Orthogonality constrained inverse regression to improve model selectivity and analyte predictions from vibrational spectroscopic measurements

Skou, P.B.; Hosseini, E.; Ghasemi, J.B.; Smilde, A.K.; Eskildsen, C.E.

DOI
10.1016/j.aca.2021.339073

Publication date
2021

Document Version
Final published version

Published in
Analytica Chimica Acta

Analytica Chimica Acta 1185 (2021) 339073

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Orthogonality constrained inverse regression to improve model selectivity and analyte predictions from vibrational spectroscopic measurements

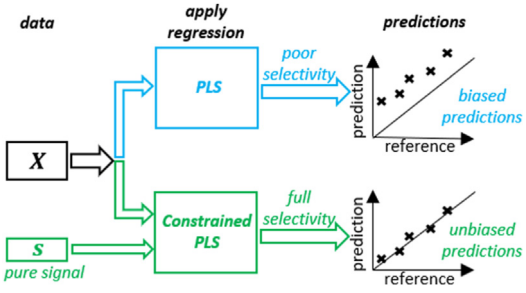
Peter B. Skou [a], Ensie Hosseini [b], Jahan B. Ghasemi [b], Age K. Smilde [c], Carl Emil Eskildsen [d, *]

[a] Arla Foods Ingredients Group P/S, DK-6920, Videbæk, Denmark
[b] School of Chemistry, College of Science, University of Tehran, IR-1417614411, Tehran, Iran
[c] Swammerdam Institute for Life Sciences, University of Amsterdam, NL-1098 XH, Amsterdam, the Netherlands
[d] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Science Park 904, NL-1098 XH, Amsterdam, the Netherlands

HIGHLIGHTS

- PLS model selectivity is improved by constrained regression vector estimation.
- The constraint handles known interfering signals explicitly.
- Predictions are improved by applying the constraint.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 16 August 2021
Received in revised form
14 September 2021
Accepted 15 September 2021
Available online 20 September 2021

Keywords:
Inverse regression
PLS
NIPALS
Selectivity
Orthogonality constraint
Infrared spectroscopy

ABSTRACT

In analytical chemistry spectroscopy is attractive for high-throughput quantification, which often relies on inverse regression, like partial least squares regression. Due to a multivariate nature of spectroscopic measurements an analyte can be quantified in presence of interferences. However, if the model is not fully selective against interferences, analyte predictions may be biased. The degree of model selectivity against an interferent is defined by the inner relation between the regression vector and the pure interfering signal. If the regression vector is orthogonal to the signal, this inner relation equals zero and the model is fully selective. The degree of model selectivity largely depends on calibration data quality. Strong correlations may deteriorate calibration data resulting in poorly selective models. We show this using a fructose-maltose model system. Furthermore, we modify the NIPALS algorithm to improve model selectivity when calibration data are deteriorated. This modification is done by incorporating a projection matrix into the algorithm, which constrains regression vector estimation to the *null*-space of known interfering signals. This way known interfering signals are handled, while unknown signals are accounted for by latent variables. We test the modified algorithm and compare it to the conventional NIPALS algorithm using both simulated and industrial process data. The industrial process data consist of mid-infrared measurements obtained on mixtures of beta-lactoglobulin (analyte of interest), and alpha-lactalbumin and caseinoglycomacropeptide (interfering species). The root mean squared error of beta-lactoglobulin (% w/w) predictions of a test set was 0.92 and 0.33 when applying the conventional and the modified NIPALS algorithm, respectively. Our modification of the algorithm returns simpler models

* Corresponding author.
E-mail address: c.e.eskildsen@uva.nl (C.E. Eskildsen).

https://doi.org/10.1016/j.aca.2021.339073

with improved selectivity and analyte predictions. This paper shows how known interfering signals may be utilized in a direct fashion, while benefitting from a latent variable approach. The modified algorithm can be viewed as a fusion between ordinary least squares regression and partial least squares regression and may be very useful when knowledge of some (but not all) interfering species is available.

1. Introduction

Vibrational spectroscopic techniques, including mid-infrared spectroscopy (MIRS), are attractive to both academia and industry, compared to traditional wet chemical analysis. Vibrational spectroscopy offers low cost and high-throughput analysis, and problematic chemicals are most often not needed [1]. Most spectroscopic applications rely on advanced inverse calibration methods such as partial least squares (PLS) regression [2]. In pursuit of obtaining predictions of multiple analytes simultaneously in complex systems, using rapid spectroscopic methods coupled with advanced inverse regression methods, the importance of calibration model selectivity becomes apparent [3–5]. If a PLS model, predicting an analyte, is not fully selective against interferences, the analyte prediction will depend on the quantities of these interferences. This will compromise model validity and robustness [6,7]. In this paper, we show how the configuration of calibration data affects PLS model selectivity, and we propose a modification of the non-linear iterative partial least squares (NIPALS) algorithm for PLS regression, which ensures model selectivity against known interfering signals. This may result in less complex PLS models with lower prediction error uncertainties and better selectivity.

A net analyte signal (NAS) is a part of an analyte signal (at unit concentration) that is orthogonal to signals of all coexisting interferences (i.e., a NAS is the part of an analyte signal that is in the *null*-space of all interferences) [8,9]. Therefore, the size of a sample measurement projected onto the NAS is directly related to the sample's analyte quantity. There are two aspects to a NAS, namely direction and size, which are affected by the pure signals of the analyte and all interfering species. If the estimated regression vector has the direction of the NAS, the regression model is fully selective against interfering species in the calibration data [10–12]. Regression model selectivity is further elaborated in section 2.1.

When calibrating a PLS model, great care must be taken during calibration data acquisition. To obtain good analyte predictions of a future sample, calibration data configuration should ideally resemble that of the future sample [4]. However, if the data configuration of future samples is inconsistent (i.e., the analyte concentration varies independent of interferences), calibration model selectivity becomes very important [4]. To calibrate a PLS model with good selectivity, a calibration data set should consist of samples with different and linearly independent combinations of quantities of analyte and interfering species. This ensures that sample measurements span a relevant space of the analyte as well as interfering species [13]. In turn, this allows estimating a PLS model with good selectivity. Nevertheless, the calibration data set may be deteriorated if the space spanned by sample measurements collapses, and this may compromise selectivity of the estimated PLS model. Several factors may deteriorate calibration data, like strong correlations between quantities of the analyte and interfering species, compounds present in quantities with relative low variation, and compounds with highly similar signals. The effect of these factors is sketched in Fig. 1. The degree to which these factors deteriorate the calibration data also depend on signal-to-noise ratio and the number of samples used for calibration.

Fig. 1 shows the *row*-space of four simulated data sets, each consisting of an analyte signal at unit concentration, $\boldsymbol{s}_y$, an interfering species signal at unit concentration, $\boldsymbol{s}_k$, the NAS and five sample *measurements,* $\boldsymbol{x}_i$. The sample *measurements* are constructed as a bilinear combination of two factor matrices, one containing compound quantities and the other containing signals, mimicking a *Beer's law* system. In Fig. 1A, the coefficient of determination, r^2, between quantities of the analyte and interferent is 0.1 and in Fig. 1B this r^2 is 0.9. Comparing Fig. 1A to B shows that increasing r^2 between quantities of the analyte and interferent will deteriorate the space spanned by $\boldsymbol{x}_i$ (i.e., the $\boldsymbol{x}_i$-space in Fig. 1B collapses and is well approximated by a one-dimensional latent space). In Fig. 1C the r^2 between quantities of the analyte and interferent is 0.1, but quantities of the interferent is only 10% as compared to data presented in Fig. 1A (i.e., the interferent has lower variation in Fig. 1C). Comparing Fig. 1A–C shows that a compound present in quantities with relatively low variation will also deteriorate the $\boldsymbol{x}_i$-space. Last, in Fig. 1D the quantities of both the analyte and interferent are identical to Fig. 1A. However, in Fig. 1D $\boldsymbol{s}_k$ has higher similarity with $\boldsymbol{s}_y$, in contrast to Fig. 1A. Comparing Fig. 1A–D demonstrates that increased similarity between signals will deteriorate the $\boldsymbol{x}_i$-space.

The signals are identical in Fig. 1A–C. Therefore, the NAS and the regression vector are identical for the systems presented in Fig. 1A–C, even though $\boldsymbol{x}_i$ differ between Fig. 1 A-C. Sample measurements in Fig. 1B and C contain less variation in the NAS-direction, as compared to samples measurements in Fig. 1A. Therefore, the regression vector may be estimated with larger uncertainties when calibrating a regression model using sample measurements presented in either Fig. 1B or 1C, as compared to samples measurements in Fig. 1A. In turn, this may compromise regression model selectivity. Using a fructose-maltose model system, we will show how PLS model selectivity is affected by increasing r^2 between quantities of the analyte and interferent (Fig. 1B) and by lowering the variance of the interferent (Fig. 1C). In Fig. 1D, the interferent signal is changed, while the analyte signal is kept constant, as compared to Fig. 1A-C. Consequently, the NAS and thereby the regression vector change direction and size. However, in this paper we will not deal with changing interfering signals.

