The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/174465

Please be advised that this information was generated on 2019-10-25 and may be subject to change.
Article 25fa pilot End User Agreement

This publication is distributed under the terms of Article 25fa of the Dutch Copyright Act (Auteurswet) with explicit consent by the author. Dutch law entitles the maker of a short scientific work funded either wholly or partially by Dutch public funds to make that work publicly available for no consideration following a reasonable period of time after the work was first published, provided that clear reference is made to the source of the first publication of the work.

This publication is distributed under The Association of Universities in the Netherlands (VSNU) ‘Article 25fa implementation’ pilot project. In this pilot research outputs of researchers employed by Dutch Universities that comply with the legal requirements of Article 25fa of the Dutch Copyright Act are distributed online and free of cost or other barriers in institutional repositories. Research outputs are distributed six months after their first online publication in the original published version and with proper attribution to the source of the original publication.

You are permitted to download and use the publication for personal purposes. All rights remain with the author(s) and/or copyrights owner(s) of this work. Any use of the publication other than authorised under this licence or copyright law is prohibited.

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please contact the Library through email: copyright@ubn.ru.nl, or send a letter to:

University Library
Radboud University
Copyright Information Point
PO Box 9100
6500 HA Nijmegen

You will be contacted as soon as possible.
Weight randomization test for the selection of the number of components in PLS models

Thanh Tran1,2 | Ewa Szymańska2,3 | Jan Gerretzen2,3,8 | Lutgarde Buydens2 | Nelson Lee Afanador2,4 | Lionel Blanchet5,6,7

1 Center for Mathematical Sciences, Merck, Sharp & Dohme, Oss, The Netherlands
2 Institute for Molecules and Materials, Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
3 TI-COAST, Science Park 904, 1098 XH Amsterdam, The Netherlands
4 Center for Mathematical Sciences, Merck, Sharp & Dohme, West Point, PA, USA
5 Department of Pharmacology and Toxicology, School of Nutrition, Toxicology and Translational Research in Metabolism (NUTRIM), Maastricht University Medical Center, Maastricht, The Netherlands
6 Thayer School of Engineering, Dartmouth College, Hanover, NH, USA
7 Top Institute Food and Nutrition (TIFN), Wageningen, The Netherlands

Correspondence
Thanh Tran, Center for Mathematical Sciences, Merck, Sharp & Dohme, Oss, The Netherlands.
Email: thanh.tran@merck.com, thanh.tran@science.ru.nl

The selection of the optimal number of components remains a difficult but essential task in partial least squares (PLS). Randomization tests have the advantage of being automatic and they make use of the entire dataset, in contrary with the widely used cross-validation approaches. Partial least squares modeling may include component(s) with a large amount of irrelevant data variation, and this might affect the model, depending on the assigned y-loading (which is the regression coefficient in the latent domain). This has recently been indicated by us in the basic sequence framework with respect to the underlying theory of the PLS algorithm and presented to the chemometrics society. We will show in this work that this irrelevant data variation is the root cause of the difficulty in current methods for selecting the optimal number of components. For randomization tests, PLS models with nonsignificant components may result in false positive tests because of the incorrect assumption that “the components enter the model in a natural order”.

In this work, we introduce a new randomization test, weight randomization test, selection of the optimal number of components in PLS in light of the underlying theory of the PLS algorithm. In the proposed method the null distribution is well characterized and efficiently determined taking into account a newly defined model quality metric: the number of consecutive non-significant components (CNC). We illustrate the effectiveness of weight randomization test in optimization of preprocessing as well as in classification models, where results are compared with the double cross-validation procedure for the latter. This is an important step towards the full automation of PLS model development and routine updates.

KEYWORDS
number of components, partial least squares, randomization test

1 | INTRODUCTION

Partial least squares (PLS) is arguably one of the most popular algorithms in the chemometrics community because it rapidly became a reference solution for regression and classification problems. The popularity of this algorithm spread to other scientific fields such as sensory analysis, econometrics, and resulted in many successful applications in chemical, petrochemical, pharmaceutical, food processing, and industrial processes. Practitioners across multiple disciplines are familiar not only with the benefits of PLS but also with one of its most troublesome features: the need to decide how many components (#Comp) should be retained in the model. Our objective here is to propose a statistical evaluation that aids in deciding what the optimal #Comp is.
A high number of components will come with the risk of overfitting the model, thereby losing its generality, resulting in a loss of predictive power as measured via an independent test set. On the other hand, underestimating the number of components leads to lower predictive power. Procedures such as cross-validation (CV), leverage correction, and evaluation of the size of the (PLS) eigenvalues are currently in use for determining the optimal #Comp. Until now, the selection of the #Comp by CV has been the most widespread practice in chemometrics. During a CV procedure, a subset of samples (k-folds) is iteratively removed from the dataset and subsequently used to assess the performance of PLS models with different #Comp via the root mean square error of CV (RMSECV). Besides the difficulty on deciding the number of k-folds, a robust cutoff value in a performance measure curve is hard to obtain. It is often not at the RMSECV minimum, and there is no consensus of approaches for defining a plateau area within the RMSECV curve for the final #Comp. Hence, alternative methods are in high demand by the PLS user community.

Contrary to CV approaches, randomization tests (RTs), also known as permutation tests, have the advantage of being an automatic method that makes use of the entire dataset. Van der Voet and Wiklund et al proposed RTs to assess the statistical significance of each individual component in PLS. This was done by assessing each component against its corresponding permutation null distribution and using this to test whether the observed component could be reasonably observed under the null hypothesis. Unfortunately, currently used RT methods tend to be sensitive to irrelevant data variation because of, eg, improper data preprocessing. In this case, a component becomes nonsignificant and RT fails to report an optimal #Comp.

