Analytica Chimica Acta 869 (2015) 21–33

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Sum of ranking differences (SRD) to ensemble multivariate calibration model merits for tuning parameter selection and comparing calibration methods

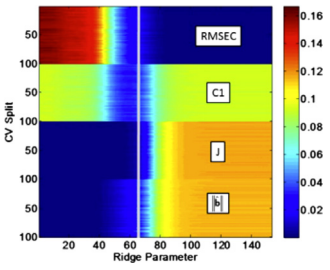
John H. Kalivas [a,*], Károly Héberger [b], Erik Andries [c,d]

[a] Department of Chemistry, Idaho State University, Pocatello, ID 83209, USA
[b] Research Centre for Natural Sciences, Hungarian Academy of Sciences, Pusztaszeri út 59-67, 1025 Budapest, Hungary
[c] Center for Advanced Research Computing, University of New Mexico, Albuquerque, NM 87106, USA
[d] Department of Mathematics, Central New Mexico Community College, Albuquerque, NM 87106, USA

HIGHLIGHTS

- Sum of ranking differences (SRD) used for tuning parameter selection based on fusion of multicriteria.
- No weighting scheme is needed for the multicriteria.
- SRD allows automatic selection of one model or a collection of models if so desired.
- SRD allows simultaneous comparison of different calibration methods with tuning parameter selection.
- New MATLAB programs are described and made available.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 February 2014
Received in revised form 16 October 2014
Accepted 9 December 2014
Available online 7 February 2015

Keywords:
Sum of ranking differences
Multivariate calibration
Partial least squares
Ridge regression
Model comparison

ABSTRACT

Most multivariate calibration methods require selection of tuning parameters, such as partial least squares (PLS) or the Tikhonov regularization variant ridge regression (RR). Tuning parameter values determine the direction and magnitude of respective model vectors thereby setting the resultant predication abilities of the model vectors. Simultaneously, tuning parameter values establish the corresponding bias/variance and the underlying selectivity/sensitivity tradeoffs. Selection of the final tuning parameter is often accomplished through some form of cross-validation and the resultant root mean square error of cross-validation (RMSECV) values are evaluated. However, selection of a "good" tuning parameter with this one model evaluation merit is almost impossible. Including additional model merits assists tuning parameter selection to provide better balanced models as well as allowing for a reasonable comparison between calibration methods. Using multiple merits requires decisions to be made on how to combine and weight the merits into an information criterion. An abundance of options are possible. Presented in this paper is the sum of ranking differences (SRD) to ensemble a collection of model evaluation merits varying across tuning parameters. It is shown that the SRD consensus ranking of model tuning parameters allows automatic selection of the final model, or a collection of models if so desired. Essentially, the user's preference for the degree of balance between bias and variance ultimately decides the merits used in SRD and hence, the tuning parameter values ranked lowest by SRD for automatic selection. The SRD process is also shown to allow simultaneous comparison of different

* Corresponding author. Tel.: +1 208 282 2726; fax: +1 208 282 4373.
E-mail address: kalijohn@isu.edu (J.H. Kalivas).

http://dx.doi.org/10.1016/j.aca.2014.12.056
0003-2670/

calibration methods for a particular data set in conjunction with tuning parameter selection. Because SRD evaluates consistency across multiple merits, decisions on how to combine and weight merits are avoided. To demonstrate the utility of SRD, a near infrared spectral data set and a quantitative structure activity relationship (QSAR) data set are evaluated using PLS and RR.

1. Introduction

Multivariate calibration for quantitative purposes is becoming ever more important in diverse fields such as on-line process monitoring for product yield and quality, medical diagnostics, the pharmaceutical industry, and agriculture and environmental monitoring just to name a few. Many of the multivariate calibration processes, such as partial least squares (PLS) or the Tikhonov regularization (TR) variant known as ridge regression (RR) require selection of appropriate respective tuning parameter (meta-parameter) values [1–3]. Specifically, a model vector must be selected from a set of tuned models developed by a particular calibration method. The number of model vectors generated depends on the number of tuning parameter values for the respective method. For PLS, the number of potential models is the number of latent variables (LVs) determined by the data pseudo-rank. The number of ridge parameters (number of RR models), is essentially unlimited since the ridge parameter is continuously varied.

Using one of several cross-validation (CV) processes [4–8], the final model vector (tuning parameter) is typically chosen to predict with "acceptable" accuracy (low bias) based on the one model merit root mean square error of CV (RMSECV) [1,2]. However, when RMSECV values are plotted against the tuning parameter value, the plot can resemble a RMSE of calibration (RMSEC) plot and thus, choosing a tuning parameter value on this one model merit is then not obvious [9]. One of the data sets evaluated in this paper has such a difficulty. Other single model merits have been developed and compared for model selection [10–19].

A primary consideration in choosing a suitable tuning parameter value is obtaining a model not under- or over-fitted (good predictability in conjunction with proper model complexity also known as the bias/variance tradeoff). In this case, bias is the degree of prediction accuracy obtained from a model and variance is related to the extent of uncertainty in the prediction [20–23]. Methods such as RR and PLS are biased methods and hence a tradeoff in the degree of under- and over-fitting is mandatory to form a model with an "acceptable" bias/variance balance [3,21–23]. Models with acceptable bias/variance tradeoffs were recently shown to also balance the intrinsic model selectivity and sensitivity [23]. Selectivity is a measure of the level of unique analyte information in measurements, e.g., spectra, and is often identified with the net analyte signal (NAS) [13]. Sensitivity refers to the degree of change in signal relative to a change in the quantity of analyte, e.g., in analytical chemistry, a system is sensitive if a small change in analyte concentration generates a large change in signal [13,17]. It follows then that at least two model merits, each trending in opposite directions, should be simultaneously evaluated in order to characterize the balance between under- and over-fitting [20,21,24–28].

Different tactics have been used to combine two model merits. One is a graphical approach forming L-curves by plotting RMSEC (or RMSECV) against a model complexity or variance measure with the better models residing in the corner region of the resultant L shaped curve [3,20,21,24–26,29]. The RMSEC (or RMSECV) values have been scaled and combined with scaled model complexity values or variance measures to convert L-curves to U-curves allowing automatic model selection [20,28]. Different combinations of RMSEC with RMSECV values have been plotted against model complexity or variance measures to form other U-curves [23]. Variations are possible by combining respective R^2 values slopes, or intercepts from plotting model predicted values against reference values. While these most recent approaches have expanded the number of model merits simultaneously evaluated, there are many more model merits that can participate in the tuning parameter selection process [13–19]. The difficult part in using a collection of model merits is how to actually combine them. Multicriteria desirability functions are possible but these require tuning in themselves [30]. Recent work developed a two-step sequential process to first select the number of latent variables for each data preprocessing method, and then from these, select the best preprocessing method [31]. Several model merits have been used in a consensus approach, but again, empirical data set dependent merit threshold values were needed [32]. Essentially, the user's preference for the degree of balance between bias and variance ultimately decides the merits used (and potential weights) in any multicriteria process and hence, the tuning parameter values deemed best.

This paper shows that the sum of ranking differences (SRD) [33–37] is a simple objective process to ensemble multiple model merits for ranking models (tuning parameters) allowing automatic selection of a consensus model or set of models. When CV is used to generate model merits, then SRD allows the models merits computed on each data split to be evaluated, not just the mean values as in the standard CV process to select a tuning parameter. Because SRD evaluates consistency across multiple merits, decisions on how to combine and weight merits are avoided. If desired for a specific data set, the flexibly of the SRD process allows concurrent comparisons of modeling methods in combination with tuning parameter selection. Only a few of the possible model merit combinations with SRD are studied in this paper and only model vectors estimated by PLS and RR are compared. As noted above for any tuning parameter selection processes, it is further verified in this paper that the user's preference and choice of model merit(s) used can affect the tuning parameter value selected.

The current versions of SRD are in Excel [38] and have data size limitations due to constraints imposed by Excel and other restrictions on the input SRD matrix exist. Developed for this paper is MATLAB code removing these restrictions [39]. The new algorithm attributes are described in the section overviewing SRD. Before overviewing SRD, the calibration methods and model merits used are briefly described.

2. Calibration processes

The multivariate calibration model for this paper is expressed by

$$\boldsymbol{y} = X\boldsymbol{b} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ specifies the $m \times 1$ vector of quantitative values of the property to be predicted for m calibration samples, $\boldsymbol{X}$ symbolizes the $m \times p$ calibration matrix of p predictor variables, and $\boldsymbol{b}$ represents the $p \times 1$ vector of calibration model coefficients to be estimated. The $m \times 1$ vector $\boldsymbol{e}$ denotes normally distributed errors with mean zero and covariance matrix $\sigma^2\boldsymbol{I}$. The relationship in Eq. (1) is common to many disciplines. However, the prediction property and

predictor variables are quite varied across respective disciplines. A frequent situation in spectroscopic analysis is where $\boldsymbol{y}$ contains analyte concentrations and the measured p variables are wavelengths. Usually $m \ll p$ with spectroscopic data and hence, methods such as PLS or RR are needed. If $m \geq p$, then multiple linear regression (MLR) can also be used. There are many other methods of modeling processes, but only PLS and RR are evaluated here.

Extensive explanations of PLS and RR are available [1–3] and only key minimization expressions are shown emphasizing respective tuning parameters. Tuning parameter values establish the bias/variance tradeoff and the corresponding model selectivity/sensitivity balance [23]. For least squares, there is no tradeoff (unless variable selection is involved) and the minimization is expressed as determining a $\boldsymbol{b}\,(\hat{\boldsymbol{b}})$ such that $\min(\| \boldsymbol{y} - \mathbf{X}\boldsymbol{b} \|^2)$ is satisfied where the double brackets $\| \cdot \|$ indicate the L_2 norm (vector 2-norm or Euclidian norm) that defines the model vector magnitude. The methods of PLS and RR minimize related expressions.

