Chemometrics and Intelligent Laboratory Systems 188 (2019) 79–86

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Number of components and prediction error in partial least squares regression determined by Monte Carlo resampling strategies

Olav M. Kvalheim [a,*], Bjørn Grung [a], Tarja Rajalahti [a,b]

[a] Department of Chemistry, University of Bergen, Bergen, Norway
[b] Research and Innovation, Førde Health Trust, Førde, Norway

ARTICLE INFO

Keywords:
Model selection
Component selection
Model validation
Prediction error
Monte Carlo resampling

ABSTRACT

Using a metabolomics data set with 1057 serum samples, we designed and assessed different procedures based on Monte Carlo resampling schemes to determine the optimal number of components to be included in partial least squares (PLS) regression models. Corresponding estimates of prediction error were calculated and compared in a single algorithm comprising i) a single loop Monte Carlo approach repeatedly and randomly splitting samples into calibration and validation samples, ii) a double loop validation splitting samples into calibration/validation and prediction sets, and, iii) independent sample sets in a third loop. In order to mimic the common situation with only a moderate number of samples available for building the model, only a fraction of the 1057 samples analyzed was randomly selected from the total sample set and used in the algorithm. The results show that if the samples available for modelling are representative for the future samples to be predicted from the model, the single loop Monte Carlo procedure consistently provides the same estimates of prediction errors as double loop resampling procedures and for 75% of the cases these estimates are the same as for independent prediction sets. This has important implications for optimal use of a training set for component selection and estimation of prediction error. Two methods were developed and compared for selecting the optimal number of PLS components defined as the number where no statistically significant improvement in prediction error is observed when additional components are included in the model. Both methods determine a probability measure and provide similar results for model selection in this application.

1. Preface – Collaboration and friendship with Yi-zeng Liang

This paper is dedicated to Liang to pay tribute to a lifelong scientific collaboration and personal friendship. The work was prepared for the memorial session for Liang at the Chemometrics in Analytical Chemistry (CAC) conference in Halifax in June, but due to overlap with an important soccer tournament where my youngest son participated, I had to withdraw my presentation.

Liang came to Bergen in June 1990 on a sunny, warm day. I met him at the airport and the first thing I noticed was his smile and characteristic laughter that proved to be his trade mark. He came with a suitcase and 100 U$ in cash. On the way to his apartment, we stopped at a supermarket and I bought him the food he needed before he would get his first salary payment in Norway. Liang was employed on a post doc project financed by the Norwegian Research Council. After the first year, I applied for extra funding so that his wife and 7-years old son could come and live with him. After half a year in the elementary school, his son spoke Norwegian fluently with the typical Bergen dialect.

Liang was extremely hard-working and this virtue, combined with his excellent mathematical and programming skills, made his years in Bergen very productive. Together we developed several methods and applied them in many problem areas; sometimes in collaboration with other pioneers in chemometrics, such as Luc Massart and Richard Brereton. The most important development was the HELP method which together with evolving factor analysis (EFA) developed by Marcel Maeder and co-workers created a good theoretical and practical foundation for resolution of two-way multicomponent data. Before submitting the work, we realized that we needed a striking acronym for the method. Bjørn Grung, who was one of my PhD students at that time, and I had been active in rock and roll bands and we sometimes performed at karaoke bars when the research group was out partying. My own favorite performance was HELP with the Beatles and this became the acronym for the method. Heuristic evolving latent projections fitted to this acronym and it was also a good description of the method.

* Corresponding author.
E-mail address: Olav.Kvalheim@uib.no (O.M. Kvalheim).

https://doi.org/10.1016/j.chemolab.2019.03.006
Received 21 December 2018; Received in revised form 10 March 2019; Accepted 12 March 2019
Available online 21 March 2019
0169-7439/© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

After his post doc project, Liang returned to China with his family to take up a position at Hunan University at the department where his PhD supervisor, Ru-Qin Yu, was working. He soon returned to Bergen as a visiting scientist for a couple of months, first in 1994 and then in 1996-97. In 1994, he wrote and defended a thesis for the Norwegian doctor degree at the University of Bergen with another pioneer, Bernard Vandeginste, as first opponent. Afterwards, Liang and I travelled together to China and he took me on a trip to Zhangjiajie, at that time a 12 h drive by car on a narrow twisted mountain road from his home town Changsha in the Hunan province. This scenic region is famous for the Panda bear, the Yellow Dragon Cave and, last, but not least, the forest of peaks; hundreds of separated peaks. We decided that this was a perfect location for a conference, and with funding from both China and Norway, we organized the 1st Chemometric Conference in China together in 1997 with many distinguished chemometricians from both North America and Europe attending. On the excursion to one of the peaks, we had to walk a path consisting of several thousand steps of stone. A few delegates hired Chinese carriers, but it must have been a nerve breaking experience to be carried on the steep path.

My collaboration with Liang also resulted in an official agreement of student exchange between our universities. Three of his students took their PhD degrees at the University of Bergen and a few others worked for a year or so as post docs.

My collaboration and friendship with Liang brought me to mainland China and Hong Kong 15–20 times. Sometimes I came alone, but many times I was accompanied by collaborators, friends or family. My first visit with Liang to China was in 1991. It was my first travel to China ever and my first encounter with Beijing duck and many other exotic meals Liang wanted me to taste. In Beijing, we visited the Forbidden City, the Summer Palace and the favorite restaurant of Chairman Deng Xiaoping to enjoy Sichuan spicy food. Afterwards we travelled to Changsha where I was appointed consulting professor at Hunan university. At that time the infrastructure, roads and buildings, and the research facilities were poor. But each time I returned to Changsha, I was astonished by how quickly the infrastructure, the working conditions and financial opportunities for researchers, and their standard of living improved. On a trip to Zhangjiajie in 2009, Liang was driving us in his big Japanese car. He was very proud, but the Chinese driving culture was not the same as in the Western part of the world. So when we arrived the five stars hotel Liang had booked, I needed to take a shower before dinner. And Zhangjiajie was also completely changed with millions of visitors coming every year compared to the handful of tourists we observed during the conference in 1997. And the path we had walked in 1997 was no longer used. Lifts brought us swiftly up to the top of the peaks.