We will extend this discussion and explain the root of the difficulties observed in both CV and RT frameworks in light of the unified framework in Section 2. With this understanding, we propose a novel RT for the distribution of randomized model weight vector lengths, referenced in the present paper as the weight randomization test (WRT), which can be seen as an extension of Wiklund's method. The use of weight vector lengths in WRT instead of the covariance between scores and the response variable, as is done by Wiklund, enables a better understanding of the corresponding distribution and also enables enhanced practical implementation of the WRT method on a larger scale. A more extensive description of the WRT test for the selection of the number of components in PLS is given in Section 2.

In Sections 3 and 4, we will illustrate the use of WRT in 2 different application areas of PLS models: selection of optimal data preprocessing in regression models and model optimization in classification. In the regression example, the quality of the data preprocessing is assessed using an experimental spectroscopy dataset. It involves the selection of the most optimal “preprocessing strategy” from almost 5000 different reasonable preprocessing methods and their combinations (these are called preprocessing strategies). Model performance and model interpretation are discussed and compared with a CV strategy and #Comp selection proposed by Engel et al and van der Voet, respectively.

In the second application, WRT tests are compared with the double CV (2CV) procedure, the most common procedure for model optimization and validation, which combines model optimization in an inner loop with model validation in an outer loop. In PLS with WRT (or WRT-PLS) tests, the CV in the inner loop is replaced by WRT tests for the selection of #Comp. The CV in the outer CV loop for the estimation of model performance is kept untouched. Different PLS models with WRT tests and a 2CV procedure are built and evaluated for several 2-class datasets with different levels of known class differences. Model complexity, performance, and significance assessed with additional permutation tests, as earlier reported in 1 study, are included in the comparison.

2 | METHODOLOGY

2.1 | Current methods to select #Comp in PLS and their properties

According to 1 study, a PLS component may be dominated by irrelevant data variation because of “the rotation towards the first eigenvector property.” The rotation may not be relevant to the regression, such as in a situation when the data is influenced by improper data preprocessing or the presence of nonlinearity. In such cases, the rotation may bring irrelevant data variance into a PLS component. Thus, the selection of #Comp using PLS eigenvalues as in SIMCA becomes a nonrobust procedure, since the variance captured in the X-data may be driven by irrelevant data variance and hence impacting the PLS eigenvalue; this biases the PLS eigenvalue as a relative ratio of the captured variance in the X-data and the total X-variance at each component. In prediction, this rotation property may not create an issue for PLS regression when the component corresponds to a very small y-loading and hence its contribution to the PLS regression is negligible. This is actually a positive property of PLS with regard to removal of irrelevant data variance. However, it can result in a nonrobust PLS model in cases where there is a high irrelevant data variation coupled with a high y-loading.

In CV, the “optimal” #Comp in PLS regression is investigated on the basis of the RMSECV curve. When an RMSECV minimum exists, one can use the corresponding component as optimal #Comp, or a backward penalty approach, which starts at #Comp corresponding to the minimum RMSECV and from there removes components as long as the corresponding RMSECV is not larger than the predefined penalty (%RMSECV). One subjectively defines
the penalty %RMSECV value, but this value may not statistically define the natural plateau period as illustrated in Figure 1 at #Comp = 5 (see RMSECV error curve). Additionally, a minimum in RMSECV may not exist at an acceptable #Comp or the curve may vary for different CV-subset samples in k-folds potentially leading to nonrobust selection and suboptimal model performance.28

In the example given in Figure 1 (see RMSECV curve), minimum RMSECV is reached at #Comp = 7. However, RMSECV values from #Comps 4 to 7 define a plateau period and one can select #Comp = 4 as a better choice for the final #Comp. Note that we also have a local plateau area in between #Comps 2 and 3. These observations are purely empirical and subjective to the operator. Objective statistical testing is therefore highly desired here.

In significance tests by way of RTs, the methods in general can be traced back to both Fisher29 and Pitman,30 but it is only recently, with the advent of high-performance computing, that permutation tests have moved beyond theory and into practice.31 Van der Voet21 and Wiklund et al14 proposed the application of an RT for selection of #Comp in PLS. The current RTs assess for each component whether it is significantly different compared to any random component with respect to a regression of the response variable. These tests continue until the first nonsignificant component is detected.

In an RT, there is no equivalent curve to the root mean square error curve as used in CV. The test generates a list of significant components as a binary array such as [1 1 0 1 0 0 0 0 0] for the example in Figure 1 (1, a significant component, is represented by a filled square ■ and 0, a nonsignificant component, by a nonfilled square □). In this case, #Comp 3 may be driven by a dominant X-data variation, which is irrelevant to the regression of y. Indeed, the predictive performance is not improved at this component as shown by the RMSECV curve. Hence, in case of RT, this component is not significantly different from random components and the current RT methods are sensitive to the first nonsignificant component as in this case. Wiklund et al acknowledge this issue and associate it to the case when improper preprocessing is used.14 However, preprocessing is not necessarily able to remove all irrelevant information and nonsignificant components in the RT can appear in any component with the presence of high irrelevant data variation (ie, higher than relevant information), including nonlinearities. In the situation in Figure 1, current RT methods result in #Comp = 2, which is similar to the local optimum obtained by CV because of the local plateau situation.