Even though it is desirable to have calibration data, which span the space of the analyte and interfering signals, this cannot necessarily be accomplished. Consider, for example, industrial productions, where unwanted (interfering) variation is often minimized, and some industrial processes may induce covariance structures between the analyte and interferents [14]. Calibration data may also be deteriorated when e.g., monitoring chemical cascade reactions with spectroscopic measurements. In such reactions, quantities of reactants, intermediates, and products will depend on each other and spectroscopic signals of those products may be highly similar, as previously observed for photo degradation of crystal violet [15]. Moreover, strong covariance structures are also found in naturally occurring biological systems [11,16,17].

Traditionally, the problem of strong correlations between quantities of analyte and interfering species, and compounds present in low variation, is augmented with spiked samples to obtain a

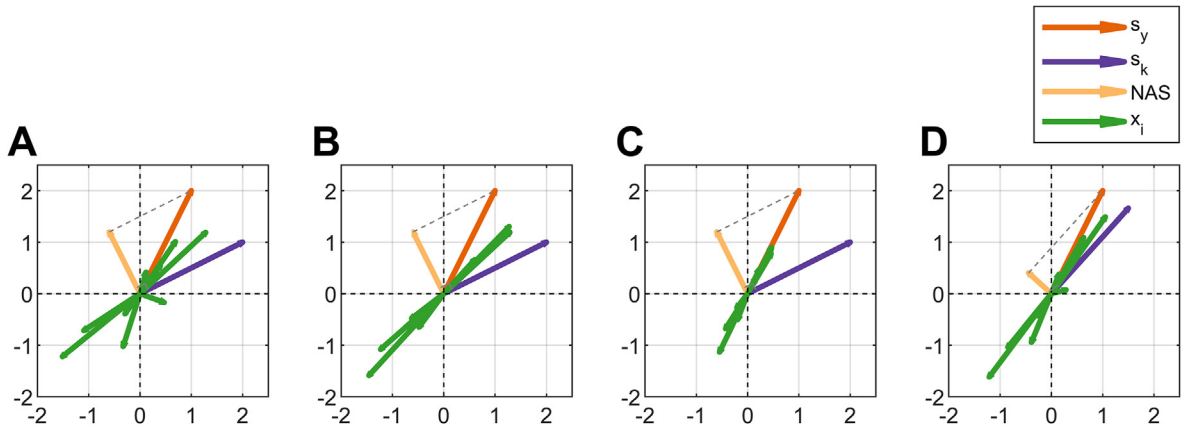


Fig. 1. Deterioration of sample measurement space (simulated data). **A**) Reference data set, **B**) increased covariance between quantities of the analyte and interferent, **C**) decreased variation of interfering quantities, and **D**) increased similarity of signals. Analyte and interfering signals at unit concentration are denoted by $\boldsymbol{s}_y$ and $\boldsymbol{s}_k$, respectively. The net analyte signal is denoted by *NAS* and $\boldsymbol{x}_i$ are combinations of $\boldsymbol{s}_y$ and $\boldsymbol{s}_k$, which mimic sample *measurements*.

better sample representation of the space spanned by the pure signals. Nevertheless, spiked samples are manipulated and may not reflect true samples in terms of background and concentration ranges, which, in turn, may compromise calibration model validity. Furthermore, the spiking procedure may be laborious. Therefore, alternatives to the spiking procedure are of our interest.

Substantial knowledge about interferents is commonly available in industrial processes and academia. This knowledge could, for example, be compound identity, pure signal, and typical concentration range. Several studies show how to utilize this information through projection-based strategies to correct the calibration data for effects of interferences prior to PLS modeling. Wold et al. [18], Andersson [19] and Trygg and Wold [20] all split the *column*-space variation into a part related to analyte quantities and a part orthogonal to analyte quantities. These methods are effective in removing variation independent from the analyte and may provide simpler models with better interpretability. However, if quantities of the analyte and interferents covaries, such corrections of the *column*-space may not be satisfying. Hansen [21] suggested a correction of the *row*-space. This correction requires acquisition of an additional data set without analyte variation. This additional data set is used to model the interference space by principal component analysis, and the calibration data with analyte variation is then projected onto the *null*-space of interferences prior to PLS modeling. An identical approach, though implemented differently, was presented by Ferré and Brown [22], and Roger et al. [23] suggested a similar approach with an alternative way of modeling the interference space. Whereas Hansen [21] predicted acetone in milk from MIRS measurements, both Ferré and Brown [22], and Roger et al. [23] used the approach to minimize the effect of temperature on spectroscopic measurements by recording a calibration data set under varying temperature conditions. These approaches work well for correcting for interferences, when it is possible to obtain the additional data that are needed.

In this paper we work along the lines of Hansen [21], Ferré and Brown [22], and Roger et al. [23] in the sense that we also control the *row*-space, used for analyte prediction, by defining a *null-space* of interference. Rather than collecting an additional data set of interference, we take advantage of known interfering signals, as also suggested by Ferré and Brown [22], and we show how the estimated regression vector can be constrained in the *null*-space of these known interferents, while latent variable (*LV*) estimation accounts for unknown interferences. Instead of implementing this as a preprocessing step, as suggest by Hansen [21], Ferré and Brown [22], and Roger et al. [23], we incorporate a projection matrix into the NIPALS algorithm, projecting all *row*-space calculations onto the *null*-space of the known interfering signal(s). This way, we improve model selectivity towards known interferents. This ultimately ensures robust models with improved analyte predictions.

We outline the orthogonality constrained inverse regression method and compare it to the conventional NIPALS PLS algorithm. We do this by using both simulated data, as well as industrial process data obtained from a whey protein fractionation process.

2. Theory

Consider a *Beer's law* system with two constituents. The independent variables are given by $\boldsymbol{X}(I \times J)$ in (1),

$$\boldsymbol{X} = \boldsymbol{c}_y \boldsymbol{s}_y^T + \boldsymbol{c}_k \boldsymbol{s}_k^T + \boldsymbol{E} \tag{1}$$

True concentrations of an analyte and an interferent are given by $\boldsymbol{c}_y(I \times 1)$ and $\boldsymbol{c}_k(I \times 1)$, respectively, analyte and interferent signals (at unit concentration) are given by $\boldsymbol{s}_y(J \times 1)$ and $\boldsymbol{s}_k(J \times 1)$, respectively, and $\boldsymbol{E}(I \times J)$ is noise. The reference measurements of the analyte (i.e., the response values) are given by $\boldsymbol{y}(I \times 1)$ in (2),

$$\boldsymbol{y} = \boldsymbol{c}_y + \boldsymbol{e} \tag{2}$$

where $\boldsymbol{e}(I \times 1)$ is noise. The predictions of $\boldsymbol{y}$, $\hat{\boldsymbol{y}}(I \times 1)$, are obtained in (3) by multiplying $\boldsymbol{X}$ and a regression vector, $\hat{\boldsymbol{b}}(J \times 1)$, estimated using calibration data.

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{b}} \tag{3}$$

2.1. Regression model selectivity

A regression model is fully selective against an interferent if analyte predictions are not affected (i.e., biased) by quantities of the interferent [4,5,7]. Substituting (1) into (3) returns,

$$\hat{\boldsymbol{y}} = \boldsymbol{c}_y \boldsymbol{s}_y^T \hat{\boldsymbol{b}} + \boldsymbol{c}_k \boldsymbol{s}_k^T \hat{\boldsymbol{b}} + \boldsymbol{E}\hat{\boldsymbol{b}} \tag{4}$$

which shows that analyte predictions are made up of a vector sum with contributions from the analyte ($\boldsymbol{c}_y \boldsymbol{s}_y^T \hat{\boldsymbol{b}}$), the interferent ($\boldsymbol{c}_k \boldsymbol{s}_k^T \hat{\boldsymbol{b}}$), and $\boldsymbol{X}$-noise ($\boldsymbol{E}\hat{\boldsymbol{b}}$). Hence, the regression model is fully selective against the interferent if the inner relation between the interferent

signal and the regression vector ($\boldsymbol{s}_k^T\widehat{\boldsymbol{b}}$) equals zero. This inner relation is zero when $\widehat{\boldsymbol{b}}$ is in the *null*-space of $\boldsymbol{s}_k$ (i.e., when $\widehat{\boldsymbol{b}}$ is orthogonal to $\boldsymbol{s}_k$). Furthermore, (4) shows that the inner relation between analyte signal (at unit concentration) and the regression vector ($\boldsymbol{s}_y^T\widehat{\boldsymbol{b}}$) is close to one for a good regression model [7,10]. If the regression model is not fully selective against the interferent (i.e., $\boldsymbol{s}_k^T\widehat{\boldsymbol{b}}\neq 0$), then the terms $\boldsymbol{s}_y^T\widehat{\boldsymbol{b}}$ and $\boldsymbol{s}_k^T\widehat{\boldsymbol{b}}$ are balanced (i.e., $\boldsymbol{s}_y^T\widehat{\boldsymbol{b}}\neq 1$) during calibration, to return predictions of the calibration data, which are on average unbiased. Here, $\boldsymbol{E}$ is assumed relatively small and random. Hence, $\boldsymbol{E}\widehat{\boldsymbol{b}}\approx\boldsymbol{0}$, where $\boldsymbol{0}(I\times 1)$ is a vector of zeros, and the term $\boldsymbol{E}\widehat{\boldsymbol{b}}$ in (4) is neglected.

