Visualizing indirect correlations when predicting fatty acid composition from near infrared spectroscopy measurements

C.E. Eskildsen,[a,*] T. Næs,[a,b] J.P. Wold,[a] N.K. Afseth[a] and S.B. Engelsen[b]

[a]Nofima, Norwegian Institute for Food and Fisheries Research, NO-1433 Ås, Norway
[b]Department of Food Science, University of Copenhagen, DK-1958 Frederiksberg, Denmark. E-mail: carl.eskildsen@nofima.no

In recent years, vibrational spectroscopy has been used to predict detailed sample composition like protein and fatty acid profiles. This study shows that fatty acid predictions from near infrared measurements in food stuffs rely on covariance structures amongst the fatty acids. These covariance structures, in turn, vary with factors like breed, age, feed, season etc. and therefore they are not likely to remain constant. Consequently, the robustness and validity of the developed calibration models will be compromised.

Introduction

The food industry rapidly moves toward circular economy with optimal exploitation of waste streams and increasing productivity while retaining quality and safety demands. Analysis of raw materials, real-time process control and end-product quality evaluations are crucial steps in reaching the desired product quality in a cost-effective way. Collecting the right information sufficiently fast is key for increasing production throughput. Near infrared spectroscopy (NIRS) has great potential for monitoring of food processes and commonly we find *Near Infrared Spectroscopy at Work in the Food Industry*.[1]

For decades, NIRS has been used to quantify bulk protein, fat etc. in the food industry. However, in recent years, requests for more detailed information have increased. For example, in cheese making, the protein composition is essential[2] and in a similar manner, the fatty acid composition is important to the sliceability of bacon.[3] Several studies have suggested vibrational spectroscopy as a successful tool for providing such detailed information.

A number of studies have reported good predictions of individual fatty acids from vibrational spectroscopic measurements. However, these predictions are almost exclusively the result of strong covariance structures in the collected data. Strong covariance structures among sample properties are likely to exist in biological samples. For instance, an increase of individual fatty acids causes an increase of total fat content (%FAT). This may enable predictions of individual fatty acids, from NIRS measurements, through an indirect relationship with %FAT, as sketched in Figure 1.

The problem is that model estimates of individual fatty acids will contain variation dictated by %FAT and possibly from

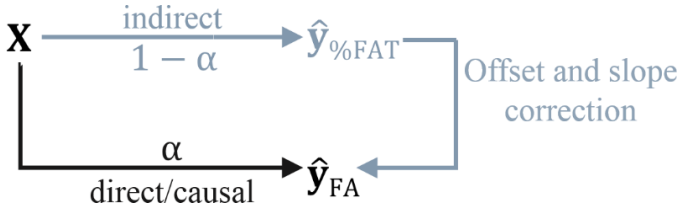

Figure 1. Prediction of a given fatty acid, $\hat{y}_{FA}$ from near infrared measurements, X. Variation in $\hat{y}_{FA}$ can be described through a causal relationship with X or through an indirect relationship with the fat percentage, $\hat{y}_{\%FAT}$. The amount of variation in $\hat{y}_{FA}$ remaining from the causal relationship with X is given by α. Figure modified from Eskildsen *et al.*[6]

Correspondence
C.E. Eskildsen (carl.eskildsen@nofima.no)
doi: 10.1255/nir2017.039
Citation: C.E. Eskildsen, T. Næs, J.P. Wold, N.K. Afseth and S.B. Engelsen, "Visualizing indirect correlations when predicting fatty acid composition from near infrared spectroscopy measurements", in *Proc. 18th Int. Conf. Near Infrared Spectrosc.*, Ed by S.B. Engelsen, K.M. Sørensen and F. van den Berg. IM Publications Open, Chichester, pp. 39–44 (2019). https://doi.org/10.1255/nir2017.039

ISBN: 978-1-906715-27-4

some variation related to groups of fatty acids (unsaturated fats, conjugated fats, branched fats etc.). Hence, model estimates will largely depend on %FAT. Therefore, fatty acid estimates do not contain much information on fatty acid composition (variation orthogonal to %FAT) and certainly not information about all single fatty acids, which most often is of interest.

Whereas problems relating to covariances amongst independent variables (i.e. spectral variables) are well understood,[4,5] covariances amongst dependent variables (i.e. reference variables or sample properties) have only received minor attention.[6–9] This study discusses issues of regression modeling when strong covariance structures exist in the reference data and how to visualize (diagnose) when these covariance structures become dominating during regression modeling.