In the example presented in Figure 1, the PLS model with #Comp = 4 may be acceptable under the condition that the irrelevant data variation captured in component 3 will not deviate in the future from the structure represented in the training set. The irrelevant variation included in the PLS model should be reproducible, which is hardly possible in real-world applications. Including nonsignificant components generally reduces model robustness as well. For this reason, we propose in this work a new metric for the quality of PLS models: the number of consecutive nonsignificant components (CNCs) in the final PLS model. It is a crucial parameter in the weight RT introduced in the next section.

The CNC is proportional to the irrelevant data variation captured in the PLS model, which may impact model robustness. A higher CNC results in a model capturing more irrelevant data variation and hence less robust. In this way, CNC defines a suitable model quality. This parameter can even be defined a priori and may depend only on the application and less on specific data at hand.

### 2.2 Weight randomization test

Figure 2 provides a general scheme for the WRT for a single PLS component, a. This construction is common for RTs for a single component in PLS. The response variable is permuted into M randomized response variables \( \tilde{y}_1, \ldots, \tilde{y}_M \) from which the null distribution can be generated according to a transformation function. To do this, Van der Voet21 used the model residuals, while Wiklund et al in 1 study14 took the covariance between the score and response of the randomized response models for the null distribution. Finally, a one-sided test is performed by comparing the observed value from the original PLS model (ie, the model residual in Van der Voet's method and the correlation value in Wiklund's method) with a cutoff value corresponding to a predefined significance level \( \alpha \), eg, 5%. The cutoff value is estimated by different ways: a one-sided \( F \) test was used in Van der Voet's method, while an empirical inverse Gaussian distribution was fitted and used in Wiklund's method. In all cases, an exact test can be obtained by using the upper quantile from the permutation null distribution that corresponds with the desired
alpha level. A nice property of using the quantiles is that they are transformation invariant, meaning that the observed quantile in the actual permutation null distribution will remain the same after any monotonic transformation of the permutation null distribution. However, this traditional approach requires a larger number of permutations and sorting to better define the end points of the permutation null distribution, hence, it requires higher computation time.

In WRT, the length of PLS weight vector is used for the null distribution and a statistical test is applied to the observed weight length against the permutation weight lengths null distribution (Figure 2). It can be shown that the length of the PLS weights (||w||) is proportional to the covariance between the score vectors and response variable, \((t^\prime y)/N\), with N number of samples, used in the work of Wiklund et al\(^{14}\) as follows, with scores \(t\), and response variable \(y\):

\[
||w|| = \frac{||w||^2}{||w||} = w^\prime w \left( \frac{1}{||w||} \right) = \left( \frac{1}{||w||} \right) (w^\prime x) y = \left( \frac{xw}{||w||} \right) y = t^\prime y. \tag{1}
\]

The WRT tests use the null distribution of weight lengths to assess if a PLS component is statistically significant at a predefined significance level \(\alpha\), and it is crucial to properly estimate a cutoff value corresponding to a significance level \(\alpha\) (red line in Figure 2).

In RTs, generating the sampling distribution under the null distribution is often computationally intensive. In earlier works,\(^{14,21}\) the PLS model is refit for thousands of permutations, which involve the calculation of all essential PLS metrics. In this work, the thousands of permutation weights vectors, \(\tilde{w}_i\), can be described by 1 matrix calculation, which is much less computationally expensive:

\[
[\tilde{w}_1, \ldots, \tilde{w}_M] = X_{\alpha}[\tilde{y}_1, \ldots, \tilde{y}_M]. \tag{2}
\]

As mentioned earlier, RTs should not necessarily stop at the first nonsignificant component and the RT should “not assume the components to enter the model in a natural order.” The PLS may allow a number of nonsignificant components, which is related to the model quality aspect as discussed earlier.

Regarding the input parameters of WRT, besides a predefined significance level \(\alpha\), eg, 5%, WRT requires as a maximum of \(k\)-CNC for the lowest acceptable quality of PLS model, for example \(k\)-CNC = 3 components. The final component is the last significant component before \(k\)-CNC. This parameter represents PLS model quality since a model with more \(k\)-CNC brings more irrelevant data variation into the model. In the example in Figure 1, the #Comp is 4 after accepting 1 nonsignificant component (ie, no. 3) and before 3-CNC.

With regard to the limitations of WRT-PLS, since the method involves permutation of the response variable, extreme outliers in the X-matrix might affect the results of the permutation test. This impact is not unique to the method being presented here, and similarly to what would be done in a traditional analysis setting, it is essential to inspect potential abnormalities in the X-data. Finally, when a parametric test is desirable, in lieu of more computationally expensive permutation tests to better define the distribution end points, one could consider a lognormal transformation of the permutation distribution. This well-understood transformation is known.

---

\(\text{FIGURE 2} \quad \text{Illustration of WRT: a randomization scheme, permutation distribution and 1-tailed test for each component at the 5\% significance level. WRT, weight randomization test}\)
for its variance stabilizing properties resulting in approximately normal distributions for data that tends to be skewed to the right. This is the approach that we demonstrate in the subsequent sections and, as we show in this paper on various applications, it supports the reasonableness of this approach in parametrically testing for significant components. Moreover, our simulation study demonstrates that it is indeed a challenging task to decide which distribution is the most suited for our purpose, and further research into this area may be warranted. As an alternative, when it is computationally feasible with a single use of a PLS model, the exact test nonparametric method discussed earlier based on a permutation distribution quantile can be confidently used.

