Available online at www.sciencedirect.com

SCIENCE DIRECT®

Analytica Chimica Acta 516 (2004) 1–5

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Identification of micro-organisms by dint of the electronic nose and trilinear partial least squares regression

Sven Serneels, Maarten Moens, Pierre J. Van Espen*, Frank Blockhuys

Department of Chemistry, University of Antwerp, Antwerp, Belgium

Received 5 March 2004; received in revised form 20 April 2004; accepted 20 April 2004

Available online 5 June 2004

Abstract

Ventilator-associated pneumonia is one of the most lethal infections occurring in intensive care units of hospitals. In order to obtain a faster method of diagnosis, we proposed to apply the electronic nose to cultures of the relevant micro-organisms. This allowed to halve the time of the analysis. In the current paper, we focus on the application of some chemometrical tools which enhance the performance of the method. Trilinear partial least squares (tri-PLS) regression is used to perform calibration and is shown to produce satisfactory predictions. Sample specific prediction intervals are produced for each predicted value, which allows us to eliminate erroneous predictions. The method is applied to an external validation set and it is shown that only a single observation out of 22 is being wrongly classified, so that the method is acceptable for inclusion in the clinical routine.

Keywords: Ventilator-associated pneumonia; Electronic nose; Trilinear partial least squares regression; Tri-PLS; Prediction error

1. Introduction

Pneumonia is one of the most severe diseases which occur as secondary infections in hospitals. Moreover, the risk of infection for patients increases significantly if the latter are subject to insufflation, in which case the disease is referred to as ventilator-associated pneumonia (VAP). In contrast to infections of other organs, for which the mortality rates do only seldomly exceed 4[1]. It should be clear that special care should be taken for both the detection and treatment of ventilator-associated pneumonia in hospitals. Whereas the latter is beyond the scope of the present article, the need for a fast and accurate detection method for several micro-organisms which may cause VAP is the necessity which led to the work presented here.

It has been shown that ventilator-associated pneumonia can be caused by about 15 different micro-organisms [1], all of which can cause a serious inflammation if present in the airways or the deeper lung tissue. Moreover, some of these micro-organisms may also cause serious inflammation of other organs, such as the urinal tracts. A correct treatment of the patient relies on the choice of the antibiotics, which in the first stage of treatment, is purely based upon the analysis of the symptoms. Treatment with wrong antibiotics automatically leads to a significant increase in the probability of mortality. Of course, the correct choice of antibiotics can only be made if the pathogenic micro-organism has been identified. Currently, this identification is carried out as follows: a sample is taken from the patient, it is plated on a nutrient, incubated and finally analysed by classical microbiological testing. The identification process takes about 36 h. During this time, the patient may be treated with the wrong antibiotics.

The advantage of an electronic nose over the classical microbiological technique resides in the fact that the time needed for the analysis can be reduced by about 50%, i.e. the analysis by dint of the electronic nose can be completed in 17 h. Furthermore, the analysis itself only takes about 10 min; the vast majority of those 17 h are needed for the incubation of the micro-organisms.

This application of the electronic nose is based on the observation that each of the micro-organisms involved produces different gaseous excrements; hence these

* Corresponding author. Present address: Departement Scheikunde, Universiteit Antwerpen, Universiteitsplein 1, 2610 Antwerpen, Belgium. Tel.: +32-3-8202358; fax: +32-3-8202376.

E-mail address: piet.vanespen@ua.ac.be (P.J. Van Espen).

0003-2670/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2004.04.027

gaseous excrements can be used as a "fingerprint" of the micro-organism [2]. A well-configured electronic nose reacts differently to each of these "fingerprints" and should hence be able to identify the micro-organism. For application in clinical practice, the misclassification probability should not exceed 5%. In practice, the electronic nose is calibrated by means of a set of calibration samples, for which the corresponding species is known. For each sample, the gaseous excrements are lead over 10 different resistors. For each resistor, the resistance is measured during 60 s. Hence, in the chemometrical sense, the calibration of the electronic nose is a typical "three-way" problem, the calibration data matrix being of dimensions $n \times 60 \times 10$.

