Ensemble Methods and Partial Least Squares Regression[*]

Bjørn-Helge Mevik[†]   Vegard H. Segtnan
Tormod Næs
Matforsk, Osloveien 1, N-1430 Ås, Norway.

May 10, 2005

Abstract

Recently, there has been an increased attention in the literature on the use of ensemble methods in multivariate regression and classification. These methods have been shown to have interesting properties both for regression and classification. In particular, they can improve the accuracy of unstable predictors.

Ensemble methods have so far, been little studied in situations that are common for calibration and prediction in chemistry, i.e., situations with a large number of collinear x-variables and few samples. These situations are often approached by data compression methods such as principal components regression (PCR) or partial least squares regression (PLSR).

The present paper is an investigation of the properties of different types of ensemble methods used with PLSR in situations with highly collinear x-data. Bagging and data augmentation by simulated noise are studied. The focus is on the robustness of the calibrations. Real and simulated data is used.

The results show that ensembles trained on data with added noise can make the PLSR robust against the type of noise added. In particular, the effects of sample temperature variations can be eliminated. Bagging does not seem to give any improvement over PLSR for small and intermediate number of components. It is, however, less sensitive to over-fitting.

Keywords

ensemble methods; bootstrap aggregating (bagging); data augmentation; noise addition; partial least squares regression (PLSR)

1 Introduction

Recently, there has been an increased attention in the literature on the use of so-called ensemble methods in multivariate regression and classification These are methods which by the use of perturbed versions of the original data set, generate a large number of alternative predictors which are combined either by averaging or by majority vote strategies.

The most well known method in this class of techniques is perhaps *bootstrap aggregating* (bagging) [5]. This method generates a large number of perturbed data sets by sampling with

---

replacement from the original data. Each data set is used to form a predictor and finally the predictors are combined as described above. Bagging has been a source of inspiration for development of similar and also more refined techniques.

Another simple way of generating a large number of data sets for producing independent predictors to be averaged is to add various types of random noise to the data. This can be done in many different ways, also for the purpose of mimicking possible future changes of data structure. For instance in spectroscopy where sample preparation effects and instrumental drifts and changes can affect the measurements substantially, such methods can be important. Combining noise augmentation with bagging is also possible.

Ensemble methods have been shown to have interesting properties both for regression and classification. They can lead to increased accuracy of the predictors and also better stability. In particular, ensemble methods have good properties when used in connection with regression and classification trees [3,5,16], but they have also given promising results for other methods [4,5,17,24,26].

The ensemble methods have so far, however, been little studied in situations that are common for calibration and prediction in chemistry. These situations will typically have a very large number of collinear x-variables (several hundred or thousand) and the number of samples is relatively limited (typically between 50 and a few hundred). Such situations are frequently approached by data compression methods such as principal components regression (PCR) or partial least squares regression (PLSR) [20].

Conlin *et al.* [11] investigated some properties of PLSR in combination with adding different types of simulated noise to the data, and concluded that the combination had good properties. There are, however, still many open questions and more research is needed before some more reliable conclusions can be drawn.

The present paper is an investigation of the properties of different types of ensemble methods used in connection with PLS regression. We will consider situations with highly collinear x-data, which are the situations PLS regression was originally developed for. We will study both the effects of bagging and the effects of generating new data sets with various types of noise added. Prediction performance as measured by mean squared error of prediction (MSEP) will be investigated. In particular we will study the effects of ensemble methods on the robustness of PLSR. Real and simulated data will be used. A better understanding of reasons for possible improvements by the use of ensemble methods will also be given attention.

2   Ensemble methods

2.1   Bagging

Bagging was introduced by Breiman [5]. The idea is to generate variants of a base predictor by training it on bootstrap samples from the original data set. When a new observation is to be predicted, it is first predicted with all the trained predictors. The bagged prediction is then the average (for regressions) or majority vote (for classifications) of the individual predictions.

Bagging was developed to reduce the variability of unbiased (or nearly unbiased), but variable, predictors, and this is the kind of predictors it seems best suited for. The inspiring cases for bagging were variable selection and tree-based predictors [6].

Bagging has been very popular in the machine learning literature, and has especially been used for improving unstable classifiers. See e.g. [3, 5, 17, 26]. It has, however, also been successfully applied to regressions such as projection pursuit regression (PPR), multivariate

adaptive regression splines (MARS), local learning based on recursive covering (DART), support vector machines (SVM), regression trees and and linear regression with forward variable selection [2, 4, 5, 24].

2.2 Data augmentation by simulated noise

There are at least two (related) aspects of this approach. The first is the aim of providing a predictor which is more robust than the original one. By adding a small amount of noise to the original data, one hopes to get data which cover the future sample population in a better and broader way. The second aim is to be able to create predictors that are insensitive to various types of drifts or changes in the population of prediction samples. In for instance spectroscopy, temperature changes and sample preparation procedures are known to create this type of structure [27]. Information about what types of effects this causes is often available, at least approximately. One therefore hopes that adding this type of noise to the data will create a predictor which is more robust, at least when used in an ensemble context.

Conlin *et al.* [11] generate perturbed data sets by adding noise to the original training data set. Raviv and Intrator [21] first generate large datasets by sampling with replacement from the original data set, and then add noise to the resampled data sets. In both cases, Gaussian noise is used, and the level of the noise added is chosen by cross-validation. Lee and Cho [18] suggest a method for automatically choosing the level of noise in this procedure.

Along a slightly different line, Despagne *et al.* [12] add noise to the test data in order to estimate the optimal number of components in a PLSR model.

In the present paper, perturbed data sets will be generated by adding various types of noise to the training data set. The idea is to make the predictor more robust and less sensitive to certain types of noise in future prediction data.

2.3 Related methods

Boosting [22], or arcing (*a*daptive *r*esampling and *c*ombin*ing*) [8] works by sequentially applying a classification algorithm to reweighted versions of the training data, and then taking a weighted majority vote of the sequence of classifiers. It has been adapted to regressions (see [2] for a description of several versions).

Stacking regressions [7] works by forming linear combinations of different predictors to give improved prediction ability. Cross-validation and least squares under non-negativity constraints are used to determine the coefficients of the combination.