### 2.3 Distributions of weight lengths in WRT

When an exact test via quantiles is not preferable because of computation time limitations, one can parametrically estimate the null distribution via the use of some theoretical distribution, eg, the F-distribution in Van der Voet's method or the inverse Gaussian distribution in Wiklund's method. As stated earlier, our experimental results, coupled with what we subsequently present regarding the distribution of weight vectors, supports the use of a lognormal distribution as a simple, yet robust, transformation for the observed permutation distribution.

Within the context of this work, the projection of a random vector $y$ onto $X$ creates weight vectors that follow a multivariate normal distribution. In the case of uncorrelated weight vectors, the sum-of-squares of these vectors follow a chi-square distribution, and its square root (L-2 norm) for bivariate variables follows a Rayleigh distribution, both of which are known to be positively skewed when the centrality parameter is under the null. In the case of correlated weight vectors, the null distribution $\text{diag} \left( \bar{W}^T C^{-1} \bar{W} \right)$, with a nonsingular (regular) covariance matrix $C = \bar{W}\bar{W}^T$, follows a chi-square distribution. Note that for uncorrelated weights, it becomes the sum-of-squares of the weight vectors with the identity matrix $C$. Although weight vectors belong to the Krylov sequence, they do not form a Krylov subspace and hence the covariance matrix $C$ can be either singular or nonsingular. In the latter case, matrix $C$ is not invertible hence an analytical solution for the null distribution by means of chi-square distribution is not applicable and another approximation approach is needed. Alternatively, the weight lengths in bivariate variables are known to follow a Nakagami-m distribution, and it has been used in various applications, where the Rayleigh distribution can be considered a special case of the Nakagami-m distribution.

In the case of higher dimensional data, without a closed analytical form for the null distribution of weight lengths, either empirical approaches using inverse Gaussian or lognormal distributions are often used to model the null distribution. These 2 empirical approaches are an alternative to the chi-square, Rayleigh, or Nakagami distributions. Both the inverse Gaussian and the lognormal distributions belong to the exponential family of distributions and have shown good results in practice. We take the same approach in this work in implementing an empirical approach by acknowledging the positive skewness that is present in the aforementioned distributions, and we apply a natural log transformation to the resulting null distribution as obtained via WRT. This transformation can account for situations in which the distribution tail becomes less heavy for higher components with smaller eigenvalues, hence a reduced effect due to the Krylov rotation. In this instance, a lognormal distribution is a good estimation of the inverse Gaussian for the one-sided test, further supporting our use of the lognormal transformation.

For illustration purposes, Figure 3 presents the null distributions for simulated scenarios with 2 and 4 correlated $X$-predictor variables, respectively. In both cases, the simulation confirms that the diagonal elements of $\bar{W}^T C^{-1} \bar{W}$ follow a chi-square distribution as shown in the left panels of Figure 3 given invertible the covariance matrices, $C$. It also reconfirms that the Nakagami-m distribution works better for 2-dimensional data. With higher dimensional data, both the inverse Gaussian and lognormal distributions approximate the null distribution well, largely overlapping on the tail of the distribution.

### 2.4 The WRT-PLS implementation considerations

For implementation, the permutation test can be introduced at the same time with the calculation of the PLS weight vector which leads to the WRT-PLS algorithm given in pseudo code in Table 1. The Matlab script is given in the appendix, and the algorithm is available in the mvdalab R-package.

### 3 WEIGHT RANDOMIZATION TEST VS VAN DER VOET METHOD IN PREPROCESSING OPTIMIZATION

For the preprocessing example, we follow the approach taken by Engel et al. Herein, the authors identified the most commonly applied preprocessing methods for spectroscopic data and made almost 5000 reasonable combinations of them (called preprocessing strategies) as shown in in Figure 4. All those 5000 strategies were applied to a spectroscopic dataset and subsequently, PLS models were constructed to evaluate which of all strategies led to the highest prediction accuracy. However, this procedure is not recommended for preprocessing selection because it is very time consuming
(eg, each differently preprocessed dataset requires separate CV). With the novel WRT-PLS approach, this process will be less time consuming and we expect WRT-PLS to be able to select a more appropriate preprocessing strategy. In this work, the preprocessing quality is evaluated by both the root mean square error of prediction (RMSEP, ie, model performance based on an independent test set) and by evaluating the relevant variables using the significant multivariate correlation (sMC) measure,\(^ {45}\) (compared to the known true relevant variables, see Section 3.1). Note that the sMC method is sensitive to the quality of the PLS model with respect to the presence of irrelevant data and overfitting. As a consequence, an overfitted PLS model, ie, containing irrelevant variance, will lead to the selection of an inadequate subset of variables. For this reason, the combination of WRT-PLS and sMC is a good approach for assessing the

**TABLE 1** Pseudo code of WRT-PLS

*Input:* \(X\), and \(y\), Alpha (default = 0.05) and number of CNC (default = 3)

**Step 0. Initialization:** \(X_{\text{def}} = X\) and \(y_{\text{def}} = y\)

**Step 1. For each component \(a\)**

// Ordinary steps of NIPALS PLS algorithm
weight: \(v = X_{\text{def}}y_{\text{def}}\), unit variance normalized weight, \(w\)
score: \(t_a = X_{\text{def}}w_a\) and loading: \(p_a = X_{\text{def}}t_a / (t_a \cdot t_a)\)

// WRT for this component \(a\)
Calculate weight lengths, \(||w||\)
Generate and null distribution \(||w_1||, \ldots, ||w_M||\) from Equation 1 and Equation 2
Calculate one-sided upper-quantile or log-normal limit UPPLIM
if \(||w|| \geq UPPLIM\), Reject \(H_0\) hypothesis and component \(a\) is significant
Otherwise, Accept \(H_0\), component \(a\) is non-significant

**Step 2. IF** the number of consecutive non-significant components reaches CNC value, **stop** and report the final PLS significant component, **ELSE** go back to **step 1**.

