Received: 4 January 2021 | Revised: 18 May 2021 | Accepted: 25 June 2021

DOI: 10.1002/cem.3366

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Diagnosing indirect relationships in multivariate calibration models

Carl Emil Eskildsen[1,3] | **Søren B. Engelsen**[2] | **Katinka R. Dankel**[3] | **Lars Erik Solberg**[3] | **Tormod Næs**[3]

[1]Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, The Netherlands

[2]Department of Food Science, University of Copenhagen, Frederiksberg, Denmark

[3]Nofima, Norwegian Institute of Food, Fisheries and Aquaculture Research, Ås, Norway

Correspondence
Carl Emil Eskildsen, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, NL-1098 XH Amsterdam, The Netherlands.
Email: c.e.a.eskildsen@uva.nl

Funding information
Netherlands Organization of Scientific Research, Grant/Award Number: 15506; Norwegian Agricultural Food Research Foundation, Grant/Award Number: 262308/F40

Abstract

Problems concerning covariance among independent variables are well understood and dealt with by inverse regression methods like partial least squares regression. However, covariance between dependent variables has only received minor attention. Biological samples are often complex mixtures of multiple covarying compounds. During multivariate calibration, analyte predictions may be mediated through relationships with interfering compounds, which implies that the calibration model is not providing a direct link between the multivariate measurements and the analyte of interest. This compromises robustness and validity of the calibration model—important aspects when applying the model to future samples and data sets. This study discusses issues of calibration modeling when strong covariance structures exist among the analyte of interest and interfering compounds.

We propose a projection-based method to diagnose whether indirect covariance structures dominate the calibration model. The proposed method is tested on a two-constituent *Beer's* law system consisting of 20 aqueous samples with covarying amounts of fructose (analyte of interest) and riboflavin (interfering compound). Transmission measurements are obtained on all samples in the visual and near-infrared wavelength ranges. Riboflavin has strong absorption in the visual region, whereas fructose exclusively absorbs in the near-infrared region. Hence, predictions of fructose concentrations, obtained from the visual wavelength range only, are fully mediated through riboflavin, whereas fructose predictions obtained from the near-infrared wavelength range may be obtained independent of riboflavin.

KEYWORDS

diagnostics, indirect correlations, multivariate calibration, selectivity

1 | INTRODUCTION

The purpose of multivariate calibration is often to estimate analyte concentrations by a linear combination of the multivariate measurements. Collinearity among the independent variables is a well-known issue in, for example, spectroscopy and is (easily) handled by data compression methods like partial least squares (PLS) regression.[1] However, issues related to collinearity between the dependent variables (collinearity between quantities of the analyte of interest and interfering compounds) appears to be less known. Although Brown,[2] Brown and Ridder,[3] and Ridder et al[4] presented extensive theoretical and practical considerations for the collinearity between dependent variables, this still appears to receive insufficient attention in multivariate calibration.

The quantitative information (for the analyte) provided by the calibration model may be mediated through an interfering compound as illustrated in Figure 1, where α defines the proportion of analyte predictions estimated through the indirect relationship with the interfering compound. In this paper, we discuss issues related to indirect predictions and propose a method to diagnose the degree to which analyte predictions are mediated through an interfering compound.

Biological samples, for example, foodstuffs, are often complex multicomponent samples consisting of water, fats, proteins, carbohydrates, etc.[5] The concentrations of these components may be highly collinear, like previously observed for dry matter and fat content in cheese.[6] One may be interested in quantifying the dry matter content from near-infrared (NIR) measurements. Nevertheless, due to collinearity between dry matter and fat content, dry matter predictions may be (fully or partly) mediated through an NIR spectral basis related to fat. In such case, the calibration model is not providing a direct link between the NIR measurement and dry matter content. Though such model may provide dry matter estimates with small errors in the calibration data, the regression model is based on an indirect relationship to fat content. This may be less problematic if the indirect relationship found in the calibration data is conserved in a new sample to which the regression model is applied. Brown and Ridder[3] and Kalivas et al[7] even show how such an indirect relationship may support the model in providing a smaller prediction error. However, as soon as the indirect relationship in the calibration data is not representative for the new sample, which may happen due to several reasons including seasonal changes when dealing with biological samples, calibration validity may be compromised.[3,7,8] Model validity is easier compromised when a model is built on indirect relationships rather than direct relationships. Therefore, a tool to diagnose indirect relationships in multivariate regression modeling is useful.

To obtain an analyte prediction, which is independent of interfering compounds, the regression vector should be in the net analyte signal (NAS) space.[9–11] The NAS is the part of the analyte signal that is in the null space of signals of all interference.[9] Hence, the magnitude of an NIR spectrum projected onto the true NAS is insensitive to interfering signals, and the prediction obtained from such projection is independent of interfering compounds.[9–11]

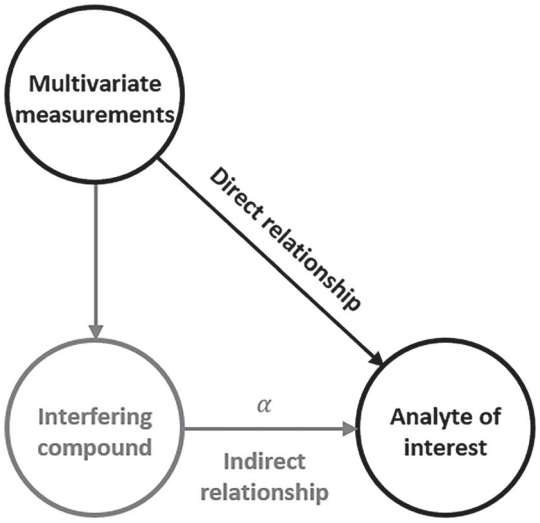

FIGURE 1 Prediction of an analyte of interest from multivariate measurements. Predictions may be based on a direct relationship between the multivariate measurements and the analyte of interest or mediated through an interfering compound, where α determines the degree to which predictions are mediated through the interfering compound

When applying data compression methods for regression, like PLS, the prediction error uncertainty is composed of a random error, an estimation error (variance contribution), and a model error (bias contribution).[12,13] It is well known that the random error is constant, whereas the estimation error will increase and the model error will decrease with increasing model complexity. This is also known as the bias/variance tradeoff.[12,13] An analyte prediction is independent of interference if the model error equals zero (i.e., the regression vector is in the true NAS space). However, as the optimal PLS model (the one providing the minimum mean squared error) is found by balancing the variance and the bias contribution, a model error (i.e., bias) will in most practical situations be present when using PLS for regression. This is especially true if the analyte and interferent are highly, but not perfectly correlated in sample concentrations of calibration data. In NAS theory, the analyte and interferent are two compounds, but depending on the measurement noise, they may, in practice, be indistinguishable and span a one-dimensional space.[2] In such case, the model is not selective toward the interferent, but the bias contribution to the mean squared error may be relatively small due to collinearity among concentrations of the analyte and interferent. Therefore, it may not be cost effective (in terms of mean squared error) to improve model selectivity (i.e., decreasing the model error) at the expense of increasing the variance contribution (i.e., estimation error), and the PLS model will not be fully selective.

