The permutation analysis prior to variable selection greatly enhances robustness of OPLS analysis in small cohorts

Marika Ström[a#], Nicole Wagner[a#], Iryna Kolosenko and Åsa M. Wheelock[a, b,*]

[a] Respiratory Medicine Unit, Department of Medicine Solna and Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden;

[b] Department of Respiratory Medicine and Allergy, Karolinska University Hospital, Stockholm, Sweden

Shared authorship

[*] Corresponding author

Correspondence should be addressed to: Åsa Wheelock, E-mail: asa.wheelock@ki.se

Wheelock lab, CMM, L8:02, Respiratory Medicine Unit, Department of Medicine, Karolinska Institutet, SE-171 76, Stockholm, Sweden.

Research funding source

The Swedish heart-lung foundation (HLF), Swedish Research Council (VR), Karolinska Institutet Faculty funding for doctoral students (KID)

Conflict of interest statement

The authors declare that they have no competing interests.

Abstract

The R-workflow roplspvs (**R o**rthogonal **p**rojections of **l**atent **s**tructures with **p**ermutation over **v**ariable **s**election) facilitates variable selection, model optimization and significance testing using permutations of OPLS-DA models, with the scaled loadings (p[corr]) as the main metric of significance cutoff. Permutations including (over) the variable selection procedure, prior to (sans), as well as post variable selection are performed. The resulting p-values for the correlation of the model (R^2) and the cross-validated correlation of the model (Q^2) sans-, post- and over- variable selection are provided as additional model statistics. These model statistics are useful for determining the true significance level of OPLS models, which otherwise have proven difficult to assess particularly for small sample sizes. Furthermore, we propose a means for estimating the background "noise" level based on permutated false positive rates of R^2 and Q^2. This novel metric is then utilized to calculate an adjusted Q^2 value. Using a publicly available metabolomics dataset, the advantage of performing permutations over variable selection was demonstrated for small sample sizes. Iteratively reducing the sample sizes resulted in overinflated models with increasing R^2 and Q^2, and permutations post variable selection indicated falsely significant models. In contrast, the adjusted Q^2 was marginally affected by sample size, and represents a robust estimate of model predictability, and permutations over variable selection showed true significance of the models.

Background

In clinical studies, particularly those involving invasive sampling, multivariate statistical modeling is key to characterizing disease mechanisms. Small sample sizes combined with a multitude of analytes in omics studies increase the risk of false discoveries.[1].

Partial least squares (PLS) is a multivariate modelling method [2] used for correlating two data blocks. When one block (generally the Y-block) is set to a single variable, PLS can be used as a supervised method to evaluate group separation in a multivariate fashion, as well as to extract features driving the separation. Orthogonal projections to latent structures (OPLS) [3] is a development of PLS that rotates the principal components in order to separate the variance *within* the groups of interest (orthogonal components) from the variance *between* the groups of interest (predictive components), and thereby structure the model to better facilitate the extraction of features driving the separation between the groups. Similar to PLS, the y-block needs to be reduced to variables determinant of group belonging to facilitate efficient supervised modeling of group separation.

OPLS is an efficient tool for omics dataset analysis [4]. The model statistics provided by the original algorithms in the SIMCA software (REF) are primarily designed to provide significance of group separation through cross-validated ANOVA (CV-ANOVA), and the predictive power of group separation through the cross validated correlation of the original Y-block with the model (Q^2Y, a.k.a. Q^2). The resulting OPLS model can help reduce the large number of variables in an omics dataset to a subset of key features that drive group separation, which may be of interest as potential biomarkers or molecular targets. This process, known as variable selection or feature selection, is often conducted using methods such as random forest, support vector machine learning, or an alpha-cutoff based on t-statistics corrected for multiple testing.

In OPLS, two main variable selection metrics are commonly used: Variable Importance in the Projection (VIP) and scaled loadings p(corr). VIP ranks variables based on their contribution to group separation, with 1.0 representing the average contribution. However, the relative nature of VIP and its dependence on the number of variables make it not optimal for variable selection, especially when comparing models. On the other hand, p(corr) is a more robust metric because it is related to the loadings of the predictive variable and is scaled as a correlation coefficient ranging from -1 to 1, facilitating comparison between models. Here, we propose using p(corr) as the primary metric for variable selection in OPLS.

