Chemometrics and Intelligent Laboratory Systems 213 (2021) 104311

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

# Cage of covariance in calibration modeling: Regressing multiple and strongly correlated response variables onto a low rank subspace of explanatory variables

Carl Emil Eskildsen [a,b,*], Tormod Næs [b], Peter B. Skou [c], Lars Erik Solberg [b], Katinka R. Dankel [b], Silje A. Basmoen [b], Jens Petter Wold [b], Siri S. Horn [b], Borghild Hillestad [d], Nina A. Poulsen [e], Mette Christensen [f], Theo Pieper [f], Nils Kristian Afseth [b], Søren B. Engelsen [c]

[a] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Science Park 904, NL-1098 XH, Amsterdam, the Netherlands
[b] Nofima AS, Norwegian Institute of Food, Fisheries and Aquaculture Research, Muninbakken 9-13, NO-9291, Tromsø, Norway
[c] Department of Food Science, University of Copenhagen, Rolighedsvej 26, DK-1958, Frederiksberg, Denmark
[d] SalmoBreed AS, Sandviksboder 3A, NO-5035, Bergen, Norway
[e] Department of Food Science, Aarhus University, Agro Food Park 48, DK-8200, Aarhus N, Denmark
[f] Frontmatec A/S, Hassellunden 9, DK-2765, Smørum, Denmark

ARTICLE INFO

*Keywords:*
Cage of covariance
Regression
Indirect models

ABSTRACT

In analytical chemistry, multivariate calibration is applied when substituting a time-consuming reference measurement (based on e.g. chromatography) with a high-throughput measurement (based on e.g. vibrational spectroscopy). An average error term, of the response variable, is often used to evaluate the performance of a calibration model. However, indirect relationships, between the response and explanatory variables, may be used for calibration. In such cases, model validity cannot necessarily be determined solely by the average error term. One should also consider the use of the models, as well as the validity of the indirect relationships in future samples. If the analyte of interest is partly quantified from signals of interfering compounds, then these interfering compounds will play a hidden role in the calibration. This hidden role may affect future use of the calibration model as strong covariance relationships between analyte estimates and interfering compounds may be imposed. Hence, such model cannot detect changes in the relationship between the analyte and interfering compounds. The problem is called the *cage of covariance*. This paper discusses the concept *cage of covariance* and possible consequences of applying models exposed to this issue.

## 1. Introduction

The desire to rapidly extract large amounts of sample information is natural. Unfortunately, this desire has recently led to misuse of vibrational spectroscopy returning misleading results. Several studies have investigated the possibilities of applying vibrational spectroscopy to acquire detailed sample information, which, normally, is only available through advanced chromatographic separation techniques. Examples from food research are e.g. prediction of fatty acid (FA) composition in bovine milk [1] or determination of biochemical quality parameters in fermented cocoa [2]. In both studies [1,2], more than 30 response variables (reference variables) were estimated from the explanatory variables (spectroscopic measurements). In such cases, it is relevant to ask whether these multiple response variables are estimated independently of one another, or whether the response variables are estimated from identical chemical features in the spectroscopic measurements.

The purpose of establishing a calibration model is to use the model (instead of performing reference analysis) in the future. Normally, it is preferable that calibration models are based on direct relationships between the response variables and the explanatory variables, but it is important to underline that linear regression models represent only covariance relationships. Hence, non-causal relationships may be used to fit regression models.

In the case where multiple response variables have a large amount of variation in common (i.e. high covariance), and this common variation also linearly relates to the explanatory variables, then regression models

* Corresponding author. Institute for Biodiversity and Ecosystem Dynamics, Science Park 904, NL-1098 XH, Amsterdam, the Netherlands.
*E-mail address:* c.e.eskildsen@uva.nl (C.E. Eskildsen).

https://doi.org/10.1016/j.chemolab.2021.104311
Received 4 January 2021; Received in revised form 25 March 2021; Accepted 3 April 2021
Available online 20 April 2021

(regressing the individual responses onto the explanatory variables) may appear good at first glance (small errors). However, if it is solely the variation, which the response variables have in common that is explained by the models, linear dependencies will be forced among the predicted response variables. Consequently, the covariance among the predicted response variables will be higher than the covariance among the *true* response variables. These altered covariance structures, among the predicted response variables, are conserved by the regression models. When applying the models to future data sets, predictions will forever be trapped in a *cage of covariance* and will be linearly dependent of each other [3–5]. This may compromise calibration robustness, validity and for sure interpretation [3].

Intercorrelations (biological, chemical or technical) may exist between response variables and (consistently) lead to very high correlations and excellent predictions even under the strictest validation criteria. This may be called the *cage of biological covariance*. If two response variables are predicted from the same feature in the explanatory variables, the models may be useful if the relationship between the two response variables stays the same (i.e. the *cage of biological covariance* is conserved). However, it is important to be aware that the response variables as such do not independently relate to the observed explanatory variables. Therefore, models solely provide information on how the two response variables are similar in the calibration data. The information on how the two response variables differ from each other is not retained by the models. Hence, this information is lost. This is true also for future predictions.

This is very problematic if predictions are to be used in e.g. an optimization. Consider for instance a breeding program, which aims at breeding cows producing milk with a given ratio between two different FAs. If the FA predictions, estimated from e.g. spectroscopic measurements, depend on each other, it will be impossible to use these predictions in a breeding program pursuing to alter the covariance of the two FAs [3,4]. This will be demonstrated, for the example, in the Results and Discussion section.

