


Unbiased prediction errors for partial least squares regression models: Choosing a representative error estimator for process monitoring

Journal of Near Infrared Spectroscopy
2023, Vol. 31(4) 186–195
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09670335231173139
journals.sagepub.com/home/jns


Peter B Skou¹, Margherita Tonolini² , Carl Emil Eskildsen^{3,4}, Frans van den Berg² and Morten Arendt Rasmussen²

Abstract

Partial least squares (PLS) regression is widely used to predict chemical analytes from spectroscopic data, thus reducing the need for expensive and time-consuming wet chemical reference analysis in industrial process monitoring. However, predictions via PLS by definition carry sample-specific errors, and estimation of these errors is essential for correct interpretation of results. To increase trust in PLS regression-based predictions, reliable prediction error estimates must be reported. This can be achieved by determining realistic sample-specific prediction errors using an unbiased mean squared prediction error estimate. This work provides a guide for estimating sample-specific prediction errors, showing the importance of choosing an appropriate error estimator prior to deploying PLS models for industrial applications. We reviewed recent and established methods for estimating the sample-specific prediction error and test them through simulation studies. The methods were subsequently applied for estimating prediction errors in two real-life datasets from the food ingredients industry, where near-infrared spectroscopy was used to quantify i) urea in process water and ii) individual protein concentrations in ultrafiltration retentates from a protein fractionation process. Both the simulations and real data examples showed that the mean squared error of calibration is *always* a downward biased estimator. Although leave-one-out-cross-validation performed surprisingly well in the data analysed in this work, this paper demonstrated that the appropriate choice of error estimator requires the user to make an informed, data-centered decision.

Keywords

Partial least squares regression, process monitoring, error estimators, unbiased prediction error, analytical chemistry

Received 22 August 2022; accepted 1 April 2023

Introduction

Process monitoring, control, and optimization of industrial productions are often guided by quantitative predictions of key parameters, which should reflect the true behavior of the production process. The prediction of these parameters is done by applying a predictive model to process measurements. However, if the measurements or the predictive model do not reach a sufficient level of precision, the predicted parameters may not reflect the true behavior of the process, and process improvement could fail. In other words, if the predictions carry large uncertainties, despite on average giving good estimates, it may not be justified to react to the perceived behavior of the process.¹ Adjusting a process based on predictions without knowing the prediction error was referred to as tampering by Deming already in 1982, exemplified by his famous funnel experiment²; this *modus operandi* is unfortunately still all too common today.

Indirect measurement techniques such as nearinfrared (NIR) spectroscopy are usually accompanied by advanced, inverse regression methods such as partial least squares (PLS). Although other regression methods exist, PLS is

especially popular within industrial use of spectroscopic data for process monitoring and has been previously used in applications similar to the ones targeted in the this text.³ In these methods, model complexity and performance are typically determined by resampling error values such as the mean squared error of cross-validation (MSECV). These error values are useful in model selection but should not be interpreted as a surrogate of the expected predictive ability towards a future observation, as this also depends on the

¹Arla Foods Ingredients Group PS, Viby, Denmark

²Department of Food Science, University of Copenhagen Faculty of Science, Frederiksberg, Denmark

³Nofima AS, Norwegian Institute of Food, Fisheries and Aquaculture Research, Muninbakken 9-13, NO-9291, Tromsø, Norway

⁴Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Science Park 904, NL-1098 XH, Amsterdam, the Netherlands

Corresponding author:

Morten Arendt Rasmussen, Department of Food, University of Copenhagen Faculty of Science, Rolighedsvej 26, Frederiksberg 1958, Denmark.

Email: mortenr@food.ku.dk

location of the future sample in the distribution of the calibration data. Furthermore, as the cross-validation error is used for model selection, this error rate is potentially optimistic and biased downwards.

To illustrate the issue at hand, we will show the practical implications of using different error estimates when measuring the beta-lactoglobulin concentration (%w/w) in re-tentates from a protein fractionation process by the use of NIR spectroscopy. Samples were measured with NIR spectroscopy and beta-lactoglobulin was predicted with a PLS model calibrated on samples from different batches and fractionation steps. The beta-lactoglobulin concentration, in this case, is an indicator of the performance and selectivity of membranes over time. Therefore, choosing the appropriate error estimator is crucial for distinguishing real changes in the system from background noise also known as measurement uncertainty. An additional challenge frequently encountered in industrial production processes is datasets with a high degree of sample clustering, since samples coming from the same step of the process will be very similar, not only regarding their analyte content but also their entire chemical profile. This is the case for the protein fractionation process discussed in this work, where samples withdrawn over several days from a specific step of the process have a near-constant characteristic chemical profile. Consequently, the prediction error varies depending on the location of the samples in the distribution of the calibration data, and samples with extreme analyte concentrations have high leverage which influences the sample prediction accuracy.³ The need for sample-specific prediction error estimation is a consequence of the first-order advantage,⁴ namely to observe interfering species (chemical or physical). It cannot be expected that two samples that are equal in the constituent of interest but different in the degree of interference are predicted equally well. The uncertainty will increase for sample clusters if their chemical profile has less support in the calibration model. This is problematic when deploying the PLS model to predict new samples from a specific phase in the process where expected analyte and nuisance interferences are higher compared to the calibration set.

In the following, we briefly present basic cross-validation and bootstrapping and show one modification for each which will subsequently be tested through simulations and on real-world data from two industrial applications. For the industrial examples, the consequence of choosing an appropriate MSE for the calculation of sample-specific prediction error estimates is shown and evaluated.

Theory

The prediction error is defined here as the uncertainty associated with the prediction of a new sample (i.e. the concentration of the analyte in the sample), from measured explanatory variables (i.e. the spectrum for this sample). In this work, bias is used in relation to the difference between the estimated and true mean squared prediction error.

The random error σ^2 is often approximated by the mean squared error of calibration (MSEC; also referred to as the apparent error) for sample-specific prediction error

estimation. While MSEC is appropriate from a theoretical viewpoint when having large amounts of independent samples compared to variables, this apparent error is an optimistic (or naïve) estimate of the *true* σ^2 . Finding an unbiased estimator of the mean squared error of prediction (MSE), i.e. an estimate of σ^2 , has been the topic of many investigations. The random error can be determined in a variety of ways (i.e. leave-one-sample-out, leave-k-fold-out, leave-one-experimental-condition-out, etc.) but is seldom used in lieu of MSEC for sample-specific squared prediction error estimation.