To evaluate the performance of models, R^2Y (hereafter R^2) is the established model statistic that describes the variation of the dataset explained by the model. Q^2Y (hereafter Q^2) describes the predicted R^2 assessed by cross validation, and is a more robust measure of the model performance [5].

OPLS models are prone to overfitting [6]. Adding additional principal components to a model increases R^2 and can increase group separation but at the same time the predictability Q^2 decreases indicating over-fitting of the model. Reducing the ratio between the number of samples and features leads to greater separation, potentially resulting in perfect segregation of a dataset when the ratio is low. This threshold has been estimated to be 1:2 or lower [7]. Also, during variable selection R^2 and Q^2 increase, and the extent of the increase is partly due to the overinflation of the model statistics. CV-ANOVA is a common method for OPLS testing; in this context accuracy is defined as the proportion of the subject group correctly predicted by the model.

One way to test the significance of model statistics is by permutation tests [8]. In this approach, sample group labels are scrambled prior to modeling the dataset. The procedure is iterated, and

the results represent model statistics unaffected by chance. Permutation tests, such as SIMCA software (Sartorius) [9], have been implemented in tools for multivariate analysis and in the R package ropls. This procedure is commonly performed either prior to variable selection (by including all variables in the full dataset to calculate the significance of R^2 and Q^2), or post-variable selection (by including only the final selected variables of interest in the permutation test). While the latter approach assures that the selected variables will not produce a significant separation of groups of subjects randomly picked from cohort, it brings the caveat of only controls for the variables of interest being considered and does not assess alterations represented by random variation.

A more robust method to assess significance involves permutation testing with variable selection, also known as permutation test over variable selection. In this procedure, permutations are first conducted similarly to standard permutation testing. Then, variable selection is applied to each permuted dataset before fitting a model to the permuted data. This approach, which incorporates variable selection into permutations, has been explored in OPLS models [10] and has been integrated into the MUVR package for PLS models [11].

To estimate the significance of the model, we compare the model statistics of permuted models to those of the unpermutated model under examination. This is achieved by calculating the ratio between permuted models that exhibit statistics above and below those of the unpermutated reference model, yielding a nonparametric p-value. This approach, implemented in the ropls package, provides p-values for R^2 and Q^2.

The objective of this study was to develop an R script, ropls-pvs (**R O**rthogonal **P**rojections of **L**atent **S**tructures with **P**ermutation over **V**ariable **S**election), utilizing the Bioconductor package ropls for OPLS modeling. The script incorporates variable selection and permutations both before and after variable selection. Additionally, we aimed to establish a significance level for the predictive power of Q^2 in models post-variable selection by comparing it to the estimated

overinflation caused by reducing the number of variables. This overinflation serves as a threshold for Q^2, with models achieving a Q^2 above this threshold considered statistically significant. Subsequently, an adjusted Q^2 was calculated by subtracting the estimated threshold, resulting in a significance cutoff of 0.

In this study, we utilized the script to examine the impact of decreasing sample sizes on R^2, Q^2, overinflation, and the proposed adjusted Q^2 using a publicly available metabolomics dataset. Furthermore, we investigated whether permutations without, after, and over variable selection could distinguish between a significant model and a model generated from random data at various sample sizes. Given that OPLS models are prone to overfitting, particularly after variable selection, the development of a tool to identify and avoid publishing insignificant models is crucial. Such a tool could help address the well-established issue of irreproducibility in research.

Methods

The ropls-pvs Package

The ropls-pvs workflow involves an R package designed to streamline semi-automated pairwise group comparisons within a dataset using OPLS modeling, based on primary group assignment (e.g., treatment vs. placebo groups). The package offers the flexibility to stratify the analysis based on additional variables of interest (e.g., gender, smoking status). To ensure robust and reliable model optimization, thorough significance testing is included via permutations, which incorporate variable selection.

OPLS analysis is conducted using the ropls R package. Additionally, for generating HTML reports with plots, the package relies on several dependencies including Rmarkdown, tools,

ggplot2, ggrepel, kableExtra, gridExtra, ggpubr, matrixStats, stringr, tryCatchLog, devtools, DescTools, precrec, pROC, and rstatix.

The ropls-pvs R workflow is available at https://github.com/pulmonomics-lab/roplspvs.