Brown and Ridder[3] estimated selectivity (i.e., the degree to which an analyte prediction depends on the concentration of an interferent) by investigating the relationship between the sample-specific prediction bias and the concentration of the interfering compound. This approach works well when concentrations of the analyte and interferent are orthogonal (or if the sensitivity equals one). When concentrations of the analyte and interferent covary, the selectivity could be poor, but this may not manifest as a proportional prediction bias due to the collinearity between concentrations of the analyte and interferent.[3] However, if the selectivity toward an interferent is poor, the covariance between the analyte predictions and concentrations of the interferent will increase as compared to the covariance between reference values of the analyte and interferent.[14] In this paper, we will use this altered covariance to estimate how analyte predictions depend on an interfering compound.

In a previous study, Eskildsen et al[15] suggested to split the explained analyte variation into a direct and indirect part by projection/orthogonalization. But only the analyte variation orthogonal to the variation of the interfering compound was found to be direct. However, an analyte and an interfering compound may covary while being estimated from independent chemical information. Therefore, the approach[15] is a simplification and may provide misleading results. To diagnose indirect relationships in multivariate regression models, it is necessary to consider the relationship between the regression vector (related to the analyte) and the pure signals of both the analyte and interfering compounds.[2,3] This is further elaborated in the subsequent section of this paper. Eskildsen et al[15] additionally suggested to calculate the coefficient of determination (between concentrations of the analyte and interfering compound) using the reference values as well as the predicted values. If, for example, the analyte is modeled through indirect correlation to an interfering compound, then predictions of the analyte and the interfering compound will originate from a similar linear combination of the multivariate measurements. In that case, the magnitude of the covariance between the analyte and interfering compound will be higher among the predicted values as compared to the reference values. This approach was also applied by Eskildsen et al.[8]

In a situation with high covariance between quantities of the analyte and an interfering compound, Rinnan et al[16] calibrated the regression model on a subset of the available samples. The subset was selected so the analyte variation was as close to orthogonal as possible to the variation of the interfering compound. This approach avoids that indirect relationships are being incorporated into the regression model due to covariance between quantities of the analyte and interfering compound but will not avoid effects of overlapping signals. As this approach[16] is restricted to calibrate on a subset of the available samples, the number of calibration samples as well as the concentration range of both analyte and interfering compound may be limited. The approach[16] is a pragmatic explorative method and may provide information on whether an analyte can be modeled independent of an interfering compound.

Romano et al[17] and Aben et al[18] used principles from partial correlation analysis to investigate structures among multiple data sets. Consider three data sets, A, B, and C. Partial correlation analysis provides, for example, information on how much variation of C can be explained by A conditioned on B. In this present paper, we wish to resemble the idea of partial correlation analysis into a context of diagnosing indirect relationships in regression models. We aim at understanding how much analyte variation can be explained from the multivariate measurements conditioned on interfering compounds. The general idea, limitations, and prerequisites of the proposed method are presented in the subsequent section.

2 | THEORY

Having fitted the multivariate linear calibration model between reference values of the analyte of interest, $\boldsymbol{c}_1(n \times 1)$, and the multivariate measurements, $\boldsymbol{X}(n \times m)$, the predictions are obtained by

$$\hat{\boldsymbol{c}}_1 = \boldsymbol{X}\hat{\boldsymbol{b}} \tag{1}$$

where $\hat{\boldsymbol{c}}_1(n \times 1)$ contains predicted analyte quantities and $\hat{\boldsymbol{b}}(m \times 1)$ contains the estimated regression coefficients. Both $\boldsymbol{X}$ and $\boldsymbol{c}_1$ are assumed column-wise mean centered.

From classical least squares, we get

$$\boldsymbol{X} = \boldsymbol{C}\boldsymbol{S}^T \tag{2}$$

where $\boldsymbol{C}(n \times r)$ contains (mean-centered) quantities of analyte and interfering compounds and $\boldsymbol{S}(m \times r)$ contains the pure signals at unit concentration of analyte and interfering compounds. For simplicity, in Equation 2, we neglect the error term containing, for example, random instrumental noise.

Substituting Equation 2 into Equation 1 returns,

$$\hat{\boldsymbol{c}}_1 = \boldsymbol{C}\boldsymbol{S}^T\hat{\boldsymbol{b}} \tag{3}$$

If we consider a two-constituent system, consisting of the analyte of interest and an interfering compound, Equation 3 is expressed as

$$\hat{\boldsymbol{c}}_1 = [\boldsymbol{c}_1, \boldsymbol{c}_2]\begin{bmatrix}\boldsymbol{s}_1^T \\ \boldsymbol{s}_2^T\end{bmatrix}\hat{\boldsymbol{b}} = \boldsymbol{c}_1\boldsymbol{s}_1^T\hat{\boldsymbol{b}} + \boldsymbol{c}_2\boldsymbol{s}_2^T\hat{\boldsymbol{b}} \tag{4}$$

where $\boldsymbol{c}_2(n \times 1)$ contains concentrations of the interfering compound and $\boldsymbol{s}_1(m \times 1)$ and $\boldsymbol{s}_2(m \times 1)$ are pure signals at unit concentration (i.e., the chemical bases) of analyte and interfering compound, respectively.

