Honouring John S. Shenk

A new method to improve the accuracy of the near infrared models: noise addition partial least squares method

P. Dardenne[a,*] and J.A. Fernández Pierna[a,b]

[a]*Walloon Agricultural Research Centre (CRA-W), Quality of Agricultural Products Department, Chaussée de Namur no. 24, 5030 Gembloux, Belgium*

[b]*Collaborateur Scientifique FNRS Unité de Statistique et Informatique, Faculté Universitaire des Sciences Agronomiques, Avenue de la Faculté 8, B-5030 Gembloux, Belgium*

The accuracy of the y **vectors estimated by near infrared spectroscopic models depends on the quality of the reference method. In this paper the reference data values are augmented with noise. The level of noise added ranged from 0 to 20% of the variability for the mean** y **values. Partial least squares models are then calculated for each addition of noise resulting in many regression coefficient vectors that are used to produce the final calibration model. This final model is selected as the one with the highest R^2 with the median of all these coefficient vectors. This model, applied on unchanged independent test sets, produced a root mean square error of prediction that leads to a clear improvement over the original value without noise. The ability of this technique is investigated on a simulated and on an industrial data set.**

Keywords: NIR spectroscopy, noise addition, PLS, b coefficients

Introduction

Linear regression relates two data matrices, $\mathbf{X}$ $(k \times m)$ and $\mathbf{Y}$ $(k \times n)$, to each other. For a single $\mathbf{y}$ vector, and if $\mathbf{X}$ and $\mathbf{y}$ are mean-centred, the regression model can be expressed by $\mathbf{y}=\mathbf{Xb}+\mathbf{e}$. In this equation, the predictive model is defined by $\mathbf{b}$ $(k \times 1)$, the matrix of regression coefficients and by $\mathbf{e}$ $(\mathrm{k} \times 1)$, the residual vector. The matrix of unknown parameters, $\mathbf{b}$, is solved as $\mathbf{b}=(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{y}$ where the superscript $\mathbf{T}$ denotes the matrix transpose and the superscript $^{-1}$ denotes the matrix inverse.[1,2]

Interpretation of regression coefficients has been described widely in the literature[3–5] and the main conclusion is that a good estimate for these regression coefficients is required because they represent the most important part of the regression model: it must provide both a good fit to $\mathbf{y}$ and good predictions for future unknown samples. However, since the beginning of the technique of near infrared (NIR) spectroscopy it has been known that the accuracy of the regression coefficients, and therefore the accuracy of the analyses (vector $\hat{\mathbf{y}}$) estimated by NIR models, depends on the quality of the reference method. Both the $\mathbf{y}$ and the $\mathbf{X}$ values are subject to random error. Noise in $\mathbf{X}$ and $\mathbf{y}$ will affect the outcome of any prediction model, i.e. the quality of the predictions of $\mathbf{y}$ will be reduced with an increase of noise. Only few studies have addressed the effects of the noise on calibrations and most of them are based on noise addition on $\mathbf{X}$.[6,7] However, NIR can ultimately only be as accurate as the reference methods that provide primary calibration data. The performance of the reference laboratory methods limits the reliability of the NIR calibrations and any increased standard error of prediction (*SEP*) can be due to inaccuracies in NIR or differences in the reference method analytical procedure; in other words, if the variation of the reference method is large, the accuracy of the NIR models expressed as standard error of cross-validation (*SECV*) or *SEP* will be large. As explained by Sørensen,[8] this effect may be negligible when the imprecision of the reference data is lower than the true accuracy of the NIR method. A crucial assumption in multivariate calibration is that these reference values are sufficiently precise and the accuracy of the NIR models is never demonstrated to be better than the reference method.[9]

© NIR Publications 2006, ISSN 0967-0335

However, as discussed by Fernández *et al.*[10] and Faber *et al.,*[11] this assumption is certainly not always true; methods like octane rating or classical Kjeldahl have shown that often the prediction is even better than the reference value.