To visualize the *cage of covariance*, we use three data sets as examples. The data sets consist of near-infrared (NIR) or mid-infrared (MIR) spectroscopic measurements as explanatory variables and individual FAs as response variables. Individual FAs will absorb both NIR and MIR radiation. However, the signals are almost identical and in a complex sample matrix it will, in practice, be very hard to extract unique spectral features of the individual FAs. Nevertheless, quantities of the individual FAs covary and the individual FAs can therefore be predicted from spectral features associated with total fat content [4].

In such situation, the subspace of the spectroscopic measurements, used for prediction, is relative low rank whereas the matrix containing the FAs is regarded full rank, but with high covariance between individual FAs. We show how individual FAs are predicted with relatively small errors, while the predictions of the different FAs are not independent of each other. Hence, if applying the model in the future to obtain predictions of a given FA, then other FAs will play a hidden role in that prediction. This phenomenon is called the *cage of covariance*. This paper discusses the *cage of covariance* in detail as well as the possible consequences of applying models exposed to the *cage of covariance*.

## 2. Background

In a classical least squares perspective, spectroscopic measurements, $\boldsymbol{X}(n \times m)$, of multi-component samples are viewed as the outer product of the component concentration profiles, $\boldsymbol{C}(n \times q)$, the pure component signals at unitary concentration, $\boldsymbol{S}(m \times q)$, and an error term, $\boldsymbol{E}(n \times m)$, where $n$ is the number of samples, $m$ is the number of measured variables and $q$ is the number of chemical components. This is formalized in Equation (1).

$$\boldsymbol{X} = \boldsymbol{C}\boldsymbol{S}^T + \boldsymbol{E} \qquad \text{Equation 1}$$

In $\boldsymbol{X}$, the number of linearly independent columns is equal to the number of linearly independent rows and this number determines the rank of $\boldsymbol{X}$, $r(\boldsymbol{X})$. Due to measurement noise, the mathematical rank may be close to full. However, the chemical rank is determined by $q$, given that $q \leq \min(n, m)$ and $r(\boldsymbol{C}) = r(\boldsymbol{S}) = q$. Independent information of the $q$ components is not directly available from $\boldsymbol{X}$ if $r(\boldsymbol{X}) < q$.

Say, a high number, $k$, of response variables, $\boldsymbol{Y}(n \times k)$, are regressed onto $\boldsymbol{X}$ (i.e. through $k$ individually fitted regression models). This would return $k$ individual regression vectors, $\widehat{\boldsymbol{B}}(m \times k)$, and the estimate of $\boldsymbol{Y}$, $\widehat{\boldsymbol{Y}}(n \times k)$ is given by Equation (2). Columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are assumed centered around zero.

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{B}} \qquad \text{Equation 2}$$

It may be hard to determine whether response variables are predicted independent of each other ($r(\widehat{\boldsymbol{Y}}) = k$) and through direct relationships with $\boldsymbol{X}$. An obvious prerequisite for this is that $\boldsymbol{Y}$ has full rank ($r(\boldsymbol{Y}) = k$). If $r(\boldsymbol{Y}) < k$, columns of $\boldsymbol{Y}$ are linearly dependent and so are columns of $\widehat{\boldsymbol{Y}}$, i.e. $r(\widehat{\boldsymbol{Y}}) < k$.

Furthermore, for the $k$ response variables to be estimated independently of one another, the individual regression vectors (i.e. columns of $\widehat{\boldsymbol{B}}$) must each describe unique features (directions) in the row space of $\boldsymbol{X}$. In fact, the regression vector for a given response variable should describe the part of the response' signal that is orthogonal to all other signals in $\boldsymbol{X}$ (first-order calibration) [6,7]. In principle, an infinite number of regression models may be fitted to $\boldsymbol{X}$. However, $r(\boldsymbol{X})$ determines the number of independent directions in the $\boldsymbol{X}$-space, and thereby determines the number of independent regression models that possibly can be fitted using $\boldsymbol{X}$ as explanatory variables. As shown in Equation (2), $\widehat{\boldsymbol{Y}}$ is a linear combination of $\boldsymbol{X}$. Therefore, $\widehat{\boldsymbol{Y}}$ will always be in the $\boldsymbol{X}$-space and independent estimates of $k$ response variables cannot be based on explanatory variables, which total a rank less than $k$.

Moreover, when extracting information from $\boldsymbol{X}$ the signal-to-noise ratio is important. This is especially true when $q$ is high and columns of $\boldsymbol{C}$ and/or $\boldsymbol{S}$ covary calling for complex partial least squares (PLS) regression models fitted with many latent variables. Before estimating a new latent variable, $\boldsymbol{X}$ is deflated by the variation explained by the previous latent variable [8]. This will deteriorate the signal-to-noise ratio of the data and subsequent latent variables are estimated with larger uncertainties. For complex data with a poor signal-to-noise ratio, this may result in underfitted PLS regression models. A regression vector from an underfitted PLS regression model may not be orthogonal to the signals of all interfering compounds. Therefore, analyte predictions will partly depend on signals from interfering compounds. This will, in turn, force linear dependencies among the predictions of the analyte and interfering compounds. Hence, future predictions are (forever) trapped in a *cage of covariance*. This may compromise calibration robustness and validity.