Wagging [3] (*w*eight *agg*regat*ing*) is a variant of bagging, where the perturbed data sets are generated by giving random weights to the observations of the original sample.

Iterated bagging [10] is intended to reduce bias as well as variance. It works by running bagging repeatedly, using the out-of-bag residuals from each step as the response values for the next.

A random forest as discussed in [9] is an ensemble of trees grown on bootstrap samples. In addition, at each node in the trees, only a fraction of the variables, randomly selected from the complete set of variables, are used to search for the best split. Randomness is thus introduced both into the sample set and variable set of each tree.

3 Why do ensemble methods work?

Most ensemble methods have been designed to reduce the variance of predictors, and have been most successful with unbiased, but variable predictors. There are, however, applications of ensemble methods to reduce bias; see for instance Breiman [10]. This paper will consentrate on the reduction of variance.

It is not yet fully understood why ensemble methods work so well in many cases, but an important aspect of the explanation is that when many predictors are combined (i.e. averaged), their average may sometimes become more stable than each individual predictor separately.

3.1 General considerations

There are several informal arguments why ensemble methods in general work. Let f_i be the individual predictors, for $i = 1, \ldots, N$, and let $\bar{f}$ be their mean. We assume that all f_is are identically distributed, with common variance V, but not necessarily independent. If each f_i is unbiased, $\bar{f}$ will also be unbiased. Raviv and Intrator [21] notes that

$$
\begin{aligned}
\mathrm{Var}(\bar{f}) &= 1/N^2 \sum_{i=1}^{N} \mathrm{Var}(f_i) + 2/N \sum_{i<j} \mathrm{Cov}(f_i, f_j) \\
&= V/N + 2/N \sum_{i<j} \mathrm{Cov}(f_i, f_j).
\end{aligned}
\tag{1}
$$

If all f_is are equal, then $\mathrm{Var}(\bar{f}) = V$, and one gains nothing by averaging. If they are uncorrelated, however, then $\mathrm{Var}(\bar{f}) = V/N$. Averaging can therefore be expected to work well when $\mathrm{Cov}(f_i, f_j)$ are small and V is not too high compared to the variance of the original predictor. There will usually be a trade-off between variance and correlation.

There are other, bagging-specific, arguments. Breiman [5] presents the following argument: Let f_L be a predictor trained on a training data set L. Then

$$
\mathsf{E}_L[\mathrm{MSEP}(f_L)] \geq \mathrm{MSEP}(\mathsf{E}_L\, f_L),
\tag{2}
$$

where the expectations are taken over all possible training data sets of the same size. He argues that when the f_Ls are variable, the inequality can be large. Now a bagged predictor $\bar{f} = 1/N \sum_i f_i$ is an approximation to $\mathsf{E}_L\, f_L$, so if the approximation is close enough, and the inequality large enough, we could expect

$$
\mathrm{MSEP}(\bar{f}) < \mathsf{E}_L[\mathrm{MSEP}(f_L)],
\tag{3}
$$

i.e. the MSEP of the bagged predictor is lower than the expected MSEP of the original predictor.

Friedman & Hall [14] argue that bagging mainly works by substituting the non-linear part of a predictor by an estimate of its expectation, while leaving the linear part unaffected.

Elder [13] demonstrates that the generalised degrees of freedom (GDF) [29] used by a bagged regression tree can be less than the GDF used by the individual regression trees. This supports the observation that bagged predictors are often less prone to over-fitting than their base predictors.
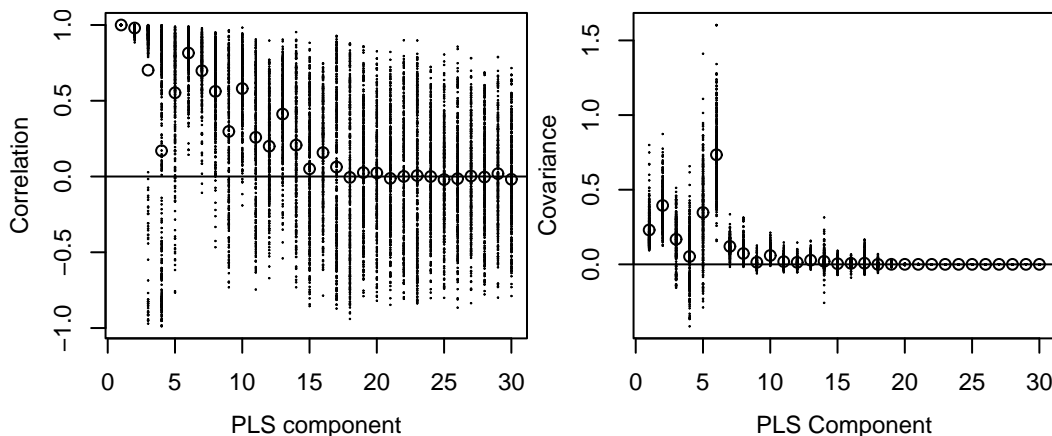
Figure 1: Left panel: correlations between scores; right panel: covariances between predicted responses. The dots show the individual pairwise correlations (covariances) and the circles their mean.

3.2   Data compression methods

In this paper we will focus on prediction based on the PLSR method. This is a method which is based on compressing the information in the x-data down to the most dominating and important dimensions and only use information along these dimensions in the regression. In this way, problems associated with inflation of variance due to the directions with small variability are solved, and stable regressions are obtained [20].

The regression step for the data compression methods PCR and PLSR can be written as

$$y = \sum_{i=1}^{C} q_i t_i + e \qquad (4)$$

where the t_i correspond to the scores along the most dominating dimensions, and C is the number of components. The coefficients are estimated using regular least squares regression. The directions of main variability are estimated with high precision [20]. This means that in an ensemble situation, the most dominating directions will be almost identical for all cases. A consequence of this is that t-values along the most dominating directions are also almost identical for most replicates. The predictions from the different bootstrap or simulated replicates will therefore be highly correlated and the above aggregating advantage less pronounced than for other methods. As the number of components increases, however, the similarity between the t-values in the various replicates will be smaller and smaller. In other words, the correlation is small and the aggregation advantage becomes larger. It is therefore reason to believe that the bagging and noise simulation will have a regularising effect of the components further out.

