Statistics and Computing (2000) **10,** 209–229

On the use of cross-validation to assess performance in multivariate prediction

P. JONATHAN*, W. J. KRZANOWSKI† and W. V. MCCARTHY*

* *Shell Research Ltd., Thornton, Chester CH1 3SH, UK*
† *School of Mathematical Sciences, University of Exeter, Laver Building, North Park Rd., Exeter EX4 4QE, UK*

Received December 1998 and accepted September 1999

We describe a Monte Carlo investigation of a number of variants of cross-validation for the assessment of performance of predictive models, including different values of k in leave-k-out cross-validation, and implementation either in a one-deep or a two-deep fashion. We assume an underlying linear model that is being fitted using either ridge regression or partial least squares, and vary a number of design factors such as sample size n relative to number of variables p, and error variance. The investigation encompasses both the non-singular (i.e. $n > p$) and the singular (i.e. $n \leq p$) cases. The latter is now common in areas such as chemometrics but has as yet received little rigorous investigation. Results of the experiments enable us to reach some definite conclusions and to make some practical recommendations.

Keywords: cross-validation, ridge regression, partial least squares, prediction, assessment of predictive models

1. Introduction

1.1. *Assessment of predictive models*

We consider the classical problem of modelling the behaviour of a dependent variable Y by a linear function of explanatory variables $X_1, X_2, \ldots, X_p$, i.e.

$$Y = a_0 + a_1X_1 + \cdots + a_pX_p + \varepsilon$$

where ε is a $N(0, \sigma^2)$ random departure term, and the values of the X_i are observed without error. Typically we have a random sample of $(p+1)$-tuples $\{x_{i1}, x_{i2}, \ldots, x_{ip}, y_i\}_{i=1}^n = \{\underline{x}_i, y_i\}_{i=1}^n$ as the *training* or *design* set of data. Multiple regression analysis uses the principle of least squares to obtain estimates $\hat{a}_i$ of the constants a_i $(i = 0, 1, \ldots, p)$ and hence to define the predictor of Y as

$$\hat{Y} = \hat{a}_0 + \hat{a}_1X_1 + \cdots + \hat{a}_pX_p.$$

A fundamental question is how best to assess the performance of this predictor. One way is to obtain a *test* set of m further observations from the same population, $\{\underline{x}_i, y_i\}_{i=n+1}^{n+m}$ say, and to compare the observed y_i in this set with their predictions $\hat{y}_i$ from the model using a single criterion measure such as the *standardised prediction sum of squares*

$$sPRESS = \sum_{i=n+1}^{n+m} (y_i - \hat{y}_i)^2 \Bigg/ \sum_{i=n+1}^{n+m} (y_i - \bar{y})^2$$

where $\bar{y} = \frac{1}{m}\sum_{n+1}^{n+m} y_i$ is the mean of the dependent variable in the test set.

When no test set is available, we need to base assessments on the training set data only. The simplest idea is *resubstitution*, i.e. comparing the predictions $\{\hat{y}_i\}_{i=1}^n$ for the individuals in the training set with their counterparts y_i. However, this will give an optimistically biased value of *sPRESS*, because least squares perforce finds those $\hat{y}_i$ that are closest to the y_i in the training set and such close matching will not occur for independently gathered data. For a formal demonstration, compare the expected *sPRESS* for a test set in equation (3.10) on page 45 of Brown (1993) with its training set counterpart as deduced from the third equation on page 41 of the same reference.

A favoured alternative is *cross-validation* (Lachenbruch and Mickey 1968, Stone 1974). Here we divide the training data into g equal-sized groups and conduct g separate operations. Each group is omitted in turn from the data, the model is fitted to the remaining $(g-1)$ groups, and the predictions $\hat{y}_i$ are obtained for the omitted group. This yields n predictions $\hat{y}_i$, none of which has used the corresponding y_i as part of the modelling

0960-3174 © 2000 Kluwer Academic Publishers

stage, so the *sPRESS* formed from them should not be optimistically biased. The number of individuals in each omitted group is $k = n/g$, so this method of assessment will be termed *leave-k-out*. One complication is caused by the fact that there are $n!/(g!(k!)^g)$ ways of dividing the training set into g groups each of size k, and different partitions may yield very different performance assessments. One solution is to average *sPRESS* values over different partitions to arrive at an overall assessment. Another is to take $g = n$ and hence $k = 1$, in which case we have *leave-one-out* assessment and an unambiguous partitioning into n groups each containing a single individual.

1.2. *Model selection and tuning*

Complicated predictive models often depend on parameters that can only be optimised (estimated) through data-based inspection, in addition to ones estimated analytically. Two such models are those for ridge regression (RR) and partial least squares (PLS). The former introduces bias into the linear model estimator through an extra "ridge" parameter, with the aim of achieving lower mean square error of estimation, but selection of this parameter must be done before application of least squares (Montgomery and Peck 1982, p. 310). The latter reformulates the model in terms of a set of components optimally related to the dependent variable, but requires a prior selection of the number of such components before fitting the model (Garthwaite 1994). Such prior model selection can itself be done using cross-validation (Stone 1974). For example, the number of PLS components to include in the model can be chosen as the number that yields the lowest *sPRESS* when successively fitting one, two, three components and so on. We call this process tuning.

Assessment of performance of such tuned models on test data is again the best approach, but what if we don't have a test set? The *sPRESS* value for the chosen model is clearly an optimistically biased assessment, because the model has been *chosen* to give the lowest *sPRESS* on the training data. For unbiased assessment, we need a second layer of cross-validation: leave out each group of individuals in turn, use cross-validation on the remaining individuals to both tune and fit the model, and then predict the y_i values for the omitted individuals using the fitted model. This process involves a *two-deep* cross-validation and associated *sPRESS*, as opposed to the *one-deep* cross-validation described earlier.

1.3. *Cross-validation and high dimensionality*

The aim of cross-validation is to mimic the prediction of *future* individuals from the population. This will be achieved if the training data fully represents the sample space and each omitted individual can lie anywhere in this space. Large samples and small dimensionality generally satisfy these requirements. Many modern application areas such as spectroscopy, however, give rise to data of very high dimensionality with severe restrictions on the size of samples (e.g. Krzanowski *et al.* 1995). With such small samples and high dimensionality, the training data are likely to fall in a very small fraction of the sample space (the "curse of dimensionality"; Bellman 1961), and any omitted unit from the training set will only come from this restricted area. Cross-validation may therefore fall far short of replicating the conditions of a test set, so that as dimensionality increases the method may become less reliable.

1.4. *Previous work*

Cross-validation was developed for error rate estimation in discriminant analysis, with theoretical contributions by Hills (1966) and Lachenbruch and Mickey (1968) and an application to the classification of the Federalist papers by Mosteller and Tukey (1968). These papers all used one-deep leave-one-out assessment of error of a classification rule. Stone (1974) considered the full range of situations as outlined above and coined the term "two-deep" for cross-validatory assessment of a model itself chosen by cross-validation. Efron (1982) gave a theoretical formulation, including asymptotic theory and connections with other data resampling methods such as the jackknife and bootstrap. A recent extension to multivariate predictions using multiple regression by Breiman and Friedman (1997) included cross-validatory estimation of shrinkage and ridge parameters. Of the various offshoots of cross-validation, the method of *generalized* cross-validation (Golub, Heath and Wahba 1979) is noteworthy in that it permits analytical approximation to the numerical process.

Leave-*k*-out cross-validation was popularised in the chemometrics area, where it is often termed "*g*-fold" cross-validation, through applications such as that by Wold (1978) into cross-validatory choice of number of components in principal component analysis. Mertens *et al.* (1995) give some recent developments in the context of principal component regression. Theoretical and computational investigations have been conducted into the influence of k on results. Shao (1993) established that consistency improves as k increases, while Altman and Leger (1997) came to similar conclusions in respect of asymptotic optimality.

Another consideration is numerical speed and accuracy. Rannar *et al.* (1995) compared the numerical accuracy of several algorithms for partial least squares, which involves cross-validation as an integral issue. Breiman (1996) provides a full discussion of the problems of instability in model selection, including suggestions for improving stability in the presence of cross-validation.

The necessity for two-deep cross-validation has been stressed by Ganeshanandam and Krzanowski (1989) whenever discriminant rules are constructed by optimising cross-validation error rates, and by Krzanowski (1995) whenever selection of variables is based on cross-validated error rates. Despite such warnings, there is often still a reluctance to use the two-deep variant.

Many other references to cross-validation may be found in the literature, but we have focussed above purely on its use in assessment of performance of predictive rules.

1.5. *Objectives*

We wished to investigate the efficacy of different variations of cross-validation as methods of assessing performance in multivariate prediction, especially when tuning is necessary to obtain the predictive model. Using simulation, it is always possible to generate a test set against which to assess predictions made by any given model, so the benchmark is the test-set *sPRESS* value for any given combination of experimental factors. Against this benchmark we computed *sPRESS* values for both one-deep and two-deep cross-validation over a range of values of k in leave-k-out, in order to investigate the properties of the methods and arrive at some general recommendations. We also looked at different random partitions of the training data when $k > 1$ in leave-k-out in order to investigate their effects.

The data were generated from a specific linear model (but see 3.7), and certain factors were varied over the experiment. Sample size n was in the range [10, 100] with dimensionality p set at 50 as a typical value in practice, covering both the non-singular ($n > p$) and the singular ($n \leq p$) cases. Also, we applied both RR and PLS to all the data sets. The presumption was that they would perform comparably, but we wished to discover whether any particular conditions were more favourable to one or other method. We describe the simulation experiment in detail in Section 2, give the results from it in Section 3, and present the conclusions together with discussion in Section 4.

2. Description of simulation experiment

2.1. *The covariance structure*

An orthonormal basis $\{\underline{\gamma}_j\}_{j=1}^p$ for Euclidean p-space $\mathbf{R}^p$ together with an eigenvalue decay profile $\{\lambda_j\}_{j=1}^p$ defines the covariance structure of the explanatory-variable space, $\underline{\Sigma} = \Gamma\Lambda\Gamma^T$ in obvious notation. Observations $\underline{x}$ can then be drawn at random from the multivariate normal distribution $N(\underline{0}, \underline{\Sigma})$.