## 3. Materials and methods

### 3.1. Bovine milk samples

Eight hundred ninety milk samples from individual Jersey and Holstein cows were included. Mid-infrared measurements were obtained using MilkoScan FT2 (Foss Analytical A/S, Hillerød, Denmark). The spectroscopic measurements originated from Eskildsen et al. (2014) [4]. Each sample was measured in triplicates and the average spectrum was used for further analysis. The spectral regions from 2,968 $cm^{-1}$ to 2,802 $cm^{-1}$, 1,773 $cm^{-1}$ to 1,692 $cm^{-1}$ and 1,604 $cm^{-1}$ to 925 $cm^{-1}$ were included [4].

In order to approximately obey *Beer's law*, the MIR spectra were transformed from transmittance (T) units to absorbance ($A \approx \log(1/T)$) and preprocessed by Savitzky-Golay [9,10] first derivative (window size of 9 points and second-order polynomial), as suggested by Eskildsen et al. (2014) [4].

As response variables, FAs ($k = 12$) were quantified by gas chromatography of FA methyl esters (GC-FAME) as described by Poulsen et al. (2012) [11]. All FAs were expressed in units of g of FA/100 g of milk. For a comprehensive sample description, see Poulsen et al. (2012) [11]. Furthermore, total fat, protein and lactose content were quantified using MilkoScan FT2 and the commercial calibration models (FOSS Analytical A/S, Hillerød, Denmark). Hence, total fat, protein and lactose were linear combinations of the MIR spectra. To create a more realistic relationship between the MIR spectra and these three response variables, white Gaussian noise was added to the observed values, $\boldsymbol{y}^{obs}$, of total fat, protein and lactose (Equation (3)).

$$\boldsymbol{y} = \boldsymbol{y}^{obs} + \Delta\boldsymbol{y}, \Delta\boldsymbol{y} \sim N\left(0, 0.1 \cdot var\left(\boldsymbol{y}^{obs}\right)\right) \quad \text{Equation 3}$$

### 3.2. Atlantic salmon muscle samples

Six hundred sixty-three samples from the Norwegian Quality Cut of the filets obtained from individual Atlantic salmon of the SalmoBreed population were included. Atlantic salmon samples originated from Horn et al. (2018) [12]. The samples were homogenized and measured in mini sample cups (FOSS Analytical A/S, Hillerød, Denmark) using FOSS NIR Systems XDS Rapid Content™ Analyzer (FOSS Analytical A/S, Hillerød, Denmark). The NIR measurements were obtained in reflectance mode with 32 scans per spectrum. Samples were measured in triplicates and the average spectrum was used for further analysis. An internal ceramic standard was used as reference. The spectral range was from 1,100 nm to 2,500 nm. In order to approximately obey *Beer's law*, the NIR spectra were transformed from reflectance (R) units to absorbance ($A \approx \log(1/R)$) and preprocessed using extended multiplicative signal correction (EMSC) [13,14]. The EMSC was performed using the mean spectrum and a second-order polynomial as suggested by Eskildsen et al. (2019) [7].

As response variables, 22 FAs ($k = 22$) were quantified using GC-FAME as described by Horn et al. (2018) [12]. Total fat was quantified as stated in Horn et al. (2018) [12] and all FAs were expressed in units of g of FA/100 g muscle. For a comprehensive sample description, see Horn et al. (2018) [12].

### 3.3. Pork samples

One hundred slaughter pig carcasses from an industrial abattoir (Danish Crown, Herning, Denmark) were included. Approximately 30 min after slaughter, each carcass was measured with the NitFom™ device (Frontmatec A/S, Smørum, Denmark) to obtain transmission measurements of the subcutaneous fat layer. The measurements were made in the neck region of the back fat next to the shoulder blade – approximately 7 cm from the split line. The NitFom™ consisted of stainless-steel twin-probes mounted in a probe house and spaced by 2 mm [15]. Each probe was knife-tipped and designed to penetrate 3 cm into the carcass through the skin. Optical fibers connected the emitter probe to a light source while the receiver probe was connected to a NIR-spectrometer. The two probes have windows facing each other allowing light to be transmitted through the adipose tissue. As the probe head was ejecting itself from the carcass, NIR transmission spectra were recorded at several depths. The spectral range was from 1,100 nm to 1,900 nm. A built-in algorithm facilitated differentiation of tissue types (meat vs. fat). Only spectra recorded in adipose tissue were used.

In order to approximately obey *Beer's law*, the spectra were transformed from transmittance (T) into absorbance ($A \approx \log(1/T)$) and preprocessed using EMSC. The EMSC was done using the mean spectrum and a second-order polynomial as suggested by Sørensen et al. (2012) [16].

As response variables, FAs ($k = 4$) were quantified by GC-FAME at Danish Technological Institute - Danish Meat Research Institute (Taastrup, Denmark). All FAs were expressed in units of g of FA/100 g of fat tissue.

### 3.4. Data analysis

Data were analyzed using MATLAB version R2016b (9.1.0.441655, MathWorks Inc., Natick, MA, USA). Prior to modeling, the spectroscopic measurements were preprocessed and mean centered. Fatty acids were mean centered and scaled to unit variance. The non-linear iterative partial least squares algorithm was used for PLS regression [8,17]. 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 significance test [18] as described in the subsequent paragraph. Only well predicted FAs are included in this study. A threshold was decided at 60% explained variation (i.e. only FAs with $R^2 \geq 0.6$ between measured and predicted FA, are included in this study). To investigate the covariance structures in the data sets, data were decomposed by singular value decomposition (SVD) and the percent explained variation as a function of number of latent variables was used to determine whether the covariance structures were altered in the predicted response variables as compared to the reference values.