During these years, Liang came several times to Bergen. And we usually drove to my farm at the Fjord where we had small meetings on joint projects. During one of these seminars, in 2008, we had a Fjord excursion with an authentic copy of a Norwegian Viking ship picking us up and then ten of us, including Liang, had to row. In 2009, Liang spent time with us at the farm before the Scandinavian symposium on chemometric (SSC11) in Loen which I organized together with my wife Tarja Rajalahti and others from the University of Bergen. Bruce Kowalski was also staying with us and together we enjoyed a barbeque with good red wine before driving up to Loen the next day. One of the most exotic travels I did with Liang, was when we, surrounded by reindeer in a frozen and snowcovered landscape, visited Santa Claus together in Finnish Lappland in the winter of 2011.

In November 2013, our mutual friend and collaborator, Prof. Foo-Tim Chau at Hong Kong Polytechnic University, invited us to a conference on Chinese Medicine in Hong Kong. Sadly, this conference became our last meeting. When Liang passed away in October 2016 he had friends all over the world and a huge scientific production, both in terms of papers and students.

2. Introduction

Building a calibration model with good predictive performance requires many considerations and decisions that are critical for successful implementation [1] (and refs. therein). Among the tasks that need to be verified is that the method used to measure the predicted variables (the responses) provides reliable values over the range projected for future samples and that the method chosen to characterize the samples provides variable profiles that possess the information needed for satisfactory predictions. Often signal from interferences overlaps signal related to the predicted variables and therefore samples with such characteristics must be included in the calibration set to obtain reliable predictions from future samples.

After a representative sample set has been collected and analyzed, the modeling, i.e. establishing the mathematical relation between predictive and predicted variables, can be addressed. In analytical applications, the variables used for prediction are usually either instrumental continuous profiles or list of peaks that may comprise hundreds or even thousands of variables. At the same time, the number of available samples for building the calibration model is usually limited due to analysis costs and/or availability of samples. Therefore, the number of calibration samples is often much less than the number of variables used for prediction. This constraint can be handled by regression methods using linear combinations of the original variables to construct components, termed latent variables [2], to construct a pseudo inverse that can be used to predict the desired variables for future samples. Partial least squares (PLS) regression [3] has become one of the preferred choices for this purpose.

A critical task in PLS modelling is the selection of the number of PLS components to construct a pseudo inverse which is optimal for prediction. Too few components imply underfitting and too many lead to overfitting of data. Both outcomes may have a negative impact on the predictive performance. In order to guard against these situations, validation procedures simulating real prediction are routinely used in the modeling process. Validation is commonly done by resampling techniques, e.g., Monte Carlo [4–7] where a part of the calibration samples is randomly kept out, latent variable regression models built from the remaining samples with 1,2, …, A components, and the response variables predicted for the kept-out validation samples from the models. The procedure is repeated many times and measures derived from the mean or median root-mean-square error (RMSE) are subsequently used for model selection. The model with lowest RMSE [5], or the lowest RMSE after adjusting according to some heuristic [4] or statistical measure of significance [6,7], is selected as the model with the best predictive ability for future samples.

In repeated cross validation [6], also termed double cross validation [8], the samples are split into calibration/validation samples and prediction samples. In a repetitive outer loop prediction samples are randomly selected under the constraint that all samples in the outer loop are kept out once and only once for each repetition. The remaining samples are utilized in an inner loop where they are randomly partitioned into different groups. Each group is kept out once, and only once in each repetition in the outer loop and used as validation samples in the inner loop for determination of number of components, i.e. model selection. The RMSE values in the outer loop are utilized for estimation of prediction error.

The reason for splitting samples into an outer and inner loop is to separate the estimation of prediction error from the determination of the optimal number of components. This requires more available samples than for single loop procedures. Furthermore, all samples are actually used both as calibration/validation samples to find the optimal model and as prediction samples to estimate the prediction error although samples in inner and outer loop are not exchanged in a single repetition. An alternative approach would be to keep a part of the samples completely out and use that part to estimate the prediction error from the model obtained from the remaining samples. However, this may require even more extensive sampling since it requires a representative independent sample set for estimating the prediction error in addition to a representative sample set for modeling. The question is if such extensive sampling is necessary or if the prediction error can be estimated from a

single loop procedure just as well as from a double loop approach with exchange of calibration and prediction samples or from an independent sample set. Using Monte Carlo resampling on smaller sized sample sets randomly selected from a sample set with more than 1000 samples, we examine different strategies for model selection and calculation of prediction error with the aim to assess if outcomes differ significantly between the procedures. Although many other methods and strategies exist for these aims [1] (and refs. therein), the scope of this work is limited to Monte Carlo resampling strategies.

3. Theory

3.1. Design used to partition samples

By using smaller sized sample sets randomly selected from a set with more than 900 samples, Martens and Dardenne [4] examined Monte Carlo resampling procedures for model selection and calculation of prediction error. They derived RMSE for not only validation samples and prediction samples, but also for samples not included in the calibration or validation/prediction sets. We follow a similar idea in this investigation, but use a more systematic approach to design sample sets in the different loops. Thus, the algorithm used in this work for estimating the optimal number of components and the prediction error of the PLS regression models splits the samples into three parts. First, samples to be used for inner and outer loop are randomly selected from the total sample set. The remaining samples are used as a pool to construct independent sample sets. The samples selected for the inner and outer loop are repeatedly split with one third of the samples in the outer loop and two thirds in the inner loop. The samples in the inner loop is further split repeatedly half and half between calibration and validation samples. Using a probability measure p to assess the optimal number of PLS components, we showed recently [7], that the same number of components was obtained by splitting a sample set half and half as with a ratio of nine to one between calibration and validation samples. This was ascribed to the fact that the variation in RMSE values for the validation sets increased with reduction of validation samples and that this effect counterbalanced the improved model description obtained by simultaneously increasing the number of calibration samples. Therefore, splitting a sample set equally between calibration and validation set is reasonable as long as a sufficient number of samples is available. As pointed out by Xu et al. [13] this may not be optimal for smaller sized sample sets. In such cases, the prediction ability tends to be underestimated by Monte Carlo resampling methods since too few samples are available for the calibration step.