This is illustrated in Figure 1, which shows results from a bagged PLSR, trained on 50 observations from the data set used in the simulation below, and tested on the remaining 208 observations. The left panel shows pairwise correlations of predicted scores (t_i) for each component. The right panel shows the corresponding covariances of the contributions from each component to the predicted response (i.e., $q_i t_i$). Both the average correlation between

score values and the average covariance between responses decrease with higher components. Also the individual covariances tend to 0. This indicates that the contribution of the higher components to the variance of the predicted response (and thus the MSEP) is low.

4   Simulations

This paper will give main attention to ensemble methods obtained by adding various types of noise to the data and then averaging predictors obtained for the simulated data. These will be called *noise ensemble PLSR*s. A real near-infrared (NIR) data set will be used as a basis for the simulations. The data set consists of measurements of 100 variables for 258 samples in the spectral range 850–1050 nm.

4.1   The different types of noise added

The various types of simulated noise are generated in order to correspond to well-known effects that can influence this type of spectroscopic readings. In particular we will study four different types of noise.

Local shift: For a peak in the spectrum, a random multiple of the 1. derivative of the peak is added to the spectrum of each observation. This corresponds to a local shift with a fix-point. This is a realistic shift simulation for NIR spectra, where the bands are broad and represent a number of different molecular species. A shift in the NIR is normally the result of a shift in the concentration of different states of a molecule, thus leaving an isosbestic point and very often a band with a different shape. This is the case for spectra of water or high moisture samples at different temperatures [23]. Spectral shifts are often simulated by moving a band to higher and lower wavelengths, but such shifts are rarely seen in NIR spectra.

Multiplicative noise: The spectra are multiplied by log-normal noise with mean 1. The same value is multiplied with all variables of an observation. The use of log-normal noise ensures that the multiplicative noise is positive, and on the log scale it corresponds to adding zero mean Gaussian noise. These effects correspond to scatter effects due to for instance sample packing or particle size differences.

Additive noise: Gaussian noise with mean 0 is added to the spectra. The same value is added to all variables of an observation. This effect corresponds to baseline shifts or background differences of the spectra.

Intensity-dependent noise: Independent Gaussian noise with mean 0 is added to each variable of an observation. The size of the noise depends on the signal amplitude x, in that the standard deviation of the noise is proportional to $\exp(0.18x^2 + 0.37x - 5.19)$. This equation was found empirically by taking near infrared spectra of mixtures of carbon black and sulphur (0–100 %), and extracting the noise from the individual spectra representing different absorbance levels This noise thus depends on the intensity of the signal, giving optimal signal-to-noise ratios around 0.5 absorbance units.

Combination: The combination of local shift, additive, multiplicative and intensity-dependent noise. The order of the combination was chosen to follow the photons from the sample to

the detector. The spectral shift is simulated first because this is a phenomenon related to the chemical state of the sample. This is followed by simulation of multiplicative and additive noise, because these phenomena very often represent differences in sample preparation. The intensity-dependent noise is added last because it normally represents detector limitations.

Of course, the larger the noise, the larger the problems will be. In the present paper, the level of the noise added was adjusted such that for an ordinary PLSR trained on unperturbed data, the minimum MSEP when tested on data with noise added should be 1.5 times the minimum MSEP when tested on unperturbed data. More precisely, the level was chosen such that the median of min{MSEP(test data with noise)}/min{MSEP(unperturbed test data)} was 1.5.

While the focus of the present paper is to examine the properties of ensemble methods with PLSR, it should be noted that other techniques have been proposed in the literature to handle similar problems with unwanted variation, such as Extended Multiplicative Signal Correction (EMSC) [19], Orthogonal Signal Correction (OSC) [28] and Direct Orthogonalisation [1]. Also, another type of noise is outliers in training or prediction data, and methods based on for instance least median regression can be robust against such noise.

4.2   Setup of the simulation

The following calibrations were tested: Ordinary PLSR (trained on the original training data), PLSR with multiplicative scatter correction (MSC) [20] of the data, bagged PLSR, and noise ensemble PLSRs for each type of noise described above.

Noise ensemble PLSRs were generated in the following way: 25 perturbed training data sets were created by adding noise of the given type to the original training data. PLSRs were trained on each data set, and then averaged. The ensembles will be denoted $ePLSR_{shift}$, $ePLSR_{int.dep}$, $ePLSR_{add}$, $ePLSR_{mult}$ and $ePLSR_{comb}$. Bagged PLSR was generated by averaging 25 PLSR equations trained on bootstrap samples from the original training data set.

The real NIR data set was split 50 times at random into a training set of size 50 and test set of size 208. Each time the calibrations above were developed for the training data set and tested on the test data set.

In addition to testing all calibrations on unperturbed test data, perturbed test data sets were generated for each type of noise, by adding noise to the original test data set. The noise ensemble PLSRs were tested on test data with 'their' type of noise added. The other calibrations were tested on all perturbed test data.

MSEP was used as criterion for the quality of the predictors. All calculations were performed with R version 1.8.1 [25].

