Research Article

JOURNAL OF CHEMOMETRICS

Received: 31 October 2008, Revised: 19 March 2009, Accepted: 13 April 2009, Published online in Wiley Interscience: 2009

(www.interscience.wiley.com) DOI: 10.1002/cem.1243

Canonical partial least squares—a unified PLS approach to classification and regression problems

Ulf G. Indahl[a,*], Kristian Hovde Liland[b] and Tormod Næs[c]

We propose a new data compression method for estimating optimal latent variables in multi-variate classification and regression problems where more than one response variable is available. The latent variables are found according to a common innovative principle combining PLS methodology and canonical correlation analysis (CCA). The suggested method is able to extract predictive information for the latent variables more effectively than ordinary PLS approaches. Only simple modifications of existing PLS and PPLS algorithms are required to adopt the proposed method. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: canonical correlation analysis; partial least squares; regression with several responses; discriminant analysis; powered partial least squares

1. INTRODUCTION: DATA COMPRESSION FOR CLASSIFICATION AND MULTI-RESPONSE REGRESSION PROBLEMS

This paper presents a collection of new ideas and possibilities in the ongoing research and development of PLS methodology. The underlying motivation of our work is twofold. We believe that for modeling with latent variables, more emphasis should be put on the possibilities of

1. using *additional* information such as
   - weights (individual weights or weighting of groups of observations related to for instance their relative frequencies),
   - additional measurements (e.g., reference measurements, design factors etc.) not necessarily available for prediction of future samples.
2. deriving *alternative* approaches to extraction of components so that
   - fewer components are required for good predictions,
   - interpretations of the associated models are simplified.

Both aspects have been emphasized in recent work; the difference between multi-response PLS (PLS2) with dummy coded responses indicating group membership, and the PLS discriminant analysis (PLS-DA) discussed in References [1,2], and [3] is an important example emphasizing the weighting aspect. In these papers the idea of using PLS for discriminant analysis is put in a framework where more natural and theoretically satisfying weights are assigned to the different groups of observation as compared to the straight forward application of PLS2 with dummy coded responses. The second aspect is an important part of the powered PLS (PPLS) methodology (see References [4] and [5]). PPLS is a modification of PLS useful for providing more parsimonious models in terms of both the number of components needed to obtain good predictions and the complexity of these components.

In the present paper we will propose an alternative multi-response PLS methodology called *canonical* PLS (CPLS), which combines classification and regression in a joint framework and which emphasizes both aspects introduced above. The method combines PLS and canonical correlation analysis (CCA), and it will be demonstrated to have additional and favorable properties related to

- incorporating information from additional variables (not to be considered as predictors or responses) to improve predictions or interpretations,
- simultaneous utilization of several available responses for the purpose of predicting one particular of these responses as well as possible.

Applications to several data sets indicate that CPLS (in comparison to existing PLS methodology) is able to extract

* *Correspondence to: U. G. Indahl, Department of Mathematical Sciences and Technology and Center for Integrative Genetics, Norwegian University of Life Sciences, N-1432 Ås, Norway.*
E-mail: ulf.indahl@umb.no

a *U. G. Indahl*
Department of Mathematical Sciences and Technology and Center for Integrative Genetics, Norwegian University of Life Sciences, N-1432 Ås, Norway

b *K. H. Liland*
Section for Biostatistics, Norwegian University of Life Sciences, N-1432 Ås, Norway

c *T. Næs*
Nofima Mat AS, Oslovegen 1, NO-1430 Ås, Norway

Copyright © 2009 John Wiley & Sons, Ltd.

more information in the first few components. CPLS is also advantageous since it provides a theoretical framework including a number of PLS based methods. After presenting the basics of CPLS we introduce some extensions and generalizations with direct reference to the underlying motivation described above. Generalization to the framework of PPLS (see References [4] and [5]) provides a class of methods called *canonical* PPLS (CPPLS), encompassing CPLS, PPLS, and single response PLS as sub-methodologies.

2. BACKGROUND: PLS AND CANONICAL CORRELATION ANALYSIS

2.1. Notational conventions

In the following, scalars will be denoted by lower case italicized characters, e.g., $c \in \mathbb{R}$. With p and q being positive integers, vectors will be denoted by lower case bold italic characters, i.e., the p-dimensional $\boldsymbol{u} \in \mathbb{R}^p$. Matrices will be denoted by upper case bold roman characters, i.e., the $p \times q$ matrix $\mathbf{W} \in \mathbb{R}^{p \times q}$.

2.2. The present PLS approaches to regression and classification problems

Given the $n \times p$ matrix $\mathbf{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \ldots\ \boldsymbol{x}_p]$ of predictors and the $n \times q$ matrix $\mathbf{Y} = [\boldsymbol{y}_1\ \boldsymbol{y}_2\ \ldots\ \boldsymbol{y}_q]$ of responses, a PLS2 component is found by maximization of the covariance between $\mathbf{X}$ and $\mathbf{Y}$. More precisely; for the regression situation (continuous response variables) we seek unit vectors $\boldsymbol{u} \in \mathbb{R}^p$ and $\boldsymbol{v} \in \mathbb{R}^q$ so that the expression

$$f_1(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^t\mathbf{X}^t\mathbf{Y}\boldsymbol{v} = \boldsymbol{u}^t\mathbf{W}\boldsymbol{v} \tag{1}$$

is maximized. A solution to this problem is provided by the dominant left and right singular vectors (with unit length) obtained by singular value decomposition (SVD) of

$$\mathbf{W} = \mathbf{X}^t\mathbf{Y} \tag{2}$$

where both the predictor matrix $\mathbf{X}$ and the response matrix $\mathbf{Y}$ are assumed to be centered. A pair of dominant unit vectors $\boldsymbol{a} \in \mathbb{R}^p$ and $\boldsymbol{b} \in \mathbb{R}^q$ maximizing Equation (1) and the associated maximal singular value s corresponds to a rank 1 SVD approximation $\mathbf{W}_{(1)} = s\boldsymbol{a}\boldsymbol{b}^t$ of $\mathbf{W}$. The function $f_1(\boldsymbol{u},\boldsymbol{v})$ is a scaled version of the covariance between the vectors $\mathbf{X}\boldsymbol{u}$ and $\mathbf{Y}\boldsymbol{v}$, i.e., $\mathrm{cov}(\mathbf{X}\boldsymbol{u},\mathbf{Y}\boldsymbol{v}) = \frac{1}{n}f_1(\boldsymbol{u},\boldsymbol{v})$, and its relationship to the CCA problem will be explained further below.

PLS2 aims at simultaneous prediction of several responses based on a joint set of latent variables (see Reference [6]). Despite the fact that single response PLS models (PLS1) often require fewer latent variables for prediction of a particular response variable compared to a PLS2 model, PLS2 is the recommended approach when a common context is required for interpretation of the prediction models (see References [7] and [8]).