Only the samples in the inner loop are used for determining the number of components, while the prediction error is estimated in three ways: i) from the validation samples in the inner loop, ii) the prediction samples in the outer loop, and, iii) from the samples randomly selected from the independent data set. With a large number of samples available, this algorithm can be used both to mimic sample sets of different size and to compare the results of the three different strategies for estimation of prediction error.

3.2. Determining the optimal number of components

In our recent work [7], we developed a strategy to determine the optimal number of PLS components, A_{opt}, based on the distributions of $\{RMSE_{val,m,a},\ m = 1,2\ \ldots,\ M;\ a = 1,2\ ..,\ A\}$ obtained by Monte Carlo resampling, by repeatedly splitting a sample set between calibration and validation samples. The index m runs over the number of repetitions in a single round in the inner loop with M being the total number of repetitions, while the index a runs over the number of PLS components with A representing the maximum number of calculated components. The subscript *val* implies that only samples in the validation sets are used for the calculation of RMSE.

Let $n_{val,m}$ be the number of validation samples. For a single selection of validation samples the mean squared error (MSE) for the segment m with a PLS components in the inner loop is then calculated as

$$MSE_{val,m,a} = \frac{\sum_{i=1}^{n_{val,m}} \left(y_{val,m,a,i} - \hat{y}_{val,m,a,i}\right)^2}{n_{val,m}} \quad m = 1, 2, \ldots, M; \quad a = 1, 2, \ldots, A \tag{1}$$

The root mean square values $\{RMSE_{val,m,a}\}$ for the validation samples is then:

$$RMSE_{val,m,a} = \sqrt{MSE_{val,m,a}}\ m = 1,2, \ldots, M;\ a = 1, 2, \ldots, A \tag{2}$$

Eq. (2) provides a distribution of M RMSE values for each model dimension a.

The median ($RMSE_{val,a}$) of the distribution of $\{RMSE_{val,m,a},\ m = 1, 2, \ldots, M\}$ is located for each PLS model dimension a. Median is preferred to arithmetic mean because the distribution of RMSE values for a model cannot *a priori* be assumed normally distributed. The model with the lowest median RMSE, with A_{min} components, determines the starting point for a backward selection to decide the optimal number of PLS components.

In our previous work [7], we calculated the fraction $f_{val,Amin}$ for which the $\{RMSE_{val,m,Amin},\ m = 1, 2, \ldots, M\}$ values are larger or equal to the median RMSE for the PLS model with $A_{min} - 1$ components, i.e.

$$\text{median}(RMSE_{val,Amin\text{-}1}) \le RMSE_{val,m,Amin}\ m = 1, 2, \ldots, M \tag{3}$$

The fraction $f_{val,Amin}$ obtained from Eq. (3) represents a measure p that can assist in the decision whether component A_{min} is significant or not and thus define the optimum number of PLS components, A_{opt}. If the decision is that the component is not significant, i.e. the fraction $f_{val,Amin}$ is smaller than a preselected threshold (see below), we continue with the previous component and repeat the procedure by comparing the distribution of values $\{RMSE_{val,m,Amin\text{-}1},\ m = 1, 2, \ldots, M\}$ with the median ($RMSE_{val,m,Amin\text{-}2}$). This approach quantifies the common heuristic visual procedure of trying to locate from a plot of RMSE values vs. number of PLS components when the change in RMSE is levelling off.

We have also implemented an alternative test where we compare the median RMSE of the model with A_{min} components with models with successively one less component:

$$\text{median}(RMSE_{val,Amin\text{-}q}) \le RMSE_{val,m,Amin}\ m = 1, 2, \ldots, M;\ q = 1, 2, A_{min} \tag{4}$$

This approach resembles the procedure used by Filzmoser et al. [6] where an error term is added to the RMSE of the A_{min} component before comparison with the mean RMSE of the previous components to locate A_{opt}. A possible drawback of this approach is that it may lead to acceptance of components in an almost "flat" region, i.e. when several consecutive PLS components each provides a small decrease in RMSE that adds up to exceed the threshold for acceptance. This may lead to extra components compared to the pairwise comparison (Eq. (3)) and thus make the model selection more vulnerable to overfitting.

We recently showed [7] that the probability p defined as the fraction obtained from Eq. (3) could be used to design a formal significance test. Thus, we choose a critical p-value, p_{upper}, and test the null hypothesis that component A_{min} is not significant, i.e. that p calculated by comparing the distribution of RMSE-values for A_{min} components with the median RMSE of the PLS model with one less component using Eq. (3), is larger than p_{upper}. If p is less or equal to p_{upper} the null hypothesis is rejected and the model with A_{min} components is selected as optimal. This probability enables the opportunity to take an informed decision of optimum number of components where the user can balance the risk of overfitting against the risk of underfitting through the choice of p_{upper}. The test is non-parametric since no distributional assumptions are necessary. Similarly, Eq. (4) can be used to design a similar significance test by comparing the distribution of RMSE values on component A_{min} with the medians of the previous components.

