Analytica Chimica Acta xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Comparison of various chemometric approaches for large near infrared spectroscopic data of feed and feed products

J.A. Fernández Pierna[a,*], B. Lecler[a], J.P. Conzen[b], A. Niemoeller[b], V. Baeten[a], P. Dardenne[a]

[a] Walloon Agricultural Research Centre (CRA-W), Valorisation of Agricultural Products Department, Food and Feed Quality Unit (U15), Henseval building, Chaussée de Namur 24, 5030 Gembloux, Belgium
[b] BRUKER OPTIK GmbH, NIR & Process Technology, Rudolf-Plank-Str. 27, 76275 Ettlingen, Germany

ARTICLE INFO

Article history:
Received 30 November 2010
Received in revised form 2 February 2011
Accepted 9 March 2011
Available online xxx

Keywords:
NIR
Feed
Chemometrics
PLS
ANN
LS-SVM

ABSTRACT

In the present study, different multivariate regression techniques have been applied to two large near-infrared data sets of feed and feed ingredients in order to fulfil the regulations and laws that exist about the chemical composition of these products. The aim of this paper was to compare the performances of different linear and nonlinear multivariate calibration techniques: PLS, ANN and LS-SVM. The results obtained show that ANN and LS-SVM are very powerful methods for non-linearity but LS-SVM can also perform quite well in the case of linear models. Using LS-SVM an improvement of the RMS for independent test sets of 10% is obtained in average compared to ANN and of 24% compared to PLS.

1. Introduction

In most of the countries that produce important amounts of feed, feed ingredients, fresh silages or soil products, regulations and laws exist about the chemical composition of these products. Normally these regulations lay down some kind of limits (minimum water content, etc.) in the final product in order to guarantee that it meets their legitimate expectations and fulfils the good manufacturing practice. However, manufacturers want to produce at minimal costs and try to arrange their formulations in a way, that the chemical composition of the products is approaching those limiting values. In order to do so, the chemical composition of the raw materials must be known. This requires considerable analytical methods, which are expensive and require the use of chemical reactive. Near infrared spectroscopy (NIRS) is a good alternative to these analytical techniques [1–5]. NIRS is the most widely used non-destructive technology in the feed industry and official control laboratories to determine qualitative parameters of feed ingredients and feeding stuffs. The high throughput of the method, the capacity to determine in one single analysis large panoply of parameters and the possibility to build network of spectrometers made this technique very attractive for the feed sector. The fact that this method can be also used on-line in a feed production plant made this technique even more attractive.

The recent proliferation of fast computers and chemometric algorithms has boosted the use of NIR instruments and it became possible to test everyday foods, feeds and different agricultural products routinely for quality control [6]. Multivariate regression is the process which maps spectra onto the chemical composition of materials by using statistical and mathematical methods, and includes the analysis of data with many observed variables, as well as the study of systems with many important types of variation [7,8]. Multivariate regression establishes a model between component concentrations or properties and the spectral absorbance (log $1/R$) measured for a set of samples at different wavelengths [9]. The performance of multivariate regression is based on several important steps. In a first step we need to acquire as wide a range of samples of one type as possible with an emphasis on covering even extreme samples, which could be samples that are either rare or undesirable because they are poor quality, deteriorated or out of specification. As soon as possible after scanning with the NIR instrumentation the selected samples are submitted for reference analyses. This part should be done with care and precision, ideally with duplicates so that the reference method precision is known and any tests can be repeated if gross errors are encountered. Then, the modelling itself can begin using a number of chemometric algorithms [10,11]. In many spectroscopic applications, multiple linear regression (MLR) [6,7] and partial least squares (PLS) [7,12] are used

[*] Corresponding author. Tel.: +32 81 620352; fax: +32 81 620388.
E-mail address: fernandez@cra.wallonie.be (J.A. Fernández Pierna).

