Received: 16 October 2019 | Revised: 5 December 2019 | Accepted: 6 December 2019

DOI: 10.1002/cem.3211

SPECIAL ISSUE - RESEARCH ARTICLE

WILEY Journal of CHEMOMETRICS

Variable importance: Comparison of selectivity ratio and significance multivariate correlation for interpretation of latent-variable regression models

Olav M. Kvalheim

Department of Chemistry, University of Bergen, Bergen, Norway

Correspondence
Olav M. Kvalheim, Department of Chemistry, University of Bergen. Bergen, Norway.
Email: Olav.Kvalheim@kj.uib.no

Abstract

This work examines the performance of significance multivariate correlation (sMC) and selectivity ratio (SR) for ranking variables according to their importance in latent-variable regressions (LVRs) models. Both indices are based on target projection (TP) of a validated LVR model obtained by partial least squares (PLS). The matrix of explanatory *x*-variables is projected on the normalized regression vector to obtain a score vector that is proportional to the vector of predicted values for the response variable *y*. sMC for each *x*-variable is calculated by dividing the squared variance explained by the decomposition obtained from these two vectors on the squared residuals. This is similar to how SR is calculated except that for SR, the regression vector is replaced by the loading matrix obtained by projecting the data matrix of *x*-variables onto the score matrix obtained by TP. The two indices for variable importance are compared for three different applications with data representing instrumental profiles from liquid chromatography, infrared spectroscopy, and proton nuclear magnetic spectroscopy. Results show that SR outperforms sMC for interpretation and biomarker selection. The main drawback of sMC appears to be the mixing of predictive and orthogonal variation resulting from the direct use of the normalized regression vector in the calculation. SR uses a loading vector that is proportional to the covariances between *x*-variables and the predicted response variable.

KEYWORDS

biomarkers, selectivity ratio, significance multivariate correlation, variable importance, variable selection

1 | INTRODUCTION

Measures for variable importance are crucial for interpretation and biomarker selection using partial least squares (PLS)[1] or any other method based on latent-variable regression (LVR) modeling.[2,3] Many measures and visualizations[4] have been proposed and assessed for their appropriateness and usefulness to rank variables according to importance. The most used approach in chemometrics is without doubt variable influence in projections (VIPs).[1] Other methods such as selectivity ratio (SR)[5,6] and the related method significance multivariate correlation (sMC)[7] as well as several modifications of VIP adapted to orthogonal PLS[8] have been proposed and compared in various investigations.[9-11] The results of comparisons, however, are not conclusive. One reason for confusion is that the results partly depend on the purpose of the modeling, but also erroneous implementation of methods has occurred. Thus, Andries et al[11] calculated SR as the ratio of variance explained to residual variance in the "standard" PLS model. This leads of course to wrong conclusions. The need for indices of variable of importance is most obvious for applications searching for patterns of biomarkers. Commonly, samples are analyzed on an instrument capable of providing a profile of the composition that potentially can have contributions from hundreds of chemical constituents, and data analysis is performed to reveal the important components. Typical examples are the search for bioactive components in complex extracts of natural products[12-15] and for disease patterns from samples of body fluids.[16] The first application area usually involves a continuous response variable measuring some kind of bioactivity of the total extract, while disease patterns are typically revealed by using a binary response variable describing the condition as healthy or ill, the so-called PLS discriminant analysis (PLS-DA). Despite relevant criticism,[17] PLS-DA is still preferred over classical methods for classification and discrimination in major application areas, ie, metabolomics. In this work, the aim is to assess variable importance as portrayed by SR and sMC. We first provide the algorithm to calculate SR and sMC and then aim at examining their performance for model interpretation and for biomarker and variable selection. Three datasets with continuous response variables are investigated, embracing applications within mixture analysis, process analysis, and metabolomics.

2 | THEORY

Assume that a validated LVR model has been obtained by, eg, PLS.[1] The regression model can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{LVR} + \mathbf{f} \tag{1}$$

Here,

$$\mathbf{b}_{LVR} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X} = \mathbf{X}_{LVR}{}^{+} \tag{2}$$

In Equations (1) and (2), the matrix $\mathbf{X}$ and the vector $\mathbf{y}$ represent the preprocessed and mean-centred values of the explanatory variables and the response variable, respectively, while $\mathbf{f}$ is the vector of y-residuals and $\mathbf{b}_{\mathrm{LVR}}$ the regression vector connecting the response to the explanatory variables. Superscript T implies the operation of transposing a vector or matrix, superscript -1 implies the inverse, and superscript $+$ implies the generalized inverse obtained from the specific LVR algorithm chosen.

