Journal of Chemometrics

WILEY

Journal of CHEMOMETRICS

RESEARCH ARTICLE OPEN ACCESS

Adjusted Pareto Scaling for Multivariate Calibration Models

Kurt Varmuza | Peter Filzmoser

Institute of Statistics and Mathematical Methods in Economics, Research Unit Computational Statistics, TU Wien, Vienna, Austria

Correspondence: Kurt Varmuza (kurt.varmuza@tuwien.ac.at)

Received: 24 April 2024 | **Revised:** 26 June 2024 | **Accepted:** 6 July 2024

Keywords: multivariate calibration | repeated double cross validation (rdCV) | variable scaling

ABSTRACT

The performance of multivariate calibration models $\hat{y} = f(\boldsymbol{x})$ for the prediction of a numerical property y from a set of x-variables depends on the type of scaling of the x-variables. Common scaling methods are autoscaling (dividing the centered x by its standard deviation s) and Pareto scaling (dividing the centered x by s^P with $P = 0.5$). The adjusted Pareto scaling presented here varies the exponent P between 0 (no scaling) and 1 (autoscaling) with the aim of obtaining an optimum prediction performance for $\hat{y}$. Related scaling methods based on the variable spread are range scaling and vast scaling; while level scaling is based on the location (central value) of the variable. These scaling methods and robust versions are compared for models created by partial least-squares (PLS) regression. The applied strategy repeated double cross validation (rdCV) evaluates the model performance for test set objects and considers its variability. Results with three data sets from chemistry show: (a) the efficacy of the different scaling methods depends on the data structure; (b) optimization of the Pareto exponent P is recommended; (c) range scaling or vast scaling may be better than adjusted Pareto scaling; (d) in general a heuristic search for the best scaling method is advisable. Overall, the consideration of different variants of scaling allow for a flexible adjustment of the variable contributions to the calibration model.

1 | Introduction

Empirical, multivariate models $\hat{y} = f(\boldsymbol{x})$ for calibration or classification are well established in chemometrics and other areas of data science. The performance of such models for predicting a numerical property y from a set of x-variables may heavily depend on the data pretreatment, particularly on the scaling of the x-variables. Widely used are scaling methods based on the spread of the x-variables, and usually the variables are processed separately but by the same method. Autoscaling uses the ratio of the centered x and the standard deviation s of x; thus, the scaled variables have equal spread ($s = 1$) and mathematically an equal influence on the model. Pareto scaling uses the ratio of the centered x and $s^{0.5}$; thus, the strong influence of variables with a high variance is reduced but not eliminated. Here, we suggest a generalization of this type of scaling (calling it "adjusted Pareto scaling") defined by using the ratio of the centered x and s^P with the Pareto exponent P varied between 0 (no scaling) via 0.5 (standard Pareto scaling) to 1 (autoscaling). An optimum value of P for best model performance can be found by evaluating models made from data scaled with P varying between 0 and 1, for instance in steps of 0.1. Related scaling methods [1] based on the spread are range scaling (using the ratio of the centered variable and the range of the variable) and vast scaling (using the ratio of the centered variable and the relative standard deviation). Furthermore, level scaling uses the ratio of the centered variable and the center of the variable (for instance the median), thus not considering the spread. Details of these scaling methods are discussed in Section 2.1, and applications in multivariate calibration in Section 3.

Linear calibration models for modeling a numerical property y by a set of x-variables are calculated here by partial least-squares (PLS) regression. The model performances are compared by using the criterion standard error of prediction (SEP), which is defined as the standard deviation of prediction errors of test set

objects. PLS is combined with the strategy repeated double cross validation (rdCV) [2, 3], which is distinguished by (a) separating the estimation of an optimum model complexity (number of PLS components) from the estimation of the prediction performance, and (b) estimating the variability of SEP for different random splits of the data into calibration sets and test sets. Details of the use of PLS and rdCV are given in Section 2.2; a summary of rdCV is in Appendix 2.

The above mentioned scaling methods are applied for making calibration models from three data sets from chemistry (Section 2.3) with the results presented in Section 3. Aim of the study was demonstrating effects of adjusting Pareto scaling and comparing it with range scaling, vast scaling and level scaling, including robust method versions. Computations were performed within the programming environment R [4], particularly using the packages *pls* [5, 6] and *chemometrics* [7, 8]. The used symbols and basic definitions are given in Appendix 1.

2 | Methods and Data Sets

2.1 | Centering and Scaling

The notation used and the method descriptions are mainly based on publications by van den Berg et al. [9] and by Walach et al. [1], dealing with multivariate classification in metabolomics; this work, however, refers to calibration with examples from chemical technology, analytical chemistry and quantitative structure–property relationships.

Centering of the x-variables is required for the scaling methods applied here. The centered value of an x-variable is $x_C = x - c$ with the central value c of the variable given by the classical arithmetic mean, c_{MEAN}, or by the robust estimation median, c_{MED}. Note that the central value may be misleading for asymmetrically distributed variables; for instance a variable with more than 50% zeros gives a median of zero, as for instance is the case for 44 molecular descriptors in the used data set PAC-RI (Section 2.3).

Autoscaling is widely applied in chemometrics with the scaled variable given by $x_{AUTOSCALED} = x_C/s$. The spread (dispersion) measure s is typically the classical empirical standard deviation, s_{SD}, of variable x. Alternatively, a robust measure can be used, for instance derived from the interquartile distance, IQR as $s_{IQR} = 0.7413$ IQR (see Appendix 1). Autoscaling eliminates the influence of different units of the variables; autoscaled variables have all unit variance (unit variance scaling) and are in this sense equal important for model building. A disadvantage of autoscaling is a blow-up of variables with small values possibly originating from noise.

Pareto scaling is similar to autoscaling, however, uses the square root of the spread measure as scaling parameter with the scaled variable defined as $x_{PARETO} = x_C/s^P$, with $P = 0.5$. Thus, the scaling effect is weaker than with autoscaling, noise is less amplified, and variables with a high original variance retain part of their higher importance for the model. For a robust Pareto scaling x_C may be calculated as x_{MED} instead of x_{MEAN}, and s may be s_{IQR} instead of s_{SD}. Pareto scaling is widely used in biomarker identification [1, 10, 11], multivariate classification of metabolomics data [12], and has been proposed for exploratory data analysis [13, 14]. Pareto scaling was established 1993 by Svante Wold et al. [15] and named in honor of the scientist and economist Vilfredo Pareto (1848–1923), as reported elsewhere [16–18].