For approximately normally distributed RMSE values, p_{upper} has a one-to-one correspondence to the standard deviation around the median

RMSE. Thus, p_{upper} of 0.159, 0.308 and 0.401 corresponds to 1.0, 0.5 and 0.25 standard deviations, respectively, while $p_{upper} = 0.5$ corresponds to choosing the model with number of components corresponding to minimum median, i.e. median ($RMSE_{Amin}$), as optimal. Lowering the threshold reduces the risk of overfitting, but increases the risk of underfitting. Although our test is nonparametric and does not assume normally distributed RMSEs, we use p_{upper} equal to 0.401 and 0.308 for model selection in this work since the distributions of RMSEs may often be approximately normally distributed for a well-designed sample set.

In this work, component selection in the inner loop is performed multiple times, i.e. in correspondence with the number of repetitions in the outer loop. This makes it possible to use the distribution of p to assess how robust A_{opt} is with respect to the number of samples available for the resampling procedure and the number of repetitions in the outer loop. Furthermore, the model evaluation in this work provides indications of the sensitivity of the model selection to the choice of p_{upper} for the two criteria (Eqs. (3) and (4)) for backward comparison to decide A_{opt} discussed above.

3.3. Estimating the prediction error

When the aim is not to use the regression model for predictions for future samples, but only to reveal underlying structures and association patterns that are predictive and not just descriptive in the collected sample set, it is sufficient to determine the dimension A_{opt} of the model. However, when a model is built to be used for prediction for future samples, predictive performance must be quantified and for this purpose an estimate of expected prediction error is required.

The design used for partitioning the total sample set in this work, provides multiple opportunities to estimate the prediction error. Let k be the index running over the outer loop and K the total number of repetitions in the outer loop. For each repetition, we predict the y-variables for the samples in the outer loop and calculate the RMSEs for different dimensions a of the PLS models, i.e. $\{RMSE_{outer,k,a}, k = 1, 2, \ldots, K; a = 1,2, \ldots, A\}$. Simultaneously we calculate $\{RMSE_{indep,k,a}, k = 1, 2, \ldots, K; a = 1,2, \ldots, A\}$ for the same number of samples randomly drawn from the total pool of independent samples for each repetition in outer loop. The subscript *indep* is used to imply that RMSE is calculated from samples drawn from this pool. After execution of the K repetitions, the optimal model dimension A_{opt} is obtained from the inner loop and the medians for the distribution of RMSEs for prediction samples in outer loop and independent samples are located, i.e. the medians median ($RMSE_{outer,k,Aopt}$) and median ($RMSE_{indep,k,Aopt}$). These medians represent expected prediction errors in similar future samples. The mean and standard deviation of these medians for the K repetitions can be used to compare the prediction performance estimated from samples in the outer loop and samples drawn from the pool of independent prediction samples.

We can calculate corresponding estimates for the prediction performance of the regression model for the validation samples in the inner loop and compare these with the results obtained for the outer loop and the independent samples. However, during a single repetition in the outer loop, the Monte Carlo resampling is repeated M times in the inner loop and this provides M sets of validation samples that can be used to obtain the median and error bounds of RMSE. This provides a distribution of medians for RMSE for each repetition, i.e. $\{median\ (RMSE_{val,k,Aopt}), k = 1, 2, \ldots, K\}$. This narrows the range of medians in the inner loop compared to the outer loop and independent samples and thus leads to a smaller standard deviation of the medians around the mean prediction error calculated for the inner loop.

4. Experimental

4.1. Sampling protocol

Blood samples were collected over a period of 6 weeks in early autumn. Sampling was done between 8 and 10 a.m. after overnight fasting for a cohort of 1057 ten year old Norwegian children from the rural Fjord region of Western Norway. This is an area with a homogeneous population of ethnic Norwegians. Serum was obtained according to a standardized protocol [9], split into 0.5 ml aliquots, and stored in cryo tubes at −80 °C. At this temperature, the lipoproteins are stable for several years [10].

4.2. Analyses

Total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C) and total triglyceride (TG) were quantified by two different analytical methods: i) proton nuclear magnetic resonance (NMR) spectroscopy using nuclear Overhauser effect spectroscopy (NOESY) [11], and, ii) the standard protocol for analysis of blood samples with LDL-C estimated from TC, HDL-C and TG by the Friedewald equation [12]. The NMR analyses were performed at NTNU (Trondheim, Norway), while the standard analyses were performed at the Endocrine Laboratory of the VU University Medical Center (VUmc; Amsterdam, the Netherlands). The samples were transported to the analysis sites in boxes with dry ice to keep the temperature stable at minus 78.5 °C. During and after transportation the amount of dry ice was controlled, and, if necessary, dry ice was added.

4.3. Selection of shift regions and data pre-treatment

After baseline correction and spectral alignment using the Alanine doublet at approx. 1.5 ppm, the NMR profiles for the shift region 1.53–0.60 ppm, comprising 2041 spectral variables, were examined for use as possible explanatory variables. This shift region embraces the lipoprotein triglyceride peak at 1.3 ppm and lipoprotein cholesterol peak at approx. 0.88 ppm. These regions contain quantitative information about TC, LDL-C, HDL-C and TG. The number of spectral variables was further reduced by eliminating regions with low intensities or dominated by interferents. The remaining 1172 variables were selected as descriptors for the triglyceride and cholesterol lipoproteins. Without any further pretreatment, these profiles were selected as explanatory variables to the PLS modelling with TC, HDL-C, LDL-C and TG concentrations determined by the standard method [12] as response variables. Reproducibility was checked by replication of the total analytical procedure at several timepoints from the first until the last sample analyzed.