Trilinear partial least squares regression, or tri-PLS [3], has recently become one of the most popular techniques to perform calibration for three-way data which are used to predict a dependent variable. Its success is both based on the sound statistical underpinning of the method and on various successful applications [4]. The application of tri-PLS for the calibration of the electronic nose is very appealing: once the calibration stage has been completed, the prediction of the dependent variable for a new sample (this would be a real patient) only consists of a single (matrix) multiplication. Moreover, very recently, Faber and Bro [5] have proposed a method to compute sample specific prediction errors for the predictions made by tri-PLS [5]. The relevance of prediction errors on the predictions made by tri-PLS is obvious: it may lead to the conclusion that the possible presence of (eventually a second) micro-organism can or cannot be excluded. In earlier work [7], a combination of genetic algorithms, neural networks and the k-nearest neighbour algorithm were used to perform calibration and prediction. The emphasis of the current paper is on the applcation of trilinear partial least squares to the electronic nose, as it improves the transparancy and practicability of the method, and above all, allows to compute a sample-specific prediction error.

2. Trilinear partial least squares regression

Before we can proceed with the description of the application of tri-PLS to data measured by the electronic nose, we will first introduce the reader briefly to the methodology of tri-PLS.

Let $\boldsymbol{X}$ and $\boldsymbol{y}$ denote the calibration data matrices. Let $\boldsymbol{X} \in \mathbb{R}^{n\times p\times q}$ and $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$, respectively, where n is the number of samples at hand and $\boldsymbol{y}$ is the vector to be predicted. Furthermore, $X \in \mathbb{R}^{n\times pq}$ denotes the so-called unfolded data matrix, where each of the slabs of $\boldsymbol{X}$ have been aligned next to each other. Matrices will always be denoted be upper-case letters. The columns of a matrix will be denoted by the corresponding lower-case bold-face letter. Three-way matrices will always be denoted by bold-face upper-case letters. Let $\text{vec}(\cdot)$ denote the vectorization operator, which vertically stacks the colums of its argument underneath each other. Hence, if A is a $p\times q$ matrix, $\text{vec}(A)$ will be a $pq \times 1$-vector. Let $\text{vec}^{-1}_{p,q}$ denote the operator which re-shapes a pq-vector into a $p \times q$ matrix, such that $\text{vec}^{-1}_{p,q}(\text{vec}(A)) = A$.

Trilinear partial least squares regression is a natural extension of partial least squares regression [6] to three-way data. Both univariate and multivariate tri-PLS algorithms exist (referred to as tri-PLS1 and tri-PLS2, respectively). In the current article, we will limit ourselves to tri-PLS1 regression.

The success of partial least squares regression for—often multicollinear—two-way data is easily understood if one considers the fact that PLS is a latent variable regression technique, which first summarizes the often high-dimensional data matrix into a small number of uncorrelated latent variables, upon which regression is carried out. The benefits of PLS over other latent variable regression techniques for prediction are mainly caused by the fact that in PLS latent variables are defined according to a maximization criterion of the covariance between X and $\boldsymbol{y}$. Hence, the latent variables summarize the fraction of the total variance which is relevant for the prediction of $\boldsymbol{y}$ which is not the case for other latent variable techniques such as principal component regression. The same maximization criterion is maintained for three-way data in tri-PLS, carefully extending the PLS methodology to three-way data and respecting this three-way structure of the data.