Adjusted Pareto scaling is proposed here as a generalization of the standard Pareto scaling by systematically varying the Pareto exponent P between 0 and 1 as discussed by I. Noda [18]. $P = 0$ means no scaling, $P = 0.5$ is standard Pareto scaling, and $P = 1$ is autoscaling. Here, we vary P between 0 and 1 in steps of 0.1 for finding an optimum value for P.

Range scaling relates the centered value x_C to the range d of the variable as $x_{RANGE} = x_C/d$; with the range defined as the difference between appropriate high and low values of the x-variable (x_{HIGH}, x_{LOW}) [19]. Using the minimum and maximum of x for the range borders is sensitive to outliers. Here, a specific low quantile q_{MIN} of the x-variable is used for x_{LOW} and $1 - q_{MIN}$ for x_{HIGH}. An optimum range can be found by varying q_{MIN} between for instance 0 and 0.1 in steps of 0.01; a robust version of range scaling uses the median for x_C. A disadvantage of range scaling is a blow-up of variables with small values possibly originating from noise.

Vast scaling [20] (variable stability scaling) is an extension of autoscaling by using the relative standard deviation (coefficient of variation, ratio of spread measure and central value) of the variable, defined as $c_V = s/c$. The scaled variable is $x_{VAST} = x_C/(s.\ c_V)$; again the robust estimations c_{MED} and s_{IQR} can be used for central value and spread, respectively. Variables with a low c_V are considered more "stable" and obtain an increased importance by vast scaling.

Level scaling relates the centered value x_C to the central value c of the variable as $x_{LEVEL} = x_C/c$; and thus, x_{LEVEL} is, in contrary to the other scaling methods used here, not based on a spread measure but on a location (size) measure [9]. Because the method is sensitive to outliers, the median is preferably used as central value; however, a very asymmetric distribution of the data may cause a zero median making this approach not applicable. Level scaling is recommended for searching variables possessing a high importance for the model, for instance biomarkers [1].

2.2 | Model Creation and Evaluation

Multiple linear calibration models for the prediction of a dependent variable y from a set of m independent variables x_1 to x_m are created here by partial least-squares (PLS) regression [8] as implemented in the R software package *pls*, using the function *plsr* [5, 6]. The applied strategy repeated double cross validation [2] (rdCV) separates the estimation of the optimum model complexity (optimum number of PLS components, A_{OPT}), from the estimation of the prediction performance for new cases.

The performance criterion used is the standard deviation of prediction errors for test set objects, usually called standard error of prediction (SEP). Because the prediction errors are often approximately normal distributed, the range ± 2 SEP estimates a 95% confidence interval for predicted values of y.

The parameters of the rdCV used are as follows: The number of segments for the outer CV loop (split into calibration sets and test sets) is $z_{TEST}=3$, of the inner CV loop (estimation of A_{OPT} for a training set) is $z_{CALIB}=7$. For PLS we use the maximum number of components min(15, m). The number of repetitions is $n_{REP}=50$. We obtain $n_{REP}\times z_{TEST}=150$ estimations of the optimum number of PLS components; the most frequent value, A_{OPT}, is used for calculating a SEP value for each repetition (each from n prediction errors obtained from objects in test set). Boxplots of the obtained 50 SEP values allow a comparison and evaluation of the applied variable scalings.

2.3 | Data

Three data sets have been used for comparing the scaling methods.

Data set HEAT with $\boldsymbol{X}(122\times13)$ is from biomass technology. A set of $n=122$ biomass samples with different origin (e.g., wood, grass, rye) is characterized by the contents of the elements C, H, and N (mass %), giving three basic x-variables. Further x-variables are mathematically derived as squared terms, cross terms, and logarithms, resulting in a total of $m=13$ x-variables [8, 21]. The property y to be modeled is the heating value of the biomass samples (HHV, higher heating value, enthalpy of complete combustion; determined by bomb calorimetry, range 15,719–25,948 kJ/kg, standard deviation 1415 kJ/kg).

Data set GLU-NIR with $\boldsymbol{X}(166\times221)$ is from analytical chemistry for bioethanol fermentation experiments [22]. A set of $n=166$ cereal samples (wheat, rye, corn, barley, and triticale flours) were fermented and NIR absorbance spectra were measured on centrifuged, clear mash samples. The first derivative of the spectral data is used (1100–2300 nm, 5 nm intervals, Savitzky–Golay smoothing and differentiation with a second order polynomial and a window of seven data points); giving 235 x-variables; $m=221$ remain after cleaning as described below. The property y to be modeled is the concentration of glucose in the mash samples, determined by the reference method HPLC (range 0.32–54.4 g/L, standard deviation 14.2 g/L).

Data set PAC-RI with $\boldsymbol{X}(209\times2290)$ is from modeling the relationship between chemical structure data and a physical property (QSPR, quantitative structure–property relationship [23]) as used before [24]. For a set of $n=209$ polycyclic aromatic compounds (PAC, molecular formulae range C_{8-24} $H_{6-24}N_{0-2}$ $O_{0-2}S_{0-2}$) a set of 2661 molecular descriptors is calculated by the software Dragon [25, 26] using approximated 3D-structures with explicit H-atoms; $m=2290$ variables remain after cleaning as described below. The property y to be modeled is a gas chromatographic retention index, experimentally determined by Lee et al. [27] (range 197.0–503.9, standard deviation 80.8). This index is based on the reference values 200, 300, 400, and 500 for the compounds naphthalene, phenanthrene, chrysene, and picene containing 2, 3, 4, and 5 condensed rings in the chemical structure, respectively.

A cleaning procedure with variable eliminations is performed for the data sets GLU-NIR and PAC-RI with the aim to avoid arithmetic errors in some scaling methods: (1) Elimination of variables with spread measure $s_{IQR}<s_{LOW}$. The cutoff s_{LOW} is obtained from preliminary tests using a value close to the 0.05 quantile of the s_{IQR} values of the variables; in particular $s_{LOW}=0.0003$ for the GLU-NIR data, and 0.02 for the PAC-RI data. (2) Elimination of variables with less than 10 different values. (3) Elimination of almost constant variables - that are variables being constant except of a maximum of 10 values.