From Equation 4, it is clear that $\hat{\boldsymbol{c}}_1$ is made up of a vector sum with contributions from both the analyte of interest ($\boldsymbol{c}_1\boldsymbol{s}_1^T\hat{\boldsymbol{b}}$) and the interfering compound ($\boldsymbol{c}_2\boldsymbol{s}_2^T\hat{\boldsymbol{b}}$). To evaluate the calibration model, it is common to plot $\hat{\boldsymbol{c}}_1$ against $\boldsymbol{c}_1$. The average error term is obviously important. But the slope, a of the linear least squares fit between $\boldsymbol{c}_1$ and $\hat{\boldsymbol{c}}_1$, describing the relation between $\boldsymbol{c}_1$ and $\hat{\boldsymbol{c}}_1$ (Figure 2A), is also important. The slope a is given by $\frac{\boldsymbol{c}_1^T\hat{\boldsymbol{c}}_1}{\|\boldsymbol{c}_1\|^2}$. Substituting the expression for $\hat{\boldsymbol{c}}_1$, given by Equation 4, into $\frac{\boldsymbol{c}_1^T\hat{\boldsymbol{c}}_1}{\|\boldsymbol{c}_1\|^2}$ returns,

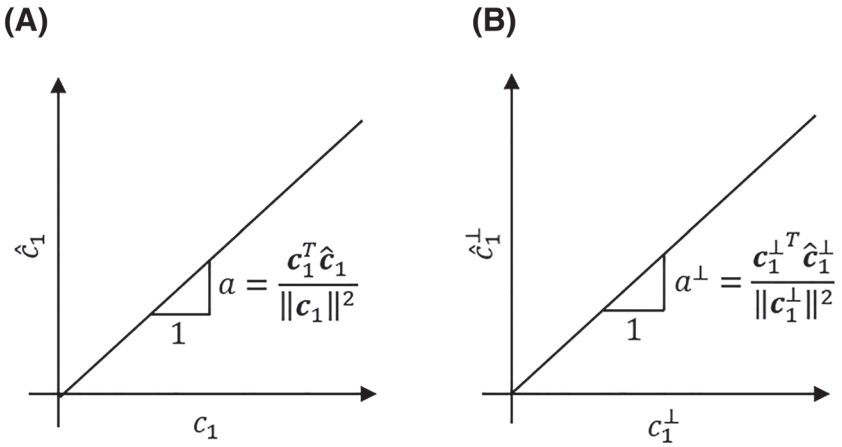

FIGURE 2 (A) Relationship between predicted ($\hat{\boldsymbol{c}}_1$) and measured ($\boldsymbol{c}_1$) analyte quantities. (B) Relationship between predicted ($\hat{\boldsymbol{c}}_1^\perp$) and measured ($\boldsymbol{c}_1^\perp$) analyte quantities orthogonalized with respect to an interfering compound

$$\frac{\boldsymbol{c}_1^T\widehat{\boldsymbol{c}}_1}{\|\boldsymbol{c}_1\|^2}=\frac{\boldsymbol{c}_1^T}{\|\boldsymbol{c}_1\|^2}\left(\boldsymbol{c}_1\boldsymbol{s}_1^T\widehat{\boldsymbol{b}}+\boldsymbol{c}_2\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}\right)=\boldsymbol{s}_1^T\widehat{\boldsymbol{b}}+\frac{\boldsymbol{c}_1^T\boldsymbol{c}_2}{\|\boldsymbol{c}_1\|^2}\boldsymbol{s}_2^T\widehat{\boldsymbol{b}} \tag{5}$$

Hence, the relationship between $\boldsymbol{c}_1$ and $\widehat{\boldsymbol{c}}_1$ is given by two model dependent constants, namely, $\boldsymbol{s}_1^T\widehat{\boldsymbol{b}}$ and $\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}$, as well as a data set dependent constant, $\frac{\boldsymbol{c}_1^T\boldsymbol{c}_2}{\|\boldsymbol{c}_1\|^2}$ Equation 5.

To obtain good direct predictions, $\widehat{\boldsymbol{b}}$ should have the direction of $\boldsymbol{s}_1$ while being (nearly) orthogonal to any interfering signals in the calibration data (i.e., $\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}=0$). This ensures the model to, exclusively, fit on the chemical basis related to $\boldsymbol{c}_1$ and the relationship between $\boldsymbol{c}_1$ and $\widehat{\boldsymbol{c}}_1$ would be given by $\boldsymbol{s}_1^T\widehat{\boldsymbol{b}}$ (Equation 5). For a good and direct model, the $\left\|\widehat{\boldsymbol{b}}\right\|$ is scaled so $\boldsymbol{s}_1^T\widehat{\boldsymbol{b}}$ is close to 1.[5,19] These observations are also evident from Equations 4 and 5.

Hence, the proportion to which $\widehat{\boldsymbol{c}}_1$ is mediated through $\boldsymbol{c}_1$ is given by $\boldsymbol{s}_1^T\widehat{\boldsymbol{b}}$, and the proportion to which $\widehat{\boldsymbol{c}}_1$ is mediated through $\boldsymbol{c}_2$ is given by $\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}$. Whereas both the proportions ($\boldsymbol{s}_1^T\widehat{\boldsymbol{b}}$ and $\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}$) are model-dependent constants and therefore conserved when the model is applied to future data sets, $\boldsymbol{c}_2$ is data set dependent.

If $\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}=0$, then $\widehat{\boldsymbol{c}}_1$ is not influenced by $\boldsymbol{c}_2$ (Equation 4). However, evaluating this is difficult in practice because $\boldsymbol{s}_2$ (as well as $\boldsymbol{s}_1$) is not necessarily known. An alternative approach of assessing this, which will be pursued here, is to use orthogonalization with respect to $\boldsymbol{c}_2$. This is done by multiplying both sides of Equation 4 with $(\boldsymbol{I}-\boldsymbol{P})$ where $\boldsymbol{I}(n\times n)$ is the identity matrix and $\boldsymbol{P}(n\times n)$ is the projection matrix given by $\boldsymbol{c}_2\left(\boldsymbol{c}_2^T\boldsymbol{c}_2\right)^{-1}\boldsymbol{c}_2^T$. We then obtain

$$\widehat{\boldsymbol{c}}_1^{\perp}=\boldsymbol{c}_1^{\perp}\boldsymbol{s}_1^T\widehat{\boldsymbol{b}}+\boldsymbol{0}\boldsymbol{s}_2^T\widehat{\boldsymbol{b}} \tag{6}$$

where $\widehat{\boldsymbol{c}}_1^{\perp}$ (n $\times$ 1) and $\boldsymbol{c}_1^{\perp}$ $(n\times 1)$ are $\widehat{\boldsymbol{c}}_1$ and $\boldsymbol{c}_1$ orthogonalized with respect to $\boldsymbol{c}_2$, respectively, and $\boldsymbol{0}(n\times 1)$ is a vector of zeros (the contribution from the interfering compound is canceled in Equation 6).