Workflow

The workflow of ropls-pvs entails running a main file (Run.R) that executes the opls-pvs function (refer to Figure 1). Users can customize a subset of settings through two configuration files:

1. The Configure_Get_Started.R file contains basic parameter settings, including information about the data matrix (e.g., variable names, preprocessing details) and observations (e.g., sample IDs, primary group assignments). Additionally, it specifies the metadata to be used for stratification in OPLS analyses. This file also allows users to define the number of permutations, the significance level for the p(corr) cutoff, and the order of groups for directional interpretation in OPLS score plots (e.g., healthy vs. diagnosis or diagnosis vs. healthy). Users can also specify which pairwise comparisons to run, with the default being all groups in the sample ID file.

2. The Configure_Advanced.R file sets default directories and folder names for output files, which can be adjusted if an alternative folder structure is preferred. This file also enables users to customize variable selection settings, set the maximum number of allowed orthogonals, define criteria for selecting the best performing model, and adjust the length and format of variable names displayed in OPLS loadings plots. Users can select modelling strategies (default: 0-5) and choose whether to generate both comparisons and summary output files.

The ropls-pvs script first generates a table (model_table_to_analyse) containing information about each pairwise comparison to be performed. Then, it passes parameters from the model_table_to_analyse and the two configure files to the Rmarkdown files. The data matrix,

metadata file, and functions from roplspvs_Functions.R are incorporated into the Rmarkdown files to conduct the modelling, resulting in an Rdata file and an HTML output file for each comparison. The HTML output includes score plots, loading plots, permutation plots before and after variable selection, and Receiver Operating Characteristic (ROC) curves. Additionally, HTML files summarizing model statistics for all pairwise comparisons are generated, along with variable lists displaying associated p(corr) values for variables selected by the respective modeling strategies. Figure 1 provides an overview of the input and output files of ropls-pvs.

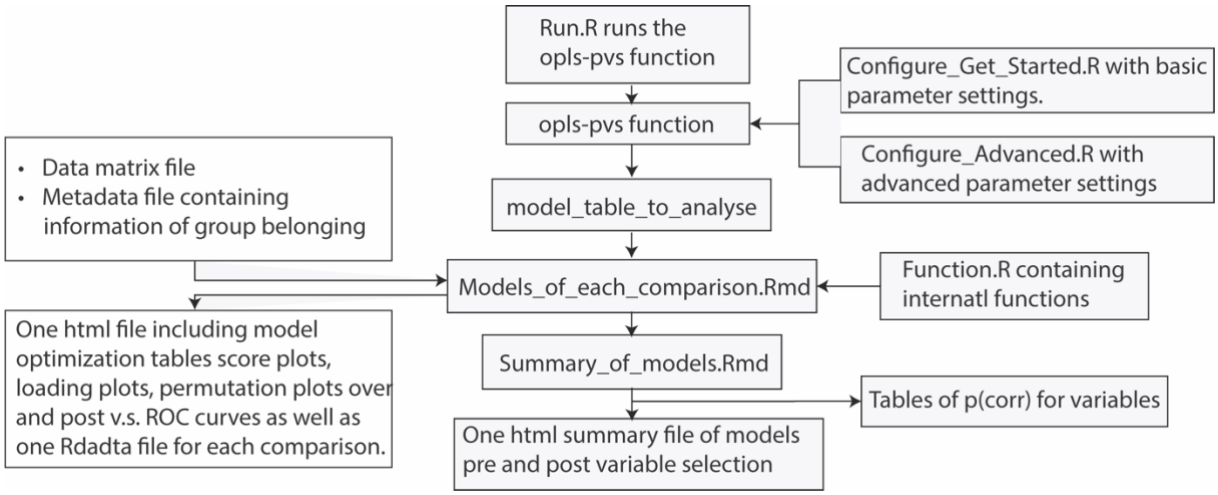

Figure 1. *Schematic of the ropls-pvs workflow and package.*

Preprocessing of Dataset

The dataset undergoes preprocessing as outlined in Figure 2. Preprocessing is iteratively conducted for each pairwise comparison to optimize the variables considered. Options include replacing zero values with NA or the lower limit of detection (LLD), followed by optionally replacing NA with LLD. Users can manually set the LLD or estimate it as 1/3 of the minimum value in the dataset. Additionally, data filtering removes variables based on a user-defined missing value tolerance level for each group in the comparison.