**FIGURE 3** Distribution of \(\text{diag}(\hat{W}^T C^{-1} \hat{W})\) (left panels) and length of randomized weights (right panels), and fittings for 2 (top panels) and 4 (bottom panels) correlated \(X\)-predictor variables, respectively

**FIGURE 4** Schematic overview of method combinations for different preprocessing strategies
relevant variables in a PLS model of acceptable quality i.e. with less or no nonsignificant components.

The results of WRT-PLS are compared with those of CV-PLS. For selection of #Comp for CV-PLS, we have integrated an automated #Comp selection algorithm, initially developed by van der Voet.\(^2^1\) A 10-fold CV procedure was used to optimize #Comp for CV-PLS. The WRT-PLS was used with settings (5 CNC and \(\alpha = 0.01\)). A maximum number of 20 components were set for both CV-PLS and WRT-PLS.

### 3.1 | Data

A dataset about latex samples is used to illustrate the application of WRT-PLS for preprocessing selection. It consists of \(n = 196\) near infrared spectra of latex samples, recorded at \(p = 1037\) wavenumbers, measured under aqueous conditions. The spectra are measured in the wavenumber range 4000-12 000 cm\(^{-1}\) (Figure 5). The dataset contains 2 response variables, namely, the levels of styrene (S) and butyl acrylate (BA) in the latex samples. Both are measured in parts per million units. The vinylic bands for BA and S are found at 6160 and 6145 cm\(^{-1}\), respectively.\(^4^6,4^7\) The dataset was randomly split in 150 training samples and 46 test samples to test the predictive performance of the different preprocessing strategies.

### 3.2 | Results

We first discuss the results of using CV-PLS with the van der Voet method as #Comp selection to select an appropriate preprocessing strategy and assess the performance of a WRT-PLS model using the preprocessing strategy found using CV-PLS. Next, we discuss preprocessing selection using WRT-PLS and similarly the performance of a model on the basis of CV-PLS using the preprocessing strategy selected with WRT-PLS. This is illustrated on the basis of prediction of levels of styrene (S); the results for prediction of butyl acrylate (BA) can be found in Data S1.

#### 3.3 | Preprocessing selection based on CV-PLS

The left top panel of Figure 6 shows the model performance (ie, RMSEP) and complexity of the model (ie, #Comp) for each of the 5000 differently preprocessed datasets for prediction of S, on the basis of CV-PLS. The preprocessing strategy leading to the lowest RMSEP (ie, the red square at an RMSEP value of 1225) equals smoothing using the Savitzky-Golay algorithm (window width 11 px, 2nd-order polynomial) and level scaling.\(^4^8\) This is a considerable improvement over a CV-PLS model on the basis of the raw data, of which the performance is indicated by the black triangle.

If the obtained preprocessing strategy is indeed appropriate, there should be no or just a few nonsignificant #Comp in the model, and hence, there should not be a large difference in model performance and complexity compared with WRT-PLS on the same data. In this case, the result of WRT-PLS applied to data preprocessed with the obtained preprocessing strategy, shown in the left top panel of Figure 6 with the green dot, is a relatively large difference in both model performance and model complexity compared to CV-PLS. This is an indication that a CV-PLS model may not be optimal because of nonsignificant components, leading to overfitting or a nonrobust model. For prediction of BA, the differences are even larger (see Data S1).

A further investigation indicated that from the 13 components required in the WRT-PLS model, 6 are nonsignificant (components 2, 3, 7, 8, 9, and 10). This is a relatively high number, and the fact that components 2 and 3 are already nonsignificant also indicates that the selected preprocessing strategy is probably suboptimal.

A similar conclusion can be drawn on the basis of the important variables indicated by sMC as shown in the left center and bottom panels in Figure 6. The left center and the left bottom panels show the important variables for the CV-PLS and WRT-PLS models, on the basis of the selected preprocessing strategy by CV-PLS, respectively. These selected variables are clearly not inside the known relevant region. This mismatch indicates that the selected region.
preprocessing strategy is not appropriate and the optimal PLS model—according to the ordinary CV procedure—has led to a poor quality model, which still contains high irrelevant data variation because of an improper preprocessing strategy used.

3.4 | Preprocessing selection using WRT-PLS

The complete preprocessing selection procedure is repeated using WRT-PLS instead of CV-PLS. The right top panel of Figure 6 shows model performance and model complexity for WRT-PLS models built on an approximate of 5000 differently preprocessed datasets. In this case, the model performance and model complexity for the WRT-PLS model with the lowest RMSEP is indicated with a blue dot. As a result, the preprocessing strategy corresponding to this model is baseline correction using asymmetric least squares, followed by smoothing (window 11 px, 2nd order) and level scaling (RMSEP: 1354). This RMSEP value is in the same order of magnitude as the one found with the CV-PLS–based preprocessing strategies in the previous section. However, in this case, the model performances of the CV-PLS and WRT-PLS model with the same preprocessing strategy are very similar as indicated in the top-right panel by moving from the blue circle to the pink square (also for BA prediction—see Data S1).