2.2. *The underlying model*

A model of the form $y = \sum_{j=1}^p \beta_j \underline{x}^T \underline{\gamma}_j + e$ was taken, where $\{\beta_j\}_{j=1}^p$ specifies the regression relationship in terms of the principal components $\{\underline{x}^T \underline{\gamma}_j\}_{j=1}^p$. (The model is equivalent to $y = \sum_{j=1}^p \underline{x}^T \underline{\alpha}_j + e$ for $\underline{\alpha}_j = \beta_j \underline{\gamma}_j$, $j = 1, 2, \ldots, p$.) The error term e, independent of $\underline{x}$, was taken to be normally-distributed $N(0, \varepsilon^2 \sigma_0^2)$ for a constant $\varepsilon \in [0, 1]$, where σ_0^2 is the variance of the "error-free response" $m = \sum_{j=1}^p \beta_j \underline{x}^T \underline{\gamma}_j$, easily shown to be $\sigma_0^2 = \mathrm{var}(\sum_{j=1}^p \beta_j \underline{x}^T \underline{\gamma}_j) = \sum_{j=1}^p \beta_j^2 \lambda_j$, a constant. Response variance σ^2 is therefore given by $\sigma^2 = (1 + \varepsilon^2)\sigma_0^2$, so that the proportion of σ^2 due to error is specified via ε.

2.3. *Estimating the predictive model and assessing its performance*

For a given realisation $\{\underline{x}_i\}_{i=1}^n$ of a sample of explanatory data, a corresponding set of responses $\{y_i\}_{i=1}^n$ was generated using the model $y_i = \sum_{j=1}^p \beta_j \underline{x}_i^T \underline{\gamma}_j + e_i$, $i = 1, 2, \ldots, n$. If we write $m_i = \sum_{j=1}^p \beta_j \underline{x_i}^T \underline{\gamma}_j$ then $s_0^2 = \frac{1}{n-1} \sum (m_i - \bar{m})^2$ is an estimate of σ_0^2. We drew $\{e_i\}_{i=1}^n$ randomly from $N(0, \varepsilon^2 s_0^2)$ rather than from $N(0, \varepsilon^2 \sigma_0^2)$ in order to keep a fixed ratio between the amount of information in the error-free and error components for each realisation. PLS and RR were then used to develop a predictive model for the response in terms of the explanatory data. Tuning was performed using leave-k-out cross-validation, for some value of k.

There were three measures of assessment for each simulation:

1. The *sPRESS* value on applying the chosen model to an independent sample (size 1000) of test data; this is the *one-deep external* assessment.
2. The optimal value of *sPRESS* obtained during model tuning; this is the *one-deep internal* assessment.
3. The *sPRESS* value following a nested cross-validation strategy; this is the *two-deep internal* assessment.

2.4. *The simulations*

nReal realisations of the training sample were generated for each factor combination, and values $k = \{1, 2, 5, 10\}$ were used for leave-k-out cross-validation, except for the smallest sample sizes. For $k > 1$, the data must be partitioned into g groups, where g is the smallest integer greater than or equal to n/k. Here n has been chosen as a multiple of k. To quantify the effect of partitioning on the variability of *sPRESS*, *nPart* $= 20$ different partitions of the data were taken for each realisation.

Elements of the orthonormal basis $\{\underline{\gamma}_j\}_{j=1}^p$ were obtained by Gram-Schmidt orthonormalisation of vector elements randomly sampled from $N(0, I)$. For most simulations, a linear eigenvalue decay profile $\{\lambda_j\}_{j=1}^p = \{50, 49, \ldots, 1\}$ was chosen; some variations from this are described in 3.7. In the underlying model, the form $\{\beta_j\}_{j=1}^p = \{1, 0, 1, 0, 1, 0, 0, \ldots, 0\}$ was taken, so that the response is defined in terms of a linear combination of the first, third and fifth principal components of the explanatory data. Further, simulations for each of the values $\varepsilon = \{0.3, 0.4, 0.6\}$ were undertaken. Again, some variations are described in 3.7.

3. Results

For each combination of sample size n, error factor ε, method (RR or PLS), value of k in leave-k-out cross-validation, and assessment method (1-deep internal, 1-deep external or 2-deep internal), we report the median, mean and standard error of the mean for *sPRESS* estimated from modelling *nReal* realisations of the sample. To allow for the *nPart* random partitions of the data when $k > 1$, we use a repeated measures-type estimate for the standard error:

$$\text{Standard error} = \sqrt{\sum_{i=1}^{nReal} \frac{(\bar{r}_{i\bullet} - \bar{r}_{\bullet\bullet})^2}{nReal\,(nReal - 1)}}$$

Table 1. *A brief guide to the simulation experiments reported*

Description	Section	Figure
Original model *Error $\varepsilon = 0.3$; Sample size $n = 10, 20, 30, 40, 70, 100$;*	2	
Methods = RR, PLS; Leave-out-$k = 1, 2, 5, 10$; 20 random data partitions		
Results (for $n = 100, 70, 40, 30, 20, 10$ respectively)	A.1–A.6	
• Effect of sample size on *sPRESS*	3.2	2
• Empirical densities for *sPRESS*	3.2	3
• Comparison of PLS and RR	3.3	4
• Effect of k in leave-k-out	3.5	6
• Variability due to choice of random partition	3.6	
• Effect of n on optimal ridge parameter/number of PLS factors	4.2	7
Original model *Error $\varepsilon = 0.4, 0.6$; Sample size $n = 40$;*		
Methods = RR, PLS; Leave-out-$k = 1, 2, 5, 10$; 20 random data partitions		
Results (for $\varepsilon = 0.4$ and 0.6 respectively)	A.7–A.8	
• Effect of error variance	3.4	5
Modified models (i) "PC 5 + 7 + 9" and (ii) "quickly decaying eigenspectrum" *Error $\varepsilon = 0.3$; Sample size $n = 20$;*		
Methods = RR, PLS; Leave-out-$k = 1, 2, 5, 10$; 20 random data partitions		
Results (for modified models (i) and (ii) respectively)	A.9–A.10	
• Effect of varying the model specification	3.7	

where r_{ij} is the value for *sPRESS* for the ith realisation of the data and the jth random partition, and dots represent averages over the corresponding suffix.

We are particularly interested in comparing different assessment methods. For this reason, values of the differences in *sPRESS* between the 3 assessment methods are also provided. To allow comparison of RR and PLS performance, differences in *sPRESS* between RR and PLS are given for each assessment method (see Section 3.3).

Simulations consisting of 100 realisations were performed for sample sizes of 100, 70, 40 and 30. As sample size decreases, the variability of estimates for the location and spread of distributions of *sPRESS* (and differences in *sPRESS*) tend to increase. For this reason, 1 000 realisations were simulated for sample size 20, and 10 000 for sample size 10. Section 3.2 examines the effect of sample size more generally in these simulations. For sample size 40, simulation results are reported for different error variance factors $\varepsilon = 0.3$, 0.4 and 0.6 (see Section 3.4 for further detail).

Table 1 provides a concise overview of the simulation experiments undertaken:

Comprehensive tables of results are included as an appendix (Appendix A.1–A.10). In the sections below, we extract the more salient features from those tables to illustrate the main findings of the simulations.

3.1. *Overview for sample size $n = 70$*

Figure 1 gives a flavour for the results obtained, using sample size $n = 70$ and $\varepsilon = 0.3$ for illustration (see Appendix A.2 for fuller results). Figure 1(a) shows the values of *sPRESS* obtained for RR prediction using leave-1-out cross-validatory assessment. *sPRESS* values corresponding to 1-deep external (1de, 'o') and 2-deep internal assessment (2di, '*') are shown. Values range from about 0.1 to about 0.3, and the mean *sPRESS* for 2-deep internal assessment appears to be somewhat larger that the corresponding value for 1-deep external assessment. If we take the latter as characteristic of true performance, this observation implies that 2-deep internal assessment provides slightly pessimistic estimates of future predictive performance on average for this simulation. Figure 1(b) shows the corresponding very similar results for PLS modelling. The difference between 1-deep external and 2-deep internal values for each simulation is shown in Fig. 1(c), for RR ('o') and PLS ('*') modelling. The mean value of this difference is negative for both RR and PLS. The results in Fig. 1(d) suggest, on average, that PLS performs slightly more poorly than RR for this simulation, since the mean value of the difference between *sPRESS* values for PLS and RR is positive for both 1-deep external and 2-deep internal assessment (see Section 3.3).

Estimates for the mean and median *sPRESS* (and differences in *sPRESS*) values, and the standard error of the mean, are shown in Appendix A.2. In particular, we note that the mean difference ('1di-1de') between 1-deep internal and 1-deep external *sPRESS* values for PLS using leave-1-out assessment ('PLS lo1') is not significant (mean = −0.004, standard error = 0.004), whereas both the corresponding mean differences between 2-deep internal and 1-deep external ('2di-1de', mean = 0.016, standard error = 0.005), and 1-deep internal and 2-deep internal ('1di-2di', mean = −0.020, standard error = 0.001) are significant. Differences in estimates of predictive performance from different assessment methods are highlighted in Sections 3.2 and 3.3.