#### 3.4.1. Number of latent variables in PLS models

In order to evaluate statistical significance of each individual latent variable that enters the PLS models, the procedure proposed by Wiklund et al. (2007) was used [18]. In short, a PLS model is fitted between $\boldsymbol{X}$ and $\boldsymbol{y}$ and the test statistics is calculated as the covariance between the scores of the latent variable and the $\boldsymbol{y}$-values. Then the null-distribution (i.e. the distribution that holds when the latent variable is not significant) is created by permuting $\boldsymbol{y}$. A PLS model is fitted between $\boldsymbol{X}$ and the permuted $\boldsymbol{y}$ and the test statistics is calculated. This is repeated 500 times to create the null-distribution. The critical value is derived as the value exceeded by 5% of the values in the null-distribution. Finally, the test statistics obtained for the original data is compared to the critical value, to evaluate whether the latent variable is significant or not. Both $\boldsymbol{X}$ and $\boldsymbol{y}$ are deflated before evaluating the subsequent latent variable. See Wiklund et al. (2007) for comprehensive description and computational details [18].

## 4. Results and discussions

Fig. 1 presents the preprocessed spectral data of the three datasets. The MIR spectra obtained on bovine milk (Fig. 1A) primarily contain information on fat, protein and carbohydrates since the strong water band regions are excluded [3,4,19]. The NIR spectra obtained on Atlantic salmon muscle (Fig. 1B) primarily contain information on water, protein and fat. The NIR spectra obtained on the porcine adipose tissue (Fig. 1C) primarily contain information on fat and water [20].

Total fat, protein and lactose content are known to be well predicted from MIR measurements obtained on milk [19]. This is also found in this study. Table S1 (Supporting Information) presents the descriptive statistics for total fat, protein and lactose in bovine milk and Table S2 (Supporting Information) presents the results from PLS models of the milk data for total fat, protein and lactose. Total fat, protein and lactose give rise to distinct spectral features in the MIR spectra [19]. Therefore, these parameters may be estimated independent of each other (even though total fat and protein content are fairly correlated in this present bovine milk data [3]).

Fig. 2 presents a heat map showing coefficients of determination ($R^2$) between total fat, protein and lactose in the milk dataset. The $R^2$-values between columns of $\boldsymbol{Y}$ are presented below the diagonal, whereas $R^2$-values between $\widehat{\boldsymbol{Y}}$ are presented above the diagonal. The $R^2$-values between reference and predicted values (i.e. explained variation of the regression models, also presented in Supporting Information, Table S2) are on the diagonal. Fig. 2 shows that the correlation structures in $\boldsymbol{Y}$ is not altered in $\widehat{\boldsymbol{Y}}$. Hence, total fat, protein and lactose are estimated independent of each other from different chemical bases in the MIR measurements.

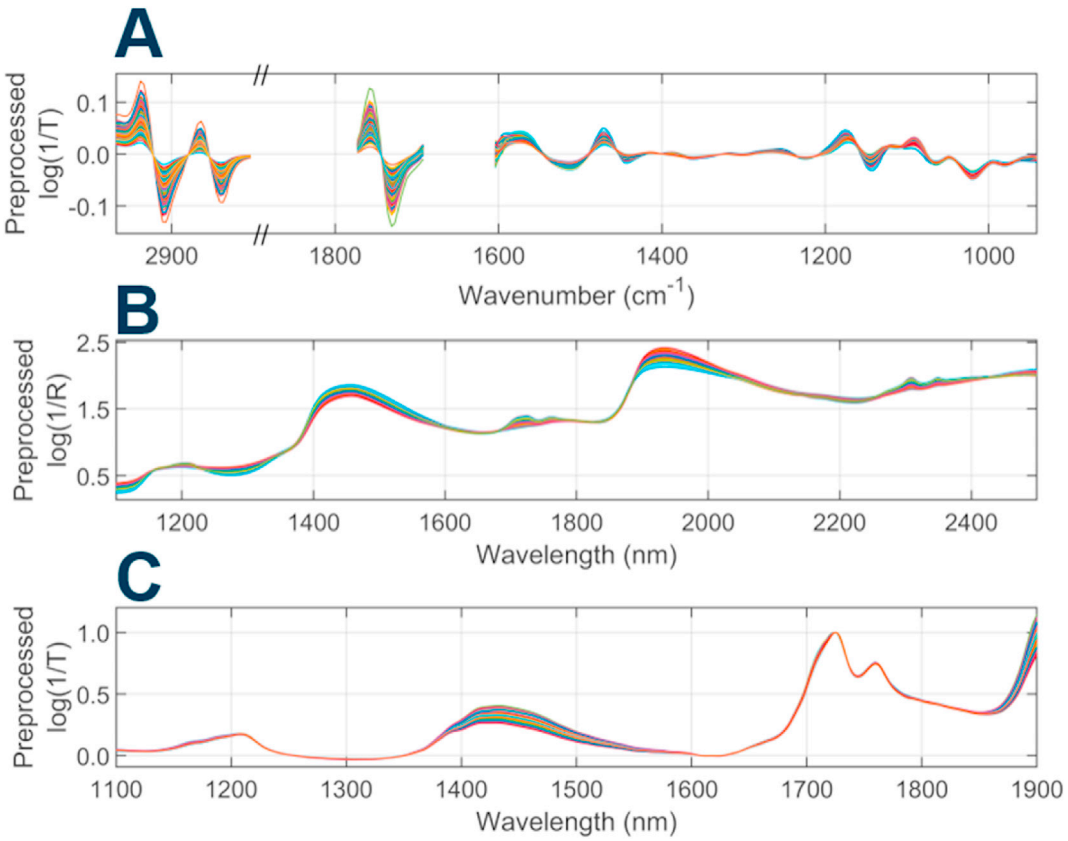


**Fig. 1.** Preprocessed spectral measurements. A) Bovine milk, B) Atlantic salmon muscle, C) Porcine adipose tissue. R = reflection, T = transmission.

