JOURNAL OF CHEMOMETRICS
J. Chemometrics 2006; **20**: 425–435
Published online 5 February 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/cem.1017

A modification of canonical variates analysis to handle highly collinear multivariate data

Lars Nørgaard[1]*, Rasmus Bro[1], Frank Westad[2] and Søren Balling Engelsen[1]

[1]Department of Food Science, Quality and Technology, Chemometrics Group, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
[2]Norwegian Food Research Institute, Osloveien 1, 1430 Ås, Norway

Received 31 March 2006; Revised 2 July 2006; Accepted 15 July 2006

A modification of the standard Canonical Variates Analysis (CVA) method to cope with collinear high-dimensional data is developed. The method utilizes Partial Least Squares regression as an engine for solving an eigenvector problem involving singular covariance matrices. Three data sets are analyzed to demonstrate the properties of the method: a two-group problem with near infrared spectroscopic data consisting of 60 samples and 376 variables, a multi-group problem with fluorescence spectroscopic data (1023 variables) consisting of 83 samples from six groups and a three-group problem with physical-chemical data (10 variables) consisting of 41 samples from three groups. It is demonstrated that the modified CVA method forces the discriminative information into the first canonical variates as expected. The weight vectors found in the modified CVA method possess the same properties as weight vectors of the standard CVA method. By combination of the suggested method with, for example, Linear Discriminant Analysis (LDA) as a classifier, an operational tool for classification and discrimination of collinear data is obtained. Copyright © 2007 John Wiley & Sons, Ltd.

KEYWORDS: canonical variates; discriminant analysis; classification; collinear; spectroscopy

1. INTRODUCTION

Multi-collinear data including data sets with more variables than samples are often met, for example, in chemometric data analysis particularly due to the development of modern instrumental methods such as nuclear magnetic resonance, near infrared, infrared, fluorescence, Raman and mass spectroscopy. For exploratory data analysis of such data, latent variable methods, for example, Principal Component Analysis (PCA) [1–3], have proven to be highly beneficial. When it comes to the application of classification and discrimination methods on highly collinear data the procedure is often to perform a dimension reduction (e.g., by PCA or Discriminant-PLS (DPLS) [4]) of the data prior to the application of standard discrimination methods on the scores values (e.g., Fisher's Linear Discriminant Analysis (LDA) [5,6]). This dimension reduction is necessary because standard discrimination methods are based on full rank non-singular data; that is, data with low collinearity.

*Correspondence to: L. Nørgaard, Department of Food Science, Quality and Technology, Chemometrics Group, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.
E-mail: lan@kvl.dk

One problem in using PCA is that the components calculated might not necessarily be the components relevant for discrimination and then the subsequent use of, for example, LDA on the score values might not give the optimal result or that so many components are needed that exploratory analysis and data-mining is needlessly complicated. As an alternative to PCA a possibility is to use DPLS as a pre-processing technique and then apply the classifier on the obtained PLS scores. By the application of a dependent matrix containing information about the groups, the PLS loadings and thereby the scores will be more relevant for discriminative purposes [7]. Common to the dimension reduction-based methods is that they produce intermediate scores that subsequently are analyzed by the chosen classifier. This can make interpretation of the results with respect to the original high-dimensional data more difficult since the interpretation is first performed in the reduced space and then in the original data space through the model applied in the pre-processing. An alternative to the above-mentioned methods is to apply DPLS alone [4], that is, the predictions in a DPLS model are used to classify an unknown sample. In DPLS the interpretation can be performed with respect to the original high-dimensional data space but DPLS suffers from poor performance in situations not unlikely to occur in real data [8].

Canonical Variates Analysis (CVA) [6,9] is a method for estimation of directions in space that maximize the differences between the groups in the data according to a well-defined optimization criterion. A drawback of CVA is that it cannot deal with highly collinear data, for example, spectroscopic data tables where the number of variables is larger than the number of samples. Methods have been developed to compensate for this problem by replacing the singular matrices involved in the maximization criteria [10–12]. Related to these methods is the principal discriminant method developed by Jiang *et al*. [13] which is a sort of regularized discriminant analysis [14] bridging the gap between PCA and CVA by testing a continuum of methods. Also CVA-based ranking of principal components have been developed [15] as well as weighted PCA of the group means [16]. The reader is referred to Naes and Indahl [17] that provides a unified description of classification methods for multi-collinear data.

In this study an alternative method is suggested to solve the problem of singular matrices that results when analyzing collinear data with CVA. The method is based on the standard CVA and by a transformation of an eigenvector problem to a regression problem it is possible to use PLS in the inner part of CVA thereby allowing for the analysis of collinear data. The suggested method calculates canonical variates directly in the original high-dimensional space making it possible to interpret the model in relation to this space and perform outlier tests directly without an intermediate model. The suggested method preserves the properties of the standard CVA method.

2. THEORY

We start by outlining the standard method of CVA which is the basis for the developed method. In Reference [6] a description of CVA is given and this reference together with Reference [18] are used throughout as the references for the basic theory. Please note that the following also is presented as a part of Fisher's LDA in many texts (e.g., [18]).

Assume a data matrix $\mathbf{X}$ $(n \times v)$ where the samples are divided into g different groups with n_i samples in the ith group $\left(n = \sum_{i=1}^{g} n_i\right)$.

The within-group covariance matrix is defined as

$$\mathbf{S}_{\text{within}} = \frac{1}{(n-g)} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' \tag{1}$$

and the between-group covariance matrix is defined as

$$\mathbf{S}_{\text{between}} = \frac{1}{(g-1)} \sum_{i=1}^{g} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \tag{2}$$

where $\mathbf{x}_{ij}$ is the jth sample in the ith group (represented as a column vector), $\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$ is the mean vector in the ith group, and $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} \mathbf{x}_{ij} = \frac{1}{n} \sum_{i=1}^{g} n_i \bar{\mathbf{x}}_i$ is the overall mean vector. Note that the dimensions of $\mathbf{S}_{\text{within}}$ and $\mathbf{S}_{\text{between}}$ are both $v \times v$.

It is now possible to define CVA as the problem of finding a direction, $\mathbf{w}$, that maximizes

$$J(\mathbf{w}) = \frac{\mathbf{w}' \mathbf{S}_{\text{between}} \mathbf{w}}{\mathbf{w}' \mathbf{S}_{\text{within}} \mathbf{w}} \tag{3}$$

The solution to this can be written as an eigenvector equation

$$\mathbf{S}_{\text{between}} \mathbf{w} = \lambda \mathbf{S}_{\text{within}} \mathbf{w} \tag{4}$$

If $\mathbf{S}_{\text{within}}$ is non-singular we have the solution

$$\mathbf{S}_{\text{within}}^{-1} \mathbf{S}_{\text{between}} \mathbf{w} = \lambda \mathbf{w} \tag{5}$$

which is an eigenvalue problem, where λ represents the eigenvalue and $\mathbf{w}$ is the eigenvector.