In this study, regression model selectivity is determined by calculating the inner relation between the pure signals at unit concentrations (both analyte and interfering species) and the estimated regression vector. Furthermore, the root mean squared error (*RMSE*) of test set predictions is used to evaluate model predictive performance.

2.2. Orthogonality constrained regression algorithm

As mentioned, to obtain good and valid predictions of $\boldsymbol{y}$, the regression vector should have the direction of the NAS in the J-dimensional $\boldsymbol{X}$-space. However, if $\boldsymbol{s}_k$ is poorly represented in $\boldsymbol{X}$, the NIPALS algorithm (presented in Appendix A) may not be able to successfully estimate the regression vector in the *null*-space of $\boldsymbol{s}_k$. A way to constrain the regression vector estimation into this *null*-space, which will be pursued here, is by incorporating a projection matrix into the NIPALS algorithm. This projection matrix projects all *row*-space calculations of the NIPALS algorithm onto the *null*-space of $\boldsymbol{s}_k$. This ultimately ensures that the regression vector also will be in this *null*-space. Hence, $\boldsymbol{s}_k^T\widehat{\boldsymbol{b}}=0$ and the regression model is fully selective against the interferent. The modified NIPALS algorithm is outlined in the following paragraphs and presented in Appendix B.

In (5) the projection matrix, $\boldsymbol{Proj}(J\times J)$, projecting onto the *null*-space of $\boldsymbol{s}_k$ (i.e., the space orthogonal to $\boldsymbol{s}_k$), is calculated.

$$\boldsymbol{Proj}=\boldsymbol{I}-\boldsymbol{s}_k\left(\boldsymbol{s}_k^T\boldsymbol{s}_k\right)^{-1}\boldsymbol{s}_k^T \tag{5}$$

Where $\boldsymbol{I}(J\times J)$ is an identity matrix.

A vector of weights, $\boldsymbol{w}(J\times 1)$ are calculated in (6). Here a multiplication by $\boldsymbol{Proj}$ ensures that $\boldsymbol{w}$ is in the *null*-space of $\boldsymbol{s}_k$.

$$\boldsymbol{w}=\frac{\boldsymbol{ProjX}^T\boldsymbol{y}}{\boldsymbol{ProjX}^T\mathbf{y}} \tag{6}$$

Scores, $\boldsymbol{t}(I\times 1)$ are calculated in (7) following the NIPALS algorithm.

$$\boldsymbol{t}=\boldsymbol{Xw} \tag{7}$$

In (8) a vector of loadings, $\boldsymbol{p}(J\times 1)$ relating to $\boldsymbol{X}$ is calculated. Again, multiplication by $\boldsymbol{Proj}$ ensures that the loading vector is in the *null*-space of $\boldsymbol{s}_k$. In principle, multiplication by $\boldsymbol{Proj}$ can be omitted in (8), as this will not affect the regression vector estimation. However, if $\boldsymbol{p}$ is used for model interpretation, it may be advantageous to include $\boldsymbol{Proj}$ in (8).

$$\boldsymbol{p}=\frac{\boldsymbol{ProjX}^T\boldsymbol{t}}{\boldsymbol{t}^2} \tag{8}$$

In (9) a loading value, $q(1\times 1)$ relating to $\boldsymbol{y}$ is calculated following the NIPALS algorithm.

$$q=\frac{\boldsymbol{t}^T\boldsymbol{y}}{\boldsymbol{t}^2} \tag{9}$$

Finally, $\boldsymbol{X}$ is deflated in (10) following the NIPALS algorithm.

$$\boldsymbol{X}=\boldsymbol{X}-\boldsymbol{tp}^T \tag{10}$$

Step (6) to (10) are iterated for the number of LV. For each iteration in the algorithm, $\boldsymbol{w}$, $\boldsymbol{p}$ and q are stored to generate a matrix of weights, $\boldsymbol{W}(J\times LV)$, a matrix of loadings, $\boldsymbol{P}(J\times LV)$ relating to $\boldsymbol{X}$ and a vector of loadings, $\boldsymbol{q}(LV\times 1))$ relating to $\boldsymbol{y}$, and $\widehat{\boldsymbol{b}}$ is calculated in (11) following the NIPALS algorithm.

$$\widehat{\boldsymbol{b}}=\boldsymbol{W}\left(\boldsymbol{P}^T\boldsymbol{W}\right)^{-1}\boldsymbol{q} \tag{11}$$

Due to the incorporation of $\boldsymbol{Proj}$ in (6) and (8), columns of $\boldsymbol{W}$ and $\boldsymbol{P}$ are orthogonal to $\boldsymbol{s}_k$ and consequently $\widehat{\boldsymbol{b}}$ will also be orthogonal to $\boldsymbol{s}_k$. Hence, $\widehat{\boldsymbol{y}}$, obtained using (3), is independent of the interferent. If signals of more interfering species are known, $\boldsymbol{s}_k$ is substituted with $\boldsymbol{S}_k(J\times K)$ in (5), where $\boldsymbol{S}_k$ is a matrix of K known signals concatenated horizontally.

It is important to note that the additional projections suggested here operate in the *row*-space of $\boldsymbol{X}$. As $\boldsymbol{y}$ is in the *column*-space of $\boldsymbol{X}$, there is no need for processing future $\boldsymbol{x}$-measurements as a consequent of the additional projections. This is also pointed out by Ferré and Brown [22]. In principle, the projection could be implemented as a preprocessing of $\boldsymbol{X}$, as suggested by Hansen [21], Ferré and Brown [22], and Roger et al. [23], using $\boldsymbol{XProj}$, rather than $\boldsymbol{X}$, as input to the conventional NIPALS algorithm. This would return the same regression vector as using the algorithm presented above. However, as a projection is a vector-wise operation, there is a computational advantage of incorporating $\boldsymbol{Proj}$ into the algorithm if the number of estimated $LV\ll I/2$. MATLAB code for the algorithm is presented in Appendix B.

3. Materials and methods

3.1. Fructose-maltose model system

A two-constituent model system was made to investigate how calibration data configuration affect PLS model selectivity. The model system was prepared with Fructose as analyte of interest and maltose as interfering species.

Two stock solutions were prepared: one stock solution with D-(−)fructose (Chem Lab NV, Zedelgem, Belgium) and one stock solution with maltose monohydrate (Merck, Darmstad, Germany). Fructose and maltose were dissolved in water to make the two stock solutions of concentrations 1.7 and 0.87 mol/L, respectively. Samples were prepared by mixing varying amounts of the two stock solutions and adding water to a total volume of 2 ml. In total, 75 samples were included and divided into three different calibration data sets (Fig. 2).

Each of the three calibration data sets consist of 25 samples (i.e., $I=25$). The descriptive statistics for the data sets are presented in Table 1.

The mean values for both fructose and maltose, and the variance of fructose are comparable across the three calibration data sets. However, calibration data set 2 has higher r^2 between quantities of fructose and maltose, as compared to calibration data set 1 and 3. Moreover, calibration data set 3 has lower variance of maltose (interfering compound), as compared to calibration data set 1 and 2 (Table 1). Hence, calibration data set 2 and 3 are expected to be deteriorated as compared to calibration data set 1 (recall Fig. 1A–C).

Mid-infrared spectroscopic measurements were obtained on all

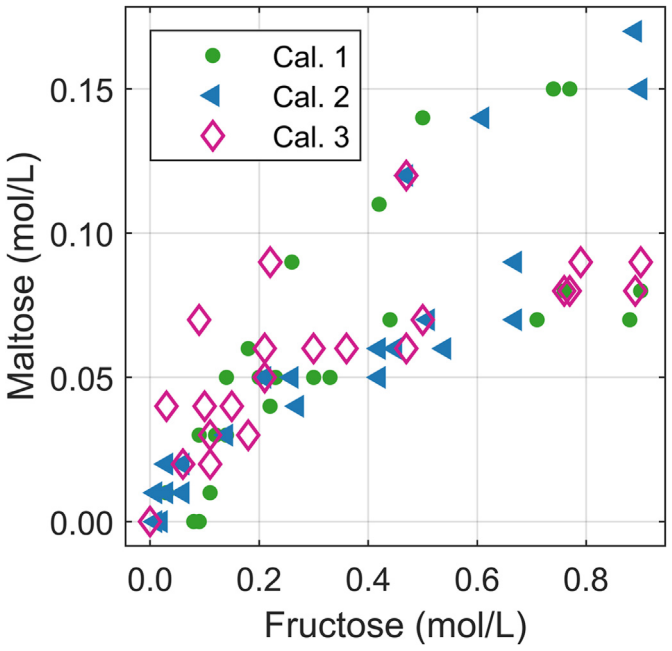


Fig. 2. Reference values for calibration data, fructose-maltose model system.

samples with a WQF-520 FT-IR spectrometer (Beijing Rayleigh Analytical Instrument Corporation, Beijing, China) using a 6 cm attenuated total reflection cell. Furthermore, MIRS measurements of the stock solutions were obtained and normalized to unit concentration. The spectroscopic measurements were obtained in duplicates. The average spectra of duplicates were used for further analysis. The spectroscopic measurements were obtained in the wavenumber range from 4,000 to 400 cm^{-1} with a spectral resolution of 2 cm^{-1}. However, only the region from 1200 to 1000 cm^{-1} was included.