An illustration of the problem

The following section is an illustration of the underlying problem using a simple two constituent *Beer's law* model.

In Figure 2a, the sample signal (spectrum), $\mathbf{x}(1\times 2)$, is composed by analyte signal, $\mathbf{s}_1(1\times 2)$ and interfering compound signal, $\mathbf{s}_2(1\times 2)$. Here, $\mathbf{s}_1$ and $\mathbf{s}_2$ are at unitary concentration and $\mathbf{x}$ is given by,

$$\mathbf{x} = c_1\mathbf{s}_1 + c_2\mathbf{s}_2 \tag{1}$$

where $c_1(1\times 1)$ and $c_2(1\times 1)$ are concentrations of the analyte and interfering compound, respectively.

Any linear regression model has the form,

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \tag{2}$$

where $\hat{\mathbf{y}}(n\times 1)$ is concentration estimates (centered), $\mathbf{X}(n\times m)$ is spectra (pre-processed and centered) and $\mathbf{b}(m\times 1)$ is the true regression vector. Concentration estimates are simply dot products of sample spectra and the regression vector. Hence, the estimate of c_1, $\hat{c}_1$, is,

$$\hat{c}_1 = \mathbf{x}\mathbf{b}_1 = |\mathbf{x}| \bullet |\mathbf{b}_1| \bullet \cos(\theta) \tag{3}$$

where $\mathbf{b}_1$ is the true regression vector for the analyte, $|\mathbf{x}|$ and $|\mathbf{b}_1|$ are lengths of $\mathbf{x}$ and $\mathbf{b}_1$, respectively, and θ is the angle between $\mathbf{x}$ and $\mathbf{b}_1$ (Figure 2a). Since,

$$\cos(\theta) = \frac{|\mathbf{proj}\,\mathbf{x}|}{|\mathbf{x}|} \tag{4}$$

where $|\mathbf{proj}\mathbf{x}|$ is the length of the projection of $\mathbf{x}$ onto $\mathbf{b}_1$,

$$\hat{c}_1 = |\mathbf{b}_1| \bullet |\mathbf{proj}\,\mathbf{x}| \tag{5}$$

In order for $\hat{c}_1$ to be independent of c_2, $|\mathbf{proj}\mathbf{x}|$ should not be affected by varying c_2. This is ensured by fitting $\mathbf{b}_1$ in the direction of $\mathbf{s}_1$, while being orthogonal to $\mathbf{s}_2$,[10] as shown in Figure 2a. Figure 2b shows the exact same two constituent system as Figure 2a. However, in Figure 2b the incorrectly estimated regression vector, $\hat{\mathbf{b}}_1$, is partly pointing in the direction of $\mathbf{s}_2$. As shown in Figure 2b, c_2 is then affecting the length of the projection of $\mathbf{x}$ onto $\hat{\mathbf{b}}_1$. From Equation 5 it becomes clear that analyte estimates depend on c_2, if the regression vector is estimated partly in the direction of $\mathbf{s}_2$.