If $\mathbf{S}_{\text{within}}$ is singular it is not possible to left multiply by the inverse of $\mathbf{S}_{\text{within}}$ and this is the cause of the breakdown of standard CVA when analyzing, for example, multi-collinear data.

2.1. The new approach for the two group situation

The new method suggested is the following: for the two-group situation Equation (4) can be rewritten as [18]

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{w} = \lambda \mathbf{S}_{\text{within}} \mathbf{w} \tag{6}$$

$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{w}$ is a scalar, k, so the equation can be written as

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) k = \lambda \mathbf{S}_{\text{within}} \mathbf{w} \tag{7}$$

Equation (7) is now transformed into a standard multivariate regression problem

$$\mathbf{y} = \mathbf{R}\mathbf{b} + \mathbf{f} \tag{8}$$

where $\mathbf{y} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ is the dependent variable, $\mathbf{R} = \mathbf{S}_{\text{within}}$ contains the independent variables and $\mathbf{b} = \mathbf{w}$ is the regression vector. $\mathbf{f}$ contains the residuals. Since k and λ are constants they do not change the solution of Equations (7) and (8), or more precisely the direction of $\mathbf{w}$. In this context it is suggested that Equations (7) and (8) are solved with a PLS regression method but other regression techniques are also applicable. The developed method can be considered as an alternative way of estimating a pseudo-inverse for $\mathbf{S}_{\text{within}}$ by PLS. By multiplication of the mean-centered data matrix, $\mathbf{X}_{\text{MC}}$ with the weight vector, $\mathbf{w}$, the canonical variates, $\mathbf{t}_{\text{CV}}$, are obtained ($\mathbf{t}_{\text{CV}} = \mathbf{X}_{\text{MC}} \mathbf{w}$). The mean centering is performed using the mean vector of all calibration samples (the same mean vector as used in the calculation of $\mathbf{S}_{\text{between}}$). The calculated canonical variates can be used directly in a classifier which in this study is an LDA method and as such the method becomes an LDA method for dealing with collinear data.

2.2. Multi-group CVA

When more than two groups are considered more directions may be needed to represent the data adequately. These directions can be estimated from Equation 4 as there will generally be more than one eigenvalue/eigenvector pair:

$$\mathbf{S}_{\text{between}} \mathbf{w}_a = \lambda_a \mathbf{S}_{\text{within}} \mathbf{w}_a \tag{9}$$

where a indicates the number of directions. In general Equation (9) has $a = \min(v, g\text{-}1)$ non-zero eigenvalues and the maximum dimensionality for the canonical space is thus a. Analyzing high-dimensional data (large v) implies that the maximum number of canonical variates is $g - 1$ (the number of groups minus one).

DOI: 10.1002/cem

2.3. The new approach for the multi-group situation

By analogy to the two-group situation we suggest that the regression equation to be solved for the multi-group case is

$$\mathbf{Y} = \mathbf{RB} + \mathbf{F} \tag{10}$$

where $\mathbf{Y}$ contains as columns the differences $(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})$, that is, the difference between each group mean and the overall mean, $\mathbf{R}$ is $\mathbf{S}_{\text{between}}$ and the columns of $\mathbf{B}$ are $\mathbf{w}_a$ (designated as $\mathbf{W}$ in the following). $\mathbf{F}$ is the residual matrix. The dimension of both $\mathbf{Y}$ and $\mathbf{B}$ (and $\mathbf{F}$) is $v \times g$.

PLS2 is used as the regression technique in Equation (10). The number of weights calculated corresponds to the number of groups and the weights are sorted in descending order according to their values when inserted one-by-one in the optimization criterion (Equation (3)). The weight with the lowest value is left out before application of the classifier because there is an intrinsic rank-deficiency due to the closure properties of the dependent variables. The properties of PLS2 will ensure that the space spanned by the retained $g-1$ weights cover the full space of the solution which is all that is needed. An alternative solution would be to modify $\mathbf{Y}$ such that the rank-deficiency was removed before performing the regression but this would imply an unequal scaling of the groups which is not desirable [7]. If we alternatively used g PLS1 models, rather than one PL2 model, it would not be possible to remove one of these since they would span a g dimensional space and furthermore the solution provided would not be consistent with traditional CVA.

By multiplication of the mean-centered data matrix, $\mathbf{X}_{\text{MC}}$, with the canonical weights matrix, $\mathbf{W}$, the canonical variates, $\mathbf{T}_{\text{CV}}$, are obtained ($\mathbf{T}_{\text{CV}} = \mathbf{X}_{\text{MC}}\mathbf{W}$) and LDA (or other classifiers) can then be applied on these. The advantage of the presented method is that no dimension-reducing step is necessary before the classifier is applied to the canonical variates; the discriminative directions are estimated directly in the original multidimensional space which is of interest, for example, spectroscopic applications. Residuals $\mathbf{E}$, for example, for outlier control, can be found as $\mathbf{E} = \mathbf{X}_{\text{MC}} - \mathbf{X}_{\text{MC}}\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$.

2.4. Relation between standard CVA and the suggested method

If the data considered are full rank and non-singular the method suggested and standard CVA give identical solutions.

For multi-group data there is a rotational ambiguity for the $\mathbf{W}$ matrix. For standard CVA the following restrictions are usually imposed for any i, j where $i \neq j$ [19]:

$$\mathbf{W}_i'\mathbf{S}_{\text{within}}\mathbf{w}_i = 1 \tag{11}$$

$$\mathbf{W}_i'\mathbf{S}_{\text{within}}\mathbf{w}_j = 0 \tag{12}$$

Since the new method might result in a rotated $\mathbf{W}$, the restrictions of Equations (11) and (12) are not in general fulfilled immediately. Equation (11) can be fulfilled by a suitable scaling of the weight vectors and Equation (12) can be fulfilled by a rotation of $\mathbf{W}$ using the eigenvectors of $\mathbf{W}'\mathbf{S}_{\text{within}}\mathbf{W}$ as a rotation matrix. This scaling and rotation is implemented in the algorithm used in this paper.

2.5. LDA classifier

A LDA that fits a multivariate normal density to each group with a pooled estimate of covariance [6] was used as the classifier in the present study. The discriminant function for the canonical variates is

$$L_i(\mathbf{t}) = \log(\pi_i) - \frac{1}{2}(\mathbf{t} - \bar{\mathbf{t}}_i)'\mathbf{S}^{-1}_{\text{within},\mathbf{T}_{\text{CV}}}(\mathbf{t} - \bar{\mathbf{t}}_i) + \log|\mathbf{S}_{\text{within},\mathbf{T}_{\text{CV}}}| \tag{13}$$

where i is a group index $(1, \ldots, g)$, $\mathbf{t}$ contains the canonical variates (as a column vector) for the sample to be classified, $\bar{\mathbf{t}}_i$ is the mean vector of the canonical variates for group i, $\mathbf{S}_{\text{within},\mathbf{T}_{\text{CV}}}$ is the pooled covariance matrix for the canonical variates (an analog to $\mathbf{S}_{\text{within}}$ for the raw data presented above). The prior, π, was selected as equal probabilities, for example, if six groups are analyzed the prior is 1/6 for each group. The sample is classified to the group that gives the highest value of L_i.