In theory, it is possible to determine a regression model's *true* predictive performance with a test dataset consisting of sufficient/many samples where the reference values have been determined with high accuracy (i.e. practical elimination of the reference error). In reality, this is often too costly or technically impossible (e.g. off-line calibration of in-line NIR probes). Usually, a sample set of modest size is available and from this, a model plus a good estimate of future prediction performance must be obtained. Analog to a large dataset, data could be split into a calibration set and an independent test set (also known as a holdout set). However, if the sample set is small or the experimental domain is wide (e.g. a broad temperature or pH range) the partitioning strategy will affect both the bias and confidence interval, which will both increase with a smaller calibration set.⁵

Consider the matrix \mathbf{X} ($N \times J$; *samples \times independent variables*) and the corresponding reference vector \mathbf{y} ($N \times 1$). PLS relates \mathbf{X} to \mathbf{y} (both assumed column-wise centered around zero) via the regression vector \mathbf{b} ($J \times 1$)

$$\mathbf{y} = \mathbf{X} \cdot \mathbf{b} + \mathbf{e} \quad (1)$$

where \mathbf{e} ($N \times 1$) contains the residuals. The MSEC, or apparent error, can be estimated from the residuals as

$$MSEC = \frac{\sum_{i=1}^N e_i^2}{N - r - 1} = \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{N - r - 1} \quad (2)$$

Where y_i is the reference value of calibration sample i ($i = 1, 2, \dots, N$), N the total number of samples, \hat{y}_i is the predicted value of sample i also present in the calibration dataset and r is the model complexity or the number of components – the most important optimization parameter in PLS regression.

For least squares linear regression, the sample-specific squared prediction error is calculated, as suggested by Næs and Mevik,⁶ using the formula

$$s_{PE_i}^2 = \left(\frac{1}{N} + h_i + 1 \right) \cdot \sigma^2 \quad (3)$$

where N is the number of calibration samples, σ^2 is the random error of the linear regression model and h_i represents the leverage of the sample of interest calculated as

$$h_i = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \quad (4)$$

where \mathbf{x}_i represents the measurement of interest and \mathbf{X} represents the measurements used for calibration.

When data compression methods like PLS regression, where the data is represented by a few components, are used, the sample-specific squared prediction error in line

with equation (3) needs to include a (squared) systematic deviation. This systematic deviation is introduced because of a user-selected model complexity, the famous bias-variance tradeoff. It is a consequence of how the specific sample is located in the space spanned by the omitted (PLS) components. This is what makes the systematic deviation sample-specific as discussed in detail for principal component regression by Eskildsen and Næs.⁷

When dealing with real data, it is in general not possible to estimate the (squared) systematic deviation introduced by omitted components.⁸ Indeed, this can be detrimental when using under-fitted models on new samples falling outside the calibration range. However, when the calibration data span the future domain for prediction, this systematic deviation is likely to be relatively small for the optimal model.^{7,9} A pragmatic approach to obtain sample-specific squared prediction errors in PLS regression is to use equation (3) but substitute σ^2 with MSEC and calculate leverage (Equation (4)) with component scores rather than measurements. This approach, which will also be followed in this paper, builds on the error-in-variable concept for principal component regression and PLS introduced by Faber and Kowalski¹⁰ and later modified by other authors^{3,11}

$$s_{PE_i}^2 = \left(\frac{1}{N} + h_i + 1 \right) \cdot MSEC \quad (5)$$

The focus of this study is to find the most appropriate squared prediction error estimate, i.e. a replacement for the current MSE estimate, MSEC. In the following the *true* (squared) prediction error will be referred to as $Err^{(true)}$, while *estimates* will be denoted by the method used to obtain them (e.g. for cross-validation we will write $Err^{(CV)}$).

Materials and methods

Datasets

Simulations. The data used in the simulations for comparing alternative unbiased prediction variance estimates in PLS regression was generated as follows. The signal matrix, \mathbf{X} ($N \times 100$), was constructed as a bilinear combination of two factor matrices of rank five \mathbf{T} ($N \times 5$), \mathbf{P} (100×5), and the reference, \mathbf{y} , was a linear combination of the columns of \mathbf{T} defined by a weighting vector \mathbf{q} (1×5). \mathbf{T} , \mathbf{P} , and \mathbf{q} are all drawn from a normal distribution centered around zero with variance 1 ($N(0,1)$)

$$\mathbf{X} = \mathbf{TP}^T \quad (6)$$

The true (unknown) reference values, \mathbf{y} , were constructed as

$$\mathbf{y} = \mathbf{Tq}^T \quad (7)$$

While the observed reference values, $\tilde{\mathbf{y}}$, were obtained by adding white noise

$$\tilde{\mathbf{y}} = \mathbf{y} + \mathbf{e}_{\Delta y}, \mathbf{e}_{\Delta y} \sim N(0, 1) \quad (8)$$

Similarly, the observed signals, $\tilde{\mathbf{X}}$, were obtained by adding white noise

$$\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{E}_{\Delta X}, \mathbf{E}_{\Delta X} \sim N(0, 1) \quad (9)$$

Subsequently, calibration and test data were denoted as $\tilde{\mathbf{X}}_{cal}$ plus $\tilde{\mathbf{y}}_{cal}$ and $\tilde{\mathbf{X}}_{test}$ plus $\tilde{\mathbf{y}}_{test}$, respectively.

The effect of the number of calibration samples ($N = 32, 64, 128, 512$) and the model complexity was investigated. For all calibration dataset sizes, the complexity was set to the known rank $r_{opt} = 5$ of the system and for $N = 128$ two additional runs were made with too few components or latent variables ($r = 3$, deliberately under-fitting the data), as well as too many components ($r = 9$, over-fitting). The split into calibration and validation sets was done randomly.