2.1. PLS

The PLS approach to regression can be expressed as the minimization of $(\| \boldsymbol{y} - \mathbf{X}\boldsymbol{b} \|^2)$ subject to the constraint $\boldsymbol{b} \in K_d\left(\boldsymbol{X}^T X, X^T y\right)$ where $K_d\left(X^T X, X^T y\right) = \text{span}\left(X^T y, X^T X X^T y, \ldots, \left(X^T X\right)^{d-1} X^T y\right)$ is the span of the Krylov subspace based on d PLS basis vectors (latent variables (LVs)) and the superscript T indicates the matrix algebra transpose operation. In the process of forming the model vector, it has been shown that the magnitude of the estimated model vector, expressed as $\| \hat{\mathbf{b}} \|$, increases as more PLS LVs are used, i.e., the model complexity or effective rank increases [40–42]. Another measure recently studied to characterize model complexity is the jaggedness of the model vector [28] defined by

$$J_i = \sqrt{\sum_{j=2}^{p} \left(\hat{b}_{ij} - \hat{b}_{ij-1}\right)^2} \tag{2}$$

Jaggedness is also computed for the ith model in this paper. The number of PLS LVs is the tuning parameter that regulates the model vector direction and size and the underlying tradeoffs.

2.2. RR

The minimization expression for the TR variant RR [24,43–45] is $\min(\| \boldsymbol{y} - \mathbf{X}\boldsymbol{b} \|^2 + \eta^2 \| \boldsymbol{b} \|^2)$ where η symbolizes the regularization tuning parameter controlling the penalty given to the second term and is in the range $0 \leq \eta < \infty$. The value of η regulates the model vector direction and size of the corresponding estimated model vector. The greater the value, the smaller $\| \hat{\mathbf{b}} \|$ is. Other modifications of TR have been recently reviewed [45].

3. Model prediction and model evaluation (selection) merits

With an estimate of $\boldsymbol{b}\,(\hat{\mathbf{b}})$, the amount of the calibrated property present in a new measured $p \times 1$ sample vector $\boldsymbol{x}$ is predicted by $\hat{y} = \boldsymbol{x}^T \hat{\mathbf{b}}$. Thus, the degree of accuracy of the predicted value depends on the magnitude and direction of the estimated model vector which are determined by the tuning parameter. Because actual reference values of new samples are not known, model merits relative to the calibration samples are evaluated as proxies to assist in selecting respective model tuning parameters to hopefully ensure acceptable predictions of new samples.

The L-curve for selecting tuning parameters [3,20,21,24–27,29] can be formed by plotting mean RMSEC or RMSECV against a model variance or complexity measure. Models in the corner region of the L-curve represent acceptable compromises for the bias/variance tradeoff, i.e., least risk of over- and under-fitting. These models have been found to correspond to the underlying model selectivity/sensitivity balance. Studied in this paper is using SRD to rank models based on model tradeoffs characterized by the CV split-wise values of RMSEC, RMSECV, $\| \hat{\mathbf{b}} \|$, J, and others.

As noted in Section 1, approaches have been developed to remove the potential ambiguity in determining the corner region of an L-curve by forming U-curves with the best tuning parameter value at the minimum allowing automatic tuning parameter selection [20,23,28]. Two specific merits to be evaluated with SRD in this study are

$$C1_i = \left(\frac{\| \hat{\mathbf{b}}_i \| - \| \hat{\mathbf{b}} \|_{\min}}{\| \hat{\mathbf{b}} \|_{\max} - \| \hat{\mathbf{b}} \|_{\min}}\right) + \left(\frac{\text{RMSEC}_i - \text{RMSEC}_{\min}}{\text{RMSEC}_{\max} - \text{RMSEC}_{\min}}\right) \tag{3}$$

and

$$C2_i = \frac{\text{RMSEC}_i + \text{RMSECV}_i}{\text{RMSEC}_i / \text{RMSECV}_i} \tag{4}$$

where values in C1 for the ith model are range scaled from 0 to 1. The RMSECV values can be substituted for RMSEC in C1 as can J be substituted for $\| \hat{\mathbf{b}} \|$. Unless noted otherwise, C1 expressed by Eq. (3) is used with SRD. The goal with C2 is to minimize the numerator and maximize the denominator to favor the CV merit. In this way, the calibration and validation samples are predicted similarly with a bias toward predicting validation samples with a smaller error. Respective R^2 values obtained by plotting predicted calibration values ($\hat{y}_{\text{cal}}$) or the CV predicted sample values such as $\left[\left(1 - R^2_{\text{cal}}\right) + \left(1 - R^2_{\text{cv}}\right)\right]$/(RMSEC/RMSECV) are possible. Unless noted otherwise, C2 is used with SRD as written in Eq. (4).

Various other merits have been proposed and evaluated to select model tuning parameters when the merit values are used univariately. For example, Mallow's C_p criterion [46], generalized CV (GCV) [47], AIC [48], BIC [49], trace $(\boldsymbol{X}^T\boldsymbol{X})^+$ [21], and others [12,18,19]. These merits were not used in this paper, but their usages with SRD are also feasible. Instead, SRD rankings are reported using the CV split-wise combinations of RMSEC, RMSECV, respective R^2, slopes, and intercepts, $\| \hat{\mathbf{b}} \|$, J, C1, and C2. For comparison, SRD rankings are presented from just using the RMSECV model merit. The mean L- and U-curves are also plotted for comparison to SRD rankings.

4. SRD

The SRD algorithm is a simple, powerful, general process to determine similarities between variables by ranking the variables (columns of the SRD input matrix) across objects (rows of the SRD input matrix) relative to respective object reference (target) ranking values [33–37]. The method is well described in the literature and hence, only briefly outlined here. First a brief statement is given on using SRD for tuning parameter selection. This statement is followed by a brief but detailed description of the general SRD process. This section concludes with further details on how SRD will be used for tuning parameter selection.

As a simple example of SRD being used for tuning parameter selection, a CV process can be used to produce the number of rows (n) of the SRD input matrix with RMSECV values enumerated in the rows and variables (columns) are, in the case of PLS, number of LVs.

Similarly, RR ridge parameters can be placed as column headings for the input matrix and the rows are the particular CV splits.

With SRD, target reference values are required for the each object and these can be the minimum, maximum, median, or mean of respective rows or known reference values can be used. For each row (object) of the input SRD matrix, the value closest to the corresponding row target is identified. A target vector is created with these values sorted (ranked) from low to high and the respective row indexes are noted. The SRD input matrix is rearranged to this target row index sort and all values in each respective column (variable) are ranked from low to high. The absolute value of the difference between the target row ranking and each column ranking of the reordered rows is computed and summed for each column to form the column-wise vector of the final SRD ranked columns. The closer an SRD value is to zero, the closer the ranking of that column (variable) to the row (object) targets, and the better the variable is for that particular SRD evaluation. The proximity of SRD rank values shows which variables are similar. Groupings of variables can also be observed. The SRD rankings can also be considered dissimilarity assessments with the greater the SRD rank value, the more dissimilar the variable is to the object targets. Recently, SRD has been related to the inversion number [50] and SRD has been advanced to handle observations with ties [37].

A process has been established to validate the SRD ranking results. The validation involves determining if the SRD rankings are no different than random rankings [34]. The process is named the comparison of ranks by random numbers (CRRN). For CRRN, distributions are generated for random numbers and are used to evaluate how far the SRD ranked values are from being ranked randomly. Random numbers are used for a small number of objects (less than 13, or 9 if ties are present) and the normal distribution is used as the approximate for a large number of objects (13 or greater). The CRRN process is not the validation focus in this paper and the reader is referred to Ref. [34] for the details of CRRN.

Instead of CRRN, and as originally developed and available in the Excel SRD version [38], a CV process of the input SRD matrix can also be used with the SRD algorithm to further validate results. With the Excel version a 7-fold CV is used on the SRD input matrix to estimate uncertainties in the SRD rankings of the variables. In this situation, one-seventh of the objects are left out and the SRD algorithm is run on the remaining six-seventh of the objects to obtain the SRD rank values. The process is repeated seven times and the variation of the SRD rankings across the folds can be evaluated by assigning uncertainties to the individual SRD ranks and by using a boxplot to visualize. With the CV of the SRD input matrix, the Wilcoxon matched pair or sign tests [51] can be used to provide statistical significance between SRD rankings. While both validation process are evaluated in this paper, graphical results are primarily presented using CV on the SRD input matrix, i.e., boxplots are mostly shown.

Typically, object measures (model merits for this paper) being used in the SRD input matrix are not measured on the same scale. For SRD to function correctly, SRD input values must be scaled to have similar magnitudes. Numerous scaling approaches are possible such as range scaling inclusively between 0 and 1, autoscaling (or standardization) to mean 0 and standard deviation 1, and others [36,52]. Normalizing each row (vector) of the SRD input matrix to unit length is used in this study.

The SRD process has been useful in a large number of varied situations [35,36 and references therein]. For example, in one study, SRD was used to compare the rankings of two different methods for rapidly screening the comprehensive two-dimensional liquid chromatographic analysis of wine [53]. Different data sets were used for the comparison. In other recent studies, SRD was used to compare rankings of sensory models relative to panel scores [54,55], different curve resolution and classification methods were compared using a variety of performance merits [56,57]. Lastly, among the diverse applications, SRD has been used to compare several modeling methods to compare and form quantitative structure activity relationship (QSAR) models [34,58].