2.6. Validation and data pre-processing

All models presented are validated using fivefold segmented cross-validation [20], that is, if the data set consists of 60 samples, the samples left out in the first segment are 1, 6, 11, 16, 21, 26, 31, 36, 41, 46, 51, 56. It should be stressed that for the models including LDA as the classifier both the suggested CVA model and the LDA step are included in cross-validation scheme.

Data that are normally mean centered (e.g., spectroscopic data) should *not* be mean centered prior to application of the suggested CVA method. The calculation of covariance matrices has a built-in mean centering (Equations (1) and (2)) for each group and a mean centering including all samples. Data that are normally autoscaled should only be scaled (i.e., not mean centered) prior to application of the method due to the same reasons.

2.7. Method name

As described in the introduction several methods have been developed to deal with the collinearity problem in CVA and we suggest this class of methods is called Extended Canonical Variates Analysis (ECVA) since the methods are capable of handling extended data sets. Our specific suggested method we name ECVA in the following. Used for classification with LDA as the classifier it is called ECVA-LDA.

3. EXPERIMENTAL

3.1. Software

MATLAB Version 7.2.0.232 with the Statistics Toolbox Version 5.2 (R2006a) (The MathWorks, Inc., Natick, MA, USA), the PLS_Toolbox 4.0 (Eigenvector Research, Inc., Manson, WA, USA) and LatentiX 1.05beta (Latent5, Denmark, www.latentix.com) were used for the calculations.

Software for the suggested method was developed by the authors and can be downloaded at http://www.models.kvl.dk (ECVA Toolbox) together with the data sets analyzed. Note that it is possible to apply other discriminant analysis classifiers in the toolbox than presented in this paper; for

DOI: 10.1002/cem

example, quadratic discriminant analysis, and also the use of non-uniform priors is an option.

3.2. Data sets

3.2.1. Near infrared spectroscopy data set—two groups

In total 60 samples of blood from slaughter pigs from two slaughterhouses were measured by near infrared spectroscopy. NIR spectra were collected in the interval 1100–1850 nm with a 2 nm step. Thirty samples are from slaughterhouse A and 30 samples are from slaughterhouse B, so the dimension of **X** is 60×376. The data are a subset of data from a larger study [21] where the purpose was to investigate if differences in pig stunning methods could be observed in the blood.

3.2.2. Fluorescence spectroscopy data set—six groups

In total 83 white sugar samples from six sugar factories were measured in solution by fluorescence spectroscopy. The number of samples from the six factories is 13, 14, 15, 14, 15, and 12, respectively. The fluorescence emission spectra are recorded at four excitation wavelengths: 230 nm (emission 275–560 nm), 240 nm (emission 275–560 nm), 290 nm (emission 311–560 nm), and 340 nm (emission 361–560 nm) and the spectra are concatenated before analysis. The dimension of **X** is 83×1023. For a thorough description of the sugar data set see Reference [22].

3.2.3. Physical-chemical data set—three groups

In total 41 white sugar samples from three sugar factories were analyzed with respect to 10 quality parameters: ash content, color, turbidity, grainsize Mk, grainSize s, SO_2, invert, floc, residue, and amino N. The number of samples in the three groups is 15, 14, and 12, respectively. The dimension of **X** is 41×10. For a thorough description of the sugar data set see Reference [22].

4. RESULTS AND DISCUSSION

4.1. Two-group model—data set one

The first example to be analyzed is the two-group case with 60 samples and 376 variables (Figure 1). The number of canonical directions is always one less than the number of groups in the data set and for the two-group case this means that the solution is one-dimensional.

In Figure 2A the canonical weight vector for the ECVA method is shown. This weight is based on eight PLS components in the inner relation as given in Equation (8) (see below for a justification of the number of components). In Figure 2B the corresponding canonical variates are shown, and it is observed that the discrimination seems promising. In the original spectroscopic data it can be seen when inspecting the raw and mean-centered data (not shown) that the largest differences between the two groups of spectra are in the wavelength range from 1400 to 1580 nm; this matches the canonical weight which has large values in this spectral range.

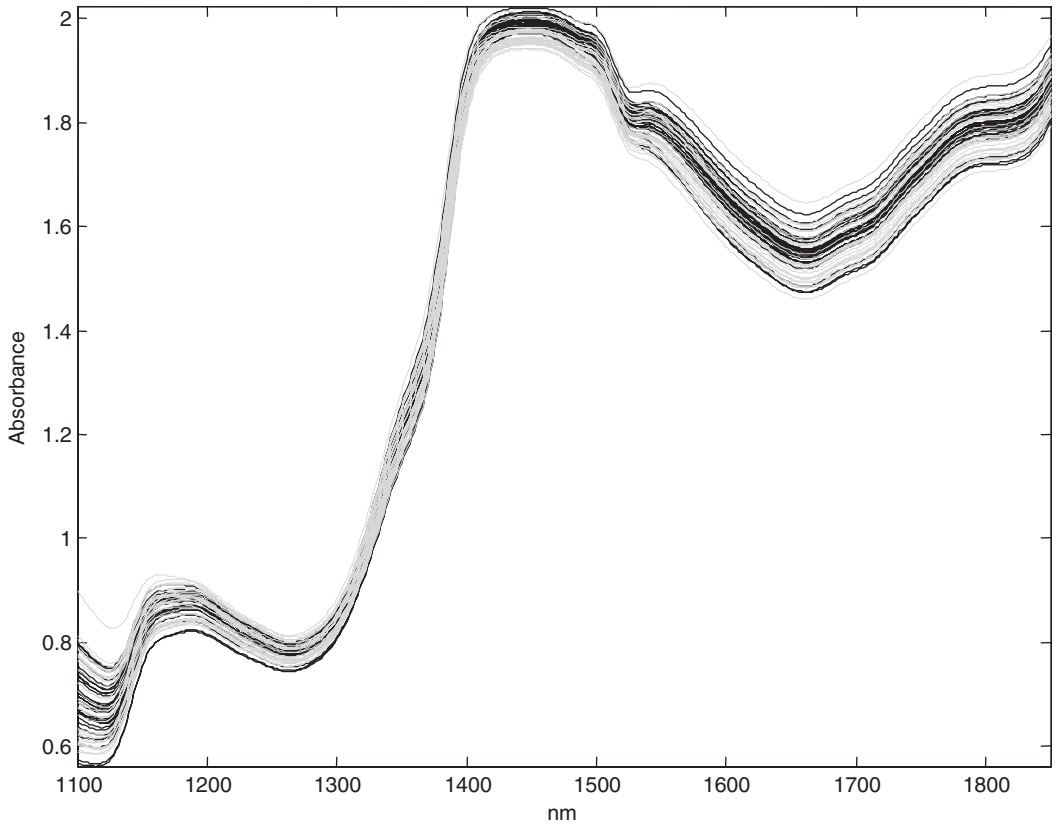

Figure 1. Near infrared spectra in the range 1100–1850 nm of 60 blood samples from two slaughterhouses A (black) and B (gray). The spectra are severely overlapping.