0003-2670/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2011.03.023

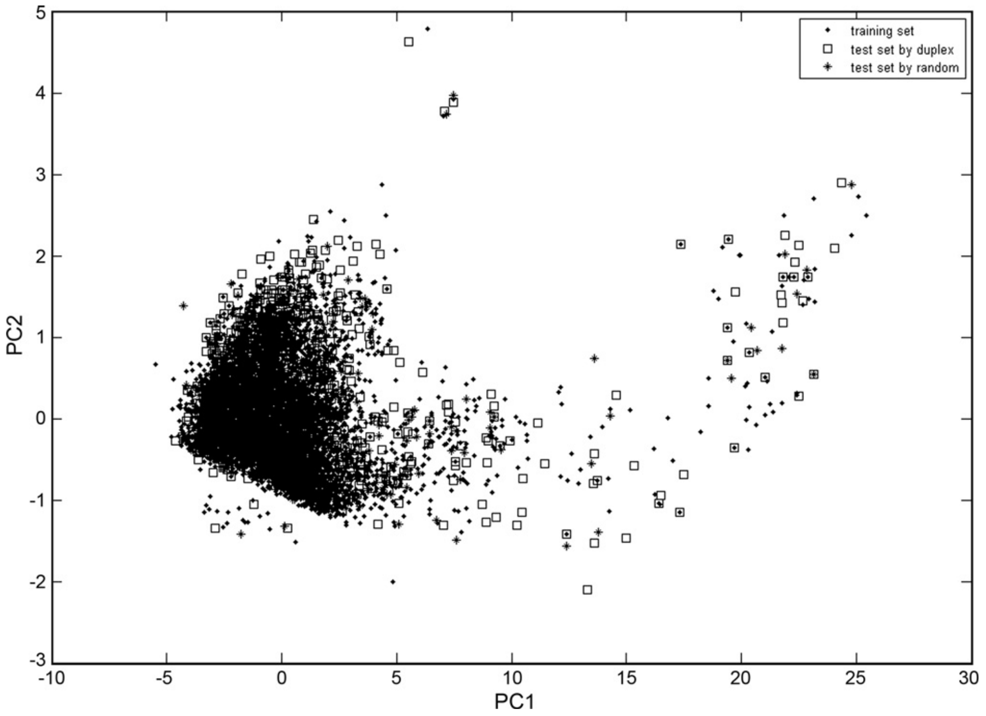


Fig. 1. Example of data distribution (feed data set).

to make regression models because of their simplicity to use, speed and good performances. However, as explained by Thissen et al. [13] nonlinear relations are quite often found which can only be modelled in a limited way by taking into account more latent variables. Then, alternative methods have to be used; techniques as the artificial neural network (ANN) [14,15] or the least-squares support vector machines (LS-SVM) [16–20] have the ability to model nonlinear relationships. LS-SVM, which deals with high dimensional input vectors, has the advantage to ANN that it builds a global model. Previous studies have compared the performance of different multivariate calibration techniques (PCR, PLS and ANN) applied to NIR data in situations of interpolation and extrapolation as well as in the case where the new spectra are affected by instrumental perturbations not accounted for in the calibration set [11,21,22].

The main goal of the present work is to present an effective and complete procedure for multivariate calibration of large NIR data sets from feed samples including diagnostics, feature reduction/selection, modelling and validation of models. This is also done by comparing several supervised multivariate regression methods but on the basis of their performances to handle such large datasets. For this, two large datasets of NIR spectra of feed and feed ingredients with known wet chemical information for several parameters were used. The chemometric methods used in this work were well know techniques as PLS or ANN and the more recent technique LS-SVM.

2. Materials and methods

2.1. Samples

Two NIR spectral datasets have been obtained using a Bruker MPA instrument (Karlsruhe, Germany) working in the range between 10,000 cm^{-1} and 4000 cm^{-1}, having a spectral resolution of 8.0 cm^{-1}. The fist dataset consisted of 26,652 spectra (with 700 wavelengths) of feed ingredients with reference values for ash, fat, fibre and protein. The second dataset contained a total of 28,676 spectra (with 700 wavelengths) of feed with reference values for ash, fat, fibre, starch and protein. The spectral data were pre-processed by the standard normal variate transform followed by detrend [23] and 1st derivative Savitzky–Golay treatment (2nd degree polynomial and a window of 15) in order to remove the scattering effects and to smooth the spectra [24].