The decomposition of $\mathbf{X}$ can be written as a product of scores $\mathbf{T}$ and loadings $\mathbf{P}$ matrices plus a residual matrix:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}_{LVR} = \mathbf{X}_{LVR} + \mathbf{E}_{LVR} \tag{3}$$

The predictive part of $\mathbf{X}_{\mathrm{LVR}}$ can be partitioned into a predictive part $\mathbf{X}_{\mathrm{LVR,pred}}$ and an orthogonal part $\mathbf{X}_{\mathrm{LVR,orth}}$:

$$\mathbf{X}_{LVR} = \mathbf{X}_{LVR,pred} + \mathbf{X}_{LVR,orth} \tag{4}$$

The predictive part of the matrix $\mathbf{X}$ can be obtained by target projection (TP)[18,19] using the following algorithm:

Select the normalized regression coefficients as weights:

$$\mathbf{w}_{TP} = \mathbf{b}_{LVR}/||\mathbf{b}_{LVR}|| \tag{5}$$

Calculate the predictive target-projected score vector by projecting on $\mathbf{X}$:

$$\mathbf{t}_{TP} = \mathbf{X}\mathbf{w}_{TP} = \mathbf{X}\mathbf{b}_{LVR}/||\mathbf{b}_{LVR}|| = \hat{\mathbf{y}}/||\mathbf{b}_{LVR}|| \tag{6}$$

Equation (6) shows that the target-projected score vector is proportional to the predicted response vector $\hat{\mathbf{y}}$. Thus, the TP algorithm produces the scores of the predictive latent variable with maximum covariance to the predicted y-variable.

The predictive target-projected loadings can subsequently be calculated as follows:

$$\mathbf{p}_{TP} = \mathbf{X}^T\mathbf{t}_{TP}/\left(\mathbf{t}_{TP}{}^T\mathbf{t}_{TP}\right) = \mathbf{X}^T\hat{\mathbf{y}}/\left(||\mathbf{b}_{LVR}||||\mathbf{t}_{TP}||^2\right) \tag{7}$$

Equation (7) shows that the target-projected loadings are proportional to the covariances between the exploratory x-variables and the predicted y-variable. The predictive part of the LVR model, ie, the TP model, can be written as

$$\mathbf{X} = \mathbf{X}_{TP} + \mathbf{E}_{TP} = \mathbf{t}_{TP}\mathbf{p}_{TP}{}^T + \mathbf{E}_{TP} \tag{8}$$

From Equation (8), we can calculate explained predictive variance $v_{\mathrm{expl},i}$ and residual variance $v_{\mathrm{res},i}$ for each x-variable i. The ratio of explained to residual variance of a variable after TP defines the selectivity ratio SR_i for each spectral variable:

$$SR_i = v_{expl,i}/v_{res,i} = \left\|\mathbf{t}_{TP}p_{TP,\mathrm{i}}{}^T\right\|^2/\left\|\mathbf{E}_{TP,\mathrm{i}}\right\|^2 \quad i = 1,\ 2,\ 3,\ \ldots \tag{9}$$

If the x-variables represent a spectral profile, the SRs can be displayed and interpreted in a similar manner. Just like a relatively high intensity in a spectrum implies high concentration, a relatively high value of SR means that the corresponding x-variable has a strong association to the predicted y-variable. Thus, the SRs can be used quantitatively to rank the x-variables′ associations to the predicted y-variable. By multiplying SR_i with the sign of the target-projected loading p_i for variable x_i, information about the direction of the correspondence between the variable $x_{\mathrm{TP},i}$ and the predicted y-variable can be visualized in an SR plot.

Tran et al[7] proposed an index called the sMC. This index is calculated from TP but without performing the orthogonalization step, Equation (7):

$$\mathbf{X} = \mathbf{X}_{sMC} + \mathbf{E}_{sMC} = \mathbf{t}_{TP}\mathbf{w}_{TP}{}^T + \mathbf{E}_{sMC} \tag{10}$$

From Equation (10), sMC is obtained analogously to SR as the ratio

$$sMC_i = \left\|\mathbf{t}_{TP}w_{TP,\mathrm{i}}{}^T\right\|^2/\left\|\mathbf{E}_{sMC,\mathrm{i}}\right\|^2 \quad i = 1,\ 2,\ 3,\ \ldots \tag{11}$$

Since the orthogonalization step is not performed, the residual matrix $\mathbf{E}_{\mathrm{sMC}}$ is not orthogonal to the predictive part of $\mathbf{X}$ but may approach orthogonality when $\mathbf{p}_{\mathrm{TP}} \approx \mathbf{w}_{\mathrm{TP}}$. Visualization of profiles of sMCs can be obtained similar to SR plots. Note that by using the regression vector as loadings in the decomposition described by Equation (10), sMC is expected to be strongly impacted by the regression vector, as opposed to SR that will be mostly impacted by the covariance between the x-variables and the predicted response. Thus, by simple matrix algebra, we can derive the following expression for SR:

$$SR_i = cov(x_i/\hat{y})^2/\left[var\ (x_i)^2\left(1 - cov\ (x_i,\hat{y})^2\right)\right] i = 1,\ 2,\ 3,\ \ldots \tag{12}$$

For standardized variables, Equation (12) simplifies to

$$SR_i = cor(x_i/\hat{y})^2/\left(1 - cor\ (x_i,\hat{y})^2\right) \qquad i = 1,\ 2,\ 3,\ \ldots \tag{13}$$

Statistical tests have been developed for selecting important variables for both SR[6] and sMC.[7] For SR, the discriminating variable (DIVA) plot was developed as an additional tool for biomarker (variable) selection by which the user can take into account different data characteristics and aims of applications.[6] The DIVA plot is possible when the y-variable is not a continuous variable but represents the group belonging of samples. The DIVA plot sorts the variables according to correct classification rate. This makes it possible to define an SR threshold for selection of biomarker candidates. For continuous y-variables, SR ranks the variables. Experience has shown that the threshold is application and data dependent and it has to be chosen by the user according to aim. Automated variable selection is possible by first ranking the variables from high to low SR and then either exclude or include variables by backward or forward selection; see, eg, Sinkov et al.[20] This approach guarantees inclusion also of variables with low association to the response variable, but still important for prediction, eg, when signal from interfering constituents is overlapping with signal from "true" biomarkers.