5   Results from the simulation

Figure 2 shows results of the repeated calibrations of ordinary PLSR trained on unperturbed data and tested on data with the various types of noise added. (For clarity, only a random sample of 15 calibrations are shown.) As can be seen, PLSR performs well on unperturbed test data and also reasonably well for an intermediate number of components for the intensity dependent and shift types of noise. For the other two situations, however, PLSR breaks down. Especially for the additive noise, the MSEP curves behave wildly. By design of the simulation, the median of the lowest MSEP value for each curve, is roughly equal for all noise
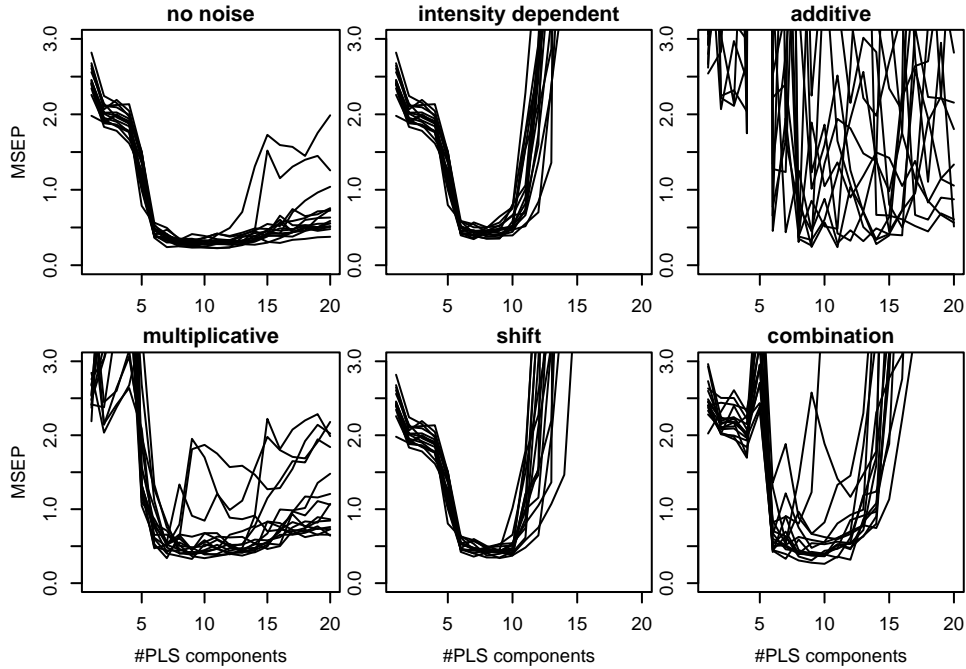
Figure 2: MSEP curves of PLSR trained on original data and tested on perturbed data. Panel titles indicate the type of noise added to the test data.

types. However, the predictions become very unstable for these noise types, leading to a high variability of the MSEP.

Figure 3 shows the corresponding results of noise ensemble PLSRs trained on the various types of noisy data and tested on test data with the corresponding noise. The top left panel, which shows ordinary PLSR tested on unperturbed test data, is the same as the top left panel of Figure 2. One can see that noise ensemble PLSR performs well in all cases. The shift situation, however, shows sign of strong over-fitting for a larger number of components. Also the variability of the MSEP is eliminated for all noise types.

For each type of noise, we tested whether the minimum MSEP for noise ensemble PLSR was lower than the minimum MSEP for ordinary PLSR when tested on the perturbed data. Let $\text{minMSEP}_{\text{NE}}$ be the minimum MSEP (over the number of components in the model) of the noise ensemble PLSR, and similarly $\text{minMSEP}_{\text{Ord}}$ the minimum MSEP of the ordinary PLSR. We tested the null hypothesis that $\text{median}(\text{minMSEP}_{\text{NE}} - \text{minMSEP}_{\text{Ord}}) = 0$ against the alternative $\text{median}(\text{minMSEP}_{\text{NE}} - \text{minMSEP}_{\text{Ord}}) < 0$. Because normality (or even symmetry) could not be assumed, the sign test [15] was used. The resulting p-values are shown in Table 1. As can be seen, for additive, multiplicative and combined noises, the difference was highly significant.

Average results for the different predictors tested on data with various types of noise are given in Figures 4–8.

In Figure 4 are given the results obtained for the *additive noise*. Both the ordinary PLSR, the noise ensemble PLSR trained on data with additive noise and the bagged PLSR give good results when tested on unperturbed data (left panel). The right panel shows that when noise is added to the test data, the only method that keeps its power is the noise ensemble PLSR.
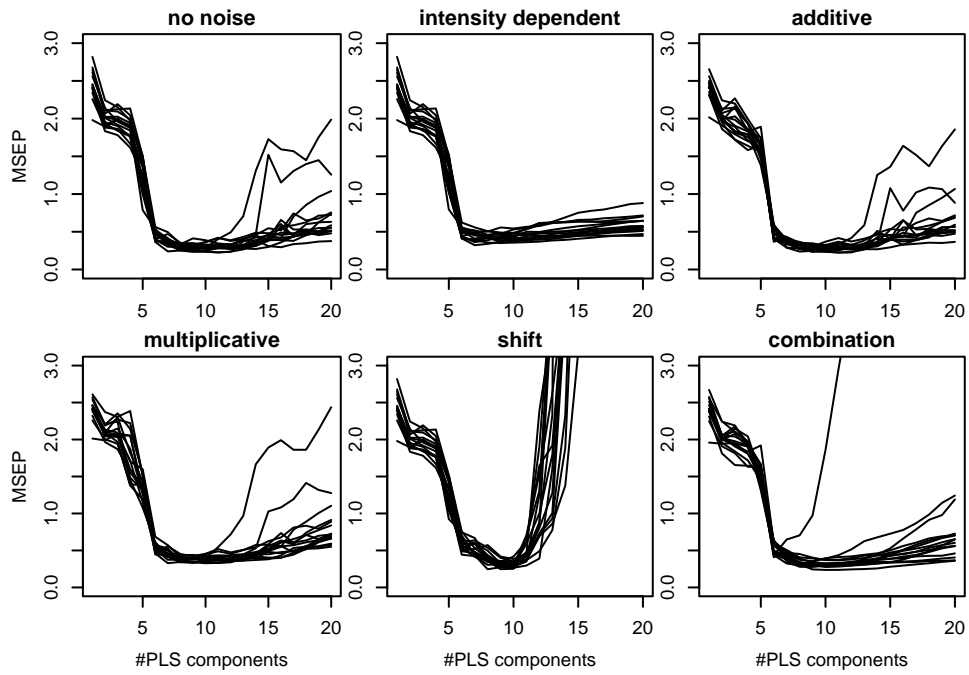
Figure 3: MSEP curves of noise ensemble PLSRs trained on perturbed data and tested on the corresponding perturbed test data. The top left panel shows curves for ordinary PLSR, and is identical to the top left panel of Figure 2. Panel titles indicate the type of noise added to training and test data.