3 | Results

The influence of variable scaling on the performance of multivariate calibration models is compared for 40 scaling methods (see Section 2.1) applied to three data sets (see Section 2.3) with the scaling parameters and results summarized in Table 1. The scaling methods are from six groups. Group P consists of 11 versions of adjusted Pareto scaling with the Pareto exponents $P=0$, 0.1, 0.2, ..., 1, and central value and spread given by the classical estimations c_{MEAN} and s_{SD}. Group Q is as group P, using the robust estimations c_{MED} and s_{IQR}. Group R consists of seven versions of range scaling with the central value c_{MEAN} and applying the low range border x_{LOW} as the quantiles 0, 0.01, 0.02, 0.03, 0.05, 0.07 and 0.1. Group S is as group R using the robust estimations c_{MED} for the central value. Group V contains a classical version of vast scaling, based on c_{MEAN} and s_{SD}, and a robust version based on c_{MED} and s_{IQR}. Group L contains a classical version of level scaling, based on c_{MEAN}, and a robust version based on c_{MED}.

Table 1 contains for each scaling method the applied parameters and the results from rdCV, namely SEP (median of 50 repetitions, in units of y), and A_{OPT}, the estimated final optimum number of PLS components. Note that the robust version of level scaling was not applicable to data set PAC-RI, because the median of several variables (given by molecular descriptors) is zero and level scaling requires division by the central value.

Figure 1 presents the obtained SEP values in boxplots, each for one scaling and originating from $n_{REP}=50$ repetitions of rdCV; the sequence of the 40 boxplots from left to right corresponds to the rows in Table 1. A visual interpretation is summarized as follows: (1) The method of scaling and the applied parameter influence the prediction performance (SEP); however, typically only moderately. (2) The variation of SEP caused by random splits of the objects in cross validation is often in the same range as for varying the scaling method or parameter. (3) The influence of scaling method and parameter depends on the data set; however, no obvious relation between data structure and recommended scaling appears for the three data sets. (4) For these data sets, robust methods achieved similar results as the corresponding classical approaches; probably, the used data sets do not contain many, severe outliers. (5) For obtaining a calibration model with high prediction performance, an exhaustive search is recommended by applying several scaling methods with varying parameters, and selecting the scaling method giving low SEP values preferably with a low variation.

More specific, adjusted Pareto scaling is clearly best for data set HEAT with $P=0.9$ to 1 (autoscaling), is best for data set NIR-GLU with $P=0.4$ to 0.6 (only a minor effect compared to

TABLE 1 | Scaling methods applied and results for data sets HEAT, GLU-NIR, and PAC-RI.

						HEAT		NIR-GLU		PAC-RI	
Id	**Group**	**Code**	**Robust**	**P**	**q_MIN**	**SEP**	**Aopt**	**SEP**	**Aopt**	**SEP**	**Aopt**
1	P	P0.0	N	0.0	NA	417.95	1	7.16	9	9.27	11
2	P	P0.1	N	0.1	NA	416.55	1	7.04	9	8.81	12
3	P	P0.2	N	0.2	NA	414.65	1	7.02	9	8.68	12
4	P	P0.3	N	0.3	NA	411.50	1	6.97	9	8.43	12
5	P	P0.4	N	0.4	NA	408.27	1	6.93	9	8.31	12
6	P	P0.5	N	0.5	NA	406.28	1	6.91	9	9.07	8
7	P	P0.6	N	0.6	NA	410.21	1	6.89	9	8.88	8
8	P	P0.7	N	0.7	NA	407.37	2	6.88	9	9.14	7
9	P	P0.8	N	0.8	NA	421.71	2	7.16	8	9.14	6
10	P	P0.9	N	0.9	NA	385.54	5	7.06	8	9.03	6
11	P	P1.0	N	1.0	NA	387.28	5	7.07	8	9.07	6
12	Q	Q0.0	Y	0.0	NA	417.95	1	7.16	9	9.27	11
13	Q	Q0.1	Y	0.1	NA	416.82	1	7.08	9	9.90	9
14	Q	Q0.2	Y	0.2	NA	415.34	1	7.06	9	8.82	12
15	Q	Q0.3	Y	0.3	NA	413.40	1	7.02	9	8.55	12
16	Q	Q0.4	Y	0.4	NA	410.49	1	7.03	9	8.79	11
17	Q	Q0.5	Y	0.5	NA	408.63	1	7.03	9	8.80	11
18	Q	Q0.6	Y	0.6	NA	408.95	1	7.04	9	9.21	9
19	Q	Q0.7	Y	0.7	NA	419.19	1	7.02	9	9.04	9
20	Q	Q0.8	Y	0.8	NA	408.29	2	7.26	8	8.61	9
21	Q	Q0.9	Y	0.9	NA	388.87	5	7.25	8	8.35	9
22	Q	Q1.0	Y	1.0	NA	390.26	5	7.28	8	8.36	9
23	R	R 0	N	NA	0.00	382.54	6	6.23	11	8.59	7
24	R	R 1	N	NA	0.01	387.86	5	6.34	11	9.14	6
25	R	R 2	N	NA	0.02	387.73	5	6.37	11	9.77	5
26	R	R 3	N	NA	0.03	387.02	5	6.43	11	9.40	6
27	R	R 5	N	NA	0.05	384.40	5	6.35	12	9.48	6
28	R	R 7	N	NA	0.07	382.74	5	6.42	12	8.87	6
29	R	R10	N	NA	0.10	385.17	5	6.49	12	8.40	6
30	S	S 0	Y	NA	0.00	382.54	6	6.23	11	8.59	7
31	S	S 1	Y	NA	0.01	387.86	5	6.34	11	9.14	6
32	S	S 2	Y	NA	0.02	387.73	5	6.37	11	9.77	5
33	S	S 3	Y	NA	0.03	387.02	5	6.43	11	9.40	6
34	S	S 5	Y	NA	0.05	384.40	5	6.35	12	9.48	6
35	S	S 7	Y	NA	0.07	382.74	5	6.42	12	8.87	6
36	S	S10	Y	NA	0.10	385.17	5	6.49	12	8.40	6
37	V	Vc	N	NA	NA	379.89	3	8.20	5	8.29	14

(Continues)

TABLE 1 | (Continued)

Id	Group	Code	Robust	P	q_MIN	HEAT		NIR-GLU		PAC-RI	
						SEP	Aopt	SEP	Aopt	SEP	Aopt
38	V	Vr	Y	NA	NA	371.84	3	8.20	6	10.43	9
39	L	Lc	N	NA	NA	411.02	4	6.80	14	15.58	7
40	L	Lr	Y	NA	NA	404.87	5	6.88	14	NA	NA

Note: Column group: P, adjusted Pareto scaling; Q, robust adjusted Pareto scaling; R, range scaling; S, robust range scaling; V, vast scaling; L, level scaling. Column code: Names for boxplots in Figure 1A–C. Column robust: N, classical (arithmetic mean and standard deviation); Y, robust (median and spread measure based on interquartile range). Column P: Pareto exponent. Column q_MIN: quantile for low border in range scaling. Column SEP: median of the obtained SEP values for 50 repetitions. Column Aopt: optimum number of PLS components.
Abbreviation: NA, not applicable/available.

no scaling or autoscaling), and for data set PAC-RI is clearly best with $P = 0.3$ to 0.4. The often a priori claimed high performance of the standard Pareto scaling with $P = 0.5$ has to be questioned.