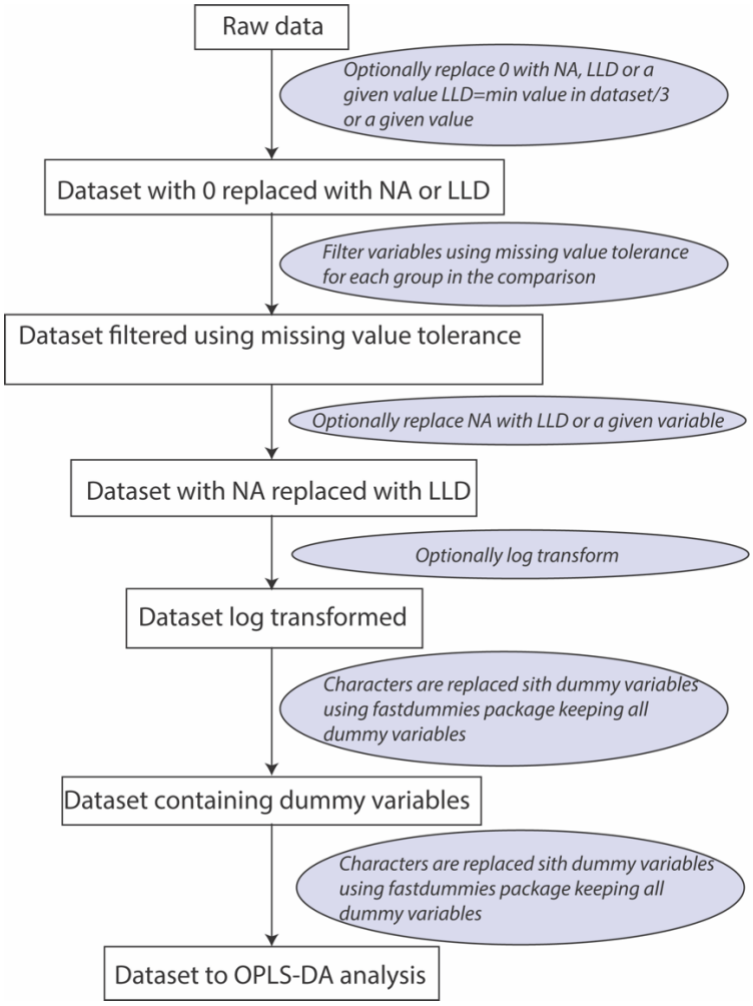

Figure 2. *Schematic of data pre-processing options in the ropls-pvs package.*

OPLS and PCA Modeling

The Bioconductor R package ropls (https://bioconductor.org/packages/release/bioc/html/ropls.html) facilitates initial PCA and OPLS modeling [28]. By default, OPLS employs NIPALS (Nonlinear Iterative Partial Least Squares) to handle missing data, if NAs are present in the dataset. Mean centering and scaling to unit variance (UV) are applied, enabling variables with low abundance or amplitude to contribute to the model.

The process begins with the generation of PCA models before variable selection for each comparison. This encompasses creating PCA models that integrate all observations in the comparison, as well as PCA models for each group individually, aiding in robust outlier identification. Subsequently, OPLS models are constructed for each comparison, both without variable selection (encompassing all variables meeting the preprocessing QC criteria) and after variable selection (comprising solely variables selected by the respective modeling strategy, as elaborated in the section below). Model evaluation entails reporting R^2, Q^2, and RMSE model statistics, alongside significance assessment through permutations.

Variable Selection

Feature selection, also termed variable selection, operates using five distinct modeling strategies and/or stringency levels. The default variable selection procedures in all model strategies hinge on p(corr), with the possibility of integrating Variable Importance in Projection (VIP) as an additional variable selection metric. Different p(corr) cutoffs for variable selection are applied across the five strategies. The determination of the p(corr) cutoff primarily stems from the iterative optimization of the predictive power, Q^2, of the model post-variable selection. This iterative process aims to thwart overfitting, thus maintaining the maximum predictive power of the model after variable selection. Figure 3 furnishes an overview of the available modeling strategies.