3 | EXPERIMENTAL

Three datasets are analyzed in this work. Details of sampling and work-up have been provided in previous publications,[4,12,21] so only brief descriptions are provided in this work.

3.1 | Dataset 1: Chromatographic profiling of extracts of herbs

Seventy-eight extracts of a Chinese herb were characterized using high-performance liquid chromatography, and their corresponding antioxidant activities were measured.[12] The retention time (RT) region 9.01 to 31.39 minutes with a data point resolution of 0.02 minute was selected for the chromatographic profiles that thus were represented by 1120 RTs. Background correction was performed using the asymmetric least squares method.[22] Alignment to minimize RT shift was executed by the method of Wong et al.[23]

3.2 | Dataset 2: Infrared profiling of three-component mixtures

Three chemical species, methylcyclohexane (MCH), *di*-butylether (DBE), and ethylbenzene (EB), were mixed according to a mixture design to prepare 34 mixtures that span the whole range of fractions between 0 and 1 for all constituents.[4] Infrared (IR) profiles were acquired, and the wave numbers from 2990 to 2780 cm^{-1} were selected as explanatory variables to represent the CH stretch region. With a data point resolution of 1 cm^{-1}, the selected range is described by a total of 211 variables. The fractions of the three chemical constituents were selected as response variables. No preprocessing of data was done.

3.3 | Dataset 3: Proton nuclear magnetic resonance profiling of lipoproteins in serum

One hundred forty-seven healthy volunteers, equally divided between women and men, were recruited among the inhabitants of a rural community in the Fjord region of Western Norway. Inclusion criteria were age 18 to 62 years and body mass index (BMI) 18.5 to 30.0. Exclusion criteria were pregnancy, smoking, drug abuse, use of lipid-lowering drugs and established cardiovascular disease (CVD), diabetes type 2, or cancer.

Blood samples were collected between 8 and 9 AM after overnight fasting. Serum was obtained according to a standardized protocol consisting of the following steps: (a) Blood plasma was collected in 5-mL tubes with gel (Vacuette Serum Gel with activator, G456073). (b) Tubes were carefully turned upside-down five times and placed vertically for coagulation. (c) After 30 minutes, the sample was centrifuged at 2000 *g* for 10 minutes. Serum was then visually inspected for residues, and centrifugation was repeated if residue was present. (d) The serum tube was kept in refrigerator at 4°C before pipetting 0.5 mL into cryo tubes. (e) The cryo tubes were then stored at −80°C until the analysis was performed.

Reference values for total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), and total triglyceride (TG) in serum were quantified for the samples by the standard method for analysis of blood samples. In the reference method, LDL-C is estimated from the measured TC, HDL-C, and TG concentrations by the Friedewald equation[24]

Proton nuclear magnetic resonance (NMR) spectroscopy using nuclear Overhauser effect spectroscopy (NOESY) was performed for the same samples according to a procedure described in previous work.[21] The NMR profiles were aligned to the lactate doublet at approximately 1.32 ppm, and the shift regions embracing the lipoprotein methylene peak, 1.30 to 1.19 ppm, the lipoprotein methyl peak, 0.90 to 0.78 ppm, and the peak located at 0.70 to 0.62 ppm were selected as explanatory variables. These regions are known to contain quantitative information about TC, LDL-C, HDL-C, and TG. The three regions provided 1383 spectral variables without any other pretreatment than normalization using standards as described in earlier work.[21] These NMR profiles were selected as explanatory variables to the regression modeling with TC, HDL-C, LDL-C, and TG concentrations determined by the standard method as response variables.

3.4 | Regression modeling

The relations between the explanatory variables, ie, the instrumental descriptors, and the response variables, ie, the reference values in the three datasets, were obtained by calculating a PLS regression model for each response variable separately. The predictive power of the models was optimized by using the significance test based on Monte Carlo resampling procedure developed by Kvalheim et al.[25,26] The resampling was performed with 100 repetitions with the samples split half and half between calibration and validation samples.