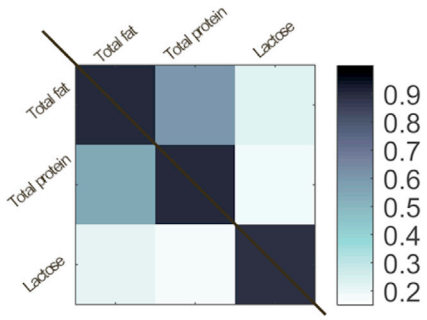


**Fig. 2.** Heat map showing coefficients of determination ($R^2$) between total fat, protein and lactose. Below the diagonal: $R^2$ between reference values. On the diagonal: $R^2$ between reference and predicted values. Above the diagonal: $R^2$ between predicted values. Data originate from the bovine milk data set.

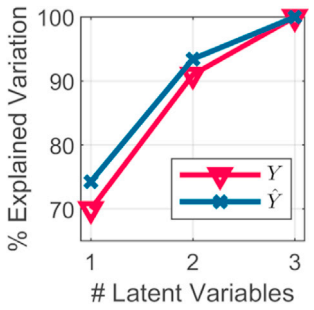


**Fig. 3.** Cumulative % explained variation as a function of latent variables included in the model. Singular value decomposition performed on matrices containing total fat, protein and lactose ($\boldsymbol{Y}$) and the predicted total content of fat, protein and lactose ($\widehat{\boldsymbol{Y}}$). Data originate from the bovine milk data set.

This observation is confirmed in Fig. 3, which shows the results from SVD of total fat, protein and lactose (milk data set). The SVD is performed on the reference values, $\boldsymbol{Y}$ and the predicted values of total fat, protein and lactose, $\widehat{\boldsymbol{Y}}$. Fig. 3 shows that columns of $\boldsymbol{Y}$ are correlated. The first latent variable describes approximately 70% of the total variation in $\boldsymbol{Y}$. The SVD of $\widehat{\boldsymbol{Y}}$ reveals that the covariance structures of $\widehat{\boldsymbol{Y}}$ are very similar to the covariance structures of $\boldsymbol{Y}$, as the individual latent variables approximately describe the same amount of relative variation. This indicates that total fat, protein and lactose are estimated independent of one another. Hence, predictions are not trapped in the *cage of covariance*. The total explained variation and the covariance structures in $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$ are directly comparable. Therefore, the pattern, in Fig. 3, for $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$ must be similar if columns of $\widehat{\boldsymbol{Y}}$ independently relate to $\boldsymbol{X}$.

The descriptive statistics for FAs reference data presented in Table S3 (Supporting Information) and the results from PLS models are presented in Table S4 (Supporting Information). The complexities of individual PLS models are determined by the permutation procedure described in the methods section [18]. It was observed that a given PLS latent variable may appear insignificant, whereas successive latent variables are significant. Nevertheless, PLS models are fitted with a complexity corresponding to the last significant latent variable (i.e. all successive latent variables are insignificant). In this way, under-fitting the PLS models is avoided. Under-fitting the individual PLS models may decrease $r(\widehat{\boldsymbol{Y}})$ and thereby force a severe *cage of covariance* among the predicted FAs (i.e. columns of $\widehat{\boldsymbol{Y}}$). On the other hand, over-fitting the individual PLS models may include random noise in $\widehat{\boldsymbol{Y}}$ and thereby increase $r(\widehat{\boldsymbol{Y}})$ making it difficult to detect a possible *cage of covariance* among the columns of $\widehat{\boldsymbol{Y}}$. However, in this study the risk of over-fitting the PLS models (by including insignificant latent variables) is accepted in order to be confident that the models are not under-fitted. Therefore, a high number of

latent variables are included in the models (Supporting Information, Table S4) as compared to e.g. the PLS models in Eskildsen et al. (2014) [4]. In general, calibration models are fitted with the purpose of predicting external samples. This purpose is clearly compromised when over-fitting the models. Therefore, prediction errors presented in Table S4 are likely to be optimistic.

At first glance, FAs, presented in Table S4 (Supporting Information), appear to be well predicted. In general, we observe better predictions of FAs with higher variation. Examples from the Atlantic salmon muscle data set are C16:0 and C18:1n9. Vibrational spectroscopy is generally sensitive to functional groups, but cannot be used to distinguish between molecular species with similar functional groups (i.e. methine, methylene, methyl, carbonyl, C=C). It is thus not possible to distinguish between FA species in a complex mixture. Hence, the FA predictions depend on how well FA quantities covaries with the spectral features of the functional groups. The FAs with higher variation will dominate the ratio changes of the functional groups and thereby also the variation of the spectral features associated with these groups. Therefore, those FAs are likely to be better predicted.

Fig. 4 presents heat maps showing $R^2$ between the individual FAs of the three datasets. Fig. 4A presents the bovine milk data, Fig. 4B presents the Atlantic salmon data and Fig. 4C presents the porcine adipose tissue data. The $R^2$-values between reference FAs ($\boldsymbol{Y}$) are presented below the diagonal, whereas $R^2$-values among the predicted FAs ($\widehat{\boldsymbol{Y}}$) are presented above the diagonal. The $R^2$-values between reference and predicted FAs (i.e. explained variation of the regression models, also presented in Supporting Information, Table S4) are on the diagonal.