With range scaling the low border was varied between quantile 0 and 0.1 with no remarkable influence on the SEP values for data sets HEAT and NIR-GLU. However, for the HEAT data set, range scaling is slightly better than adjusted Pareto scaling. For data set NIR-GLU, range scaling is clearly better than Pareto scaling. For data set PAC-RI, a low border at quantile 0 or 0.1 is better than at 0.02 to 0.05; however, the performance is similar as with Pareto scaling.

Vast scaling (robust version) is the best approach for data set HEAT, however, is the worst choice for data set NIR-GLU. For data set PAC-RI, vast scaling performs similar as adjusted Pareto scaling or range scaling.

Level scaling is clearly worst for data set PAC-RI (the robust version is not applicable as discussed above), and has no advantage for data sets HEAT or NIR-GLU.

4 | Discussion

The results presented in the previous section revealed clear differences in prediction performance, depending on the scaling method used. However, there is no clear winner for all data sets, and it might depend on the characteristics of a data set which method leads to better or worse results. It can be assumed that the number of variables plays a certain role, in particular the number of noise variables being not relevant for the prediction model. Also, outliers in single variables might be important for performance differences. In the following, we want to investigate in more detail the role of the scaling methods within PLS regression.

It is well known that the goal of PLS regression is to construct latent variables, and the corresponding scores are replacing the matrix of explanatory variables in the regression model. Consider a response y and explanatory variables $x_1, ..., x_m$ which are supposed to be mean centered. For constructing the first latent variable, the goal is to find a normed vector of weights (loadings) $\boldsymbol{w} = (w_1, ..., w_m)$, $\|\boldsymbol{w}\| = 1$, which maximizes the covariance $\mathrm{cov}(y, x_1 w_1 + ... + x_m w_m)$. The resulting linear combination of weights with the explanatory variables are the scores, and subsequent latent variables are found in the same way, but with an additional constraint, e.g., uncorrelatedness of the new scores to all previous ones. We have the identity $x_i w_i = (x_i/s_i)\, s_i w_i$, for $i = 1, ..., m$, for any non-zero values s_i. Suppose that s_i represents any of the proposed forms of scaling for the ith variable. For example, it could be the standard deviation for the ith variable, or the standard deviation with a power according to the adjusted Pareto transformation, or the (robust) range of the ith variable. Denote $\boldsymbol{v} = (v_i, ..., v_m) = (s_1 w_1, ..., s_m w_m)$. If $\boldsymbol{w}$ is the solution of the maximization problem for the original (mean centered) variables, $\boldsymbol{v}/\|\boldsymbol{v}\|$ is the solution maximizing the covariance of y with the transformed variables $x_i/\, s_i$, thus with the scores $(x_1/s_1)\, v_1/\|\boldsymbol{v}\| + ... + (x_m/s_m)\, v_m/\|\boldsymbol{v}\|$. It is crucial to see the role of $\|\boldsymbol{v}\|$, which adjusts the entries of the score vector by a quantity which originates from the transformations applied to all variables, since $\|\boldsymbol{v}\|^2 = (s_1 w_1)^2 + ... + (s_m w_m)^2$. This can also be seen as a kind of shrinkage term, similar to shrinkage estimators such as Ridge regression [28], where the squared Euclidean norm of the regression coefficients is constrained. However, the way how shrinkage is performed here depends not only on the "optimized" weights $\boldsymbol{w}$, but on individual adjustments of the components of $\boldsymbol{w}$, which are depending on the type of scaling used. Thus, scaling the individual variables is more flexible and adjustable to differences in variance/range/level of the single variables. However, there is no guarantee that this higher flexibility leads to better prediction models.

As an illustration, Figure 2 compares for the PAC-RI data set the regression coefficients from the PLS models based on different adjusted Pareto transformations (left) with those from Ridge regression by varying the Ridge parameter, leading to different values of the norm of the Ridge coefficients (right). The optimized solution is indicated by the vertical dashed line. For Ridge regression, a generalized cross-validation procedure was used for a grid of 100 values of the tuning parameter.

Although the procedures work differently, it is worthwhile to compare the regression coefficients. This is done in Figure 3, by computing the correlations between all regression coefficients from PLS and Ridge regression. The rows of this matrix correspond to the powers of the adjusted Pareto transformation (from 0 at top to 1 at the bottom), and the columns to the 100 values of the tuning parameter. We can see that the correlation gets very high for higher powers and less shrinkage in Ridge regression. In spite of this similarity of the coefficients, the final models still differ. Here, the optimized Ridge model leads to a SEP value of 7.61.