Other recent works investigating processes to combine rankings of variables based on a set of measured objects have recently been published [59,60]. In these studies, the focus is ranking molecules in a data base to a user defined target reference structure. The rankings are based on multiple intermolecular structural similarity measures. Specifically, a matrix of similarity values is formed where the columns (variables) are the molecules and rows (objects) are the similarity measures. For each row similarity measure, the columns are numerically ranked from 1 to the number of columns relative to the magnitude of that particular similarity measure. A rank of 1 is for the column molecule most similar to the target reference structure. The ranks in each column are summed and the columns are sorted to the respective rank sums. The lower the rank sum, the more similar the column molecule is to the sought reference structure. Other combinations of the ranked matrix besides the sum were studied. The method is applicable to tuning parameter selection and other areas where a subset variables need to be selected from a collection of variables. This approach can be considered unsupervised while the SRD process is supervised (a target vector is used). The SRD approach could also be used with molecular matching studies.

4.1. New SRD features with the MATLAB code

At the time of this writing, there are Excel versions to perform SRD with CRRN, SRD with 7-fold CV, and SRD to handle ties. In all cases, the number of objects for the SRD input matrix has been tested to 1400 and the number of possible variables is 250. These

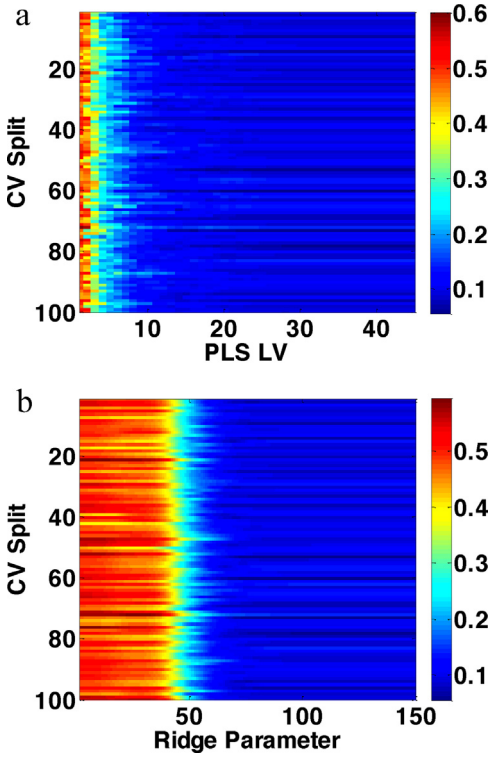


Fig. 1. Corn data images of (a) PLS and (b) RR CV split-wise RMSECV values for the 100 LMOCVs and respective tuning parameters. Ridge values range from 68 at ridge parameter 1 to 6.7×10^{-7} at ridge parameter 150.

Excel versions with sample input and output files are available for downloading [38]. The Excel versions require the same target values for each object.

For this work, MATLAB code was developed to work in the same format as the Excel versions as well as additional formats, albeit there is no MATLAB version of the Excel SRD developed for ties [33]. With MATLAB, the only limitation to the size of the SRD input matrix is the memory available on the computer performing the SRD computations. The MATLAB code including a demo is available for downloading [39].

The MATLAB code allows for multiple blocks of model merits. For example, an SRD input matrix can be composed of a block of RMSECV rows with each row being the corresponding CV split of RMSECV values and another block of rows with the corresponding CV split-wise model R^2 values. The target reference values for the RMSECV block would be row minima and target reference values of row maxima for the R^2 block. Regardless, all values in model merit blocks need to be scaled to similar magnitudes (or rank transformed) prior to analysis by SRD. The MATLAB code is flexible to allow SRD computations based on single object rows (considered one block and the only block for the SRD input matrix) or blocks of separate objects with equal or unequal number of rows in each block.

For validation of the SRD rankings, a similar CRRN process applied in the Excel versions is used in the MATLAB code. For CV of the SRD input matrix, the MATLAB code allows the option of using n-fold CV or leave multiple out CV (LMOCV) processes to obtain a boxplot as previously described [34] for the Excel SRD version. With n-fold CV, the user specifies a value for n and this value is used for each block of model merit CV values in the SRD input matrix. For LMOCV, the user specifies the percent to be randomly left out of each model merit block of CV values and how many times each block is to be split. As noted in Section 4, if the SRD input matrix is based on only single object rows, then the SRD input matrix is

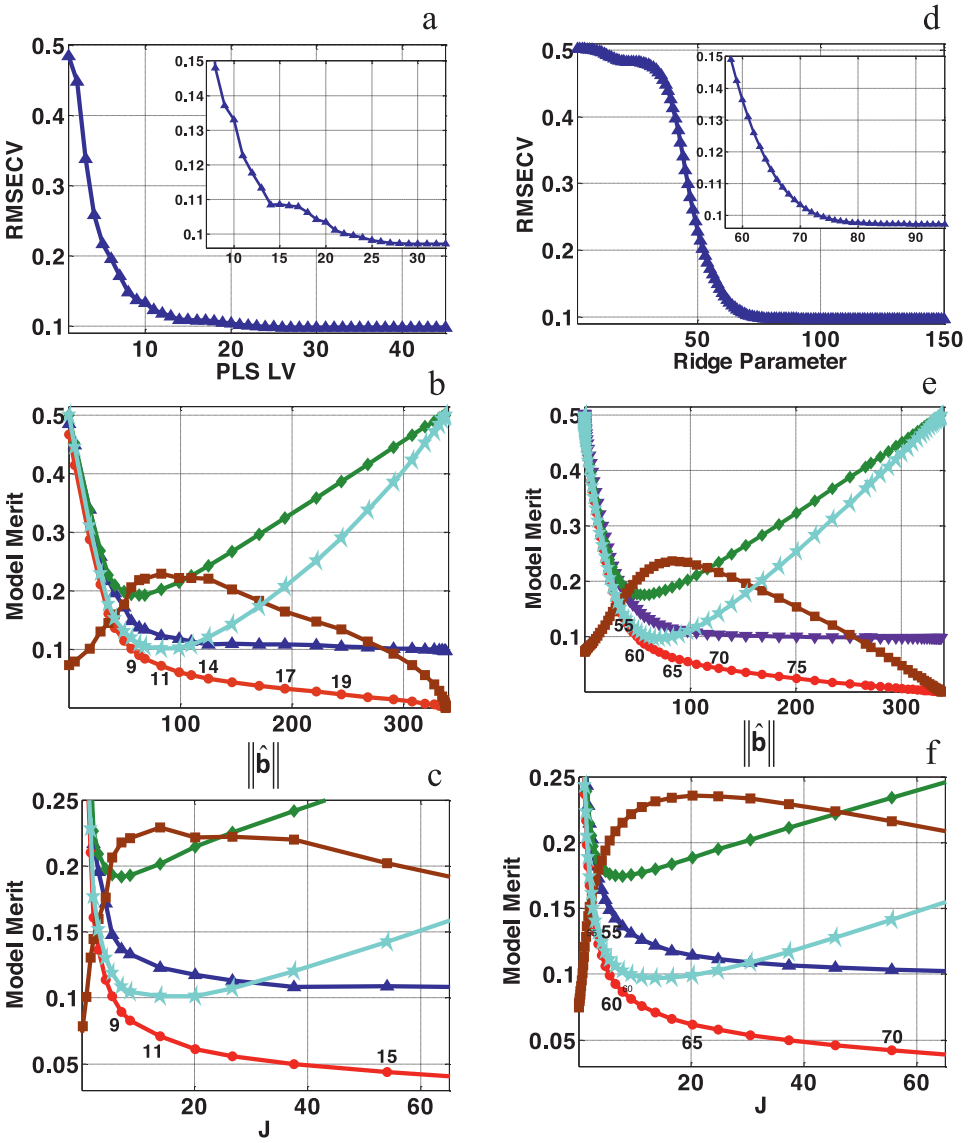


Fig. 2. Mean corn model merit graphics for PLS plotting (a) RMESCV against LVs and (b) and (c) are model merit values plotted against the model L_2 norm and J values, respectively. For both (b) and (c), RMSECV (blue triangles), RMSEC (red circles), C1 (green diamonds), C1 with J replacing the L_2 norm (cyan stars), and C2 inverted (brown squares). Values plotted in (b) and (c) are scaled to fit in the plots. Numbers in PLS plots correspond to number of LVs. Also shown are the corresponding mean RR model merit graphics for (d) RMESCV against ridge parameters, (e) merits plotted against the model L_2 norm values and (f), against the J values. Numbers in the RR plots correspond to ridge parameter number. Ridge values range from 68 at ridge parameter 1 to 6.7×10^{-7} at ridge parameter 150 in (d) and the same range trends are shown from left to right, respectively, in (e), and (f). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

considered one block for the SRD CV purpose to obtain the boxplot. In this case, all SRD input values in each row need to transformed to one common target value such as minimization.

4.2. SRD setup for tuning parameter selection and comparisons of modeling methods

The SRD input matrix is objects by variables and it is best to have at least seven rows to avoid a random ranking of the variables. In order to build up the number of rows, a CV process is used in this study. For example, if the goal is to select the number of PLS LVs using *n*-fold CV to form RMSECV values, the SRD input matrix would then be *n* by number of PLS LVs. Each row of this SRD input matrix would contain the corresponding RMSECV fold values for that particular split at the respective LVs. The input reference target RMSECV values for the SRD algorithm would be the row minima. The SRD algorithm uses this input matrix to rank the PLS LVs (models) relative to meeting target minima and presents model rankings providing the user with an automatic process to select the most consistent model(s). The closer a LV SRD value is to zero, the closer the ranking is to reference minima values. The PLS models (LVs) with similar SRD values are models predicting similarly. As noted, SRD values can also be considered as a dissimilarity measure and the greater the value, the more dissimilar to the reference minima values. To validate SRD results, the CRNN and CV processes described in Section 4 can be used (in this example situation, CV of the PLS RMSECV rows in the SRD input matrix).