The relationship between $\boldsymbol{c}_1^{\perp}$ and $\widehat{\boldsymbol{c}}_1^{\perp}$ is again given by the slope term $a^{\perp}$ of the linear least squares fit (Figure 2B). The slope term $a^{\perp}$ is given by $\frac{\boldsymbol{c}_1^{\perp\,T}\widehat{\boldsymbol{c}}_1^{\perp}}{\|\boldsymbol{c}_1^{\perp}\|^2}$. Substituting the expression for $\widehat{\boldsymbol{c}}_1^{\perp}$, given by Equation 6, into $\frac{\boldsymbol{c}_1^{\perp\,T}\widehat{\boldsymbol{c}}_1^{\perp}}{\|\boldsymbol{c}_1^{\perp}\|^2}$, returns,

$$\frac{\boldsymbol{c}_1^{\perp\,T}\widehat{\boldsymbol{c}}_1^{\perp}}{\|\boldsymbol{c}_1^{\perp}\|^2}=\frac{\boldsymbol{c}_1^{\perp\,T}}{\|\boldsymbol{c}_1^{\perp}\|^2}\left(\boldsymbol{c}_1^{\perp}\boldsymbol{s}_1^T\widehat{\boldsymbol{b}}+\boldsymbol{0}\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}\right)=\boldsymbol{s}_1^T\widehat{\boldsymbol{b}} \tag{7}$$

Now, the term $\frac{\boldsymbol{c}_1^T\boldsymbol{c}_2}{\|\boldsymbol{c}_1\|^2}\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}$ (the degree to which $\boldsymbol{c}_2$ affects the relationship between $\boldsymbol{c}_1$ and $\widehat{\boldsymbol{c}}_1$) is calculated by subtracting Equation 7 from Equation 5 (i.e., subtracting the two slope terms, sketched in Figure 2, from each other).

$$\frac{\boldsymbol{c}_1^T\widehat{\boldsymbol{c}}_1}{\|\boldsymbol{c}_1\|^2}-\frac{\boldsymbol{c}_1^{\perp\,T}\widehat{\boldsymbol{c}}_1^{\perp}}{\|\boldsymbol{c}_1^{\perp}\|^2}=\frac{\boldsymbol{c}_1^T\boldsymbol{c}_2}{\|\boldsymbol{c}_1\|^2}\boldsymbol{s}_2^T\widehat{\boldsymbol{b}} \tag{8}$$

Hence, if the two slope terms presented in Figure 2 are identical, then Equation 8 equals zero and $\boldsymbol{c}_2$ is not impacting the relationship between $\boldsymbol{c}_1$ and $\widehat{\boldsymbol{c}}_1$. However, if Equation 8 is different from zero, then $\boldsymbol{c}_2$ has an impact on the relationship between $\boldsymbol{c}_1$ and $\widehat{\boldsymbol{c}}_1$. The degree to which $\widehat{\boldsymbol{c}}_1$ is mediated through $\boldsymbol{c}_2$, $\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}$ is obtained by multiplying each side of Equation 8 by $\frac{\|\boldsymbol{c}_1\|^2}{\boldsymbol{c}_1^T\boldsymbol{c}_2}$. If $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are orthogonal, which could be the case when applying an experimental design, it is evident that $\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}$ cannot be estimated from the proposed procedure. In such case, $\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}$ should be estimated by the procedure of Brown and Ridder.[3]

When multiple interfering compounds are present, Equations 4–8 are expanded with extra terms corresponding to the number of extra interfering compounds. The influence of multiple interfering compounds may be investigated by deflating the interfering compounds one at a time. In such case, expressions corresponding to Equation 8 will be archived for each interfering compound. Hence, in the case of two interfering compounds, it would be a matter of solving two equations with two unknowns.

For the method to be valid, data are assumed to follow the principles of *Beer's* law, as also indicated by Equation 2. Furthermore, $\boldsymbol{X}$ must be measured, and quantitative information on the analyte of interest as well as (all) interfering compound(s) is required. Reference measurements of the analyte and interfering compounds should be reasonable with independent and identically distributed noise.

In many industrial (and academic) systems, extensive knowledge of both the analyte and interfering compounds often exists. However, in systems of biological nature, this kind of extensive knowledge may not be available. Nevertheless, if only (one or) a few interfering compounds are known, the orthogonalization procedure may still be applied. One could still evaluate the difference in the two slope terms presented in Figure 2. If the two slope terms are different, it is a consequence of the analyte being mediated through variation associated with the interfering compound. If the two slope terms are identical, this could indicate that the model is fitting solely on the analyte information in $\boldsymbol{X}$. However, it could also be the case that the analyte is mediated through an interfering compound for which quantitative information is not available.

3 | MATERIALS AND METHODS

3.1 | Model system

A simple two-constituent model system was prepared to test the projection procedure outlined in the previous section. The model system was made with fructose as analyte of interest and riboflavin as interfering compound. The model system was constructed with high covariance between quantities of fructose and riboflavin to allow fructose concentrations to be modeled through its correlation with riboflavin concentrations. Spectroscopic measurements in the visual and NIR wavelength ranges were obtained as independent variables. First, fructose concentrations were modeled from the visual range of the spectroscopic measurements containing only signals from riboflavin, returning an indirect model. Then, fructose concentrations were modeled from the visual and NIR range of the spectroscopic measurements containing both signals from riboflavin and fructose in order to facilitate a direct model.

3.2 | Sample preparation

Two stock solutions were prepared: one stock solution with D(−)-fructose (VWR International, Leuven, Belgium) and one stock solution with riboflavin (Sigma-Aldrich, St. Louis, MO, USA). Fructose and riboflavin were dissolved in water to make the two stock solutions of concentrations 10% w/V and $3.5 \cdot 10^{-7}$ M for fructose and riboflavin, respectively.