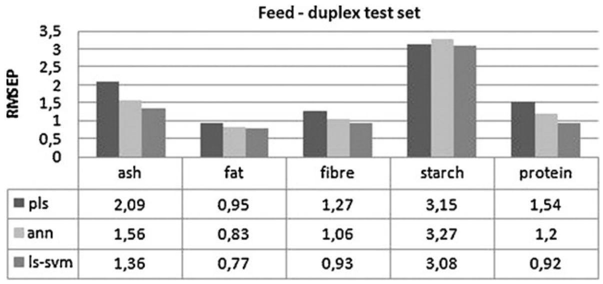
2.2. Brief description of the chemometric methods

2.2.1. Partial least squares (PLS)

PLS is a very well known regression technique [7,12]. The fact that the studied system or process is driven by a small number of latent variables, and that these latent variables are weighted averages of the observed variables, is the fundamental theory of any PLS model. Therefore, PLS is a method for the indirect observation of the latent variables.

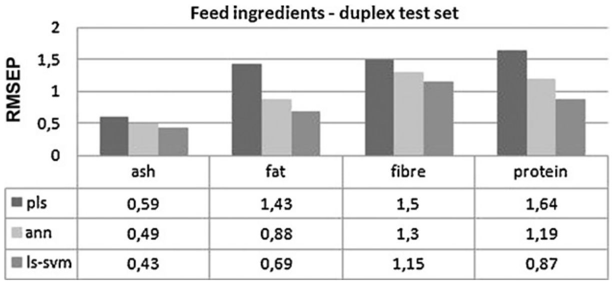
2.2.2. Artificial neural networks (ANN)

Artificial neural networks are designed to mimic functions of the human brain [14,15]. Information processing occurs at many simple elements called neurons. Neurons are connected via synapses (connection links), that modulate signals passing through them. Each synapse has an associated weight w. The net input N is the function of all transmitted signals x_i and their corresponding weights w_i in a neuron: $N = \Sigma(w_i x_i)$ (sum of weighted input signals). Each neuron applies an activation (transfer) function to its net input N in order to provide an output signal for each neuron. A neural network is characterized by its architecture or its pattern of connections between the neurons. Neurons are arranged in several layers: input layer that receives the inputs, hidden layer(s) which transforms the input representation into a new 'hidden' representation and output layer whose units send the predicted values out. Input data are signals x_i of the input layer. Initial weights are random values. Processing the data from input layer to output layer provides the output, which is the result. A neural network is also characterized by its learning algorithm, i.e. its method of determining the weights on the connections. Before using a network for prediction it must be trained with known data. This is necessary to ensure that the net-



	ash	fat	fibre	starch	protein
■ pls	2,09	0,95	1,27	3,15	1,54
■ ann	1,56	0,83	1,06	3,27	1,2
■ ls-svm	1,36	0,77	0,93	3,08	0,92

Fig. 2. Summary of the results for the test set selected by duplex in terms of RMS for the feed data set.



	ash	fat	fibre	protein
■ pls	0,59	1,43	1,5	1,64
■ ann	0,49	0,88	1,3	1,19
■ ls-svm	0,43	0,69	1,15	0,87

Fig. 4. Summary of the results for the test set selected by duplex in terms of RMS for the feed ingredient data.

work provides useful results. The mostly used learning algorithm is based on the 'back-propagation of errors'. While learning, the network compares its output with observed (known) output values of learning data. The effectiveness is usually determined in terms of the root mean square (RMS) error between the actual and the desired outputs averaged over the learning data. After comparison, the network changes weights backwards from output layer to input layer with respect to the output error. A neural networks is characterized also by its activation (transfer) function, which determines its output. In this study a hyperbolic tangent function was chosen as transfer function [25].