DOI: 10.1002/cem

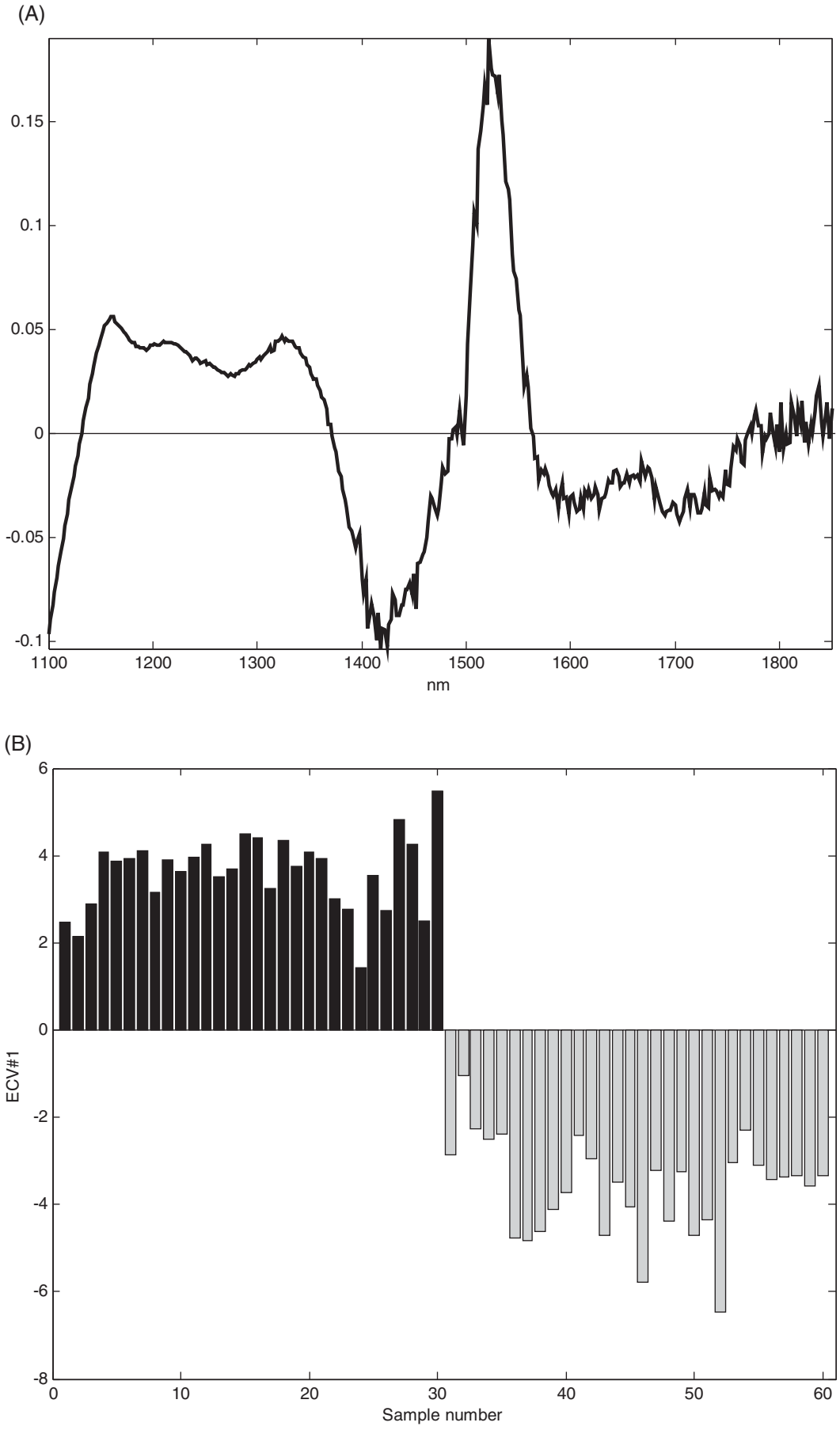

Figure 2. (A) The extended canonical weight vector for the two-group NIR problem. (B) The corresponding extended canonical variates for group 1 (black) and 2 (gray).

DOI: 10.1002/cem

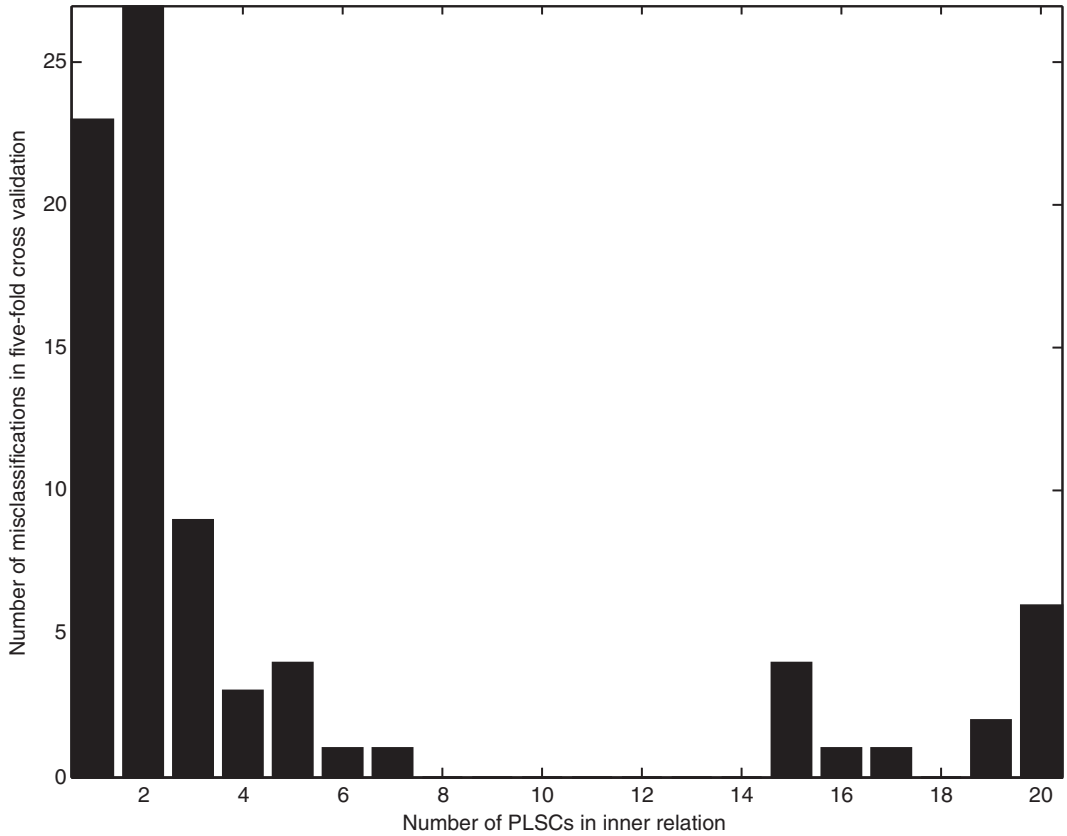

Figure 3. Number of misclassifications as a function of the number PLS components used in the inner PLS relation of ECVA on the NIR data.