4.4. Modelling

By randomly selecting 150 and 300 samples, respectively, from the total pool of samples available, we created two sample sets to be used in the inner and outer loop in the Monte Carlo resampling algorithm. Separate PLS models for TC, LDL-C, HDL-C and TG measured by the standard method (response) and NMR (exploratory variables) were calculated for the two sample sets with *1*, *2*, …, *10* PLS components using the resampling procedure described in section 2.1. 100 repetitions were performed in the outer loop repeatedly splitting samples randomly between outer and inner loop with one third of samples in the outer loop and two-thirds of the samples in the inner loop. For each repetition in outer loop, 100 repetitions were performed in the inner loop splitting samples randomly half and half into calibration and validation samples. Furthermore, in order to mimic the situation with independent sample sets, for each repetition in the outer loop, we randomly draw sample sets of the same size as in the outer loop from the remaining samples in the pool. This creates an independent sample set with 100 samples and 50 samples for each repetition in the outer loop for the sample sets of size 300 and 150, respectively. This design allows us to assess different approaches to model selection and estimation of prediction error as well as the effect of number of samples available on these estimates.

4.5. Model selection

The two criteria defined by Eqs. (3) and (4) were used for selection of the optimal model for each response for the two sample sizes defined in section 3.4. For each repetition in the outer loop, the RMSE values for $a = 1, 2, \ldots, 10$ components for all the 100 repetitions in the inner loop were calculated for the validation samples using Eqs. (1) and (2). From the distributions of 100 RMSE values, the medians were determined for each model dimension. The minimum median was located, i.e. median ($RMSE_{Amin}$) where A_{min} implies the number of PLS components in the model with minimum median for the *10* calculated components. Subsequently, we calculated probabilities for models using the two criteria defined by Eqs. (3) and (4). These probability measures were compared with the threshold, p_{upper}. In this work, we assess p_{upper} equal to 0.308 0.401. With 100 repetitions in the outer loop, we get a distribution of 100 *p*-values in the inner loop. These distributions provided us with the possibility to check the stability of *p*-values by plotting their distributions for different choices of *a*.

4.6. Estimation of prediction errors

For all models, the mean prediction error and standard deviation was estimated for the outer loop, the independent data sets and the validation samples in the inner loop from the distributions of median prediction errors as described in the theoretical section.

5. Results and discussion

5.1. Model selection

Table 1 shows the outcome of the 100 repetitions for model selection for the four responses using the two criteria defined by Eqs. (3) and (4), respectively, with p_{upper} chosen as 0.401 and 0.308 corresponding to 0.25 and 0.5 standard deviation around the mean *p*-value, respectively, for normally distributed RMSE-values. Our experience is that when the majority of the systematic variation has been accounted for by the first PLS components in the models, the RMSE distributions approximate normal distribution. So although the test is nonparametric and does not require normally distributed RMSE-values, choices of p_{upper} corresponding to fractions of the standard deviation for a normal distribution has the advantage of connecting p_{upper} to a well-known statistical metric.

Using the majority count for model selection, Table 1 reveals that the two tests for model selection defined by Eqs. (3) and (4) provide the same A_{opt} in 13 out of the 16 cases defined by the 4 responses based on either 50 or 100 validation samples in the inner loop and with p_{upper} at two levels. In the remaining three cases pairwise comparison (Eq. (3)) implies one PLS component less than comparison of the median RMSE of the A_{min} model with models with successively one less component (Eq. (4)). This behavior is to be expected when the median RMSE corresponding to A_{min} is situated in a rather "flat region" of RMSE for adjacent PLS components. Thus, as expected, the pairwise comparison seems more robust in guarding against overfitting in this situation.

Table 1 shows that doubling the number of calibration samples from 50 to 100 has little impact on the model selection with $p_{upper} = 0.308$ for both tests. For pairwise comparison, 3 out of 4 cases provide the same A_{opt} for the two sample sizes, while for HDL-C doubling of samples increases the number of components from 7 to 9. The same increase in model dimension for HDL-C is observed for the other test. In addition one more component becomes significant for LDL-C and one less for TG for this test. For $p_{upper} = 0.401$ doubling the sample size seems to have a larger effect: Only for TG is there no change in model dimension. In 5 cases the model dimension increases from 7 to 10 and in one case from 8 to 10. The effect on model dimension of increasing p_{upper} is similar for both tests.

Increased model dimension accompanying larger sample size is probably caused by better coverage of the sample variation which provides calibration and validation sets with increased similarity. This raises the odds of minor PLS components to become significant. With an improvement of 1–2% in R2 for the measured and predicted responses for

Table 1
Distributions of selected models for the four responses using Eqs. (3) and (4) for 50 and 100 validation/outer loop/independent samples at the two probability levels 0.308 and 0.401 for p_{upper}. Majority models are highlighted. Models with less than 5 counts for all 8 combinations are not listed.

		$p_{upper} = 0.308$				$p_{upper} = 0.401$			
		50 samples		100 samples		50 samples		100 samples	
Var.	nLV	Eq. 3	Eq. 4	Eq. 3	Eq. 4	Eq. 3	Eq. 4	Eq. 3	Eq. 4
TC	6	74	73	82	47	32	31	25	18
	7	26	27	8	21	66	67	20	17
	9	0	0	0	21	0	0	2	12
	10	0	0	10	10	0	0	51	51
LDL-C	6	78	49	49	26	17	9	3	2
	7	19	44	35	38	53	52	37	30
	8	2	4	0	4	23	25	0	3
	9	0	2	3	19	5	12	6	11
	10	0	0	13	13	2	2	54	54
HDL-C	5	21	0	0	0	0	0	0	0
	6	30	13	0	0	6	2	0	0
	7	33	40	7	0	30	15	0	0
	8	10	30	19	1	27	37	0	0
	9	5	17	42	67	27	36	15	15
	10	0	0	32	32	10	10	85	85
TG	4	8	7	7	4	1	1	0	0
	5	48	28	80	76	11	11	32	30
	6	38	62	10	18	56	56	56	55
	7	3	3	1	1	32	31	7	6
	9	0	0	0	1	0	1	4	6

TC: total cholesterol; LDL-C: low-density lipoprotein cholesterol; HDL-C: high-density lipoprotein cholesterol; TG: total triglyceride; nLV: number of PLS components.

these models, the choice between models have minor practical consequences for our application, but such improvements may of course be important in other applications.