After removal of two outliers, PLS regression modeling of bioactivity from chromatographic profiles (dataset 1) gave a model with four PLS components for $P < .3$ and five PLS components for $P < .4$ using the significance test based on Monte Carlo resampling.[25,26] For MCH, DBE, and EB (dataset 2), PLS regression with the IR profiles as explanatory variables resulted in models for relative concentrations with two, three, and five components, respectively, with $P = .3$ for the significance testing. The models explained 99.8% to 99.9% of the relative concentrations for all three constituents. After removal of one outlier, the models using NMR profiles to predict TC, LDL-C, HDL-C, and TG gave the following results for number of PLS components: For both TC and LDL-C, the significance test gave eight components with $P < .4$ and seven with $P < .3$, explaining 97.8% and 97.6% of the reference TC values and 98.0% and 97.9% of the reference values for LDL-C. For HDL-C, 11 components were significant with $P < .4$ and eight with $P < .3$, explaining 96.0% and 94.6% of the reference HDL-C values. Finally, for TG, we obtained seven components with both $P < .3$ and $P < .4$ and explained variance 98.1%. It was decided to select all models on the basis of $P < .3$ since the additional explained variance in the reference values with increased model complexity was small for all four responses. From these models, SRs and sMCs were calculated and presented as SR and sMC plots. For the bioactivity data (dataset 1), SR and sMC were calculated for both the four- and five-component model in order to check for robustness towards minor overfitting or underfitting in the model selection step. Note that for both indices, all profiles were normalized so that the largest number in every profile is always 1. This was necessary to make visual comparison possible. We have also multiplied sMC for the explanatory variables by the sign of the corresponding loading to be able to simplify visual comparison of the two indices.

4 | RESULTS

4.1 | Dataset 1: Bioactivity of individual constituents in multicomponent extracts

Figure 1 shows SR and sMC for the antioxidant activity modeled from chromatographic profiles with four and five PLS components as input to TP.

SR reveals three regions in the chromatograms with high antioxidant activity with maxima at RT 12.53, 13.29, and 15.63 minutes. The correlations between bioactivity and the three maxima in the SR plot are .863, .797, and .778, respectively, indicating that constituents with strong antioxidant activity are present in these regions. On the other hand, Figure 1 shows that for sMC, only two of these constituents stand out. The peak at 15.63 minutes is weak in the

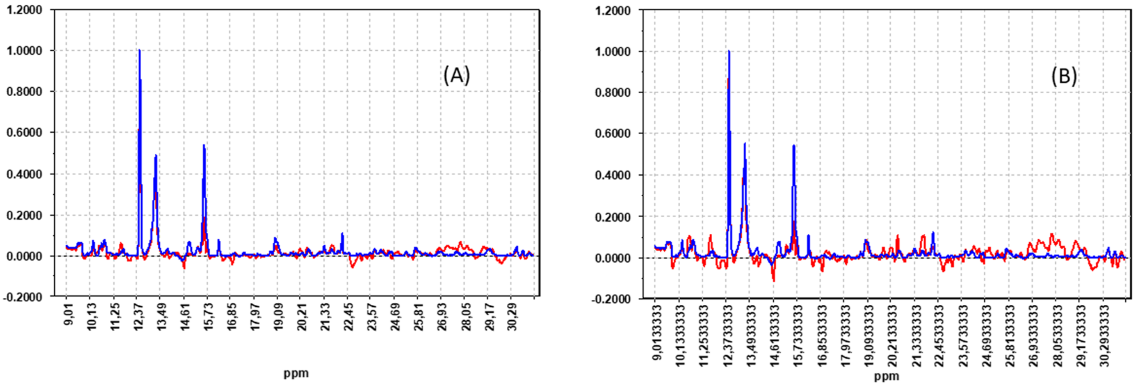

FIGURE 1 Selectivity ratio (blue) and significance multivariate correlation (red) for chromatographic profiles of bioactivity data. A, model with four partial least squares (PLS) components used as input to target projection. B, model with five PLS components used as input to target projection

sMC plots with both four and five PLS components even if the mean chromatographic size of this peak is the second largest of all peaks in the chromatographic profiles.

4.2 | Dataset 2: Relative concentrations of constituents from infrared profiles of mixtures

Figure 2 displays the SR and sMC profiles for the three models.

For MCH (Figure 2A), the two profiles are close to identical. The most important variable and the second most important variable for sMC are found at 2973 and 2019 cm^{-1}, respectively, and for SR at 2972 and 2019 cm^{-1}, respectively. The most important variable for both sMC and SR has a correlation of −.982 to the fraction of MCH implying that at this wavelength, the absorbance of MCH is small compared with the two other constituents in the mixtures. This observation shows that it is important to display the directions of the associations of the *x*-variables with the response. For the second most important variable, the correlation between IR absorption and fraction of MCH is .970 implying

