Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR)*

Bjørn-Helge Mevik[†‡] Henrik René Cederkvist[§]

April 8, 2005

Abstract

The paper presents results from simulations based on real data, comparing several competing mean squared error of prediction (MSEP) estimators on principal components regression (PCR) and partial least squares regression (PLSR): leave-one-out cross-validation, K-fold and adjusted K-fold cross-validation, the ordinary bootstrap estimate, the bootstrap smoothed cross-validation (BCV) estimate and the 0.632 bootstrap estimate.

The overall performance of the estimators is compared in terms of their bias, variance and squared error. The results indicate that the 0.632 estimate and leave-one-out cross-validation are preferable when one can afford the computation. Otherwise adjusted 5- or 10-fold cross-validation are good candidates because of their computational efficiency.

Keywords

Mean squared error of prediction (MSEP); cross-validation; adjusted cross-validation; bootstrap; 0.632 estimate; principal component regression (PCR); partial least squares regression (PLSR)

1 Introduction

The mean squared error of prediction (MSEP), or its square root, is frequently used to assess the performance of regressions. It is also used for choosing the optimal number of components in principal components regression (PCR) [1] and partial least squares regression (PLSR) [1].

The MSEP of a regression can be estimated by applying the regression to an independent test set. Often, a (large enough) test set is not available. In such situations, the MSEP has to be estimated from the learning data, i.e., the data used to train the regression. The leave-one-out cross-validation [2, called the 'U method'] is perhaps the most widely used internal estimator. It is nearly unbiased, and is easy to implement and understand.

*This is a preprint of an article published in *Journal of Chemometrics* 2004; **18**(9): 422–429. URL: `http://www.interscience.wiley.com/`

[†]Corresponding author. E-mail: bjorn-helge.mevik@matforsk.no

[‡]Matforsk – Norwegian Food Research Institute, Ås, Norway

[§]Agricultural University of Norway, Ås, Norway

The leave-one-out cross-validation can be used to estimate other performance measures, such as misclassification rate for classifiers. It has, however, been criticised for being variable [3–6], and alternative estimators, often based on K-fold cross-validation [7] or the bootstrap [3], have been proposed to reduce this variability.

Several comparisons of cross-validation- and bootstrap-based estimators have published, see for instance [6, 8–15]. However, most theoretical results regarding the properties of these estimators have been developed either for performance measures such as the misclassification rate, or under the assumption that the number of variables is small compared to the number of observations. Also, most empirical comparisons have been performed with such data. (One exception is Denham [15], who compares (among other estimators) leave-one-out cross-validation, the ordinary bootstrap and the 0.632 estimate (see below) with PLSR on a dataset with 51 observations and 700 variables. Also, Wehrens and van der Linden [16] estimate the prediction error of a PCR model with both leave-one-out cross-validation, the ordinary bootstrap and the 0.632 estimate; however, the main focus is on bootstrap methods for confidence intervals of regression coefficients, and model selection.)

It is not obvious whether these results are valid when estimating the MSEP in situations where PCR and PLSR are commonly used, i.e., when there are more variables than observations. For chemometricians and statisticians using PLSR and PCR, it is important to know how variable the leave-one-out cross-validation is, and whether it might be better to use an alternative estimator. It is especially interesting to know if K-fold cross-validation is less variable than leave-one-out cross-validation.

There are several quality criteria for MSEP estimators, such as the ability to select the correct number of components in the model. The focus in this paper is on the overall closeness of the estimated MSEPs to the true MSEP for 'interesting' model sizes. This is important when one wants to use the estimated MSEP as a measure of the performance of the fitted model.

The present paper investigates the performance of MSEP estimators based on cross-validation or the bootstrap. The estimators are tested using PLSR and PCR on several real data sets, in a simulation where the real data sets are repeatedly split into learning and test data sets. Test set estimates are used as the 'truth', and the estimators are compared in terms of their bias, standard deviation and squared error.

The paper is organised as follows: Section 2 presents the estimators to be tested and some notation. In Section 3 the simulation is described. The results are discussed in Section 4.

2 MSEP estimators

We assume that we have a learning data set $L = \{(\boldsymbol{x}_i, y_i)\}$ of n_L observations, and a predictor f_L trained on L. In the present paper, this will be PLSR or PCR. For the simulations, we also assume that we have a test data set $T = \{(\boldsymbol{x}_{T,i}, y_{T,i})\}$ of size n_T. Both L and T are assumed to be random samples from a common distribution.

The following sections describe the estimators. Their computational costs are summarised in Table 1.

Estimator	# fits	# predictions
$\text{MSEP}_{\text{test}}$	1	n_T
MSEP_{app}	1	n_L
$\text{MSEP}_{\text{cv.K}}$	K	n_L
$\text{MSEP}_{\text{adj.cv.K}}$	$K+1$	$2n_L$
$\text{MSEP}_{\text{naive}}$	R	Rn_L
$\text{MSEP}_{\text{boot}}$	$R+1$	$(R+1)n_L$
MSEP_{BCV}	R	$\approx 0.368Rn_L$
$\text{MSEP}_{0.632}$	$R+1$	$\approx (0.368R+1)n_L$

Table 1: Computational costs of estimators. # fits are the number of times the predictors must be fit (trained). # predictions are the number of (single observation) predictions that must be performed. Usually, the cost of fitting is much higher than the cost of predicting. K and R are described together with the corresponding estimators.