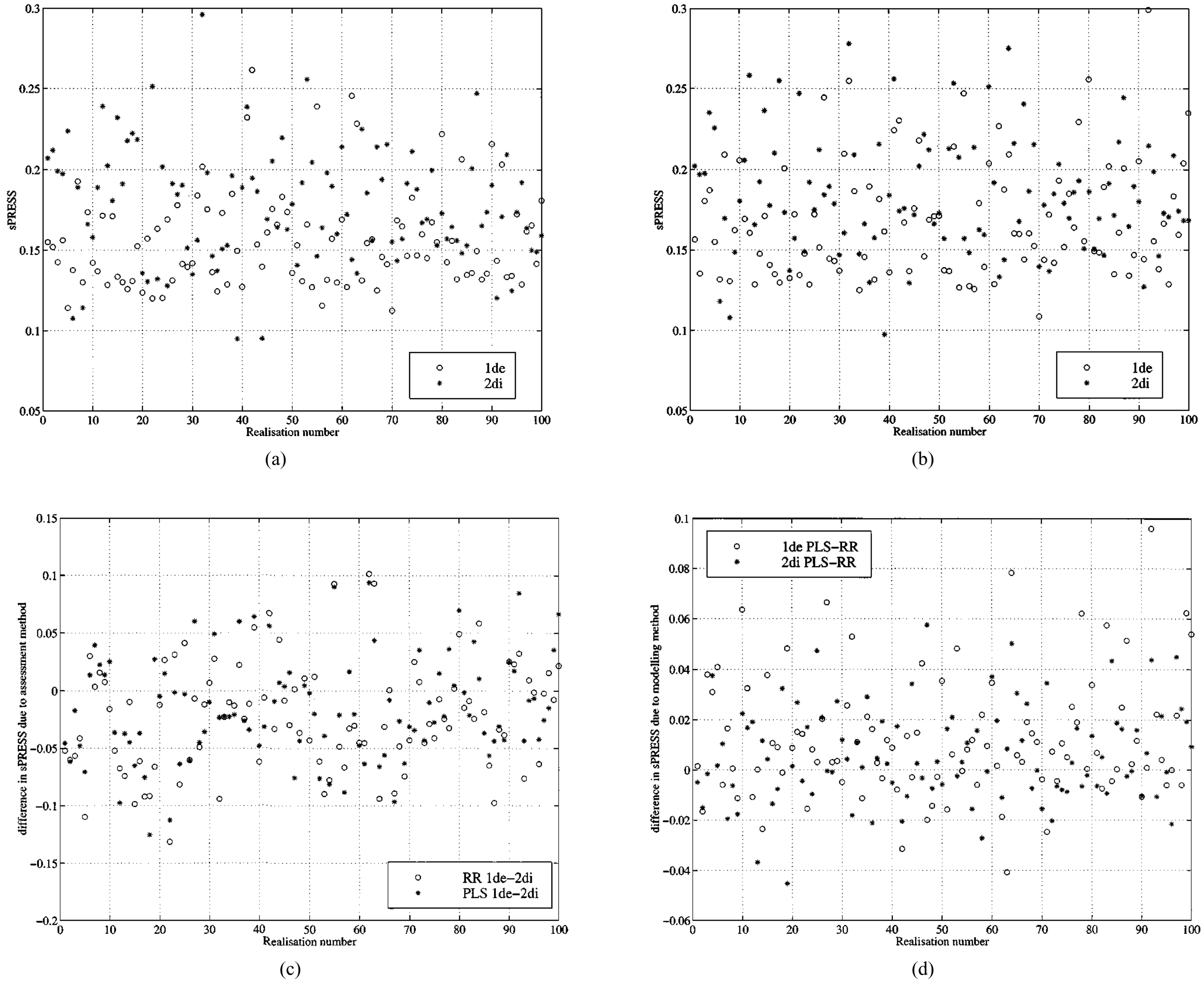


Fig. 1. *(a) sPRESS values for 100 realisations of $n = 70$ RR leave-1-out (b) sPRESS values for 100 realisations of $n = 70$ PLS leave-1-out (c) Difference in sPRESS values for 100 realisations of $n = 70$ leave-1-out (d) Difference in sPRESS values for 100 realisations of $n = 70$ leave-1-out*

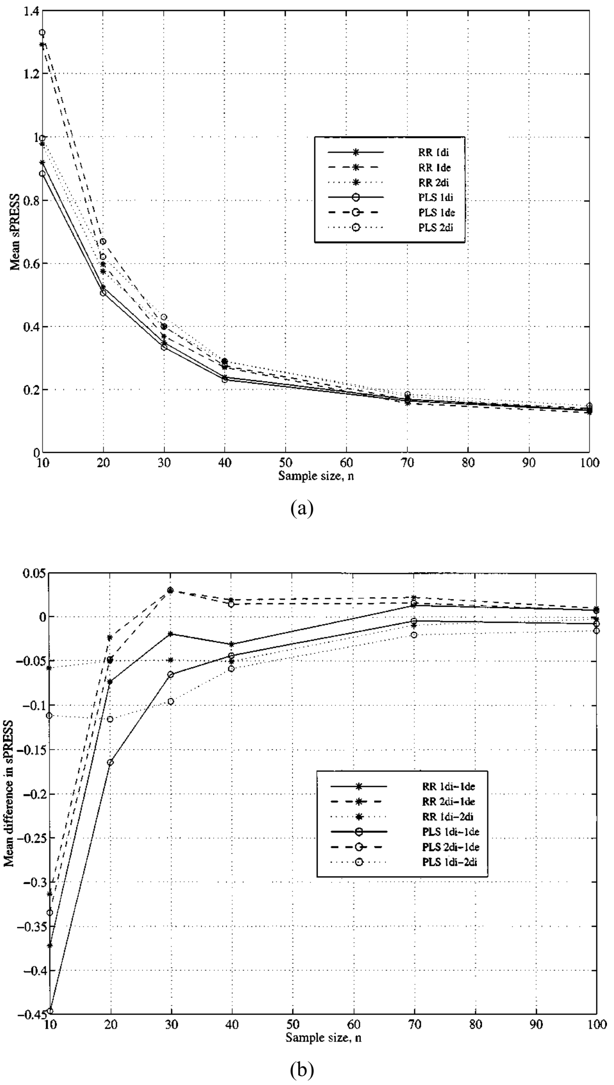
Also included with the results for 'PLS lo1' are median, mean and standard error estimates for the optimal number of PLS factors chosen for 1-deep internal and 2-deep internal assessment ('nFac1di', 'nFac2di' respectively). For leave-1-out ridge regression ('RR lo1'), the corresponding optimal shrinkage factors 'Shri1di' and 'Shri2di' are given. In the case of 2-deep assessment, the number of PLS factors or the value of the shrinkage factor obtained in each run was in fact the average over the values obtained for each unit omission in the outer loop. The summary statistics in Appendix A.2 are therefore means, medians and standard errors of these averages. Averaging means gives smaller standard errors for the 2-deep assessments (column 8) than for the 1-deep assessments (column 7) where only single values are averaged. The shrinkage factor is expressed as a fraction of the total sample variance; for 'RR lo1', the median optimal shrinkage factor for 1-deep internal assessment is 0.400, with a mean value of 0.387 and a standard error of 0.015. That is, a shrinkage factor of approximately 0.4 times the total sample variance is typically used in this case. The effect of sample size and assessment method on the choice of the optimal number of PLS factors and RR shrinkage factor is discussed further in Section 4.

Appendix A.2 also illustrates how the choice of k in leave-k-out affects predictive performance. Typically, as the value of k increases, the mean (and median) value of *sPRESS* increases for internal assessment methods using both RR and PLS. The choice of k has less effect on performance measured using 1-deep external assessment. These issues are explored more fully in Section 3.5.

The final block of results in Appendix A.2 compares the performance of PLS with respect to RR for this simulation. Values for the median and mean (and its standard error) difference in *sPRESS* values for each of the 3 assessment approaches are quoted, for each leave-out possibility. For sample size $n = 70$, 1-deep internal assessment suggests that PLS will perform slightly better than RR. In contrast, the more reliable 1-deep external and 2-deep internal assessment methods suggest the converse (see Section 3.3).

3.2. *The effect of sample size on assessment method*

Intuition suggests that predictive performance will become progressively poorer as design sample size decreases. Simulation results confirm this. Figure 2(a) shows mean *sPRESS* for leave-1-out assessment as a function of sample size. For large sample sizes ($n = 100$, 70) there is good agreement between mean values for 1-deep internal, 1-deep external, 2-deep internal assessment for both RR and PLS predictions. As sample size decreases, however, the bias of 1-deep internal assessment becomes clear, with PLS suffering slightly more than RR. In contrast, 2-deep internal assessment compares well with the 1-deep external data for sample sizes of 30 and larger. For the very smallest sample sizes ($n = 20$, 10) however, it appears that even estimates of predictive performance based on 2-deep internal assessment are biased with respect to the 1-deep external results. Inspection of Appendices A.5 and A.6 (for $n = 20$, 10) demonstrates this effect further. Figure 2(b) gives mean differences in *sPRESS* between the three assessment methods, for each of RR and PLS analysis, and shows clearly the bias of 1-deep internal and 2-deep internal values with respect to 1-deep external.



(a)

(b)

Fig. 2. *(a) Effect of sample size on mean sPRESS for leave-1-out (b) Differences in sPRESS between assessment methods for leave-1-out*

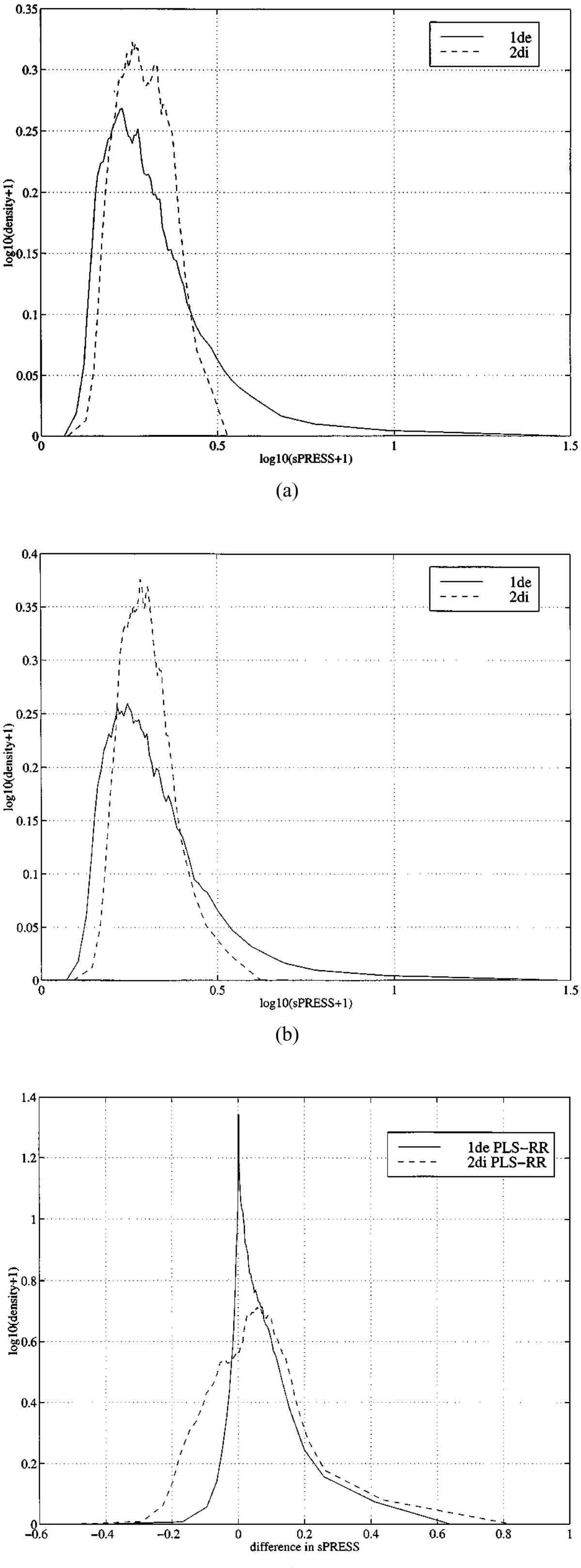
Interestingly, the mean difference between *sPRESS* values corresponding to 1-deep internal and 2-deep internal assessment is still relatively small, even for small sample sizes. This can be perhaps attributed to the fact that during cross-validation the sample space is bounded by the training data, but an external test set may easily produce data far outside these bounds.