PLS associated with classification problems usually includes an $n \times g$ dummy coded group membership matrix $\mathbf{Y}$ (a matrix of zeros and ones arranged so that each column indicates membership according to the corresponding group) where $g \geq 2$ is the number of groups considered. Rather than maximizing Equation (1), Barker and Rayens [1] and Nocairi *et al*. [2] showed that a slightly different expression is the logical choice for PLS modeling of classification problems. Maximization of covariance corresponds to finding the dominant unit eigenvector $\boldsymbol{u}$ of the associated *between-groups sum of squares and cross-products matrix* $\mathbf{B}_\mathbf{P} = n\overline{\mathbf{X}}_g^t\mathbf{P}\overline{\mathbf{X}}_g$ where $\mathbf{P}$ is a $g \times g$ diagonal weighting matrix with diagonal entries $p_k = \frac{n_k}{n}$. This is equivalent to maximization of the function

$$f_2(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^t\mathbf{X}^t\mathbf{Y}(\mathbf{Y}^t\mathbf{Y})^{-\frac{1}{2}}\boldsymbol{v} = \boldsymbol{u}^t\mathbf{W}_\Delta\boldsymbol{v} \tag{3}$$

restricted to unit vectors, where

$$\mathbf{W}_\Delta = \mathbf{W}(\mathbf{Y}^t\mathbf{Y})^{-\frac{1}{2}} \tag{4}$$

The factor $\overline{\mathbf{X}}_g = (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{X}$ is a $(g \times p)$-matrix of group means. According to Indahl *et al*. [3], the dominant left singular unit vector $\boldsymbol{a}$ obtained by SVD of $\mathbf{W} = \mathbf{X}^t\mathbf{Y}$ is also a dominant eigenvector of the weighted between-groups sum of squares and cross-products matrix $\mathbf{B}_\mathbf{Q} = n\overline{\mathbf{X}}_g^t\mathbf{Q}\overline{X}_g$. Here $\mathbf{Q}$ is a $g \times g$ diagonal weighting matrix with diagonal entries q_k that are non-negative and proportional to n_k^2 (the square of group sizes n_k, $k = 1, \ldots, g$) and scaled so that $\sum_1^g q_k = 1$. The dominant left singular unit vector $\boldsymbol{u}$ of the scaled version $\mathbf{W}_\Delta$ coincides with a dominant eigenvector of $\mathbf{B}_\mathbf{P}$ defined above. In the following sections we refer to the latter choice of weighting as PLS-DA, in short for PLS discriminant analysis.

Implementations of PLS are often based on the NIPALS algorithm or the SIMPLS algorithm (see References [9] and [10]). SIMPLS extracts components based on the dominant left singular unit vector of $\mathbf{W} = \mathbf{X}^t\mathbf{Y}$ followed by deflation of $\mathbf{W}$ before extraction of the next component. NIPALS also extracts the dominant left singular unit vector of $\mathbf{W}$, but deflates the entire $\mathbf{X}$ matrix before recomputing $\mathbf{W}$ according to Equation (2). de Jong [9] and Burnham and Viveros [10] give detailed descriptions of the most popular PLS algorithms. As noted in Reference [9], NIPALS and SIMPLS lead to similar but not identical models when the $\mathbf{Y}$-matrix has two or more columns (multiple responses). In the case of PLS-DA either algorithm can be applied with $\mathbf{W}_\Delta$ of Equation (4) replacing $\mathbf{W}$ of Equation (2).

2.3. Canonical correlation analysis

CCA addresses the problem of maximizing the correlation between the $n \times p$ matrix $\mathbf{X}$ and the $n \times q$ matrix $\mathbf{Y}$ in the sense of finding vectors $\boldsymbol{a} \in \mathbb{R}^p$ and $\boldsymbol{b} \in \mathbb{R}^q$ so that the correlation between $\mathbf{X}\boldsymbol{a}$ and $\mathbf{Y}\boldsymbol{b}$ becomes as large as possible (see Mardia *et al*. [11]). Assuming that $\mathbf{X}$ and $\mathbf{Y}$ are centered, CCA maximizes

$$\mathrm{corr}(\mathbf{X}\boldsymbol{u}, \mathbf{Y}\boldsymbol{v}) = \frac{\boldsymbol{u}^t\mathbf{X}^t\mathbf{Y}\boldsymbol{v}}{\sqrt{\boldsymbol{u}^t\mathbf{X}^t\mathbf{X}\boldsymbol{u}}\sqrt{\boldsymbol{v}^t\mathbf{Y}^t\mathbf{Y}\boldsymbol{v}}} \tag{5}$$

over all possible choices of $\boldsymbol{u} \in \mathbb{R}^p$ and $\boldsymbol{v} \in \mathbb{R}^q$. It is straightforward to show that maximization of Equation (5) is equivalent to maximization of the function

$$f_3(\boldsymbol{r}, \boldsymbol{t}) = \boldsymbol{r}^t(\mathbf{X}^t\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^t\mathbf{Y}(\mathbf{Y}^t\mathbf{Y})^{-\frac{1}{2}}\boldsymbol{t} \tag{6}$$

over unit vectors $\boldsymbol{r}$ and $\boldsymbol{t}$. The problem is solved by choosing $\boldsymbol{r} = \boldsymbol{r}_0$ and $\boldsymbol{t} = \boldsymbol{t}_0$ where $\boldsymbol{r}_0$ and $\boldsymbol{t}_0$ are the unit vectors corresponding to the largest singular value (s_0) in the SVD of the matrix $(\mathbf{X}^t\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^t\mathbf{Y}(\mathbf{Y}^t\mathbf{Y})^{-\frac{1}{2}}$. The unit vector $\boldsymbol{r}_0$ obtained by maximization of f_3 is also an eigenvector of the matrix $\mathbf{T}^{-\frac{1}{2}}\mathbf{B}\mathbf{T}^{-\frac{1}{2}}$ corresponding

 Copyright © 2009 John Wiley & Sons, Ltd.

to its dominant eigenvalue $\lambda = s_0^2$ with the definitions $\mathbf{T} = \mathbf{X}^t\mathbf{X}$ and $\mathbf{B} = \mathbf{X}^t\mathbf{Y}(\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{X}$. By defining $\boldsymbol{a} = \mathbf{T}^{-\frac{1}{2}}\boldsymbol{r}_0$ and $\boldsymbol{b} = (\mathbf{Y}^t\mathbf{Y})^{-\frac{1}{2}}\boldsymbol{t}_0$ a corresponding maximum of Equation (5) is obtained.

Note that $\boldsymbol{a}$ is also a dominant eigenvector of the matrix $\mathbf{T}^{-1}\mathbf{B}$ with λ as its associated eigenvalue. For classification problems where $\mathbf{Y}$ is the uncentered dummy coded group membership matrix, the above definition of $\mathbf{B}$ corresponds to the between groups sum of squares and cross-products matrix (see Indahl *et al*. [3]). When $\mathbf{X}$ is centered, the computations of both $\mathbf{B}$ and $\mathbf{T}^{-1}\mathbf{B}$, as well as the eigenvalues and eigenvectors of the latter, are unaffected with respect to $\mathbf{Y}$-centering or not (because the factor $\mathbf{Y}(\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t$ of $\mathbf{B}$ corresponds to the projection mapping onto the column space $Col(\mathbf{Y})$, and the centered columns of $\mathbf{X}$ are already orthogonal to the associated subspace spanned by the constant vector $\mathbf{1} \in Col(\mathbf{Y})$). The dominant eigenvector $\boldsymbol{a}$ defines the canonical loadings and the corresponding dominant canonical variate $\boldsymbol{z} = \mathbf{X}\boldsymbol{a}$ of Fisher's canonical discriminant analysis (FCDA) and maximizes the two associated Rayleigh quotients

$$r_1(\boldsymbol{u}) = \frac{\boldsymbol{u}^t\mathbf{B}\boldsymbol{u}}{\boldsymbol{u}^t\mathbf{V}\boldsymbol{u}} \quad \text{and} \quad r_2(\boldsymbol{u}) = \frac{\boldsymbol{u}^t\mathbf{B}\boldsymbol{u}}{\boldsymbol{u}^t\mathbf{T}\boldsymbol{u}}$$

where $\mathbf{V} = \mathbf{T} - \mathbf{B}$ is the within groups sum of squares and cross products matrix associated with the optimization of FCDA (see Reference [3]). This important relationship between CCA and FCDA, well known from the literature, was first recognized by Bartlett [12].

Finally we note that with $\mathbf{X}^t\mathbf{X}$ proportional to the $p \times p$ identity matrix, maximization of f_3 in Equation (6) is equivalent to maximization of f_2 in Equation (3), and if also $\mathbf{Y}^t\mathbf{Y}$ is proportional to the $q \times q$ identity matrix, maximization of f_3 simplifies to solving the original PLS problem, i.e., maximization of f_1 in Equation (1).