2.2.3. *Least squares support vector machines (LS-SVM)*

Support vector machines (SVM) is a powerful methodology for solving problems in nonlinear classification, function estimation and density estimation which has also led to many other recent developments in kernel based methods in general [20].

The LS-SVM method focuses on the least squares version of SVM, whose main advantage is that it is computationally more efficient than the standard SVM method. In this case training requires the solution of a linear equation set instead of the long and computationally hard quadratic programming problem involved by the standard SVM. In the LS-SVM solution, the training samples are mapped to a kernel space, where a hyperplane is fitted on these points. In this case all the training samples are used to achieve a result, which consequently is not sparse as was the case for the classical SVM that select only some of them called the support vectors.

2.3. *Software*

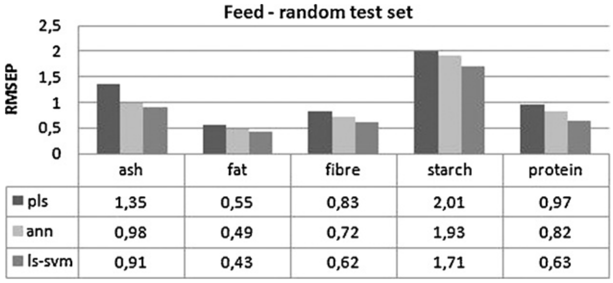
All computations, chemometric analyses, and graphics were carried out with programs developed in Matlab v7.4.0. (The Mathworks, Inc., Natick, MA, USA). For PLS, the PLS toolbox v. 4.11 (Eigenvector Research, Inc.,Wenatchee, WA, USA) was used; for ANN, the Neural Network Toolbox Version 5.0.2 (The Mathworks, Inc., Natick, MA, USA) and for LS-SVM the programs indicated in Ref. [18].

3. Results

A regression model for each property and for each data set is determined, which can then be applied to classify new (unknown) samples. A full analysis including diagnostics, feature reduction/selection, modelling and validation of models has been performed. For the application of all the different regression methods, the data set was summarized by an *nxm* matrix $\boldsymbol{X}$ where n denoted the number of samples and m the number of variables (absorbances at different wavelengths). The references values for such data set were summarized in $\boldsymbol{y}_i$. Such observations constituted the *training set*. The calibration models constructed using the training set were then applied to *two test sets*. These two test sets have been selected in the following way:

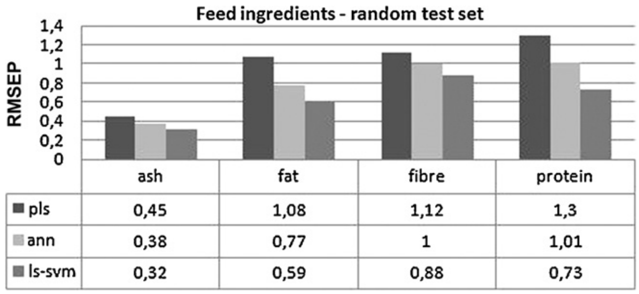
- A first test set was selected by using the duplex design proposed by Snee [26]. This method starts by selecting the two points furthest from each other and puts them both in a first set (training). Then the next two points furthest from each other are put in a second set (test), and the procedure is continued by alternatively placing pairs of points in the first or second set. As a result, 10% of the total number of samples was used as test set. This technique yields a test set including the most diverse samples.
- With the remaining samples, another 10% of the samples were randomly selected as second test set.

