma'\mathrm{var}_c(x)^{-1}$. We will show that $\gamma'\mathrm{var}_c(x)^{-1}\mathrm{var}(x)$ in the denominator is proportional to γ', and then the numerator will cancel an identical factor of the denominator. Again we use the binomial inverse theorem, now on $\mathrm{var}_c(x)^{-1}$:

$$\mathrm{var}_c(x)^{-1} = (\theta^2\gamma\gamma' + \Sigma_\delta)^{-1} = \left(\Sigma_\delta^{-1} - \frac{\theta^2\Sigma_\delta^{-1}\gamma\gamma'\Sigma_\delta^{-1}}{1+\theta^2\gamma'\Sigma_\delta^{-1}\gamma}\right). \tag{45}$$

Multiplication of (45) from the left by γ' and from the right by $\mathrm{var}(x) = \gamma\gamma' + \Sigma_\delta$ yields $\gamma'(1+\gamma'\Sigma_\delta^{-1}\gamma)$. Combination of these formulae yields the desired result.

A3. Combined effects

We here indicate how the approximate formula (27) for the small-sample overdispersion shrinkage factor in Section 4.2 can be derived. If the derivation in Appendix A1 is reconsidered for an overdispersed calibration sample, we first replace $\widehat{b}$ by $\widehat{b}_c$ in (35) (three times). The constant term is no longer 1, but is replaced by the factor (43) of Appendix A2, valid for very large sample sizes. The second order terms are more complicated, and for simplicity we restrict here to the limiting case $\theta \to \infty$, in which we can apply the following lemma:

Lemma: In the limit as $\theta \to \infty$,

$$\mathrm{var}_c(x)^{-1} = \Sigma_\delta^{-1} - \frac{\Sigma_\delta^{-1}\gamma\gamma'\Sigma_\delta^{-1}}{\gamma'\Sigma_\delta^{-1}\gamma}, \tag{46}$$

$$b_c = \Sigma_\delta^{-1}\gamma\frac{\beta}{\gamma'\Sigma_\delta^{-1}\gamma}, \tag{47}$$

and the matrix $\mathrm{var}_c(x)^{-1}\mathrm{var}(x)$ is a projection matrix with trace $\dim(x)-1$, that projects b_c on zero.

Proof of Lemma: The limiting form of $\mathrm{var}_c(x)^{-1}$ is obvious from (45). The limiting form of $b_c = \beta\theta^2\mathrm{var}_c(x)^{-1}\gamma$ requires a little more care, but is found by again using the form (45). Finally, the properties of $\mathrm{var}_c(x)^{-1}\mathrm{var}(x)$ are easily checked by using the limiting form of $\mathrm{var}_c(x)^{-1}$, but they are also natural

from the construction of the calibration sample, being imagined as stretched out indefinitely in one direction. Inserting these formulas for $\mathrm{var}_c(x)^{-1}$, in particular where it replaces $\mathrm{var}(x)^{-1}$ in (39), and for b_c where it replaces b in (35) and subsequent formulas, the desired result (27) appears. For example, the Lemma makes (37) vanish and makes $\dim(x)$ in the numerator of (40) be replaced by $\dim(x) - 1$.

Acknowledgement

I am grateful to Eli Vibeke Olsen of the Danish Meat Research Institute for providing access to the pigs grading data.

References

Andersen, C.M., Bro, R. and Brockhoff, P.B. (2003). Quantifying and handling errors in instrumental measurements using the measurement error theory. *Journal of Chemometrics*, **17**, 621–629.

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

Burnham, A.J., MacGregor, J.F. and Viveros, R. (1999). Latent variable multivariate regression modelling. *Chemometrics and Intelligent Laboratory Systems*, **48**, 167–180.

Burnham, A.J., MacGregor, J.F. and Viveros, R. (2001). Interpretation of regression coefficients under a latent variable regression model. *Journal of Chemometrics*, **15**, 265–284.

Copas, J.B. (1983). Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society, Series B*, **45**, 311–354.