3. NEW DEVELOPMENTS

3.1. Canonical PLS (CPLS)

From Reference [3] (Section 3.1) it follows that maximization of the expressions $\boldsymbol{u}^t\mathbf{X}^t\mathbf{Y}\boldsymbol{v}$ and $\boldsymbol{u}_0^t\mathbf{Z}^t\mathbf{Y}\boldsymbol{v}$ where $\mathbf{Z} = \mathbf{X}\mathbf{W}$ are equivalent. This equivalence follows directly by definition of the dominant right singular vector $\boldsymbol{v} = \boldsymbol{b}$ as the dominant unit eigenvector in the eigendecomposition of $\mathbf{W}^t\mathbf{W}$ with s^2 as the corresponding maximal eigenvalue. The associated dominant left singular unit vector $\boldsymbol{u} = \boldsymbol{a}$ is given by $\boldsymbol{a} = s^{-1}\mathbf{W}\boldsymbol{b}$. Thus, optimization of $\mathbf{X}$, $\mathbf{Y}$-covariance and optimization of $\mathbf{Z}$, $\mathbf{Y}$-covariance is equivalent and the solution of the last can be found by SVD (or eigendecomposition) of

$$\mathbf{Z}^t\mathbf{Y} = \mathbf{W}^t\mathbf{X}^t\mathbf{Y} = \mathbf{W}^t\mathbf{W}$$

For any $n \times p$ predictor matrix $\mathbf{X}$, presumably with $p \gg n$ and an $n \times q$ response matrix $\mathbf{Y}$ with $q \ll n$, the dimensionality and/or potential multi-collinearity problems associated with modeling directly based on the $\mathbf{X}$ data are much avoided when replaced by $\mathbf{Z} = \mathbf{X}\mathbf{W}$ of dimension $n \times q$. In particular, according to ordinary PLS theory, each column of $\mathbf{Z}$ corresponds to the direction of maximum sample covariance with the corresponding column of $\mathbf{Y}$. Note that this direction also corresponds to the dominant component of ordinary principal component analysis (PCA) applied to $\mathbf{W}^t\mathbf{W}$. SVD of $\mathbf{W}$ (or equivalently PCA of $\mathbf{W}^t\mathbf{W}$) finding the PLS2 component does not take any further advantage of the available $\mathbf{Y}$ information. It acts *unsupervised* on $\mathbf{W} = \mathbf{X}^t\mathbf{Y}$ with respect to the computation of a linear combination of the $\mathbf{W}$-columns. CCA on the other hand relates $\mathbf{Z}$ to $\mathbf{Y}$ by considering the $\mathbf{Y}$-information a second time. In this respect the PLS2 covariance maximization based on Equation (1) may be considered as an unnecessarily modest optimization criterion.

As an improvement we suggest CPLS where maximization of $\mathbf{Z}$, $\mathbf{Y}$ covariance is replaced by maximization of the $\mathbf{Z}$, $\mathbf{Y}$ canonical correlation according to Equation (5) with $\mathbf{Z}$ replacing $\mathbf{X}$, and the associated maximizing vectors denoted by $\boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^q$. We define the unit vector $\boldsymbol{w}$ of CPLS loading weights as

$$\boldsymbol{w} = \frac{\mathbf{W}\boldsymbol{c}}{\|\mathbf{W}\boldsymbol{c}\|}$$

with the corresponding score vector $\boldsymbol{t} = \mathbf{X}\boldsymbol{w}$. For extraction of subsequent CPLS components we suggest deflation of the $\mathbf{X}$-matrix (NIPALS) or $\mathbf{W}$-matrix (SIMPLS) based on the definitions of $\boldsymbol{w}$ and $\boldsymbol{t}$. Other modifications are not required. Thus the largest possible number of extracted components coincides with the rank of the centered $\mathbf{X}$-matrix.

Note that in comparison to PLS2, CPLS can be considered as finding a *supervised* linear combination of columns from $\mathbf{W}$. In particular, the CCA finds vectors $\boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^q$ so that the correlation between $\mathbf{Z}\boldsymbol{c}(= \mathbf{X}\mathbf{W}\boldsymbol{c})$ and $\mathbf{Y}\boldsymbol{d}$ is maximized. The corresponding CPLS loading weights $\boldsymbol{w} = \mathbf{W}\boldsymbol{c}$ can be normalized without affecting the optimized correlation. Thus, compared to PLS2 and PLS-DA, the loading weights defined by CPLS aim more aggressively toward prediction of the $\mathbf{Y}$-data in the same sense that linear regression aims more aggressively toward prediction when compared to PCA. Hence, compared to the traditional PLS solutions, one should expect that the resulting CPLS models require fewer components and lead to simplified interpretations from the two dimensional substructures associated with the model. Because solving the CCA problem in multi-variate regression with dummy responses is equivalent to solving the FCDA problem in classification, maximization of Equation (5) with $\mathbf{Z}$ replacing $\mathbf{X}$ avoids the PLS2/PLS-DA inconsistency.

3.2. Extensions

3.2.1. *Weighted CPLS*

Indahl *et al*. [3] suggested inclusion of prior probabilities in the calculations of PLS-DA components, and used this to motivate weighted generalizations of PLS-DA and FCDA. A similar weighted extension of ordinary CCA is straightforward. If the $n \times n$ diagonal weighting matrix $\mathbf{D}$ assigns non-negative weights for each of the n individual observations associated with the $\mathbf{Y}$ data, a slight modification of Equation (5) implies maximization of the weighted correlation, i.e.,

$$\text{wcorr}(\mathbf{Z}\boldsymbol{u}, \mathbf{Y}\boldsymbol{v}, \mathbf{D}) = \frac{\boldsymbol{u}^t\mathbf{Z}^t\mathbf{D}\mathbf{Y}\boldsymbol{v}}{\sqrt{\boldsymbol{u}^t\mathbf{Z}^t\mathbf{D}\mathbf{X}\boldsymbol{u}}\sqrt{\boldsymbol{v}^t\mathbf{Y}^t\mathbf{D}\mathbf{Y}\boldsymbol{v}}} \tag{7}$$

Maximization of Equation (7) is equivalent to maximization of the function

$$f_4(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^t(\mathbf{Z}^t\mathbf{D}\mathbf{X})^{-\frac{1}{2}}\mathbf{Z}^t\mathbf{D}\mathbf{Y}(\mathbf{Y}^t\mathbf{D}\mathbf{Y})^{-\frac{1}{2}}\boldsymbol{v} \tag{8}$$

Copyright © 2009 John Wiley & Sons, Ltd.

restricted to unit vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. From a maximizing pair $\boldsymbol{a}_0, \boldsymbol{b}_0$ of Equation (8) the corresponding maximizing pair $\boldsymbol{a}, \boldsymbol{b}$ of Equation (7) is given by $\boldsymbol{a} = (\mathbf{Z}^t\mathbf{D}\mathbf{X})^{-\frac{1}{2}}\boldsymbol{a}_0$ and $\boldsymbol{b} = (\mathbf{Y}^t\mathbf{D}\mathbf{Y})^{-\frac{1}{2}}\boldsymbol{b}_0$. Here we assume that $\mathbf{Z}$ and $\mathbf{Y}$ are centered by subtraction of weighted means according to the diagonal elements of $\mathbf{D}$.

For estimation of the loading weights of *weighted* CPLS we maximize the weighted canonical correlation between $\mathbf{Z} = \mathbf{X}\mathbf{W}$ and $\mathbf{Y}$ in Equation (7) analogously to the maximization of the ordinary canonical correlation in the CPLS. The weighted CPLS is appropriate for:

- Classification problems with group priors (corresponding to identical weights of the samples within each group, but different weights associated with different groups).
- Regression and classification problems with individual weighting of the observations (possibly unique weights for all n observations).

3.2.2. *CPLS with mixed responses*

When applying PLS in situations where the response matrix $\mathbf{Y}$ contains a mixture of categorical and continuous columns (variables) a theoretical problem emerges because of the PLS2/PLS-DA inconsistency regarding maximization of covariance (see Section 2.2). Consequently either choice between the two variants leaves the practitioner with an ad hoc strategy for extraction of components (except in the special cases of equally sized groups for the categorical data or if all the columns of $\mathbf{Y}$ are individually standardized). With CPLS (weighted or unweighted) a mixture of categorical and continuous columns will not cause such problems because scaling of the responses is not required (scaling has no effect on maximization of canonical correlation).