In total, 20 samples were prepared by mixing varying amounts of the two stock solutions and adding water to a total volume of 10 ml. The samples are presented in Figure 3. The added volume of each stock solution was used as reference values.

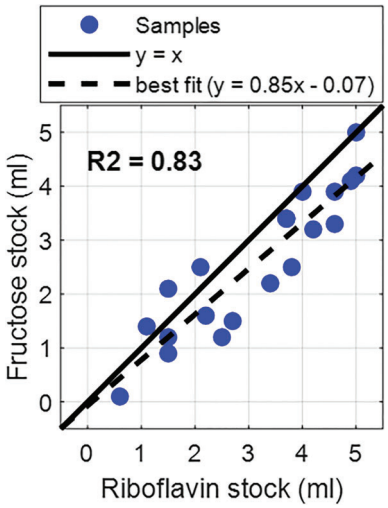

FIGURE 3 Relationship between riboflavin and fructose (reference values) in calibration data

3.3 | Spectroscopic measurements

Spectroscopic measurements were obtained using a quartz cuvette with a path length of 2 mm using a FOSS NIRSystems XDS Rapid Liquid™ Analyzer (FOSS Analytics A/S, Hillerød, Denmark). The spectral range was from 400 to 2500 nm, and measurements were obtained at every 0.5 nm. However, only the spectral range from 410 to 2315 nm was included in the study. A total of 32 scans were obtained per measurement, and samples were measured in triplicates. Triplicates were averaged, and the average spectrum was used for further analysis.

3.4 | Data analysis

Data were analyzed using MATLAB Version R2019a (9.6.0.1072779, MathWorks Inc., Natick, MA, USA). In order to obey *Beer's law*, spectroscopic measurements were converted from transmission % into absorbance (A = log1/T). Prior to modeling, the spectroscopic measurements were preprocessed by Savitzky–Golay first derivative (window size of 41 data points, corresponding to 10 nm on each side of the center point, and second-order polynomial)[20,21] and mean centered. Furthermore, riboflavin and fructose concentrations were mean centered. The nonlinear iterative partial least squares (NIPALS) algorithm[22] was used for PLS regression. All PLS models were built with univariate reference values (i.e., $\boldsymbol{y}$-block). The number of latent variables included in each PLS model was determined by a 10-fold random subset cross-validation.

4 | RESULTS AND DISCUSSION

The preprocessed spectroscopic measurements (colored by riboflavin concentration) are presented in Figure 4. The region from 1830 to 2130 nm was removed due to noise.

Riboflavin stock solution is yellow and therefore absorbs light in the region from 400 to 500 nm.[23] This is also clearly observed in Figure 4. In contrast, the fructose stock solution is transparent and thus will not absorb light in the visual region. Due to overtones of molecular vibrations (primarily third overtones of the O–H stretching vibrations from the hydroxy groups), fructose absorbs electromagnetic radiation between 900 and 1000 nm in the shortwave NIR region.[24] It should however be noted that also the hydroxy groups of riboflavin will absorb in the same region. Furthermore, fructose will absorb at 2270 nm due to a combination tone of O–H and C–C stretching.[25]

Constructing calibration models using the visual region (410–700 nm) will result in direct and robust predictions of riboflavin content. Predictions of fructose content, obtained from the visual region, will however be fully mediated through the relationship with riboflavin. However, constructing calibration models using both the visual and NIR region should result in direct and robust predictions of both riboflavin and fructose concentrations. In the following, this will be investigated using the projection-based diagnostics method outlined in the theory section of this paper.

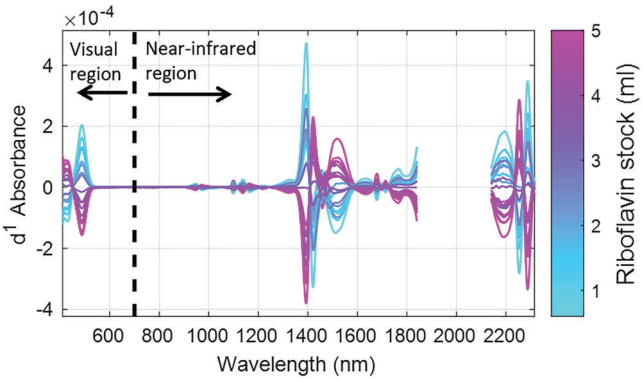

FIGURE 4 Spectroscopic measurements preprocessed by Savitzky–Golay first derivative and mean centered. Spectra are colored by riboflavin concentration

Figure 5 shows the results when fitting calibration models using the visual region of the spectroscopic measurements only. The visual region is only affected by riboflavin absorption. Hence, the spectra in this region are in a one-dimensional space, and consequently, the two PLS models (predicting riboflavin and fructose, respectively) are both constructed using just one latent variable. As expected, almost perfect predictions are obtained for riboflavin concentrations (Figure 5A and Table 1). At first glance, fructose predictions seem good (Figure 5B and Table 1). However, a closer look at the pattern between reference and predicted fructose concentrations (Figure 5B) reveals similarities to the pattern between reference riboflavin and fructose concentrations (Figure 3). This indicates that fructose is being modeled through the indirect correlation to riboflavin. Moreover, regression coefficients (normalized to unit length) for riboflavin and fructose coincide (Figure 5C). Hence, concentrations of riboflavin and fructose are predicted from the same subspace (the chemical basis related to riboflavin) of the spectroscopic measurements when the visual region is used only.

This is confirmed in Figure 6A, which shows a perfect least squares fit between predicted riboflavin and fructose concentrations. When applying the orthogonalization procedure (Equation 6), Figure 6B shows that the predictions of fructose are completely deflated (this is also observed from Table 2). The slope between reference and predicted values (Figure 6B) is zero. Hence, predictions of fructose concentrations are fully mediated through riboflavin concentrations when predictions are obtained using the visual region of the spectroscopic measurements only (Table 2). The term $\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}$,