The *estimated* MSE was compared to the *true* MSE to evaluate the bias. The true MSE was defined as follows ($N_{test} = 500$), as was done in Kohavi⁵ (1995). First, a PLS model is built on the observed calibration data

$$\tilde{\mathbf{y}}_{cal} = \tilde{\mathbf{X}}_{cal} \hat{\mathbf{b}} \quad (10)$$

which was subsequently used to predict the observed test set

$$\hat{\mathbf{y}}_{test} = \tilde{\mathbf{X}}_{test} \hat{\mathbf{b}} \quad (11)$$

The *true* MSE was defined as the mean squared difference between the predicted test samples and the noiseless reference values

$$Err^{(true)} = \frac{\sum_{i=1}^{N_{test}} \left(\hat{\mathbf{y}}_{test i} - \mathbf{y}_{test i} \right)^2}{N_{test}} \quad (12)$$

The different MSE estimators were calculated as the mean squared difference between the predictions obtained from the given estimation method, and the observed reference values. Since the predictions ($\hat{\mathbf{y}}_{test}$) carry the reference uncertainty ($e_{\Delta y}^2$) and the true values (\mathbf{y}_{test}) do not, all MSE estimates were reduced by this reference uncertainty

$$Err^{(\dots)} = \frac{\sum_{i=1}^{N_{cal}} \left(\hat{\mathbf{y}}_{cal i} - \tilde{\mathbf{y}}_{cal i} \right)^2}{N_{cal} - r - 1} - e_{\Delta y}^2 \quad (13)$$

This is done to ensure a fair comparison between the true and the estimated MSE. The bias of each MSE estimate was evaluated as the difference between the estimated MSE (from the calibration stage) and the *true* MSE (based on the test samples). This was done for each draw

$$Bias = Err^{(true)} - Err^{(\dots)} \quad (14)$$

For each combination of model selection strategy and size of the calibration set, 100 simulations were performed, with independent datasets generated for each iteration. The reference uncertainty ($e_{\Delta y}^2$) was equal to zero in the simulations (while in real-life data it would normally be estimated from e.g. duplicate reference determinations, ring-test or Gage repeatability and reproducibility studies)¹⁶ 500.

Urea data. In total 68 process-water samples, split into 32 calibration samples and 36 test samples, each with a NIR absorption spectrum measured in transmission mode and urea reference value are available (ppm levels). Samples were collected from a whey protein ingredients production plant over time and the calibration/test split is made not

random but at a specific calendar time to emulate the end-stage of model building followed by validation. Each sample was measured with an ABB Bomem MB series FT-NIR spectrometer (Quebec, QC, Canada) with a custom-made, temperature-controlled sample flow cell from 14 285 – 4000 cm^{-1} (700–2500 nm) with a spectral resolution of 8 cm^{-1} and 128 scans averaged. Further spectroscopic and modeling details can be found at Skou et al.³ (2017). For this dataset, we fix calibration- and test samples and use the MSE as the *true* MSE. This NIR data presents less data clustering and a simpler chemical background (water) than whey protein retentate, making it complementary to the Ultrafiltration retentate dataset.

Ultrafiltration retentate data. Seventy-two whey protein concentrate samples were collected from different production lines and at different steps of protein fractionation processes at Arla Food Ingredients (Nr. Vium, Denmark). A primary set of retentate samples was collected from randomly selected extraction points in the protein fractionation process line of interest, while a secondary sampling campaign was compiled by extraction of retentate samples from a specific step of the process. Eighteen whey concentrate samples were spiked with whey protein powders to decrease the covariance between whey proteins concentrations present in the original matrix and to increase the concentration span in the sample set with a similar spiking methodology as presented in Tonolini et al.¹⁷ NIR spectra were collected using the same spectrometer described under *Urea data*. Spectra were acquired in transmittance mode with a 1 mm path length cuvette (HELLMA Macro-Cuvette 100-QS 1 mm Quartz Glass 100-1-40, Hellma Materials GmbH, Jena, Germany). Spectra were measured in the range 1000–2500 nm, a total of 64 scans were recorded and averaged for each sample and the spectral resolution was 8 cm^{-1} . Beta-lactoglobulin concentration (%w/w) was measured with RP-HPLC using in-house routines for whey protein quantification at Arla Food Ingredients and was used as the reference variable. NIR spectral measurements were used as independent variables. Variable selection and preprocessing were reproduced from Tonolini et al.¹⁸

A dataset containing both process and spiked samples ($N = 68$) was used for calibration and process samples from a specific step in the process over different days were used as validation ($N = 32$). The optimum number of components was found for the PLS model based on 5-fold cross-validation repeated 100 times. The prediction set's protein concentration has a limited concentration span (4–8 % w/w) compared to the calibration set (0–8% w/w). This prediction set was selected to illustrate the *practical problems* encountered when deploying a regression model for process monitoring, where clusters of data with very similar chemical profiles (due to the highly standardized processes found in the industry) result in insufficient support for some samples in the validation set.

Prediction error estimation strategies

***K*-fold cross-validation.** The basic idea in cross-validation is to leave out a part of the sample set, build a model on the

remaining sample set, and predict the left-out samples. Most often data is divided by K -fold non-overlapping splits where K is varied from 2, thus halving the sample set, up to N , resulting in the leave-one-out cross-validation (LOOCV) scheme.

Formally, for each split k of size N_k , we remove the k^{th} part of \mathbf{X} , fit a model $\hat{\mathbf{b}}^{(-k)}$, and predict the left-out k^{th} part, $\hat{\mathbf{y}}^{(k)}$. If D_k contains the indices of the measurements in the k^{th} split the sum of the squared prediction error for each split becomes

$$Err_k^{(CV)}(r) = \frac{1}{N_k} \sum_{i \in D_k} (y_i^k - \hat{y}_i^k(r))^2 \quad (15)$$

Averaging over all K splits results in the K -fold cross-validation error rate

$$Err^{(CV)}(r) = \frac{1}{K} \sum_{k=1}^K Err_k^{(CV)}(r) \quad (16)$$

Estimating model complexity is often done by scanning a range of r (i.e. testing PLS with 1 to r components in the model) to find that r_{opt} which minimizes $Err^{(CV)}$.

CV corrections. Tibshirani and Tibshirani (2009)¹² report that this minimum error rate found at r_{opt} often is too optimistic to function as an estimate for the true MSE. In other words, the minimum has a downward bias, due to its dual role in model selection, and must be corrected via the bias-corrected cross-validation (bcCV)

$$Err^{(bcCV)} = Err^{(CV)}(r_{opt}) + Bias \quad (17)$$

To estimate this bias, we can consider the mean squared error of each k^{th} part of the data, containing N_k samples, as shown in equation (15). Assuming that all reasonable values of r have been evaluated for all K splits of the data, we can easily find the r -value that minimizes the mean squared error for each k^{th} part separately, denoted here as r_k . The concept of the correction is that the bias between the (global) minimum MSEC_V – based on all K folds – and the local k^{th} part minimum mean squared error will mimic the bias between the true MSE and the minimum MSEC_V. We can thus obtain the bias for the MSEC_V as follows

$$\widehat{Bias} = \frac{1}{K} \sum_{k=1}^K (Err_k^{(CV)}(r_{opt}) - Err_k^{(CV)}(r_k)) \quad (18)$$

The estimated bias for the model is the average difference in MSE obtained at the global minimum (r_{opt}) and at the local minima (r_k) for all K splits. Note that the \widehat{Bias} is expressed in the same unit as the variances, following the theory of Tibshirani & Tibshirani (2009)¹² and that this bias contribution for all resampling trials could (theoretically) be zero

$$\widehat{Err}^{(bcCV)} = Err^{(CV)}(r_{opt}) + \widehat{Bias} \quad (19)$$

In this paper, we use the bcCV method to obtain unbiased MSE estimates, rather than decide the complexity of the model and estimate the expected error. This means that the

bias is not estimated as the average difference between the minimum MSE and MSE_k , but the MSE at the chosen number of components and minimum MSE_k .