In order to have an idea of data distribution after splitting, explorative data analysis, as the principal component analysis (PCA) technique, can be performed [6]. PCA is an unsupervised method that transforms the set of observed variables into a new set of uncorrelated variables, expressed as a linear combination of the observed ones, with lower dimensions. This gives an insight of possible outliers, clusters and other data structures. PCA has been applied to the data after splitting as previously explained in order to check the distribution of the samples in the different subsets. The results are shown in Fig. 1, in this case for the feed data set (PC1 vs. PC2). As expected, it can be seen that the samples selected by the duplex method are distributed in a homogeneous way and covered



	ash	fat	fibre	starch	protein
■ pls	1,35	0,55	0,83	2,01	0,97
■ ann	0,98	0,49	0,72	1,93	0,82
■ ls-svm	0,91	0,43	0,62	1,71	0,63

Fig. 3. Summary of the results for the test set selected by random in terms of RMS for the feed data.



	ash	fat	fibre	protein
■ pls	0,45	1,08	1,12	1,3
■ ann	0,38	0,77	1	1,01
■ ls-svm	0,32	0,59	0,88	0,73

Fig. 5. Summary of the results for the test set selected by random in terms of RMS for the feed ingredient data.

Please cite this article in press as: J.A. Fernández Pierna, et al., Comparison of various chemometric approaches for large near infrared spectroscopic data of feed and feed products, Anal. Chim. Acta (2011), doi:10.1016/j.aca.2011.03.023

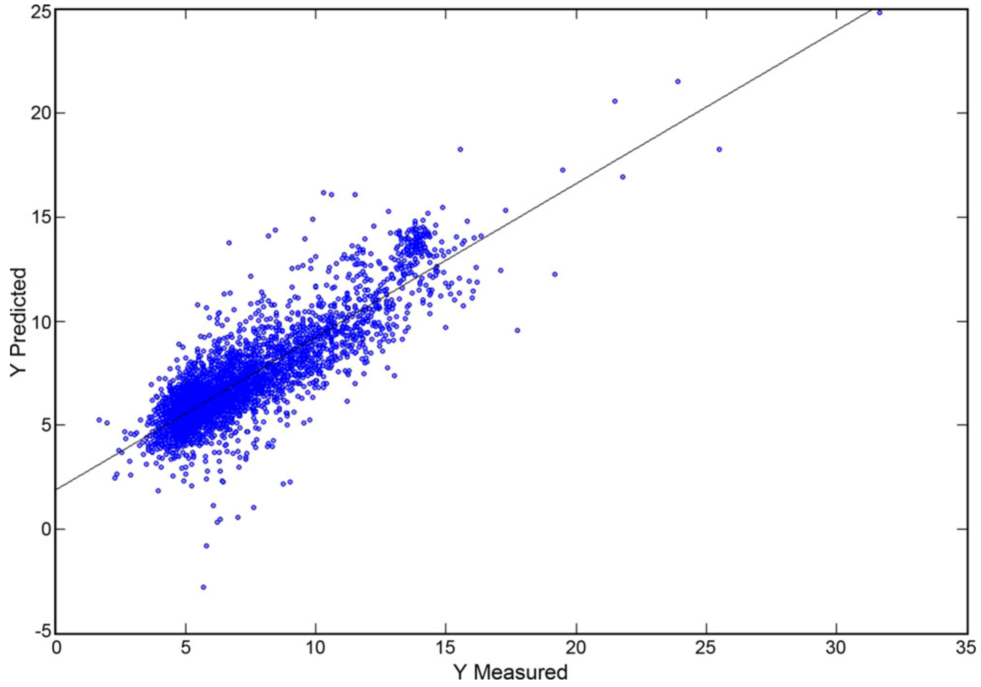


Fig. 6. Y measured vs. Y predicted for feed – ash showing a non-linearity.

the whole range in the PC space. However, the random selection, in this case, contained less extreme samples than the test set selected by duplex.

Regression models have been constructed using PLS, ANN and LS-SVM using the training set (train). For each model, different parameters have to be optimized: the optimal number of latent variables in the case of PLS, the optimal number of inputs and nodes for ANN and the best combination of C and sigma in the case of LS-SVM. C is a regularisation parameter that allows adjusting the trade-off between error minimization, and maximal margin estimation. Sigma corresponds to the width of the Gaussian function, and can be used to adjust the degree of generalization [27].