The rows of this example SRD PLS RMSECV input matrix can be augmented with a second block of the corresponding CV split-wise $\| \hat{\mathbf{b}} \|$ values. The target reference values for this block would be row minima. Additional model merits can be augmented as other blocks. A similar tuning parameter selection process can be used to rank and select a RR model or a pool of models as well as ranking and selecting other tuning parameter dependent modeling methods. Regardless, the SRD process ranks the tuning parameters relative to the consistency of meeting the respective target values across the merits being assessed. Ultimately, the final tuning parameter rankings are affected by what type of model merits the user has selected to use for rows in the SRD input matrix. To simultaneously compare modeling methods in conjunction with tuning parameter selection for a particular data set, the SRD input matrix is column-wise augmented with the corresponding model tuning parameters.

5. Experimental

5.1. Algorithms

MATLAB 8.1 (The Math Works, Natick, MA) algorithms for RR, PLS, CV, SRD, and all model merits were written by the authors. The SRD Excel versions are downloadable [38] as is the MATLAB version [39]. In all cases, the SRD input matrix was row-wise normalized to unit length.

5.2. Cross-validation to form PLS and RR models

In order to assess model tradeoffs within a modeling process as well as between modeling methods, the LMOCV format was used. For each data set, 100 splits were used and on each split, a random 60% of the samples went to form the calibration set and the remaining 40% were used for validation. On each split, values for model merits such as vector L_2 norm, J, RMSECV, etc. were computed for each tuning parameter value. The maximum number of PLS LVs was determined by the respective data sets mathematical ranks (min(*m*,*p*)). The number of RR tuning parameters and actual values differ per data set and are specified in the following data set descriptions. On each CV split, all samples were column-wise mean-centered to the calibration set before forming respective models and predictions.

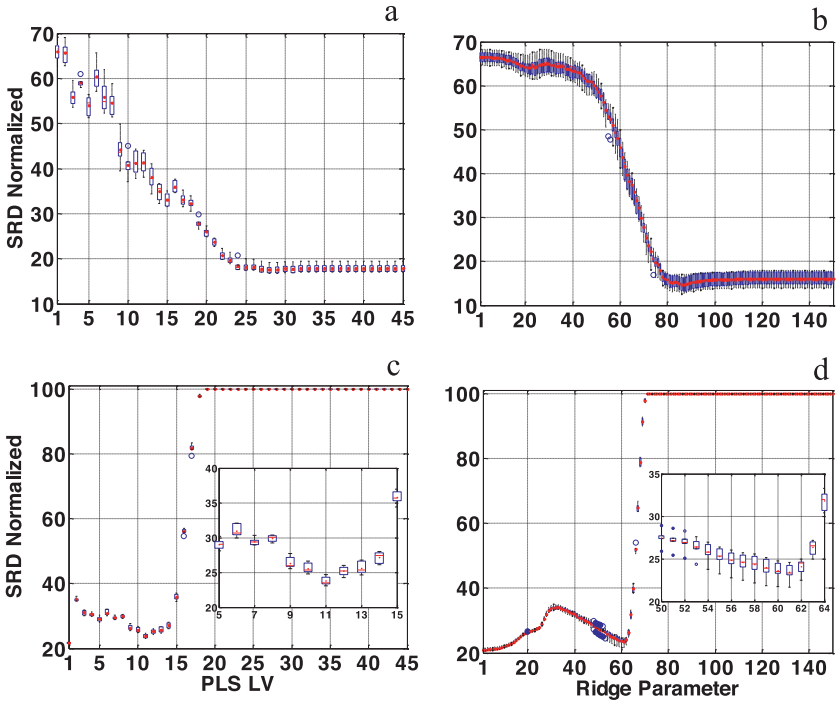


Fig. 3. Corn data SRD boxplots using 7-fold CV on the (a) PLS 100 LMOCV RMSECV block in Fig. 1a, (b) respective RMSECV RR block in Fig. 1b, (c) PLS RMSECV and L_2 norm blocks, and (d) respective RR RMSECV and L_2 norm blocks.

5.3. SRD validation

The SRD CRRN results were inspected to ensure models of interest were not randomly ranked. A graphical example is presented for the corn data. In this case, the SRD input matrix is composed of mean merit values across the 100 LMOCV as single rows. Otherwise, graphical results displayed are boxplots from using SRD in the 7-fold CV mode for each block of model merits.

5.4. NIR corn data

Eighty samples of corn were measured from 1100 to 2498 nm at 2 nm intervals for 700 wavelengths on three near infrared (NIR) spectrometers designated m5, mp5 and mp6 [61]. Reference values are provided for oil, protein, starch and moisture content. Presented are the protein results using m5. The η RR tuning parameter values exponentially decrease from 68 to 6.7×10^{-7} for 150 values.

5.5. Quantitative structure activity relationship (QSAR) data

The QSAR data consist of 142 compounds with 63 molecular descriptors [62]. The compounds were assayed for inhibition of the three carbonic anhydrase (CA) isozymes CA I, CA II, and CA IV. Carbonic anhydrase contributes to production of eye humor which with excess secretion, causes permanent damage and diseases (macular edema and open-angle glaucoma). Results are presented for CA I. The η RR tuning parameter values exponentially decrease from 11,383 to 1.2×10^{-4} for 80 values.

6. Results and discussion

6.1. Corn

Shown in Fig. 1 are images of the PLS and RR CV split-wise RMSECV results for the 100 LMOCVs. Plotted in Fig. 2 are the mean PLS and RR RMESCV plots against the respective tuning parameters as well as PLS and RR graphics plotting mean RMSECV and RMSEC values against the mean model L_2 norm and J values. Also plotted are C1 and C2 (where C2 has been inverted for maximization). The images in Fig. 1 show the discrete nature of PLS versus the continuous aspect of RR. This difference is further exemplified in the corresponding plots shown in Fig. 2. From the expanded mean RMSECV plots in Fig. 2a and d, it is observed that empirically selecting appropriate tuning parameter values is not obvious. Fig. 2b for PLS shows that by plotting the mean RMSECV or RMSEC values against the model complexity measure L_2 norm, the tradeoff becomes discernible in the corner regions of the L-curves assisting in selecting the number of LVs. Note that in Fig. 2b, the models are no longer equally spaced across the *x*-axis compared to Fig. 2a. While models in the corner regions are those balancing the tradeoff, the plots of C1 and C2 allow automatic selection with 9 LVs chosen using C1 and 11 LVs from the C2. These two models are in the corner regions of the L-curves. Using J values (jaggedness or roughness) of the model vectors instead of the L_2 norms does not provide any additional insight in the graphics other than the early LV models change little in jaggedness while the other model merits are adjusting. A similar discussion can be formed for Fig. 2d–f. From the mean C1 and C2 plots, ridge parameters 60 ($\eta = 1.0 \times 10^{-2}$) and 65 ($\eta = 4.8 \times 10^{-3}$) are chosen.

While mean C1 and C2 values are useful in selecting a tuning parameter for PLS and RR, these composite merits are limited in the number of specific model merits evaluated and the individual CV values are not assessed. Using SRD can alleviate these restrictions. Evaluated first are the SRD 7-fold CV results using the split-wise PLS and RR RMSECV matrices imaged in Fig. 1 as the

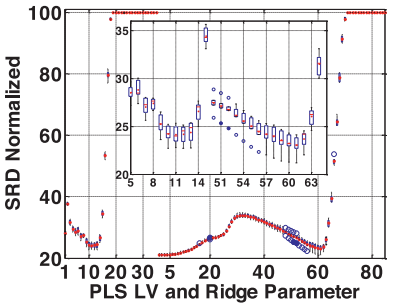


Fig. 4. Corn data SRD boxplots from combing the PLS and RR RMSECV and L_2 norm values into one SRD.

SRD input matrix. These results are presented as boxplots in Fig. 3a and b. It is not surprising from the mean RMSECV plots in Fig. 2a and d that using row minima as the SRD targets results in lower SRD rank values starting at 25 LVs and the 80th ridge parameter ($\eta = 5.6 \times 10^{-4}$). Thus, additional model merits are needed as these models are overfitted. Including the block of respective 100 LMOCV L_2 norm results in the SRD 7-fold CV boxplots presented in Fig. 3c and d. By including the L_2 norm for a model complexity and variance indicator, the SRD process now ranks 11 LVs the lowest for PLS (ignoring the 1 LV model) and ridge parameter 61 ($\eta = 8.8 \times 10^{-3}$) for RR (ignoring approximately the first twenty ridge parameters). Substituting J for the L_2 norm results in similar plots to Fig. 3c with no change in the lowest ranked PLS model and the lowest ranked RR model is now ridge parameter 68 ($\eta = 3.1 \times 10^{-3}$). Results from combining the RMSECV and L_2 norm CV blocks for PLS and RR into one SRD are displayed in Fig. 4. These plots indicate that PLS and RR are modeling equivalently.