3.2.3. *The general formulation: CPLS with primary and additional responses*

Although modeling and prediction of the selected responses from the predictors associated with $\mathbf{X}$ is our primary goal, additional information associated with each sample is often available as a by-product the data generation process. By assumption, we consider this information not to be available for prediction of future samples. Examples of additional information may include additional reference measurements, the design factors from an experimental design used to generate the data, or something more exotic such as the fitted values of a model predicting the responses based on a set of predictors not supposed to be included in $\mathbf{X}$.

Assume that the $n \times q_1$ matrix $\mathbf{Y}_{\text{prim}}$ represent our *primary* response data (to be modeled for later predictions based on $\mathbf{X}$-data), and that a corresponding set of additional information is available as the $n \times q_2$ matrix $\mathbf{Y}_{\text{add}}$. Although (by assumption) the variables corresponding to $\mathbf{Y}_{\text{add}}$ will not be available for prediction of future samples, we can still take advantage of this information when building the desired model. Combine the primary and additional information into the $n \times q$ ($q = q_1 + q_2$) super matrix $\mathbf{Y} = [\mathbf{Y}_{\text{prim}}, \mathbf{Y}_{\text{add}}]$ where prediction of $\mathbf{Y}_{\text{prim}}$ is the main task. With $\mathbf{Y}$ composed of the primary and additional blocks we compute $\mathbf{W} = \mathbf{X}^t\mathbf{Y}$ and the corresponding transformed data $\mathbf{Z} = \mathbf{X}\mathbf{W}$ followed by *maximization of the canonical correlation between* $\mathbf{Z}$ *and* $\mathbf{Y}_{\text{prim}}$.

Thus the background information represented in $\mathbf{Y}_{\text{add}}$ is used to add extra columns to $\mathbf{Z}$. To the extent that $\mathbf{Y}_{\text{add}}$ contains information relevant for prediction of $\mathbf{Y}_{\text{prim}}$ that is also present in the $\mathbf{X}$-predictors, $\mathbf{Y}_{\text{add}}$ will contribute more emphasis on this information enabling it to be more efficiently extracted into the CPLS components. Algorithm 1 shows the general formulation of CPLS including additional responses:

1. Calculate $\mathbf{W} = \mathbf{X}^t\mathbf{Y}$ with $\mathbf{Y} = [\mathbf{Y}_{\text{prim}}, \mathbf{Y}_{\text{add}}]$.
2. Transform the $\mathbf{X}$-data to $\mathbf{Z} = \mathbf{X}\mathbf{W}$.
3. With $\mathbf{Z}$ replacing $\mathbf{X}$ and $\mathbf{Y}_{\text{prim}}$ replacing $\mathbf{Y}$, obtain the unit vectors $\boldsymbol{a} \in \mathbb{R}^q$, $\boldsymbol{b} \in \mathbb{R}^{q_1}$ from maximization of Equation (5), or Equation (7) if a weighting matrix $\mathbf{D}$ is available, and calculate the optimal unit loading weight vector $\boldsymbol{w} = \mathbf{W}\boldsymbol{a}/\|\mathbf{W}\boldsymbol{a}\|$.
4. Use the calculated loading weight vector $\boldsymbol{w}$ to find scores, $\boldsymbol{p}$-loadings and do other required computations according to the preferred algorithm (such as SIMPLS, NIPALS etc.).
   (a) Stop or
   (b) Deflate the data set before calculation of the next component (steps 1–5).

Algorithm 1: The general formulation of CPLS.

Note that inclusion of additional variables $\mathbf{Y}_{\text{add}}$ in the general formulation of CPLS corresponds exactly to the first bullet point of the introduction: • 'incorporating information from additional variables (not to be considered as predictors or responses) to improve predictions or interpretations.' From the general formulation with multiple responses $\mathbf{Y} = [\boldsymbol{y}_1, \cdots \boldsymbol{y}_q]$ it is also straightforward to focus one particular response $\boldsymbol{y}_i$ (if the response is continuous) or an $n \times g$ sub-matrix of associated dummy columns $\mathbf{Y}_g$ (if these responses represents the group labeling of a g-group classification problem) as the primary block $\mathbf{Y}_{\text{prim}}$. The remaining columns should be assigned into the additional block $\mathbf{Y}_{\text{add}}$ as indicated in step 1 above. This corresponds to the second bullet point in the introduction: • 'simultaneous utilization of several available responses for the purpose of predicting one particular of these responses as well as possible.'

3.2.4. *Canonical PPLS*

The PPLS methods introduced in References [4] and [5] extend the ordinary PLS methodology. An important feature of PPLS is its ability to reduce or eliminate the influence of predictors less important to prediction. We complete our description of extensions based on canonical correlation by presenting the *canonical* PPLS (CPPLS).

in Reference [4], flexible trade-offs between the element wise correlations and variances defining the loading weights were defined by

$$\boldsymbol{w}(\gamma) = k_\gamma \cdot \Big[s_1 \left|\text{corr}\left(\boldsymbol{x}_1, \boldsymbol{y}\right)\right|^{\frac{\gamma}{1-\gamma}} \cdot \text{std}\left(\boldsymbol{x}_1\right)^{\frac{1-\gamma}{\gamma}}, \ldots, s_p \left|\text{corr}\left(\boldsymbol{x}_p, \boldsymbol{y}\right)\right|^{\frac{\gamma}{1-\gamma}} \cdot \text{std}\left(\boldsymbol{x}_p\right)^{\frac{1-\gamma}{\gamma}} \Big]^t \tag{9}$$

where the power parameter γ is ranging from 0 to 1, s_k denotes the sign of the kth correlation and k_γ is a scaling constant assuring unit length of $\boldsymbol{w}(\gamma)$. For a corresponding parameterization of CPLS we need a generalized reformulation of the transformation matrix $\mathbf{W} = \mathbf{X}^t\mathbf{Y}$.

The columns of $\mathbf{W}$ essentially correspond to PLS1 loading weights whose directions maximize the covariance between $\mathbf{X}$ and the associated columns in $\mathbf{Y}$. Hence we can factorize $\mathbf{W} = \mathbf{S}_x\mathbf{C}\mathbf{P}$, where $\mathbf{S}_x$ is a diagonal $(p \times p)$ matrix containing the column

 Copyright © 2009 John Wiley & Sons, Ltd.

wise standard deviations of $\mathbf{X}$, $\mathbf{C}$ is a $(p \times q)$ matrix containing the pairwise correlations between the columns of $\mathbf{X}$ and $\mathbf{Y}$, and $\mathbf{P}$ is a $(q \times q)$ diagonal scaling matrix with entries proportional to the column wise standard deviations of $\mathbf{Y}$. Because the factor $\mathbf{P}$ only contributes to $\mathbf{Z} = \mathbf{XW}$ by scaling the columns of $\mathbf{XS}_x\mathbf{C}$, i.e., $\text{span}(\mathbf{Z}) = \text{span}(\mathbf{XS}_x\mathbf{C})$, it is sufficient to consider the canonical correlation between $\mathbf{Z}_0 = \mathbf{XS}_x\mathbf{C}$ and $\mathbf{Y}_{\text{prim}}$.