The robustness of a model can be improved by introducing "stabilisation" spectra into the data set, such as including more samples over time to reflect the changing conditions, like spectra at different temperatures, or by creating artificial variation, for example, two grindings for each sample. Dardenne *et al.*[12] have proved that by introducing these matrix effects it is possible to prepare a NIR calibration with fewer samples and much less wet chemical analysis effort whilst retaining a robust calibration. Other approaches have been proposed to deal with robust models.[13,14] Some use statistical methods to compute whether a sample is contained within the variance already described by some initial samples or not, in such a way the sample can be added to the calibration model allowing the development of a more robust calibration using a low percentage of the samples. In all cases, these methods affect the matrix **X** and lead to more robust models.

This paper aims to prove that one can still have good calibrations with quite poor reference values or with a large quantity of noise added to **y**. For this reason, a method compatible with the classical chemometric techniques and based on an iteration noise addition model has been proposed. This method, called noise addition partial least squares (NAPLS), computes "stabilised" *b* coefficients for the construction of more robust models.

The algorithm

Before applying the method, the data set has to be split into three subsets: a calibration set, a validation set and a test set. The validation set will be used for internal optimisation and the test set is completely independent.

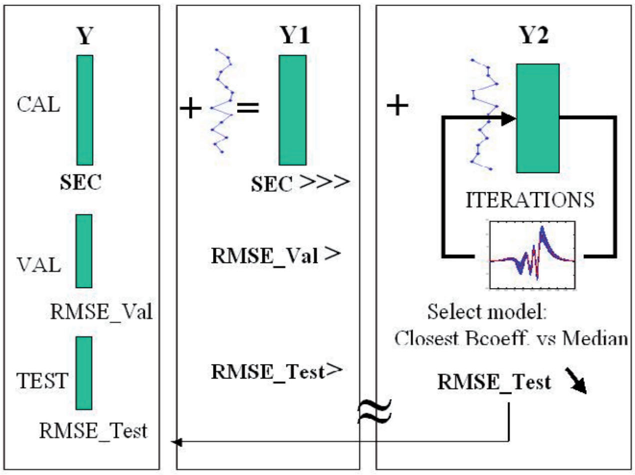

Figure 1. NAPLS method.

Figure 1 shows an overview of the method and for more details the method is described in Annex 1.

The first step is to construct a partial least squares (PLS) model for the calibration data set (CAL) in order to obtain the standard error of calibration (*SEC*). This model is applied in both the validation (VAL) and the test (TEST) set in order to have an estimation of the root mean square error (*RMSE*) for both subsets. Even if the data contains some inherent noise (associated with the reference method), these results could be considered as "noise free" results.

The method proposed in this paper is an iterative method; in a first step, 10% of random noise is added on the **y** matrix of the CAL set and a PLS model is constructed and applied to the VAL and TEST sets. This is done in order to record a first impact of poor reference methods on the CAL and TEST sets. In a second step, a second vector of random noise (10% again) is added onto the **y** value of the already noisy CAL set and a new PLS model is constructed and applied to the VAL set. This second step is repeated 500 times. After each iteration the *b* coefficients of the PLS models are kept in a matrix. Inside this 500 run loop, a test is used to check whether the root mean square error of prediction (*RMSEP*)_Val for a certain PLS model is lower or not than the previous one (i.e. the goal is to find the minimum). When a smaller *RMSEP*_Val is found, the "noisy" **y** values are kept as such, replacing the original ones and the noise added is smaller for the next iteration. In other words, when a better solution is found, a reduction factor is applied to the noise vector to decrease its standard deviation. The median of the 500 vectors of the *b* coefficients is computed and the model which has its *b* coefficients with the highest coefficient of determination (R^2) with the median is selected as the final model. The median has been selected as being representative of the central tendency of the sample set, i.e. it expresses better the common run due to the fact that it is less affected by an excessively high or low figure (and, therefore, outliers) than the mean value. Then, this model is applied to the TEST set in order to obtain the final *RMSE*_Test. The whole procedure can be repeated ten or more times to obtain an average value.