The bias of 2-deep internal assessment, with respect to 1-deep external, is most marked for sample size $n = 10$. To quantify the effect as precisely as possible, a simulation of 10 000 realisations of samples size 10 was performed. Empirical density functions for the distributions of *sPRESS* from RR and PLS analyses (for both 1-deep external and 2-deep internal assessment) are given in Fig. 3(a) and (b). The distributions are all seen to be skewed (note the log scales on the axes), with the distributions for 1-deep external assessment extremely so. The largest values of 1-deep external *sPRESS* are in the region of 30, whereas the largest values of 2-deep internal *sPRESS* are around 3, for both RR and PLS. It is important to note, however, that the realisations which yield large values for 1-deep external *sPRESS* for RR also yield large values for PLS analysis. This can be seen from inspection of Fig. 3(c) which gives empirical density functions for the differences between *sPRESS* values for PLS and RR, for each of 1-deep external and 2-deep internal assessment (note log scale on ordinate axis only). The distributions in Fig. 3(c) have much lower spread than those in Fig. 3(a) and (b). The minimum and maximum values recorded, for 1-deep external and 2-deep internal differences respectively, are −0.43, 0.64 and −0.51, 0.83.

For small sample sizes, Figs. 2 and 3 back the suppositions made towards the end of Section 1.3: the training data will necessarily fall in a small fraction of the data space, so are not truly representative of the population and hence the 2-deep internal assessment will occasionally be far more optimistic than reality as characterized by 1-deep external assessment. The fact that choice of tuning parameter is based on $n - 2$ observations and a prediction model is based on $n - 1$ observations will add to this effect.

3.3. *Comparison of RR and PLS*

Results in Sections 3.1 and 3.2 have already alluded to some small difference in the performance of PLS and RR in

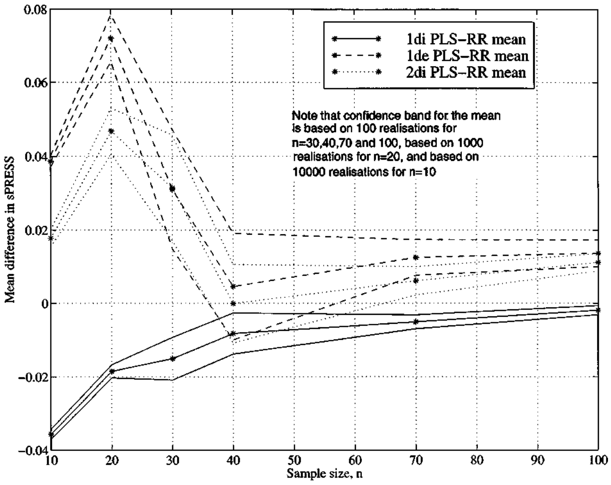


(a)

(b)

(c)

Fig. 3. *(a) Empirical density for sPRESS: n = 10 RR leave-1-out (b) Empirical density for sPRESS: n = 10 PLS leave-1-out (c) Empirical density for difference in sPRESS: n = 10 leave-1-out*



Note that confidence band for the mean is based on 100 realisations for n=30,40,70 and 100, based on 1000 realisations for n=20, and based on 10000 realisations for n=10

Fig. 4. *Mean differences in sPRESS between PLS and RR for leave-1-out*

simulations. In general, results (see Appendix) suggest that for 1-deep internal assessment (regardless of sample size or choice of k in leave-k-out), values of mean and median *sPRESS* for PLS are smaller than for RR. This trend is illustrated in Fig. 4 for leave-1-out analyses, which also shows the corresponding trends for 1-deep external and 2-deep internal assessment.

Approximate (point-wise) 95% confidence bands for each of the three trends are also shown, but it should be noted that simulations for sample sizes $n = 10$ and $n = 20$ involve different numbers of realisations compared with the remainder of sample sizes explored.

For 2-deep internal assessment, Fig. 4 suggests that PLS performs more poorly than RR, especially for small sample sizes. This behaviour is largely confirmed by the 1-deep external curve. The latter does however suggest that 2-deep internal assessment understates the difference between PLS and RR performance. (It is not our intention to establish superiority of either method, and we don't believe that there are good reasons why RR should perform better than PLS for our simulation set-up).

3.4. *The effect of error variance*

Data in Appendices A.3, A.7 and A.8 permit us to compare simulation results for sample size $n = 40$ for different error variance factors $\varepsilon = 0.3$, 0.4 and 0.6 respectively. For leave-1-out assessment, values for mean *sPRESS* are illustrated in Fig. 5. It is interesting to compare Fig. 5 (for increasing ε) with Fig. 2(a) (for decreasing sample size, n); there is a strong similarity in the trends observed. In general, as ε increases, both 1-deep internal and 2-deep internal assessment tend towards over-optimistic assessment of future predictive performance. This is rather like the behaviour reported in Section 3.2 above for decreasing sample size, and in fact the two effects are probably related: as error variance increases, cross-validation is unable to capture the true variation in the population values.

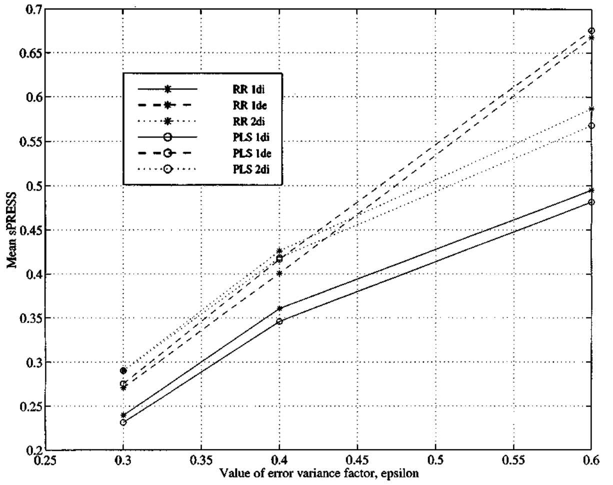


Fig. 5. *Effect of model error variance on mean sPRESS for leave-1-out analysis on n = 40*

3.5. *The value of k in leave-k-out*

For sample size $n = 70$, the behaviour of mean *sPRESS* for RR and PLS models, assessed using each of 1-deep internal, 1-deep external and 2-deep internal approaches is illustrated in Fig. 6(a). It is apparent that the mean 1-deep external *sPRESS* is relatively unaffected as the value of k in leave-k-out increases, for both PLS and RR. For both internal assessment approaches, however, the mean *sPRESS* increases gradually with increasing k.

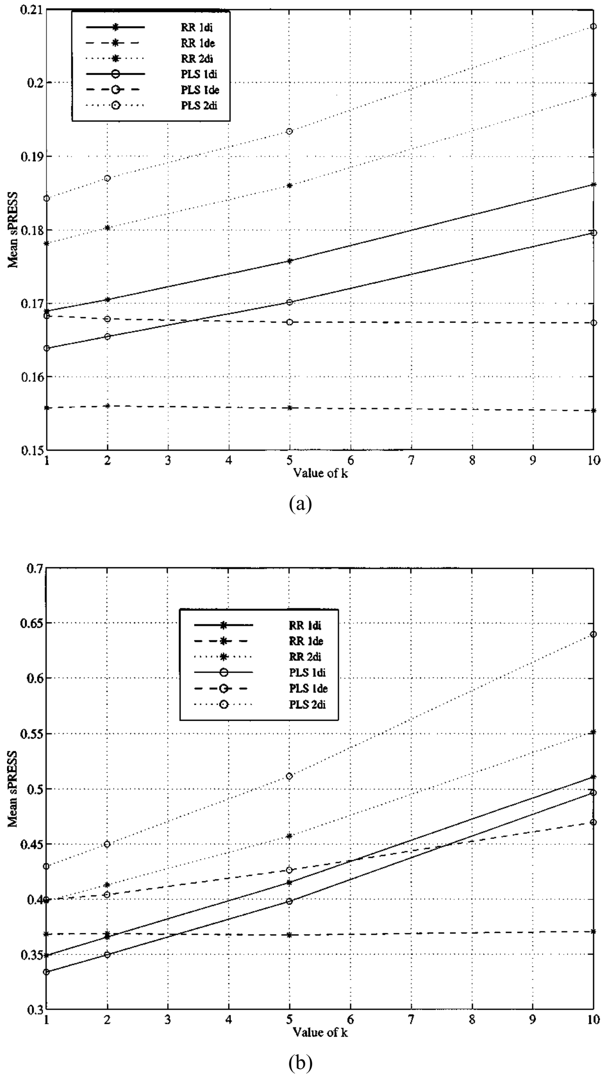
It is interesting to note (Appendix A.2) that the mean optimal RR shrinkage factor decreases with increasing k (and the corresponding mean optimal number of PLS factors increases).

The corresponding result for $n = 30$ is given in Fig. 6(b). The trends observed are quite similar to those in Fig. 6(a). In this situation (Appendix A.4), however, the mean optimal number of PLS factors decreases with increasing k, and no obvious trend is apparent for the size of the RR shrinkage factor.

The trend in these figures has the following possible explanation. All 1-deep external estimates are based on models that use n observations regardless of k, which accounts for the (relative) flatness of the corresponding plots. However, the choice of tuning parameter for 1-deep and 2-deep internal estimates is based on models using $n - k$ and $n - 2k$ observations respectively. Prediction models are each based on $n - k$ observations. The increasing *sPRESS* for these methods reflects the decreasing amount of data available.

3.6. *Variability attributable to the choice of random partition*

For leave-k-out cross-validation, with $k > 1$, arbitrary partitioning of the data is necessary to facilitate assessment of predictive performance. Partitioning therefore provides an extra source of variation for *sPRESS*. In the simulations reported here, predictive rules were generated based on $nPart = 20$ different random partitions for each realisation of the training sample. Simulation results therefore allow us to estimate the proportion of total variance in *sPRESS* (for any particular case) attributable to the partitioning. Results are given in Table 2 below for RR models; the corresponding results for PLS are very similar.



(a)

(b)

Fig. 6. *(a) Effect of k in leave-k-out on mean sPRESS for n = 70 (b) Effect of k in leave-k-out on mean sPRESS for n = 30*

The table explores the effect of sample size and error variance on the percentage of *sPRESS* variance attributable to partitioning. We see that, for 1-deep internal assessment, percentage variance attributable to partitioning decreases with increasing sample size, increasing error variance and decreasing value of k in leave-k-out. Results for 2-deep internal assessment were very similar to those for 1-deep internal assessment so are not given. For 1-deep external assessment, the proportion of variance attributable to partitioning is much smaller. Note however that the percentage variance attributable to partitioning now increases with increasing error variance.