The MIRS measurements were transformed from reflectance (R) units to absorbance (A) units ($A \approx \log(1/R)$) and preprocessed by Savitzky-Golay first derivative with a second order polynomial fitting and a window size of 31 data points [24,25]. The preprocessed spectral variables, as well as fructose quantities, were centered prior to calibrating PLS models. Model complexity was determined by random subset cross-validation with five data splits and 50 iterations. The minimum *RMSE*, averaged over all iterations, was used to identify optimal model complexity.

3.2. Simulated data

MATLAB code for data simulation is presented in Appendix C.

Data are simulated to mimic a *Beer's law* system with two constituents (one analyte and one interferent). One calibration set and two test sets are simulated. The calibration data consist of 100 samples (i.e., $I = 100$), while the two test sets each consist of 25 samples (i.e., $I = 25$). The simulated data were used as a proof of concept. Therefore, the number of calibration samples was intentionally kept relatively low to challenge model calibration and force poor selectivity of the conventional NIPALS algorithm.

Concentrations of the analyte, $\boldsymbol{c}_y(I \times 1)$ are randomly drawn from a uniform distribution on the open interval (0,1) and subsequently centered around zero and scaled to unit variance. One random draw defines $\boldsymbol{c}_y$ in the calibration set and another random draw defines $\boldsymbol{c}_y$ in both test sets. Hence, $\boldsymbol{c}_y$ is identical for both test sets.

Concentrations of the interferent, $\boldsymbol{c}_k(I \times 1)$ are simulated by randomly drawing I numbers from a uniform distribution on the open interval (0,1). These numbers are then adjusted (for computational details, see Appendix C) to generate a r^2 of 0.85 between $\boldsymbol{c}_y$ and $\boldsymbol{c}_k$. Hence, the calibration data are deteriorated (recall Fig. 1). For the calibration data, $\boldsymbol{c}_k$ is centered around zero. For the two test sets, $\boldsymbol{c}_k$ is generated based on the same draw but for test set 1, $\boldsymbol{c}_k$ is centered around zero, whereas for test set 2, $\boldsymbol{c}_k$ is centered around one. For all data sets, $\boldsymbol{c}_k$ is scaled to unit variance. Hence, quantities of the interferent are, as compared to the calibration set, unbiased for test set 1, whereas they are biased for test set 2. For a summary of the descriptive statistics, see Table 2.

To generate the response, $\boldsymbol{y}$ noise, $\boldsymbol{e}$ is added to $\boldsymbol{c}_y$ for all data sets following (2), where $\boldsymbol{e} \sim N(0, 0.2)$. Identical noise is added to the two test sets making $\boldsymbol{y}(I \times 1)$ of the two test sets identical.

Signals of the analyte, $\boldsymbol{s}_y(J \times 1)$ and the interferent, $\boldsymbol{s}_k(J \times 1)$ are gaussian peaks ($J = 150$) with a peak width $\sigma = 20$ and centered around 70 and 80 for $\boldsymbol{s}_y$ and $\boldsymbol{s}_k$, respectively. Both $\boldsymbol{s}_y$ and $\boldsymbol{s}_k$ are scaled to unit length.

Spectral measurements are simulated following *Beer's law* as presented in (1), where $\boldsymbol{E} \sim N(0, 0.02)$. Again, identical noise is added to the two test sets.

Partial least squares regression models were calibrated using the calibration set and tested using the two test sets. Complexity of the models were predetermined. The simulated data consist of two constituents. Hence, for conventional NIPALS PLS regression the model complexity was predetermined to two *LV*. For the orthogonality constrained regression approach, the pure signal of the interferent is actively used in the algorithm. Hence, the model complexity for the orthogonality constrained regression was predetermined to one *LV*.

Table 2
Descriptive statistics of reference values, simulated data.

Calibration data	$\mu_{\boldsymbol{c}_y}$	$\sigma^2_{\boldsymbol{c}_y}$	$\mu_{\boldsymbol{c}_k}$	$\sigma^2_{\boldsymbol{c}_k}$	$\boldsymbol{r}^2$
Calibration data ($I = 100$)	0	1	0	1	0.85
Test set 1($I = 25$)	0	1	0	1	0.85
Test set 2 ($I = 25$)	0	1	1	1	0.85

I = number of calibration samples; μ = mean value; σ^2 = variance; r^2 = coefficient of determination between quantities of analyte ($\boldsymbol{c}_y$) and interferent ($\boldsymbol{c}_k$).

Table 1
Descriptive statistics of reference values, fructose-maltose model system.

Calibration data	$\mu_{fructose}(mol/L)$	$\sigma^2_{fructose}$ (mol^2/L^2)	$\mu_{maltose}$ (mol/L)	$\sigma^2_{maltose}$ (mol^2/L^2)	$\boldsymbol{r}^2$
Calibration data 1 ($I = 25$)	0.3	0.08	0.06	$2 \cdot 10^{-3}$	0.5
Calibration data 2 ($I = 25$)	0.3	0.08	0.06	$2 \cdot 10^{-3}$	0.8
Calibration data 3 ($I = 25$)	0.3	0.08	0.06	$8 \cdot 10^{-4}$	0.5

I = number of calibration samples; μ = mean value; σ^2 = variance; r^2 = coefficient of determination between quantities of fructose and maltose.

3.3. Industrial process data

Thirty-two process samples were collected from an industrial whey protein fractionation process. Beta-lactoglobulin (β-LG) acts as analyte of interest, with alpha-lactalbumin (α-LA) and caseinoglycomacropeptide (cGMP) as interfering species. Reference quantification of β-LG (as well as α-LA and cGMP) was done using a routine in-house HPLC based method (Arla Foods Ingredients Group P/S). Eighteen of the 32 process samples were, after being measured as is, divided into three parts and spiked with a pure in-house standard (Arla Foods Ingredients Group P/S) of either α-LA, β-LG or cGMP returning 54 spiked samples. Mid-infrared transmittance measurements of all 86 (32 process plus 54 spiked) samples were obtained using a MilkoScan FT1 (Foss Analytical A/S, Hillerød, Denmark). Furthermore, MIRS measurements of pure standards of β-LG, α-LA and cGMP were obtained in duplicates and normalized to unit concentration. Each MIRS measurement was ratioed against a water background measurement. The spectroscopic measurements were obtained in the wavenumber range from 4,996 to 929 cm^{-1} with a spectral resolution of 4 cm^{-1}. However, only the region from 3,011 to 929 cm^{-1} was included. Furthermore, wavenumbers relating to *the dead sea* from 2,795 to 1,804 cm^{-1} and a region from 1,696 to 1,585 cm^{-1} (relating to O-H bend of water) were removed.

The MIRS measurements were transformed from transmittance (T) units to absorbance units ($A \approx \log(1/T)$) and preprocessed by Savitzky-Golay second derivative with a second order polynomial fitting and a window size of 11 data points [24,25]. The preprocessed spectral variables, as well as β-LG quantities, were centered prior to calibrating PLS models.

Fig. 3 shows the reference values of the industrial process data for the three proteins, β-Lg (analyte of interest), α-La and cGMP (interfering species). The industrial process data were divided into two calibration sets and one test set. The first calibration set consists of 25 process samples. The second calibration set consists of the same 25 process samples with the addition of 15 spiked samples (i.e., a total of 40 samples). The test set consists of seven process samples and 39 spiked samples.

Model complexity was determined by random subset cross-validation with five data splits and 50 iterations. The minimum *RMSE*, averaged over all iterations, was used to identify optimal model complexity. Hereafter, the optimal models were tested using the test set.

Note that samples with elevated α-La concentrations, in general, are not found in the calibration samples. Especially, the configuration of low β-Lg concentration and high α-La concentration is only found in the test set (Fig. 3A). Hence, if a PLS model predicting β-Lg is not selective towards α-La then it is likely that these samples will show biased predictions.

3.4. Software

Data were analyzed using MATLAB version R2019a (9.6.0.1072779, MathWorks Inc., Natick, MA, USA). The conventional NIPALS algorithm [26], presented in Appendix A, was compared to the orthogonality constrained NIPALS algorithm, outlined in section 2 and Appendix B.

4. Results

4.1. Fructose-maltose model system

The fructose-maltose model system is used for investigating how calibration data configuration affect PLS model quality when using the conventional NIPALS algorithm. Fig. 4A shows the preprocessed pure signals (normalized to unit concentration) of fructose and maltose, and Fig. 4B shows the preprocessed sample measurements. Fructose is a monosaccharide, whereas maltose is a disaccharide. The preprocessed pure spectra of fructose and maltose deviate in the region from 1,180 cm^{-1} to 1,140 cm^{-1} (Fig. 4A), due to the absorption of a C-O-C glycosidic bridge present in maltose [27]. Furthermore, the preprocessed pure spectra of fructose and maltose deviate in the region from 1,080 cm^{-1} to 1,000 cm^{-1} (Fig. 4A), which is due to differences in C-C and C-O stretching vibrations [28].