2.1 Test set estimate

The *test set estimate* is the generally admitted criterion of quality. It is defined as

$$\text{MSEP}_{\text{test}} = \frac{1}{n_T} \sum_{i=1}^{n_T} \left(f_L(\boldsymbol{x}_{T,i}) - y_{T,i}\right)^2, \tag{1}$$

where the sum is taken over the test set T. The estimate is unbiased, and its standard deviation given f_L can be estimated by $\sqrt{V_T/n_T}$, where V_T is the sample variance of the squared prediction errors $\{(f_L(\boldsymbol{x}_{T,i}) - y_{T,i})^2\}$. The test set estimate is often simply denoted 'MSEP' in the literature. In this paper, it will be denoted $\text{MSEP}_{\text{test}}$, to separate it from the true MSEP.

2.2 Apparent MSEP

The *apparent MSEP*, also called *mean squared error of calibration (MSEC)*, *mean squared error of estimation (MSEE)* or *resubstitution estimate*, uses the learning data set L as a test set:

$$\text{MSEP}_{\text{app}} = \frac{1}{n_L} \sum_{i=1}^{n_L} \left(f_L(\boldsymbol{x}_i) - y_i\right)^2, \tag{2}$$

where the sum is taken over L. The estimate is in general biased downwards, and the bias increases when more variables or components are added to the model. For ordinary least squares regression (OLSR), if the usual assumptions are correct, the bias of MSEP_{app} is $-2q\sigma^2/n_L$, where $\sigma^2 = \text{Var}(y|\boldsymbol{x})$ and q is the number of parameters in the model [4, p. 292]. This bias can be eliminated by adjusting for the degrees of freedom (df) used in the model (using $n_L - \text{df}$ as divisor). For PCR and PLSR, however, there is no simple expression for the degrees of freedom. Due to its large bias, the apparent MSEP should never be used as an estimate on its own. It is included here because it is part of other estimates.

2.3 Cross-validation

Divide the learning data set L randomly into K segments L_k , $k = 1, \ldots, K$, of roughly equal size. Let f_k be the predictor trained on $L \setminus L_k$, i.e., all observations not in L_k. The *K-fold cross-validation* estimate is

$$\mathrm{MSEP}_{\mathrm{cv.K}} = \frac{1}{n_L} \sum_{k=1}^{K} \sum_{i \in L_k} \left(f_k(\boldsymbol{x}_i) - y_i \right)^2 , \tag{3}$$

where the inner sum is taken over the observations in the kth segment [4]. The estimate is sometimes denoted *mean squared error of cross-validation (MSECV)*. Some authors define the K-fold cross-validation as $1/K \sum_{k=1}^{K} 1/\#L_k \sum_{i \in L_k} \left(f_k(\boldsymbol{x}_i) - y_i \right)^2$, where $\#L_k$ is the size of the kth segment (see for instance [7]), however, the difference is small if the segments are of roughly equal size. The bias of $\mathrm{MSEP}_{\mathrm{cv.K}}$ is of order $(K-1)^{-1} n_L^{-1}$ [7].

The *leave-one-out cross-validation* or *full cross-validation* is the K-fold cross-validation with $K = n_L$. This is a nearly unbiased estimate: The bias is $O(n_L^{-2})$ [7] (i.e., the bias is approximately c/n_L^2 for large n_L and some finite constant c). It has been shown that under reasonable conditions, leave-one-out cross-validation is asymptotically optimal for choosing the best model in OLSR [17], in the sense that the MSEP of the chosen model is close to the minimal MSEP. It is however not asymptotically consistent for selecting variables, in that it tends to include too many variables [18]. This is because the MSEP usually increases only slightly when a few unnecessary variables are included, but increases a lot when any important variables are removed.

The leave-one-out cross-validation has been reported to be rather variable for classifications [8, 9, 11–14]. It has also been argued that this can be expected, at least in the case of classification [3–6]. Possible reasons for large variance is that the fitted values do not depend smoothly on the learning data or that the error estimator is not continuous. It has also been argued that K-fold cross-validation will reduce the variance, at the cost of higher bias, and $K \approx \sqrt{n_L}$ or $K \approx 10$ have been proposed as good compromises between variance and bias. On the other hand, Burman [6] shows that for OLSR, $\mathrm{Var}(\mathrm{MSEP}_{\mathrm{cv.K}} - \mathrm{MSEP}) > \mathrm{Var}(\mathrm{MSEP}_{\mathrm{adj}} - \mathrm{MSEP}) > \mathrm{Var}(\mathrm{MSEP}_{\mathrm{loo}} - \mathrm{MSEP})$, where MSEP is the true MSEP, $\mathrm{MSEP}_{\mathrm{adj}}$ is the adjusted cross-validation defined below, and $\mathrm{MSEP}_{\mathrm{loo}}$ is the leave-one-out cross-validation of MSEP; but that the difference is negligible when K is large.