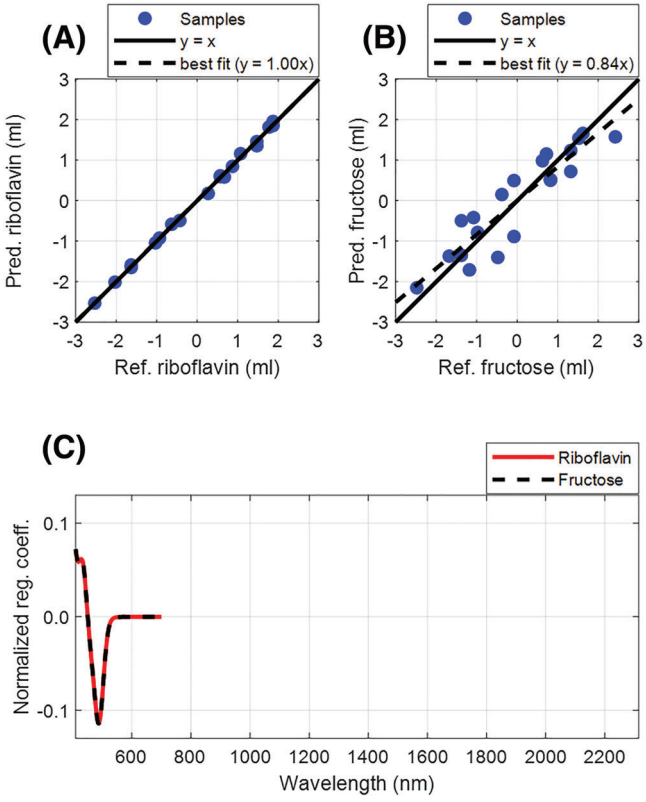

FIGURE 5 Predictions of (A) riboflavin and (B) fructose from preprocessed spectroscopic measurements obtained in the visual region. (C) Normalized regression coefficients. Data are mean centered

TABLE 1 Results from calibration models

	Visual region		Visual + NIR region	
	#LV	*MSEC*	*#LV*	*MSEC*
Riboflavin (interferent)	1	$2.9 \cdot 10^{-3}$	3	$2.4 \cdot 10^{-3}$
Fructose (analyte)	1	$2.7 \cdot 10^{-1}$	3	$3.0 \cdot 10^{-4}$

Abbreviations: *#LV*, number of latent variables included in the calibration model; *MSEC*, mean squared error of calibration.

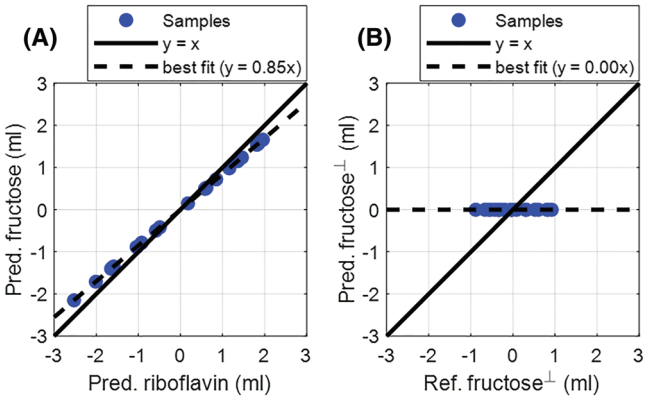

FIGURE 6 (A) Relationship between predicted riboflavin and fructose and (B) relationship between reference and predicted fructose for orthogonalized data. Predictions are obtained from the visual region of spectroscopic measurements and data are mean centered

TABLE 2 Elements used for calculating $\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}$ (i.e., the degree to which estimates of fructose are influenced by riboflavin)

Wavelength region	$\|\boldsymbol{c}_1\|^2$	$\boldsymbol{c}_1^T\boldsymbol{c}_2$	$\boldsymbol{a}$	$\boldsymbol{a}^\perp$	$\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}$
Visual region	33.2	32.6	0.84	0.00	0.85
Visual + NIR region			1.00	0.98	0.02

Notes: Reference values of fructose (analyte of interest), $\boldsymbol{c}_1$, and riboflavin, interfering compound), $\boldsymbol{c}_2$. Slope terms between reference and predicted fructose for original and orthogonalized data, a and $a^\perp$, respectively.

which is the proportion fructose is mediated through riboflavin, is equal to 0.85 (Table 2). This term is conserved if this model is used in the future.

Figure 7 shows the results when fitting calibration models using the visual and NIR region of the spectroscopic measurements. Again, perfect predictions are obtained for riboflavin content (Figure 7A and Table 1), and improved predictions are obtained for fructose content (Figure 7B and Table 1). Now, the spectra are more complex, and both PLS models are built with three latent variables. From Figure 7B, it is observed that the slope between reference and predicted fructose is 1.00. The regression coefficients (normalized to unit length) for riboflavin and fructose are shown in Figure 7C. The regression coefficients no longer coincide (like in Figure 5C), and the riboflavin PLS model is also picking up information in the NIR region presumably due to the native O-H vibrations of riboflavin. Nevertheless, the regression vector for fructose must be orthogonal to the riboflavin signal in the spectroscopic measurements. Otherwise, fructose predictions are not independent of riboflavin. This is difficult to judge even in this very simple model system. To investigate whether fructose predictions are independent of riboflavin, the orthogonalization procedure (Equation 6) is once again carried out.

Figure 8A shows that the relationship between predicted riboflavin and fructose content is very similar to the relationship between reference riboflavin and fructose content (Figure 3). This is, of course, a reflection of the fact that concentrations of both riboflavin and fructose are well predicted when using both the visual and NIR region. Furthermore, it indicates that predictions of fructose concentrations are obtained with no or little influence of riboflavin. Figure 8B shows the predictions of fructose content when applying the orthogonalization procedure with riboflavin as interfering compound (Equation 6). It is found that the slope between reference and predicted fructose concentrations, in the orthogonalized data, is 0.98 (Figure 8B). This is a slight decrease as compared with the non-orthogonalized data (Figure 7B). Hence, less variation is left in the predictions, as compared with the reference values, after orthogonalization with riboflavin. This reveals that the fructose predictions are still (slightly) modeled by the chemical basis related to riboflavin in the spectroscopic measurements, even though signals for both fructose and riboflavin are present in the spectral data. When calculating the degree to which fructose predictions are mediated through the indirect correlations to riboflavin, $\boldsymbol{s}_2^T\widehat{\boldsymbol{b}}$, it suggests that the fructose predictions (Figure 7B) are mediated by a factor of 0.02 through the riboflavin content (Table 2).