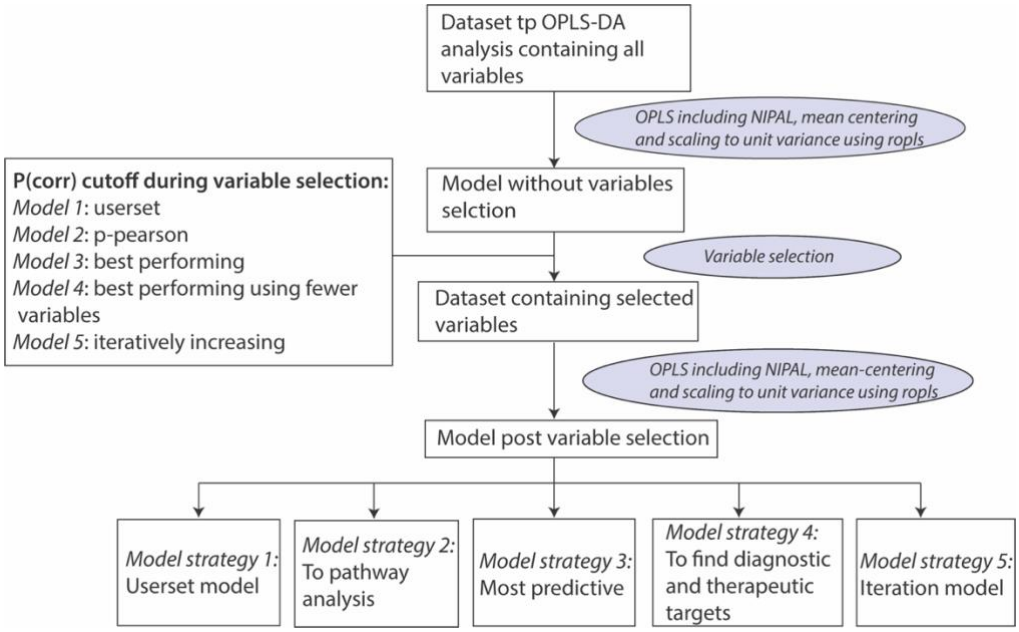

Figure 3*. Overview of the five different Modelling strategies available in the roplspvs package. The p(corr) cutoff used in the respective modelling strategies are indicated in the text box to the left.*

Modeling Strategies

In Modeling Strategy 1, users have the option to set the p(corr) cutoff, offering flexibility to adjust model settings or evaluate models optimized in other software using the permutation analyses in the ropls-pvs script.

In Modeling Strategy 2, the p(corr) cutoff is determined by incorporating all variables with significant correlation to the score, established by a user-defined p-value corresponding to the Pearson's correlation coefficient for the group size. This cutoff also serves as the default for Modeling Strategy 1 if no other cutoff is specified.

Modeling Strategy 3 is recommended for users aiming to identify the best performing model with high Q^2, minimal difference between R^2 and Q^2, and low permutation p-values (p[R2_perm_post_vs] and p[Q2_perm_post_vs]). This strategy prioritizes predictive power while avoiding overfitting and selecting models not significantly better than random. Users can

adjust the balance between these criteria using the variable prefered_pR2_and_pQ2_permutated_post_vs.

Modeling Strategy 4 minimizes the number of selected variables while preserving predictability. The p(corr) cutoff is adjusted to achieve the best performing model post-variable selection, gradually increasing the cutoff as long as Q^2 is not reduced by more than 1%. Users can control the reduction in variables by setting the pcorr_diff parameter. Modeling strategies 3 and 4 are summarized in Figure 4.

Modeling Strategy 5 employs an iterative approach by incrementally increasing the p(corr) cutoff and refining the model, eliminating variables with the least correlation. The cutoff is raised as long as Q^2 post-variable selection is not reduced by more than 1%.

To facilitate the evaluation of p(corr) cutoff optimization, a plot of Q^2 versus the number of variables using different cutoffs is generated. This, coupled with a model table, enables users to confirm the selection of the best performing model.