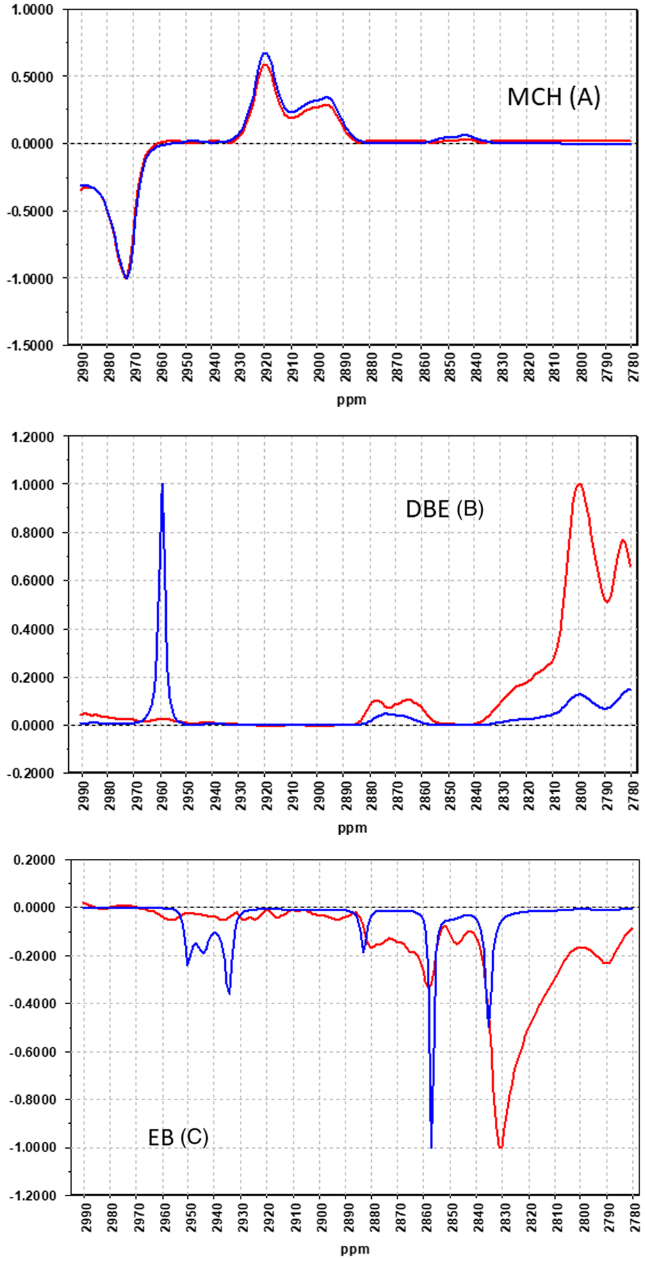

FIGURE 2 Selectivity ratio (blue) and significance multivariate correlation (red) for infrared profiles of mixture data. A, Methylcyclohexane. B, *di*-Buthylether. C, Ethylbenzene

that in this region, the absorption from MCH is dominating over the two other constituents that act as interferents with low absorbance in the region around 2019 cm^{-1}.

For DBE (Figure 2B), the picture is different for SR and sMC. The SR profile implies 2959 cm^{-1} as the most important variable and 2780 cm^{-1} as the second most important, while sMC selects 2799 cm^{-1} as the most important and 2781 cm^{-1} as the second most important. The correlations between the fraction of DBE and the absorbances at 2959 and 2799 cm are, respectively, .994 and .957. So for this case, sMC picks a wavelength that is considerably more impacted by absorbance from the two other constituents than the one selected by SR. The correlation at 2959 cm^{-1} is so high that this wavelength is almost selective for DBE. The two indices agree on the second most important wavelength that has a correlation to DBE of .962.

For EB (Figure 3C), SR reveals 2857 cm^{-1} as the most important and 2835 cm^{-1} as the second most important variable. The absorbances at these two wavelengths have correlations with EB of −.997 and −.995 implying that the absorbance of EB is small compared with the two other constituents at these wavelengths. sMC picks 2831 and 2858 cm^{-1} as the most and second most important variable. The corresponding absorbances have correlations of −.970 and −.991 to EB. Thus, the two indices produce different results, and for sMC, the selected most important variable has lower correlation to EB than the second most important. However, both indices find the most important variables at wavelength where the correlation of absorbances to fraction EB is large and negative. The negative correlation of total absorbance with fraction of EB shows that EB has lower absorbance than the two other constituents at 2857 cm^{-1} so an increase in fraction of EB is accompanied by a decrease in total absorbance. However, the correlation of −.997 implies that the wavelength found by SR is almost selective for EB. This strange observation is caused by 2857 cm^{-1} representing an isobestic point for the two interferents, MCH and DBE.[4] At an isobestic point, the absorbance does not change with changing fractions of two constituents, thus reducing the local chemical rank by one. Since closure in the three-component mixture further reduces the local chemical rank by one at this wavelength, the signal is selective for EB.

4.3 | Dataset 3: Concentrations of lipoproteins from NMR profiles

Figure 3 shows the variable importance profiles obtained by predicting TC, LDL-C, HDL-C, and TG from NMR.

For TC (Figure 3A), both SR and sMC detect the most important variables in the shift region corresponding to the cholesterol peak. However, the maximum is obtained at 0.846 ppm for SR and 0.841 ppm for sMC with corresponding correlations to TC reference values of .953 and .929, respectively. However, we observe a shoulder in the variable importance profile for sMC coinciding with the maximum in the SR profile. Both indices show a shoulder in the cholesterol