Accordingly, a parametric version of the simplified $\mathbf{W}_0 = \mathbf{S}_x\mathbf{C}$ that corresponds to Equation (9) is given by

$$\mathbf{W}_0(\gamma) = \mathbf{S}_x(\gamma)\mathbf{C}(\gamma) \tag{10}$$

where

$$\mathbf{S}_x(\gamma) = \begin{bmatrix} \text{std}\,(\boldsymbol{x}_1)^{\frac{1-\gamma}{\gamma}} & & 0 \\ & \ddots & \\ 0 & & \text{std}\,(\boldsymbol{x}_p)^{\frac{1-\gamma}{\gamma}} \end{bmatrix} \tag{11}$$

$$\mathbf{C}(\gamma) = \begin{bmatrix} s_{11}|\text{corr}(\boldsymbol{x}_1, \boldsymbol{y}_1)|^{\frac{\gamma}{1-\gamma}} & \cdots & s_{1q}|\text{corr}(\boldsymbol{x}_1, \boldsymbol{y}_q)|^{\frac{\gamma}{1-\gamma}} \\ \vdots & & \vdots \\ s_{p1}|\text{corr}(\boldsymbol{x}_p, \boldsymbol{y}_1)|^{\frac{\gamma}{1-\gamma}} & \cdots & s_{pq}|\text{corr}(\boldsymbol{x}_p, \boldsymbol{y}_q)|^{\frac{\gamma}{1-\gamma}} \end{bmatrix} \tag{12}$$

and s_{ik} is the sign of the correlation $\text{corr}(\boldsymbol{x}_i,\boldsymbol{y}_k)$. A unit loading weight vector $\boldsymbol{w}(\gamma) = \mathbf{W}_0(\gamma)\boldsymbol{a}_\gamma / \|\mathbf{W}_0(\gamma)\boldsymbol{a}_\gamma\|$ is obtained as in ordinary CPLS by replacing the $\mathbf{Z}$ matrix with the parameterized version $\mathbf{Z}(\gamma) = \mathbf{XW}_0(\gamma)$ followed by maximization of Equation (5) (or Equation (7) if a non-trivial weighting matrix $\mathbf{D}$ is included). With $\gamma \to 1$ and $\gamma \to 0$, selection of the $\mathbf{X}$-variable most correlated (in absolute value) to any of the columns in $\mathbf{Y}$ and selection of the $\mathbf{X}$-variable with the largest standard deviation is approached, respectively. Note that the special case of CPPLS with $\gamma = 0.5$ (fixed) is equivalent to CPLS.

Due to the relationship between FCDA and CCA, the powered PLS-DA method described in Reference [5] can be considered as a special case (dummy-coded categorical response matrix $\mathbf{Y} = \mathbf{Y}_{\text{prim}}$ with no additional responses and a weighting matrix $\mathbf{D}$ with entries corresponding to the associated group sizes) of CPPLS. The numerical optimization required to compute the γ-values defining the components in CPPLS is as described for PPLS-DA in Reference [5].

4. MODELING REAL DATA WITH CPLS AND EXTENSIONS

4.1. Classification with prior probabilities

To illustrate an application of classification with prior probabilities, we use a data set from the resource web-page (http://www-stat.stanford.edu/~tibs/ElemStatLearn/data.html) accompanying Hastie *et al*. [13]. The data were originally extracted from the TIMIT database (TIMIT Acoustic-Phonetic Continuous Speech Corpus, NTIS, US Dept of Commerce) and contains 4509 labeled samples of log-periodograms with length 256 representing the five phonemes:

- group 1: 'aa' (695 samples)
- group 2: 'ao' (1022 samples)
- group 3: 'dcl' (757 samples)
- group 4:'iy' (1163 samples)
- group 5: 'sh' (872 samples)

As in Indahl *et al*. [3] we impose an 'artificial' context for this data set by introducing the prior distribution

$$\Pi = \{\pi_1 = 0.47, \pi_2 = 0.47, \pi_3 = 0.02, \pi_4 = 0.02, \pi_5 = 0.02\} \tag{13}$$

The priors are chosen to increase focus on the two phonemes that are hardest to distinguish from one another. Accordingly the data set is split into a training set of 4009 samples and a test set of 500 samples. The test set contains 235 samples from each of the groups 1 and 2, and 10 samples from each of the groups 3, 4, and 5, to reflect the specified prior distribution of Equation (13). From the training data we compute models with up to 15 components according to the following four different strategies:

1. PLS-DA based on empirical prior probabilities from the training data.
2. PLS-DA based on the specified priors (Equation 13).
3. CPLS based on the specified priors (Equation 13).
4. CPPLS based on the specified priors (Equation 13).

For all strategies we applied linear discriminant analysis (LDA) with the corresponding scores as features and inclusion of the priors in Equation (13). Tenfold cross-validation and validation by the test set were used for all models. To calculate the cross-validation success rates, the contribution of each correctly classified phoneme was weighted according to the specified prior probability of its corresponding group. The results of the four strategies are shown in Figure 1.

By including five or more components there is not much difference between the methods. However, for sparse models with less than five components, the feature extraction based on CPLS is clearly preferable to PLS-DA. Already with two components a fairly good model (78.8% correct classification on test data) is obtained by CPLS. CPPLS looks second best among the methods, but an extra component (three in all) is required for comparable performance (79.0% correct classification on test data). PLS-DA using the specified prior probabilities for extraction of the components requires four components (79.0% correct on the test data). Finally PLS-DA only using empirical priors for extraction of the components requires five components (79.8% correct classification on test data).

4.2. Classification with additional responses

To illustrate the modeling with additional responses we have chosen a data set consisting of raw NIR measurements (351 wavelengths, 1100–2500 nm in steps of 4 nm) taken on 54 samples of mayonnaise based on six different oil types (soybean, sunflower, canola, olive, corn, and grapeseed). The samples were produced according to a 2^{4-1}, res IV design, with center points, varying the amounts of oil, stabilizer, eggs, and sugar. The resulting 54 samples were measured in triplicates, resulting in $54 \times 3 = 162$ different spectra. A classification study for this data set was first presented in Indahl *et al*. [14]. Here we have randomly split the data into a training set of $40 \times 3 = 120$ samples and a test set of $14 \times 3 = 42$ samples (triplicates always together) with the following group distribution:

Copyright © 2009 John Wiley & Sons, Ltd.

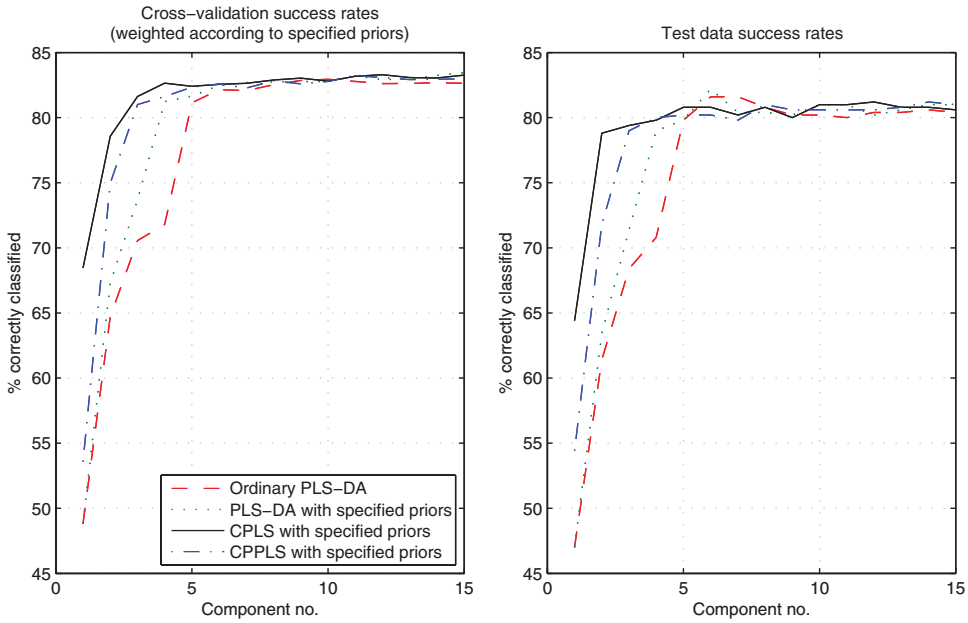

Figure 1. [Speech data] Classification results for components extracted by PLS-DA using empirical priors (dashed), PLS-DA using specified priors (dotted), CPLS using specified priors (solid), and CPPLS using specified priors (dot-dashed). This figure is available in color online at www.interscience.wiley.com/journal/cem

- group 1: 'Soybean' (30 training samples and 12 test samples)
- group 2: 'Sunflower' (18 training samples and 6 test samples)
- group 3: 'Canola' (15 training samples and 9 test samples)
- group 4: 'Olive' (12 training samples and 12 test samples)
- group 5: 'Corn' (24 training samples and 0 test samples)
- group 6: 'Grapeseed' (21 training samples and 3 test samples)

In the modeling, we have compared the LDA-success rates based on up to 20 components obtained by:

1. PLS-DA with empirical priors (without inclusion of additional responses).
2. CPLS with empirical priors (without inclusion of additional responses).
3. CPLS with empirical priors (including the five design parameters as additional responses).
4. CPPLS with empirical priors (without inclusion of additional responses).
5. CPPLS with empirical priors (including the five design parameters as additional responses).