Table 3 summarizes results obtained from PLS models (conventional NIPALS algorithm) fitted using the three calibration data sets. Calibration data set 2 is deteriorated by increased r^2 between

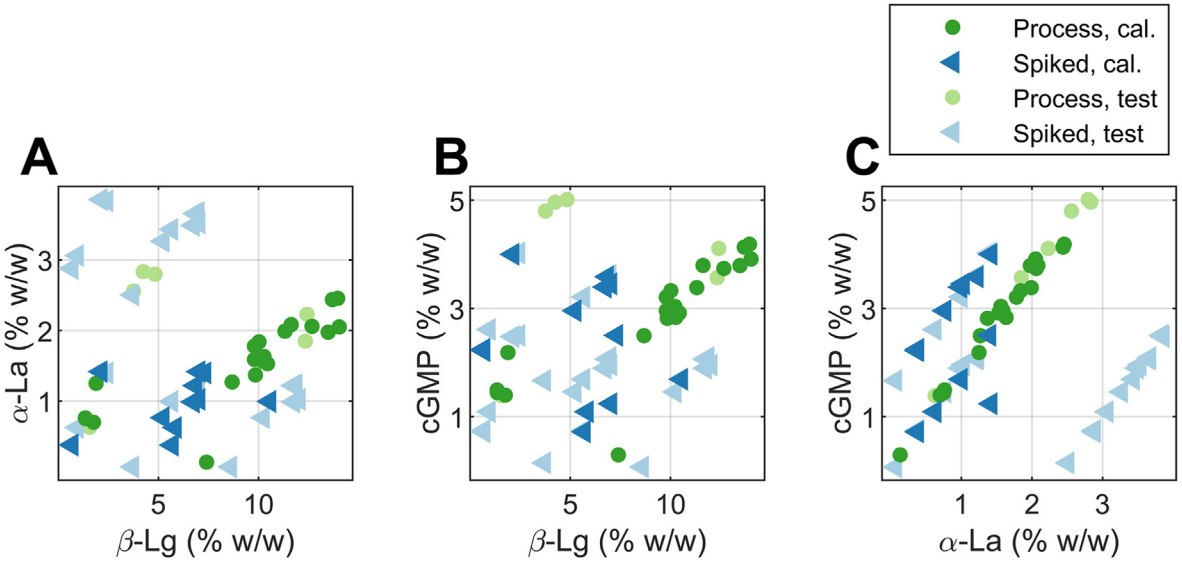


Fig. 3. Reference values (% w/w), industrial data. **A**) α-La against β-Lg. **B**) cGMP against β-Lg. **C**) cGMP against α-La. Process samples are marked with circles and spiked samples are marked with triangles. Calibration samples are darker color and test samples are lighter color. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

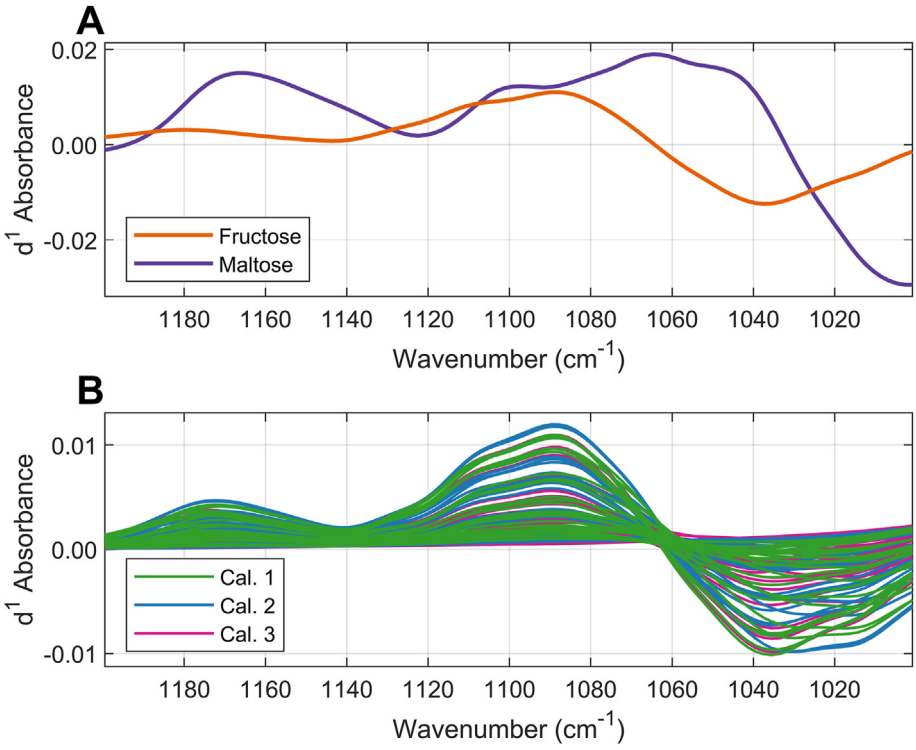


Fig. 4. Preprocessed spectra, fructose-maltose model system. **A**) fructose (analyte) and maltose (interferent) normalized to unit concentration and **B**) calibration samples.

Table 3
Regression results from fructose-maltose model system.

Calibration data	Model complexity (# *LV*)	$\boldsymbol{s}_{fructose}^{T}\hat{\boldsymbol{b}}$	$\boldsymbol{s}_{maltose}^{T}\hat{\boldsymbol{b}}$
Calibration data 1 ($I = 25$)	3	1.12	− 0.02
Calibration data 2 ($I = 25$)	2	0.91	0.51
Calibration data 3 ($I = 25$)	1	0.96	0.94

I = number of calibration samples; # *LV* = number of latent variables; $\boldsymbol{s}_{fructose}^{T}\hat{\boldsymbol{b}}$ = inner relation between fructose (analyte) signal and the estimated regression vector; $\boldsymbol{s}_{maltose}^{T}\hat{\boldsymbol{b}}$ = inner relation between maltose (interfering) signal and the estimated regression vector.

quantities of fructose and maltose, and calibration data set 3 is deteriorated by decreased variation of maltose (interferent). Recall Table 1 (section 3.1) and Fig. 1.

In principle, the fructose-maltose model system is a rank two system, but a background signal could be present in the spectroscopic measurements even after preprocessing. This may be the reason why the optimal PLS model, fitted using calibration data 1, consists of three *LV*. The optimal PLS model fitted using calibration data 2 and 3, consists of two and one *LV*, respectively. This is likely because the sample measurement spaces collapse in the calibration data 2 and 3, as also sketched in Fig. 1. Therefore, calibration data 2 and 3 may be well approximated by fewer *LV*, as compared to calibration data 1.

The inner relation between the pure fructose signal and the estimated regression vector is close to one for all models (Table 3). However, the inner relation between the pure maltose signal and the regression vector is almost zero for the model fitted to calibration data 1, whereas it is remarkably larger for models fitted to calibration data 2 and 3 (Table 3). Hence, only the model fitted to calibration data 1 is selective against maltose. It is likely that the measurement space of calibration data 1 represents the space spanned by the pure fructose and maltose signals better, as compared to the measurement spaces of calibration data 2 and 3 (recall Fig. 1). This is believed to enable regression vector estimation in the *null*-space of the maltose signal when using calibration data 1, as compared to calibration data 2 and 3.

4.2. Simulated data

The simulated data are used to compare the conventional and the orthogonality constrained NIPALS algorithms when calibration data have deteriorated. The two algorithms are compared in terms of model selectivity and hence robustness. Table 4 shows regression results obtained from the simulated data. The inner relation between the analyte signal and the estimated regression vector is 0.71 and 0.99 for the conventional and constrained NIPALS algorithm, respectively (Table 4). Furthermore, the inner relation between the interfering signal and the estimated regression vector is 0.30 and 0.00 for the conventional and constrained NIPALS algorithm, respectively (Table 4). Thus, the regression vector obtained from the constrained NIPALS is fully selective against the interferent, whereas this is not the case for the regression vector obtained by the conventional NIPALS algorithm. For the model fitted by the conventional NIPALS algorithm, the inner relations between the pure signals and the regression vector are balanced (recall section 2.1). However, this balance (and thereby the model) is solely valid when data are configurated similarly to the calibration data. Therefore, the regression model obtained by conventional NIPALS is less robust, as compared to the model obtained by the orthogonality constrained NIPALS.

This is also confirmed when evaluating the *RMSE* values of the two test sets (Table 4). When applying the conventional model, the *RMSE* value of test set 1, i.e., the test set that resamples the calibration set, is 0.20, whereas the *RMSE* value for test set 2, i.e., the test set with elevated interferent quantities, is 0.36. However, when

Table 4
Regression results from simulated data.

Calibration data	Algorithm	Model complexity (# *LV*)	$s_y^T\hat{b}$	$s_k^T\hat{b}$	Test set 1 prediction error (*RMSE*)	Test set 2 prediction error (*RMSE*)
Simulated ($I = 100$)	Conventional NIPALS	2	0.71	0.30	0.20	0.36
Simulated ($I = 100$)	Orthogonality constrained NIPALS	1	0.99	0.00	0.23	0.23

I = number of calibration samples; # *LV* = number of latent variables; $s_y^T\hat{b}$ = inner relation between analyte signal and the estimated regression vector; $s_k^T\hat{b}$ = inner relation between interfering signal and the estimated regression vector; *RMSE* = root mean squared error.

applying the constrained model, the *RMSE* values of test set 1 and 2 are identical (Table 4).