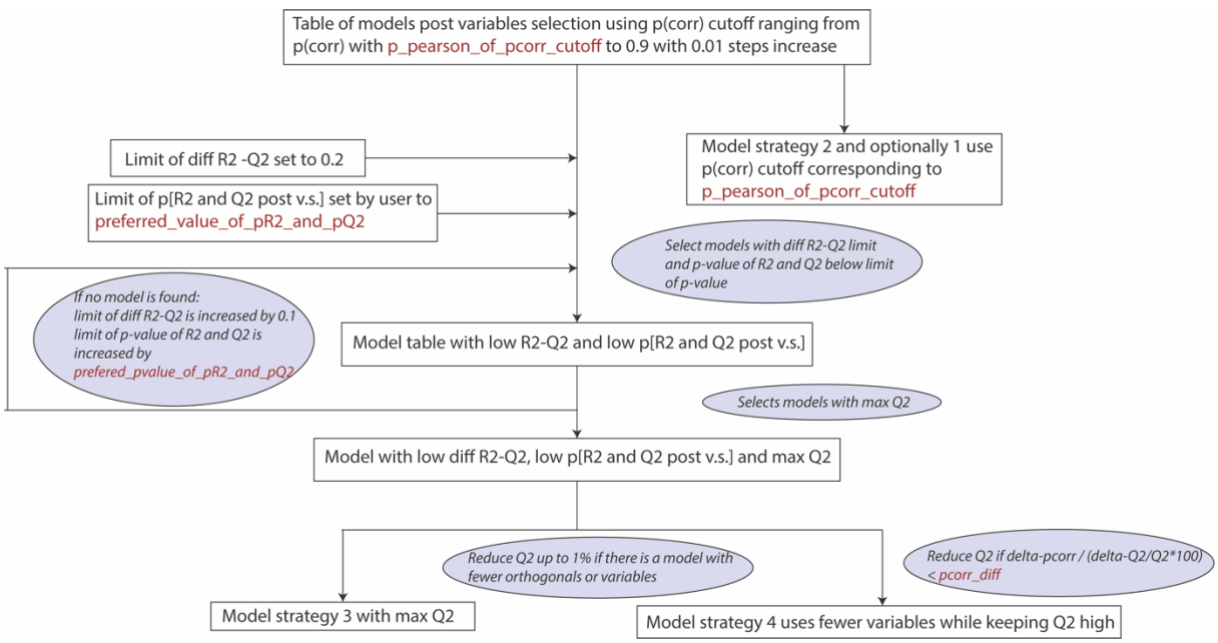

Figure 4. *Details for selecting p(corr) cutoff of Modelling strategies 3 and 4. Parameters that are user-defined in ropls-pvs are marked in red.*

Selection of Number of Orthogonal Variables

The user can define the number of orthogonals for Modeling Strategy 1, with the default set to 0. For Mode**ling** Strategies 2-5, the number of orthogonals is determined using the ropls default method, which adds an orthogonal as long as Q^2 increases by 1% in the currently evaluated model. Users can also set the maximum number of orthogonals for both models without and after variable selection. Additionally, the script checks that no strategy produces a better performing model post-variable selection using fewer orthogonals.

Permutations

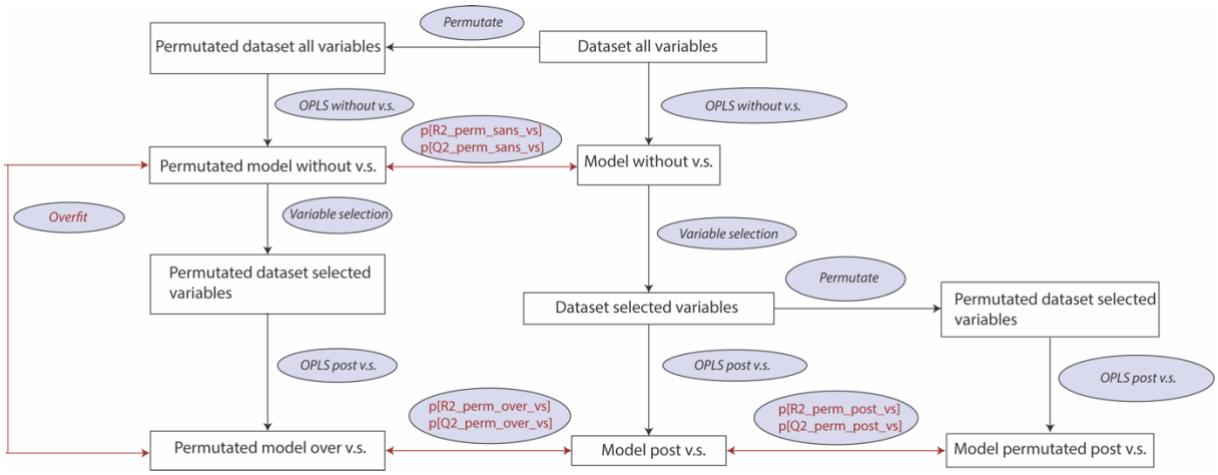

Figure 5. *Estimations of the significance level of each model statistics is provided based on permutation test at three levels; prior to variable selection (sans v.s.), after variable selection (post v.s.), as well as over variable selection (over v.s.) p[R2 and Q2 for permutations sans, post and over variable selection] is estimated by comparing permutated models to unpermutated model under investigation. The comparisons of models for the estimated p-values as well as the overfit are shown in red.*

We assess the significance of models using permutations without, after, and over variable selection, as illustrated in Figure 5.