In PLS, the latent variables are computed as a linear combination of the original predictor variables, i.e. $\boldsymbol{t}_i = X\boldsymbol{w}_i$. The $\boldsymbol{w}_i$ are called the weighting vectors and are defined respecting the aforementioned maximization criterion. In tri-PLS, separate weights $\boldsymbol{w}_i^p$ and $\boldsymbol{w}_i^q$ have to be computed for the different dimensions of $\boldsymbol{X}$, which are afterwards combined into overall weights $\boldsymbol{w}_i$. This is seen from the tri-PLS algorithm, which is given by ($i \in 1\ldots k$, $\boldsymbol{e}_0 = \boldsymbol{y}$):

$$Z_i = \text{vec}^{-1}_{p,q}(X^{\mathrm{T}}\boldsymbol{e}_{i-1}) \tag{1a}$$

$$\boldsymbol{w}_i^p, \boldsymbol{w}_i^q = \text{dominant singular vectors of } Z_i \tag{1b}$$

$$\boldsymbol{w}_i = \boldsymbol{w}_i^q \otimes \boldsymbol{w}_i^p \tag{1c}$$

$$\boldsymbol{t}_i = X\boldsymbol{w}_i \tag{1d}$$

$$b_i = (T_i^{\mathrm{T}}T_i)^{-1}T_i^{\mathrm{T}}\boldsymbol{y} \tag{1e}$$

$$e_i = [I_n - T_i(T_i^{\mathrm{T}}T_i)^{-1}T_i^{\mathrm{T}}]\boldsymbol{y} \tag{1f}$$

$$\boldsymbol{\beta}_i = W_i\boldsymbol{b} \tag{1g}$$

In Eq. (1c), ⊗ denotes the Kronecker product.

The algorithm stated above was first introduced by de Jong [8], who reported it to outperform previous versions in terms of computational properties.

In the Section 1, we heeded that prediction for new samples is completed by a single matrix multiplication. Indeed, the predicted response υ for an unfolded new sample $\boldsymbol{\xi}$ is given by:

$$\upsilon = \boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\beta}_i \tag{2}$$

A sample specific estimate of the prediction error for the new sample can now be computed as follows. A "score vector" $\boldsymbol{\tau}$ is defined for each new sample as:

$$\boldsymbol{\tau} = \boldsymbol{\xi} W_i \tag{3}$$

This score vector is used to compute the sample leverage:

$$h = \boldsymbol{\tau}^{\mathrm{T}} (T_i^{\mathrm{T}} T_i) \boldsymbol{\tau} \tag{4}$$

Let now $\hat{\boldsymbol{y}}$ denote the vector of predictions for the samples in the training set, then the mean squared error of calibration is defined as (k denotes the number of latent variables used):

$$\mathrm{MSEC} = \frac{(\hat{\boldsymbol{y}} - \boldsymbol{y})^{\mathrm{T}} (\hat{\boldsymbol{y}} - \boldsymbol{y})}{(n - k)} \tag{5}$$

Finally, the sample-specific prediction error for the new sample is given by:

$$\mathrm{PE}(\boldsymbol{\xi}) = [(1 + h)\mathrm{MSEC}]^{1/2} \tag{6}$$

It has been reported that the estimates of prediction error obtained by Eq. (6) may be slightly pessimistic [9]. A correction is possible if some knowledge is available about what is called the variance of the reference method, i.e. some knowledge about the uncertainties of the elements of $\boldsymbol{y}$. However, in our application in the next section this variance will be negligible. Furthermore, note that Eq. (6) is identical to the sample specific prediction error estimate that is used by the American Society for Testing and Materials (ASTM) for PLS [10] and in general it can be stated that the estimate obtained by Eq. (6) is satisfactory in many a practical application.

The last important question to address is the correct number of latent variables to use for prediction (previously denoted k). In practice, it is most frequently estimated by means of cross-validation. Briefly, cross-validation comes down to randomly omitting an arbitrary number of observations from the calibration data set and then predicting these observations from the remaining observations. This is repeated in an arbitrary number of iterations, whereafter an overall root mean squared error of prediction can be computed. Ideally, this number should be minimal at the optimal number of latent variables. Different names are given to the method of cross-validation depending on the number of observations left out at atime and their location in the original data matrix. If one observation is being left out each time, the method is called leave one out cross-validation. This type of cross-validation has been reported to under-estimate the true number of underlying components in the case of partial least squares regression [11]. A better, frequently applied alternative is venetian blinds cross-validation, where the pattern of observations being left out resembles venetian blinds. This type of cross-validation has been used throughout the application in the next Section.