These observations are further confirmed when investigating the predictions of both test sets obtained from the conventional model (Fig. 5A) and the orthogonality constrained model (Fig. 5B). Fig. 5A shows that predictions of test set 2 are biased upwards, as compared to test set 1 when applying the conventional model. Contrary, Fig. 5B shows that predictions of the test sets 1 and 2 are identical when applying the orthogonality constrained model. Hence, the constrained model is robust towards changes in the interferent concentration, whereas this is not the case for the conventional model.

Table 4 shows that the model fitted by conventional NIPALS produces a slightly lower *RMSE* value for the test set 1, as compared to the model fitted by the constrained NIPALS algorithm. However, it should be noted that when simulations are repeated, it is interchangeable whether the conventional or constrained model produces lower *RMSE* (results not shown). Nevertheless, the *RMSE* values for the test set 1 are always in the same neighborhood for the two models. Yet, the results presented in Table 4 for test set 1 confirm that an interferent may provide support in the regression model and improve the predictions, as also outlined by Brown and Ridder [4] and Kalivas et al. [29].

4.3. Industrial process data

Fig. 6A shows the preprocessed signals of the pure proteins (normalized to unit concentration), and Fig. 6B shows the sample measurements. In Fig. 6A the difference in secondary structure can be observed from the amide II band (1,480–1,575 cm^{-1}), where α-helix information is found around 1,545 cm^{-1}, while β-sheet information is found around 1,530 cm^{-1} [30]. Since α-La consists mainly of α-helices and β-Lg of β-sheets, a small shift towards lower energy can be observed for α-La as compared to β-Lg. This is also in agreement with recent findings [31].

Table 5 shows the regression results obtained on the industrial process data. Three models are fitted. The first model is fitted by applying the conventional NIPALS algorithm to the 25 process calibration samples. The second model is fitted by applying the conventional NIPALS algorithm to the 40 process and spiked calibration samples. The third model is fitted by applying the orthogonality constrained NIPALS algorithm to the 25 process calibration samples. All models are tested on the same test set, consisting of 46 process and spiked samples.

Model complexity, estimated by cross-validation, is not consistent among the three models (Table 5). When using the conventional NIPALS algorithm, 5 *LV* are optimal for the model fitted to solely process samples, whereas 6 *LV* are optimal when both process and spiked samples are used for calibration. The spiked samples will introduce variation. Therefore, it is likely that space spanned by the pure protein signals is better represented by the calibration data consisting of both process and spiked samples, compared to the calibration data consisting of process samples only. Consequently, more *LV* can be estimated. This was also

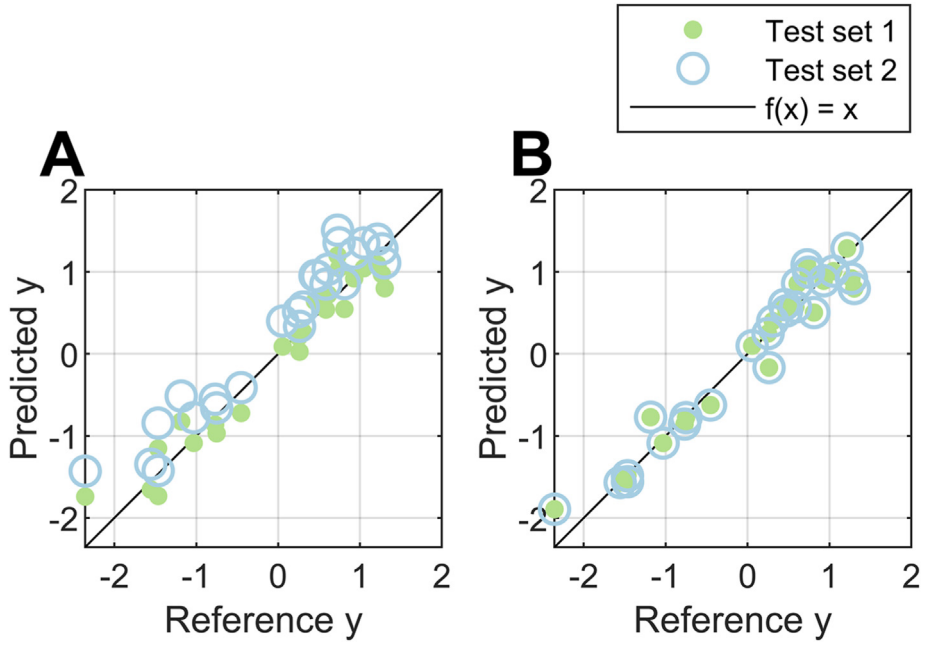


Fig. 5. Test set predictions, simulated data. Predictions obtained from **A**) conventional and **B**) orthogonality constrained NIPALS regression model.

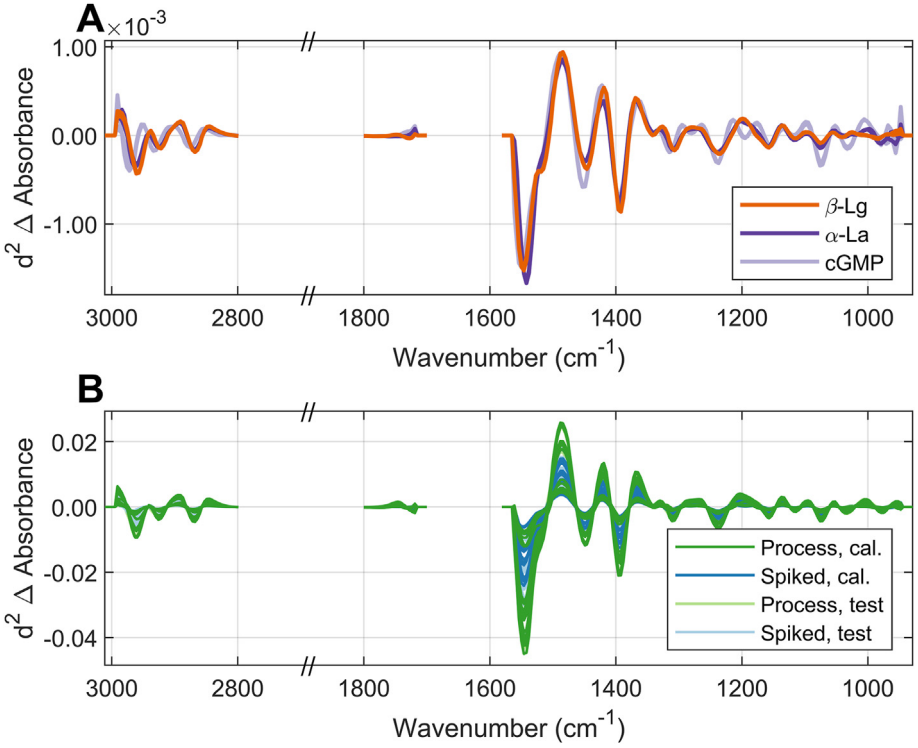


Fig. 6. Preprocessed spectra, industrial data. **A**) β-Lg (analyte), α-La (interferent) and cGMP (interferent) normalized to unit concentration and **B**) samples.

Table 5
Regression results from industrial process data.

Calibration data	Algorithm	Model complexity (# *LV*)	$\boldsymbol{s}_{\beta-Lg}^{T}\hat{\boldsymbol{b}}$	$\boldsymbol{s}_{\alpha-La}^{T}\hat{\boldsymbol{b}}$	$\boldsymbol{s}_{cGMP}^{T}\hat{\boldsymbol{b}}$	Test set prediction error (*RMSE*, % *w*/*w*)
Process ($I = 25$)	Conventional NIPALS	5	0.85	0.47	− 0.26	0.92
Process + spiked ($I = 40$)	Conventional NIPALS	6	0.87	0.26	− 0.06	0.58
Process ($I = 25$)	Orthogonality constrained NIPALS	4	0.85	0.00	0.00	0.33

I = number of calibration samples; # *LV* = number of latent variables; $\boldsymbol{s}_{\beta-Lg}^{T}\hat{\boldsymbol{b}}$ = inner relation between β-Lg (analyte) signal and the estimated regression vector; $\boldsymbol{s}_{\alpha-La}^{T}\hat{\boldsymbol{b}}$ = inner relation between α-La (interferent) signal and the estimated regression vector; $\boldsymbol{s}_{cGMP}^{T}\hat{\boldsymbol{b}}$ = inner relation between cGMP (interferent) signal and the estimated regression vector; *RMSE* = root mean squared error.

observed for the fructose-maltose model system (Table 3, section 4.1). For the constrained model, the pure spectra of α-La and cGMP are actively used. Hence, these dimensions will not be handled by including more *LV*, and therefore, the complexity of the constrained model is lower.

The inner relation between the β-Lg signal and the regression vector is approximately the same for the three models. This inner relation, which should ideally be one, is slightly higher for the model fitted to both process and spiked samples (Table 5). The reasons why this inner relation is less than expected could be multiple. For the conventional models, this inner relation is balanced with the inner relations between pure spectra of α-La and cGMP (interfering species) and the regression vector (recall section 2.1). However, this is not the case for the constrained model. The signals could also be exposed to matrix effects in the samples as well as noise, which plays a role. The intensity of the pure β-Lg signal could be underestimated. Furthermore, the least-squares effect bias will also contribute to this inner relation being less than one [32]. Another plausible explanation could be that the models find other unknown support in the calibration data, as also explained by Brown [3].