In order to optimize the different model parameters, cross-validation was used by selecting a subset of samples within the training set, using Venetian blinds (every *n*-th sample together). In all the cases the aim was to obtain the model with the minimal error. In order to have a good indication of the adequacy of the model, the results were presented as RMSEC (root mean squared error for calibration) and RMSEV (root mean squared error for validation) that represents the model and the validation errors, respectively. The model with smallest standard error for the validation data set was assumed to have the best fit to the data and then both independent test sets were predicted using the optimal parameters. RMSEP (root mean squared error for prediction) described the model predictive ability calculated on both independent test sets.

Figs. 2 and 4 show a summary of the results for the test sets selected by duplex expressed in terms of RMS for the feed data and the feed ingredients data set respectively. Figs. 3 and 5 show the summary for the test sets selected by random for the feed data and the feed ingredients data set, respectively. All the models obtained showed reasonable RMS values and considerable high determination coefficients R^2 (not shown) between the predicted and the real values. The best results are those obtained with the LS-SVM technique, with slightly lower RMS values than those obtained with the rest of the regression methods, in some cases the error is almost half of the one obtained using PLS. PLS gives in general poor results, in some cases it can be due to the fact that the relationship between the predicted values and the actual concentration is not linear. This can be checked by visual examination of the method response versus the analyte concentration. This usually works well but it is subjective and open to different interpretations. Then we can also use some statistical methods as for instance the ones based on the study of the correlation of residuals. The residuals are the differences between observed and expected values and they are assumed to be zero. The Run test [28,29] and the Durbin–Watson test [30] look for a correlation (non randomness) of the residuals. In the case of the feed for the determination of the ash concentration, for instance, both tests were applied showing a certain correlation between the predicted and expected values. This is not always the case for the data studied here and in most of the cases, like the ash case (see Fig. 6) this is mainly due to some possible outliers. For this work outlier detection methods have been not applied and all the samples have been used for the study.

4. Conclusion

This work has shown that ANN and LS-SVM are very powerful methods for non-linearity but LS-SVM is also performing quite well in the case of linear models. Models based on PLS failed badly in most of the cases; however ANN and mainly LS-SVM provided good generalization performance. Using LS-SVM an improvement of the RMS for both test sets of 10% is obtained in average compared to ANN and of 24% compared to PLS. Although no general rules could be extracted from the results, it is clear that in the case of large datasets, LS-SVM outperforms classical techniques as PLS or ANN. However, a drawback of the technique remains the selection of the LS-SVM parameters, which is a key point in the training process of these models when applied to regression problems. In general this selection is implemented using grid searches, which grow as the size of the build data set increases, and then the computational cost of the LS-SVM training process also increases considerably. For a small training set the solution of the LS-SVM problem is obtained straightforward, however with huge datasets, the memory space will increase with the level of the number of the

Please cite this article in press as: J.A. Fernández Pierna, et al., Comparison of various chemometric approaches for large near infrared spectroscopic data of feed and feed products, Anal. Chim. Acta (2011), doi:10.1016/j.aca.2011.03.023

training points. There is no upper limit on the number of samples, the only constraints are those imposed by hardware.