Bootstrapping. The bootstrap methodology is well described in several publications including the seminal work *An Introduction to the Bootstrap*.^{12–15} Just like cross-validation, bootstrapping is based on a resampling strategy. When performing bootstrap estimates, the uncertainty of a parameter is mimicked by creating p new bootstrap sample sets with N samples each, drawn from the original dataset, with replacement. For each bootstrap sample a model is built and a prediction error estimate, e.g. MSEC, is calculated thus providing p estimates of the prediction error estimate. The uncertainty can be assessed from this information (by e.g. taking the mean of the p MSEC estimates). This procedure is also known as the naïve bootstrap due to the simple nature of the procedure, only relying on the (often unrealistic) assumption that the samples are independent and equally relevant.

A variation on the naïve bootstrap is the leave-one-out bootstrap^{14,15} or bootstrap smoothed cross-validation¹⁶ estimate. Assume we draw p bootstrap samples from a sample size of N ; for each build a model on that draw and predict the – on average $(1-1/N) \wedge N = 0.368\%$ – samples not included in the draw. Let e_i contain the residual of the N_p (size of the bootstrapping data split) predictions for sample i . Note that N_p may be different across the original (full) sample set, but will by definition be between zero and p . Taking the mean squared error per sample and averaging this across all samples gives the leave-one-out bootstrap estimate, $Err^{(LOBS)}$

$$Err^{(LOBS)} = \frac{1}{N} \sum_{i=1}^N \frac{1}{N_p} \sum_{p=1}^{N_p} e_{i,p}^2 \quad (20)$$

0.632 bootstrapping. Efron (1983)¹⁴ argues that while the apparent error – or in this case the MSEC – underestimates the true MSE (since the predicted samples are also part of the calibration set), the leave-one-out bootstrap, in turn, overestimates the true MSE because a given sample has a probability of 63.2% of ending in the training set. To correct for this (upward) bias he proposes to compute a weighted average of the apparent and bootstrapped mean squared error

$$Err^{(0.632)} = 0.368 \cdot MSEC + 0.632 \cdot Err^{(LOBS)} \quad (21)$$

0.632+ bootstrapping. Efron & Tibshirani (1997)¹² remark that the 0.632 bootstrap estimates will be downward biased when using severely over-fitted models such as in nearest-neighbor classification. To counter this, they propose the 0.632+ bootstrap estimator, which implements a weighting regime to the estimator so that it can compensate if over-fitting occurs. Over-fitting is evaluated here as the difference between the apparent error and the leave-one-out bootstrap mean squared error estimate. To adjust the over-fit model error correctly, the so-called *no-information* (squared) error must also be determined.

The no-information (squared) error, γ , is the expected (squared) error for the model based on the original dataset

given that there is no relation between X and y . This can be determined by predicting all samples with a global model and calculating the (squared) error for all (correct and incorrect) combinations of the reference and the prediction, normalized by the number of entries, subtracting the combinations that are truthfully related (hence, where $i = j$) by subtracting MSEC

$$\gamma = \frac{\sum_{j=1}^N \sum_{i=1}^N (y_i - \hat{y}_j)^2}{N^2} - MSEC \quad (22)$$

A relative over-fit, F , is now defined as

$$F = \frac{Err^{(LOBS)} - MSEC}{\gamma} \quad (23)$$

which in turn is used to define a weight, ω :

$$\omega = \frac{0.632}{1 - 0.368 \cdot F} \quad (24)$$

The 0.632+ estimator is then defined as

$$Err^{(0.632+)} = (1 - \omega) \cdot MSEC + \omega \cdot Err^{(LOBS)} \quad (25)$$

If the relative over-fit rate is zero, the 0.632+ estimator coincides with the 0.632 estimator.

Independent test set. Finally, the predictive performance of an independent test set is quantified. If instead of resampling the calibration data a truly independent set of samples is set aside and used for testing different models the squared prediction error will be reported as the mean squared error of prediction (MSEP), here from a test set of I samples

$$MSEP = \frac{\sum_{i=1}^I (\hat{y}_i - y_i)^2}{I} \quad (26)$$

Results

Simulations

The simulation results were evaluated as the difference between the true and estimated MSE, where the true MSE was based on 500 test samples. We presume that if the difference is zero, the MSE estimator is unbiased. If the difference is positive it underestimates the MSE and is therefore optimistic, while if the difference is negative it overestimates and gives a pessimistic estimate of the MSE.

The simulation study representing the simplest situation – namely where the complexity of the model is known – is shown in Figure 1. Naturally, the spread around the difference between true and estimated MSE decreased as the number of calibration samples increased. This is true for all MSE estimators. The MSEC was optimistic across all calibration set sizes and the leave-one-out bootstrap slightly pessimistic. Unexpectedly, the 0.632 and 0.632+ bootstrap procedures were optimistic. Regarding the K -fold cross-validation strategies, 2-fold CV (also known as split-half) was too pessimistic, but this reduced with increasing K , eventually leading to unbiased estimates at 8- and 16-fold CV. The bias correction applied to the K -fold CV turned out to increase in size when increasing the number of splits and generally makes the estimates pessimistic. LOOCV