Table 2. *Percentage of sPRESS variance in ridge regression attributable to sample partitioning for leave-k-out (k > 1) cross-validation*

	Varying sample size, n (Constant $\varepsilon = 0.3$)					Varying ε (Constant $n = 40$)		
	20	30	40	70	100	0.3	0.4	0.6
1-deep internal assessment								
lo2	7.9	6.7	4.7	2.4	1.5	4.7	3.8	3.3
lo5	27.4	20.9	16.4	9.6	6.6	16.4	13.7	12.6
lo10	–	37.9	30.4	19.5	13.4	30.4	25.2	24.8
1-deep external assessment								
lo2	0.9	0.5	2.8	0.9	0.6	2.8	5.0	12.9
lo5	2.3	2.0	4.4	1.4	1.1	4.4	6.6	4.5
lo10	–	4.6	6.6	2.1	1.9	6.6	7.8	8.2

3.7. *The effect of varying the model*

All the above results came from the same underlying model, as described in Sections 2.2 and 2.4. At the instigation of one of the referees, we did some further simulations to check on consistency of results when varying the model. Specifically, we looked at the case $n = 20$, $\varepsilon = 0.3$, *nReal* $= 500$ and *nPart* $= 20$ and conducted two variations of the previous model:

(i) we defined the response in terms of a linear combination of the fifth, seventh and ninth principal components of the explanatory data instead of the previous first, third and fifth components;

(ii) we used a "quickly decaying" eigenspectrum $\{\lambda_j\}_{j=1}^{p} = \{50, 45, 40, \ldots, 10, 5, 0, \ldots, 0\}$ in place of the previous regular one $\{\lambda_j\}_{j=1}^{p} = \{50, 49, \ldots, 1\}$.

Results for (i) and (ii) are shown in Appendix A.9 and A.10 respectively; these results may be compared most directly with those in Appendix A.5. Although the *sPRESS* values are larger in A.9 than in A.5, and those in A.10 are smaller than in A.5, the trends within both A.9 and A.10 are consistent with those within A.5. Moreover, since the quickly decaying eigenstructure of A.10 implies a dimensionality $10 < n = 20$, the results in A.10 mimic those in A.1 and A.2 for $n > p$. It would thus appear that our results show consistency across model changes.

4. Conclusions, discussion and recommendations

4.1. *Conclusions*

The following conclusions can be drawn from the results of the simulation experiments:

(i) In the presence of tuning one-deep internal assessment provides biased estimates of predictive performance. The extent of the bias increases with decreasing sample size and increasing error variance, and is related to the precision with which tuning parameters are estimated.

(ii) Two-deep internal assessment provides reasonable estimates of predictive performance for all cases except when n is much smaller than p. For the smallest sample sizes, two-deep internal (as well as one-deep internal) assessment over-estimates predictive performance, due predominantly to the design set not being sufficiently representative of the population. In such circumstances, the estimate of predictive performance is valid only for predictions whose regressor variables fall in the subspace defined by the regressor variables in the design set. Thus any future test data should be checked for concordance with the design set before admission for testing.

(iii) The value of k in leave-k-out influences internal estimates of predictive performance, the size of the effect increasing with decreasing sample size and increasing value of k. This effect may be due to the fact that only $n - k$ (for one-deep assessment) or $n - 2k$ (for two-deep assessment) individuals remain for tuning, and $n - k$ in both cases for model construction. The size of the effect for 2-deep internal assessment does not differ markedly from the size of the effect for 1-deep internal assessment. External estimates of predictive performance are less sensitive to the choice of k.

(iv) The random partitioning of data is a source of variation in leave-k-out ($k > 1$) internal estimates of predictive performance. The proportion of variation in internal estimates for *sPRESS* attributable to random partitioning increases with increasing k and decreasing sample size. As might be expected, external estimates are less sensitive to the choice of random partition.

(v) Ridge regression performs slightly better than partial least squares regression in general. RR appears to give one-deep internal estimates for *sPRESS* which are slightly less biased than those produced by PLS, in general, for the same assessment method. There is no obvious reason, from consideration of the simulation set-up, why RR should outperform PLS. (It should be noted, however, that a grid search method was used for RR. A finer or coarser grid might have yielded slightly different results.)

4.2. *Optimal tuning parameters*

A peripheral issue in the results quoted above, but a central one as regards operation of the predictive model, is the optimal choice of tuning parameters (i.e. number of PLS factors or RR shrinkage constant). In all the results above we have simply assumed that this prior choice has been made, and then concentrated on the resulting *sPRESS* values. However, the chosen values of the tuning parameters also vary from run to run of the experiment, and this variation merits consideration.

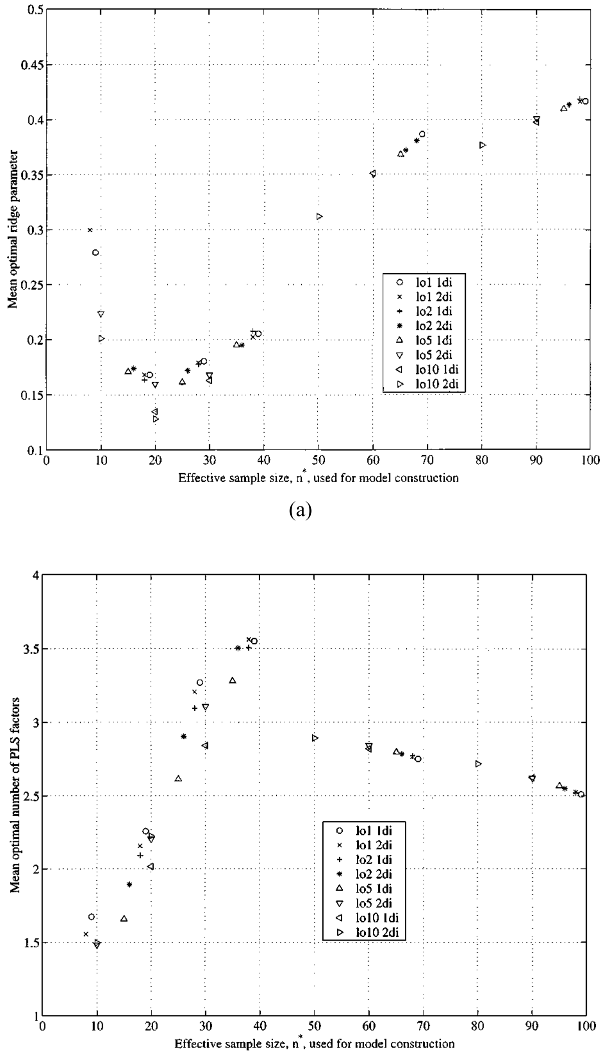
For each sample size n and each value of k in leave-k-out, we have recorded the value of the optimal tuning parameters for both 1-deep and 2-deep internal assessment. Since these tuning parameters are selected using an "effective sample size" n^* of $n - k$ and $n - 2k$ observations respectively, by plotting the value of the optimal tuning parameter against n^* we can obtain the optimal tuning parameter value as a function of the actual numbers of observations used for tuning. Results are shown in Fig. 7 below.

For ridge regression (Fig. 7a) there is a minimum shrinkage factor corresponding to an effective sample size of about 20 or 30, while for PLS (Fig. 7b) there is a maximum number of factors corresponding to an effective sample size of about 30 or 40. These results need some explanation, and we might conjecture as follows.

As n^* increases beyond 30, then there is enough data to identify the X subspace of principal components 1, 3 and 5 corresponding to variation in the response, so RR shrinkage increases (thereby reducing the relative influence of principal components with small sample variance). As n^* decreases below 20, however, then the estimates for the variance associated with any principal component have large variability (with severe overestimation of large eigenvalues and underestimation of small ones), so larger and larger shrinkage is required to yield a more realistic balance between the variances of the components. For PLS, on the other hand, as n^* decreases below 40, there is less and less correlation with the response in successive factors so that the optimal number of PLS factors decreases too, but as n^* increases above 40 then PLS is able to capture progressively more concisely the subspace of X variation corresponding to variation in the response. Equilibrium between the two effects is achieved when n^* is about 40.

4.3. *Recommendations*

The following general recommendations can be made, based on the simulation experiments reported here. In the simulations, the underlying dimensionality of the explanatory space is equal to the number of explanatory variables p. In general, it is possible (e.g. for spectroscopic data from analytical chemistry) that the underlying dimensionality is actually less than p. For this reason, the recommendations below, although referring to the number of explanatory variables p, should be understood as referring to the actual underlying dimensionality.



(a)

(b)

Fig. 7. *(a) Effect of sample size on mean optimal ridge parameter (b) Effect of sample size on mean optimal number of PLS factors*

(i) *We recommend the use of* 2*-deep internal assessment for cases when the number of explanatory variables exceeds the number of observations* ($n < p$). However, if $n \ll p$ then the warning in 4.1(ii) should be borne in mind and caution exercised when interpreting calculated values.

(ii) For large sample sizes (e.g. $n = 100$ in this work), the choice of k in the domain [1, 10] is largely arbitrary. As n decreases, however, the effect of k on internal estimates of predictive performance increases, due to random partitioning and the fact that only $n - k$ (for one-deep assessment) and $n - 2k$ (for two-deep assessment) individuals remain for tuning. For $n = 20$, simulation suggests that there is little difference between the choices $k = 1$ and $k = 2$ in terms of estimation of predictive performance. Further, the variability of *sPRESS* for $k = 2$ increases only slightly due to random partitioning. In addition, leave-2-out is computationally considerably faster that leave-1-out. *For this reason we recommend the use of leave-2-out cross-validation for all sample sizes in the interval* [20, p].

(iii) If a leave-k-out assessment approach is adopted with $k > 4$, then results here suggest that the estimate of predictive performance can be quite sensitive to the choice of random partition. For this reason, *reanalysis of a number of different random partitions of the design sets should always be considered.* This will give some indication of variation in predictive performance over random partitions, and enable confidence intervals to be constructed for this quantity.