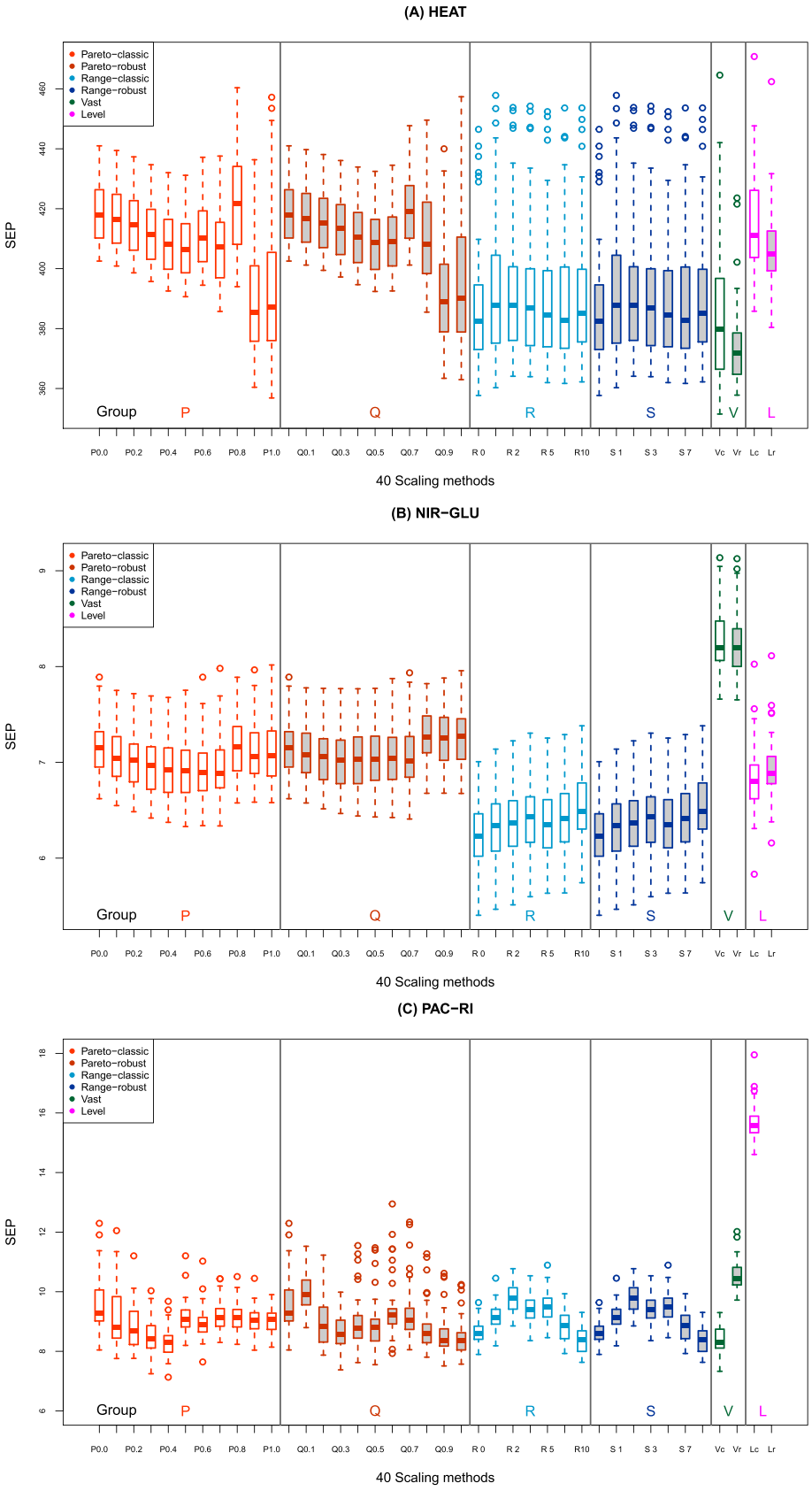

FIGURE 1 | Boxplots for 50 SEP values from the repetitions in rdCV for 40 scaling methods (see Section 2.1). Notations and numerical results are given in Table 1; data sets are described in Section 2.3.

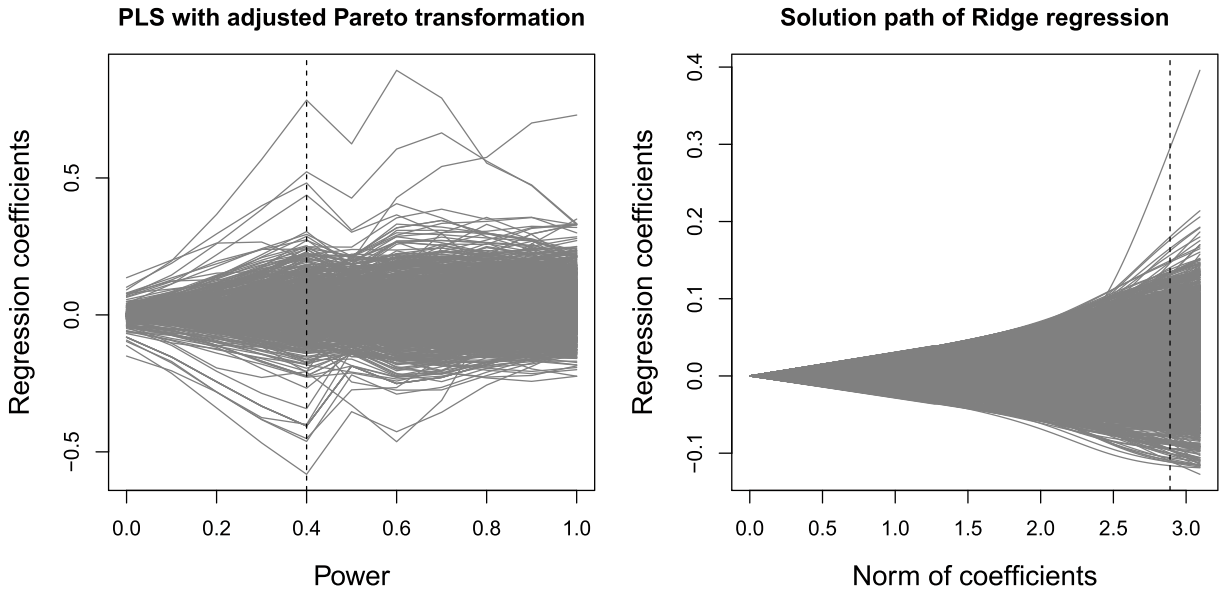

FIGURE 2 | Comparison of the regression coefficients for the PAC-RI data set, resulting from the optimized PLS models for the adjusted Pareto transformation (left) and from Ridge regression with varying tuning parameter (right). The best models are indicated by the vertical dashed line.

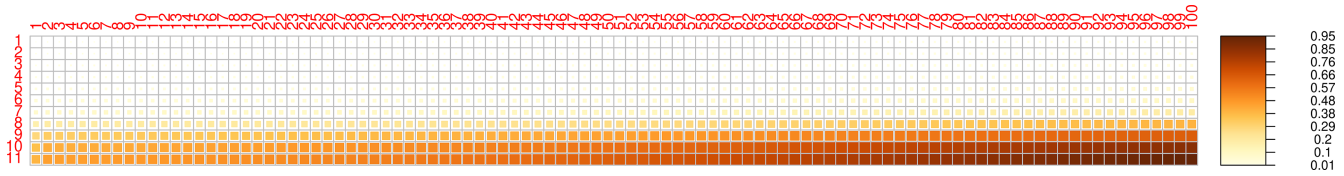

FIGURE 3 | Correlations between the regression coefficients of the PLS models from the adjusted Pareto transformed data and those from Ridge regression with varying tuning parameter. The rows are for the 11 values used for the Pareto exponent, *P*, from 0 on top to 1 at the bottom; the columns are for a gradually decreasing tuning parameter.