Table 1 allows an assessment of the impact of the choice of p_{upper} on the model selection. For both tests the number of valid components is increasing by one for 5 out of 8 cases when increasing the p_{upper} from 0.308 to 0.401. For one case for each test, there is no change, while for the remaining two cases for each test, the number of valid components increases by 3 or 4. The largest changes are observed for cases with the largest sample size. Thus, increasing p_{upper} causes a similar effect as increasing the sample size by leading to acceptance of more of the minor components. Increasing p_{upper} affects both tests in the same way for our application.

The overall picture observed in Table 1 is that both tests imply the same number of valid components. Only for 3 out of the 16 cases did the test based on Eq. (4) find one additional component compared to pairwise comparison probably caused by a flat region of RMSE values adjacent to A_{min}. Further investigations on many data sets are necessary to decide if one test performs better than the other in the long run. In most cases, increase of p_{upper} resulted in at least one additional PLS component for the models.

The choice of p_{upper} may be important for predictive performance. One way to determine an adequate threshold for p_{upper} is to examine the distributions of the obtained p-values by plotting them as histograms for the 100 repetitions performed. Fig. 1 displays these distributions for pairwise comparison for different number of PLS components for the TC with 50 (Fig. 1, upper) and 100 (Fig. 1, lower) calibration samples. By comparing the fractions of p-values below 0.5 with the fraction above 0.5, it is obvious that a 7-component model is the optimal choice with 50 calibration samples and 10 components is the optimal choice with 100 calibration samples. This complies with the results based on the majority count of p-values for $p_{upper} = 0.401$, while $p_{upper} = 0.308$ results in underfitting for TC. Fig. 2 and Supplementary material (S-1 and S-2) display similar plots for the three other responses. Comparing the fractions of p-values below 0.5 with the fraction above 0.5 for these responses, it is clear that for all cases $p_{upper} = 0.308$ results in underfitting, while $p_{upper} = 0.401$ for all cases except HDL-C with a sample size of 50 implies the same number of components using the two tests for model selection (Table 1) or the histograms.

For HDL-C with 50 samples, the distribution of p-values (Fig. 2) implies 9 valid components implying that even $p_{upper} = 0.401$ may lead to underfitting by losing one or two minor components using Eq. (3) or 4, respectively. Note, however, that using Eq. (4), the number of counts is almost identical for the 8- and 9-component model for HDL-C, 37 compared to 36 counts. Thus, the results from Eq. (4) and the corresponding histogram do not contradict each other. The results from the histograms may imply that the threshold p_{upper} can be increased for the pairwise test (Eq. (3)) compared to the one defined by Eq. (4).

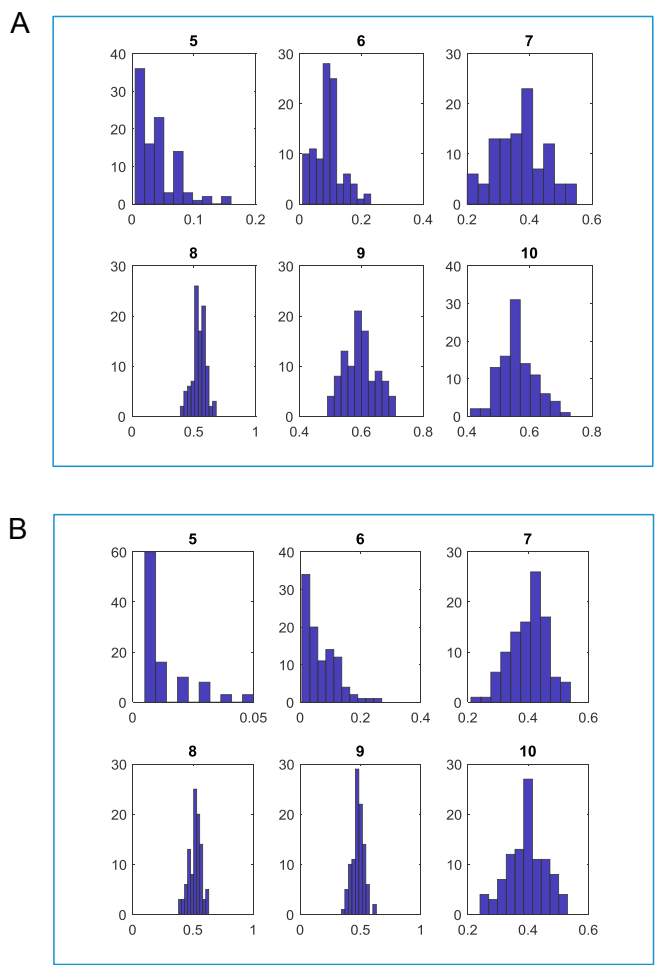

Fig. 1. Histograms displaying the number of models (y-axis) plotted versus the p-values (x-axis) calculated by pairwise comparison (Eq. (3)) for models with 10, 9, …, 5 PLS components. Total cholesterol (TC) is response variable and 100 estimations were performed with total number of validation samples in inner loop as 50 (A) and 100 (B).