Appendix: Full tables of simulation results

A.1. *sPRESS for sample size n = 100, error factor ε = 0.3, nReal = 100 realisation, nPart = 20 partitions, k = {1, 2, 5, 10}*

```
Basic results

PLS lo 1   1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.131  0.135  0.145 -0.007   0.009  -0.014   2.000   2.238
mean   :  0.133  0.141  0.148 -0.007   0.008  -0.015   2.510   2.525
sterr  :  0.002  0.003  0.003  0.003   0.003   0.001   0.056   0.055
PLS lo 2   1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.131  0.135  0.146 -0.005   0.010  -0.014   2.000   2.392
mean   :  0.134  0.140  0.149 -0.006   0.009  -0.015   2.521   2.550
sterr  :  0.002  0.003  0.002  0.003   0.003   0.001   0.055   0.053
PLS lo 5   1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.133  0.134  0.149 -0.002   0.014  -0.015   3.000   2.714
mean   :  0.136  0.138  0.152 -0.002   0.013  -0.016   2.568   2.620
sterr  :  0.002  0.002  0.002  0.003   0.002   0.001   0.049   0.046
PLS lo10   1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.137  0.132  0.152  0.004   0.019  -0.015   3.000   2.818
mean   :  0.140  0.137  0.156  0.003   0.019  -0.016   2.624   2.717
sterr  :  0.002  0.002  0.002  0.002   0.003   0.001   0.045   0.043
 RR lo 1   1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.132  0.126  0.134  0.007   0.009  -0.002   0.400   0.399
mean   :  0.135  0.127  0.137  0.008   0.010  -0.002   0.417   0.417
sterr  :  0.002  0.001  0.002  0.002   0.002   0.000   0.012   0.012
 RR lo 2   1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.133  0.126  0.135  0.009   0.011  -0.002   0.400   0.398
mean   :  0.136  0.127  0.138  0.009   0.011  -0.002   0.419   0.414
sterr  :  0.002  0.002  0.002  0.002   0.002   0.000   0.012   0.011
 RR lo 5   1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.135  0.126  0.138  0.011   0.014  -0.002   0.400   0.393
mean   :  0.138  0.127  0.141  0.011   0.014  -0.003   0.410   0.401
sterr  :  0.002  0.001  0.002  0.002   0.002   0.000   0.011   0.011
 RR lo10   1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.140  0.126  0.143  0.016   0.019  -0.003   0.400   0.373
mean   :  0.142  0.127  0.146  0.016   0.019  -0.003   0.398   0.377
sterr  :  0.002  0.001  0.002  0.002   0.002   0.000   0.011   0.010

Comparison of PLS and ridge regression

                     lo 1  PLS1di PLS1de PLS2di
                           -RR1di -RR1de -RR2di
                 median : -0.001  0.008  0.009
                 mean   : -0.002  0.014  0.011
                 sterr  :  0.001  0.002  0.001
                     lo 2  PLS1di PLS1de PLS2di
                           -RR1di -RR1de -RR2di
                 median : -0.002  0.008  0.009
                 mean   : -0.002  0.013  0.011
                 sterr  :  0.001  0.002  0.001
                     lo 5  PLS1di PLS1de PLS2di
                           -RR1di -RR1de -RR2di
                 median : -0.002  0.007  0.010
                 mean   : -0.002  0.011  0.011
                 sterr  :  0.001  0.001  0.001
                     lo10  PLS1di PLS1de PLS2di
                           -RR1di -RR1de -RR2di
                 median : -0.002  0.005  0.009
                 mean   : -0.002  0.011  0.010
                 sterr  :  0.001  0.001  0.001
```

A.2. *sPRESS for sample size $n = 70$, error factor $\varepsilon = 0.3$, nReal = 100 realisations, nPart = 20 partitions, $k = \{1, 2, 5, 10\}$*

```
Basic results

PLS lo 1    1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.158  0.160  0.179  0.000   0.022  -0.019   3.000   2.958
mean   :  0.164  0.168  0.184 -0.004   0.016  -0.020   2.750   2.764
sterr  :  0.003  0.004  0.004  0.004   0.005   0.001   0.082   0.078
PLS lo 2    1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.161  0.160  0.182  0.001   0.023  -0.020   3.000   2.861
mean   :  0.165  0.168  0.187 -0.002   0.019  -0.022   2.772   2.785
sterr  :  0.003  0.003  0.004  0.004   0.005   0.001   0.079   0.076
PLS lo 5    1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.165  0.160  0.188  0.006   0.030  -0.021   3.000   2.833
mean   :  0.170  0.167  0.193  0.003   0.026  -0.023   2.797   2.840
sterr  :  0.003  0.003  0.004  0.004   0.005   0.001   0.076   0.072
PLS lo10    1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.175  0.160  0.202  0.014   0.040  -0.025   3.000   2.750
mean   :  0.180  0.167  0.208  0.012   0.040  -0.028   2.818   2.893
sterr  :  0.004  0.003  0.004  0.004   0.005   0.001   0.071   0.067
 RR lo 1    1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.167  0.150  0.176  0.012   0.024  -0.007   0.400   0.381
mean   :  0.169  0.156  0.178  0.013   0.022  -0.009   0.387   0.381
sterr  :  0.003  0.003  0.004  0.005   0.005   0.001   0.015   0.015
 RR lo 2    1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.168  0.149  0.180  0.014   0.025  -0.007   0.400   0.369
mean   :  0.170  0.156  0.180  0.014   0.024  -0.010   0.381   0.373
sterr  :  0.004  0.003  0.004  0.005   0.005   0.001   0.015   0.014
 RR lo 5    1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.173  0.149  0.184  0.021   0.031  -0.008   0.400   0.345
mean   :  0.176  0.156  0.186  0.020   0.030  -0.010   0.369   0.351
sterr  :  0.004  0.003  0.004  0.005   0.005   0.000   0.014   0.013
 RR lo10    1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.184  0.148  0.196  0.032   0.043  -0.009   0.300   0.300
mean   :  0.186  0.155  0.198  0.031   0.043  -0.012   0.352   0.312
sterr  :  0.004  0.003  0.004  0.005   0.005   0.000   0.012   0.011

Comparison of PLS and ridge regression

                    lo 1  PLS1di  PLS1de  PLS2di
                          -RR1di  -RR1de  -RR2di
               median : -0.005   0.008   0.003
               mean   : -0.005   0.013   0.006
               sterr  :  0.001   0.002   0.002
                    lo 2  PLS1di  PLS1de  PLS2di
                          -RR1di  -RR1de  -RR2di
               median : -0.005   0.008   0.005
               mean   : -0.005   0.012   0.007
               sterr  :  0.001   0.002   0.002
                    lo 5  PLS1di  PLS1de  PLS2di
                          -RR1di  -RR1de  -RR2di
               median : -0.006   0.007   0.006
               mean   : -0.006   0.012   0.007
               sterr  :  0.001   0.002   0.001
                    lo10  PLS1di  PLS1de  PLS2di
                          -RR1di  -RR1de  -RR2di
               median : -0.007   0.008   0.007
               mean   : -0.007   0.012   0.009
               sterr  :  0.001   0.002   0.001
```

A.3. *sPRESS for sample size n = 40, error factor ε = 0.3, nReal = 100 realisations, nPart = 20 partitions, k = {1, 2, 5, 10}*

```
Basic results

PLS lo 1    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.223  0.258  0.279 -0.043   0.015  -0.053   3.000   3.073
mean   :  0.232  0.276  0.290 -0.044   0.015  -0.059   3.550   3.562
sterr  :  0.008  0.010  0.010  0.010   0.011   0.003   0.177   0.157
PLS lo 2    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.232  0.261  0.292 -0.039   0.016  -0.054   3.000   3.143
mean   :  0.238  0.278  0.298 -0.040   0.021  -0.060   3.506   3.504
sterr  :  0.008  0.009  0.010  0.010   0.011   0.003   0.163   0.134
PLS lo 5    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.254  0.265  0.317 -0.021   0.041  -0.058   3.000   3.000
mean   :  0.262  0.284  0.332 -0.022   0.048  -0.070   3.280   3.107
sterr  :  0.008  0.010  0.010  0.010   0.011   0.004   0.129   0.086
PLS lo10    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.303  0.276  0.412  0.017   0.120  -0.100   3.000   2.200
mean   :  0.313  0.298  0.425  0.015   0.127  -0.112   2.841   2.224
sterr  :  0.008  0.011  0.011  0.010   0.011   0.004   0.083   0.041
 RR lo 1    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.239  0.254  0.286 -0.015   0.023  -0.043   0.200   0.176
mean   :  0.240  0.271  0.290 -0.031   0.019  -0.051   0.205   0.202
sterr  :  0.008  0.010  0.011  0.012   0.014   0.004   0.017   0.016
 RR lo 2    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.244  0.255  0.288 -0.012   0.033  -0.038   0.200   0.172
mean   :  0.248  0.268  0.292 -0.021   0.023  -0.044   0.208   0.195
sterr  :  0.008  0.009  0.010  0.011   0.013   0.002   0.017   0.015
 RR lo 5    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.268  0.255  0.303  0.013   0.048  -0.031   0.200   0.150
mean   :  0.274  0.265  0.311  0.009   0.046  -0.037   0.195   0.168
sterr  :  0.008  0.008  0.009  0.011   0.012   0.002   0.015   0.012
 RR lo10    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.318  0.254  0.351  0.060   0.092  -0.027   0.100   0.102
mean   :  0.327  0.266  0.362  0.061   0.097  -0.035   0.163   0.128
sterr  :  0.009  0.009  0.010  0.010   0.011   0.002   0.012   0.009

Comparison of PLS and ridge regression

                    lo 1  PLS1di  PLS1de  PLS2di
                          -RR1di  -RR1de  -RR2di
               median : -0.011   0.004  -0.001
               mean   : -0.008   0.005  -0.000
               sterr  :  0.003   0.007   0.005
                    lo 2  PLS1di  PLS1de  PLS2di
                          -RR1di  -RR1de  -RR2di
               median : -0.013   0.004   0.000
               mean   : -0.010   0.009   0.007
               sterr  :  0.002   0.006   0.005
                    lo 5  PLS1di  PLS1de  PLS2di
                          -RR1di  -RR1de  -RR2di
               median : -0.013   0.009   0.015
               mean   : -0.013   0.019   0.021
               sterr  :  0.002   0.006   0.004
                    lo10  PLS1di  PLS1de  PLS2di
                          -RR1di  -RR1de  -RR2di
               median : -0.011   0.017   0.059
               mean   : -0.014   0.032   0.063
               sterr  :  0.001   0.008   0.005
```

A.4. *sPRESS for sample size $n = 30$, error factor $\varepsilon = 0.3$, nReal = 100 realisations, nPart = 20 partitions, $k = \{1, 2, 5, 10\}$*

```
Basic results

PLS lo 1   1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.325  0.331  0.397 -0.023   0.068  -0.081   3.000   2.935
mean   :  0.334  0.400  0.430 -0.066   0.030  -0.096   3.270   3.206
sterr  :  0.013  0.024  0.017  0.020   0.022   0.006   0.169   0.138
PLS lo 2   1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.336  0.332  0.419 -0.018   0.072  -0.086   3.000   2.750
mean   :  0.350  0.404  0.450 -0.054   0.046  -0.100   3.095   2.904
sterr  :  0.012  0.023  0.017  0.020   0.022   0.006   0.138   0.101
PLS lo 5   1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.375  0.349  0.491  0.004   0.103  -0.103   2.000   2.143
mean   :  0.398  0.426  0.512 -0.028   0.085  -0.114   2.614   2.203
sterr  :  0.013  0.023  0.015  0.020   0.021   0.005   0.088   0.053
PLS lo10   1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.478  0.378  0.631  0.055   0.191  -0.131   2.000   1.500
mean   :  0.497  0.470  0.640  0.027   0.170  -0.143   2.017   1.498
sterr  :  0.013  0.025  0.014  0.021   0.023   0.006   0.052   0.025
 RR lo 1   1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.345  0.325  0.382  0.003   0.055  -0.040   0.100   0.092
mean   :  0.349  0.369  0.398 -0.019   0.030  -0.049   0.180   0.179
sterr  :  0.013  0.019  0.016  0.018   0.019   0.004   0.025   0.023
 RR lo 2   1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.357  0.326  0.396  0.023   0.062  -0.038   0.100   0.101
mean   :  0.366  0.369  0.413 -0.003   0.044  -0.047   0.178   0.172
sterr  :  0.013  0.019  0.016  0.018   0.019   0.004   0.024   0.021
 RR lo 5   1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.396  0.325  0.435  0.065   0.102  -0.031   0.050   0.107
mean   :  0.415  0.367  0.457  0.047   0.090  -0.042   0.161   0.160
sterr  :  0.013  0.018  0.015  0.018   0.019   0.003   0.020   0.015
 RR lo10   1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.495  0.328  0.540  0.150   0.189  -0.025   0.050   0.163
mean   :  0.511  0.371  0.552  0.140   0.181  -0.041   0.135   0.201
sterr  :  0.013  0.018  0.014  0.018   0.018   0.002   0.016   0.013

Comparison of PLS and ridge regression

                     lo 1  PLS1di PLS1de PLS2di
                           -RR1di -RR1de -RR2di
                 median : -0.016  0.013  0.030
                 mean   : -0.015  0.031  0.031
                 sterr  :  0.003  0.008  0.007
                     lo 2  PLS1di PLS1de PLS2di
                           -RR1di -RR1de -RR2di
                 median : -0.013  0.014  0.033
                 mean   : -0.016  0.035  0.037
                 sterr  :  0.002  0.008  0.007
                     lo 5  PLS1di PLS1de PLS2di
                           -RR1di -RR1de -RR2di
                 median : -0.011  0.026  0.053
                 mean   : -0.017  0.059  0.054
                 sterr  :  0.002  0.009  0.006
                     lo10  PLS1di PLS1de PLS2di
                           -RR1di -RR1de -RR2di
                 median : -0.008  0.062  0.089
                 mean   : -0.014  0.099  0.088
                 sterr  :  0.001  0.011  0.006
```