5 | Conclusions

Experiments based on three data sets with very different characteristics revealed surprisingly big differences in performance. Even more, there is no clear winner of a scaling method, and performance differences will not only be based on the dimensions of the data sets, but also on more specific characteristics such as the distributions of the variables, their associations, and their relationships to the response. As outlined in the previous section, scaling methods such as adjusted Pareto scaling follow the principle of a shrinkage method (specifically in the context of PLS regression), and provide higher flexibility with adjusting the specific variables within a prediction model.

More specific, it is recommended to optimize the Pareto exponent *P* by applying values from 0 (no scaling), via 0.5 (classical Pareto scaling) to 1 (autoscaling), preferably in steps of. 0.1. Depending on the data set, other scaling methods, based on the variable spread, like range scaling or vast scaling, or level scaling, may outperform optimized Pareto scaling.

A final recommendation is probably not to always use all possible scaling methods in a model evaluation, as this might be too time-consuming or even impractical. However, it can still be advisable to optimize a parameter, such as the power for adjusted Pareto scaling. If such an evaluation is carried out, we do not recommend to identify the best performing scaling method across different data sets [29] but only for one specific data set. Consideration of different variants of scaling allow for an empirical adjustment of the variable contributions to the calibration model.

The goal of this contribution is also to make aware that the performance of PLS models can heavily depend on the type of preprocessing, which here is done by the same procedure collectively to all variables, and not adjusted to individual variables. Besides the different types of scaling methods, there was also a distinction between those that employ robust estimators and those that work with classical counterparts. As robust estimation is done for the individual variables, this allows to adjust variable contributions according to their noise levels. For example, outliers in single variables can inflate the classical scale estimation, and the effect to variable scaling will differ from a robust scale estimator; however, the consequence of the performance difference to the PLS model is not clear in advance.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. J. Walach, P. Filzmoser, and K. Hron, "Comprehensive Analytical Chemistry. Data Analysis for Omics Sciences: Methods and Applications," in *Data Normalization and Scaling: Consequences for the Analysis in Omics Sciences*, eds. J. Jaumot, C. Bedia, and R. Tauler (Amsterdam, The Netherlands: Elsevier, 2018), 165–196.

2. P. Filzmoser, B. Liebmann, and K. Varmuza, "Repeated Double Cross Validation," *Journal of Chemometrics* 23 (2009): 160–171.

3. K. Varmuza and P. Filzmoser, "Repeated Double Cross Validation (rdCV) - A Strategy for Optimizing Empirical Multivariate Models, and for Comparing Their Prediction Performances," in *Current Applications of Chemometrics*, ed. M. Khanmohammadi (New York, NY, USA: Nova Science Publishers, 2015), 15–31.

4. R, A Language and Environment for Statistical Computing, R Development Core Team, Foundation for Statistical Computing (Vienna, Austria, 2023), http://www.r-project.org.

5. B. H. Mevik and R. Wehrens, "The pls Package: Principal Component and Partial Least Squares Regression in R," *Journal of Statistical Software* 18 (2007): 1–24.

6. R. Wehrens, *Chemometrics With R* (Heidelberg, Germany: Springer, 2011), 155f.

7. P. Filzmoser, and K. Varmuza, *R Package Chemometrics* (Vienna, Austria, 2010), http://cran.at.r-project.org/web/packages/chemometrics/index.html.

8. K. Varmuza and P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics* (Boca Raton, FL, USA: CRC Press, 2009).

9. R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data," *BMC Genomics* 7 (2006): 142.

10. R. Bujak, E. Daghir-Wojtkowiak, R. Kaliszan, and M. J. Markuszewski, "PLS-Based and Regularization-Based Methods for the Selection of Relevant Variables in Non-Targeted Metabolomics Data," *Frontiers in Molecular Biosciences* 3 (2016): 35.

11. C. Li, Z. Gao, B. Su, G. Xu, and X. Lin, "Data Analysis Methods for Defining Biomarkers From Omics Data," *Analytical and Bioanalytical Chemistry* 414 (2022): 235–250.

12. P. S. Gromski, Y. Xu, K. A. Hollywood, M. L. Turner, and R. Goodacre, "The Influence of Scaling Metabolomics Data on Model Classification Accuracy," *Metabolomics* 11 (2015): 684–695.

13. G. Ivosev, L. Burton, and R. Bonner, "Dimensionality Reduction and Visualization in Principal Component Analysis," *Analytical Chemistry* 80 (2008): 4933–4944.

14. A. L. Souza, S. G. Leos, J. Naozuka, P. R. Miranda-Correia, and P. V. Oliveira, "Exploring the Emission Intensities of ICP OES Aided by Chemometrics in the Geographical Discrimination of Mineral Waters," *Journal of Analytical Atomic Spectrometry* 26 (2011): 852–860.

15. S. Wold, E. Johansson, and M. Cocchi, "PLS - Partial Least Squares Projections to Latent Structures," in *3D QSAR in Drug Design. Theory, Methods, and Applications*, ed. H. Kubinyi (Leiden, The Netherlands: ESCOM Science Publishers, 1993).

16. L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström, and S. Wold, *Multi- and Megavariate Data Analysis. Basic Principles and Applications* (Umea, Sweden: Umetrics AB, 2006).

17. M. Mecozzi, M. Pietroletti, F. Oteri, and R. Di Mento, "Principal Component Analysis - Engineering Applications," in *Applications of PCA to the Monitoring of Hydrocarbon Content in Marine Sediments by Means of Gas Chromatographic Measurements*, ed. P. Sanguansat (Rijeka, Croatia: InTech, 2012), 65–82.

18. I. Noda, "Scaling Techniques to Enhance Two-Dimensional Correlation Spectra," *Journal of Molecular Structure* 883 (2008): 216–227.

19. A. K. Smilde, M. J. Van der Werf, S. Bijisma, B. J. van der Werff-van der Vat, and R. H. Jellema, "Fusion of Mass Spectrometry-Based Metabolomics Data," *Analytical Chemistry* 77 (2005): 6729–6736.