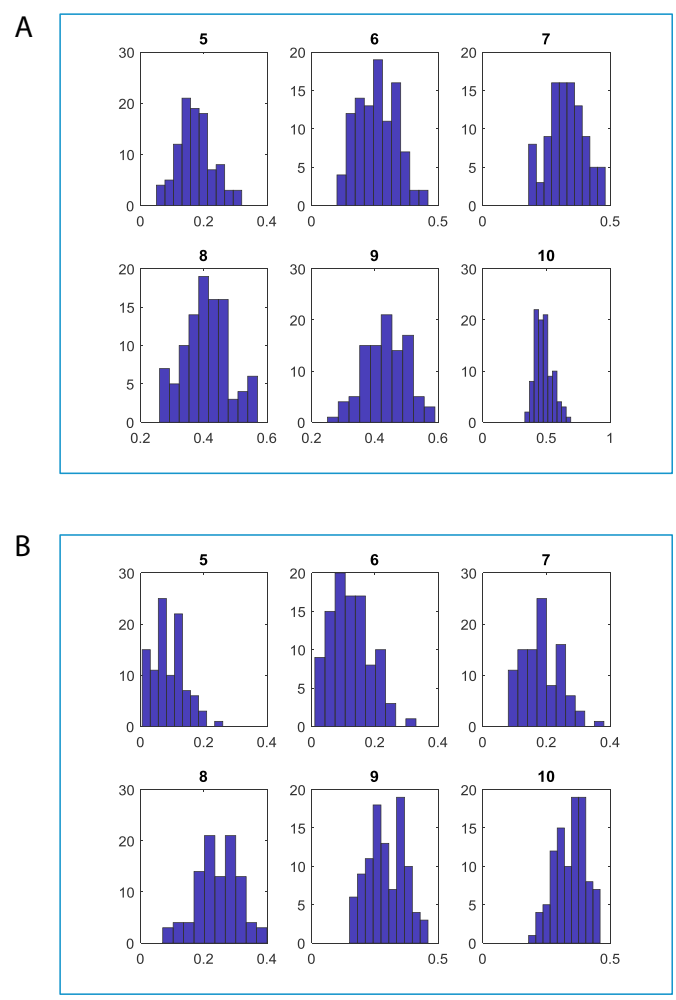

Fig. 2. Histograms displaying the number of models (y-axis) plotted versus the p-values (x-axis) calculated by pairwise comparison (Eq. (3)) for models with 10, 9, …, 5 PLS components. High-density lipoprotein cholesterol (HDL-C) is response variable and 100 estimations were performed with total number of validation samples in inner loop as 50 (A) and 100 (B).

5.2. Estimation of prediction error

Table 2 shows the estimated mean prediction errors together with their standard deviations for the validation samples from the inner loop and the prediction samples in the outer loop and the pool of independent samples outside the inner and outer loop. We show mean and standard deviation in Table 2 instead of median and upper and lower bounds to simplify the comparison and discussion of the prediction results since for the selected models the RMSE distributions are approximately normal. Pairwise comparison with $p_{upper} = 0.401$ was used for model selection.

The overall picture from Table 2 is that the prediction errors estimated from the different validation and prediction sets are very similar. The agreement is particularly striking between estimates from inner loop and outer loop sample sets. This is maybe not surprising since the resampling procedures in inner and outer loop use the same pool of samples and thus represents an exchange of samples between inner and outer loop selected from the same pool. When repeated and averaged over a relatively large number of repetitions we should expect the same prediction estimates from both loops if the samples in the calibration/validation sets (inner loop) and prediction sets (outer loop) span the same variation range. If different estimates are obtained in the two loops it may be a sign of too few samples available to span the variation range in inner and outer loop simultaneously. We observe (Table 2) that the standard deviations for the validation samples in the inner loop are always smaller than for the outer loop prediction samples. As explained in the theory section, this is expected since the median prediction errors in the inner loop span a narrower range, since they are calculated from 100 repetitions for each repetition in the outer loop. The standard deviations of these medians around their mean in the inner loop must therefore be smaller than the corresponding standard deviations in the outer loop. Comparison of the estimated mean prediction errors for the independent samples and the outer loop samples also reveals close similarity for most models. However, for two cases, LDL-C and TC with respectively 50 and 100 samples in outer loop and independent sample sets, a two-sided *t*-test shows that the estimated mean prediction errors are significantly different ($p < 0.001$) when making the reasonable assumptions of equal variance of the two means. All the other prediction errors are statistically identical ($p > 0.1$).

For TC with 100 samples and LDL-C with 50 samples, the prediction error estimated from a pool of independent samples is respectively 9–10% and 17–18% higher than the corresponding estimate from the inner and outer loop samples. These differences, however, are of little practical relevance in the application investigated here since the results for these variables are usually reported with only one significant digit after comma for the reference method. Thus, the more elaborate sampling implied to obtain a large enough pool of samples to calculate these prediction measures may not be justified in terms of the possible improvements achieved. This will probably often be the case when developing and validating calibration models. In any case, further validation and maintenance is usually necessary after putting the models at work and it may be better to spend extra resources on this step. Furthermore, independent prediction sets collected and analyzed at a later stage are important since they can disclose deficiencies in sampling or analysis that impact the performance of the calibration model.

Table 2 further shows that doubling the number of samples from 50 to 100 increases the number of valid PLS components from 7 to 10 for three responses. This is caused by better description of sample variation leading to stabilization of minor PLS components often accompanying increased number of available samples. A minor but statistically significant ($p < 0.001$ for a two-sided *t*-test) improvement in mean prediction error is observed for HDL-C by doubling the number of samples. As discussed above, this small improvement is of no practical importance for the present application. For LDL-C the prediction error calculated from the inner and outer loop with 100 samples increases, while it decreases for samples from the pool of samples kept outside. This may be caused by inclusion of samples in the inner loop increasing the variation range, thus improving the prediction of some samples in the outside pool, but at the same time leading to a more heterogeneous sample set and model and thus larger overall prediction error in the inner and outer loop.