The results of the five strategies are shown in Figures 2–5.

Both with and without inclusion of additional responses, the initial CPLS-components discriminate much better than the components derived by PLS-DA. With *five* CPLS-components including additional responses, we obtain cross-validated and test set success rates of 96.7 and 100%, respectively. Similar success rates were found for CPLS without additional responses (95.8 and 100% after *nine* components, and for PLS-DA 96.7 and 100% after *thirteen* components).

In contrast to the phoneme results, components extracted from the CPPLS approaches simplifies the classification significantly. Without additional responses, the cross-validated and test set success rates are 94.2 and 100% with *five* component modeling. With inclusion of additional responses we obtain cross-validated and test set success rates of 95.8 and 100%, respectively, by *two* component modeling. Figure 3 shows the different associated score plots.

In Figure 4, the first two vectors of loading weights for PLS-DA and CPPLS without and with additional responses are shown. The first loading weight vector from both the powered models show distinct peaks around 1700 nm and between 2200 and 2400 nm. The PLS-DA loadings on the other hand indicate no distinct focus on particular predictors. The improved focus of the CPPLS models corresponds to the fact that γ-values quite close to 0 (emphasizing wavelengths of large variance) dominated the computation of the first component. Figure 5 illustrates the γs found to be optimal for the different components of the two CPPLS approaches. The γ-values close to 0 or 1 correspond to the more focused vectors of loading weights. Regarding model interpretation, the more focused loadings resulting from the powered methodology may lead to significant simplifications compared to the traditional PLS methods.

4.3. Regression with several continuous responses

To illustrate modeling with several continuous responses we analyze a data set where the predictors are raw NIR measurements (700 wavelengths, 1100–2498 nm in steps of 2 nm) measured on biscuit dough. The calibration set has $N = 40$ samples of $p = 700$ variables. For each sample, four response variables representing percentages of *fat*, *sucrose*, *flour*, and *water*, respectively, have been measured. A corresponding set of 32 samples is reserved for testing of the candidate models. The two sets have been created and measured on different occasions. Further descriptions and modeling based on this data set are reported in Brown *et al.* [15]. We compare the Root Mean Squared Error of Cross-Validation (10-fold cross-validation) and Prediction (RMSECV and RMSEP) for different modeling strategies including up to 20 components for each of the response variables. The modeling strategies are:

 Copyright © 2009 John Wiley & Sons, Ltd.

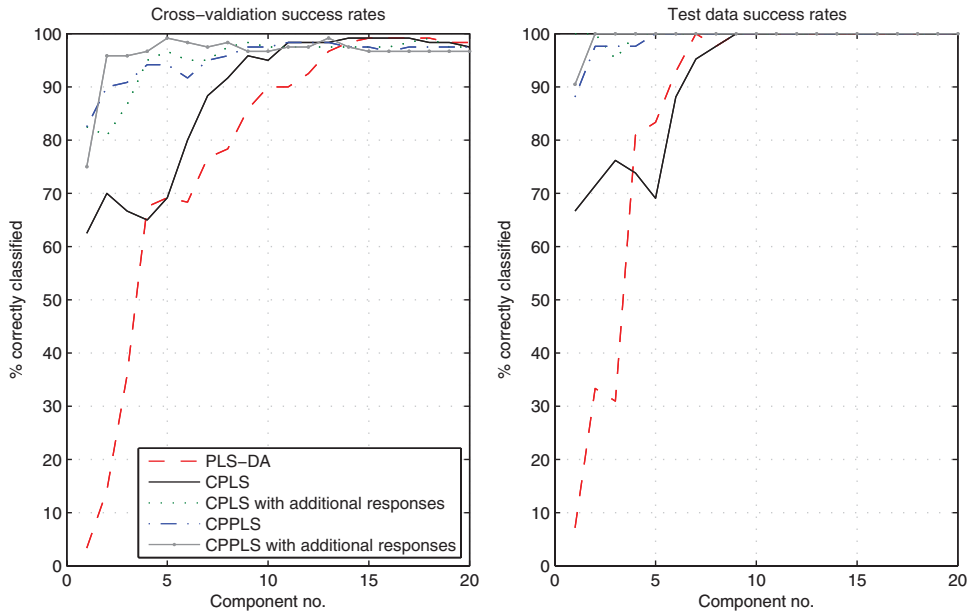

Figure 2. [Mayonnaise data] Classification results for components extracted by PLS-DA (dashed), CPLS without additional responses (solid), CPLS including additional responses (dotted), CPPLS without additional responses (dot-dashed), and CPPLS including additional responses (solid with dots). All methods use the empirical priors. This figure is available in color online at www.interscience.wiley.com/journal/cem

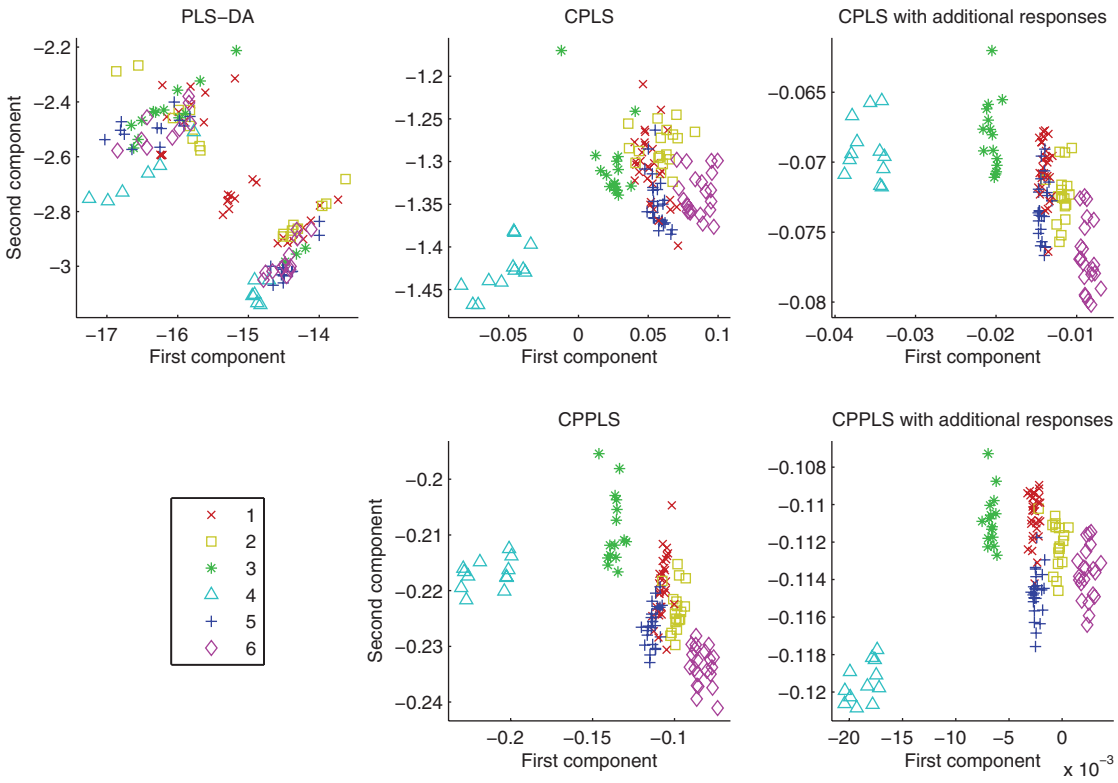

Figure 3. [Mayonnaise data] Score plots for components extracted by PLS-DA, CPLS without additional responses, CPLS including additional responses, CPPLS without additional responses, and CPPLS including additional responses. All methods use the empirical priors. This figure is available in color online at www.interscience.wiley.com/journal/cem

Copyright © 2009 John Wiley & Sons, Ltd.