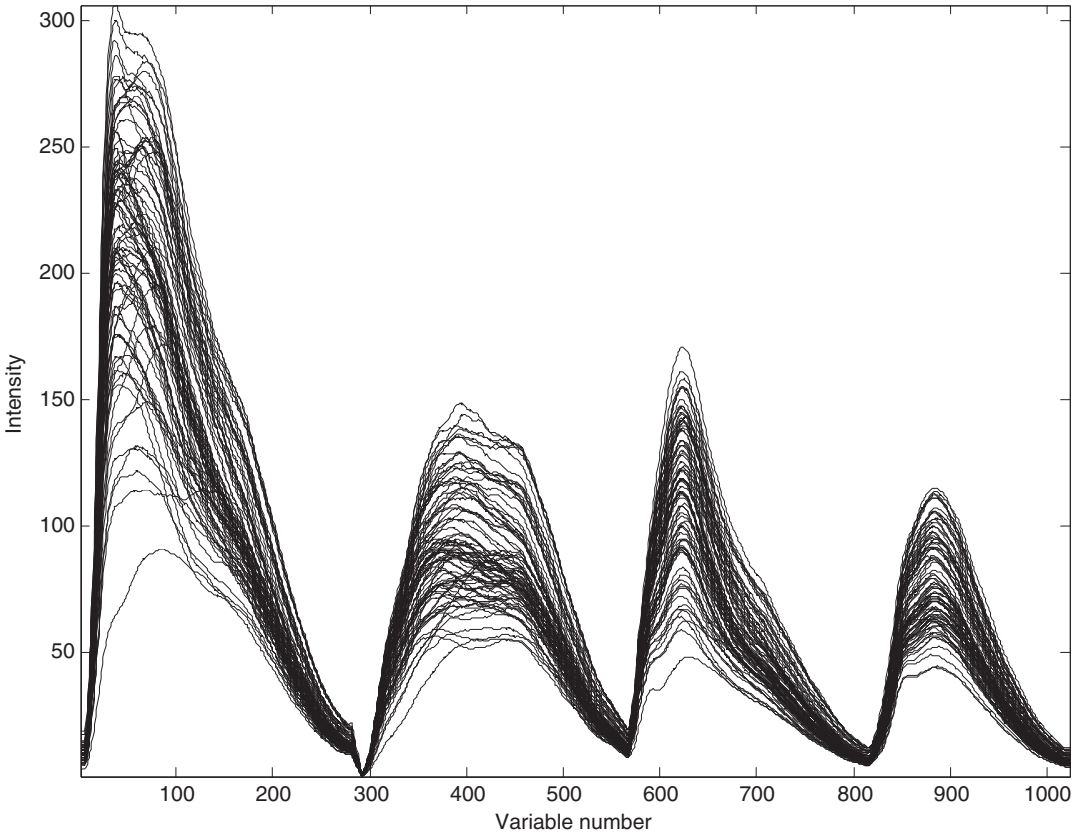

Figure 4. Concatenated fluorescence spectra of 83 water dissolved white sugar samples originating from six factories. The excitation wavelengths recorded are 230 nm, 240 nm, 290 nm, and 340 nm.

DOI: 10.1002/cem

Adding LDA as the classifier the number of misclassifications in a fivefold cross-validation as a function of the number of PLS components in the inner relation is given in Figure 3. For 8–14 components the number of misclassifications are zero so 8 is chosen as the optimal number in this case.

4.2. Multi-group model—data set two

Data set two is a more complex case including six groups; the raw data are shown in Figure 4. In Figure 5A and B the first two canonical variates and weights from an ECVA are shown. For comparison scores and loadings plots from a PCA (mean-centered data) on the same data is also given (Figure 5C and D). As expected the ECVA method focuses on discrimination in the early components compared to PCA (compare Figure 5A and C). The ECVA method has five canonical directions since the number of groups is six and if combinations of the five corresponding canonical variates plots are inspected it is seen that discrimination is pronounced compared to the same plots for the PCA.

Considering the canonical weights, the price paid for obtaining a good discrimination is that these contain more noise than the corresponding loadings in the PCA model (Figure 5B and D). The canonical weights are based on a 17 component PLS model in the inner relation and this is reflected in the noisy structure of the canonical weights.

LDA is introduced as the classifier with the canonical variates as the inputs and the number of misclassifications as a function of the number of components is shown in Figure 6. The whole model (ECVA + LDA) was fivefold cross-validated and the misclassification errors are the validated errors. The optimal number of components in the inner relation is estimated to be 17 and as for the prior example, it is seen that the cross-validated results seem quite stable even with slight overfactoring.

4.3. Multi-group model—physical-chemical data set

This data set consists of physical-chemical data and the data are scaled before analysis, that is, each element of a variable is divided by the standard deviation of the variable (note that the data should not be autoscaled).

In Figure 7A and B the canonical variates and weights scatter plots from an ECVA are shown. These plots can be interpreted from a chemical point of view, that is, the discriminating variables for factory 3 versus 1 and 2 are turbidity, grainsize Mk, and color which are known to be generally higher for factory 3 than the other ones. For factory 1 versus 2 the discriminating variables are ash content, amino N, and color.

Used in this way the ECVA method has exactly the same options with respect to interpretation as the standard CVA method.