3. Identification of micro-organisms

In the current paper, we will not focus on the experimental set-up of the electronic nose, as this has been discussed in a separate paper [7]. As stated in the Section 1, we will mainly discuss the application of trilinear partial least squares to the data generated by the electronic nose.

Ten types of micro-organisms need to be unambiguously identified. For each micro-organism, a binary response vector was created. These different response vectors will be referred to as classes of data. Furthermore, some samples in the calibration data matrix correspond to the presence of no organism at all. An 11th binary response vector corresponds to this class of data. The different classes are summarized in Table 1. It was decided that the calibration data matrix should contain about 10 samples belonging to the same class. Due to practical considerations, only 5 samples corresponding to class nine were included, so that the final calibration set consisted of 105 samples.

Tri-PLS1 regression was performed for each data class. As the problem could be considered to be multivariate (11 response variables), one could consider tri-PLS2 regression to be a viable alternative. However, application of tri-PLS2 would require a careful design of the calibration matrix, inserting calibration samples in which several of the bacteria are present at the same time. As it is more practicable to prepare samples of a single culture of bacteria, we opted only to insert samples of this type. Moreover, the tri-PLS1 routine is more transparent and more efficient in the computational sense and hence we decided to apply tri-PLS1 regression to each of the classes separately. For each data class, a vector of regression coefficients $\boldsymbol{\beta}^{\mathrm{class}\,i}$ is obtained using the tri-PLS1 algorithm (Eqs. (1)). A separate validation data set consisted of 22 observations, two observations belonging to each class. For each class, a vector of predicted responses $\hat{\boldsymbol{\upsilon}}^{\mathrm{class}\,i}$ is computed (Eq. (2)). The micro-organisms were identified as follows: if for the prediction of, e.g. class 10 a number close to one is obtained, the corresponding micro-organism is considered to belong to this class. If, on the contrary, a very small or negative number is obtained, the micro-organism is considered not to belong to this class. As the same type of calculations has to be repeated for each class of data, we

Table 1
Micro-organisms to which the different data classes correspond

Data class	Biological class	Species
1	Gram negative	*Pseudomonas æruginosa*
2	Gram negative	*Enterobacter ærogenes*
3	Gram negative	*Proteus vulgaris*
4	Gram positive	*Staphylococcus aureus*
5	Gram negative	*Escherichia coli*
6	Gram negative	*Klebsiella pneumoniæ*
7	Mould	*Aspergillus fumigatus*
8	Gram positive	*Streptococcus pneumoniæ*
9	Gram positive	*Enterococcus fæcalis*
10	Yeast	*Candida albicans*

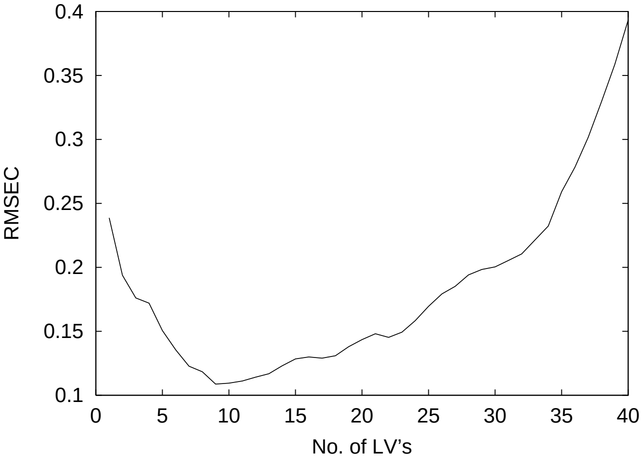

Fig. 1. Cross-validated root mean-squared error of prediction for data class 1.

will only go through the whole method for the first class, whereafter the overall result will be stated.

Let us now analyze the method in detail taking the first data class as an example. At first, the optimal number of latent variables was determined using cross-validation. From Fig. 1, it is clear that the optimal number of latent variables is found at nine.