Copas, J.B. (1987). Cross-validation shrinkage of regression predictors. *Journal of the Royal Statistical Society, Series B*, **49**, 175–183.

Faber, N.M. (1999). A closer look at the bias–variance trade-off in multivariate calibration. *Journal of Chemometrics*, **13**, 185–192.

Faber, N.M. (2000). Response to 'Comments on construction of confidence intervals in connection with partial least squares'. *Journal of Chemometrics*, **14**, 363–369.

Ganse, R.A., Amemiya, Y. and Fuller, W.A. (1983). Prediction when both variables are subject to error, with application to earthquake magnitudes. *Journal of the American Statistical Association*, **78**, 761–765.

Geladi, P., Hadjiiski, L. and Hopke, P. (1999). Multiple regression for environmental data: nonlinearities and prediction bias. *Chemometrics and Intelligent Laboratory Systems*, **47**, 165–173.

Gleser, L.J. (1991). Measurement error models. *Chemometrics and Intelligent Laboratory Systems*, **10**, 45–57.

Jones, M.C. and Copas, J.B. (1986). On the robustness of shrinkage predictors in regression to differences between past and future data. *Journal of the Royal Statistical Society, Series B*, **48**, 223–237.

Lindley, D.V. (1947). Regression lines and the linear functional relationship. *Journal of the Royal Statistical Society, Supplement*, **9**, 218–244.

Martens, H. and Næs, T. (1989). *Multivariate Calibration*. Wiley, Chichester.

Sundberg, R. (2006). Small-sample and selection bias effects in multivariate calibration, exemplified for OLS and PLS regressions. *Chemometrics and Intelligent Laboratory Systems*, **84**, 21–25.

Svensson. O., Kourti, T. and MacGregor, J.F. (2002). An investigation of orhogonal signal correction algorithms and their characteristics. *Journal of Chemometrics*, **16**, 176–188.