Experimental

Software

The algorithm described in this paper, the calculations and the graphics were executed in Matlab v7.04 (The Mathworks, Inc., Natick, MA, USA). PLS calibrations were derived using the SIMPLS algorithm included in the PLS Toolbox (Eigenvector Research, Inc., Manson, WA, USA).

Simulated data set

A matrix, **X**, was simulated by a weighted sum of four pure components with Gaussian distributions. Figure 2 shows the spectra for the four simulated components where the data points simulating wavelengths (as in a NIR spectrum) are

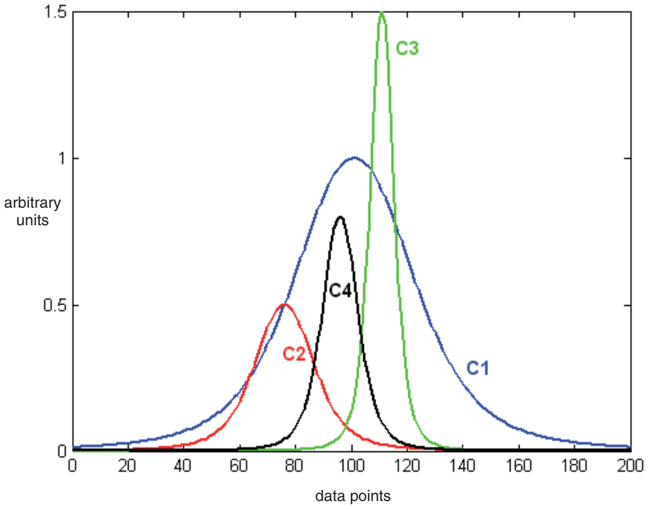

Figure 2. Spectra for four simulated components.

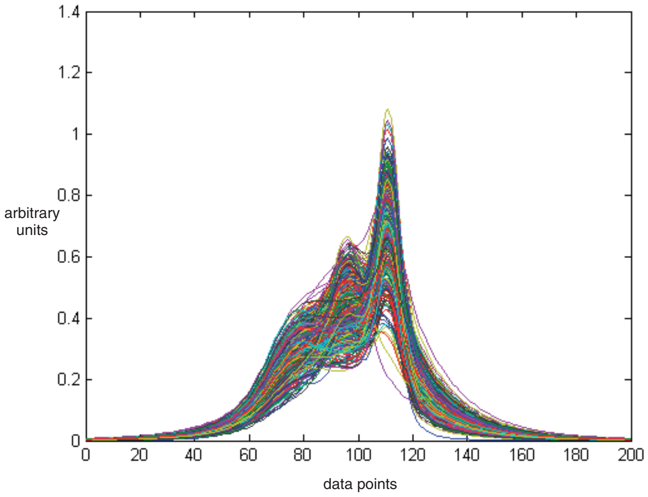

Figure 3. Spectra for four simulated components—the whole data set.

Table 1. Simulated data—range of the concentrations as a percentage of total sample.

	C1	C2	C3	C4
Mean	28.3	28.7	29.2	13.8
STD	8.7	8.8	9.7	12.6
Min.	4.8	3.6	0.0	0.1
Max.	51.8	57.3	55.8	58.6

Table 2. Simulated data—correlation matrix between the four components.

	C1	C2	C3	C4
C1	1.00	—	—	—
C2	–0.19	1.00	—	—
C3	–0.18	–0.16	1.00	—
C4	–0.42	–0.44	–0.53	1.00

plotted against absorptions (here reported as arbitrary units on the y-axis).

These four components are mixed in different proportions to create 375 mixture spectra, as shown in Figure 3.

The proportions between components are such that the sum reaches 100% for all the samples. Table 1 reports the range of the concentrations and Table 2 the correlation matrix between the four components. The reference values were exactly the percentage used to sum the "pure" spectra at each "wavelength".

The data set was randomly split into three subsets: 125 spectra for calibration, 125 as a validation set and 125 as a test set. When calibrating component 1, as expected, a PLS model with four factors expressed 100% of the variation and

Table 3. Simulated data—improvements of *RMSEP* for C1 with NAPLS.