A.5. *sPRESS for sample size $n = 20$, error factor $\varepsilon = 0.3$, nReal = 1000 realisations, nPart = 20 partitions, $k = \{1, 2, 5\}$*

```
Basic results

PLS lo 1   1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.473  0.562  0.583 -0.097   0.004  -0.103   2.000   2.095
mean   :  0.506  0.670  0.622 -0.164  -0.049  -0.116   2.258   2.157
sterr  :  0.006  0.012  0.008  0.010   0.011   0.002   0.035   0.027
PLS lo 2   1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.498  0.576  0.619 -0.084   0.024  -0.107   2.000   1.909
mean   :  0.532  0.680  0.651 -0.147  -0.028  -0.119   2.092   1.893
sterr  :  0.006  0.012  0.007  0.010   0.011   0.002   0.027   0.018
PLS lo 5   1di    1de    2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.586  0.605  0.706 -0.035   0.075  -0.103   2.000   1.400
mean   :  0.621  0.713  0.735 -0.092   0.022  -0.114   1.658   1.483
sterr  :  0.006  0.012  0.006  0.010   0.011   0.002   0.016   0.009
 RR lo 1   1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.496  0.501  0.528 -0.023   0.005  -0.032  <0.001   0.043
mean   :  0.524  0.598  0.575 -0.074  -0.023  -0.050   0.168   0.168
sterr  :  0.006  0.011  0.008  0.010   0.010   0.002   0.009   0.008
 RR lo 2   1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.519  0.502  0.559 -0.004   0.029  -0.032  <0.001   0.077
mean   :  0.550  0.598  0.598 -0.048   0.000  -0.048   0.163   0.174
sterr  :  0.006  0.011  0.007  0.010   0.010   0.001   0.008   0.007
 RR lo 5   1di    1de    2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.605  0.501  0.653  0.074   0.111  -0.032  <0.001   0.200
mean   :  0.635  0.602  0.680  0.034   0.078  -0.044   0.171   0.224
sterr  :  0.006  0.011  0.007  0.010   0.010   0.001   0.008   0.006


Comparison of PLS and ridge regression

                   lo 1  PLS1di  PLS1de  PLS2di
                         -RR1di  -RR1de  -RR2di
               median : -0.007   0.039   0.056
               mean   : -0.019   0.072   0.047
               sterr  :  0.001   0.003   0.003
                   lo 2  PLS1di  PLS1de  PLS2di
                         -RR1di  -RR1de  -RR2di
               median : -0.007   0.049   0.059
               mean   : -0.018   0.082   0.053
               sterr  :  0.001   0.003   0.003
                   lo 5  PLS1di  PLS1de  PLS2di
                         -RR1di  -RR1de  -RR2di
               median : -0.005   0.083   0.057
               mean   : -0.014   0.111   0.056
               sterr  :  0.001   0.003   0.002
```

A.6. *sPRESS for sample size $n = 10$, error factor $\varepsilon = 0.3$, nReal = 10000 realisations, nPart = 1 partitions, $k = 1$*

```
Basic results

PLS lo 1    1di     1de     2di   1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.841  1.031  0.958 -0.193   -0.088  -0.098    1.000   1.364
mean   :  0.884  1.330  0.996 -0.446   -0.335  -0.112    1.675   1.556
sterr  :  0.003  0.011  0.003  0.011    0.011   0.001    0.016   0.006
 RR lo 1    1di     1de     2di   1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.891  0.991  0.958 -0.116   -0.070  -0.056   <0.001   0.136
mean   :  0.920  1.292  0.978 -0.372   -0.314  -0.058    0.279   0.300
sterr  :  0.003  0.011  0.003  0.011    0.011   0.001    0.004   0.003

Comparison of PLS and ridge regression

                    lo 1  PLS1di  PLS1de  PLS2di
                         -RR1di  -RR1de  -RR2di
              median : -0.021   0.024   0.033
              mean   : -0.036   0.039   0.018
              sterr  :  0.001   0.001   0.001
```

A.7. *sPRESS for sample size $n = 40$, error factor $\varepsilon = 0.4$, nReal $= 100$ realisations, nPart $= 20$ partitions, $k = \{1, 2, 5, 10\}$*

```
Basic results

PLS lo 1    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.329  0.373  0.390 -0.061   0.001  -0.062   2.000   2.195
mean   :  0.346  0.417  0.418 -0.071   0.002  -0.072   2.600   2.597
sterr  :  0.011  0.015  0.014  0.014   0.015   0.005   0.156   0.132
PLS lo 2    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.337  0.376  0.407 -0.051   0.021  -0.065   2.000   2.286
mean   :  0.353  0.415  0.429 -0.062   0.014  -0.076   2.540   2.539
sterr  :  0.011  0.014  0.014  0.013   0.015   0.005   0.132   0.108
PLS lo 5    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.358  0.380  0.444 -0.028   0.047  -0.072   2.000   2.222
mean   :  0.377  0.420  0.462 -0.043   0.042  -0.085   2.358   2.306
sterr  :  0.011  0.014  0.014  0.013   0.015   0.004   0.097   0.072
PLS lo10    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.404  0.389  0.514  0.002   0.102  -0.089   2.000   1.800
mean   :  0.425  0.433  0.529 -0.008   0.096  -0.104   2.101   1.787
sterr  :  0.012  0.015  0.014  0.014   0.015   0.005   0.066   0.038
 RR lo 1    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.335  0.370  0.406 -0.026   0.022  -0.057   0.400   0.348
mean   :  0.361  0.401  0.426 -0.040   0.025  -0.066   0.414   0.401
sterr  :  0.012  0.012  0.014  0.015   0.017   0.004   0.030   0.028
 RR lo 2    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.349  0.364  0.408 -0.022   0.031  -0.049   0.400   0.326
mean   :  0.370  0.395  0.429 -0.026   0.033  -0.059   0.409   0.387
sterr  :  0.012  0.011  0.013  0.014   0.015   0.003   0.029   0.026
 RR lo 5    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.377  0.360  0.428  0.004   0.053  -0.044   0.300   0.291
mean   :  0.395  0.388  0.447  0.007   0.059  -0.053   0.389   0.346
sterr  :  0.011  0.011  0.013  0.013   0.014   0.002   0.026   0.021
 RR lo10    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.424  0.360  0.476  0.051   0.103  -0.043   0.300   0.260
mean   :  0.443  0.385  0.496  0.058   0.112  -0.053   0.349   0.287
sterr  :  0.012  0.011  0.013  0.013   0.014   0.002   0.022   0.016

Comparison of PLS and ridge regression

                    lo 1  PLS1di PLS1de PLS2di
                          -RR1di -RR1de -RR2di
               median : -0.016  0.003 -0.009
               mean   : -0.015  0.016 -0.008
               sterr  :  0.003  0.011  0.007
                    lo 2  PLS1di PLS1de PLS2di
                          -RR1di -RR1de -RR2di
               median : -0.017  0.008 -0.005
               mean   : -0.016  0.020  0.001
               sterr  :  0.003  0.009  0.005
                    lo 5  PLS1di PLS1de PLS2di
                          -RR1di -RR1de -RR2di
               median : -0.017  0.012  0.011
               mean   : -0.018  0.032  0.015
               sterr  :  0.002  0.008  0.005
                    lo10  PLS1di PLS1de PLS2di
                          -RR1di -RR1de -RR2di
               median : -0.016  0.030  0.029
               mean   : -0.018  0.049  0.033
               sterr  :  0.002  0.008  0.006
```