Other model merits can be included in the SRD process. Shown in Fig. 5 are the PLS and RR SRD results using only calibration information based on the RMSEC, C1, J, and L_2 norm values. In this case, 17 PLS LVs and ridge parameter 65 ($\eta = 4.8 \times 10^{-3}$) obtain clear lowest rankings. The change in rankings of the tuning parameters is due to including more model merits and SRD assessing a consensus in the rankings relative to row wise target values. To

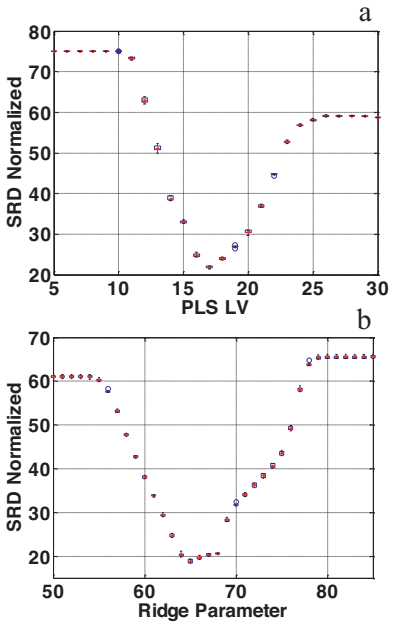


Fig. 5. Corn data SRD boxplots using model calibration merits RMSEC, C1, J, and L_2 norm for (a) PLS and (b) RR.

further characterize the consensus nature of SRD, shown in Fig. 6a is an image of the SRD input matrix for RR. This SRD input matrix sorted to the SRD rankings from low to high is imaged in Fig. 6b. From this image, the models sustaining consistency to the targets are ranked lowest. The image in Fig. 6c is the RMSECV image in Fig. 1b sorted to the SRD rankings showing that the SRD ranked tuning parameters provide consistently low RMSECV values.

Augmenting the previous calibration merits with the split-wise CV results for RMSECV and C2 provides SRD results similar to that shown in Fig. 5a and b with the lowest ranked model for PLS moving to 15 PLS LVs and the ridge parameter remained at 65. Combining these additional model merits with the previous ones into one SRD for PLS and RR showed that PLS and RR are performing consistently similar.

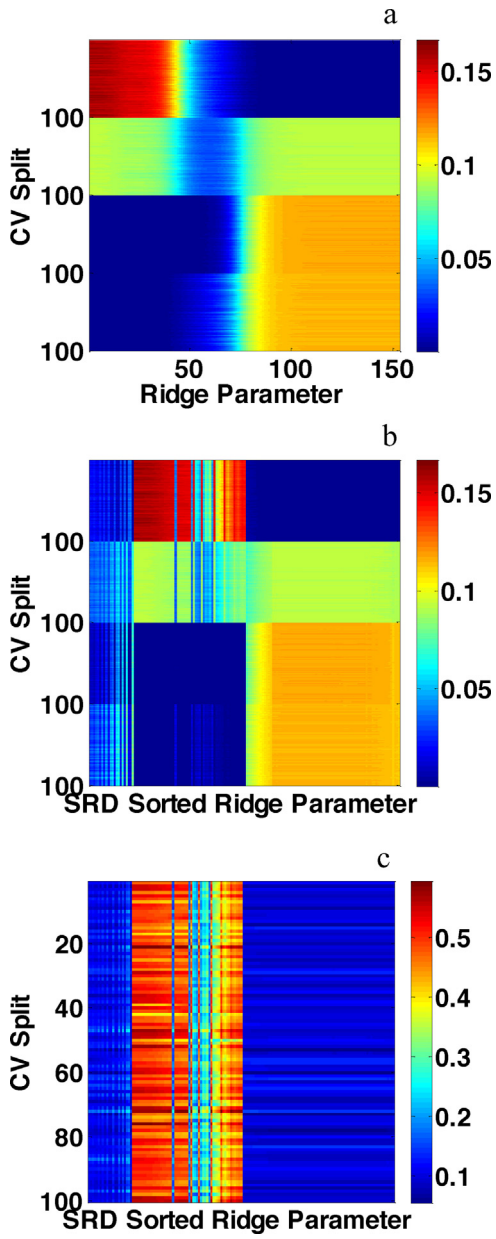


Fig. 6. Corn data images for the situation in Fig. 5 with (a) the input SRD matrix, (b) the input SRD matrix in (a) sorted to the SRD rankings from low on the left to high on the right, and (c) the RMSECV matrix in Fig. 1b sorted to the SRD rankings. For (a) and (b), the four CV blocks with 100 matching splits each are in the order RMSEC, C1, J, and L_2 norm. Each row of the SRD input matrix was scaled to unit length. The RMSECV matrix in (c) are actual values.

Another variation of the SRD input matrix generated the boxplots shown in Fig. 7 for PLS and RR. In this variation, 18 blocks of model merits were used consisting of RMSEC, R^2_{cal}, $slope_{cal}$, $intercept_{cal}$, RMSECV, R^2_{cv}, $slope_{cv}$, $intercept_{cv}$, C1, using J in C1, the corresponding two variation of C1 using RMSECV, C2, using respective R^2 values in C2, and two other variations of C2 missing R^2 with RMSE values, J, and L_2 norm. With these model merits, the 14 PLS LV model is ranked lowest and the ridge parameter model 65 ($\eta = 4.8 \times 10^{-3}$) is ranked lowest. Depending on the actual merits used in SRD, the lowest ranked models can vary, but remain in close proximity to each other indicating that there is probably not one best model and a collection of models can be useful and are essentially equivalent. The final model choice of the user depends on the tradeoffs desired for the final model. Using these 18 model merits to evaluate PLS and RR together provided similar results to that presented in Fig. 4 with the PLS and RR modeling equivalently.

Rather than using all the respective individual LMOCV results for different merit blocks in the SRD input matrix, the corresponding mean LMOCV merit values can be used as single rows provided that enough model merits are included to reduce the chance of random rankings (typically 7 or more rows for the SRD input matrix, but more are better). In this case, the SRD input matrix is considered as one block. Shown in Fig. 8 is an example of the CRRN result based on an SRD input matrix composed of one block with 18 rows with each row being the respective mean CV values of the 18 model merits previously used. As a reminder, the CRRN process involves random distributions based on random numbers for a small number of objects and the normal distribution, as used in this case, for a large number of objects. The reader is referred to reference [34] for the details of CRRN. Listed are the SRD top five rankings for PLS and RR. The results are essentially the same as those ranked best by the SRD evaluation of the same merits in block format and validated by the CV of the SRD input matrix to form the boxplots. Listed in the outlined boxes shown in Fig. 8 are the PLS LVs and RR ridge parameters followed by the SRD normalized rankings and then the probabilities. From the listed

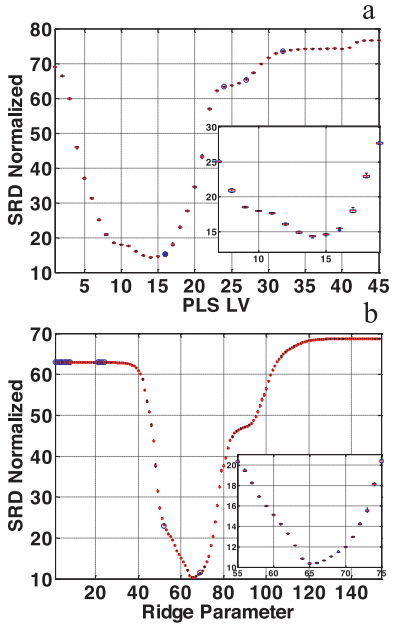


Fig. 7. Corn data SRD boxplots for (a) PLS and (b) RR using 18 blocks of model merits consisting of RMSEC, R^2_{cal}, $slope_{cal}$, $intercept_{cal}$, RMSECV, R^2_{cv}, $slope_{cv}$, $intercept_{cv}$, C1, using J in C1, the corresponding two variation of C1 using RMSECV, C2, using respective R^2 values in C2, and two other variations of C2 using R^2 with RMSE values, J, and L_2 norm.

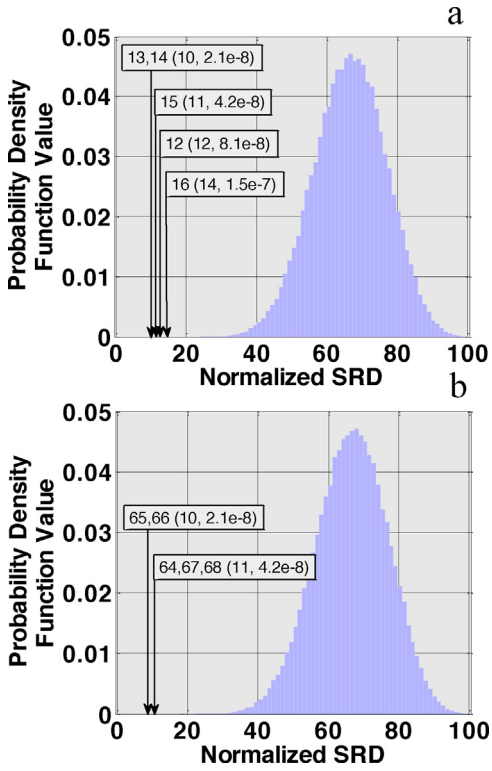


Fig. 8. Differences between random and actual corn model rankings (SRD corn CRRN plots) for (a) PLS and (b) RR with the respective five lowest rank models. For PLS, the first number in each box is the PLS LV model and the first value in the parenthesis is the SRD ranking followed by the probability density function value. It is similar for RR except the first numbers in each box are the RR ridge parameters with actual ridge values of 65 (4.8×10^{-3}), 66 (4.1×10^{-3}), 64 (5.6×10^{-3}), 67 (3.6×10^{-3}), and 68 (3.1×10^{-3}).

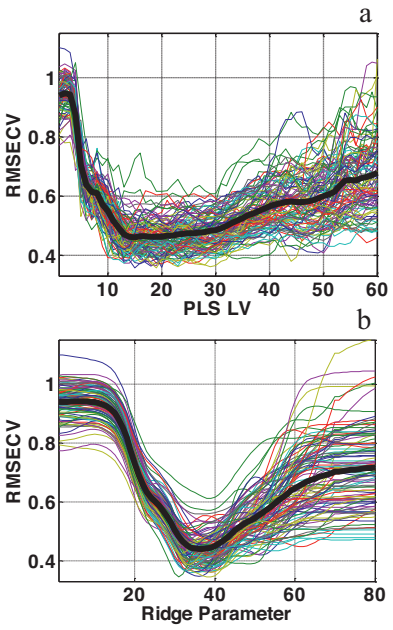


Fig. 9. QSAR CV split-wise RMSECV plots for (a) PLS and (b) RR for the 100 LMOCVs and respective tuning parameters. Starting at ridge parameter 1, 80 ridge values range from 11,383 to 1.2×10^{-4} at ridge parameter 80. Black lines are the mean RMSECV values.

probabilities in conjunction with the plotted probability functions, it can be observed that the model rankings are by no means random rankings because these SRD model rankings are not located within the plotted random distributions.