Run	*RMS*cal_noise	*RMS*val	*RMS*test	*RMS*test_final	% difference test
1	2.88	0.61	0.56	0.07	88.28
2	2.45	0.47	0.45	0.02	95.76
3	2.77	0.25	0.28	0.04	85.77
4	2.48	0.64	0.55	0.05	91.12
5	2.73	0.42	0.47	0.04	91.44
6	2.40	0.49	0.43	0.02	94.34
7	2.65	0.82	0.65	0.15	76.64
8	3.14	0.66	0.63	0.04	93.63
9	2.94	0.24	0.26	0.07	71.39
10	2.56	0.47	0.45	0.02	95.53
Ave	2.71	0.53	0.49	0.06	86.80

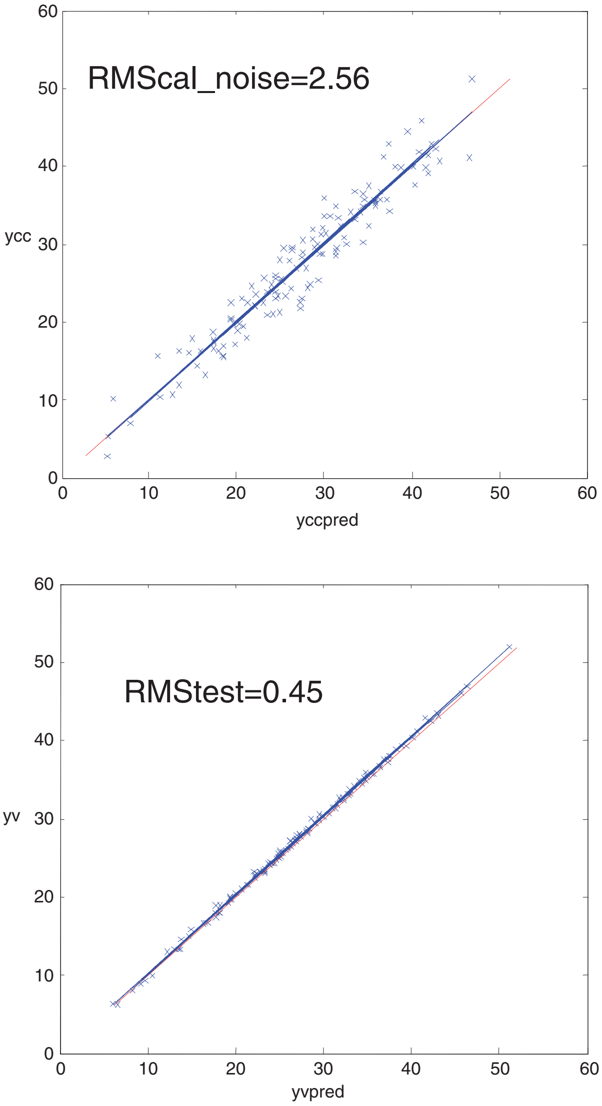

Figure 4. Spectra for four simulated components—noise addiction, run number 10, of (a) calibration and (b) test scatter plots when calibrating component 1 (C1).

SEC and *RMSE* for both validation and test sets are equal to zero. To simulate what happens with an actual reference method, which is never free of errors, a vector of random noise (mean = 0.0 and *SD* = 10% of the average value) was added to the vector, **y** (first column of the **Y** matrix), of the calibration set. The value of the validation and test sets were unchanged. A PLS model was recalculated with four factors.

The three first columns of Table 3 show the *RMSEC* for the calibration data set (*RMS*cal_noise), the validation set (*RMS*val) and the test set (*RMS*test) for 10 runs of noise addition when calibrating component 1 (C1). The averaged *RMS*cal is 2.71, while it reaches 0.53 and 0.49 for the validation and the test set, respectively. Figure 4 shows, respectively, an example (run number 10) of (a) calibration and (b) test scatter plots.