When carrying out the orthogonalization step, we recommend fitting an additional regression model to $\boldsymbol{X}$ for the interfering compound. If the reference uncertainty of the interferent is high, one may consider orthogonalization with

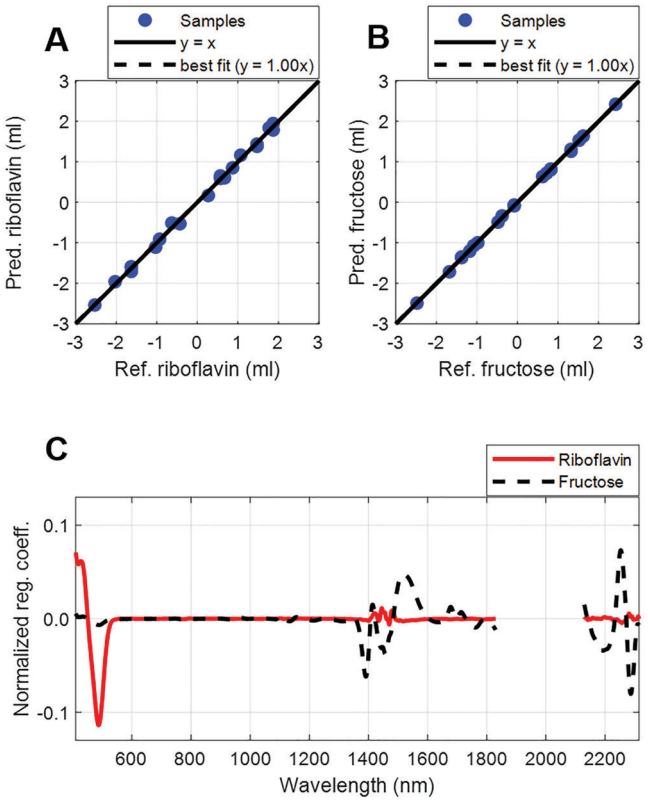

FIGURE 7 Prediction of (A) riboflavin and (B) fructose from preprocessed spectroscopic measurements obtained in the visual and near-infrared region. (C) Normalized regression coefficients. Data are mean centered

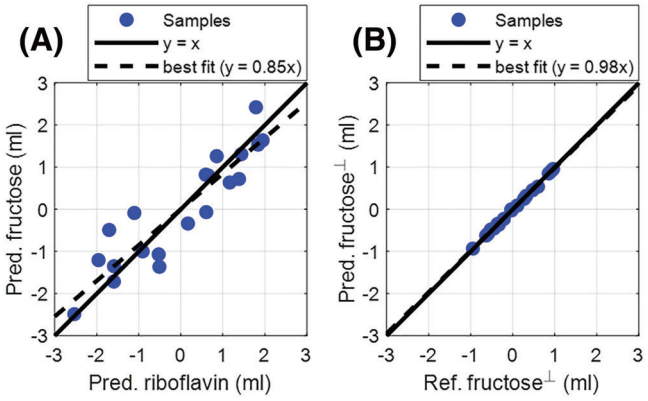

FIGURE 8 Relationship between (A) predicted riboflavin and fructose and (B) relationship between reference and predicted fructose for orthogonalized data. Predictions are obtained from the visual and near-infrared region of spectroscopic measurements and data are mean centered

respect to estimates of the interfering compound. This ensures that $\widehat{\boldsymbol{c}}_1^{\perp}$ (Equation 6) is located within the original $\boldsymbol{X}$-space. This is not necessarily the case if poor reference values of the interfering compound are used.

The method presented here uses principles from partial correlation analysis. Therefore, one needs to be careful drawing conclusions on *causal* relationships between the multivariate measurements and analyte quantities. This is especially true when the measurements are obtained from complex samples. In such situation, *hidden* compounds (i.e., compounds with no reference values measured) may be present in the samples. These *hidden* compounds could provide the indirect relationship between the multivariate measurements and the analyte. From the method presented here, it is impossible to explore such relationships. Therefore, by using this method, one can conclude whether the analyte is mediated through variation related to specific interfering compounds.

Moreover, if predictions of the interferent are used during orthogonalization, the method presented here will not differentiate on whether the analyte estimates dependent on the interfering compound or vice versa. Take the example presented in Figure 5. Are the fructose (analyte) predictions depending on riboflavin (interfering compound) or are riboflavin predictions depending on fructose? The method presented here will not provide a direct answer to this. Rather, the method tells how similar the two models are. To draw valid conclusions, one additionally needs to rely on data and model interpretation. Figure 5 shows that riboflavin content is predicted better than fructose content. This gives an indication that riboflavin is modeled from a direct relationship, whereas fructose is modeled from an indirect relationship to the multivariate measurements. Also, knowledge of absorption signals related to fructose and riboflavin should be considered.

The diagnostic method shown in this paper is based on the inner relations between the regression vector and the signals of analyte and interfering compounds, respectively. Therefore, in contrast to the method outlined by Eskildsen et al,[15] this method considers whether the analyte is modeled from a subspace of the independent variables (a chemical basis in the multivariate measurements) different from that used to model the interfering compound.

5 | CONCLUSIONS

In this paper, we show how analyte predictions, with small prediction errors, may be based (entirely or partly) on signal(s) of interfering compound(s), compromising robustness of the regression model. Furthermore, we propose a method to diagnose how indirect relationships between the analyte of interest and interfering compound(s) impact analyte predictions during regression modeling. The impact is a multiplication of two factors, namely, the concentration of the interfering compound (data set dependent) and the inner product between the signal of the interferent and the estimated regression vector for the analyte (regression model dependent). We estimate this inner product based on the concentrations without knowledge of pure signals.