2.4 Adjusted cross-validation

In K-fold cross-validation the predictors are trained on subsets of L, and can therefore be expected to perform worse than a predictor trained on all of L, especially if $K \ll n_L$. This can lead to an overestimated MSEP. The *Adjusted K-fold cross-validation* [7] tries to adjust for this. The adjustment is:

$$\mathrm{MSEP}_{\mathrm{adj}} = \mathrm{MSEP}_{\mathrm{app}} - \frac{1}{n_L} \sum_{k=1}^{K} \frac{n_k}{n_L} \sum_{i \notin L_k} \left(f_k(\boldsymbol{x}_i) - y_i \right)^2 , \tag{4}$$

where n_k is the size of the kth segment, and the inner sum is taken over $L \setminus L_k$. This is the difference in apparent MSEP between the predictor trained on all L and the weighted average of the predictors trained on $L \setminus L_k$. The adjusted cross-validation estimate is

$$\mathrm{MSEP}_{\mathrm{adj.cv.K}} = \mathrm{MSEP}_{\mathrm{cv.K}} + \mathrm{MSEP}_{\mathrm{adj}}. \tag{5}$$

The bias correction is most prominent when K is small. It can be shown that the bias of $\mathrm{MSEP}_{\text{adj.cv.K}}$ is $O\left((K-1)^{-1}n_L^{-2}\right)$ [7].

2.5 Naive bootstrap estimate

From the learning data set L, we draw R bootstrap samples L_r^*, $r = 1, \ldots, R$. Let f_r^* be the predictor trained on L_r^*.

A naive application of the bootstrap is simply to average the MSEP when the bootstrap predictors predict the learning data L:

$$\mathrm{MSEP}_{\text{naive}} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_L}\sum_{i=1}^{n_L}\left(f_r^*(\boldsymbol{x}_i) - y_i\right)^2. \tag{6}$$

This will be called the *naive bootstrap estimate*. Each bootstrap predictor is trained on a part of L, and is then tested on L. The estimate is therefore biased downwards, but usually not as much as the apparent MSEP. The estimate is not used by itself, but as part of other bootstrap estimates.

2.6 Ordinary bootstrap estimate

In general, the bootstrap is often most successful when used for estimating the bias of an estimate. The bootstrap estimate of the bias of the apparent MSEP is

$$\mathrm{Bias}_{\text{app}} = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{n_L}\sum_{i=1}^{n_L}(f_r^*(\boldsymbol{x}_i) - y_i)^2 - \frac{1}{n_L}\sum_{i=1}^{n_L}(f_r^*(\boldsymbol{x}_{r,i}^*) - y_{r,i}^*)^2\right), \tag{7}$$

where $(\boldsymbol{x}_{r,i}^*, y_{r,i}^*)$ is the ith observation of the rth bootstrap sample [3, p. 252]. Thus, for each bootstrap sample, the expression in parentheses measures the difference between predicting the complete data set and predicting the bootstrap sample. The bias estimate is the average of these differences.

Correcting the apparent MSEP with this estimate leads to the *ordinary bootstrap estimate* [8]:

$$\mathrm{MSEP}_{\text{boot}} = \mathrm{MSEP}_{\text{app}} + \mathrm{Bias}_{\text{app}}. \tag{8}$$

$\mathrm{Bias}_{\text{app}}$ can be expected to underestimate the bias of the apparent MSEP, leading to a downward biased $\mathrm{MSEP}_{\text{boot}}$. $\mathrm{MSEP}_{\text{boot}}$ has been reported to be biased, but less variable than leave-one-out cross-validation [3, 8, 9, 11]. It has been shown that for OLSR, $\mathrm{MSEP}_{\text{boot}}$ is asymptotically inconsistent for selecting the number of parameters [19], just like leave-one-out cross-validation [18].

2.7 Bootstrap smoothed cross-validation

For each $i \in \{1, \ldots, n_L\}$, let R_{-i} be the number of bootstrap samples that do not include observation i. The *bootstrap smoothed cross-validation estimate*, also called the *leave-one-out bootstrap estimate* is defined as

$$\mathrm{MSEP}_{\text{BCV}} = \frac{1}{n_L}\sum_{i=1}^{n_L}\frac{1}{R_{-i}}\sum_{r:i\notin L_r^*}(f_r^*(\boldsymbol{x}_i) - y_i)^2, \tag{9}$$

where the inner sum is taken over those bootstrap samples r that do not include observation i [4, 12]. Thus for each observation i, it smooths the leave-one-out estimate by averaging over all bootstrap predictors not trained on i. The estimate should therefore be expected to have lower variance than leave-one-out cross-validation, at least when used on unstable predictors or discrete performance measures [12]. It can be expected to have a positive bias, because on the average only 63 % distinct observations of the original learning data are used for each prediction [9, 12].

2.8 The 0.632 estimate

The *0.632 estimate* is defined as

$$\mathrm{MSEP}_{0.632} = 0.632\mathrm{MSEP}_{\mathrm{BCV}} + (1 - 0.632)\mathrm{MSEP}_{\mathrm{app}}. \tag{10}$$

The weight 0.632 was originally determined by heuristic arguments based on distance (in probability) of observations from the $\boldsymbol{x}$ in the learning data set. The number $0.632 \approx (1-e^{-1})$ is approximately the average fraction of distinct observations in each bootstrap data set. The $\mathrm{MSEP}_{\mathrm{BCV}}$ usually has positive bias, and it can be shown that for OLSR, $2/3\mathrm{MSEP}_{\mathrm{BCV}} + 1/3\mathrm{MSEP}_{\mathrm{app}}$ is unbiased to terms of order n_L^{-1} [20], i.e., its bias is $O(n_L^{-1})$. More complicated calculations suggest that 0.632 generally is a good choice [4, p. 298]. $\mathrm{MSEP}_{0.632}$ was introduced in [9] and slightly modified in [12]. The modified version is used here. The estimate has performed well in several studies [9, 13–15, 21, 22].