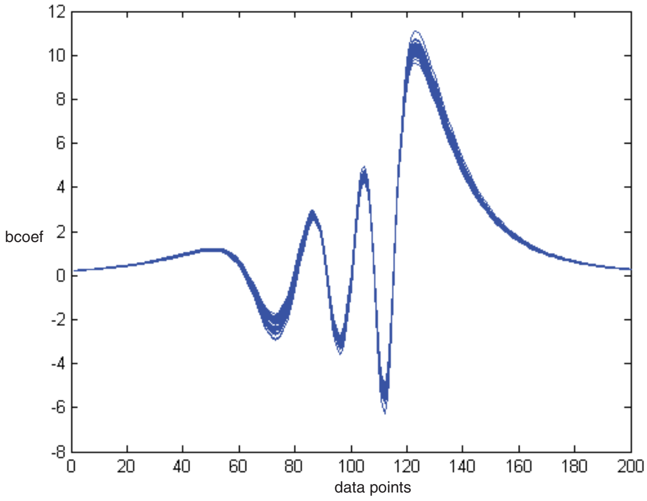

Figure 5. Simulated data: *b* coefficients after 1000 calibrations when calibrating component 1 (C1).

When increasing the level of noise, *RMS*cal changed much more than both the *RMS*val and *RMS*test. When the noise addition is repeated many times (a new random vector is added to the original values), the different PLS models produced a large variation of the *b* coefficients. This is shown in Figure 5 with 1000 calibrations. Many trials were undertaken to find a way to stabilise the coefficients from these simulations.

At the end, the final model is selected as the one with the closest *b* regression coefficients to the median of all these simulations. As shown in Figure 6, this vector of *b* coefficients shows a correlation coefficient with the original model (without any noise) close to 1.

The application of this model to the unchanged independent test set is shown in the fifth column of Table 3. The last column of this table shows, for the test set, the difference (in %) between the results obtained using the selected model and

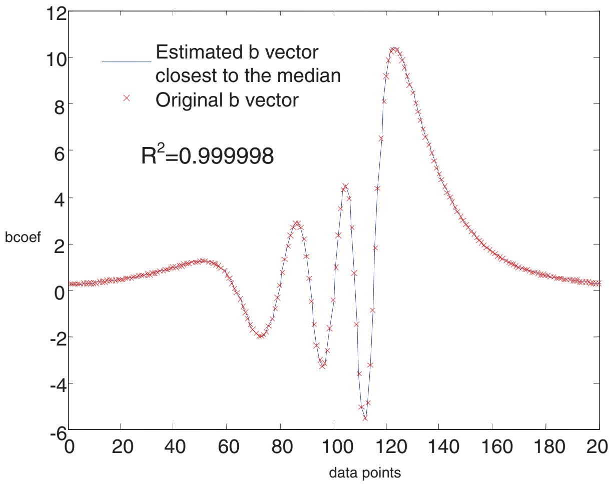

Figure 6. Simulated data: **b** vector close to the median coefficients and original **b** vector when calibrating component 1 (C1).

Table 4. Real forage data set—improvements of *RMSEP* for protein with NAPLS.

Run	*RMS*cal_noise	*RMS*val	*RMS*test	*RMS*test_final	% difference test
1	1.65	0.77	0.86	0.59	31.61
2	1.52	0.68	0.82	0.67	17.39
3	1.41	0.71	0.80	0.62	22.36
4	1.62	0.68	0.86	0.62	28.12
5	1.67	0.91	1.02	0.80	21.41
6	1.49	1.05	0.97	0.68	30.02
7	1.64	0.80	0.80	0.58	26.88
8	1.50	0.77	0.75	0.55	27.47
9	1.68	0.76	0.77	0.60	22.43
10	1.67	0.67	0.64	0.56	12.30
Ave	1.59	0.79	0.84	0.63	24.43

Table 5. Real wheat data set—improvements of *RMSEP* for protein with NAPLS.