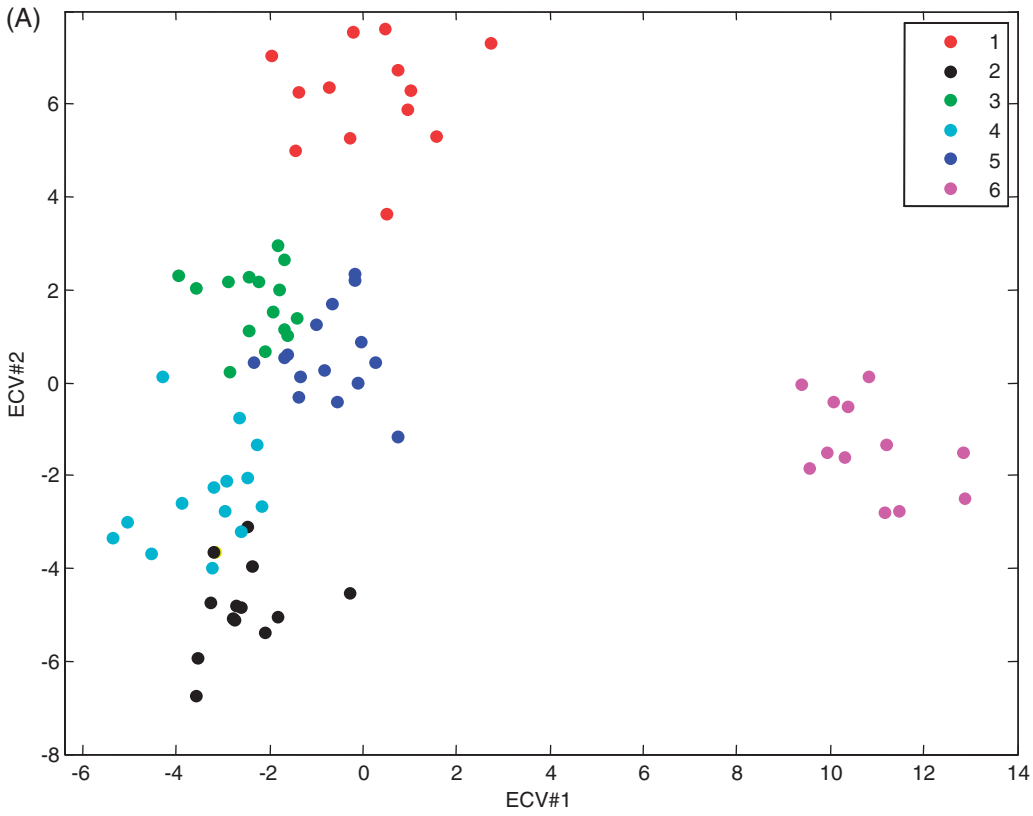

Figure 5. (A) ECV#1 versus ECV#2 for the fluorescence data consisting of six groups (1–6). (B) Corresponding weights for direction one (black) and two (gray). (C) Scores on PC#1 versus PC#2 for a PCA model on the same data. (D) PCA loadings for PC#1 (black) and PC#2 (gray).

DOI: 10.1002/cem

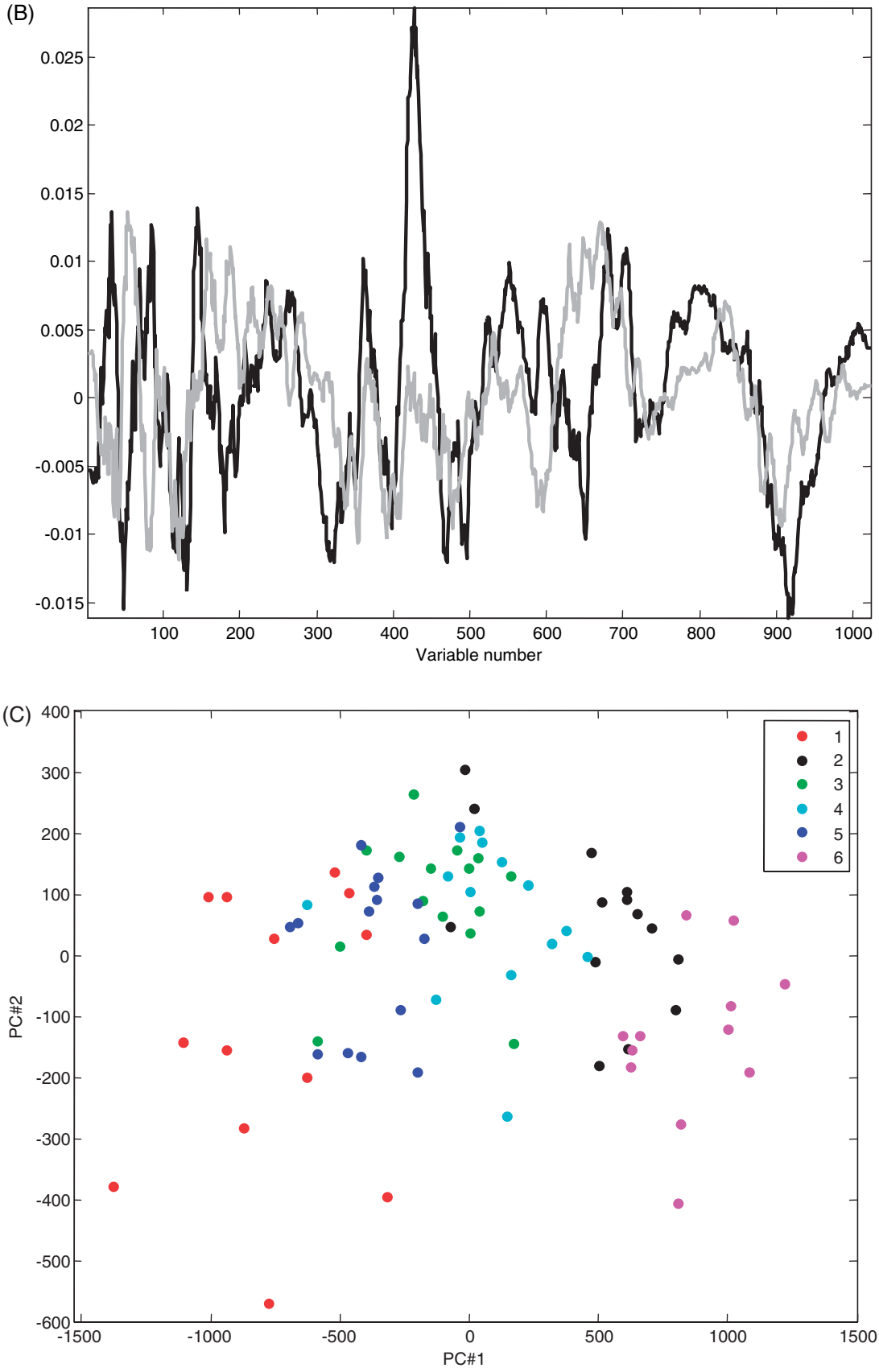

Figure 5. (*Continued*)