Noise type	p-value
intensity dependent	0.56
additive	$< 1 \times 10^{-15}$
multiplicative	$< 1 \times 10^{-15}$
shift	0.16
combination	0.00015

Table 1: P-values for paired tests for difference of minimum MSEP.
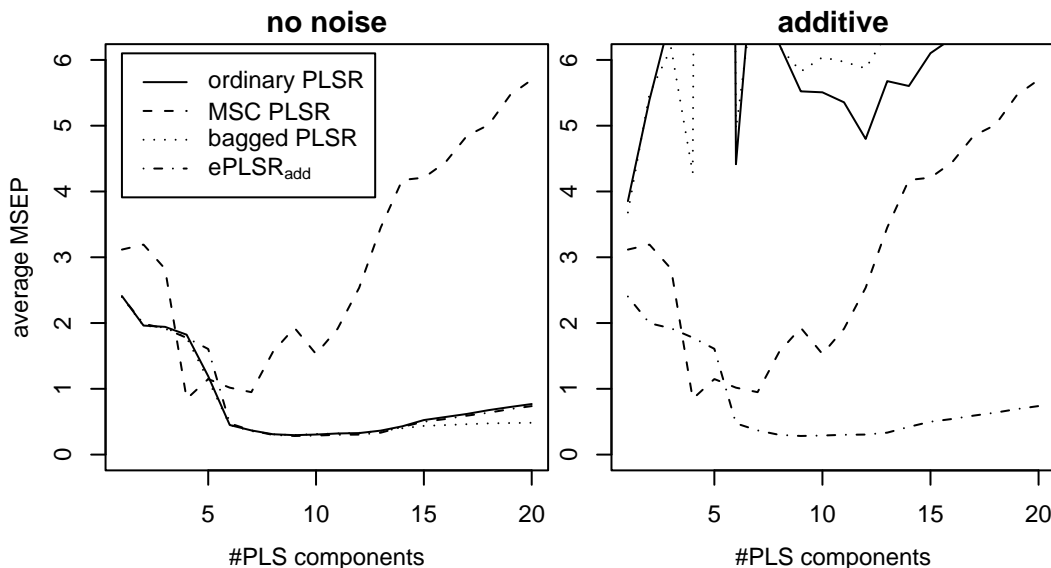
Figure 4: Average MSEP curves for additive noise. Each curve is the average over the 50 replicated calibrations. The left panel shows the results when the calibrations are tested on the unperturbed test data, and the right panel the results for test data with noise added.

It is completely insensitive to the added noise. Both the ordinary and bagged PLSRs fail completely. Even though the level of the noise was chosen so that the median of the minimal MSEP increased with 50 % when ordinary PLSR was tested on data with noise added, the mean MSEP for each model size was much larger for the additive noise. In both cases the scatter corrected data gave poor results, even though it is virtually insensitive to the added noise. This shows that the noise ensemble PLSR works well for this type of noise and is also comparable to ordinary PLSR when there is no added noise in the test set.

The corresponding results for the *intensity dependent noise* are given in Figure 5. Again the noise ensemble method is comparable to ordinary PLSR on unperturbed test data. On test data with noise, it works as good as ordinary and bagged PLSRs for small and moderate model sizes, and is much less sensitive to over-fitting for large model sizes. It is, however, not completely insensitive to the noise. Again MSC gives rather poor results.

The *multiplicative noise* is considered in Figure 6. As for the previous two types of noise, the noise ensemble method works well for the unperturbed test data and keeps most of its performance when the multiplicative noise is added. The ordinary PLSR and bagged PLSR perform reasonable but not as good at noise ensemble PLSR. PLSR obtained on MSC data performs less precise than the ordinary PLSR, but is not sensitive to the noise in the test set. This is expected since the method is designed to handle additive and multiplicative noise. The noise ensemble PLSR is about 50 % less sensitive to the noise than ordinary PLSR for all interesting model sizes.