3 Simulation

The estimators were tested in a simulation using real data sets. The simulation was performed to evaluate the overall performance of the estimators.

The following 12 estimators were tested: the apparent MSEP, leave-one-out cross-validation, K-fold and adjusted K-fold cross-validation with $K = 10$, 5 and 2, the naive bootstrap, the ordinary bootstrap, BCV and the 0.632 estimate. Each estimator was tested on six real data sets, using PCR or PLSR trained on data sets of two different sizes. All combinations were used. The six datasets were:

Wheat1: Near Infra-Red (NIR) reflection spectra and protein content measurements of 258 wheat samples. The NIR spectra had 759 wavelengths from 782nm to 2298nm. One outlying observation was deleted, leaving 257 observations in the data set.

Wheat2: A data set with the same 258 wheat samples, but with NIR transmission spectra as covariates. The spectra had 100 wavelengths from 850nm to 1050nm.

Grass: Protein content and NIR reflectance measurements on 301 samples of grass used for feed. The spectra consisted of 700 wavelengths from 1100nm to 2498nm.

Maize: NIR reflectance and cellulose content measurements on 449 maize whole plants. The NIR spectra consisted of 700 wavelengths from 1100nm to 2498nm. This is a part of a data set used in [23].

Cheese: A data set with 277 observations from cheese production, with 477 wavelengths FT-IR spectra as covariates and measured dry matter as response.

Code	Description
D	Real data set (Wheat1, Wheat2, Grass, Maize, Cheese, Beef)
n_L	Learning data set size (50 or 100)
L	Learning data set replicate (1, ..., 100)
R	Type of regression (PLSR or PCR)
A	Model size ($A_{\text{opt}} - 1, \ldots, A_{\text{opt}} + 4$)
E	MSEP estimator

Table 2: Factor codes used in formulae and ANOVA tables. The levels of E are: MSEP_{app}, leave-one-out cross-validated MSEP, $\text{MSEP}_{\text{cv.K}}$ and $\text{MSEP}_{\text{adj.cv.K}}$ with $K = 2$, 5 and 10, $\text{MSEP}_{\text{naive}}$, $\text{MSEP}_{\text{boot}}$, MSEP_{BCV}, $\text{MSEP}_{0.632}$.

Beef: A data set with 338 observations from tenderness experiments, consisting of 351 wavelengths NIR reflectance from 1100nm to 2500nm, and Warner-Bratzler shear force [24] measured on *longissimus dorsi* of beef.

The regressions were trained on data sets of $n_L = 50$ and $n_L = 100$ observations. Models with 1, 2, ..., $A_{\max} = 20$ components were trained on the small data sets, and with 1, 2, ..., $A_{\max} = 25$ components on the large data sets. However, only a subset of the number of components were used for testing and comparing the estimators (see below).

Results for the two learning data set sizes were analysed separately, because not all of the estimators are directly comparable across different sizes. For instance, a 50-fold cross-validation is the same as a leave-one-out cross-validation if $n_L = 50$, but not if $n_L = 100$. It is also interesting in itself to see the results separately.

The simulation was performed in the following manner. For a given real data set and learning data set size (n_L), the real data set was randomly divided into a learning data set (with n_L observations) and a test set (with the rest of the observations) 100 times. This creates 100 pairs of learning and test data sets. For each pair, the following was calculated. PCR and PLSR regressions with up to $A_{\max}$ components were trained on the learning data. All 12 MSEP estimators were evaluated for each of these models, using only the n_L observations in the learning data set. Finally, the test set was used to calculate the test set MSEP for each of the models.

In the simulation, the test set MSEPs are considered as the true MSEPs. The estimated MSEPs were divided by the corresponding 'true' MSEP to give a relative estimate $\widehat{\text{MSEP}}/\text{MSEP}_{\text{test}}$. The relative estimate has expectation 1 if the estimator is unbiased and we ignore the variability of the test set estimate. (The standard deviation of $\text{MSEP}_{\text{test}}$ ranged from 5.8 % to 17.6 % of the estimate for the different combinations, with an average of 12.1 %.) This facilitates comparison of estimates across model sizes, regression types and data sets.