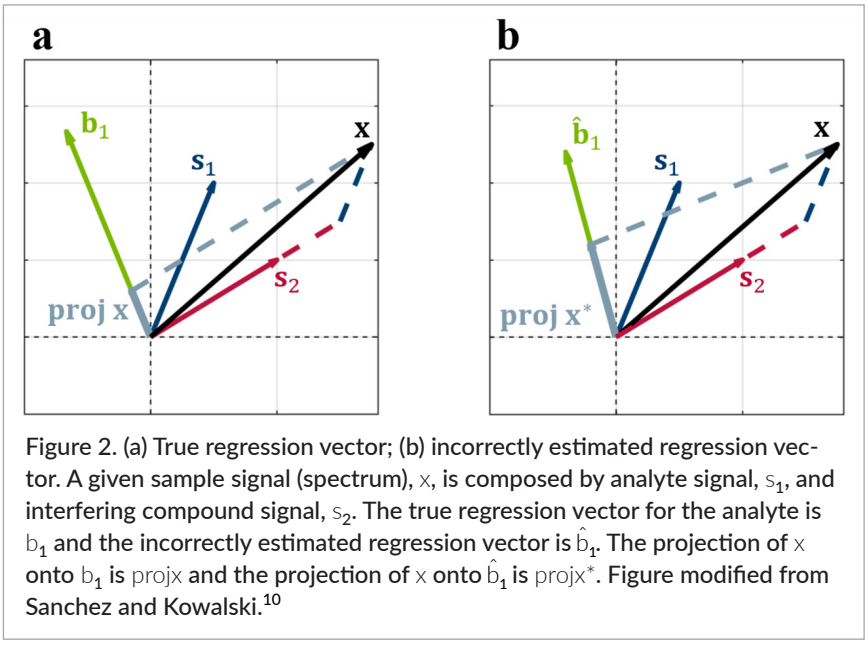

Figure 2. (a) True regression vector; (b) incorrectly estimated regression vector. A given sample signal (spectrum), x, is composed by analyte signal, s_1, and interfering compound signal, s_2. The true regression vector for the analyte is b_1 and the incorrectly estimated regression vector is $\hat{b}_1$. The projection of x onto b_1 is projx and the projection of x onto $\hat{b}_1$ is projx*. Figure modified from Sanchez and Kowalski.[10]

Estimating individual fatty acids from NIRS measurements, in a direct manner, require each fatty acid to have a unique signal. In this context, the chemical rank of the NIRS measurements is important. The chemical rank defines the number of (meaningful) orthogonal directions in the spectra, i.e. it describes how many analytes are possessing unique spectral signals. If the chemical rank is lower than the number of fatty acids, some fatty acid estimates will depend on indirect correlations to interfering compounds. In this study, we use principal component analysis (PCA) to estimate the chemical rank. The number of latent variables needed to approximate the data gives the chemical rank.

Materials and methods

Salmon samples

A total number of 240 samples from individual salmons were included. Samples were homogenized, and total lipids were extracted from homogenized muscle samples of individual fish.[11] Fatty acids were quantified using gas chromatography, following the procedure of Manson and Waller.[12] A total number of 33 individual fatty acids were included in this study. Furthermore, %FAT is included. All fatty acids and %FAT are expressed in units of g/100 g sample.

Near infrared spectroscopy measurements

The NIRS measurements were obtained in reflectance mode (32 scans) using a FOSS NIRSystems XDS Rapid Content™ Analyzer (FOSS Analytical A/S, Hillerød, Denmark). The homogenized filets were measured in mini sample cups (FOSS Analytical A/S, Hillerød, Denmark). An internal ceramic standard was used as reference. Each sample spectrum was acquired in triplicates and the average spectrum was used for further analysis. The spectral range was from 400 nm to 2500 nm with a resolution of 0.5 nm. However, the spectral range included in the present study was from 1100 nm to 2500 nm.

Data analysis

Data were analyzed using MATLAB version R2016b (9.1.0.441655, MathWorks Inc., Natrick, MA, USA). In order to obey *Beer's law*, the NIR spectra were transformed from reflectance (R) units into absorbance-like units [log(1/R)] and preprocessed by extended multiplicative signal correction.[13] Prior to modeling, the NIR spectra were additionally mean centered and fatty acids were mean centered and scaled to unit variance. The nonlinear iterative partial least squares algorithm[14] was used for partial least squares (PLS) regression. All PLS models were built with univariate reference values (i.e. **y**-block) and cross-validated using the venetian blinds method with five data splits. Data were decomposed by singular value decomposition during PCA.

Results and discussions

Figure 3a shows the raw and Figure 3b shows the pre-processed [log(1/R)] NIR spectra recorded on the minced salmon filets. The PLS models (fitted to the NIRS measurements) provided, in general, predictions of individual fatty acids and %FAT with low errors.

Figure 4a shows the root mean squared error of cross validation for the prediction of C16:0 and Figure 4b shows the relationship between measured and predicted values of C16:0, obtained from a four latent variable PLS model. Even though predictions of C16:0 appear good, the relationship between measured C16:0 and predicted C16:0 (Figure 4b) has striking similarity with the relationship between measured C16:0 and measured %FAT (Figure 5). This could indicate that predictions of C16:0 are modeled as offset and slope corrected %FAT, as sketched in Figure 1. In practice, the offset difference is handled by centering the response (Equation 2) and the slope difference is handled by scaling the length of the regression vector (Equation 5).