A.8. *sPRESS for sample size $n = 40$, error factor $\varepsilon = 0.6$, nReal = 100 realisations, nPart = 20 partitions, $k = \{1, 2, 5, 10\}$*

```
Basic results

PLS lo 1    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.471  0.668  0.583 -0.160  -0.089  -0.073   1.000   1.366
mean   :  0.482  0.676  0.568 -0.194  -0.108  -0.086   2.080   2.060
sterr  :  0.014  0.013  0.016  0.016   0.017   0.007   0.181   0.151
PLS lo 2    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.484  0.666  0.582 -0.159  -0.087  -0.080   1.000   1.524
mean   :  0.488  0.671  0.578 -0.183  -0.093  -0.090   2.004   1.984
sterr  :  0.014  0.013  0.015  0.015   0.016   0.007   0.152   0.121
PLS lo 5    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.506  0.664  0.594 -0.148  -0.078  -0.079   1.000   1.556
mean   :  0.510  0.670  0.597 -0.160  -0.073  -0.087   1.819   1.769
sterr  :  0.013  0.014  0.014  0.015   0.015   0.005   0.107   0.072
PLS lo10    1di    1de    2di  1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.541  0.664  0.630 -0.117  -0.038  -0.070   1.000   1.400
mean   :  0.552  0.675  0.636 -0.123  -0.039  -0.084   1.575   1.431
sterr  :  0.013  0.015  0.013  0.015   0.015   0.005   0.062   0.033
 RR lo 1    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.493  0.630  0.583 -0.139  -0.049  -0.082   0.600   0.644
mean   :  0.495  0.668  0.587 -0.173  -0.081  -0.092   0.587   0.585
sterr  :  0.015  0.018  0.018  0.023   0.024   0.007   0.036   0.032
 RR lo 2    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.500  0.622  0.582 -0.125  -0.050  -0.067   0.600   0.612
mean   :  0.503  0.654  0.581 -0.151  -0.073  -0.079   0.598   0.574
sterr  :  0.014  0.015  0.016  0.020   0.021   0.005   0.033   0.031
 RR lo 5    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.527  0.617  0.588 -0.103  -0.045  -0.051   0.600   0.567
mean   :  0.528  0.646  0.591 -0.119  -0.056  -0.063   0.594   0.550
sterr  :  0.014  0.014  0.014  0.018   0.018   0.003   0.031   0.026
 RR lo10    1di    1de    2di  1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.564  0.614  0.623 -0.069  -0.010  -0.048   0.600   0.500
mean   :  0.570  0.642  0.630 -0.072  -0.011  -0.060   0.574   0.498
sterr  :  0.013  0.013  0.014  0.016   0.017   0.002   0.028   0.022

Comparison of PLS and ridge regression

                   lo 1  PLS1di PLS1de PLS2di
                         -RR1di -RR1de -RR2di
               median : -0.014  0.018 -0.018
               mean   : -0.013  0.008 -0.019
               sterr  :  0.004  0.015  0.010
                   lo 2  PLS1di PLS1de PLS2di
                         -RR1di -RR1de -RR2di
               median : -0.014  0.017 -0.003
               mean   : -0.015  0.017 -0.004
               sterr  :  0.004  0.011  0.008
                   lo 5  PLS1di PLS1de PLS2di
                         -RR1di -RR1de -RR2di
               median : -0.015  0.017  0.007
               mean   : -0.017  0.024  0.007
               sterr  :  0.003  0.009  0.006
                   lo10  PLS1di PLS1de PLS2di
                         -RR1di -RR1de -RR2di
               median : -0.015  0.019  0.004
               mean   : -0.018  0.033  0.005
               sterr  :  0.002  0.008  0.007
```

A.9. ***A different underlying model, linear combination of PCs 5,7 and 9: sPRESS for sample size $n = 20$, error factor $\varepsilon = 0.3$, nReal = 500 realisations, nPart = 20 partitions, $k = \{1, 2, 5\}$***

```
Basic results

PLS lo 1  1di    1de    2di    1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.517  0.634  0.638  -0.116  -0.003  -0.109   2.000   2.000
mean   :  0.552  0.723  0.674  -0.171  -0.049  -0.122   2.136   2.064
sterr  :  0.010  0.016  0.011   0.014   0.014   0.003   0.046   0.036
PLS lo 2  1di    1de    2di    1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.542  0.645  0.670  -0.099   0.014  -0.110   2.000   1.818
mean   :  0.579  0.731  0.699  -0.152  -0.032  -0.120   1.989   1.818
sterr  :  0.010  0.016  0.010   0.014   0.014   0.003   0.036   0.025
PLS lo 5  1di    1de    2di    1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.628  0.670  0.748  -0.039   0.072  -0.100   1.000   1.400
mean   :  0.664  0.753  0.775  -0.089   0.022  -0.111   1.591   1.432
sterr  :  0.009  0.017  0.009   0.014   0.015   0.002   0.021   0.012
 RR lo 1  1di    1de    2di    1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.539  0.586  0.603  -0.050  -0.005  -0.048  <0.001   0.072
mean   :  0.571  0.662  0.631  -0.091  -0.031  -0.059   0.204   0.208
sterr  :  0.010  0.015  0.012   0.013   0.014   0.003   0.015   0.013
 RR lo 2  1di    1de    2di    1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.565  0.587  0.616  -0.025   0.018  -0.043  <0.001   0.105
mean   :  0.597  0.663  0.652  -0.066  -0.010  -0.055   0.206   0.218
sterr  :  0.010  0.015  0.011   0.013   0.014   0.002   0.014   0.011
 RR lo 5  1di    1de    2di    1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.650  0.590  0.707   0.046   0.091  -0.038  <0.001   0.200
mean   :  0.678  0.666  0.726   0.013   0.061  -0.048   0.217   0.268
sterr  :  0.009  0.015  0.010   0.013   0.013   0.001   0.012   0.010
```

```
Comparison of PLS and ridge regression

                lo 1  PLS1di  PLS1de  PLS2di
                      -RR1di  -RR1de  -RR2di
          median : -0.007   0.032   0.053
          mean   : -0.019   0.061   0.043
          sterr  :  0.001   0.004   0.004
                lo 2  PLS1di  PLS1de  PLS2di
                      -RR1di  -RR1de  -RR2di
          median : -0.007   0.040   0.054
          mean   : -0.018   0.068   0.047
          sterr  :  0.001   0.004   0.004
                lo 5  PLS1di  PLS1de  PLS2di
                      -RR1di  -RR1de  -RR2di
          median : -0.006   0.068   0.051
          mean   : -0.014   0.087   0.048
          sterr  :  0.001   0.004   0.003
```

A.10. ***A different underlying eigenspectrum, linear decay 50, 45, 40, . . . , 10, 5: sPRESS for sample size $n = 20$, error factor $\varepsilon = 0.3$, nReal = 500 realisations, nPart = 20 partitions, $k = \{1, 2, 5\}$***

```
Basic results

PLS lo 1  1di    1de    2di    1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.147  0.177  0.185  -0.034   0.003  -0.031   3.000   2.952
mean   :  0.159  0.206  0.204  -0.047  -0.002  -0.045   3.036   3.141
sterr  :  0.003  0.005  0.004   0.005   0.006   0.002   0.059   0.053
PLS lo 2  1di    1de    2di    1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.151  0.177  0.194  -0.030   0.009  -0.036   3.000   3.091
mean   :  0.164  0.205  0.216  -0.041   0.012  -0.053   3.117   3.277
sterr  :  0.003  0.005  0.004   0.005   0.005   0.002   0.054   0.050
PLS lo 5  1di    1de    2di    1di-1de 2di-1de 1di-2di NFac1di NFac2di
median :  0.171  0.177  0.244  -0.016   0.051  -0.061   3.000   3.000
mean   :  0.186  0.207  0.279  -0.021   0.073  -0.094   3.245   3.164
sterr  :  0.003  0.004  0.005   0.005   0.006   0.003   0.050   0.038
 RR lo 1  1di    1de    2di    1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.158  0.172  0.174  -0.017   0.004  -0.016   0.300   0.300
mean   :  0.167  0.190  0.191  -0.022   0.002  -0.024   0.349   0.342
sterr  :  0.003  0.003  0.004   0.004   0.005   0.001   0.011   0.010
 RR lo 2  1di    1de    2di    1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.161  0.171  0.185  -0.014   0.010  -0.019   0.300   0.282
mean   :  0.173  0.190  0.203  -0.017   0.013  -0.030   0.342   0.324
sterr  :  0.003  0.003  0.004   0.004   0.005   0.001   0.010   0.009
 RR lo 5  1di    1de    2di    1di-1de 2di-1de 1di-2di Shri1di Shri2di
median :  0.182  0.172  0.218   0.005   0.040  -0.027   0.300   0.249
mean   :  0.198  0.189  0.244   0.008   0.055  -0.047   0.313   0.278
sterr  :  0.003  0.003  0.004   0.004   0.005   0.001   0.008   0.006


Comparison of PLS and ridge regression

             lo 1  PLS1di  PLS1de  PLS2di
                   -RR1di  -RR1de  -RR2di
         median : -0.008   0.007   0.007
         mean   : -0.009   0.016   0.013
         sterr  :  0.001   0.003   0.002
             lo 2  PLS1di  PLS1de  PLS2di
                   -RR1di  -RR1de  -RR2di
         median : -0.008   0.006   0.009
         mean   : -0.010   0.014   0.013
         sterr  :  0.001   0.003   0.002
             lo 5  PLS1di  PLS1de  PLS2di
                   -RR1di  -RR1de  -RR2di
         median : -0.010   0.005   0.021
         mean   : -0.012   0.017   0.035
         sterr  :  0.001   0.002   0.002
```

Acknowledgment

We would like to acknowledge financial support and various technical discussions with colleagues within Shell Research Ltd.

References

Altman N. and Leger C. 1997. On the optimality of prediction-based selection criteria and the convergence rate of estimators. Journal of the Royal Statistical Society, Series B 59: 205–216.

Bellman R.E. 1961. Adaptive Control Processes. Princeton NJ, Princeton University Press.

Breiman L. 1996. Heuristics of instability and stabilization in model selection. Annals of Statistics 24: 2350–2382.

Breiman L. and Friedman J.H. 1997. Predicting multivariate responses in multiple linear regression (with discussion). Journal of the Royal Statistical Society, Series B 59: 3–54.

Brown P.J. 1993. Measurement, Regression and Calibration. Oxford, Clarendon Press.

Efron B. 1982. The Jackknife, the Bootstrap and Other Resampling Plans. CBSM38, SIAM, Philadelphia, Penn.

Ganeshanandam S. and Krzanowski W.J. 1989. On selecting variables and assessing their performance in linear discriminant analysis. Australian Journal of Statistics 31: 433–447.

Garthwaite P. 1994. An interpretation of partial least squares. Journal of the American Statistical Association 89: 122–127.

Golub G.H., Heath M., and Wahba G. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21: 215–223.

Hills M. 1966. Allocation rules and their error rates (with discussion). Journal of the Royal Statistical Society, Series B 28: 1–31.

Krzanowski W.J. 1995. Selection of variables, and assessment of their performance, in mixed-variable discriminant analysis. Computational Statistics and Data Analysis 19: 419–431.

Krzanowski W.J., Jonathan P., McCarthy W.V., and Thomas M.R. 1995. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. Applied Statistics 44: 101–115.

Lachenbruch P.A. and Mickey M.R. 1968. Estimation of error rates in discriminant analysis. Technometrics 10: 1–11.

Mertens B., Fearn T., and Thompson M. 1995. The efficient cross-validation of principal components applied to principal component regression. Statistics and Computing 5: 227–235.

Montgomery D.C. and Peck E.A. 1982. Introduction to Linear Regression Analysis. New York, John Wiley & Sons.

Mosteller F. and Tukey J.W. 1968. Data analysis, including statistics. In: G. Lindzey and E. Aronson (Eds.), Handbook of Social Psychology, Vol. 2. Addison-Wesley, Reading, Mass.

Rannar S., Geladi P., Lindgren F., and Wold S. 1995. A PLS kernel algorithm for data sets with many variables and few objects. Journal of Chemometrics 9: 459–470.

Shao J. 1993. Linear model selection by cross-validation. Journal of the American Statistical Association 88: 486–494.

Stone M. 1974. Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society, Series B 36: 111–147.

Wold S. 1978. Cross-validatory estimation of the number of components in factor and principal component models. Technometrics 20: 397–405.