20. H. C. Keun, T. M. D. Ebbels, H. Antii, et al., "Improved Analysis of Multivariate Data by Variable Stability Scaling: Application to NMR-Based Metabolic Profiling," *Analytica Chimica Acta* 490 (2003): 265–276.

21. A. Friedl, E. Padouvas, H. Rotter, and K. Varmuza, "Prediction of Heating Values of Biomass Fuel From Elemental Composition," *Analytica Chimica Acta* 544 (2005): 191–198.

22. B. Liebmann, A. Friedl, and K. Varmuza, "Applicability of Near-Infrared Spectroscopy for Process Monitoring in Bioethanol Production," *Biochemical Engineering Journal* 52 (2010): 187–193.

23. T. Engel and J. Gasteiger, eds., *Chemoinformatics - Basic Concepts and Methods* (Weinheim, Germany: Wiley VCH, 2018).

24. K. Varmuza, P. Filzmoser, and M. Dehmer, "Multivariate Linear QSPR/QSAR Models: Rigorous Evaluation of Variable Selection for PLS," *Computational and Structural Biotechnology Journal* 5 (2013): e201302007.

25. Dragon, Software for Molecular Descriptor Calculation, Version 6.0, Talete s.r.l (Milan, Italy, 2010), www.talete.mi.it.

26. R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics (2 Volumes)* (Weinheim, Germany: Wiley-VCH, 2009).

27. M. L. Lee, D. L. Vassilaros, C. M. White, and M. Novotny, "Retention Indices for Programmed-Temperature Capillary-Column gas Chromatography of Polycyclic Aromatic Hydrocarbons," *Analytical Chemistry* 51 (1979): 768–773.

28. A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogoal Problems," *Technometrics* 12 (1970): 55–67.

29. K. Heberger, "Sum of Ranking Differences Compares Methods or Models Fairly," *Trends in Analytical Chemistry* 29 (2010): 101–109.

30. R. G. Brereton, *Chemometrics for Pattern Recognition* (Chichester, United Kingdom: Wiley, 2009).

31. T. Hastie, R. J. Tibshirani, and J. Friedman, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction* (New York: Springer, 2009).

32. J. N. Miller and J. C. Miller, *Statistics and Chemometrics for Analytical Chemistry* (Harlow, United Kingdom: Pearson Education Ltd, 2010).

33. B. G. M. Vandeginste, D. L. Massart, L. C. M. Buydens, S. De Jong, and J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part B* (Amsterdam, The Netherlands: Elsevier, 1998).

34. E. Gurian, A. Di Silvestre, E. Mitri, et al., "Repeated Double Cross-Validation Applied to the PCA-LDA Classification of SERS Spectra: A Case Study With Serum Samples From Hepatocellular Carcinoma Patients," *Analytical and Bioanalytical Chemistry* 413 (2021): 1303–1312.

35. K. Varmuza, "Chemoinformatics - Basic Concepts and Methods," in *Data Analysis and Data Handling (QSPR/QSAR) - Methods for Multivariate Data Analysis*, eds. T. Engel and J. Gasteiger (Weinheim, Germany: Wiley-VCH, 2018), 399–437.

36. K. Varmuza, P. Filzmoser, M. Hilchenbach, H. Krüger, and J. Silen, "KNN Classification — Evaluated by Repeated Double Cross Validation: Recognition of Minerals Relevant for Comet Dust," *Chemometrics and Intelligent Laboratory Systems* 138 (2014): 64–71.

37. S. Smit, H. C. J. Hoefsloot, and A. K. Smilde, "Statistical Data Processing in Clinical Proteomics," *Journal of Chromatography B* 866 (2008): 77–88.

38. S. Smit, M. J. van Breemen, H. C. J. Hoefsloot, A. K. Smilde, J. M. F. G. Aerts, and C. G. de Koster, "Assessing the Statistical Validity of Proteomics Based Biomarkers," *Analytica Chimica Acta* 592 (2007): 210–217.

39. S. J. Dixon, Y. Xu, R. G. Brereton, et al., "Pattern Recognition of Gas Chromatography Mass Spectrometry of Human Volatiles in Sweat to Distinguish the Sex of Subjects and Determine Potential Discriminatory

Marker Peaks," *Chemometrics and Intelligent Laboratory Systems* 87 (2007): 161–172.

40. M. Forina, S. Lanteri, R. Boggia, and E. Bertran, "Double Cross Full Validation," *Quimica Analitica* 12 (1993): 128–135.

41. O. Kvalheim, R. Arneberg, B. Grung, and T. Rajalahti, "Determination of Optimum Number of Components in Partial Least Squares Regression From Distributions of the Root-Mean-Squared Error Obtained by Monte Carlo Resampling," *Journal of Chemometrics* 32 (2018): 1–12.

Appendix 1: Symbols and glossary

rdCV	Repeated double cross validation.
$\boldsymbol{X(n\times m)}$	Multivariate data (n objects, m variables).
$\boldsymbol{y}$	Property to be modeled $y_1 \ldots y_n$. $\hat{y}$ is a predicted value.
$\boldsymbol{c}$	Central value for a x-variable; c_{MEAN}, arithmetic mean; c_{MED}, median; $x_C = x - c$ is a centered variable.
$\boldsymbol{s}$	Spread measure for a x-variable; s_{SD}, classical empirical standard deviation; $s_{IQR} = 0.7413$ IQR (IQR, interquartile range).
$\boldsymbol{c_V}$	Relative standard deviation (coefficient of variation) for a x-variable used in vast scaling with $c_V = s/c$.
$\boldsymbol{P}$	Pareto exponent in interval [0, 1] for adjusted Pareto scaling x_C/s^P; $P = 0$ for no scaling; $P = 0.5$ for standard Pareto scaling; $P = 1$ for autoscaling.
$\boldsymbol{x_{LOW}}$, $\boldsymbol{x_{HIGH}}$	Low and high border defining the range of variable x for range scaling; either the minimum, x_{MIN}, and the maximum, x_{MAX}, are used or the quantiles q_{MIN} (e.g., 0.05) and $q_{MAX} = 1 - q_{MIN}$.
$\boldsymbol{A}$	Number of PLS components; A_{MAX}, maximum number used in PLS; A_{OPT}, optimum number resulting from rdCV.
$\boldsymbol{z}$	Number of segments in rdCV; z_{TEST}, in outer CV (split into calibration and test set); z_{CALIB}, in inner CV (estimation of optimum number of PLS components in calibration set).
$\boldsymbol{n_{REP}}$	Number of repetitions in rdCV.
MSE	Mean squared error, used in the estimation of A_{OPT} for a calibration set. $MSE = (1/n_{MSE}) \boldsymbol{\Sigma} (y_i - \hat{y}_i)^2$, $i = 1, \ldots, n_{MSE}$; n_{MSE} is the number of objects in the calibration set.
SEP	Standard error of prediction (standard deviation of prediction errors for test set objects). The final SEP for a data set (scaling) is the median of n_{REP} SEP values obtained in rdCV. A 95% confidence interval for predicted values of y is estimated by the range ± 2 SEP.