When using the SRD process to evaluate model tuning parameters as in this paper, it is important to have merits balancing model tradeoffs such as the bias/variance tradeoff. For example, with PLS, if the only model merits used in an SRD analysis minimize toward the maximum number of LVs (the overfitted region) such as with RMSEC, $1-R^2_{cal}$, etc., then the SRD algorithm with minima set as the target reference values will sort these overfitted models with the lowest SRD rank values.

Tabulated in Table 1 are final model merits for those models with low ranks from all the above variants of model merits with and without SRD. The "best" model with the lowest SRD ranking is going to depend on which specific models merits are used. As more model merits are included in an SRD analysis, the less variation there is in the listed model merits. For PLS, this tends to be the higher number of LVs in Table 1 and the smaller ridge parameter values for RR.

For a more specific statistical comparison between models, the uncertainties computed by the SRD CV process can be evaluated by a Wilcoxon signed rank test at a given significance level. For example, testing RR models 67 and 68 in Fig. 7b at the 5% significance level shows that there is no difference between the models. Testing models 66 and 67 results in a statistical difference. Testing the low ranked PLS models in Fig. 7a reveals that the

Table 1
Corn data mean PLS and RR LMOCV model merit values for models with low SRD rankings based on different SRD input model merits.

Method	PLS LV or ridge parameter (η)	RMSECV	R^2	Slope	Intercept	$\|\hat{\mathbf{b}}\|_2$
PLS	9	0.137	0.926	0.94	0.53	63.3
PLS	10	0.133	0.930	0.95	0.46	68.3
PLS	11	0.123	0.940	0.96	0.35	83.0
PLS	12	0.118	0.945	0.96	0.30	98.8
PLS	13	0.113	0.949	0.96	0.28	109
PLS	14	0.108	0.954	0.97	0.26	125
PLS	15	0.109	0.954	0.97	0.23	146
PLS	16	0.108	0.955	0.97	0.21	171
PLS	17	0.108	0.955	0.98	0.20	193
RR	60 (1.0×10^{-2})	0.136	0.927	0.91	0.79	56.9
RR	61 (8.8×10^{-3})	0.131	0.933	0.92	0.70	61.1
RR	62 (7.5×10^{-3})	0.126	0.937	0.93	0.63	65.8
RR	63 (6.5×10^{-3})	0.122	0.942	0.93	0.57	70.9
RR	64 (5.6×10^{-3})	0.118	0.945	0.94	0.51	76.6
RR	65 (4.8×10^{-3})	0.114	0.948	0.95	0.46	82.9
RR	66 (4.1×10^{-3})	0.111	0.951	0.95	0.42	89.8
RR	67 (3.6×10^{-3})	0.109	0.953	0.95	0.39	97.6
RR	68 (3.1×10^{-3})	0.107	0.955	0.96	0.35	106

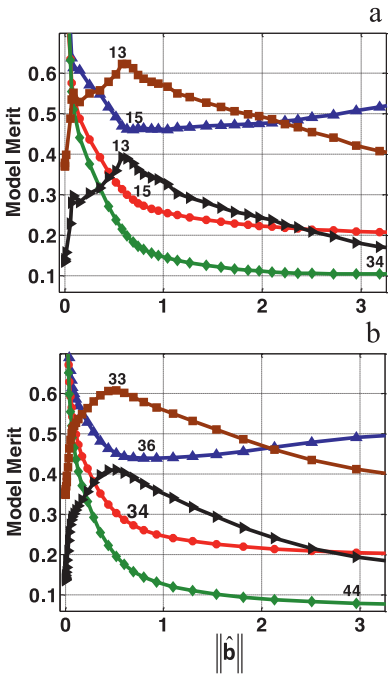


Fig. 10. Expanded QSAR model merit graphics of (a) PLS and (b) RR mean model merits plotted against the mean model L_2 norm values for RMSECV (blue triangles), RMSEC (red circles), C1 (green diamonds), C2 (brown squares), and C2 with respective R^2 values replacing the RMSE values (black right facing triangles). Values are scaled to fit in plot. Numbers in (a) correspond to number of LVs and in (b), the ridge parameters. Ridge values trend from large on left to small on the right. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

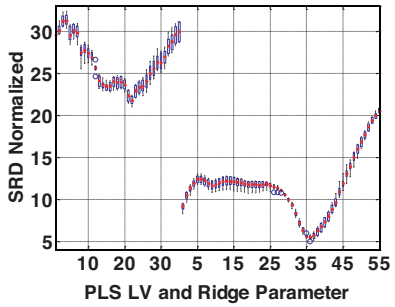


Fig. 12. QSAR data SRD boxplots from combing the PLS and RR RMSECV and L_2 norm values into one SRD.

models are all unique. While not studied in this paper, the Wilcoxon signed rank test can also be used to compare PLS models to RR models.

Models with low SRD rankings can be used in a consensus approach. To successfully utilize consensus modeling, a high degree of prediction accuracy is desired in combination with a small but noteworthy difference between the selected models (model diversity) [32,63–65]. Once a collection is selected, various methods exist to form the composite prediction from these models such as the simple approach of using the mean prediction. The collection can be a mix of PLS and RR models as well as from a single modeling method. This approach was not evaluated in this study.

6.2. QSAR

Rather than showing RMSECV blocks as images as done with the corn data, drawn in Fig. 9 are the 100 individual and mean RMSECV plots for PLS and RR. From these plots, models to select are more obvious than with the corn RMSECV graphics. Displayed in Fig. 10

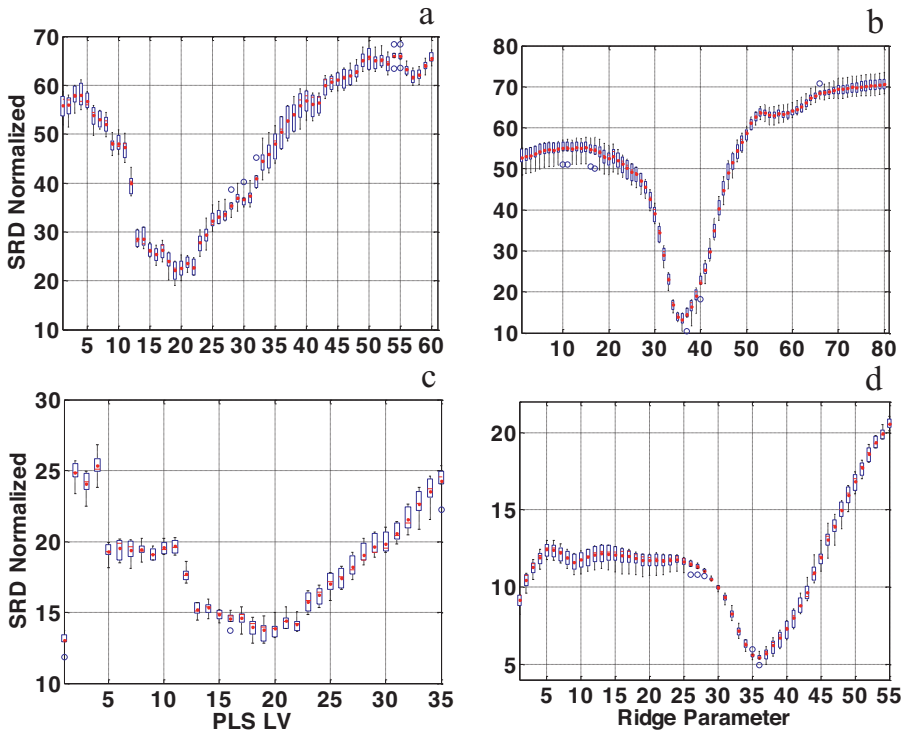


Fig. 11. QSAR data SRD boxplots using 7-fold CV on the (a) PLS 100 LMOCV RMSECV block in Fig. 9a, (b) respective RMSECV RR block in Fig. 9b, (c) PLS RMSECV and L_2 norm blocks, and (d) respective RR RMSECV and L norm blocks.

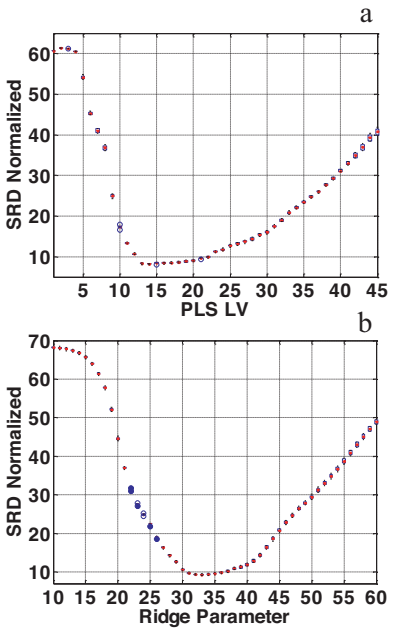


Fig. 13. QSAR boxplots of (a) PLS and (b) RR SRD results from using 18 blocks of model merits consisting of RMSEC, R^2_{cal}, $slope_{cal}$, $intercept_{cal}$, RMSECV, R^2_{cv}, $slope_{cv}$, $intercept_{cv}$, C1, using J in C1, the corresponding two variation of C1 using RMSECV, C2, using respective R^2 values in C2, and two other variations of C2 missing R^2 with RMSE values, J, and L_2 norm.

are the PLS and RR graphics plotting mean RMSECV and RMSEC values against the mean model L_2 norm. Also plotted with these graphics are mean C1 and C2 (where C2 has been inverted for maximization) as well as C2 with respective R^2 values replacing the RMSE values. Using J values instead of the L_2 norm values produces similar plots. As expected from the block of individual RMSECV values and mean plots in Fig. 9, selecting the tuning parameters from the other mean model merits results in similar selections. For PLS, the minimum RMSECV is at 15 LV and the C2 merit in both formats forms minima at 13 LVs. The range from 13 to 15 LVs is in the corner region of the RMSEC L-curve. Models based on 16 through 20 are also in the corner region. While not apparent in Fig. 10a, the mean C1 merit minimizes for PLS at 34 LVs and provides an overfitted model selection. Replacing RMSEC in C1 with RMSECV, produces a minimum at 15 LVs.