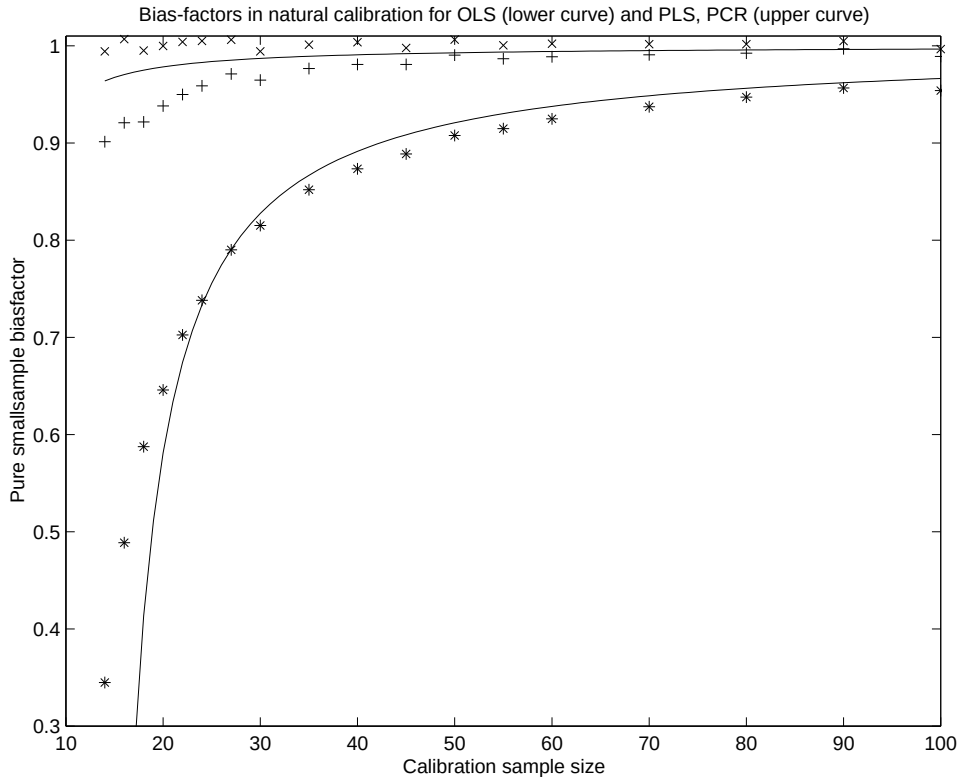

Figure 1: Small-sample bias-factor illustrated for natural calibration. Theoretical and simulated data from latent factor regression model, see Sec. 6.1.
Lower curve: OLS, formula (19)
Upper curve: PLSR/PCR with 3 latent factors, see Sec. 5.
$*$ Simulated data with OLSR for predictor construction
$+$ Simulated data with PLSR for predictor construction
$\times$ Simulated data with PCR for predictor construction
Points represent averaging over a test set of size 1000 and averaging over 1000 simulated datasets for each sample size n.

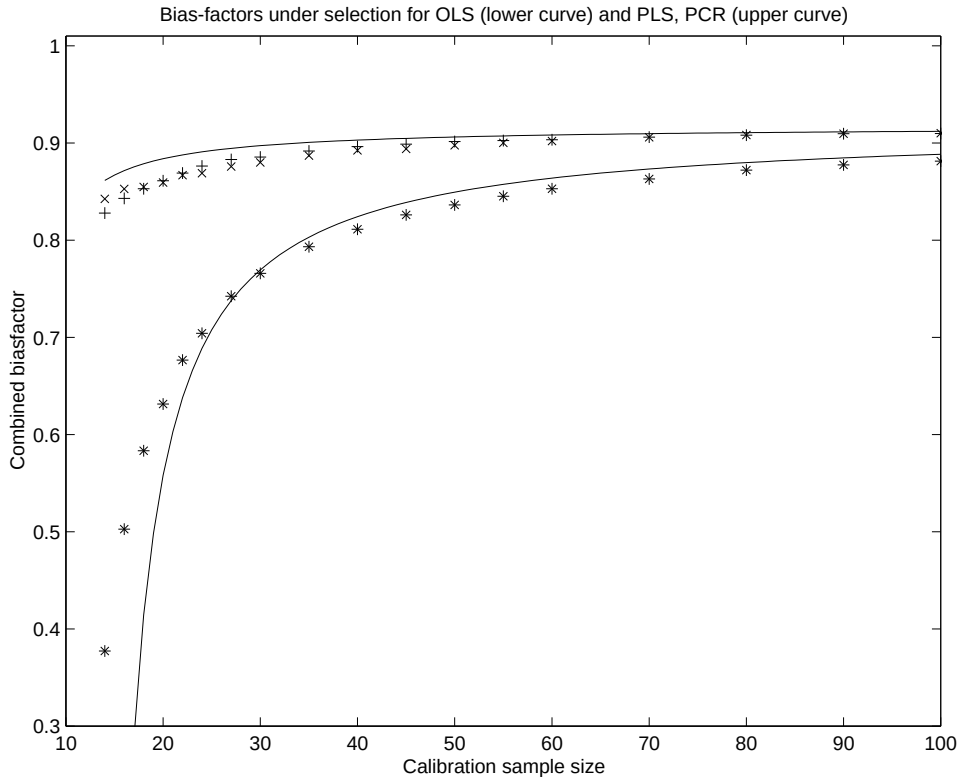

Figure 2: Combined selection and small-sample bias-factor illustrated
Theoretical and simulated data from latent factor regression model, see Sec. 6.1.
Lower curve: OLS, formula (27) Upper curve: PLSR/PCR with 3 latent factors, see Sec. 5.
$*$ Simulated data with OLSR for predictor construction
$+$ Simulated data with PLSR for predictor construction
$\times$ Simulated data with PCR for predictor construction
Points represent averaging over a test set of size 1000 and averaging over 1000 simulated datasets for each sample size n.