Appendix 2: Summary of rdCV (Repeated Double Cross Validation)

The strategy rdCV is summarized as used here together with PLS regression for creating models for the prediction of a numerical property y from a set of m x-variables; data for n objects are $\boldsymbol{X}(n\times m)$ and $\boldsymbol{y}(n\times 1)$. The performance of a model is characterized by the standard error of prediction, SEP, equivalent to the standard deviation of prediction errors of test set objects. The repetitions applied in rdCV allow an estimation of the variability of SEP and of the optimum number of PLS components, A_{OPT}, as caused by random splits of the objects into calibration and test sets. Thus, a realistic comparison of the prediction performances is possible for models made from differently scaled variable sets. Definitions of symbols are in Appendix 1; basics of cross validation (CV) and partial least-squares regression (PLS) are described elsewhere [5, 6, 8, 30–33]. Development and evaluation of PLS models with rdCV was introduced by Filzmoser et al. [7]; details and applications are described elsewhere [3, 8, 24, 34–36]. Similar strategies have been proposed for instance for binary classification in proteomics [37, 38], for the discrimination of human sweat samples [39] as well as for principal component analysis [40].

The method description starts with double cross validation (dCV) and we separate it into two parts: In part A an optimum number of PLS components, A_{OPT}, is estimated; in part B the prediction performance for test set objects, SEP, is estimated for PLS models with A_{OPT} components. From dCV a single value for SEP results; however, with the typical rather small data sets in chemistry a single estimation of SEP may be misleading because of an unbalanced random split of the objects in CV. In repeated double cross validation (rdCV) the dCV strategy is repeated n_{REP} times giving n_{REP} estimations of SEP that can be represented by a box plot, well suited here for comparisons of variable scaling.

Part A of dCV estimates an optimum number of PLS-components and consists of two nested loops. Results about prediction errors obtained in Part A are not considered in the final evaluation of the model performance.

The **outer CV loop** of part A splits the n objects into z_{TEST} segments (here, 3); thus, the loop has z_{TEST} laps, each using a **test set** built by one segment, and a **calibration set** built by the others.

The **inner CV loop** of part A is applied to each calibration set, which is split into z_{CALIB} segments (here, 7); a **validation set** consists of one of these segments, and a **training set** contains the others. In each lap of the inner CV the objects of the current training set are used for making separate PLS models with 1 to A_{MAX} (here, 15) components. These models are separately applied to the corresponding validation set giving a mean squared error, MSE (see Appendix 1). After completing the inner CV loop we have a matrix **MSE** ($z_{CALIB}\times A_{MAX}$) with $z_{CALIB} = 7$ values of MSE for each for each number of PLS-components 1 ... A_{MAX}. Next we calculate the column means of **MSE** giving MSE_{MEAN} for 1 to A_{MAX} PLS components. Usually MSE_{MEAN} decreases with increasing A, and after a more or less distinct minimum increases because of overfitting. The global minimum $_{MIN}MSE_{MEAN}$ of MSE_{MEAN} at A_{MIN} PLS components, is a first indicator for the optimum number of PLS components; however, may be too high because of overfitting. The statistically based method *one standard error rule* [31] is preferred for finding an optimum number of PLS components, A_{CALIB}, for the calibration set. The strategy starts with A_{MIN} and considers that the corresponding MSE_{MEAN} has a variation that can be characterized by the standard deviation of the corresponding MSE values (standard deviation of means), defined as $s_{MEAN} = s_{MSE}/z_{CALIB}{}^{0.5}$, with s_{MSE} for the standard deviation of the z_{CALIB} values for MSE at A_{MIN}. Thus, for a conservative model avoiding overfitting, we allow an MSE up to the limit $_{MIN}MSE_{MEAN} + s_{MEAN}$. In the described procedure A_{CALIB} is the smallest A fulfilling this condition.

From dCV we obtain n_{TEST} estimations A_{CALIB} for the optimum number of PLS components; a single value A_{OPT} for the data set is selected here by choosing the most frequent value of A_{CALIB}. Another approach, using Monte Carlo techniques, has been suggested by Kvalheim et al. [41]

Part B of dCV estimates the prediction performance for PLS models with the fixed number A_{OPT} of PLS components. Here, the same CV loop with n_{TEST} segments for all n objects is used as in part A (another splitting into segments is optional). One segment forms a test set, the others a calibration set. From each calibration set a PLS model with A_{OPT} components is made and applied to the corresponding test set. After completion of this CV we have one test set-predicted value for each object. The resulting performance criterion SEP is the standard deviation of n test set prediction errors. The strategy dCV separates the estimation of the model complexity (here, the number of PLS components) from the estimation of the prediction performance. Only a single number for the performance is obtained, not allowing a reasonable comparison with other methods or data because of potentially highly unbalanced random splits in CV. Consequently, an extension to repeated double cross validation (rdCV) is recommended as follows.

Repeated double cross validation (rdCV) repeats dCV with n_{REP} (here, 50) different random splits of the objects into segments. Following the procedure of dCV we obtain from each repetition z_{TEST} estimations for A_{CALIB}, in total $n_{REP} \times z_{TEST}$ (here 150) and the most frequent value of them is used here as a single value for A_{OPT}, thus completing Part A for rdCV.

Part B of rdCV uses the fixed A_{OPT} for estimating a SEP in each repetition, and the median is used as a single, final value for the prediction performance. An alternative would be the standard deviation of all $n \times n_{REP}$ prediction errors for test set objects, all from PLS models with A_{OPT} components. A boxplot of the n_{REP} SEP-values characterizes their distribution and is used here for comparisons of the various scalings of $\boldsymbol{X}$. Note that in general the scalings give different values for A_{OPT}. The interquartile distance of the n_{REP} SEP-values may be used as a measure of the performance variation due to random splits of the objects into segments.