ACKNOWLEDGEMENTS

For funding, we acknowledge the Norwegian Agricultural Food Research Foundation through the project FoodSMaCK—Spectroscopy, Modeling & Consumer Knowledge (no. 262308/F40). Furthermore, the position of the first author is currently funded via the TooCOLD project (grant number 15506), which is (partly) financed by the Netherlands Organization of Scientific Research (NWO) via the TTW Open Technology Programme.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/cem.3366.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Carl Emil Eskildsen https://orcid.org/0000-0003-3778-1771
Lars Erik Solberg https://orcid.org/0000-0003-0246-8064
Tormod Næs https://orcid.org/0000-0001-5610-3955

REFERENCES

1. Wold S, Ruhe A, Wold H, Dunn WJ III. The collinearity problem in regression, the partial least squares approach to generalized inverses. *SIAM J Sci Stat Comput*. 1984;5(3):735-743. https://doi.org/10.1137/0905052
2. Brown CD. Discordance between net analyte signal theory and practical multivariate calibration. *Anal Chem*. 2004;76(15):4364-4373. https://doi.org/10.1021/ac049953w
3. Brown CD, Ridder TD. Framework for multivariate selectivity analysis, part I: theoretical and practical merits. *Appl Spectrosc*. 2005;59(6):787-803. https://doi.org/10.1366/0003702054280621

4. Ridder TD, Brown CD, Steeg BJV. Framework for multivariate selectivity analysis, part II: experimental applications. *Appl Spectrosc*. 2005;59(6):804-815. https://doi.org/10.1366/0003702054280739
5. Eskildsen CE, Fvd B, Engelsen SB. Vibrational spectroscopy in food processing. In: Lindon JC, Tranter GE, Koppenaal DW, eds. *Encyclopedia of Spectroscopy and Spectrometry*. 3rd ed. Oxford, UK: Elsevier; 2017:582-589. https://doi.org/10.1016/B978-0-12-409547-2.12156-0
6. Eskildsen CE, Sanden KW, Wubshet SG, Andersen PV, Øyaas J, Wold JP. Estimating dry matter and fat content in blocks of Swiss cheese during production using on-line near-infrared spectroscopy. *J near Infrared Spectrosc*. 2019;27(4):293-301. https://doi.org/10.1177/0967033519855436
7. Kalivas JH, Ferré J, Tencate AJ. Selectivity-relaxed classical and inverse least squares calibration and selectivity measures with a unified selectivity coefficient. *J Chemometr*. 2017;31(11):1-23, e2925. https://doi.org/10.1002/cem.2925
8. Eskildsen CE, Skov T, Hansen MS, Larsen LB, Poulsen NA. Quantification of bovine milk protein composition and coagulation properties using infrared spectroscopy and chemometrics: a result of collinearity among reference variables. *J Dairy Sci*. 2016;99(10):8178-8186. https://doi.org/10.3168/jds.2015-10840
9. Booksh KS, Kowalski BR. Theory of analytical chemistry. *Anal Chem*. 1994;66(15):782-791. https://doi.org/10.1021/ac00087a001
10. Sanchez E, Kowalski BR. Tensorial calibration: I. First-order calibration. *J Chemometr*. 1988;2(4):247-263. https://doi.org/10.1002/cem.1180020404
11. Bro R, Andersen CM. Theory of net analyte signal vectors in inverse regression. *J Chemometr*. 2004;17(12):646-652. https://doi.org/10.1002/cem.832
12. Faber NM, Duewer DL, Choquette SJ, Green TL, Chesler SN. Characterizing the uncertainty in near-infrared spectroscopic prediction of mixed-oxygenate concentrations in gasoline: sample specific prediction intervals. *Anal Chem*. 1998;70(14):2972-2982. https://doi.org/10.1021/ac971270b
13. Eskildsen CE, Næs T. Sample-specific prediction error measures in spectroscopy. *Appl Spectrosc*. 2020;74(7):791-798. https://doi.org/10.1177/0003702820913562
14. Eskildsen CE, Næs T, Skou PB, et al. 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*. 2021;213:1-7, 104311. https://doi.org/10.1016/j.chemolab.2021.104311
15. Eskildsen CE, Rasmussen MA, Engelsen SB, Larsen LB, Poulsen NA, Skov T. Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: understanding predictions of highly collinear reference variables. *J Dairy Sci*. 2014;97(12):7940-7951. https://doi.org/10.3168/jds.2014-8337
16. Rinnan Å, Bruun S, Lindedam J, et al. Predicting the ethanol potential of wheat straw using near-infrared spectroscopy and chemometrics: the challenge of inherently intercorrelated response functions. *Anal Chim Acta*. 2017;962:15-23. https://doi.org/10.1016/j.aca.2017.02.001
17. Romano R, Tomic O, Liland KH, Smilde A, Næs T. A comparison of two PLS-based approaches to structural equation modeling. *J Chemometr*. 2018;33(3):1-28, e3105. https://doi.org/10.1002/cem.3105
18. Aben N, Westerhuis JA, Song Y, et al. iTOP: inferring the topology of omics data. *Bioinformatics*. 2018;34(17):i988-i996. https://doi.org/10.1093/bioinformatics/bty636
19. Eskildsen CE, Næs T, Wold JP, Afseth NK, Engelsen SB. Visualizing indirect correlations when predicting fatty acid composition from near infrared spectroscopy measurements. In: Engelsen SB, Sørensen KM, Fvd B, eds. *Proceedings, 18th International Conference on Near Infrared Spectroscopy*. Chichester, UK: IM Publications Open; 2019:39-44.
20. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedure. *Anal Chem*. 1964;36(8):1627-1639. https://doi.org/10.1021/ac60214a047
21. Steiner J, Termonia Y, Deltour J. Smoothing and differentiation of data by simplified least squares procedure. *Anal Chem*. 1972;44(11):1906-1909. https://doi.org/10.1021/ac60319a045
22. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometr Intell Lab Syst*. 2001;58(2):109-130. https://doi.org/10.1016/S0169-7439(01)00155-1
23. Hongwei Z, Zhaoxia GMZ, Wenfeng W, Guozhong W. Spectroscopic studies on the interaction between riboflavin and albumins. *Spectrochim Acta a*. 2006;65(3-4):811-817. https://doi.org/10.1016/j.saa.2005.12.038
24. Golic M, Walsh K, Lawson P. Short-wavelength near-infrared spectra of sucrose, glucose and fructose with respect to sugar concentration and temperature. *Appl Spectrosc*. 2003;57(2):139-145. https://doi.org/10.1366/000370203321535033
25. Yano T, Matsushige H, Suehara KI, Nakano Y. Measurements of the concentration of glucose and lactic acid in peritoneal dialysis solutions using near-infrared spectroscopy. *J Biosci Bioeng*. 2000;90(5):540-544. https://doi.org/10.1016/S1389-1723(01)80037-2

How to cite this article: Eskildsen CE, Engelsen SB, Dankel KR, Solberg LE, Næs T. Diagnosing indirect relationships in multivariate calibration models. *Journal of Chemometrics*. 2021;e3366. https://doi.org/10.1002/cem.3366