Thus for each combination of real data set D, size of learning data sets n_L, regression type R, model size A and estimator E, we have 100 pseudo-replicates of the relative estimate $\widehat{\text{MSEP}}/\text{MSEP}_{\text{test}}$. These will be denoted $\text{est}(D, n_L, R, A, E)_l$ for replicate $l = 1, 2, \ldots, 100$ in the formulae below. The bias, variance and squared error of these 100 replicates were calculated as

$$\text{Bias}(D, n_L, R, A, E) = \overline{\text{est}}(D, n_L, R, A, E) - 1, \tag{11}$$

$$\mathrm{Var}(D, n_L, R, A, E) = \frac{1}{99} \sum_{L=1}^{100} \left(\mathrm{est}(D, n_L, R, A, E)_l - \overline{\mathrm{est}}(D, n_L, R, A, E)\right)^2, \qquad (12)$$

$$\mathrm{sqe}(D, n_L, R, A, E) = \frac{1}{100} \sum_{L=1}^{100} \left(\mathrm{est}(D, n_L, R, A, E)_l - 1\right)^2, \qquad (13)$$

where $\overline{\mathrm{est}}(D, n_L, R, A, E)$ is the average of $\mathrm{est}(D, n_L, R, A, E)_l$ over the replicates l. Intuitively, the bias measures how much the estimator under- or over-estimates the true MSEP, and the variance measures how variable the estimator is. The squared error combines the errors of bias and variance, and indicates on the average how far (in squared distance) the estimate is from the true MSEP (or more precisely, how far $\widehat{\mathrm{MSEP}}/\mathrm{MSEP}_{\mathrm{test}}$ is from 1).

It should be noted that the 100 replicates are not independent, because the learning and testing data are drawn from from the same data set. This will influence the variance estimate (12) (and therefore the sqe). If the correlation between the $\mathrm{est}(D, n_L, R, A, E)_l$ replicates is ρ, and their true variance is σ^2, the expected value of (12) is $\sigma^2(1-\rho)$. (This can be shown in general, for n correlated observations x_i, by noting that $\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$ can be written $\sum_{i<j}(x_i - x_j)^2/(n(n-1))$ and that $\mathsf{E}(x_i - x_j)^2 = 2\sigma^2(1-\rho)$ when $i \neq j$.) Assuming that the correlation is positive, the variance will be underestimated. We believe it is reasonable to assume that the correlation is similar for the different estimators, in which case the variance estimate will be similarly affected. Also, the larger the real data set is, the smaller the correlation will be.

For each combination of real data set, size of learning data sets and regression type, the six 'most interesting' model sizes were selected by a subjective evaluation of the average of the test set estimates: First the optimal model size A_{opt}, i.e., the number of components that would have been used in practice, was identified. The 'interesting' model sizes were then chosen as $A_{\mathrm{opt}} - 1$, A_{opt}, ..., $A_{\mathrm{opt}} + 4$, i.e., from one component less than A_{opt} to four components more. For instance, for PLSR on the NIR-T data with $n_L = 50$, $A_{\mathrm{opt}} = 6$ was chosen, so the selected model sizes were 5, 6, ..., 10. Note that in general, A_{opt} was not the model size with smallest test set MSEP, but the model size that was judged as the one that would be used in practice. The reason for selecting such a subset of model sizes is that this is where it is most important for estimators to perform well. The performance of an estimator on models with 1 or 20 components is not very interesting if the optimal model size is 6.

We want to compare the estimators E for the different real data sets D and learning data set sizes n_L. In order to get average statistics for each combination of these factors, the biases, variances and squared errors above were averaged over the regression types and the selected model sizes. 'Average' standard deviations were calculated by taking the square root of the averaged variances.

4 Results and Discussion

In order to get a rough overview of the effects of the design factors on the bias of the estimators, an ANOVA was performed for each learning data set size n_L. The relative estimate is a function of quadratic prediction errors, and cannot be assumed to be normally distributed. Quantile plots and plots of residuals suggested using the logarithm in order to get a more normally distributed response. Thus, the logarithm of the relative estimate averaged over the

6 model sizes, i.e., $\ln\left(\sum_A \text{est}(D, n_L, R, A, E)_l/6\right)$, was used as response. The learning data set replicates (L) are random and nested within real data set (D). All other factors are fixed. The ANOVAs are multistratum ANOVAs with four error strata [25]. The ANOVA results are shown in Tables 3 and 4. The codes denoting the factors are listed in Table 2.

The significant effects at a 0.05 level were the main effect of estimator (E) and its interactions with the other factors. Of these, the estimator effect was by far the largest. The other significant effects were very small in comparison. This means that on the average, the estimators performed similarly (in terms of bias) on the six data sets and on PLSR and PCR. The interaction between real data set and estimator indicates a small difference between the performance of the estimators on different data sets, at least for small learning data sets.

A second set of ANOVAs were also performed, using all model sizes (1, 2, ..., $A_{\max}$ components) for $n_L = 50$ and $n_L = 100$. The results (not shown) were very similar to the shown ANOVAs.

The average bias, standard deviation and squared error of the estimates on each data set are shown in Figures 1 and 2 for the two different learning set sizes.

The apparent MSEP and the naive and ordinary bootstrap estimates had a negative bias; all other estimates had positive bias. In all cases, ordinary bootstrap, 0.632, leave-one-out cross-validation, 10-fold cross-validation and 5- and 10-fold adjusted cross-validation were close to unbiased. For the large learning data sets ($n_L = 100$), the 2-fold adjusted cross-validation was only moderately biased. In 5 of the 12 cases, the 0.632 estimate was the least biased.