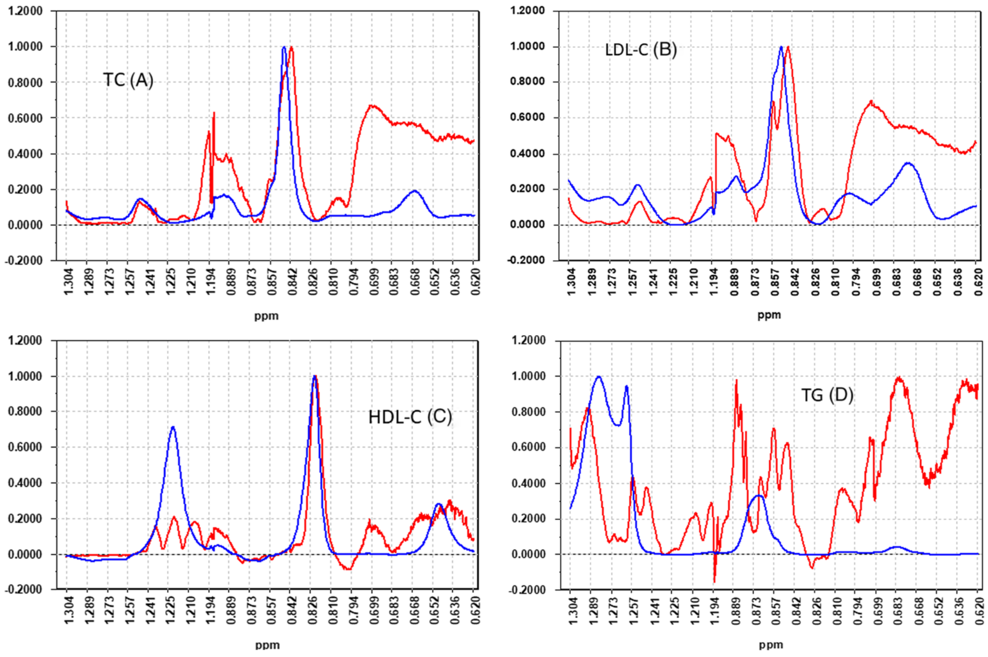

FIGURE 3 Selectivity ratio (blue) and significance multivariate correlation (red) for nuclear magnetic resonance spectroscopic profiles for lipoproteins. A, Total cholesterol. B, low-density lipoprotein cholesterol. C, high-density lipoprotein cholesterol. D, triglyceride

region at 0.857 ppm with a correlation to TC of .843. For SR, another maximum is found at 0.666 ppm with correlation to TC of .830. Two other regions are implied by sMC as important with maxima at 1.190 and 0.700 ppm with corresponding correlations to TC of .720 and .619, respectively. In summary, SR locates the most important variables more precisely than sMC.

Figure 3B shows the variable importance profiles for the two indices for LDL-C. The part of the profiles corresponding to the cholesterol region is similar to TC but shifted slightly to higher ppm, thus maxima at 0.850 ppm for SR and 0.855 ppm for sMC. The corresponding correlations with LDL-C are .894 and .855, respectively. We observe a shoulder for SR and a local maximum for sMC at 0.855 ppm. The correlation with LDL-C is .880. The third most important region is located at 0.672 ppm for SR and 0.780 ppm for sMC with corresponding correlations to LDL-C being .769 and .576, respectively. Thus, also for LDL-C, SR locates regions with higher correlations to the reference values more precisely than does sMC.

Figure 3C shows the variable importance profiles for HDL-C. Three regions with maxima at 0.823, 1.222, and 0.647 ppm, respectively, stand out for SR. The corresponding correlations with HDL-C reference values are .879, .845, and .711. The most important region for sMC coincides with the most important region obtained from SR. The maximum of the second most important region is located at 0.638 with correlation of .582. Figure 3C shows that sMC in this shift region has several maxima and the profile appears noisy compared with the SR profile. The third most important region is located in the same region as the second most important region according to SR. Maximum is almost coinciding with the maximum for SR, but several local maxima are observed for sMC at either side of the maxima of SR with almost the same importance according to sMC. Thus, also for HDL-C, SR provides the best selection of correlating regions to the reference values.

Figure 3D shows the SR and sMC profiles from NMR for TG. The SR profile implies two regions in the TG area with almost equal importance and with maximum at 1.282 ppm and 1.260. The third most important is located with maximum at 0.859. The correlations with TG reference values with regions in descending order of importance are .966, .965, and .840. For sMC, the picture is fuzzy. The profile indicates three regions of almost the same importance with maxima at 0.680, 0.628, and 0.886 ppm. The corresponding correlations with TG reference values are .663, .225, and .681. Again, the results are in favor of SR. We observe many local maxima around the regions selected by sMC that may indicate a noisy behaviour of the index.

The use of the regression vector as loadings in the decomposition of the exploratory data, Equation (10), matrix suggests that sMC may be mostly impacted by the regression vector and is therefore highlighting more or less the same variables as the regression vector itself.

Figure 4 compares the normalized regression vector with sMC for the four lipoproteins, and we see that the corresponding profiles are similar. The sign of the regression coefficients differs in some regions, but the pattern of absolute values is similar for regression coefficients and sMC. This observation may explain why instrumental regions with high correlations to the response variable often are lost when using sMC.