DOI: 10.1002/cem

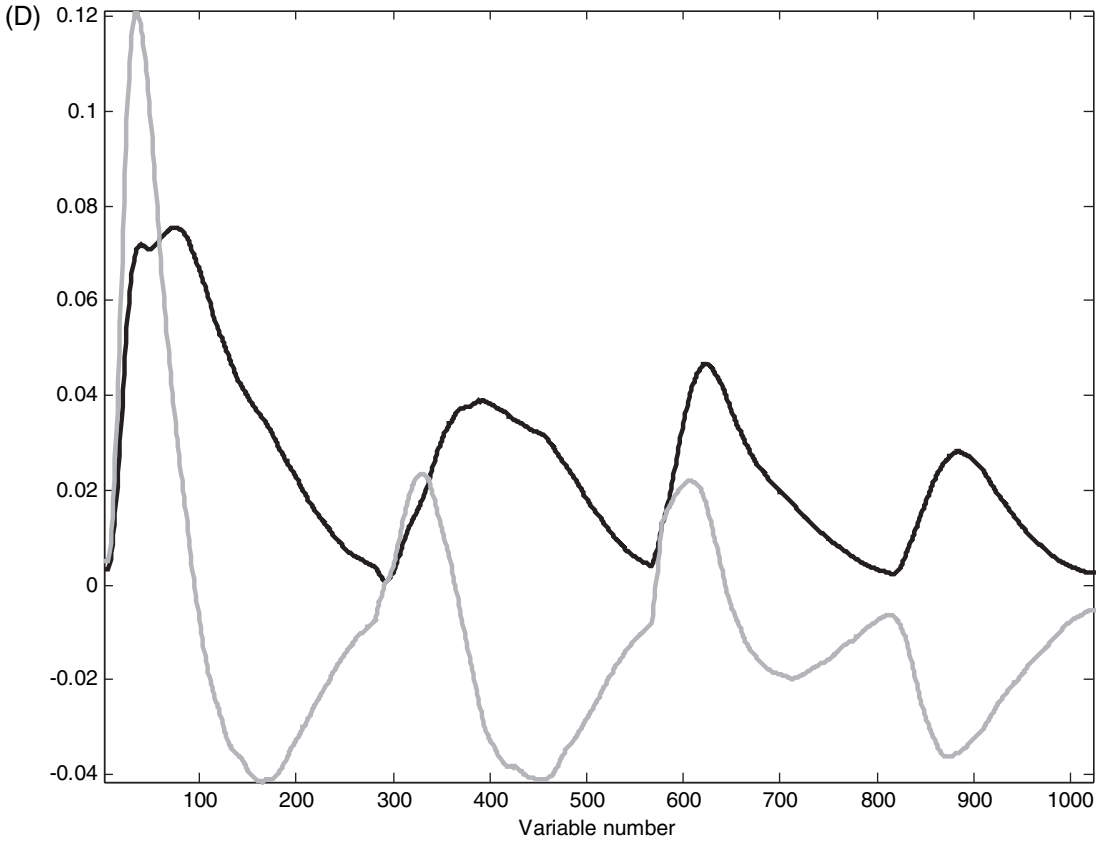

Figure 5. (*Continued*)

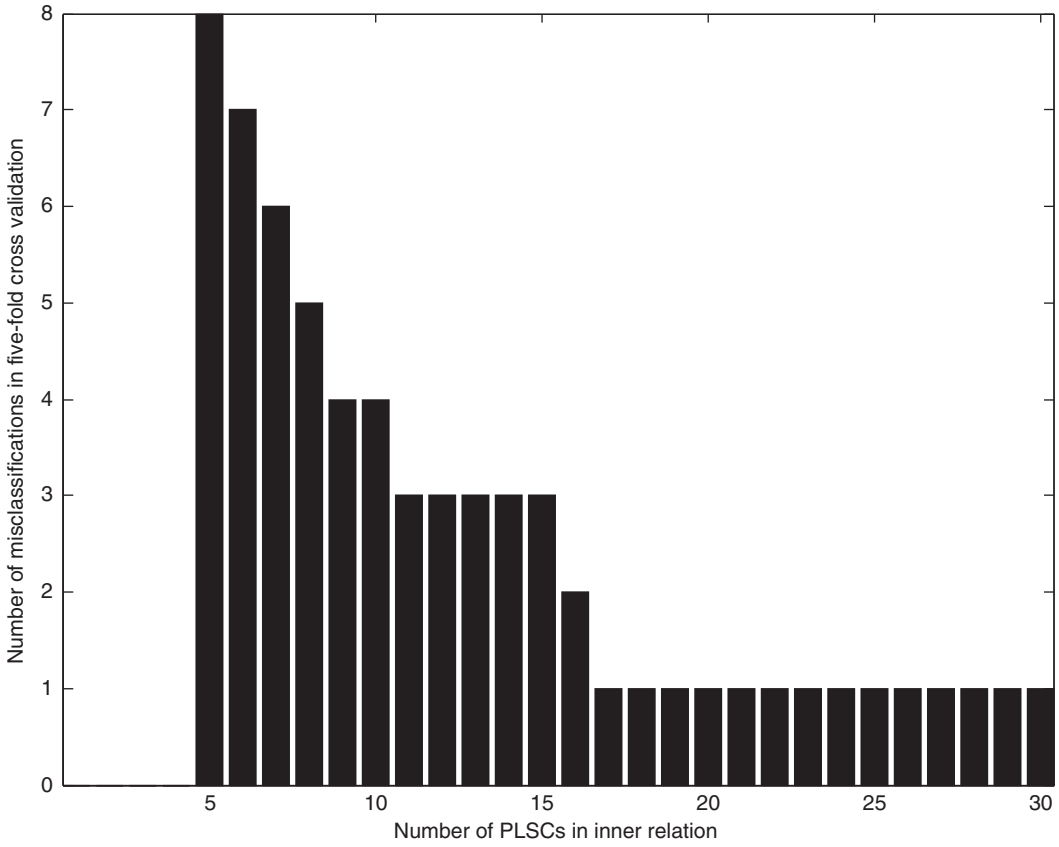

Figure 6. Number of misclassifications as a function of the number PLS components used in the inner PLS relation of ECVA on the fluorescence data. Note that at least five components have to be calculated (number of groups minus one), so the number of misclassification is missing for the first four components.