Similar trends are present for RR in Fig. 10b. Ridge parameters selected using the plots from mean RMSECV, C2, and C2 with R^2 values are 36 ($\eta = 1.6$), 33 ($\eta = 3.4$), and 33, respectively. These ridge parameters are in the corner regions of the mean RMSEC L-curve. The mean C1 merit identifies ridge parameter 50 ($\eta = 4.6 \times 10^{-2}$) at the minimum and replacing RMSEC with RMSECV in C1 ascertains ridge parameter 36 at the minimum.

Evaluated first with SRD are the 7-fold CV boxplots in Fig. 11 based on using the PLS and RR RMSECV blocks plotted in Fig. 9. Interesting that with SRD, the 19 LV model is deemed lowest rank relative to target minimization and hence, the most consistently minimized LV across the 100 LMOCV. Using the Wilcoxon signed rank test at the 5% significance level reveal no difference between LVs 19 through 22. There appears to be a local minimum from 15 through 17 LVs and this is the region identified in the single merit plots in Fig. 10. For RR in Fig. 11b, the model with ridge parameter 36 is the lowest consistently ranked model.

Including blocks of the respective model complexity measure L_2 norm for PLS and RR forms the plots shown in Fig. 11c and d. The same models are ranked the lowest as with just the RMSECV blocks, but for PLS, models 13 through 17 have similar ranks. Unlike with the corn data, when PLS and RR RMSECV and L_2 norm values are combined into one SRD, Fig. 12 shows that RR provides lower ranked models than PLS. With the corn data, PLS and RR essentially performed equivalently as portrayed in Fig. 4.

As with the corn data, other merits can be combined for an SRD evaluation. Which merits depend on what the user defines as best for their purposes. For this QSAR data set and prediction property, using only calibration merits pushes the tuning parameters to the overfitted regions. Unlike with estimating the protein prediction property with the corn data, some form of CV appears necessary in this QSAR instance. Presented in Fig. 13 are boxplots for PLS and RR from using the 18 blocks of model merits used with the corn data composed of RMSEC, R^2_{cal}, $slope_{cal}$, $intercept_{cal}$, RMSECV, R^2_{cv}, $slope_{cv}$, $intercept_{cv}$, C1, using J in C1, the corresponding two variations of C1 using RMSECV, C2, using respective R^2 values in C2, and two other variations of C2 missing R^2 with RMSE values, J, and L_2 norm. Using this mix of calibration and validation merits results in 14 LV being ranked the lowest for PLS and ridge parameter model 33 for RR. As with the corn data, the boxplot box sizes are substantially reduced indicating better regularity in the SRD rankings. Using these 18 model merits for an SRD analysis of PLS and RR simultaneously showed PLS to have a smaller SRD ranking by one unit than RR at the respective lowest ranked models of 14 LVs and ridge parameter 33.

Table 2
QSAR data mean PLS and RR LMOCV model merit values for models with low SRD rankings based on different SRD input model merits.

Method	PLS LV or ridge parameter (η)	RMSECV	R^2	Slope	Intercept	$\|\hat{\mathbf{b}}\|_2$
PLS	13	0.470	0.777	0.84	0.52	0.58
PLS	14	0.462	0.786	0.85	0.48	0.63
PLS	15	0.461	0.788	0.85	0.46	0.69
PLS	16	0.464	0.788	0.86	0.44	0.74
PLS	17	0.463	0.789	0.87	0.43	0.80
PLS	18	0.463	0.790	0.87	0.41	0.88
PLS	19	0.462	0.790	0.87	0.41	0.95
PLS	20	0.461	0.790	0.87	0.42	1.04
PLS	21	0.466	0.786	0.87	0.42	1.14
RR	30 (7.3)	0.505	0.740	0.73	0.84	0.29
RR	31 (5.7)	0.482	0.762	0.76	0.76	0.36
RR	32 (4.4)	0.464	0.779	0.78	0.68	0.43
RR	33 (3.4)	0.452	0.791	0.81	0.62	0.52
RR	34 (2.6)	0.445	0.798	0.82	0.56	0.60
RR	35 (2.0)	0.441	0.802	0.84	0.52	0.69
RR	36 (1.6)	0.440	0.804	0.85	0.49	0.80
RR	37 (1.2)	0.440	0.805	0.85	0.46	0.93
RR	38 (0.96)	0.442	0.804	0.86	0.44	1.10

Tabulated in Table 2 are final model merits for those models with low ranks from the different SRD input matrices as expressed above as well as the described signal merits. As with the corn data set, the better models listed in Table 2 are those deemed "best" by using multiple model merits compared to those models selected by single merits. As a reminder, the user can use Wilcoxon signed rank tests to evaluate uniqueness of specific models whether the goal is between different modeling methods or within a modeling method.

7. Conclusions and SRD recommendations

The goal of this paper is not to show that one modeling method is better than another, but to develop SRD as a tool for selecting tuning parameters and comparing models. Using SRD allows multiple model merits to be used for selection of model tuning parameters. The lowest ranked model can be selected or, alternatively, a collection of models with low SRD rankings can be used in a consensus approach. The collection of models can be for a single modeling method as well as a mix of different modeling methods such as PLS and RR. The SRD corresponds to the principle of parsimony and the SRD CV process to form boxplots provides uncertainties for the variables (columns) and the differences can be tested in a statistically correct way.

The better models are those having the most consistency across the different model merits evaluated. When a CV process is used to generate the model merits, then SRD allows the models merits computed on each data split to be evaluated, not just the mean values as in the standard CV proves of selecting a tuning parameter. The more model merits included to characterize the bias/variance tradeoffs, the less variation in the SRD CV boxplots for the lowest ranked models. Only a limited set of combinations of model merits were evaluated with SRD in this study. Not studied in this paper was using other model merits such as Mallow's C_p criterion, AIC [43–46], etc. to build up the number of objects for SRD. Which actual tuning parameters are ranked lowest by SRD depends on which model merits are used. As with any tuning parameter selection process, it is up to the user to decide which model merit (s) is to be used to evaluate the tuning parameters. The SRD process allows rapid comparison of the consistency of tuning parameters as model merits vary by the user.

As noted, evaluation of the consistencies of model tuning parameters can be enhanced by increasing the number of model merits. In this study only the composite split-wise merit values were used, e.g., one row of RMSECV values for each CV split. Additional SRD blocks can be included using the actual predicted values of all samples in each respective split. For example, for each RMSECV row, a block of $\hat{y}_{cv}$ values (r by number of tuning parameters for r validation samples) could be included. Target reference values would be the corresponding reference values y_{val}. Alternatively, the SRD input values could be $|\hat{y}_{cv} - y_{cv}|$ with target values of row minima. Similarly, additional blocks for the SRD input matrix could be added based on different types of CV splits as well as perturbing the data with noise and creating sets of merit blocks for each noise perturbation.

The SRD process described in this study is generic and should be applicable to other multivariate calibration methods involving selection of single tuning parameters such as the TR variant known as least absolute shrinkage and selection operator (LASSO), principal component regression (PCR), and others. Under current study is using SRD with multivariate calibration processes that involve multiple tuning parameters. The SRD process is a simple general method that is finding more uses.

With multivariate calibration, variable selection (wavelength selection with optical spectroscopic data) is often used to reduce prediction errors and improve robustness. In this paper, full wavelengths were used with the corn data and all the provided variables were used with the QSAR data. Using SRD, it is possible to select tuning parameters for models generated by variable selection processes. Various variable selected models can also be compared to full variable models by SRD. The SRD process provides a natural way to impartially compare different modeling methods.

The reader should note that SRD has two operational modes. That is, for many applications, the SRD input matrix can be transposed where the objects are now the variables and the variables are now the objects. Transposing the SRD input matrices for the situations studied in this paper was not investigated. Such an operation should allow comparison of the model merits. That is, the merits would be ranked by how consistently the respective merits meet the respective target values. The lowest ranked merits could be deemed "best".

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. CHE-1111053 (co-funded by MPS Chemistry and the OCI Venture Fund) and is gratefully acknowledged by the authors. Károly Héberger's contribution was supported by OTKA under Contract No. K112547.