Run	$n = 100$ *RMS*cal_noise	$n = 100$ *RMS*val_noise	$n = 100$ *RMS*test_noise	$n = 100$ *RMS*test_final	% difference test
1	1.34	0.50	0.47	0.39	15.51
2	1.28	0.57	0.50	0.42	16.43
3	1.25	0.63	0.53	0.36	32.70
4	1.24	0.69	0.59	0.42	28.71
5	1.27	0.64	0.55	0.41	26.22
6	1.27	0.65	0.55	0.37	33.67
7	1.36	0.59	0.53	0.36	30.94
8	1.23	0.62	0.55	0.37	32.04
9	1.21	0.53	0.47	0.38	19.64
10	1.31	0.57	0.50	0.40	20.15
RMS	1.28	0.60	0.53	0.39	26.05

the original *RMS*test. It is observed that the model produced, on average, a *RMS*test of 0.06 leading to an improvement of 86% over the original *RMS*test.

Real data sets

Forage data set

A set of 328 dried and ground forage spectra, measured using a NIRsystems 5000 in the range of 1100–2498 nm every 2 nm, with known protein content (%DM) values was used. In these data, the **y** values are subject to random error due to the reference method. First, the data set was split randomly into a calibration (110 samples), validation (109 samples) and test set (109 samples). The classical PLS (eight factors and no pre-treatment) model on the raw data led to a *RMS*cal of 0.56 and a *RMS*val of 0.50 and a *RMS*test of 0.60. A 10% noise vector increased the *RMS*cal to 1.59 and the *RMS* errors to 0.79 and 0.84 for the validation and test sets, respectively. The same procedure as above, with a second random noise of 10%, was applied 10 × 500 times and the model with the closest *b* regression coefficients to the median was selected. Applying the selected model to the test set, the *RMS*test decreased to 0.63 which is very close to the original *SEP* of 0.60 and led to an average of 25% improvement as is shown in Table 4.

Wheat data set

This data set contains 2650 samples of wheat measured using a NIRsystems 5000 in the range of 1100–2498 nm every 2 nm to calibrate the protein content. As in the previous data set, the *y* values were subject to random error due to the reference method. The data set was split into three sub-sets where the calibration set contains only 100 samples, the validation set 100 samples and test set contains 2450 samples. After SNV-detrend-first derivative a PLS model with eight factors was constructed and applied to the test set that led to an *RMS*test of 0.50. After noise addition and selection of the final model according to the NAPLS procedure the *RMS*test becomes 0.39, which supposes a reduction and an improvement of more than 26% as shown in Table 5.

Conclusion

It has to be noticed that the final *b* coefficients were not selected based on the minimum of *RMSEP*_VAL but only from the information found in the calibration set. The validation set is used to reduce the level of noise added.

The algorithm could be perfectible and must be optimised regarding the level of noise added and the reduction factor. The effect of the sample size must be tested too. It is obvious that with more samples there will be less variation of the *b* coefficients.

Nevertheless, this simple method is easy to apply and leads to significant improvements. In the real world, the first noise addition must be avoided and the "stabilised" *b* coefficients computed only with one level of noise addition. The *b* coefficients will be much more stable (more independent on the quality of **y**), although for a particular test set, the prediction could be optimum but it is impossible to prove because the reference data always have a certain level of noise.

Currently, we are doing something which leads to similar results by creating a huge data base with thousands of spectra. To cover the *X* variation, a sub-set of well selected samples would be sufficient but, for the *Y* value, the more you average samples the more robust the model.

Acknowledgement

This work was performed within the context of a FNRS collaboration with the Statistics and Informatics Department of the University of Agronomical Sciences in Gembloux, Belgium. The authors thank Professor Claustriaux and Professor Palm from that department for their valuable advice and support. We also would like to thank Dr Vincent Baeten, Dr N.M. Faber and Dr O. de Noord for their valuable comments, discussions and assistance with the manuscript.