For all three data sets, the covariances among predicted FAs is higher than the covariances between reference FAs. This is seen as more dark-colored pixels above the diagonals as compared to below the diagonals in the heat maps (Fig. 4). This indicates that the FAs are (partly) predicted from the same chemical features in $\boldsymbol{X}$. Linear dependencies are therefore imposed among columns of $\widehat{\boldsymbol{Y}}$ and FAs are not predicted independently of each other.

This is also confirmed in Fig. 5, which shows the results from singular value decomposition of $\boldsymbol{Y}$ (i.e. reference FAs) and $\widehat{\boldsymbol{Y}}$ (i.e. predicted FAs) in the bovine milk data (Fig. 5A), the Atlantic salmon muscle data (Fig. 5B) and the porcine adipose tissue data (Fig. 5C). In all three data sets, the first latent variable explains a substantial amount of the total variation in $\boldsymbol{Y}$. This is expected, as the FAs are highly collinear. However, for all three data sets, the first latent variable explains a substantial higher amount of variation when decomposing $\widehat{\boldsymbol{Y}}$. Hence, the covariance structures are stronger in $\widehat{\boldsymbol{Y}}$ than in $\boldsymbol{Y}$. This indicates that the individual PLS models, predicting individual FAs, are largely using the same information in the spectral data. Hence, the PLS models will impose stronger covariance structures among the FAs as compared to what is the actual case. These stronger covariance relationships are conserved by the models. Therefore, future predictions will always be trapped in this *cage of covariance*.

The PLS regression models do not relate the FAs to unique spectral information. Therefore, the predictions largely contain information on how the FAs are similar. Information on how the FAs differ from each other, is not preserved by the models. The calibration models provide information on how the FAs covaries in the calibration set. Hence, if the covariance structures among the FAs changes in a future dataset, this will not be reflected by the predictions. The covariance structures among future predictions will largely be the same as the covariance structures among the predictions in the calibration set - the *cage of covariance*. This is a consequence of calibrating regression models on non-causal relationships. Even though the FAs are reasonably well predicted (Supporting Information, Table S4), the linear dependencies imposed among columns of $\widehat{\boldsymbol{Y}}$ may be problematic when the regression models are used to explore future indeoendent data sets.

Imagine, for example, a breeding program aiming at altering the ratio of FAs C14:0 to C6:0 in bovine milk. Both C6:0 and C14:0 appears to be well predicted from the regression models applied to MIR spectra of bovine milk (Fig. 6A and B, respectively, and Table S4). One is interested in identifying cows producing milk with a higher C14:0 to C6:0 ratio. From the GC-FAME measurements (Fig. 6C), which are not exposed to *cage of covariance*, three milk samples are marked (ID 68, 757 and 762) as examples. These three samples are among the samples with a higher C14:0 to C6:0 ratio and are therefore interesting in the imaginary breeding program. However, if we substitute the GC-FAME measurements with MIR spectroscopy, due to the *cage of covariance* conserved by the regression models, the three samples are no longer identified as samples with a higher C14:0 to C6:0 ratio (Fig. 6D). In fact, ID 757 will be identified as samples with a C14:0 to C6:0 ratio below average. Hence, FA estimates from MIR measurements are of very limited use in the breeding program, as they simply do not provide the information that is sought for.

In this paper we use PLS with univariate reference values (i.e. $\boldsymbol{y}$-block). The *cage of covariance* problem seems to be inherent in the data and cannot necessarily be solved with other regression methods. We also applied PLS2 with multivariate reference values (i.e. $\boldsymbol{Y}$-block) to the data presented. However, it appears that PLS2 also impose the *cage of covariance* problem (results not shown). The reason is that in PLS2 the$\boldsymbol{Y}$-variables are modeled by the same latent variables, which aim at explaining common variation among the $\boldsymbol{Y}$-variables.

## 5. Conclusions

This study shows the importance of considering the rank of the subspace of explanatory variables used for prediction, the covariance structures of response variables as well as their estimates when regressing multiple response variables (e.g. FAs) onto the same explanatory variables (e.g. spectroscopic measurements). This study estimates multiple response variables from a lower rank subspace of explanatory variables. At first glance, when evaluated by e.g. percent explained variation, the response variables seem to be well predicted. Nevertheless, the covariance structures between the estimated response variables are higher than that of the *true* response variables. This is due to the fact that predicted

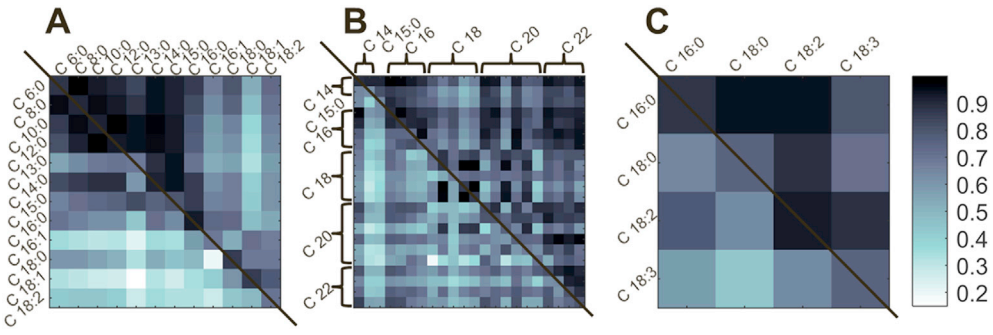


**Fig. 4.** Heat map showing coefficients of determination ($R^2$) between the fatty acids. Below the diagonal: $R^2$ between reference values. On the diagonal: $R^2$ between reference and predicted values. Above the diagonal: $R^2$ between predicted values. A) Bovine milk, B) Atlantic salmon muscle, C) Porcine adipose tissue.