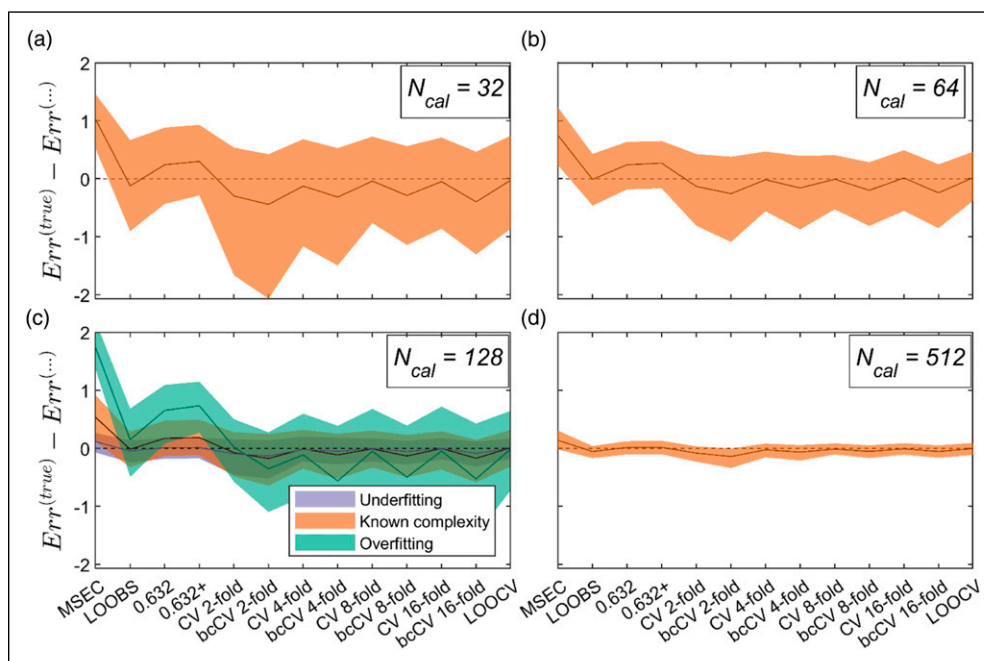


Figure 1. Simulation - (a) 32, (b) 64, (c) 128, and (d) 512 calibration samples used to simulate MSE estimates and comparing with 500 test samples when the complexity ($r_{opt} = 5$) of the system is known. For the simulation with (c) 128 calibration samples also model over-fitting ($r = 9$) and under-fitting ($r = 3$) is shown. The shaded area covers 90% of the simulations and the solid line represents the mean. The y-axis is relative to the reference uncertainty.

turned out to provide on average an unbiased estimate of the MSE, as was reported previously in the literature.^{5,15}

In Figure 1(c), the PLS model complexity was intentionally set too low ($r = 3$) and too high ($r = 9$) compared to the true rank of the system ($r_{opt} = 5$). This was done to investigate the effect of under- and over-fitting. Note that also $Err^{(true)}$ is calculated from an under-fitted and over-fitted model when investigating the effects of too low and too high complexity, respectively. When over-fitting the models, the pattern is similar to when models are fitted with correct complexity, but this is more pronounced in the case of over-fitting. MSEC is too optimistic, and the leave-one-out bootstrap was also slightly optimistic while the K -fold CV estimates were close to unbiased. The correction methods for the 0.632 and 0.632+ bootstrap reduce the pessimism and resulted in far too optimistic estimates. The bias corrections to the K -fold CV methods were too large leading to slightly pessimistic results. It is surprising to see LOOCV gave unbiased MSE estimates when the theoretically more pessimistic leave-one-out bootstrap turns out optimistic for severely over-fitted models. In the case of under-fitting the models, all $Err^{(...)}$ were unbiased and the spread around zero (no bias) is relatively small (Figure 1(c)). It is important to be aware that the under-fitted models in general perform better than models fitted with the correct complexity. Hence, $\overline{Err}_{r=3}^{(true)} < \overline{Err}_{r=5}^{(true)} < \overline{Err}_{r=9}^{(true)}$, indicating that PLS overshoots with components 4 and 5 for this specific simulated dataset. However, Figure 1(c) solely shows that the spread in differences between the true and estimated MSE is smaller for under-fitted models.

Urea data

The alternative prediction error estimation strategies suggested in the theory section were applied to the urea dataset. Two

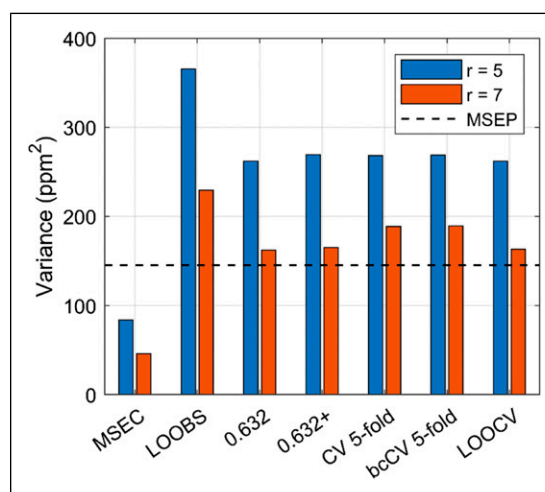


Figure 2. MSE (Variance) estimates for the urea calibration with the reported complexity ($r = 5$) and the complexity decided by 5-fold cross-validation ($r = 7$).

model complexities are evaluated, namely five PLS components as used in the original investigation,³ and seven components as suggested by computing the median of the lowest error rate of 100 iterations of a random 5-fold cross-validation on the calibration data. Interestingly, the MSEP obtained from the fixed validation set were almost identical (<1 ppm apart) for the two model complexities tested, indicating the same prediction ability. The comparison of estimated MSEs is shown in Figure 2. Usually, the parsimony principle will compel the user to choose the simpler model. However, by choosing the simpler model in this case the MSEC underestimates the MSE (too optimistic), while all the alternatives overestimate it by a considerable margin. Instead, if the more complex model

is chosen, the MSEC underestimates even more – which is in line with the simulation study – but now the alternative methods 0.632, 0.632+, LOOCV, 5-fold CV, and its bias corrected cousin give acceptable results.

The results of applying the sample-specific prediction intervals¹ for the predicted urea values, with the standard error computed via equation (5) for the seven component PLS model with either MSEC or LOOCV as MSE estimator, are shown in Figure 3 (plotted as a function of production time). Using the optimistic MSEC for the sample-specific 95% confidence intervals does not provide the needed coverage (Figure 3(a)) while utilizing LOOCV provides sufficient coverage (Figure 3(b)).