The apparent MSEP, the naive bootstrap and (sometimes) the ordinary bootstrap estimates had appreciably lower standard deviation than the other estimates. Similarly, the 2-fold and 2-fold adjusted cross-validation and the BCV had higher standard deviation that the rest. Apart from these, there were only small differences between the estimates in terms of standard deviation. The bootstrap estimates were calculated using 100 bootstrap samples. Their variance could probably be reduced slightly by increasing the number of samples, at the cost of more computation.

No reasonably unbiased estimate had substantially lower variability than the leave-one-out cross-validation. Thus it seems that the reported results for classifiers are not completely valid for PLSR and PCR on high dimensional data. In fact, the variance (as well as the bias) of K-fold cross-validation and adjusted K-fold cross-validation increases as K decreases in all instances. This is in accordance with the findings of Burman [6], who studied MSEP estimation in linear regression. Denham [15] studied MSEP estimation for PLSR, and also reports similar standard deviations for leave-one-out cross-validation, the ordinary bootstrap and the 0.632 estimate.

A heuristic, possible explanation is the following: The main situations in which the leave-one-out cross-validation could be expected to be variable is when the predictor is unstable (such as a classification or regression tree), or when the error measure is discontinuous (such as misclassification error). In both situations a small perturbation of the learning data set could result in just as large a difference in prediction error as a larger perturbation would have resulted in, i.e., the prediction error is not a smooth function of the underlying data. In leave-one-out cross-validation there are many small perturbations, whereas in K-fold cross-validation there are fewer but larger perturbations. One could therefore expect the variance of the leave-one-out cross-validation to be at least as high as K-fold cross-validation.

However, a linear regression is quite stable, and the MSEP is a continuous function, so the MSEP of a linear regression is a smooth function of the data. Therefore larger perturbations of the data would lead to larger differences in MSEP. In this situation, leave-one-out cross-

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Error: L:D					
D	5	8.06	1.61	1.53	0.1775
Residuals	594	624.22	1.05		
Error: L:D:R					
R	1	0.08	0.08	0.07	0.7918
D:R	5	8.07	1.61	1.47	0.1973
Residuals	594	651.56	1.10		
Error: L:D:E					
E	11	684.94	62.27	8212.91	0.0000
D:E	55	123.26	2.24	295.60	0.0000
Residuals	6534	49.54	0.01		
Error: Within					
R:E	11	6.41	0.58	71.14	0.0000
D:R:E	55	8.06	0.15	17.90	0.0000
Residuals	6534	53.53	0.01		

Table 3: ANOVA table for ln(average relative estimate) with $n_L = 50$. Each term is tested against the residuals within its stratum.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Error: L:D					
D	5	1.57	0.31	0.42	0.8383
Residuals	594	449.57	0.76		
Error: L:D:R					
R	1	2.83	2.83	3.70	0.0550
D:R	5	3.64	0.73	0.95	0.4469
Residuals	594	454.16	0.76		
Error: L:D:E					
E	11	247.15	22.47	8399.78	0.0000
D:E	55	42.90	0.78	291.63	0.0000
Residuals	6534	17.48	0.00		
Error: Within					
R:E	11	4.03	0.37	145.14	0.0000
D:R:E	55	4.42	0.08	31.88	0.0000
Residuals	6534	16.48	0.00		

Table 4: ANOVA table for ln(average relative estimate) with $n_L = 100$. Each term is tested against the residuals within its stratum.