The inner relations between α-La and cGMP (interfering species) and the regression vectors (Table 5) show that both models fitted by the conventional NIPALS are sensitive towards the interfering species, whereas this is not the case for the model fitted by the constrained NIPALS algorithm. This also manifests in the prediction errors of the test set. Here, the constrained model has lower *RMSE* compared to the conventional models (Table 5).

Fig. 7 shows test set predictions of β-Lg (analyte of interest). Fig. 7A shows the *RMSE* of the test set when applying the three models. When comparing the two models fitted using the conventional NIPALS algorithm, not surprisingly, it is observed that the model calibrated to both process and spiked samples performs better than the model calibrated solely to process samples. However, it is remarkable how well the constrained model performs even though this model is calibrated only to the 25 process samples (as well as utilizing the pure spectra of the interfering species). One should recall that optimal model complexity was chosen by cross-validating the calibration data. Therefore, the complexity of the conventional model calibrated to the process samples is 5 *LV*, even though Fig. 7A shows that a 3 *LV* model returns a lower *RMSE* for the test set. Similarly, a complexity of 4 *LV* was chosen for the

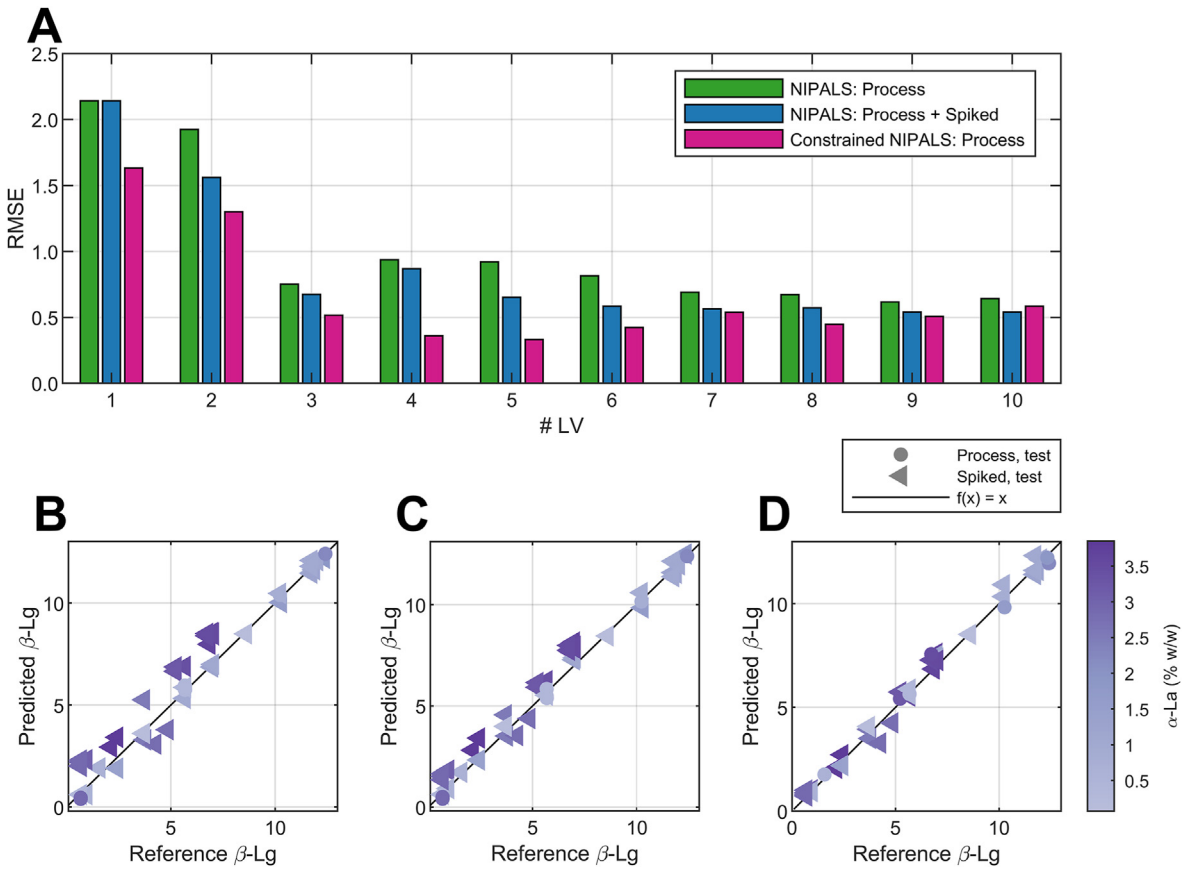


Fig. 7. Test set predictions of β-Lg (analyte), industrial data. **A**) Root mean squared error, *RMSE* of the three models. Predictions obtained by **B**) conventional NIPALS calibrated on process samples (5 latent variable, *LV* model), **C**) conventional NIPALS calibrated on process and spiked samples (6 *LV* model) and **D**) orthogonality constrained NIPALS calibrated on process samples (4 *LV* model). Predictions in **B**), **C**) and **D**) are colored by the concentration of α-LA (interferent).

constrained model, even though a 5 *LV* model returns a lower *RMSE* for the test set.

Fig. 7B–D shows the test set predictions of β-Lg (analyte of interest) colored by α-La (interfering species). Fig. 7B shows predictions by the model fitted with the conventional NIPALS algorithm calibrated using process samples. Predictions of samples with higher α-La concentrations are clearly biased upwards (Fig. 7B). Fig. 7C shows predictions by the model fitted with the conventional NIPALS algorithm calibrated using both process and spiked samples. Predictions of samples with higher α-La concentrations are still slightly biased upwards. Fig. 7D shows predictions by the model fitted with the constrained NIPALS and calibrated with the process samples. These predictions are in general unbiased. The results shown in Fig. 7 correspond well with the results for the inner products between the pure signal of α-La and the estimated regression vectors presented in Table 5. The inner relations show that the two models fitted using the conventional NIPALS are sensitive to α-La, with the model calibrated with only process samples being more sensitive. Furthermore, the constrained model is insensitive to the interfering species and this model is, therefore, to be regarded more robust.

5. Discussion

In quantitative analytical chemistry, most attention is often paid to the analyte of interest. Nevertheless, to estimate inverse regression models with good selectivity and robustness, interfering species are almost equally important. If interfering species are poorly expressed in the calibration data, the PLS NIPALS algorithm will not be able to return a regression vector in the correct *null*-space of all interfering species. Therefore, it is important to pay attention when collecting calibration data. If it is impossible to collect the calibration data, which span the space of all interfering compounds, the orthogonality constrained NIPALS algorithm proposed in this paper can be applied. The orthogonality constraint ensures the estimated regression vector to be in the *null*-space of known interfering signals.

It is important to keep in mind that the proposed orthogonality constrained algorithm is just a small and simple modification of the conventional NIPALS algorithm. Therefore, the proposed algorithm benefits from both the hard orthogonality constraint as well as the softer *LV* approach used by the conventional NIPALS algorithm. Hence, it is not necessary to know the pure signals of all interfering species to apply the orthogonality constraint. If only one or a few pure signals of interfering species are known, the constraint can be applied with the known signals, and then the latent variables will account for the remaining interferences. This also means that attention still should be paid to the selection of good calibration samples even though the orthogonality constraint is applied.

There are a few things, which must be considered when applying the orthogonality constrained algorithm. It may not always be easy to obtain good pure signals of interfering species. Nevertheless, it is important that pure interfering signals are correct and located in the sample measurement space. If the obtained interfering signals are out of the measurement space, the regression vector estimated by the orthogonality constrained NIPALS

algorithm may also be out of the sample measurement space, and interpretation of the model and predictions could be difficult.

Along these lines, applying the orthogonality constraint will give extreme importance to the pure signals of interfering species, and the entire model is adjusted according to these signals. Therefore, the entire constrained model will be exposed to uncertainties in the pure signals, whereas the conventional NIPALS gives almost equal importance (depending on leverage) to all samples. Therefore, the conventional model will be less exposed to high uncertainties in a few samples.

If the analyte signal is exposed to a poor signal-to-noise ratio, predictions obtained from the orthogonality constrained model are likely to be poor. In such situation, it may be tempting to compromise model selectivity and use any support provided by interfering species to minimize the influence of noise and thereby improve predictions, as proposed by Kalivas et al. [29].

Nevertheless, if a good signal of interfering species can be obtained and the analyte has a good signal-to-noise ratio, it may be very efficient to use the pure interfering signals more directly. As extreme importance is given to these pure signals, fewer samples are needed during calibration.

Furthermore, when applying data compression methods, like PLS, an analyte prediction is subject to both an estimation error (variance) and a model error (bias). During model calibration, increased number of *LV* will in general lead to lower model errors, but larger estimation errors. This is also known as the bias-variance tradeoff [32,33]. The model error directly relates to model selectivity [3]. Hence, during PLS model calibration, improving model selectivity comes at the expense of larger estimation errors. However, the orthogonality constrained models handle known interfering signals explicitly and *LV* will not be estimated to account for these known signals. Therefore, the constrained models have fewer *LV*, as also observed by Ferré and Brown [22]. Consequently, better model selectivity (i.e., model error) may be archived with fewer *LV*. For the deteriorated systems studied in this paper, it appears that the orthogonality constrained NIPALS PLS algorithm returns predictions with lower prediction error uncertainties, compared to the conventional NIPALS PLS algorithm.