Ultrafiltration retentate data

The importance of applying the *correct or relevant* mean squared prediction error is highlighted by the application of the previously described methods on a dataset made of industrial samples. A PLS model for prediction of beta-lactoglobulin was calibrated on samples from different batches and fractionation steps (Figure 4(a)). Beta-lactoglobulin concentrations were predicted in samples drawn from a specific loop of a protein fractionation process, collected over several days. The prediction set's protein concentration has a limited span (4–8 %w/w) compared to the calibration set and a characteristic

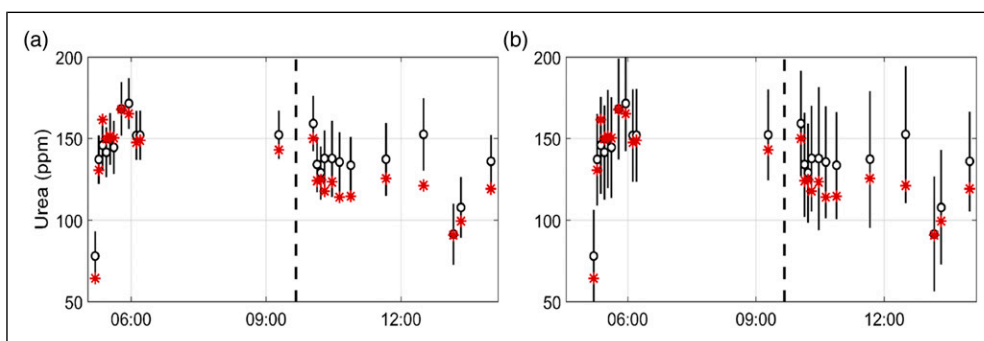


Figure 3. The difference between using (a) the mean squared calibration error (MSEC) or (b) the leave-one-out cross-validation mean squared error estimate for measurement-specific prediction error estimates for test samples with a 7 component PLS model for predicting urea concentration in process water in ppm. Predictions, open circles; reference values stars. Vertical lines show 95% sample-specific prediction intervals and the dashed vertical line indicates the calibration and validation data split.

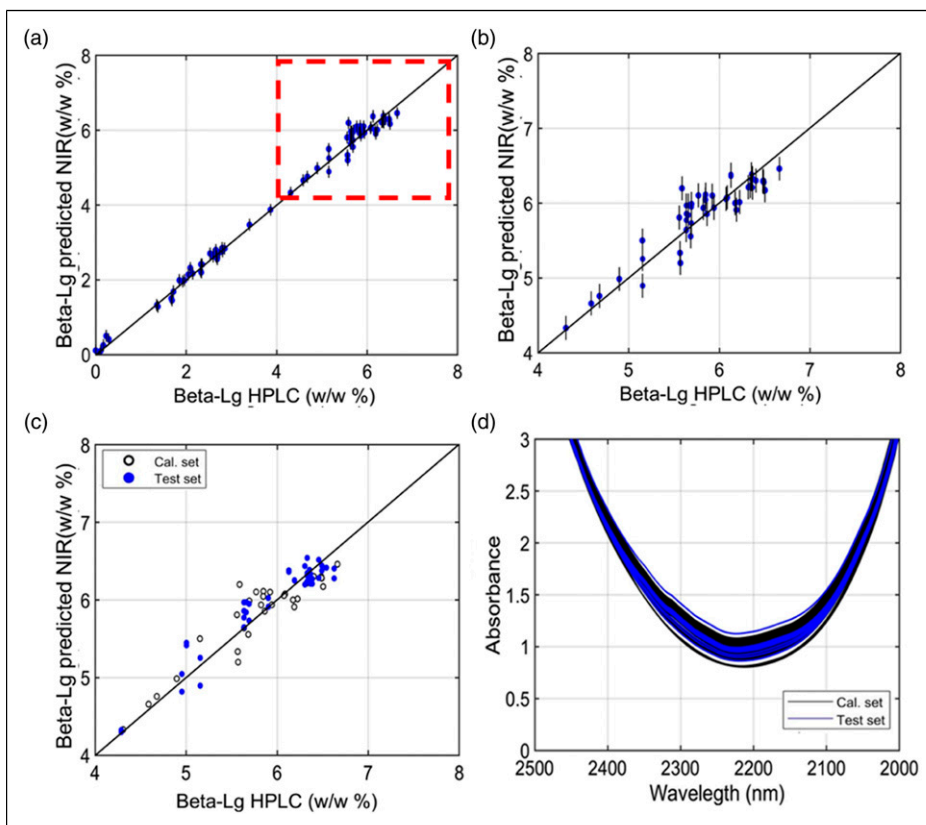


Figure 4. Prediction error intervals for a PLS model predicting Beta-lactoglobulin from NIR spectra using MSEC (0.03 %w/w) as variance estimator (a) entire calibration and (b) zoomed on a selected interval selected interval, the red dotted square highlights the concentration range of interest for the prediction set. (c) Actual versus predicted beta-lactoglobulin content for calibration and test data and (d) the raw NIR spectral range colored according to calibration and validation set.

chemical profile, as a result the model accuracy and precision are worse for high protein concentration samples (Figures 4(b) and (c)).

Here, MSE was estimated using the previously determined complexity, namely four PLS components, but also by computing the median of the lowest error rate of 100 iterations of random 5-fold cross-validation on the calibration data, which turned out to suggest six PLS components. The MSEP obtained from the fixed validation set was lower for the more complex model indicating a better prediction ability (Figure 5). Similar to the *Urea* data, if the higher model is selected, the MSEC underestimates the prediction error even more. Using the optimistic MSEC for the sample-specific 95% confidence intervals does not provide the needed coverage for either of the model complexities, as shown

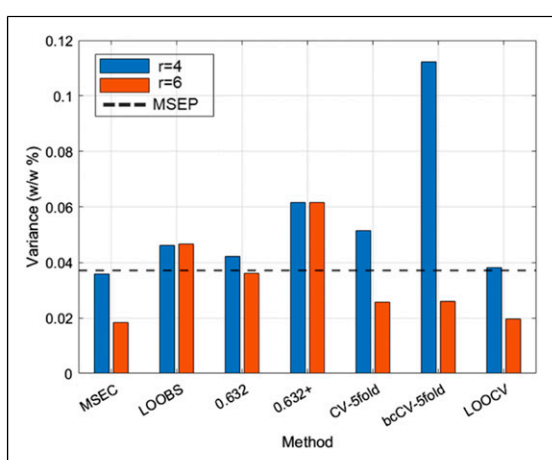


Figure 5. MSE (Variance) estimates for the ultrafiltration retentate calibration with the simpler complexity ($r = 4$) and the complexity decided by 5-fold cross-validation ($r = 6$).