Figure 6a shows explained variation from PCA of the measured fatty acids obtained from gas chromatography. This plot reveals that the chemical rank of the fatty acids is high and close to full. Figure 6b shows the explained variation from PCA of the preprocessed NIR spectra, and reveals that approximately five latent variables is sufficient to explain the systematic variation in the NIR spectra. Hence, five orthogonal directions or patterns exist in the NIR spectra. It is thus impossible to obtain independent estimates of all 33 fatty acids. This is clear when calculating a PCA model on the fatty acids predicted from the NIRS measurements (Figure 6c). Here the chemical rank is similar to the rank of the spectra and clearly lower than the chemical rank of the measured fatty acids. Hence, fatty acid estimates obtained from the NIR spectra are dependent on each other. This suggests

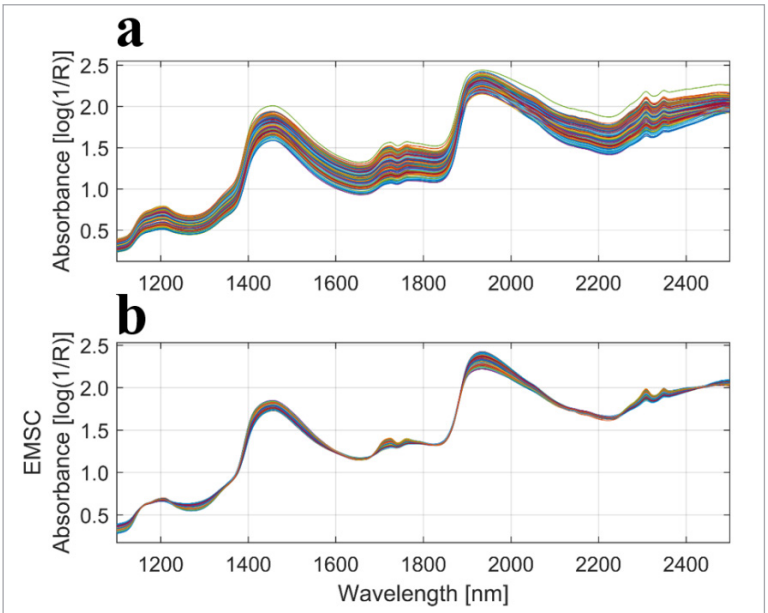

Figure 3. Near infrared spectra of salmon samples; (a) raw near infrared spectra; (b) near infrared spectra preprocessed by extended multiplicative signal correction (EMSC). R = reflectance.

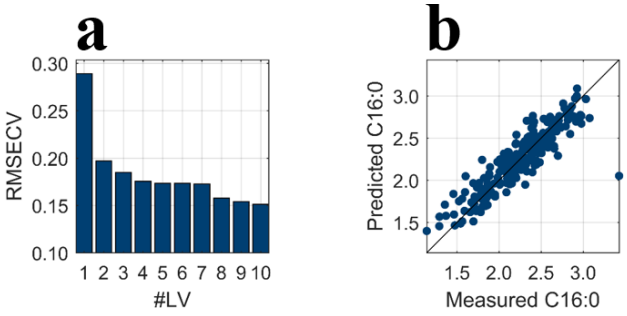

Figure 4. Prediction of C16:0 [g/100 g sample] fatty acid from near infrared measurements; (a) Root mean squared error of cross-validation (RMSECV). #LV = number of latent variables included in the model; (b) Relationship between predicted values (from a model with four latent variables) and measured values of C16:0.

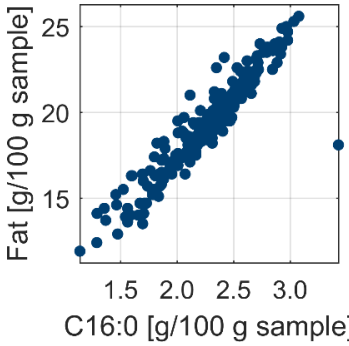

Figure 5. Relationship between measured C16:0 and measured total fat content.