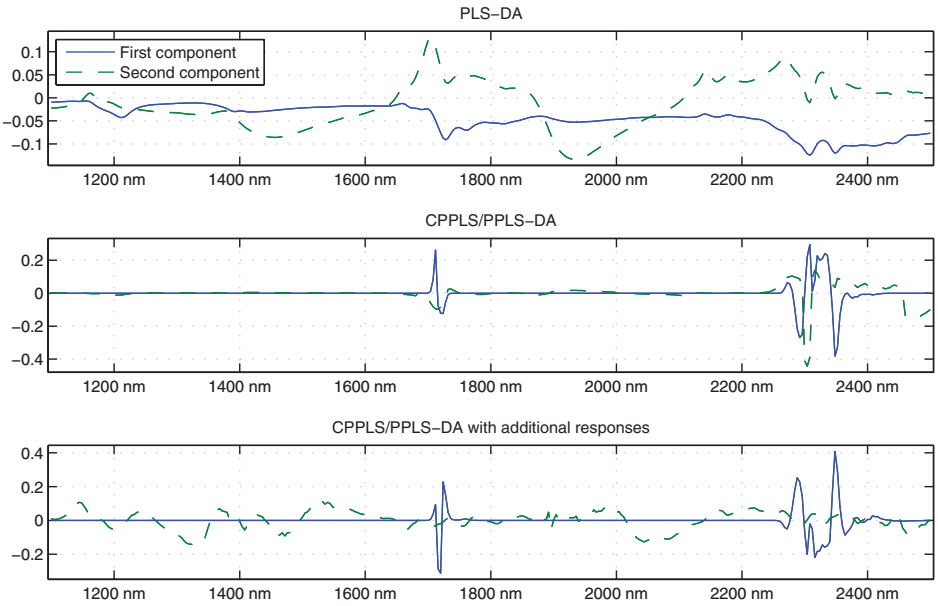

Figure 4. [Mayonnaise data] Loading plots for the first two components extracted by PLS-DA, canonical powered PLS without additional responses, and canonical powered PLS including additional responses. All methods use the empirical priors. This figure is available in color online at www.interscience.wiley.com/journal/cem

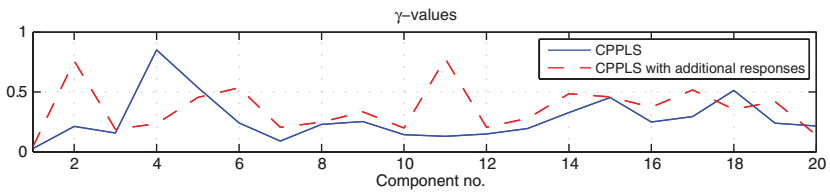

Figure 5. [Mayonnaise data] γ-values from the models for canonical powered PLS without additional responses (solid) and canonical powered PLS including additional responses (dashed). Both methods use the empirical priors. This figure is available in color online at www.interscience.wiley.com/journal/cem

1. PLS1 with separate modeling for each of the four response variables.
2. CPLS with separate modeling for each of the four response variables as the primary response and the other three response variables as additional responses.
3. CPPLS with separate modeling for each of the four response variables as the primary response and the other three response variables as additional responses. (To put focus on the predictors most correlated to the response, the domain for optimization of the power parameter γ is restricted to the interval [0.9, 1].)

Prediction results and regression coefficients for the three strategies are shown in Figures 6 and 7.

The results confirm the impression that the CPLS including additional responses in the modeling identifies good models with fewer components than ordinary PLS1. Furthermore it avoids instabilities indicated by the test data for the two component PLS1 models. With an exception for the fat response, CPPLS performs slightly better than CPLS in prediction of the test data. The regression coefficients (see Figure 7) of CPPLS are also more focused on particular predictors (from 1900 to 2100 nm and 2000 to 2200 nm) compared to the other two methods for all the responses. This is caused by restricting the γ-domain to [0.9, 1] in the CPPLS modeling. The associated powers of the standard deviation block and the correlation block are then $<\frac{1}{9}$ and >9, respectively and these restrictions forces a sharpened focus on variables in **X** with the largest correlations to the response variable.

5. DISCUSSION/CONCLUSIONS

In summary, we consider the following aspects of the CPLS methodology to have particular importance: (1) CPLS as a generalization of the traditional single response PLS. (2) The ability of CPLS to extract good components for both regression and classification problems. (3) The possibility of exploiting additional responses for more powerful modeling when such data are available. (4) Further extensions from CPLS to CPPLS.

(1) The CPLS is a genuine generalization of single response PLS (PLS1). With a single primary response (continuous or two group categorical) and no additional responses, the mathematical formulation of the CPLS simplifies to the classical PLS1 because the former is based on forming linear

 Copyright © 2009 John Wiley & Sons, Ltd.

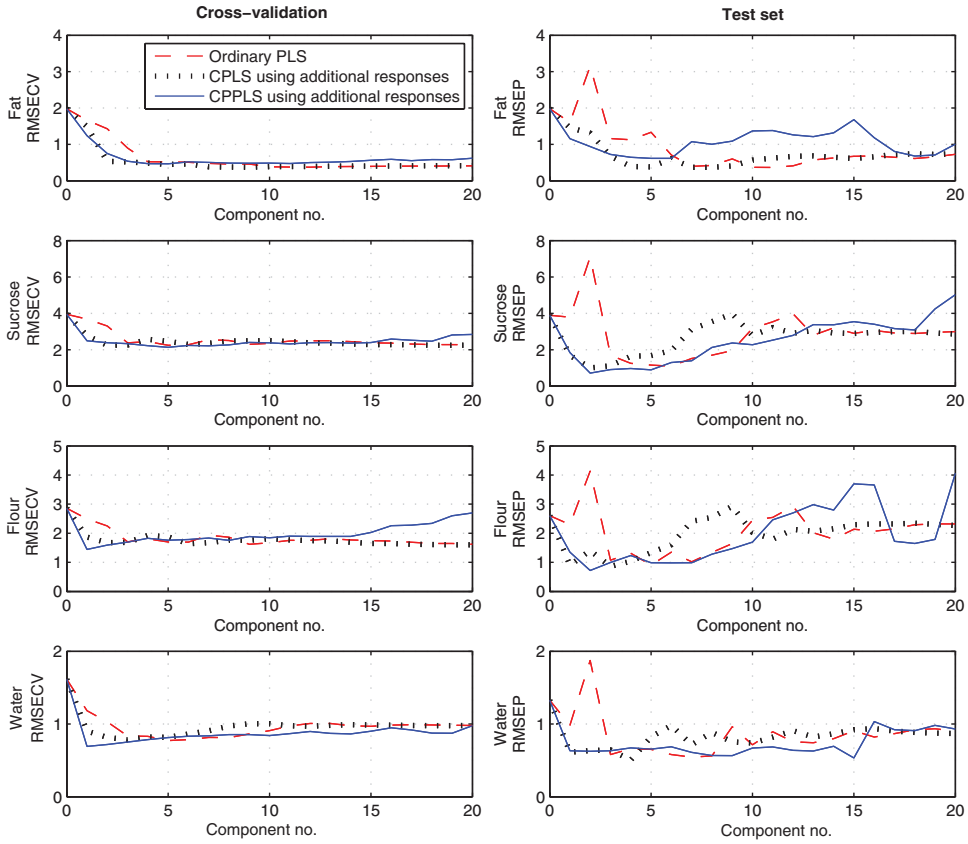

Figure 6. [Dough data] Prediction results for each of the four responses. Components extracted by PLS1 (dashed), CPLS with one primary response and inclusion of the other three reference variables as additional responses (solid), and CPPLS with one primary response and inclusion of the other three reference variables as additional responses (dotted). This figure is available in color online at www.interscience.wiley.com/journal/cem