6. Conclusions

The results in this paper highlight the importance of considering the space spanned by sample measurements when doing inverse regression modeling. The sample measurement space should be a good representation of the space spanned by the pure compound signals. The sample measurement space may be deteriorated by several factors, like strong correlations between quantities of the analyte and interfering species, and compounds present in quantities with relatively low variation. Fitting a PLS model using deteriorated calibration data may return a model with poor selectivity and robustness.

In this paper, we present a modification of the NIPALS algorithm. This modification is accomplished by incorporating a projection matrix into the NIPALS algorithm, which constrains the regression vector solution to within the *null*-space of known interfering signals. The proposed algorithm utilizes known signals of interfering compounds directly, while handling unknown interferences by estimating latent variables. This approach has the potential to improve model selectivity and thereby analyte predictions when calibration data deteriorate.

Funding

The Netherlands Organization of Scientific Research (NWO) via the TTW Open Technology Programme is acknowledged for funding C.E. Eskildsen through the TooCOLD project (grant number 15506).

Author contributions

C. E. Eskildsen is the originator of this research and is responsible for the integrity of this work, from conception and design to data analysis, interpretation of the results, and writing the manuscript. P. B. Skou, E. Hosseini, and A. K. Smilde contributed to the design, data analysis, interpretation of results, and writing of the manuscript. P. B. Skou and E. Hosseini were responsible for data acquisition. J. B. Ghasemi facilitated funding and laboratory equipment for E. Hosseini.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Elena Khachirova is acknowledged for proofreading the manuscript.

Appendices

Appendices to this article can be found online at https://doi.org/10.1016/j.aca.2021.339073.

References

[1] C.E. Eskildsen, F.v.d. Berg, S.B. Engelsen, Vibrational spectroscopy in food processing, in: J.C. Lindon, G. E Tranter, D.W. Koppenaal (Eds.), Encyclopedia of Spectroscopy and Spectrometry, third ed., Elsevier, Oxford, UK, 2017, pp. 582–589.
[2] S. Wold, A. Ruhe, H. Wold, W.J. Dunn III, The collinearity problem in regression, the partial least squares approach to generalized inverses, SIAM J. Sci. Stat. Comput. 5 (1984) 735–743.
[3] C.D. Brown, Discordance between net analyte signal and practical multivariate calibration, Anal. Chem. 76 (2004) 4364–4373.
[4] C.D. Brown, T.D. Ridder, Framework for multivariate selectivity analysis, part I: theoretical and practical merits, Appl. Spectrosc. 59 (2005) 787–803.
[5] T.D. Ridder, C.D. Brown, B.J.V. Steeg, Framework for multivariate selectivity analysis, part II: experimental applications, Appl. Spectrosc. 59 (2005) 804–815.
[6] C.E. Eskildsen, T. Næs, P.B. Skou, L.E. Solberg, K.R. Dankel, S.A. Basmoen, J.P. Wold, S.S. Horn, B. Hillestad, N.A. Poulsen, M. Christensen, T. Pieper, N.K. Afseth, S.B. Engelsen, Cage of covariance in calibration modeling: regressing multiple and strongly correlated response variables onto a low rank subspace of explanatory variables, Chemometr. Intell. Lab. Syst. 213 (2021), 104311.
[7] C.E. Eskildsen, S.B. Engelsen, K.R. Dankel, L.E. Solberg, T. Næs, Diagnosing indirect relationships in multivariate calibration models, J. Chemom. (2021), e3366.
[8] A. Lorber, Error propagation and figures of merit for quantification by solving matrix equations, Anal. Chem. 58 (1986) 1167–1172.
[9] K.S. Booksh, B.R. Kowalski, Theory of analytical chemistry, Anal. Chem. (1994) 782–791.
[10] E. Sanchez, B.R. Kowalski, Tensorial calibration: I. First-order calibration, J. Chemom. 2 (1988) 247–263.
[11] C.E. Eskildsen, T. Næs, J.P. Wold, N.K. Afseth, S.B. Engelsen, Visualizing indirect correlations when predicting fatty acid composition from near infrared spectroscopy measurements, in: S.B. Engelsen, K.M. Sørensen, F.v.d. Berg (Eds.), Proceedings, 18th International Conference on Near Infrared Spectroscopy, IM Publications Open, Chichester, 2019, pp. 39–44.
[12] R. Bro, C.M. Andersen, Theory of net analyte signal vectors in inverse regression, J. Chemom. 17 (2004) 646–653.
[13] H.F.M. Boelens, W.T. Kok, O.E.d. Noord, A.K. Smilde, Performance optimization of spectroscopic process analyzers, Anal. Chem. 76 (2004) 2656–2663.
[14] C.E. Eskildsen, K.W. Sanden, S.G. Wubshet, P.V. Andersen, J. Øyaas, J.P. Wold, Estimating dry matter and far content in blocks of swiss cheese during production using on-line near-infrared spectroscopy, J. Near Infrared Spectrosc. 27 (2019) 293–301.
[15] D. Confortin, H. Neevel, M. Brustolon, L. Franco, A.J. Kettelarij, R.M. Williams, M.R.v. Bommel, Crystal violet: study of the photo-fading of an early synthetic

dye in quesous solution and on paper with HPLC-PDA, LC-MS and FORS, J. Phys. Conf. Ser. 231 (2010), 012011.
[16] C.E. Eskildsen, T. Skov, M.S. Hansen, L.B. Larsen, N.A. Poulsen, Quantification of bovine milk protein composition and coagulation properties using infrared spectroscopy and chemometrics: a result of collinearity among reference variables, J. Dairy Sci. 99 (2016) 8178–8186.
[17] C.E. Eskildsen, M.A. Rasmussen, S.B. Engelsen, L.B. Larsen, N.A. Poulsen, T. Skov, Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: understanding predictions of highly collinear reference variables, J. Dairy Sci. 97 (2014) 7940–7951.
[18] S. Wold, H. Antti, F. Lindgren, J. Öhman, Orthogonal signal correction of near-infrared spectra, Chemometr. Intell. Lab. Syst. 44 (1998) 175–185.
[19] C.A. Andersson, Direct orthogonalization, Chemometr. Intell. Lab. Syst. 47 (1999) 51–63.
[20] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemom. 16 (2002) 119–128.
[21] P.W. Hansen, Pre-processing method minimizing the need for reference analysis, J. Chemom. 15 (2001) 123–131.
[22] J. Ferré, S. Brown, Reduction of model complexity by orthogonalization with respect to non-relevant spectral changes, Appl. Spectrosc. 55 (2001) 708–714.
[23] J.M. Roger, F. Chauchard, V. Bellon-Maurel, EPO-PLS external parameter orthogonalization of PLS application to temperature-independent measurement of sugar content of intact fruits, Chemometr. Intell. Lab. Syst. 66 (2003) 191–204.
[24] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedure, Anal. Chem. 33 (1964) 1627–1639.
[25] J. Steiner, Y. Termonia, J. Deltour, Smoothing and differentiation of data by simplified least squares procedure, Anal. Chem. 44 (1972) 1906–1909.
[26] M. Andersson, A comparison of nine PLS1 algorithms, J. Chemom. 23 (2008) 518–529.
[27] N.A. Nikonenko, D.K. Buslov, N.I. Sushko, R.G. Zhbankov, Investigation of stretching vibrations of glycosidic linkages in disaccharides and polysaccharides with use of IR spectra deconvolution, Biopolymers 57 (2000) 257–262.
[28] J. Wang, M.H. Kliks, S. Jun, M. Jackson, Q.X. Li, Rapid analysis of glucose, fructose, sucrose, and maltose in honeys from different geographic regions using Fourier transform infrared spectroscopy and multivariate analysis, J. Food Sci. 75 (2010) c208–c214.
[29] J.H. Kalivas, J. Ferré, A.J. Tevante, Selectivity-relaxed classical and inverse least squares calibration and selectivity measures with a unified selectivity coefficient, J. Chemom. 31 (2017), e2925.
[30] J. Kong, S. Yu, Fourier transform infrared spectroscopic analysis of protein secondary structures, Acta Biochim. Biophys. Sin. 39 (2007) 549–559.
[31] M. Tonolini, K.M. Sørensen, P.B. Skou, C. Ray, S.B. Engelsen, Prediction of α-lactalbumin and β-lactoglobulin composition of aqueous whey solutions using Fourier transform mid-infrared spectroscopy and near-infrared spectroscopy, Appl. Spectrosc. 75 (2021) 718–727.
[32] C.E. Eskildsen, T. Næs, Sample-specific prediction error measures in spectroscopy, Appl. Spectrosc. 74 (2020) 791–798.
[33] N.M. Faber, D.L. Duewer, S.J. Choquette, T.L. Green, S.N. Chesler, Characterizing the uncertainty in near-infrared spectroscopic prediction of mixed-oxygenate concentrations in gasoline: sample specific prediction intervals, Anal. Chem. 70 (1998) 2972–2982.