Permutations without and after variable selection involve randomizing the group assignment of subjects using the ropls package. During permutations without variable selection, we model the dataset, including all variables from preprocessing, against permuted group assignments. In

contrast, permutations after variable selection entail modeling the dataset only with variables selected during variable selection against permuted group assignments. We compare the model statistics R^2 and Q^2 from permuted models to those of the original unpermuted models, resulting in p-values for R^2 and Q^2 without variable selection (p[R2_perm_sans_vs] and p[Q2_perm_sans_vs]) and after variable selection (p[R2_perm_post_vs] and p[Q2_perm_post_vs]).

Similarly, permutations over variable selection involve modeling the dataset with all variables from preprocessing. Each permuted model undergoes variable selection using the same p(corr) cutoff as the unpermuted model under investigation. We compare the model statistics R^2 and Q^2 from permuted models after variable selection to those of unpermuted models after variable selection, resulting in p-values for R^2 and Q^2 over variable selection (p[R2_perm_over_vs] and p[Q2_perm_over_vs]).

One limitation of permutations over variable selection is that the settings of the unpermuted model are used for permuted models. This may lead to increased significance for p[R2_perm_over_vs] and p[Q2_perm_over_vs] compared to if p(corr) were optimized for permuted models as well. To address this, we consider the regression over R^2 and Q^2 of permuted models. We plot the R^2 and Q^2 of permuted models against the correlation of permuted and unpermuted group assignments. A significant positive correlation coefficient in the regression indicates a significant model, as higher correlation to the original group assignments should yield higher R^2 and Q^2.

The user can set the number of permutations, with a default of 20 permutations for initial analysis, balancing computational time with statistical power. However, increasing the number of permutations is recommended for final analysis.

Establishing a Statistically Sound Cutoff for Q^2

We determined the overinflation of OPLS models resulting from variable selection by comparing the Q^2 of permuted models after variable selection to the Q^2 of permuted models without variable selection. This difference in Q^2 represents the overinflation of Q^2, serving as a robustness threshold. Models with Q^2 above this threshold are unlikely to have occurred by chance. To account for this overinflation, we calculated an adjusted Q^2 by subtracting the estimated overinflation from the Q^2 of the model after variable selection. Consequently, models with an adjusted Q^2 above zero can be considered statistically significant and not likely to have occurred by random chance.

Figures

We utilize the ropls package to generate scores plots, as well as loading, prediction, diagnostic, and outlier plots, along with plots of permutations without and after variable selection.

Plots generated by ropls-pvs include score plots of predictive components displayed as boxplots, and p(corr) plots illustrating variables driving model separation. These plots show all variables, and if the number exceeds 50, the 50 most influential variables are displayed. We depict model optimization by comparing the number of variables to Q^2 after variable selection. Additionally, we generate plots of permutations over variable selection, as described in the section above, using ropls-pvs.

After model creation and HTML file generation, we can compare them using Shared and Unique Structures (SUS) plots. SUS plots are correlation plots of p(corr) for each variable in two models. To incorporate unique variables from each model, we create new models containing all variables from both models, with p(corr) from optimized models displayed in the resulting SUS plot. SUS plots aid in identifying variables with similar or different effects when

comparing groups, offering insights into potential interactions or confounders in the comparison.

Qualitative Variables

We transformed qualitative data in the dataset into dummy variables, following the same procedure as in SIMCA. Typically, in analyses, one dummy variable is excluded as using all of them contains redundant information. However, in this analysis, we utilized all dummy variables in the modeling. This approach offers the advantage of demonstrating how all settings affect the outcome, albeit potentially leading to slight overfitting. Including qualitative variables in the analysis is beneficial for creating models using clinical data.