References

[1] T. Næs, T. Isaksson, T. Fern, T. Davies, A User Friendly Guide to Multivariate Calibration and Classification, NIR Publications, Chichester, UK, 2002.
[2] T.J. Hastie, R.J. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer-Verlag, New York, 2009.
[3] J.H. Kalivas, Calibration methodologies, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, vol. 3, Elsevier, Amsterdam, 2009, pp. 1–32.
[4] J. Shao, Linear model selection by cross-validation, J. Am. Stat. Assoc. 88 (1993) 486–494.
[5] Q.S. Xu, Y.Z. Liang, Monte Carlo cross-validation, Chemom. Intell. Lab. Syst. 56 (2001) 1–11.
[6] K. Baumann, H. Albert, M. von Korff, A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I: search algorithm, theory, and simulations, J. Chemom. 16 (2002) 339–350.
[7] P. Filzmoser, B. Liebmann, K. Varmuza, Repeated double cross-validation, J. Chemom. 23 (2009) 160–171.
[8] S. Wold, Cross-validatory estimation of the number of components in factor and principal component models, Technometrics 20 (1978) 397–405.
[9] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, Wiley, New York, 1998.
[10] N.M. Faber, R. Rajkó, How to avoid over-fitting in multivariate calibration – the conventional validation approach and an alternative, Anal. Chim. Acta 595 (2007) 98–106.
[11] S. Wiklund, D. Nilsson, L. Eriksson, M. Sjöström, S. Wold, K. Faber, A randomization test for PLS component selection, J. Chemom. 21 (2007) 427–439.
[12] M. Wasim, R.G. Brereton, Determination of the number of significant components in LC NMR spectroscopy, Chemom. Intell. Lab. Syst. 72 (2004) 133–151.
[13] K. Booksh, B.R. Kowalski, Theory of analytical chemistry, Anal. Chem. 66 (1994) 782A–791A.
[14] W.P. Carey, B.R. Kowalski, Chemical piezoelectric sensor and sensor array characterization, Anal. Chem. 58 (1986) 3077–3084.
[15] L.L. Juhl, J.H. Kalivas, Evaluation of experimental designs for multicomponent determination by spectrophotometry, Anal. Chim. Acta 207 (1988) 125–135.
[16] J.H. Kalivas, P.M. Lang, Mathematical Analysis of Spectral Orthogonality, Marcel Dekker, New York, 1994.
[17] N.M. Faber, Multivariate sensitivity for the interpretation of the effect of spectral pretreatment methods on near-infrared calibration model predictions, Anal. Chem. 71 (1999) 557–565.
[18] A. Höskuldsson, Dimension of linear models, Chemom. Intell. Lab. Syst. 32 (1996) 37–55.
[19] F. Bauer, M.A. Lukas, Comparing parameter choice methods for regularization of ill-posed problems, Math. Comput. Simul 81 (2011) 1795–1841.
[20] R.L. Green, J.H. Kalivas, Graphical diagnostics for regression model determinations with consideration of the bias/variance tradeoff, Chemom. Intell. Lab. Syst. 60 (2002) 173–188.
[21] J.B. Forrester, J.H. Kalivas, Ridge regression optimization using a harmonious approach, J. Chemom. 18 (2004) 372–384.

[22] N.M. Faber, A closer look at the bias-variance tradeoff in multivariate calibration, J. Chemom. 13 (1999) 185–192.
[23] J.H. Kalivas, J. Palmer, Characterizing multivariate calibration tradeoffs (bias, variance, selectivity, and sensitivity) to select model tuning parameters, J. Chemom. 28 (2014) 347–357.
[24] P.C. Hansen, Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion, SIAM, Philadelphia, PA, 1998.
[25] P.C. Hansen, Analysis of discrete ill-posed problems by means of the L-curve, SIAM Rev. 34 (1992) 561–580.
[26] J.H. Kalivas, Basis sets for multivariate regression, Anal. Chim. Acta 428 (2001) 31–40.
[27] L.A. Pinto, R.K.H. Galvão, M.C.U. Araújo, Ensemble wavelet modeling for determination of wheat and gasoline properties by near and model infrared spectroscopy, Anal. Chim. Acta 682 (2010) 37–47.
[28] A.A. Gowen, G. Downey, C. Esquerre, C.P. O'Donnell, Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients, J. Chemom. 25 (2011) 375–381.
[29] F. Stout, M. Baines, J.H. Kalivas, Impartial graphical comparison of multivariate calibration methods and the harmony/parsimony tradeoff, J. Chemom. 20 (2006) 464–475.
[30] N.R. Costa, J. Lourençoa, Z.L. Pereira, Desirability function approach: a review and performance evaluation in adverse conditions, Chemom. Intell. Lab. Syst. 107 (2011) 234–244.
[31] S. Verboven, M. Hubert, P. Goos, Robust preprocessing and model selection for spectral data, J. Chemom. 26 (2012) 282–289.
[32] P. Shahbazikhah, J.H. Kalivas, A consensus modeling approach to update a spectroscopic calibration, Chemometr. Intell. Lab. Sys. 120 (2013) 142–153.
[33] K. Héberger, Sum of ranking differences compares methods or models fairly, Trends Anal. Chem. 29 (2010) 101–109.
[34] K. Héberger, K. Kollár-Hunek, Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers, J. Chemom. 25 (2011) 151–158.
[35] K. Héberger, B. Škrbić, Ranking and similarity for quantitative structure–retention relationship models in predicting Lee retention indices of polycyclic aromatic hydrocarbons, Anal. Chim. Acta 716 (2012) 92–100.
[36] B. Škrbić, K. Héberger, N. Durišić-Mladenović, Comparison of multianalyte proficiency test results by sum of ranking differences principal component analysis, and hierarchical cluster analysis, Anal. Bioanal. Chem. 405 (2013) 8363–8375.
[37] K. Kollár-Hunek, K. Héberger, Method of model comparison by sum of ranking differences in cases of repeated observations (ties), Chemom. Intell. Lab. Syst. 127 (2013) 139–146.
[38] Download address: http://aki.ttk.mta.hu/srd (assessed 26.09.14).
[39] Download address: http://www.isu.edu/chem/people/faculty/kalijohn/ (assessed 01.14).
[40] H.A. Seipel, J.H. Kalivas, Effective rank for multivariate calibration methods, J. Chemom. 18 (2004) 306–311.
[41] J.H. Kalivas, H.A. Seipel, Erratum to H.A. Seipel, J.H. Kalivas. Effective rank for multivariate calibration methods, J. Chemom. 18 (2004) 306–311, J. Chemom. 19 (2005) 64.
[42] C.M. Rubingh, H. Martens, H. van der Voet, A.K. Smilde, The costs of complex model optimization, Chemom. Intell. Lab. Syst. 125 (2013) 139–146.
[43] A.N. Tikhonov, Solution of incorrectly formulated problems and the regularization method, Soviet Math. Dokl. 4 (1963) 1035–1038.
[44] R.C. Aster, B. Borchers, C.H. Thurbe, Parameter Estimation and Inverse Problems, Elsevier, Amsterdam, 2005.
[45] J.H. Kalivas, Overview of two-norm (L_2) and one-norm (L_1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance, J. Chemom. 26 (2012) 218–230.
[46] R.H. Myers, Classical and Modern Regression with Applications, second ed., Duxbury, Pacific Grove, 1990.
[47] G.H. Golub, M. Heath, G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics 21 (1979) 215–223.
[48] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (1974) 716–723.
[49] G.E. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[50] J.A. Koziol, Sums of ranking differences and inversion numbers for method discrimination, J. Chemom. 27 (2013) 165–169.
[51] E.V. Thomas, Non-parametric statistical methods for multivariate calibration model selection and comparison, J. Chemom. 17 (2003) 653–659.
[52] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering scaling, and transformations: improving the biological information content of metabolomics data, BMC Genomics 7 (2006) 1–15. http://www.biomedcentral.com/1471-2164/7/142.
[53] H.P. Bailey, S.C. Rutan, Comparison of chemometric methods for the screening of comprehensive two-dimensional liquid chromatographic analysis of wine, Anal. Chim. Acta 770 (2013) 18–28.
[54] J.E. Wood, D. Allaway, E. Boult, I.M. Scott, Operationally realistic validation for prediction of cocoa sensory qualities by high-throughput mass spectrometry, Anal. Chem. 82 (2010) 6048–6055.
[55] L. Sipos, Z. Kovacs, D. Szollosi, Z. Kokai, I. Dalmadi, A. Fekete, Comparison of novel sensory panel performance evaluation techniques with e-nose analysis integration, J. Chemom. 25 (2011) 275–286.
[56] B. Vajna, G. Patyi, Z. Nagy, A. Bodis, A. Farkas, G. Marosi, Comparison of chemometric methods in the analysis of pharmaceuticals with hyperspectral Raman imaging, J. Raman Spectrosc. 42 (2011) 1977–1986.
[57] D. Szollosi, D.L. Denes, F. Firtha, Z. Kovacs, A. Fekete, Comparison of six multiclass classifiers by the use of different classification performance indicators, J. Chemom. 26 (2012) 76–84.
[58] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection, Chemom. Intell. Lab. Syst. 109 (2011) 146–161.
[59] P. Willett, Combination of similarity rankings using data fusion, J. Chem. Inf. Model. 53 (2013) 1–10.
[60] C.M.R. Ginn, P. Willett, J. Bradshaw, Combination of molecular similarity measures using data fusion, Perspect. Drug Discov. Des. 20 (2000) 1–16.
[61] Eigenvector Research, Inc., Wenatchee, Washington, http://www.eigenvector.com/index.htm.
[62] B.E. Mattioni, P.C. Jurs, Development of quantitative structure-activity relationships and classification models for a set of carbonic anhydrase inhibitors, J. Chem. Inf. Comput. Sci. 42 (2002) 94–102.
[63] W. Tong, H. Hong, H. Fang, Q. Xie, R. Perkins, Decision forests: combining the predictions of multiple independent decision tree models, J. Chem. Inf. Comput. Sci. 43 (2003) 525–531.
[64] A.M. Van Rhee, Use of recursion forests in the sequential screening process: consensus selection by multiple recursion trees, J. Chem. Inf. Comput. Sci. 43 (2003) 941–948.
[65] M. Hibbon, T. Evgenoiu, To combine or not to combine: selecting among forecasts and their combinations, Int. J. Forecast. 21 (2005) 15–24.