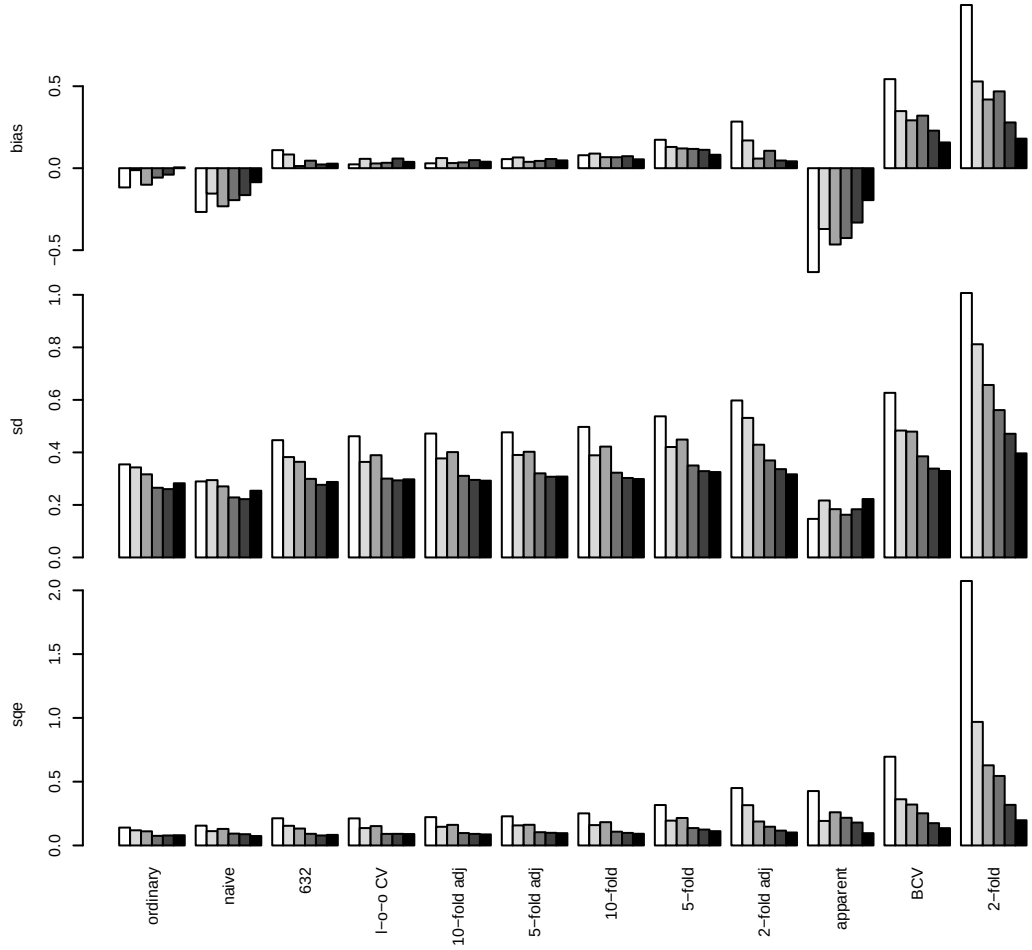

Figure 1: Average bias (top panel), standard deviation (middle panel) and squared error (bottom panel) of the estimates. Each group of bars corresponds to one estimator, and the bars within each group represent the different data sets; from left (white) to right (black): Wheat1, Wheat2, Grass, Maize, Cheese and Beef. Learning data set size 50. The estimators are sorted in order of increasing squared error.

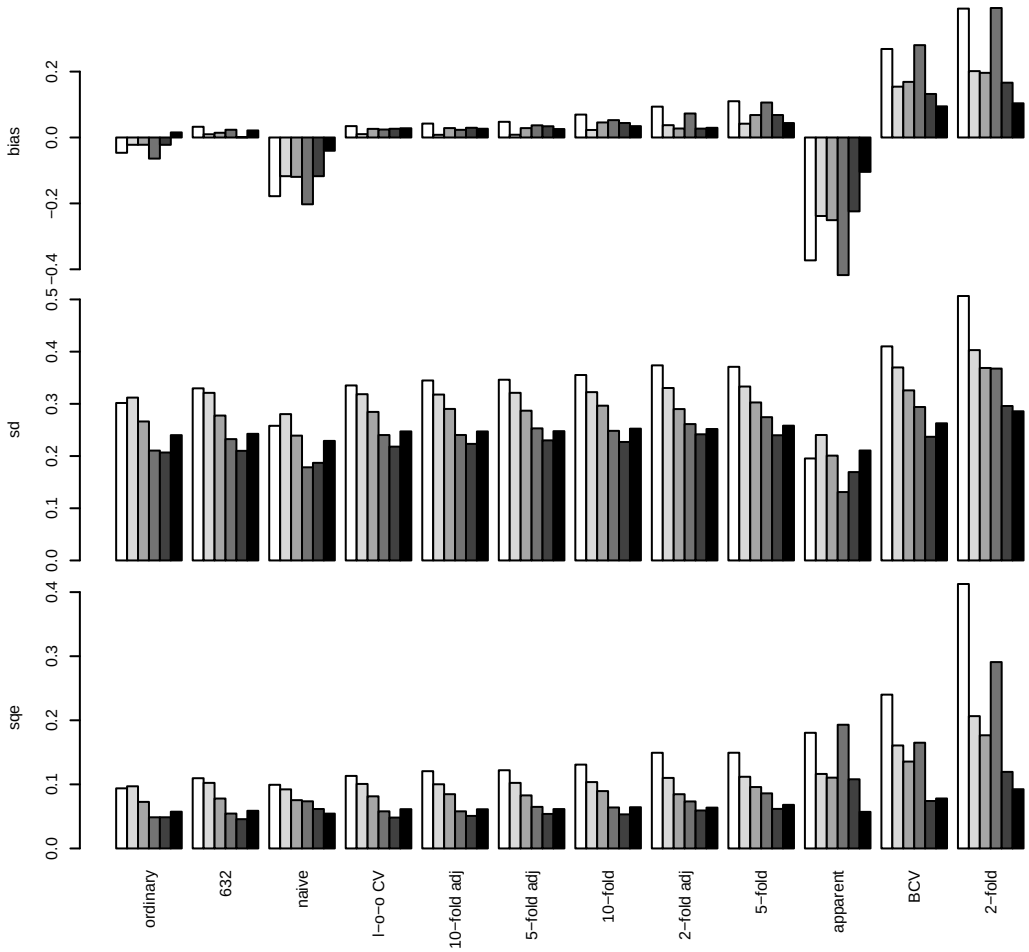

Figure 2: Average bias (top panel), standard deviation (middle panel) and squared error (bottom panel) of the estimates. Each group of bars corresponds to one estimator, and the bars within each group represent the different data sets; from left (white) to right (black): Wheat1, Wheat2, Grass, Maize, Cheese and Beef. Learning data set size 100. The estimators are sorted in order of increasing squared error.

validation is the average of many, stable values, while K-fold cross-validation is the average of a few, more variable, values. This would lead to the variance of K-fold cross-validation MSEP being higher than with leave-one-out cross-validation.