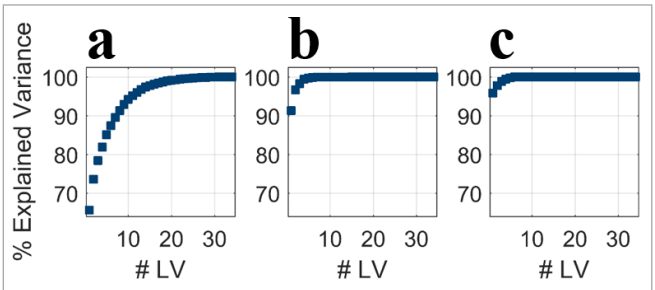

Figure 6. Output from principal component analysis; (a) measured fatty acids; (b) near infrared measurements; (c) predicted fatty acids. # LV = number of latent variables included in principal component analysis.

that the prediction of fatty acids is indirect as sketched in Figure 1.

In Figure 7, we investigate the correlation structure amongst the fatty acids (and %FAT) in a heat map. The elements below the diagonal show the correlation coefficients between measured fatty acids (and %FAT), whereas correlation coefficients between predicted fatty acids (and %FAT) are shown above the diagonal. Increasing correlation coefficients above the diagonal show that the individual fatty acids are modeled by similar linear combinations of NIR wavelengths. Hence, their estimates are not independent. The elements on the diagonal (Figure 7) show the model qualities as calculated by the correla-

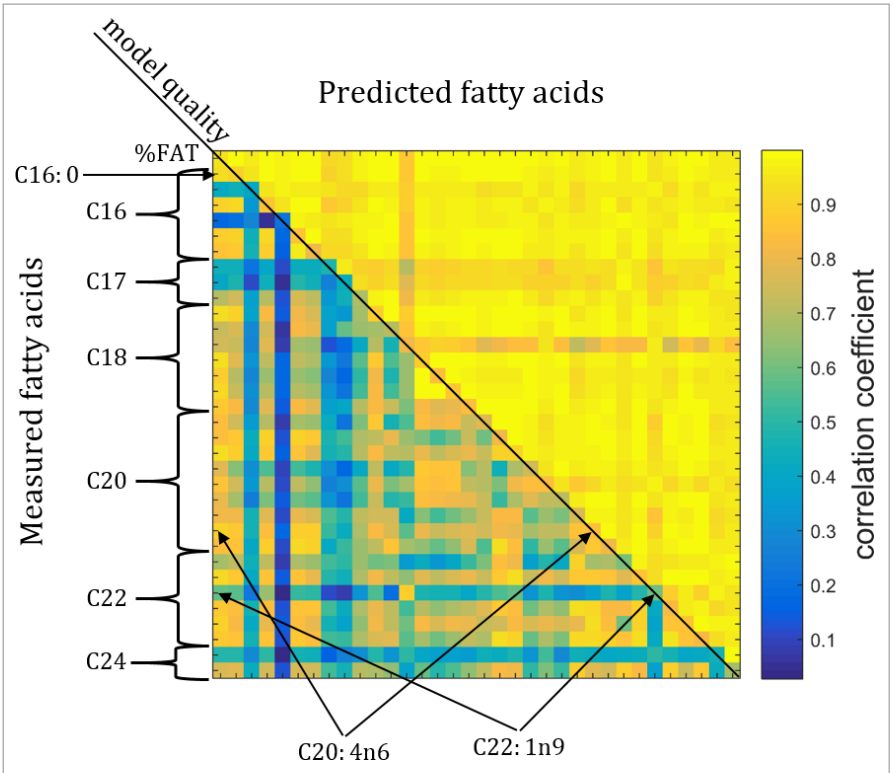

Figure 7. Heat map showing correlation coefficients between fatty acids as well as total fat content (%FAT). Below the diagonal: correlation coefficients between measured values. On the diagonal: correlation coefficients between measured and predicted values (i.e. model quality). Above the diagonal: correlation coefficients between predicted values. Tick labels on x- and y-axis are identical.

tion coefficient between measured and predicted values. Model qualities are to a large extent determined by the correlation between the individual fatty acid and the total fat concentration (%FAT). This is highlighted for C20:4n6 and C22:1n9, which indicates that fatty acids are simply modeled by the overall correlation with %FAT.