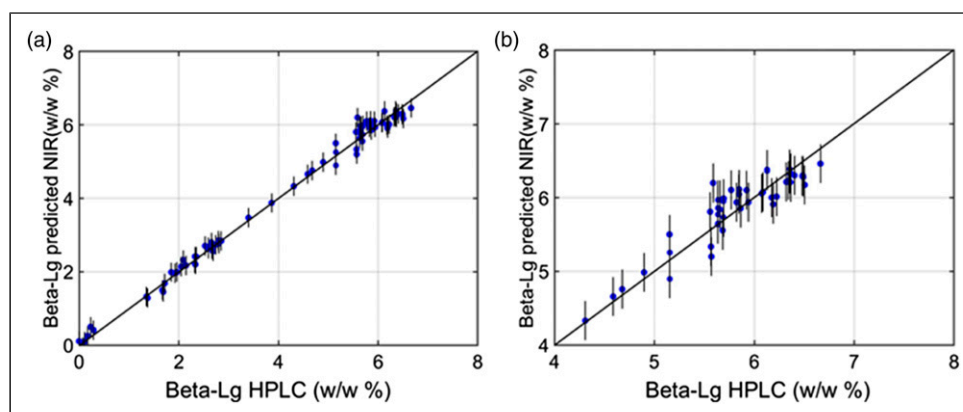


Figure 6. Prediction error intervals for a PLS model predicting Beta-lactoglobulin from NIR spectra using LOOBS variance estimator (0.07 w/w) (a) entire calibration and (b) selected interval.

Table 1. MSE (Variance) estimates for the both urea and ultrafiltration data expressed as error % of respective calibration ranges.

Dataset	#LVs	MSEC	LOOBS	0.632	0.632+	LOOCV	CV-5fold	bcCV-5fold
Urea	5	5.45	11.25	9.53	9.66	9.83	9.92	9.63
	7	4.03	9.36	7.83	7.92	7.78	7.80	7.60
Ultrafiltration	4	2.36	2.68	2.57	3.10	2.44	2.84	4.19
	6	1.69	2.70	2.38	3.10	1.75	2.01	2.02

in Figures 5 and 7. The alternative methods leave-one-out bootstrap and 0.632 bootstrap give acceptable results (Figure 5). Using LOOBS as prediction error provides better coverage than the MSEC (Figures 4(a), (b) and Figure 6) and the selected prediction error allows for 96.0% (a) and 94.5% (b) of the prediction confidence intervals to intersect with the reference-versus-predicted diagonal. The same prediction estimate (LOOBS) was selected as a prediction error estimator for interpreting a process control chart where predictions of beta-lactoglobulin concentrations over 9 days are used to monitor the state of the process (Figure 7(b)). The figure shows that days 8 plus 9 (correctly) and day 6 (incorrectly) are convincingly outside of the specification limits for this process sample point, justifying a control action. Day 2 is borderline off-spec, while process adjustments on days 1, 3, 4, 5, and 7 would classify as *tampering* under Deming's definition (Figure 6).

Table 1 summarizes the results obtained for both dataset, showing the results in error percentages instead of the respective units.

Discussion

The simulations show that the mean squared error of calibration is *always* a downward biased estimator. The improved 0.632+ bootstrap procedure turns out not to make any meaningful difference compared to the original 0.632 for our (simulated) regression task. They differ in that the former adds a correction for over-fitting. In our simulation study with over-fitting, the improved procedure does however not correct for the optimistic MSE estimate. 0.632+ was developed as a method to handle severely over-fitting models such as k-nearest neighbors' type

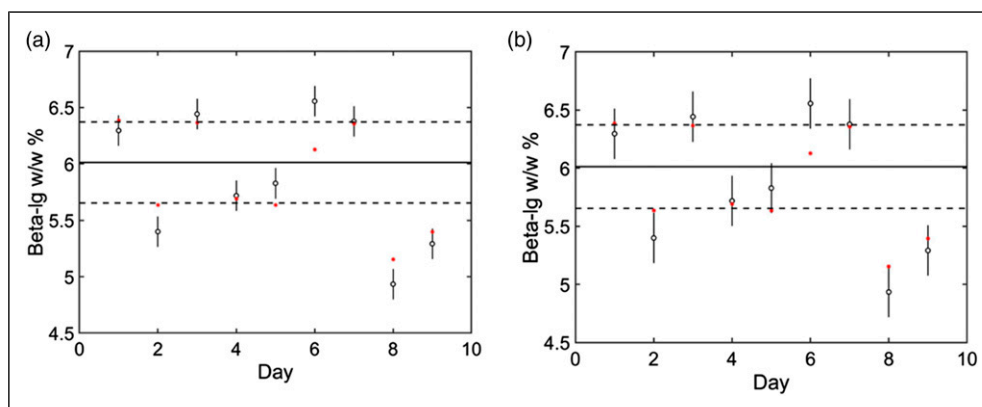


Figure 7. The difference between (a) the mean squared calibration error (MSEC) or (b) leave-one-out bootstrap mean squared error estimate for sample-specific prediction error estimates for test samples from a retentate process stream over 9 days. Predictions, open circles; reference values, stars; error bars, sample-specific 95% confidence intervals; solid line, mean concentration found at this specific point in the process; dashed lines, process control limits.

classification models¹⁵ and evidently, the over-fit in the problems presented here is not severe enough. It is worth recalling that LOOCV turned out less sensitive towards over- and under-fitting of the data in the simulation study. This fact together with the observations made above make LOOCV a very attractive alternative to MSEC in obtaining unbiased prediction uncertainties for PLS regression models, as was the case for the *Urea data* presented.

The ultrafiltration retentate dataset is far less homogenous than the presented simulation data and the urea data. It was made from *clusters* of samples coming from different steps in a production line, which can be viewed as different segments or groups. These groups have similar chemical characteristics that can however differ from one grouping to another, therefore, some groups of data (in our case the higher protein concentration samples) are less well predicted than others. Such sample-group categories (in our case process steps) can be identified in PLS models by inspecting the individual sets of scores in X and y (typically done via T vs U score plots over different components/dimensions, the so-called inner-relation). Some latent variables will (primarily) model which grouping a given sample belongs to instead of improving y -predictions. This in turn means that an *unlucky* resampling draw can underrepresent or even overlook a group entirely. It seems that optimistic MSE estimates are common in this design-scenario, probably due to the small calibration set size and the experimental design implying poor calibration data support for some of the data clusters. In principle, a stratified resampling could (partly) eliminate this *grouping issue*. In an industrial setting though many (known and unknown) potential grouping-causes might be present in the data (feeding material clusters, alternative measurement points, production days/regimes, etc.). Based on the results in Figure 7, the resampling used in the bootstrap procedures seems to be appropriate to partly compensate for this imbalance. In the case of underrepresentation, we speculate that the bias-corrected k -fold CV will also be able to compensate for this problem simply due to the internal resampling. Both dataset showed that the MSEC underestimates the prediction error (Table 1) and other estimators are more similar to the MSEP and thus more appropriate for reporting the model results.