5 | DISCUSSION

The overall picture for the three datasets is that SR points to the explanatory variables with highest positive or negative correlations to the response variable. This result is as expected since the SRs, Equation (9), use the loadings from Equation (7) to estimate the explanatory data matrix. Equation (7) shows that these loadings are proportional to the covariances between the predicted values of the response and the *x*-variables. The calculation of sMC, Equation (11), is based on the target-projected weights that are the normalized regression coefficients, Equation (5). It has been known for a long time that the regression coefficients are not very useful for interpretation and variable selection in multicollinear data. Thus, the regression coefficients are representing the cumulative effects of many factors, also the presence of interferences (orthogonal variation), and thus mix predictive with orthogonal variance. This mixing may not only confuse the interpretation but also result in the loss of the most important variables from an interpretative point of view. This is exemplified by the loss of the third most bioactive component in dataset 1 and loss of the almost selective wavelength for DBE in dataset 2. Another drawback is that sMC does not show the direction of an association. Dataset 2 discussed above show that it is crucial to know the directions of associations.

Although SR has better properties than sMC for the purpose of model interpretation, SR will often not be able alone to produce a reduced model with better performance than the model embracing all the original explanatory variables. The reason for this is that SR represents a covariance-based approach: It is designed to find explanatory variables that possess a large degree of covariance to the predicted response variable in a validated regression model. For many

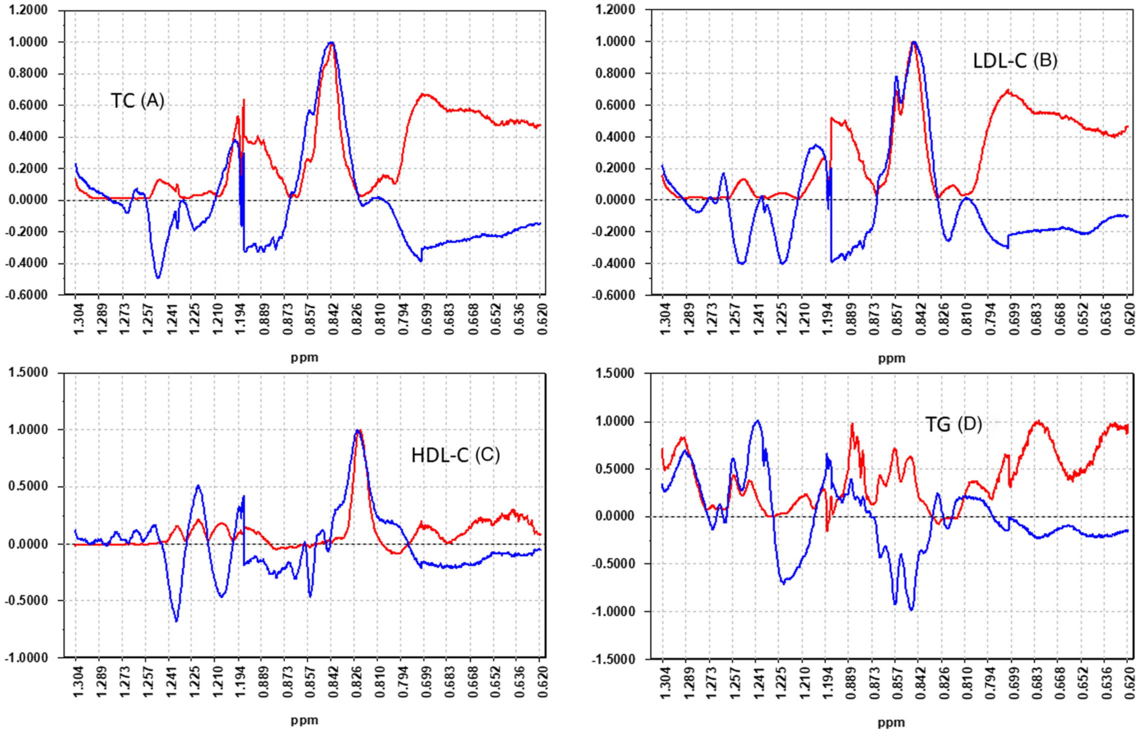

FIGURE 4 Significance multivariate correlation (red) and regression coefficients (blue) for nuclear magnetic resonance spectroscopic profiles for lipoproteins. A, Total cholesterol. B, low-density lipoprotein cholesterol. C, high-density lipoprotein cholesterol. D, triglyceride

applications, this may the only goal, for instance, when searching for biomarkers indicating bioactivity or disease in metabolomics applications. However, explanatory variables describing systematic orthogonal variation may be important for prediction performance of a model.[4] This happens, for instance, when interferences, ie, orthogonal variation, are present and overlap with regions with signal from variables that are predictive for the response. This is why we advocate to use both the SR profile and the profile of regression coefficients when doing manual variable selection to simplify models.[4] Regions with high regression coefficients can then be added to the description selected by SR. Methods that are not able to discriminate between predictive and orthogonal variance confuse the interpretation. If automated selection is the goal, then a ranking of the importance of explanatory variables can be done by SR and then, subsequently, variable selection can be done by stepwise backward or forward selection.[20]

6 | CONCLUSION

Our comparison of the two indices SR and sMC for describing variable importance has uncovered that sMC does not highlight the most important variables for interpreting models. Interpretation is crucial for understanding models and to be able to refine models in a sensible way. For some applications, such as when the goal is to reveal biomarkers, interpretation may be the most, and sometimes, only requested outcome of the analysis. The splitting of variance into a predictive part and orthogonal systematic part makes this possible. Mixing predictive and orthogonal contributions to a regression model may confuse more than enlighten an application.