The BCV is often described as a smoothed version of cross-validation. It would therefore be expected that it had lower standard deviation than leave-one-out cross-validation. In these simulations, however, it had substantially *higher* standard deviation. The reason might be the same as for the increased variance of K-fold cross-validation.

The bias, standard deviation and squared error were calculated on the six most 'interesting' model sizes, as defined in Section 3. We also performed the calculations on all model sizes 1, 2, ..., $A_{\max}$ components. In general, this gave the same results. However, the downward bias of the apparent MSEP, the naive and ordinary bootstrap estimates, was aggravated. This is due to the increased overfitting of the models when far too many components are included. Therefore, one should be careful about using the ordinary bootstrap estimate, at least with many components in the model. Also, the variance of the cross-validation estimates and the BCV increased compared to the other bootstrap estimates. This resulted in the 0.632 estimate having a somewhat smaller squared error than the other (unbiased) estimators.

The observed differences between the estimators will probably be larger for smaller learning data sets (and smaller for larger data sets). The sizes 50 and 100 were chosen because many real applications have data sets within this range of sizes.

5 Conclusions

All in all, a group of estimators seem to perform somewhat better than the others. These are the ordinary bootstrap, the 0.632 estimate, leave-one-out cross-validation, and 10-fold cross-validation and 10- and 5-fold adjusted cross-validation. In terms of squared error, they are very similar. Within this group there are only small differences and none of them can be said to significantly outperform the others.

Contrary to results for classifiers, the alternative estimators were not less variable than leave-one-out cross-validation. In particular, the variance of K-fold cross-validation increased with decreasing K. Also the bootstrap smoothed cross-validation estimate (BCV) was more variable than leave-one-out cross-validation.

On the basis of these results, it seems that the 0.632 estimate or leave-one-out cross-validation should be used for estimating the MSEP of PCR and PLSR. If computing time is a problem, the adjusted 10- or 5-fold cross-validation seem to be good choices, due to their much lower computational demand, and practically unchanged performance compared to the leave-one-out cross-validation.

Acknowledgements

The work was funded by the IBION project, which is sponsored by the Research Council of Norway (project no. 145456-130). The authors wish to thank Prof. Tormod Næs for helpful discussions, and Prof. Harald Martens, Dr. Ellen Mosleth Færgestad, Dr. Kjell Ivar Hildrum, Dr. Gustav Fystro and Mr. Kjetil Jørgensen, for permission to use their data sets.

References

[1] Harald Martens and Tormod Næs. *Multivariate Calibration*. Wiley, Chichester, 1989.

[2] Peter A. Lachenbruch and M. Ray Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10(1):1–11, 1968.

[3] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York, 1993.

[4] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK, 1997.

[5] Brian D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

[6] Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.

[7] Prabir Burman. Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā: The Indian Journal of Statistics*, 52:314–345, 1990.

[8] Bradley Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*, volume 38 of *CBMS-NFS Regional conference series in applied mathematics*. Society for industrial and applied mathematics, Philadelphia, 1982.

[9] Bradley Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

[10] Olaf Bunke and Bernd Droge. Bootstrap and cross-validation estimates of the prediction error for linear regression models. *Annals of Statistics*, 12(4):1400–1424, 1984.

[11] A. C. Davison and Peter Hall. On the bias and variability of bootstrap and cross-validation estimates of error rate in discrimination problems. *Biometrika*, 79(2):279–284, 1992.

[12] Bradley Efron and Robert J. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

[13] Geoffrey J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wileys Series in Probability and Mathematical Statistics. Wiley, New York, 1992.

[14] W. J. Krzanowski and David J. Hand. Assessing error rate estimators: The leave-one-out method reconsidered. *Australian Journal of Statistics*, 39(1):35–46, 1997.

[15] Michael C. Denham. Choosing the number of factors in partial least squares regression: Estimating and minimizing the mean squared error of prediction. *Journal of Chemometrics*, 14(4):351–361, 2000.

[16] R. Wehrens and W. E. Van der Linden. Bootstrapping principal component regression models. *Journal of Chemometrics*, 11(2):157–171, 1997.

[17] Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics*, 15(3):958–975, 1987.

[18] Jun Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494, 1993.

[19] Jun Shao. Bootstrap model selection. *Journal of the American Statistical Association*, 91(434):655–665, 1996.

[20] Peter Hall. On the biases of error estimators in prediction problems. *Statistics & Probability Letters*, 24(3):257–262, 1995.

[21] R. R. Wilcox and J. Muska. Measuring effect size: A non-parametric analogue of ω^2. *British Journal of Mathematical & Statistical Psychology*, 52:93–110, 1999.

[22] W. J. Krzanowski. Data-based interval estimation of classification error rates. *Journal of Applied Statistics*, 28(5):585–595, 2001.

[23] Harald A. Martens and Pierre Dardenne. Validation and verification of regression in small data sets. *Chemometrics and Intelligent Laboratory Systems*, 44(1–2):99–121, 1998.

[24] P. E. Bouton and P. V. Harris. Comparison of some objective methods used to assess meat tenderness. *Journal of Food Science*, 37(2):218–221, 1972.

[25] R. M. Heiberger. *Computation for the Analysis of Designed Experiments*. John Wiley and Sons, New York, 1989.