Conclusions

When constructing reliable, sample-specific prediction intervals from PLS models MSEC, or the model fit, is too optimistic a statistical metric. Several improvements are available from the literature and a selection has been tested here using simulations complemented by spectroscopic datasets from an industrial process. In these cases, different methods of calculating the prediction error can lead to variation in sample-specific prediction intervals and eventually influence how the process monitoring itself is interpreted. Simulations and spectroscopic examples show that there is no *one-size-fits-all* solution and some meta-aspects of the data should be taken into consideration.

An informed decision is essential for the successful implementation of an indirect method (such as NIR) for routine quality control analysis in the industry. Overall, if there is considerable grouping (clustering) in the calibration sample set that cannot be resolved by pre-processing, bias-corrected k -fold cross-validation and bootstrapping methods are favorable. If clustering is not present, simple leave-one-out-cross-validation performs surprisingly well. To conclude, this work presents and explains the main prediction error calculation strategies present in literature, while proving the need to integrate a data-centered decision for the calculation of a prediction error whenever a PLS model is developed and deployed for process monitoring.

Acknowledgements

We thank DRIP – *Danish partnership for Resource and water efficient Industrial food Production* as well as the Danish Dairy Research Foundation via the Auroma project for funding.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this

article: This work was supported by the Arla Food Ingredients (Videbæk, Denmark) and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199.

ORCID iD

Margherita Tonolini  <https://orcid.org/0000-0002-3688-1335>

Note

1. Assuming a Student t-distribution with $N-r-1$ degrees of freedom.

References

1. Smilde AK, van den Berg FWJ and Hoefsloot HCJ. How to choose the right process analyzer. *Anal Chem* 2002; 74: 369A–373A.
2. Deming WE. *Out of the crisis*. 1st ed. Massachusetts Institute of Technology, 1982.
3. Skou PB, Berg TA, Aunbjerg SD, et al. Monitoring process-water quality using near infrared spectroscopy and partial least squares regression with prediction uncertainty estimation. *Appl Spectrosc* 2017; 71: 410–421. DOI: [10.1177/0003702816654165](https://doi.org/10.1177/0003702816654165)
4. Booksh KS and Kowalski BR. Theory of analytical chemistry. *Anal Chem* 1994; 66: 782–791. DOI: [10.1021/ac00087a001](https://doi.org/10.1021/ac00087a001)
5. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Appear Int Jt Conf Artificial Intell* 1995; 5: 1–7. DOI: [10.1067/mod.2000.109031](https://doi.org/10.1067/mod.2000.109031)
6. Næs T and Mevik BH. Understanding the collinearity problem in regression and discriminant analysis. *J Chemom* 2001; 15: 413–426. DOI: [10.1002/cem.676](https://doi.org/10.1002/cem.676)
7. Eskildsen CE and Næs T. Sample-specific prediction error measures in spectroscopy. *Appl Spectrosc* 2020; 74: 791–798. DOI: [10.1177/0003702820913562](https://doi.org/10.1177/0003702820913562)
8. Zhang Y and Fearn T. A linearization method for partial least squares regression prediction uncertainty. *Chemometr Intell Lab Syst* 2015; 140: 133–140. DOI: [10.1016/j.chemolab.2014.11.011](https://doi.org/10.1016/j.chemolab.2014.11.011)
9. Faber NM and Rajkó R. How to avoid over-fitting in multivariate calibration—the conventional validation approach and an alternative. *Anal Chim Acta* 2007; 595: 98–106. DOI: [10.1016/j.aca.2007.05.030](https://doi.org/10.1016/j.aca.2007.05.030)
10. Faber K and Kowalski BR. Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. *J Chemom* 1997; 11: 181–238. DOI: [10.1002/\(SICI\)1099-128X\(199705\)11:3<181::AID-CEM459>3.0.CO;2-7](https://doi.org/10.1002/(SICI)1099-128X(199705)11:3<181::AID-CEM459>3.0.CO;2-7)
11. Allegrini F, Wentzell PD and Olivieri AC. Generalized error-dependent prediction uncertainty in multivariate calibration. *Anal Chim Acta* 2016; 903: 51–60. DOI: [10.1016/j.aca.2015.11.028](https://doi.org/10.1016/j.aca.2015.11.028)
12. White GH. Basics of estimating measurement uncertainty. *Clin Biochem Rev* 2008; 29: S53–S60.
13. Tonolini M, Skou PB and van den Berg FWJ. *UV spectroscopy as a quantitative monitoring tool in a dairy side-stream fractionation process*. Chemometrics and Intelligent Laboratory Systems, 2022, p. 104561. DOI: [10.1016/j.chemolab.2022.104561](https://doi.org/10.1016/j.chemolab.2022.104561)
14. Tonolini M, Sørensen KM, Skou PB, et al. Prediction of α -Lactalbumin and β -Lactoglobulin composition of aqueous whey solutions using Fourier transform mid-infrared spectroscopy and near-infrared spectroscopy. *Appl Spectrosc* 2021; 75(6): 718–727. juin 2021. DOI: [10.1177/0003702820979747](https://doi.org/10.1177/0003702820979747)
15. Tibshirani RJ and Tibshirani R. A bias correction for the minimum error rate in cross-validation. *Ann Appl Stat* 2009; 3: 822–829. DOI: [10.1214/08-AOAS224](https://doi.org/10.1214/08-AOAS224)
16. Efron B and Tibshirani RJ. *An introduction to the bootstrap*. 1st ed. 1993.
17. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc* 1983; 78: 316–331. DOI: [10.1080/01621459.1983.10477973](https://doi.org/10.1080/01621459.1983.10477973)
18. Efron B and Tibshirani R. Improvements on cross-validation: The .632+ bootstrap method. *J Am Stat Assoc* 1997; 92: 548–560. DOI: [10.1080/01621459.1997.10474007](https://doi.org/10.1080/01621459.1997.10474007)