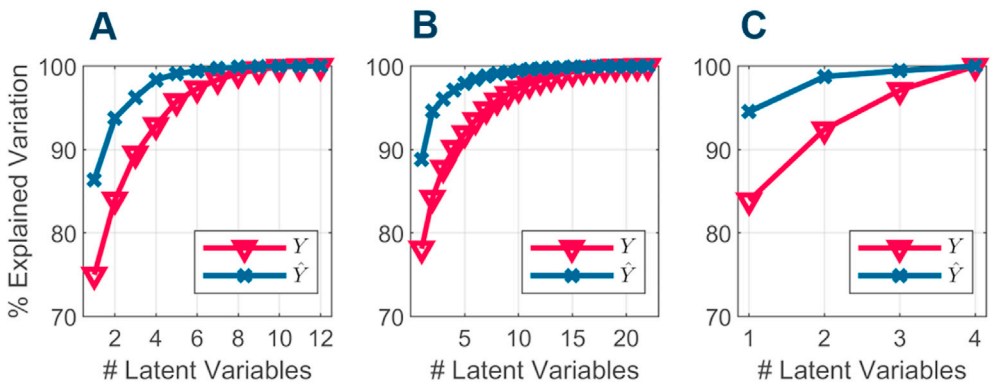


**Fig. 5.** Cumulative % explained variation as a function of latent variables included in the model. Singular value decomposition performed on matrices containing the reference fatty acidss ($\boldsymbol{Y}$) and the predicted fatty acids ($\hat{\boldsymbol{Y}}$), respectively. A) Bovine milk, B) Atlantic salmon muscle, C) Porcine adipose tissue.

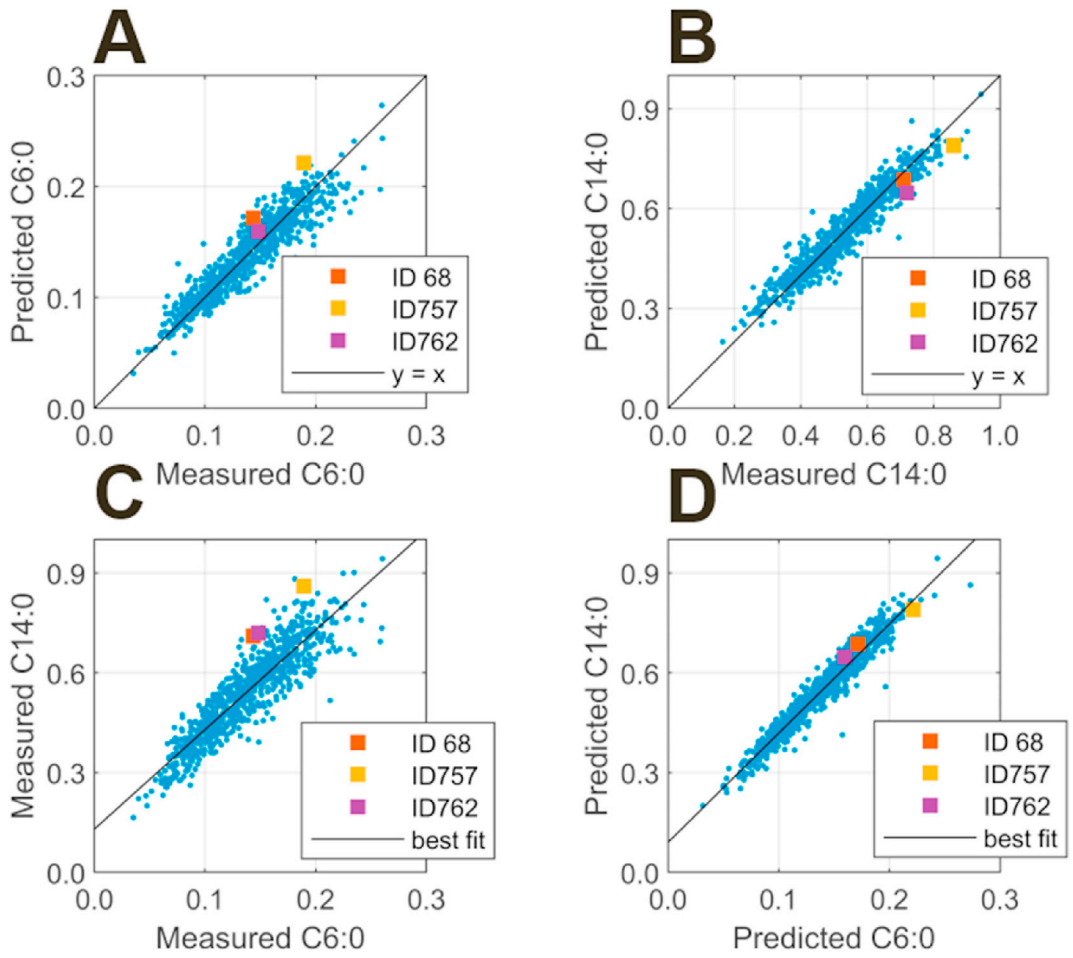


**Fig. 6.** Predictions (bovine milk data) exposed to the *cage of covariance*. A) Predicted versus measured values of C6:0. B) Predicted versus measured values of C14:0. C) Measured C14:0 versus measured C6:0. D) Predicted C14:0 versus predicted C6:0.