References

1. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi and J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics, Part A*. Elsevier, Amsterdam,The Netherlands (1997).
2. H. Martens and T. Naes, *Multivariate calibration.* John Wiley and Sons, Chichester, UK (1989).
3. M.B. Seasholtz and B.R. Kowalski, *Appl. Spectrosc.* **44,** 1337 (1990).
4. A. Lorber, K. Faber and B.R. Kowalski, *Chemometr. Intell. Lab. Syst.* **7,** 1620 (1989).
5. D.M. Haaland and E.V. Thomas, *Anal. Chem.* **60,** 1193 (1988).
6. N.M. Faber and R. Bro, *Chemometr. Intell. Lab. Syst.* **61,** 133 (2002).
7. H.R. Keller, J. Röttele and H. Bartels, *Anal. Chem.* **66,** 937 (1994).
8. L.K. Sørensen, *J. Near Infrared Spectrosc.* **10,** 15 (2002).
9. D.B. Coates, *Spectrosc. Eur.* **14,** 24 (2002).
10. J.A. Fernández Pierna, L. Jin, F. Wahl, N.M. Faber and D.L. Massart, *Chemometr. Intell. Lab. Syst.* **65,** 281 (2003).
11. N.M. Faber and B.R. Kowalski, *Appl. Spectrosc.* **51,** 660 (1997).
12. P. Dardenne, G. Sinnaeve, L. Bollen and R. Biston, *Leaping Ahead with Near Infrared Spectroscopy*, Ed by G.D. Batten, P.C. Flinn, L.A. Welsh and A.B. Blakeney. Royal Australian Chemical Institute, Melbourne Victoria, Australia, p. 154 (1995).
13. P. Vankeerberghen, C. Vandenbosch, J. Smeyers-Verbeke and D.L. Massart. *Chemometr. Intell. Lab. Syst.* **12,** 3 (1991).
14. D.L. Massart, L. Kaufman, P.J. Rousseeuw and A. Leroy. *Anal. Chim. Acta* **187,** 171 (1986).

Received: 9 September 2005
Revised: 5 January 2006
Accepted: 5 January 2006
Web Publication: 19 September 2006

Annex 1

Summary of the noise addition PLS method

1. Split the data ($\mathbf{X}$, $\mathbf{y}$) into CAL ($\mathbf{X}_{cal}$, $\mathbf{y}_{cal}$), VAL ($\mathbf{X}_{val}$, $\mathbf{y}_{val}$) and TEST ($\mathbf{X}_{test}$, $\mathbf{y}_{test}$) sub-sets.
2. Perform a PLS using $\mathbf{X}_{cal}$ and $\mathbf{y}_{cal}$.
3. Predict the validation and test sets by using the PLS model constructed in step 2.
4. Add 10% of random noise to $\mathbf{y}_{cal}$: $\mathbf{y}_c = \mathbf{y}_{cal} + \text{randn}(k,1)*\text{mean}(\mathbf{y}_{cal})*0.1$ with *k* the number of samples in the CAL set.
5. Construct a new PLS model using $\mathbf{X}_{cal}$ and $\mathbf{y}_c$.
6. Predict the validation and test sets by using the PLS model constructed in step 5 in order to obtain *RMSE*_Val and *RMSE*_Test.
7. Add 10% of random noise to $\mathbf{y}_c$ in the same way as step 4 in order to obtain $\mathbf{y}_{cc}$.

8. Construct a new PLS model using $\mathbf{X}_{\mathrm{cal}}$ and $\mathbf{y}_{\mathrm{cc}}$ and predict the VAL set to obtain a new *RMSE*_Val.
9. If the *RMSEP*_Val is lower than the previous one, then repeat steps 7 and 8 by adding 90% of the previous noise.
10. Repeat steps 7 to 9 500 times and keep the *b* coefficients for each of the PLS models.
11. Compute the median of the 500 vectors of the *b* coefficients.
12. Select as best model the one with the most correlated *b* coefficients with the median.
13. Apply this final model on the TEST set in order to obtain the final *RMSE*_Test.
14. Repeat steps 4 to 13 ten times.