The data with *random local shift* are considered in Figure 7. Again the noise ensemble method is the best for test data with noise, but differences are smaller. It is insensitive to the noise up to 10 components, but it seems that it is here more sensitive to over-fitting than for the other cases above. Ordinary and bagged PLSRs follow each other, and MSC gives poor
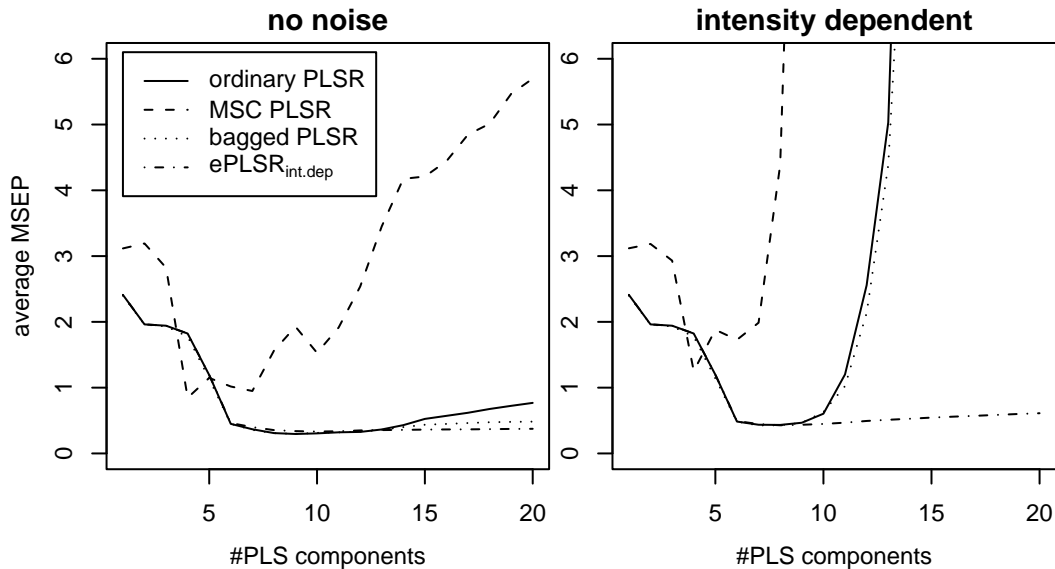
Figure 5: Average MSEP curves for intensity dependent noise. Each curve is the average over the 50 replicated calibrations. The left panel shows the results when the calibrations are tested on the unperturbed test data, and the right panel the results for test data with noise added.
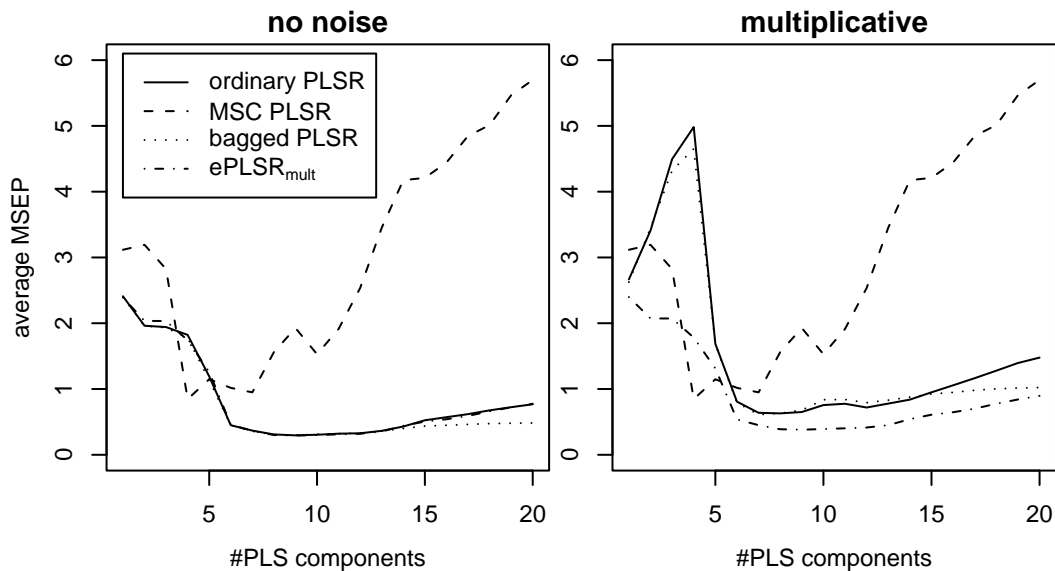

Figure 6: Average MSEP curves for multiplicative noise. Each curve is the average over the 50 replicated calibrations. The left panel shows the results when the calibrations are tested on the unperturbed test data, and the right panel the results for test data with noise added.
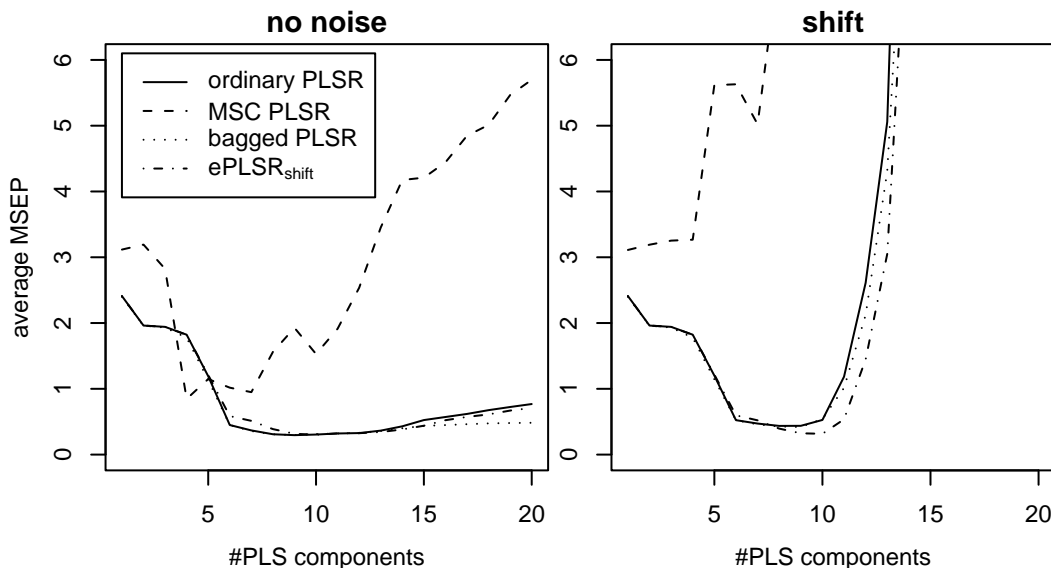
Figure 7: Average MSEP curves for random local shift. Each curve is the average over the 50 replicated calibrations. The left panel shows the results when the calibrations are tested on the unperturbed test data, and the right panel the results for test data with noise added.

results.

Figure 8 shows the results for the *combined noise*. Once again, the noise ensemble PLSR performs at least as well as ordinary or bagged PLSR on unperturbed test data, and like the ensemble PLSR trained on data with intensity dependent noise, it seems less sensitive to over-fitting. It is almost insensitive to noise up to 11 components, and for more components, the sensitivity increases only slowly. The ordinary PLSR, bagged PLSR and MSC perform as for the other noise types.

6 Example: Correcting for Sample Temperature Variation

A reduced $\{3, 9\}$ simplex lattice design containing 37 mixtures of sucrose, fructose and glucose in water was prepared. The design is shown in Figure 9. The carbohydrates were added in concentrations ranging from 0 to 1.67 % (w/w). NIR spectra were taken at 34, 36 and 38 °C for all samples using a NIRSystems 6500 instrument (Foss NIRSystems Inc., Silver Spring, MD, USA) equipped with a 1-mm quartz cell and a temperature control module. For simplicity, prediction models will be presented for sucrose only.

Two calibrations were made: An ordinary PLSR was trained on the observations taken at 36 °C, and a noise ensemble PLSR was generated from 25 perturbed data sets. A difference spectrum was calculated by averaging the difference between the spectra taken at 38 °C and 34 °C of four selected samples (the white points in Figure 9). Each of the perturbed data sets were generated by adding random, normally distributed multipla of the difference spectrum to the observations taken at 36 °C.