ORCID

Olav M. Kvalheim https://orcid.org/0000-0001-9432-8776

REFERENCES

1. Wold S, Sjöström M, Eriksson L. Chemometr. *Intell Lab Syst*. 2001;58:109-130.
2. Kvalheim OM. 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 Chemometr* 2012; 26: 210-217.

3. Stocchero M. Iterative deflation algorithm, eigenvalue equations, and PLS2. *J Chemometr*. 2019. in press
4. Kvalheim OM, Arneberg R, Bleie O, Rajalahti T, Smilde AK, Westerhuis JA. Variable importance in latent variable regression models. *J Chemometr*. 2014;28:615-622.
5. Rajalahti T, Arneberg R, Berven FS, Myhr K-M, Ulvik RJ, Kvalheim OM. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemom Intel Lab Syst*. 2009;95:35-48.
6. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr K-M, Kvalheim OM. Discriminating variables test and selectivity ratio plot—quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Anal Chem*. 2009;81(7):2581-2590.
7. Tran TN, Afanador NL, Buydens LMC, Blanchet L. Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC). *Chemom Intel Lab Syst*. 2014;138:153-160.
8. Galindo-Prietoa B, Eriksson L, Trygg J. Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *J Chemometr*. 2014;28:623-632.
9. Andersen CM, Bro R. Variable selection in regression—a tutorial, J. *Chem*. 2010;24:728-737.
10. Farrésa M, Platikanova S, Tsakovskib S, Tauler R. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. *Chem*. 2015;29:528-536.
11. Andries JPM, Vander Heyden Y, Buydens LMC. Predictive-property-ranked variable reduction in partial least squares modelling with final complexity adapted models: comparison of properties for ranking. *Anal Chim Acta*. 2013;760:34-45.
12. Chau F-T, Chan H-Y, Xu CC-Y, C-J LY, Kvalheim OM. Recipe for uncovering the bioactive components in herbal medicine. *Anal Chem*. 2009;81(17):7217-7225.
13. Kellogg JJ, Todd DA, Egan JM, et al. Biochemometrics for natural products research: comparison of data analysis approaches and application to identification of bioactive compounds. *J Nat Prod*. 2016;79:376-386.
14. Britton E, Kellogg JJ, Kvalheim OM, Cech NB. Biochemometrics to identify synergists and additives from botanical medicines: a case study with *Hydrastis canadensis*. *J Nat Prod*. 2018;81(3):484-493.
15. Caesar LK, Kellogg JJ, Kvalheim OM, Cech NB. Opportunities and limitations for untargeted mass spectrometry metabolomics to identify biologically active constituents in complex natural product mixtures. *J Nat Prod*. 2019;82(3):469-484.
16. Rajalahti T, Kroksveen AC, Arneberg R, et al. A multivariate approach to reveal biomarker signatures for disease classification: application to mass spectral profiles of cerebrospinal fluid from patients with multiple sclerosis. *J Proteome Res*. 2010;9(7):3608-3620.
17. Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. *J Chemometr*. 2014;28:213-225.
18. Kvalheim OM, Karstang TV. Interpretation of latent-variable regression models. *Chemometrics and Int Lab Syst*. 1989;7:39-51.
19. Kvalheim OM. Latent-variable regression models with higher-order terms: an alternative to response modelling by factorial design and multiple linear regression. *Chemometrics and Int Lab Syst*. 1990;8:59-67.
20. Sinkov NA, Sandercock PM, Harynuk JJ. *Chemometric Classification of Casework Arson Samples Based on Gasoline Content, Forensic Science International*. 2014;235:24-31.
21. Jones PR, Rajalahti T, Resaland GK, et al. Associations of physical activity and sedentary time with lipoprotein subclasses in Norwegian schoolchildren: the active smarter kids (ASK) study. *Atherosclerosis*. 2019;288:186-193.
22. Boelens HFM, Dijkstra RJ, Eilers PHC, Fitzpatrick F, Westerhuis JA. New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. *J Chromatogr A*. 2004;1057:21-30.
23. Wong JWH, Durante C, Cartwright HM. Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal Chem*. 2005;77(17):5655-5661.
24. Friedewald WT, Levy RL, Fredrickson DS. Estimation of the concentration of low- density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem*. 1972;18:499-502.
25. Kvalheim OM, Arneberg R, Grung B, Rajalahti T. 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 Chemometr*. 2018;32:e2993.
26. Kvalheim OM, Grung B, Rajalahti T. Number of components and prediction error in partial least squares regression determined by Monte Carlo resampling strategies. *Chemom Intel Lab Syst*. 2019;188:79-86.

How to cite this article: Kvalheim OM. Variable importance: Comparison of selectivity ratio and significance multivariate correlation for interpretation of latent-variable regression models. *Journal of Chemometrics*. 2020;34:e3211. https://doi.org/10.1002/cem.3211