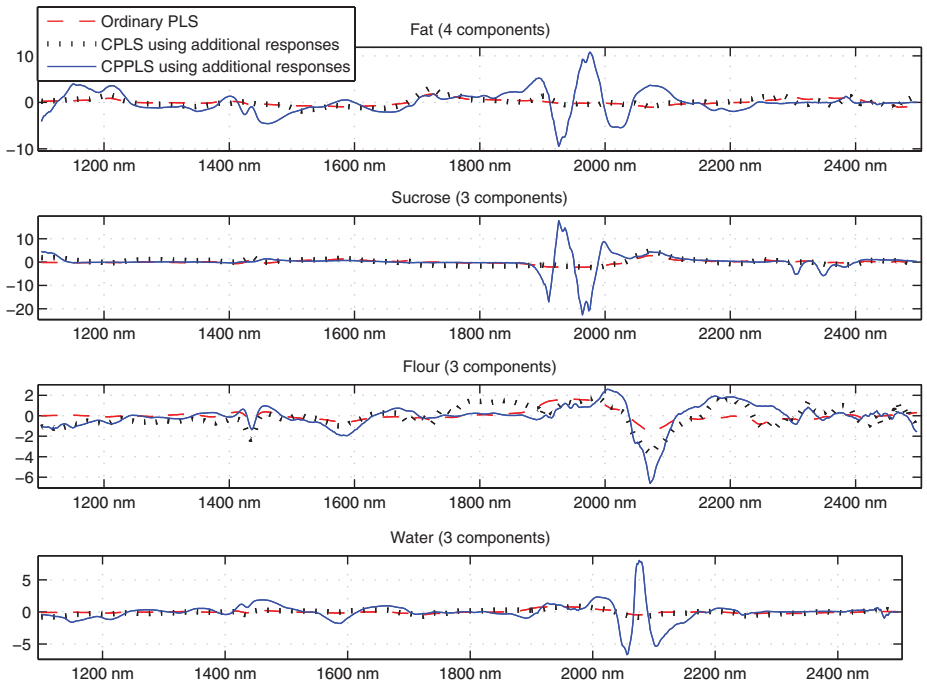

Figure 7. [Dough data] Regression coefficients for each of the four responses corresponding to the test data in Figure 6. Components extracted by PLS1 (dashed), CPLS with one primary response and inclusion of the other three reference variables as additional responses (solid), and CPPLS with one primary response and inclusion of the other three reference variables as additional responses (dotted). This figure is available in color online at www.interscience.wiley.com/journal/cem

Copyright © 2009 John Wiley & Sons, Ltd.

combinations of the columns in $\mathbf{W} = \mathbf{X}^t\mathbf{Y}$ and the fact that the number of columns in **W** and **Y** in this case are identical to 1.

(2) The temporary loading weights matrix $\mathbf{W} = \mathbf{X}^t\mathbf{Y}$ defines directions maximizing the covariances between the **X** data and the individual **Y** responses according to classical PLS1. As linear combinations of the columns in **W**, the CPLS components (and the associated loading weights) retains a close relationship to the established PLS methodology. The transformation $\mathbf{Z} = \mathbf{XW}$ temporarily maps **X** to a subspace where serious collinearity problems are avoided. The computations required for optimization can be executed effectively because the canonical loading weights of CPLS are obtained by scaling $\boldsymbol{w} = \mathbf{W}\boldsymbol{a}$ to unit length. Here the coefficient vector $\boldsymbol{a}$ is found by solving a canonical correlation with not more than q variables in either of the involved matrices. The examples analyzed above indicate that CPLS is more effective than the traditional PLS methods in the sense that simpler models (fewer components) are required to obtain good predictions. Because the method provides optimal components for both continuous and categorical response variables, it can also be applied without modifications when the columns of **Y** contain a mixture of the two types.

(3) Inclusion of additional responses for the purpose of predicting a set of primary responses is unique to CPLS. With CPLS it is possible to include directly a broader context of the particular prediction problem (such as background reference measurements, design factors from experimental designs, fitted values from prediction models of $\mathbf{Y}_{\text{prim}}$ based on measurements not included in **X**, etc.) in the model building. The examples of classification including additional responses and regression with several continuous responses indicate that using additional information from the data generation process has the potential of contributing to the building of simpler and more stable models. Note that if the matrix of available additional responses has many columns compared to the number of rows (observations), some action (PCA or other suitable data compression techniques) must be taken to prevent collinearity problems. Finally, we stress the fact that additional information (possibly unavailable or too 'expensive') is not required for the samples only to be used for testing or application of a model based on the CPLS methodology. Only (**X**, $\mathbf{Y}_{\text{prim}}$)-data is required for model testing, and for prediction only **X**-measurements are required.

(4) The extension to CPPLS incorporates the basics of the PPLS methodology for computation of prediction models. For the mayonnaise data, CPPLS including additional responses gave the most efficient models with respect to the number of required components (2) and complexity of the associated loading weights. It should also be noted that with appropriate restrictions of the most general formulation including additional responses, each of the following
   - PLS1 (fixed $\gamma = 0.5$ and $\mathbf{Y} = \mathbf{Y}_{\text{prim}} = \boldsymbol{y}$)
   - CPLS (fixed $\gamma = 0.5$)
   - PPLS ($\mathbf{Y} = \mathbf{Y}_{\text{prim}} = \boldsymbol{y}$)
   - PPLS-DA($\mathbf{Y} = \mathbf{Y}_{\text{prim}} = [\boldsymbol{y}_1 \ \ldots \ \boldsymbol{y}_g]$)

   can be considered as a special case of CPPLS.

The indicated differences between the various PLS approaches focused in this paper are partly empirical. More research is required to establish valid theoretical guidelines for selection between the alternatives. A particular issue of interest is the consequence of modifying the various PLS algorithms (deflation strategies) according to CPLS. Simulation results indicate the possibility that the CPLS versions of SIMPLS and NIPALS (with multiple responses) may lead to equivalent solutions.

REFERENCES

1. Barker M, Rayens W. Partial least squares for discrimination. *J. Chemometr.* 2003; **17**: 166–173.
2. Nocairi H, Qannari EM, Vigneau E, Bertrand D. Discrimination on latent components with respect to patterns. Application to multicollinear data. *Comput. Stat. Data Anal.* 2005; **48**: 139–147.
3. Indahl U, Martens H, Næs T. From dummy regression to prior probabilities in PLS-DA. *J. Chemometr.* 2007; **21**: 529–536.
4. Indahl U. A twist to partial least squares regression. *J. Chemometr.* 2005; **19**: 32–44.
5. Liland KH, Indahl UG. Powered PLS discriminant analysis. *J. Chemometr.* 2009; **23**(1): 7–18.
6. Martens H, Næs T. *Multivariate Calibration*. John Wiley and Sons: Chichester, UK, 1989.
7. Martens H, Martens M. *Multivariate Analysis of Quality. An Introduction*. John Wiley and Sons: Chichester, UK, 2001.
8. Wold S, Sjöström M, Eriksson L, PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab. Syst.* 2001; **58**: 109–130.
9. de Jong S. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* 1993; **42**: 251–263.
10. Burnham AJ, Viveros R. Frameworks for latenent variable multivariate regression. *J. Chemometr.* 1996; **10**: 31–45.
11. Mardia KV, Kent JK, Bibby JM. *Multivariate Analysis*. Academic Press: New York, 1979.
12. Bartlett MS. Further aspects of the theory of multiple regression. *Proceedings of the Cambridge Philosophical Society*, 1938; **34**: 33–40.
13. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer: New York, 2001.
14. Indahl U, Sahni NS, Kirkhus B, Næs T. Multivariate strategies for classification based on NIR-spectra—with application to mayonnaise. *Chemometr. Intell. Lab. Sys.* 1999; **49**: 19–31.
15. Brown PJ, Fearn T, Vannucci M. Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. Am. Stat. Assoc.* 2001; **96**: 398–408.

 Copyright © 2009 John Wiley & Sons, Ltd.