The two calibrations were cross-validated, as well as validated using all observations as test data. Care was taken in the 'test set' validation so that no observations from the same
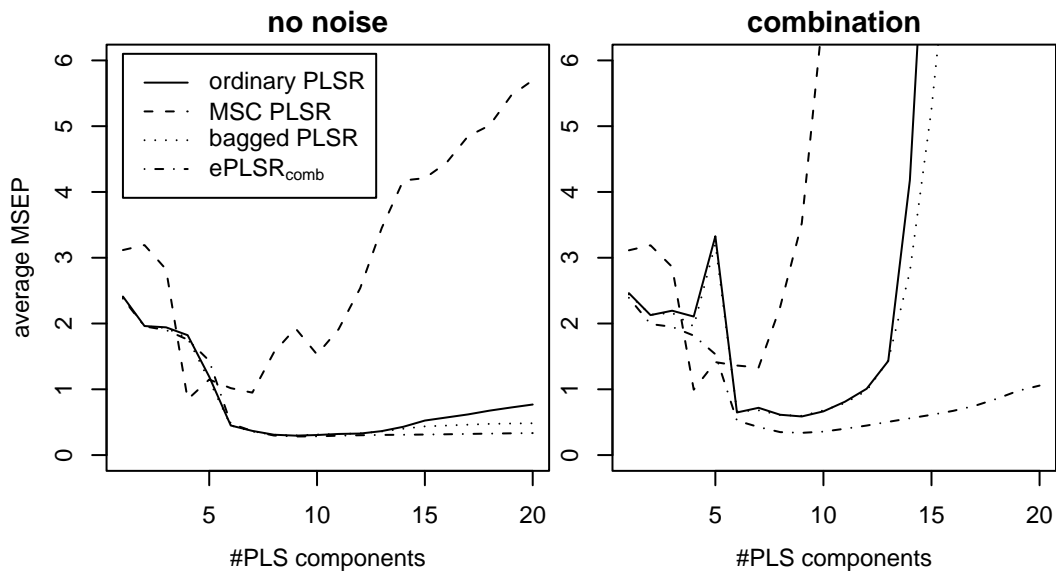
Figure 8: Average MSEP curves for the combined noise. Each curve is the average over the 50 replicated calibrations. The left panel shows the results when the calibrations are tested on the unperturbed test data, and the right panel the results for test data with noise added.
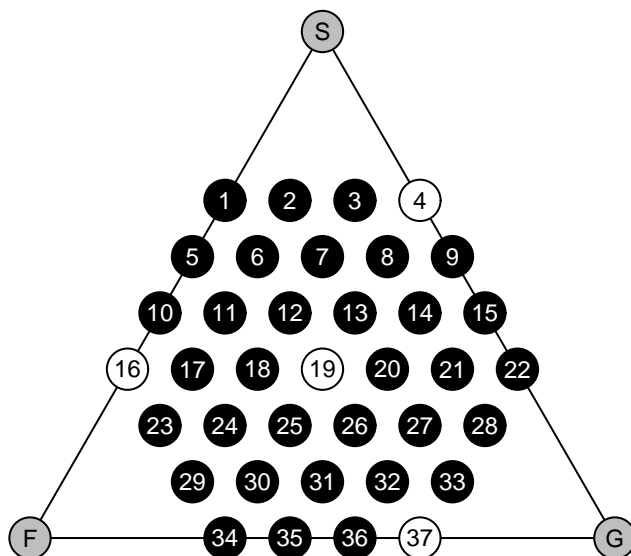

Figure 9: Design of Sucrose, Fructose and Glucose mixtures. The white points indicate the samples used to calculate the difference spectrum.
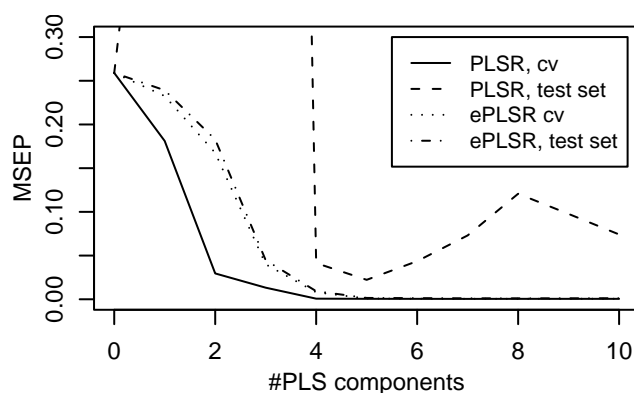
Figure 10: MSEP curves for PLSR and noise ensemble PLSR. 'cv' denotes cross-validated, and 'test set' denotes test set validated.
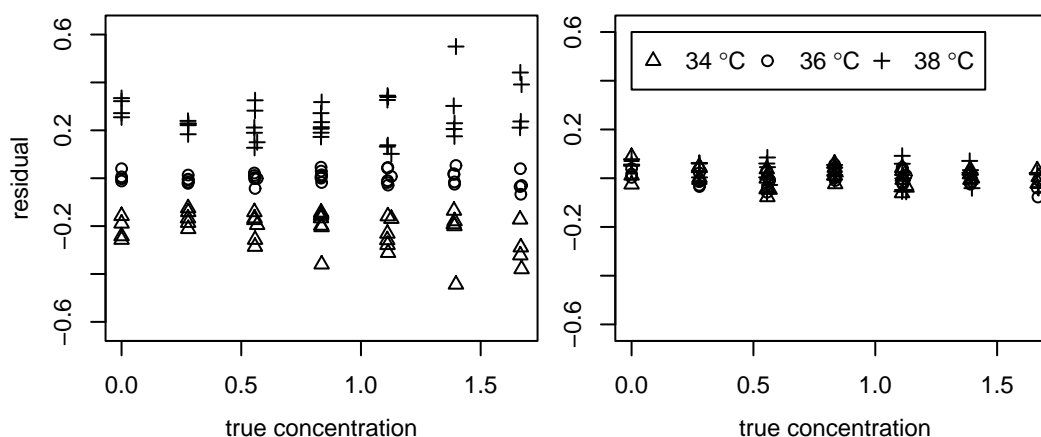

Figure 11: Test set prediction residuals, grouped by sample temperature. Left panel: ordinary PLSR; right panel: noise ensemble PLSR.

physical sample was present in test and training data simultaneously, i.e., the validation was a combination of test set and cross-validation. Also, the four physical samples used to calculate the difference spectrum, were never used as test data for the noise ensemble PLSR. The MSEP curves are shown in Figure 10.