References

[1] M. Blanco, I. Villarroya, Trends in Analytical Chemistry 21 (4) (2002) 240–250.
[2] P. Dardenne, V. Baeten, Grasas y Aceites 53 (1) (2002) 45–63.
[3] A.M.C. Davies, NIR News 10 (6) (1999) 14–15.
[4] G. Downey, Journal of Near-infrared Spectroscopy 4 (1996) 47–61.
[5] M. Manley, B.G. Gray, E. Joubert, H. Schulz, in: A. Garrido Varo, A.M.C. Davies (Eds.), Near-infrared Spectroscopy: Proceedings of the 11th International Conference, NIR Publications, Chichester, 2004, pp. 879–882.
[6] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics. Part A, Elsevier, Amsterdam, 1997.
[7] H. Martens, T. Naes, Multivariate Calibration, Wiley, Chichester, 1989.
[8] H. Martens, T. Naes, in: P. Williams, K. Norris (Eds.), Near-Infrared Technology in the Agricultural and Food Industries, 2nd ed., American Association of Cereal Chemists, MN, USA, 2001.
[9] O. Berntsson, L.-G. Danielsson, M.O. Johansson, S. Folestad, Analytica Chimica Acta 419 (2000) 45–54.
[10] P. Dardenne, G. Sinnaeve, V. Baeten, Journal of Near Infrared Spectroscopy 8 (4) (2000) 229–237.
[11] V. Centner, J. Verdu-Andres, B. Walczak, D. Jouan-Rimbaud, F. Despagne, L. Pasti, D.L. Massart, O.E. de Noord, Applied Spectroscopy 54 (2000) 608–623.
[12] P. Geladi, B.R. Kowalski, Analytica Chimica Acta 185 (1986) 1–17.
[13] U. Thissen, M. Pepers, B. Üstün, W.J. Melssen, L.M.C. Buyden, Chemometrics and Intelligent Laboratory Systems 73 (2004) 169–179.
[14] T. Naes, K. Kvaal, T. Isaksson, C. Miller, Journal of Near Infrared Spectroscopy 1 (1993) 1–11.
[15] F. Despagne, D.L. Massart, Analyst 123 (1998) 157–178.
[16] J.A. Fernández Pierna, V. Baeten, A. Michotte Renier, R.P. Cogdill, P. Dardenne, Journal of Chemometrics 18 (2004) 341–349.
[17] J.A. Fernández Pierna, P. Volery, R. Besson, V. Baeten, P. Dardenne, Journal of Agricultural and Food and Chemistry 53 (17) (2005) 6581–6585.
[18] J.A. Fernández Pierna, P. Dardenne, Chemometrics and Intelligent Laboratory Systems 91 (2008) 94–98.
[19] R.P. Cogdill, P. Dardenne, Journal of Near Infrared Spectroscopy 12 (2) (2004) 93–100.
[20] J. Valyon, G. Horváth, Proceedings of World Academy of Science, Engineering and Technology 7 (2005), ISSN 1307-6884.
[21] F. Estienne, L. Pasti, V. Centner, B. Walczak, F. Despagne, D. Jouan Rimbaud, O.E. de Noord, D.L. Massart, Chemometrics and Intelligent Laboratory Systems 58 (2) (2001) 195–211.
[22] F. Estienne, F. Despagne, B. Walczak, O.E. de Noord, D.L. Massart, Chemometrics and Intelligent Laboratory Systems 73 (2) (2004) 207–218.
[23] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Applied Spectroscopy 43 (5) (1989) 772–777.
[24] P.A. Gorry, Analytical Chemistry 62 (1990) 570–573.
[25] B.L. Kalman, S.C. Kwasny, Proceeding at the International Joint Conference on Neural Networks, vol. 1, Publisher: IEEE Neural Networks Council, Baltimore, MD, June 1992, pp. 800–805.
[26] R.D. Snee, Technometrics 19 (1977) 415–428.
[27] Z. Lin-Cheng, Y. Hui-Zhong, L. Chun-B, Natural Computation. Fourth International Conference ICNC ‘08, pp. 130–133 (2008).
[28] J.V. Bradley, Distribution-Free Statistical Tests, Prentice-Hall, Englewood Cliffs, New Jersey, 1968.
[29] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35d.htm.
[30] H. Mark, J. Workman, Spectroscopy 20 (9) (2005) 26–35.

Please cite this article in press as: J.A. Fernández Pierna, et al., Comparison of various chemometric approaches for large near infrared spectroscopic data of feed and feed products, Anal. Chim. Acta (2011), doi:10.1016/j.aca.2011.03.023