Conclusion

In a highly complex sample matrix, it is unlikely that NIRS measurements (vibrational spectroscopy) are able to provide unique signals for all the individual fatty acids. Hence, fatty acids cannot be predicted, in a direct fashion, from NIRS measurements obtained on e.g. minced salmon tissue. Even though the PLS models, predicting the fatty acids, appear good a first glance, the models do not return variation on the individual fatty acid composition, but rather on the variation in %FAT. As the models are mainly relying on correlation structures between individual fatty acids and %FAT, these correlation structures have to remain constant in order not to provide erroneous predictions. Although indirect NIRS calibrations are becoming more widespread, they are problematic in terms of accuracy and robustness of the calibration models[8] since they rely on biological covariance structures which may not remain constant over time and other external factors. There is thus a strong need for diagnosing when NIRS calibrations models rely on indirect correlations and in turn to understand the boundaries for the validity of the covariance structures.

References

1. C.E. Eskildsen, F.v.d. Berg and S.B. Engelsen, "Vibrational spectroscopy in food processing", in *Encyclopedia of Spectroscopy and Spectrometry* (3rd Edn). Elsevier, Oxford, UK, pp. 582–589 (2017).

https://doi.org/10.1016/B978-0-12-409547-2.12156-0

2. H.B. Jensen, N.A. Poulsen, K.K. Andersen, M. Hammershøj, H.D. Poulsen and L.B. Larsen, "Distinct composition of bovine milk from Jersey and Holstein-Friesian cows with good, poor, or nonco-agulation properties as reflected in protein genetic variants and isoforms", *J. Dairy Sci.* **95,** 6905–6917 (2012). https://doi.org/10.3168/jds.2012-5675
3. G.A. Teye, J D. Wood, F.M. Whittington, A. Stewart and P.R. Sheard, "Influence of dietary oils and protein level on pork quality. 2. Effects on properties and processing characteristics of bacon and frankfurter-style sausages", *J. Meat Sci.* **73,** 166–177 (2006). https://doi.org/10.1016/j.meatsci.2005.11.011
4. S. Wold, A. Ruhe, H. Wold and W.J. Dunn III, "The collinearity problem in regression, the partial least squares approach to generalized inverses", *SIAM J. Sci. Stat. Comput.* **5,** 735–743 (1984). https://doi.org/10.1137/0905052
5. P. Geladi and B.R. Kowalski, "Partial least-squares regression: a tutorial", *Anal. Chim. Acta* **185,** 1–17 (1986). https://doi.org/10.1016/0003-2670(86)80028-9
6. C.E. Eskildsen, M.A. Rasmussen, S.B. Engelsen, L.B. Larsen, N.A. Poulsen and 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,** 7940–7951 (2014). https://doi.org/10.3168/jds.2014-8337
7. D.T. Berhe, C.E. Eskildsen, R. Lametsch, M.S. Hviid, F.v.d. Berg and 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,** 18–26 (2016). https://doi.org/10.1016/j.meatsci.2015.08.009
8. C.E. Eskildsen, T. Skov, M.S. Hansen, L.B. Larsen and 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,** 8178–8186 (2016). https://doi.org/10.3168/jds.2015-10840
9. Å. Rinnan, S. Bruun, J. Lindedam, S.R. Decker, G.B. Turner, C. Felby and S.B. Engelsen, "Predicting the ethanol potential of wheat straw using near-infrared spectroscopy and chemometrics: the challenge of inherently interorrelated response functions", *Anal. Chim. Acta* **962,** 15–23 (2017). https://doi.org/10.1016/j.aca.2017.02.001
10. E. Sanchez and B.R. Kowalski, "Tensorial calibration: I. First-order calibration", *J. Chemometr.* **2,** 247–263 (1988). https://doi.org/10.1002/cem.1180020404
11. J. Folch, M. Lees and G.H.S. Stanley, "A simple method for the isolation and purification of total lipids from animal tissues", *J. Biol. Chem.* **226,** 497–509 (1956).
12. M.E. Mason and G.R. Waller, "Dimethoxypropane induced transesterification of fats and oils in preparation of methyl esters for gas chromatographic analysis", *Anal. Chem.* **36,** 583–586 (1954). https://doi.org/10.1021/ac60209a008
13. N.K. Afseth and A. Kohler, "Extended multiplicative signal correction in vibrational spectroscopy, a tutorial", *Chemometr. Intell. Lab. Syst.* **117,** 92–99 (2012). https://doi.org/10.1016/j.chemolab.2012.03.004
14. M. Andersson, "A comparison of nine PLS1 algorithms", *J. Chemometr.* **23,** 518–529 (2009). https://doi.org/10.1002/cem.1248