From the 13-components WRT-PLS model, 3 are deemed nonsignificant (components 4, 7, and 8). This is an improvement over the model where a preprocessing strategy was selected using CV-PLS, both for the number of nonsignificant components (3 vs 6). This indicates that better quality models are obtained when using WRT-PLS, with less nonsignificant components. Hence, this explains that proper preprocessing found with WRT-PLS leads to less overfitting.
illustrated by the small difference in model performance and complexity between the CV-PLS and WRT-PLS model. Furthermore, we expect the true relevant variables to be selected by both models as is confirmed in the right center and bottom panels in Figure 6: The difference between CV-PLS and WRT-PLS is negligible for relevant variables, and they align with the true relevant variables. This confirms that application of WRT-PLS has led to a preprocessing strategy that avoids overfitting and has enhanced model interpretation compared to CV-PLS.

These results show that a more appropriate preprocessing strategy can be selected by using WRT-PLS compared to the ordinary CV-PLS strategy with van der Voet method for #Comp selection. Final model performances of both approaches are in the same order of magnitude. However, the interpretation of the models is largely improved for higher-quality models (less nonsignificant components) using a preprocessing strategy selected using WRT-PLS. For those models, only the known relevant variables are selected. For models where CV-PLS was used to select an appropriate preprocessing strategy, no or wrong variables were deemed important by sMC. An inappropriate selection of variables indicates that the constructed PLS models were not reliable, because of higher irrelevant data variation after applying the selected preprocessing strategy with CV-PLS.

Moreover, the calculation time of WRT-PLS is shorter compared to CV-PLS. All 5000 preprocessing strategies can be evaluated in approximately 1 hour of calculation time on a modern PC (here, Intel Xeon W3565, octa-core at 3.20 GHz, 20.0 GB of RAM, Matlab version R2014b). This drastically reduces computation time, which may easily take or exceed a day for CV-PLS, depending on the number of iterations performed during CV.

4 | WEIGHT RANDOMIZATION TEST IN 2CV

The WRT method is incorporated into the 2CV framework (2CV-PLS) to replace the inner loop for the selection of #Comp. The schematic representation of this process is presented in Figure 7. The properties of this CV-WRT-PLS strategy are evaluated in the classification of several 2-class datasets with different levels of known differences and compared to the permutation tests in 1 study.25

4.1 | Data

The first dataset (dataset A) consists of 120 samples and 333 spectral variables selected from multicapillary column—ion mobility spectrometry measurements with a previously developed data analysis strategy.49,50 A total of 120 breath samples are collected from 67 lung cancer subjects and 53 age-matched healthy subjects.51 The second dataset (dataset B) consists of 96 samples and 101 lipid levels measured at the Netherlands Metabolomics Centre with the UPLC-MS lipidomics platform. Further details are described in Szymańska et al.52 A total of 96 samples are serum samples collected from healthy subjects before the start of a nutritional intervention study. Nutritional effects are
added to 48 randomly selected samples as described in Szymańska et al. From 10 different datasets with different magnitudes of superimposed nutritional effects, 8 are included in this study. These include the following magnitudes of superimposed nutritional effects: 1, 0.75, 0.626, 0.55, 0.5, 0.25, 0.15, and 0.1.

### 4.2 Methods

The partial least squares-discriminant analysis (PLS-DA) with a 2CV procedure, as well as PLS-DA with a single CV procedure and WRT (CV-WRT-PLS-DA), were used. Both procedures were applied 30 times (30 repetitions, ie, repeated 2CV) to dataset A and to each of 8 datasets obtained from dataset B.

During 2CV-PLS 5-fold inner and 5-fold outer CV was used. The number of components was optimized on the basis of performance of the model tested for a validation set, comparing area under the receiver operator characteristic (ROC) curve (AUROC) for models with different complexities. During CV-WRT-PLS-DA test, 5-fold CV was used. The number of components was optimized with the WRT test including 2000 randomizations. Different settings for $\alpha$ and the number of CNC were used to obtain the null distribution. For dataset A, $\alpha = 0.1, 0.05, and 0.01$ as well as 3 and 5 CNC were tested. For dataset B, $\alpha = 0.05$ as well as 3 and 5 CNC were tested.

The performance of the obtained PLS-DA models was evaluated on the basis of means of diagnostic statistics: AUROC and the number of misclassified samples (NMC) calculated across 30 repetitions. Additionally, the performance was related to the means of diagnostic statistics of permutation tests to obtain a $P$ value. A total of 3000 permutation tests were calculated for each dataset using the same procedures as described above but with permuted $y$.

### 4.3 Results

#### 4.3.1 Model performance

The performance of PLS-DA models expressed as AUROC values of 30 repetitions obtained for dataset A is shown in Figure 8A. A similar figure with NMC values is attached as Figure S1. It can be seen that good model performance was obtained with both CV-WRT-PLS-DA and 2CV-PLS strategies. Moreover, it is clearly visible that the performance of PLS-DA models with CV-WRT-PLS-DA is dependent on the WRT test settings: significance level $\alpha$ and CNC. $\alpha = 0.05$ seems to lead to models with the best performance.

Similar observations were obtained for PLS-DA models for dataset B (Figure 8B). The average performance of PLS-DA models obtained with CV-WRT-PLS-DA is 2% to 3% lower than the average performance of PLS-DA models with 2CV-PLS-DA. For datasets with a low magnitude of superimposed effects (0.25, 0.15, and 0.1) the difference is bigger. This effect is related to model significance as is explained in the methodology Section 2.
4.3.2 | Model complexity (#Comp)

The PLS-DA model performance is strongly related to the number of components selected in the inner loop of 2CV-PLS-DA or by the WRT test in CV-WRT-PLS-DA. In real data, the underlying number of components is often unknown. In this study, we have allowed the number of components to be selected from a range between 1 and 30 components.