The predicted response of the observations of the validation set for the fist data class, based on nine latent variables, is given in Table 2. Observations 10 and 15 belong to class 1; the predicted value for these observations is in both cases close to one. For the other observations, the corresponding elements of the vector of predicted responses for class 1 (column 2 in Table 2) are close to zero or even negative.

At this stage, the uncertainties start to play a rôle. Suppose, for instance, that the values close to zero would have prediction errors of about 0.01 and the two values close to one would have prediction errors of about 2, than we would not be able to draw any conclusion: there is an indication that the two observations indeed belong to class 1, but due to the immense uncertainties there might as well be no bacteria at all in the validation set which belongs to class 1 (prior knowledge will not be available in real clinical analyses).

Sample specific prediction errors were computed for the vector of predicted responses for class 1 adopting the approach described in the previous section (Eqs. (3) through (6)). The results are given in the third column of Table 2. The sample-specific prediction errors allow us in this case to draw the final conclusion: the predicted responses corresponding to observations 10 and 15 are close to one and their uncertainties are roughly about 0.15, thus we conclude that observations 10 and 15 are measurements of the excrements of the *Pseudomonas æruginosi*, which is correct.

Let us now investigate a more subtle example, where the calibration data matrix does not as well fit the model. This is the case for several of the classes. As an example, we show the results for class 5. The optimal number of latent variables was determined by means of venetian blinds cross-validation; it equals 14. The predicted responses for the validation set and their sample specific prediction errors were computed as for class 1. The results are summarized in Table 3.

Table 2
Predicted responses for class 1 and their sample specific prediction errors

True class	$\hat{\boldsymbol{\upsilon}}^{\text{class 1}}$	$\text{PE}(\hat{\boldsymbol{\upsilon}}^{\text{class 1}})$
7	0.11	0.11
2	−0.17	0.11
4	0.17	0.11
3	−0.01	0.13
11	−0.01	0.11
10	0.21	0.10
6	−0.15	0.11
8	−0.02	0.11
5	0.07	0.11
1	1.37	0.12
9	0.01	0.11
4	0.31	0.11
6	−0.05	0.11
2	−0.11	0.11
1	1.00	0.11
7	0.05	0.11
3	0.02	0.12
11	−0.01	0.11
5	−0.03	0.11
9	0.04	0.11
10	0.19	0.11
8	−0.01	0.11

The classes to which the observations belong are given in the left column for comparison.

Table 3
Predicted responses for class 5 and their sample specific prediction errors

True class	$\hat{\boldsymbol{\upsilon}}^{\text{class 5}}$	$\text{PE}(\hat{\boldsymbol{\upsilon}}^{\text{class 5}})$
7	0.21	0.20
2	0.11	0.20
4	0.16	0.26
3	0.58	0.91
11	−0.07	0.16
10	0.26	0.21
6	0.39	0.23
8	0.18	0.17
5	0.77	0.22
1	0.10	0.40
9	0.12	0.16
4	0.17	0.23
6	0.21	0.22
2	0.13	0.21
1	−0.40	0.27
7	0.18	0.19
3	0.27	0.66
11	−0.10	0.16
5	0.65	0.24
9	0.14	0.16
10	0.22	0.24
8	0.17	0.18

The classes to which the observations belong are given in the left column for comparison

Table 4
Predicted value of the forth observation in the validation set for class 5 for the forth sample based on 5, 14 and 25 latent variables, respectively

$\hat{\upsilon}_5^{\text{class 5}}(\boldsymbol{\xi}_4)$	$\hat{\upsilon}_{14}^{\text{class 5}}(\boldsymbol{\xi}_4)$	$\hat{\upsilon}_{25}^{\text{class 5}}(\boldsymbol{\xi}_4)$
−0.07	0.58	3.64

From the predicted values for class 5 (the middle column of Table 3) one would draw the incautious conclusion that three samples appertain to the class, i.e. samples 4, 9 and 19. However, the sample specific prediction errors distinguish the prediction for sample 4 from the other two: the PE is close to 1, indicating that the true value for sample 4 might as well be equal to zero. Some further investigation is required for this sample.