DOI: 10.1002/cem

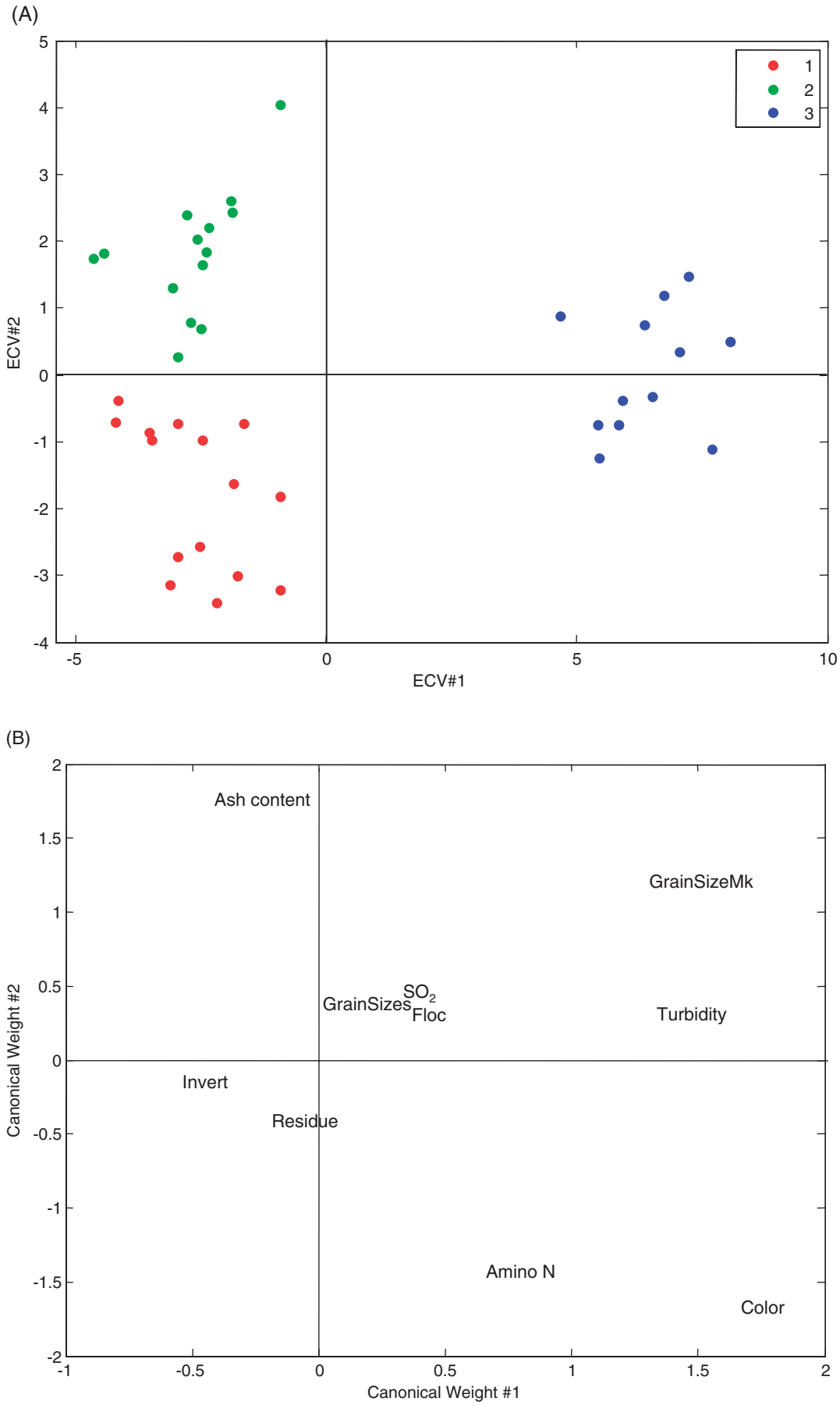

Figure 7. (A) ECV#1 versus ECV#2 for the physical-chemical data consisting of three groups (1–3). (B) Corresponding extended canonical weights scatter plot for direction one versus two.

DOI: 10.1002/cem

4.4. Comment on the inner relation PLS model

Inspecting the inner PLS model, slight non-linearities are sometimes observed (results not shown). This might be due to the fact that the $\mathbf{S}_{within}$ matrix is a covariance matrix and therefore symmetric. It is important to notice that irrespective of the number of components used in the inner PLS2 model the resulting ECVA solution always finds g-1 directions. Another topic for future research is the comparison with other multivariate classification methods and with methods utilizing different types of inverses for $\mathbf{S}_{within}$. Regarding the last aspect it will be interesting to observe if the inner PLS model can utilize the advantage normally ascribed to PLS, that is, its ability to focus on the relevant parts of the data.

5. CONCLUSIONS

In the present study the standard CVA method has been modified to cope with multi-collinear data. The modified method which is named ECVA possesses the same general properties as the standard CVA method, that is, discrimination is forced into the first canonical variates. The ECVA can be coupled with, for example, LDA as a classifier to yield an operational tool for classification of collinear data.

Acknowledgements

The authors thank Dorthe Kjær Pedersen and the Danish Meat Research Institute for carrying out the slaughterhouse measurements.

REFERENCES

1. Hotelling H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 1933; **24**: 498–520.
2. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometr. Intell. Lab. Syst.* 1987; **2**: 37–52.
3. Kvalheim OM. Latent-structure decompositions (projections) of multivariate data. *Chemometr. Intell. Lab. Syst.* 1987; **2**: 283–290.
4. Stahle L, Wold S. Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. *J. Chemometrics* 1987; **1**: 185–196.
5. Fisher RA. The Use of multiple measurements in taxonomic problems. *Ann. Eugen.* 1936; **7**: 179–188.
6. Krzanowski WJ. *Principles of Multivariate Analysis* (Revised edn). Oxford University Press: New York, 2000.
7. Barker M, Rayens W. Partial least squares for discrimination. *J. Chemometrics* 2003; **17**: 166–173.
8. Indahl UG, Sahni NS, Kirkhus B, Naes T. Multivariate strategies for classification based on NIR-spectra—with application to mayonnaise. *Chemometr. Intell. Lab. Syst.* 1999; **49**: 19–31.
9. Rao CR. *Advanced Statistical Methods in Biometric Research*. Wiley: New York, 1952.
10. Kiiveri HT. Canonical Variate Analysis of High-Dimensional Spectral Data. *Technometrics* 1992; **34**: 321–331.
11. Krzanowski WJ, Jonathan P, McCarthy WV, Thomas MR. Discriminant-analysis with singular covariance matrices—methods and applications to spectroscopic data. *Appl. Stat.-J. Royal. Stat. Soc. Series C* 1995; **44**: 101–115.
12. Jonathan P, McCarthy WV, Roberts AMI. Discriminant analysis with singular covariance matrices. A method incorporating cross-validation and efficient randomized permutation tests. *J. Chemometrics* 1996; **10**: 189–213.
13. Jiang JH, Tsenkova R, Wu YQ, Yu RQ, Ozaki Y. Principal discriminant variate method for classification of multicollinear data: applications to near-infrared spectra of cow blood samples. *Appl. Spectrosc.* 2002; **56**: 488–501.
14. Frank IE, Friedman JH. Classification: oldtimers and newcomers. *J. Chemometrics* 1989; **3**: 463–475.
15. Krzanowski WJ. Ranking principal components to reflect group-structure. *J. Chemometrics* 1992; **6**: 97–102.
16. Yendle PW, Macfie HJH. Discriminant principal component analysis. *J. Chemometrics* 1989; **3**: 589–600.
17. Naes T, Indahl U. A unified description of classical classification methods for multicollinear data. *J. Chemometrics* 1998; **12**: 205–220.
18. Duda RO, Hart PE, Stork DG. *Pattern Classification* (2nd edn). John Wiley & Sons: New York, 2001.
19. Krzanowski WJ. Orthogonal canonical variates for discrimination and classification. *J. Chemometrics* 1995; **9**: 509–520.
20. Wold S. Cross-validatory estimation of number of components in factor and principal components models. *Technometrics* 1978; **20**: 397–405.
21. Pedersen DK, Holst S, Norgaard L, Støier S, Engelsen SB. Towards objective instrumental assessment of the depth of CO_2 stunning of slaughter pigs. An exploratory study using visual and near infrared spectroscopy on blood. *in prep.*
22. Norgaard L. Classification and prediction of quality and process parameters of thick juice and beet sugar by fluorescence spectroscopy and chemometrics. *Zuckerindustrie* 1995; **120**: 970–981.

DOI: 10.1002/cem