In case of dataset A (Figure 9), it can be seen that during 2CV-PLS-DA usually models with 2 or 3 #Comp are being selected. In case of component selection by WRT, models with 2 components are most commonly obtained. The distribution of models with different complexities is shown in Figure 9. For 2CV-PLS-DA, also models with 7 to 9 components are often selected. By using WRT more skewed distributions are obtained especially with \( \alpha \) = 0.01 where models with only 1 or 2 components are selected. As a property of WRT, more complex models are obtained with higher \( \alpha \) since it is more difficult to reject the null hypothesis. The number of CNC has a less prominent effect on model complexity than \( \alpha \) but CNC = 3 is selected as the more robust option.

In case of dataset B, for the dataset with the magnitude of superimposed effects equal to 1, 2CV-PLS-DA leads to models with 7 or 8 components (see Table 2). This is slightly more than 6, the number selected in 1 study\(^\text{25}\) where 6 was the maximum number allowed in this selection. On the contrary, CV-WRT-PLS-DA leads to models with 4 or fewer components. When the magnitude of superimposed effects decreases, and thus differences between groups are more difficult to find by PLS-DA models, 2 opposite behaviors can be observed. Firstly, the complexity of models selected with 2CV-PLS-DA gradually increases to 11 components with the decrease of the magnitude of superimposed effects.

Secondly, the complexity of models selected with CV-WRT-PLS-DA decreases to 0 with the decrease of the magnitude of superimposed effects and even reaches #Comp values below 1 for the effect magnitude of 0.55.

4.3.3 | Model significance

Although an AUROC value equal to 1 or NMC equal to 0 can be thought to correspond to models with good performance and a high discriminating power, these values can be attained purely by chance because of a lucky random choice of samples in the test and training sets\(^\text{25}\). In single and 2CV procedures, additional permutation tests are often used to assess the statistical significance of the optimized models. The performance of the investigated original model (eg, its AUROC) is compared with performances of many permutation tests obtained on the same dataset, and model significance can be assessed, eg, by a \( P \) value.

For CV-WRT-PLS-DA, since the permutation test has been included in the method, by definition, the obtained #Comp from WRT-PLSDA is with a significance level of 0.05 and in theory there is no need of additional permutation tests for statistical significance assessment of the final PLS model. However, these additional tests are performed here as well to double check the significance of obtained models.

When additional permutation tests are performed for dataset A with 2CV-PLS-DA (see Table 2), permuted models are more complex than the original model: 8 to 9 components in permuted models vs on average 4.2 components in the original models. \( P \) values for AUROC and NMC are below .05 so the PLS-DA model obtained with the 2CV-PLS-DA procedure has been assumed statistically significant. When additional permutation tests are performed for models obtained with the CV-WRT-PLS-DA procedure, the permuted models result in indeed 0.05 #Comp on average (Table 2). This means that in most permutations, the WRT test was not able to find any significant components in the built PLS-DA models, not even in their first components. No valid PLS models can therefore be built and compared to the original models.

In case of dataset B, permutation tests for CV-WRT-PLS-DA are statistically significant for models with superimposed effect magnitudes of 1, 0.75, 0.626, and 0.55, whereas models with magnitudes 0.5, 0.25, 0.15, and 0.1 are not (Table 2). Permuted models always have a similar complexity of ca 11 components independent of the effect magnitude, whereas original models have different complexities as discussed earlier. Similarly, to dataset A, the permuted CV-WRT-PLS-DA models result in an average of 0.05 #Comp. A valid original model with at least 1 #Comp is in between 0.626 and 0.55 effect magnitude, which is in line with the significance level of 0.05 determined by additional permutation tests with 2CV-PLS-DA procedure.
Calculation time of single, original models with both procedures (CV-WRT-PLS and 2CV-PLS procedures) is very similar and in a range of minutes. In both cases, 3000 additional permuted models were used with calculation times of approximately 1.8 and 8.6 hours for datasets A and B, respectively. Because additional permutation tests for the CV-WRT-PLS procedure are not required to determine original model significance, they can be omitted as demonstrated above. That gives significant advantage in reducing computation time from a few hours to a few minutes when the CV-WRT-PLS procedure is used instead of the 2CV-PLS procedure.

Table 2: Complexity and significance of PLS-DA models of dataset A and datasets obtained with superimposing nutritional effects on dataset B

<table>
<thead>
<tr>
<th>Dataset/procedure</th>
<th>2CV-PLS</th>
<th>CV-WRT-PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original: Mean #Comp</td>
<td>Perm: Mean #Comp</td>
</tr>
<tr>
<td>Dataset A</td>
<td>4.23</td>
<td>8.61</td>
</tr>
</tbody>
</table>

Dataset B with an effect magnitude of:

<table>
<thead>
<tr>
<th></th>
<th>2CV-PLS</th>
<th>CV-WRT-PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original: Mean #Comp</td>
<td>Perm: Mean #Comp</td>
</tr>
<tr>
<td>1</td>
<td>7.68</td>
<td>11.25</td>
</tr>
<tr>
<td>0.75</td>
<td>8.25</td>
<td>11.09</td>
</tr>
<tr>
<td>0.626</td>
<td>9.63</td>
<td>11.32</td>
</tr>
<tr>
<td>0.55</td>
<td>9.42</td>
<td>11.33</td>
</tr>
<tr>
<td>0.5</td>
<td>9.91</td>
<td>11.42</td>
</tr>
<tr>
<td>0.25</td>
<td>11.82</td>
<td>11.29</td>
</tr>
<tr>
<td>0.15</td>
<td>10.91</td>
<td>11.44</td>
</tr>
<tr>
<td>0.1</td>
<td>11.27</td>
<td>11.21</td>
</tr>
</tbody>
</table>

PLS-DA models obtained with 2CV-PLS and CV-WRT-PLS procedure with α = 0.05 and 3 CNC are compared. Complexity is presented for original and permuted models. P value is obtained as described in Material and Methods.