response variables are located in a common lower rank subspace of the explanatory variables. Even though the *true* response variables are viewed as independent of each other, the predicted response variables cannot be viewed as independent, as they depend on the common lower rank subspace of the explanatory variables. Hence, the predictions are not based on chemical information directly associated with the individual responses. This may compromise the validity and robustness of the calibration models, as the predicted responses are caught in a *cage of covariance* with each other. This may lead to serious misinterpretation of the studied system if estimated responses are used for optimization purposes, where the target is to break the *cage of biological covariance*, like in breeding programs. This is of fundamental importance as it reduces the chance for new discoveries. In this paper, we discussed the problems of calibrating regression models on non-causal relationships in the case of multiple response variables. However, the problems caused by non-causal relationships are obviously also present in the case where only one response variable is predicted.

## 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.

## Acknowledgements

For funding we acknowledge the Norwegian Agricultural Food Research Foundation through the project FoodSMaCK – Spectroscopy, Modelling & 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.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2021.104311.

## References

[1] H.G. Olsen, T.M. Knutsen, A. Kohler, M. Svendsen, L. Gidskehaug, H. Grove, T. Nome, M. Sodeland, K.K. Sundsaasen, M.P. Kent, H. Martens, S. Lien, Genome-wide association mapping for milk fat composition and fine mapping of a QTL for de novo synthesis of milk fatty acids on bovine chromosome 13, Genet. Sel. Evol. 49 (1) (2017) 20.

[2] Krähmer, A. Engel, D. Kadow, N. Ali, P. Umaharan, L.W. Kroh, H. Schulz, "Fast and neat – determination of biochemical quality parameters on cocoa using near infrared spectroscopy", Food Chem. 181 (2015) 152–159.

[3] 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.

[4] 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 prediction of highly collinear reference variables", J. Dairy Sci. 97 (2014) 7940–7951.

[5] D.T. Berhe, C.E. Eskildsen, R. Lametsch, M.S. Hviid, F.v.d. Berg, S.B. Engelsen, Prediction of total fatty acid parameters and individual fatty acids in pork backfat using Raman spectroscopy and chemometrics: understanding the cage of covariance between highly correlated fat parameters, Meat Sci. 111 (2016) 18–26.

[6] E. Sanchez, B.R. Kowalski, Tensorial calibration: I. First-order calibration", J. Chemom. 2 (1988) 247–263.

[7] 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. van den Berg (Eds.), Proc. 18th Int. Conf. Near Infrared Spectrosc, IM Publications Open, Chichester, UK, 2019, pp. 39–44.

[8] S. Wold, A. Ruhe, H. Wold, W.J. Dunn III, The collinearity problem in regression, the partial least squares (pls) approach to generalized inverses, SIAM J. Sci. Stat. Comput. 5 (1984) 735–743.

[9] Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures", Anal. Chem. 33 (8) (1964) 1627–1639.

[10] J. Steiner, Y. Termonia, J. Deltour, Smoothing and differentiation of data by simplified least squares procedure, Anal. Chem. 44 (11) (1972) 1906–1909.

[11] N.A. Poulsen, F. Gustavson, M. Glantz, M. Paulsson, L.L. Larsen, M.K. Larsen, The influence of feed and herd on fatty acid composition in 3 dairy breeds (Danish Holstein, Danish Jersey, and Swedish Red), J. Dairy Sci. 95 (2012) 6362–6371.

[12] S.S. Horn, B. Ruyter, T.H.E. Meuwissen, B. Hillestad, A.K. Sonesson, Genetic effects of fatty acid compostition in muscle of Atlantic salmon, Genet. Sel. Evol. 50 (2018) 23.

[13] H. Martens, E. Stark, Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near-infrared spectroscopy", J. Pharmaceut. Biomed. Anal. 9 (8) (1991) 625–635.

[14] N.K. Afseth, A. Kohler, Extended multiplicative signal correction in vibrational spectroscopy, a tutorial, Chemometr. Intell. Lab. Syst. 117 (2012) 92–99.

[15] H. Marno and K.M. Sørensen. "Recording of position-specific wavelength absorption spectra". United States patent US8,530,844 B2, 2013.

[16] K.M. Sørensen, H. Petersen, S.B. Engelsen, An on-line near-infrared (nir) transmission method for determining depth profiles of fatty acid composition and iodine value in porcine adipose fat tissue", Appl. Spectrosc. 66 (2) (2012) 218–226.

[17] M. Andersson, A Comparison of nine PLS1 algorithms, J. Chemom. 23 (2009) 518–529.

[18] S. Wiklund, D. Nilsson, L. Eriksen, M. Sjöström, S. Wold, K. Faber, A randomization test for PLS component selection", J. Chemom. 21 (2007) 427–439.

[19] H. Luinge, E. Hop, E. Lutz, J. Vanhemert, E. Dejong, Determination of fat, protein and lactose content in milk using Fourier transform infrared spectrometry, Anal. Chim. Acta 284 (1993) 419–433.

[20] K.M. Sørensen, S.B. Engelsen, The spatial composition of porcine adipose tissue investigated by multivariate curve resolution of near infrared spectra: relationship between fat, the degree of unsaturation and water", J. Near Infrared Spectrosc. 25 (1) (2017) 45–53.