6. Conclusions

The two ways of calculating and using a probability measure for model selection provided identical outcome for more than 80% of the cases investigated here. For the cases with different outcome, pairwise comparison (Eq. (3)) suggested one PLS component less than the other procedure (Eq. (4)) for backward selection. The choice of p_{upper} appears to have a larger impact on model selection than the choice of test and 0.4 proved to be a good compromise for balancing the risk of overfitting against underfitting in this application. However, the "best" choice may be application dependent due to factors such as the number of samples available and, the complexity and heterogeneity in the composition of samples so it.

The large sample available here made it possible to estimate and compare alternative ways of estimating the prediction error: i) estimation from inner loop validation samples, ii) from outer loop prediction samples, and, iii) from independent prediction samples outside the pool of inner and outer loop samples. The major finding was that the double loop procedure provides the same prediction error estimates as the inner loop validation samples. Thus, our results imply that using samples resampled from the same pool in double loop procedures may not provide better estimates of prediction errors than the single loop procedure. Rather a significant difference between estimates of prediction error from inner and outer loop may imply that the number of samples is too small to ensure a satisfactory coverage of sample variations in both inner loop validation samples and outer loop samples simultaneously. This breaks with the assumption that the samples collected for calibration must cover the range and characteristics expected for the future samples to be predicted reliably from the model. A better approach is to select true prediction samples from a pool of independent samples to be used in the outer loop. This may require collection and analysis of extra samples and thus increase the cost of calibration, but will provide an estimate of prediction error from samples not used for model selection. The single loop procedure has the advantages of demanding a smaller sample set for building a calibration model and being less computer-intensive.

Table 2
Mean prediction errors and their standard deviations (SD) calculated for 50 and 100 validation samples in inner loop and the same number of prediction samples in outer loop and in the sample sets randomly sampled from the pool of samples outside both inner and outer loop. Pairwise comparison (Eq. (3)) was used for model selection. Superscript a and b implies 50 and 100 samples, respectively.

Var.	nLV	Inner loop[a]	Outer loop[a]	Independent[a]	nLV	Inner loop[b]	Outer loop[b]	Independent[b]
		Mean ± SD	Mean ± SD	Mean ± SD		Mean ± SD	Mean ± SD	Mean ± SD
TC	7	0.198 ± 0.013	0.193 ± 0.024	0.202 ± 0.040	10	0.183 ± 0.016	0.186 ± 0.026	0.202 ± 0.025
LDL-C	7	0.145 ± 0.008	0.146 ± 0.019	0.171 ± 0.027	10	0.158 ± 0.007	0.160 ± 0.015	0.156 ± 0.021
HDL-C	7	0.120 ± 0.008	0.121 ± 0.014	0.122 ± 0.027	10	0.097 ± 0.004	0.098 ± 0.009	0.099 ± 0.012
TG	6	0.055 ± 0.003	0.054 ± 0.007	0.053 ± 0.009	6	0.053 ± 0.002	0.054 ± 0.005	0.050 ± 0.005

TC: total cholesterol; LDL-C: low-density lipoprotein cholesterol; HDL-C: high-density lipoprotein cholesterol; TG: total triglyceride; nLV: number of PLS components.

Declaration of interest

None.

Acknowledgements

T. Rajalahti acknowledges financial support from The Norwegian Research Council and Førde Health Trust. Geir Kåre Resaland and his team at Western Norway University of Applied Sciences are thanked for recruiting and organizing the blood sampling of the cohort of children.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2019.03.006.

References

[1] F. Westad, F. Marini, Validation of chemometric models - a tutorial, Anal. Chim. Acta 893 (2015) 14–24.
[2] O.M. Kvalheim, History, philosophy and mathematical basis of the latent variable approach – from a peculiarity in psychology to a general method for analysis of multivariate data, J. Chemom. 26 (2012) 210–217.
[3] S. Wold, A. Ruhe, H. Wold, W.J. Dunn III, The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, SIAM J. Sci. Stat. Comput. 5 (1984) 735–743.
[4] H.A. Martens, P. Dardenne, Validation and verification of regression in small data sets, Chemometr. Intell. Lab. Syst. 44 (1998) 99–121.
[5] Q.S. Xu, Y.Z. Liang, Monte Carlo cross validation, Chemometr. Intell. Lab. Syst. 56 (2001) 1–11.
[6] P. Filzmoser, B. Liebmann, K. Varmuza, Repeated double cross validation, J. Chemom. 23 (2009) 160–171.
[7] O.M. Kvalheim, R. Arneberg, B. Grung, T. Rajalahti, Determination of optimum number of components in partial least squares regression from distributions of the root-mean-squared error obtained by Monte Carlo resampling, J. Chemom. 32 (2018) e2993.
[8] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Assessment of PLSDA cross validation, Metabolomics 4 (2008) 81–89.
[9] C. Lin, T. Rajalahti, S.A. Mjøs, O.M. Kvalheim, Predictive associations between serum fatty acid and lipoproteins in healthy non-obese Norwegians – implications for cardiovascular health, Metabolomics 12 (2016), 12:6.
[10] E.H.J.M. Jansen, P.K. Beekhof, E. Schenk, Long term stability of parameters of lipid metabolism in frozen human serum: triglycerides, free fatty acids, total-, HDL- and LDL-cholesterol, apolipoprotein-A1 and B, J. Mol. Biomark. Diagn. 5 (2014) 182.
[11] V.V. Mihaleva, D.B. van Schalkwijk, A.A. de Graaf, J. van Duynhoven, F.A. van Dorsten, J. Vervoort, A. Smilde, J.A. Westerhuis, D.M. Jacobs, A systematic approach to obtain validated partial least square models for predicting lipoprotein subclasses from serum NMR spectra, Anal. Chem. 86 (2014) 543–550.
[12] W.T. Friedewald, R.I. Levy, D.S. Fredrickson, Estimation of the concentration of low- density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge, Clin. Chem. 18 (1972) 499–502.
[13] Q.S. Xu, Y.Z. Liang, Y.P. Du, Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration, J. Chemom. 18 (2004) 112–120.