5 DISCUSSIONS AND CONCLUSION

The popularity of PLS (and PLS-DA) in chemometrics is not arguable, yet most practitioners would admit that the choice of #Comp makes the application of PLS somewhat subjective to the operator. The introduction of a RT was a major step towards the automation of #Comp selection. Yet the assumption that the components enter the model in the order of decreasing amount of relevant information only holds on “perfect” data. Again most practitioners would agree that there are no perfect data. We have shown through theoretical work that the components enter the model in order of decreasing amount of variance but that this variance is not necessarily informative.

The updated randomization procedure, WRT, is advantageous on 2 aspects. The WRT is computationally less expensive compared with Wiklund's method with the use of weight vectors instead of the covariance between the score vectors and response variable. Secondly, WRT does not suffer from the presence of irrelevant variance. Instead, the WRT method uses the irrelevant variance (in nonsignificant components) to define a new concept for the quality evaluation of PLS models. This progress opens the door for a full automation of PLS and higher-quality PLS model in less irrelevant variance.

With respect to the permutation null distribution, we provide 2 flexible approaches for testing #Comp, namely, the traditional quantile method and the lognormal transformation. In the former, an exact test can be obtained by using the upper quantile from the permutation null distribution that corresponds with the desired alpha level. In the latter, our results demonstrate the reasonableness of a lognormal transformation in parametrically testing for significant components.

We demonstrated the applicability of this approach in the selection of the most adequate preprocessing and for model selection. In application 1 (preprocessing selection), it was shown that preprocessing strategies selected by WRT-PLS had enhanced model interpretability and a similar model performance compared to strategies selected by CV-PLS. In the CV-PLS procedure, too many LVs were taken into account, thereby hampering proper interpretation of the final model. The WRT-PLS did not suffer from this issue and was able to provide an appropriate preprocessing strategy much faster than CV-PLS.

Interestingly, we observe that this approach generally leads to simpler models (ie, with fewer components and less number of nonsignificant components) in the discussed applications. We also notice that overfitted models tend to disturb further analysis by misleading variable selection procedures, such as sMC. Hence, the use of WRT might, indirectly, improve the interpretability of the PLS models by facilitating the variable selection step.
With WRT-PLS, for the first time not only #Comp but also the method can provide a quantitative measurement of model quality. This is achieved by using the amount of ‘Consecutive Non-significant Components’ (CNC), which is proportional to the irrelevant data variation captured in the PLS model. This irrelevant variation may impact model robustness. A higher CNC corresponds to a model with high irrelevant data variation and hence a less robust model and vice versa. In this way, CNC defines a suitable model quality and this parameter can even be defined a priori and may depend only on the application and less on a specific data at hand. Both the significance level, $\alpha$, and the number of nonsignificant components, CNC, define the quality of the model for the amount of informative information. As shown in the first application for the selection of an optimal preprocessing strategy, a higher-quality model was obtained using WRT-PLS and, in general, this model is expected to be also more robust to future changes when compared with CV-PLS. In many cases, the default value $\alpha = 0.05$ is acceptable as it showed the best performance in the second application in this work.

The WRT enhances model robustness by an opportunity for identifying higher quality PLS models with less irrelevant data. A combination of relevant and irrelevant variation in 1 component may also lead to less robust models and lower predictive performance, eg, responsible for declining performance over time for a predictive model for process predictive performance, eg, responsible for declining component may also lead to less robust models and lower data. A combination of relevant and irrelevant variation in 1 component may also lead to less robust models and lower predictive performance, eg, responsible for declining performance over time for a predictive model for process monitoring in real-time release applications. For this reason, in practice, the model often needs a regular checkup, eg, via monitoring of residuals, and model update or even fully and partly recalibration.53 For many industrial application areas, such as in oil54 or food industries,55 monitoring and updating hundreds of PLS models requires a lot of extra resources.

Finally, as an inherited property of the permutation, similar to several permutation-based machine learning algorithms such as random forest, WRT benefits from a built-in statistical validation mechanism. In the case of no valid PLS model, WRT results in 0 #Comp in accordance with the confidence level $\alpha$, eg, at 0.05, as illustrated in the classification example in this work. In practice, this is a useful property that helps avoiding the misuse of a PLS model even with a single component model when no additional permutation test was used, which is often the case.56,57

ACKNOWLEDGEMENTS

We would like to thank Dr Henk-Jan van Manen (AkzoNobel Supply Chain, The Netherlands) for kindly providing us with the latex data and Emma Brodick (IMSPEX, United Kingdom) for providing us with multicapillary column—ion mobility spectrometry measurements used for dataset A. Part of this work is funded by the Netherlands Organization for Scientific Research (NWO) in the framework of Technology Area COAST.

REFERENCES


30. Pitman EJG. Significance tests which may be applied to samples from any populations. Suppl J R Stat Soc. 1937;4:119-130.


44. N.L. Afanador, T.N. Tran, L. Blanchet. mvdalab: multivariate data analysis laboratory, 2016.


**SUPPORTING INFORMATION**

Additional Supporting Information may be found online in the supporting information tab for this article.