Cross-validation of the ordinary PLSR clearly suggested four components being optimal, with MSEP = 0.000716. When tested on all data, the MSEP was 0.0414, i.e., more than 50 times higher. Cross-validation of the noise ensemble PLSR suggested five components, with an MSEP of 0.000907. When this model was tested on all data, it had an MSEP of 0.00154, which is about twice the cross-validated MSEP of the ordinary PLSR, and only 1.7 times the cross-validated MSEP of the ensemble.

The left panel of Figure 11 shows the residuals from the test set validation of the ordinary PLSR (with four components). All samples measured at 34 °C are systematically predicted too low, and vice versa for 38 °C. The right panel shows the corresponding residuals for the noise ensemble PLSR (with five components). No systematic bias remains.

The noise ensemble PLSR performed almost as well on the training data as the ordinary PLSR, and was almost insensitive to the temperature variations of the samples.

7   Conclusions

Noise ensemble PLSR trained on data with added noise is able to handle most of the five types of noise in the test data significantly better than regular PLSR, and has similar properties to ordinary PLSR when tested on data without added noise. It also had reduced variability when predicting perturbed data, and for some noise types, it reduced greatly the tendency for over-fitting. Noise ensemble PLSR can be used to make PLSR robust against spectral changes such as those induced by sample temperature variations.

Bagging does not seem to give any improvement over PLSR, at least for the small and intermediate number of components. Bagging is not able to handle properly any of the noise types added. It is, however, less sensitive to over-fitting when tested on unperturbed data, as compared to regular PLSR.

Acknowledgements

References

[1] C. A. Andersson. Direct orthogonalization. *Chemometrics and Intelligent Laboratory Systems*, 47(1):51–63, 1999.

[2] Zafer Barutçuoğlu and Ethem Alpaydın. A comparison of model aggregation methods for regression. In *Artificial Neural Networks and Neural Information Processing – Ican/Iconip 2003*, volume 2714 of *Lecture Notes in Computer Science*, pages 76–83. Springer, 2003.

[3] Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning*, 36:105–139, 1999.

[4] Simone Borra and Augostino Di Ciaccio. Improving nonparametric regression methods by bagging and boosting. *Computational Statistics & Data Analysis*, 38(4):407–420, 2002.

[5] Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[6] Leo Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383, 1996.

[7] Leo Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996.

[8] Leo Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–849, 1998.

[9] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[10] Leo Breiman. Using iterated bagging to debias regressions. *Machine Learning*, 45:261–277, 2001.

[11] A. K. Conlin, E. B. Martin, and A. J. Morris. Data augmentation: an alternative approach to the analysis of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, 44(1–2):161–173, 1998.

[12] Frédéric Despagne, Désiré-Luc Massart, and Onno E. de Noord. Optimization of partial-least-squares calibration models by simulation of instrumental perturbations. *Analytical Chemistry*, 69(16):3391–3399, 1997.

[13] John F. Elder IV. The generalization paradox of ensembles. *Journal of Computational and Graphical Statistics*, 12(4):853–864, 2003.

[14] Jerome H. Friedman and Peter Hall. On bagging and nonlinear estimation, 2000. Unpublished manuscript. Available at http://www-stat.stanford.edu/~jhf/.

[15] Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric Statistical Inference*, volume 131 of *Statistics: textbooks and monographs*. Dekker, New York, 3rd edition, 1992.

[16] Tin Kam Ho. Complexity of classification problems and comparative advantages of combined classifiers. *Lecture Notes in Computer Science*, 1857:97–106, 2000.

[17] Hyun-Chul Kim, Shaoning Pang, Hong-Mo Je, Daijin Kim, and Sung Yang Bang. Constructing support vector machine ensemble. *Pattern Recognition*, 36(12):2757–2767, 2003.

[18] Shinjae Lee and Sungzoon Cho. Smoothed bagging with kernel bandwidth selectors. *Neural Processing Letters*, 14(2):157–168, 2001.

[19] H. Martens and E. Stark. Extended multiplicative signal correction and spectral interference subtraction - new preprocessing methods for near-infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, 9(8):625–635, 1991.

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

[21] Yuval Raviv and Nathan Intrator. Bootstrapping with noise: An effective regularization technique. *Connection Science*, 8(3–4):355–372, 1996.

[22] Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[23] Vegard H. Segtnan, S. Sasic, Tomas Isaksson, and Y. Ozaki. Studies on the structure of water using two-dimensional near-infrared correlation spectroscopy and principal component analysis. *Analytical Chemistry*, 73(13):3153–3161, 2001.

[24] Minghu Song, Curt M. Breneman, Jinbo Bi, N. Sukumar, Kristin P. Bennett, Steven Cramer, and Nihal Tugcu. Prediction of protein retention times in anion-exchange chromatography systems using support rector regression. *Journal of Chemical Information and Computer Sciences*, 42(6):1347–1357, 2002.

[25] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2003. http://www.R-project.org/.

[26] Giorgio Valentini, Marco Muselli, and Francesca Ruffino. Cancer recognition with bagged ensembles of support vector machines. *Neurocomputing*, 56:461–466, 2004.

[27] Philip C. Williams and Karl Norris. Variables affecting near-infrared spectroscopic analysis. In Phil Williams and Karl Norris, editors, *Near-Infrared Technology in the Agricultural and Food Industries*, pages 171–185. American Association of Cereal Chemists, Inc., St. Paul, Minnesota, USA, second edition, 2001.

[28] S. Wold, H. Antti, F. Lindgren, and J. Öhman. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 44(1-2):175–185, 1998.

[29] Jianming Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998.