The leverage of an observation is an indication for the observation to be considered as good or outlying with respect to the majority of the data. Recall that the leverages are the only contribution to the prediction error which reflects the individual differences among the samples (Eq. (6)). So indeed, if one sample has a large sample specific prediction error compared to the other samples, this sample can without doubt be classified as outlying. In order to illustrate this, we will briefly state the values of the leverages for the samples for prediction of class 5. For all samples these are approximately equal to 0.5, whereas the leverage of sample 4 equals 33, which clearly indicates that observation 4 should be considered outlying with respect to the others for the prediction of class 5, and that it should certainly not be attributed to this class. Furthermore, a computation of the predicted value for a varying number of latent variables clearly illustrates the high uncertainty on the prediction. In Table 4 the predicted values for the 4th observation in the validation set are given for 4, 14 and 25 latent variables. It is clear that this predicted value varies from −1 to 4 depending on the number of latent variables one chooses. This again is an indication that the observation does not belong to class 4 at all, but coincidentially gave such a prediction at the optimal number of LV's due to its high variability. Note that a high degree of volatility in the prediction for several numbers of latent variables cannot in se be considered as an argument to disregard the prediction, but that it is a reinforcing argument when high leverages are encountered. The aforementioned arguments considered, we decided that samples 9 and 19 were generated by the *Escherichia coli*, which is again correct. Furthermore, note that in general the prediction uncertainties for class 5 are considerable higher than for class 1, indicating that the data in this case do not as well fit the model.

As mentioned earlier, it would be tedious and unnecessary to report similar results for the other classes of bacteria and funghi. We state that for the 22 samples to be assigned, a diligent analysis of the tri-PLS results allowed us to identify all of the samples available in the calibration matrix, with one sample identified twice. This amounts to 3.5% of misclassification, which is, as heeded in the Section 1, acceptable for clinical analysis.

4. Summary and conclusions

In the current paper, we proposed a new method for the analysis of data generated by the electronic nose. We proposed to use trilinear partial least squares regression for calibration. Prediction of samples belonging to an independent validation set have been shown to be satisfactory. We have been able to provide sample specific prediction errors for the results obtained, allowing to differentiate between seemingly similar predictions. The overall method allowed us to obtain an acceptable misclassification rate. In this case, the method was applied to identify different pathogenic micro-organisms as a part of wider clinical research. However, the applicability of the method does not necessarily restrict itself to identification of bacteria, as the electronic nose itself can be used in various fields of interest.

Acknowledgements

Research financed by a Ph.D. Grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

References

[1] J. Chastre, J.Y. Fagon, Am. J. Respir. Crit. Care Med. 165 (2002) 867.
[2] M. Holmberg, F. Gustafsson, E.G. Hornsten, F. Winquist, L.E. Nilsson, L. Ljung, I. Lundstrom, Biotechnol. Tech. 12 (1998) 319–324.
[3] R. Bro, J. Chemometr. 10 (1996) 47–61.
[4] R. Bro, Multiway analysis in the food industry, Ph.D. thesis, University of Amsterdam, Amsterdam, The Netherlands, 1996.
[5] N.M. Faber, R. Bro, Chemometr. Intell. Lab. Syst. 61 (2003) 133–149.
[6] H. Wold, in: P.R. Krishnaiah (Ed.), Proceedings of an International Symposium on Multivariate Analysis, 14–19 June 1965, Dayton, OH, Academic Press, NY, 1966, pp. 391–420.
[7] M. Moens, J. Verhoeven, A. Smet, B. Naudts, M. Ieven, P. Jorens, H.J. Geise, F. Blockhuys, in press.
[8] S. de Jong, J. Chemometr. 12 (1998) 77–81.
[9] R. Bro, Å. Rinnan, N.M. Faber, in press.
[10] Annual Book of ASTM Standards, vol. 03.06, E1655, Standard Practices for Infrared, Multivariate, Quantitative Analysis, ASTM International, West Conshohocken, PA, USA, 1998.
[11] K. Baumann, H. Albert, M. von Korff, J. Chemometr. 16 (2002) 339–350.