Cohort and Dataset

We showcased the ropls-pvs package using the untargeted metabolomics dataset MTBLS136 [13], which originates from the Cancer Prevention II Nutrition Cohort [14]. This dataset is publicly available at Metabolights (http://www.ebi.ac.uk/metabolights/MTBLS136/files). It consists of blood metabolome profiles generated on the Metabolon platform for studying post-menopausal hormone treatments in women. Among the 1336 women in the cohort, 332 received estrogen therapy alone (E-only), 337 received combined estrogen and progestin treatment (E+P), and 667 women received no hormone replacement therapy (non-users). For analysis, we stratified women into age groups of 55 and under, 56-60, 61-65, 66-70, 71-75, 76-80, and 81-85. We replaced zero values in the dataset with NA, followed by filtering to allow for 25% missing values and log-transformation.

Comparing Hormone Users to Non-users Using ropls-pvs

We demonstrated OPLS modeling using ropls-pvs with the aforementioned dataset. We set a maximum of 5 orthogonals for both models without and after variable selection to mitigate the risk of overfitting caused by excessive orthogonals. We set the variable

'prefered_pR2_and_pQ2_permutated_post_vs' to 0.05 to balance between maintaining low p-values for p[R2 and Q2 post v.s.] and minimizing the difference between R^2 and Q^2 while maximizing Q^2. Additionally, we set 'Pcorr_diff' to 0.01 to limit the number of variables in Modeling strategy 4..

Effect of Small Sample Sizes on OPLS Models

The study utilized the ropls-pvs package to examine the impact of sample sizes on OPLS models using the previously mentioned dataset. To mitigate unrelated variability based on gender and age, we narrowed our focus to the 61-65 year age group, which is characterized by a relatively large number of participants (nE+ P = 129, nE-only= 91, and non-users=121) and homogeneous group nature as confirmed by PCA.

A robust model comparing estrogen plus progestin users to post-menopausal hormone non-users exhibited a Q^2 of 0.53, deemed significant by permutation tests at all levels (p[Q2_perm_ post_vs] ≤0.01, p[Q2_perm_over_vs] ≤0.01). Conversely, a weaker model with Q^2 of 0.10 also passed the significance threshold by permutation tests at all levels (p[Q2_perm_sans_vs] ≤0.01, p[Q2_perm_ post_vs] ≤0.01, p[Q2_perm_over_vs] ≤0.01). Additionally, hormone non-users were compared to the same group to verify model insignificance when no difference between groups was expected.

Subsets of subjects from the three models were drawn with consistent sample sizes for each group (spanning from 4 to 80 for each group, 4 to 60 for each non-user group), repeated 12 times for each subset. We applied Modeling strategies 3 and 4 to each subset using ropls-pvs. R^2 and Q^2 for models without and after variable selection were plotted against sample sizes, with models color-coded based on p[R2 and Q2 perm. sans, post, and over v.s.].

Furthermore, we determined the average overinflation and adjusted Q^2 for each sample size for the strong model.

Results

Comparing Hormone Users to Non-users Using ropls-pvs

The effectiveness of the ropls-pvs package demonstrated using publicly available data investigating blood metabolite composition concerning hormone replacement therapy in post-menopausal women (MTBLS136 [13]). Initially, we conducted pairwise comparisons for all contrasts among the three study groups: estrogen-and-progestin users (E+P), estrogen-only users (E only), and individuals reporting no hormone replacement use (non-users). We performed analyses both jointly for all age groups (joint) and stratified by age groups using 5-year bins. We excluded age groups 81 and above due to small sample sizes.

Table 1 summarizes the model statistics for all pairwise comparisons using Modeling strategy 4. All models across all Modeling strategies demonstrated significance in permutation post-variable selection (p[Q2_perm_post_vs]). Even with the more stringent approach of permutation testing over variable selection (p[Q2_perm_over_vs]), most comparisons remained significant across all Modeling strategies. The exceptions were the E+P versus E-only comparison in the 56-60 age group in Modeling strategy 1 and 2, and the 76-80 age group, which were insignificant using Modeling strategy 1 through 4.

Similarly, p[Q2_perm_sans_vs] indicated significance for all comparisons except the two aforementioned ones. These two insignificant comparisons also displayed insignificant p-values for correlation of permutations across Modeling strategy 1 through 4. For Modelling strategy 2 through 4, both models without and after variable selection showed low Q^2 (<0.4) for all age groups comparing E+P versus E-only, while models comparing E+P and E-only

versus non-users exhibited high Q^2 (>0.4). Adjusted Q^2 was consistently low (<0.1) for all models across all age groups comparing E+P versus E-only.

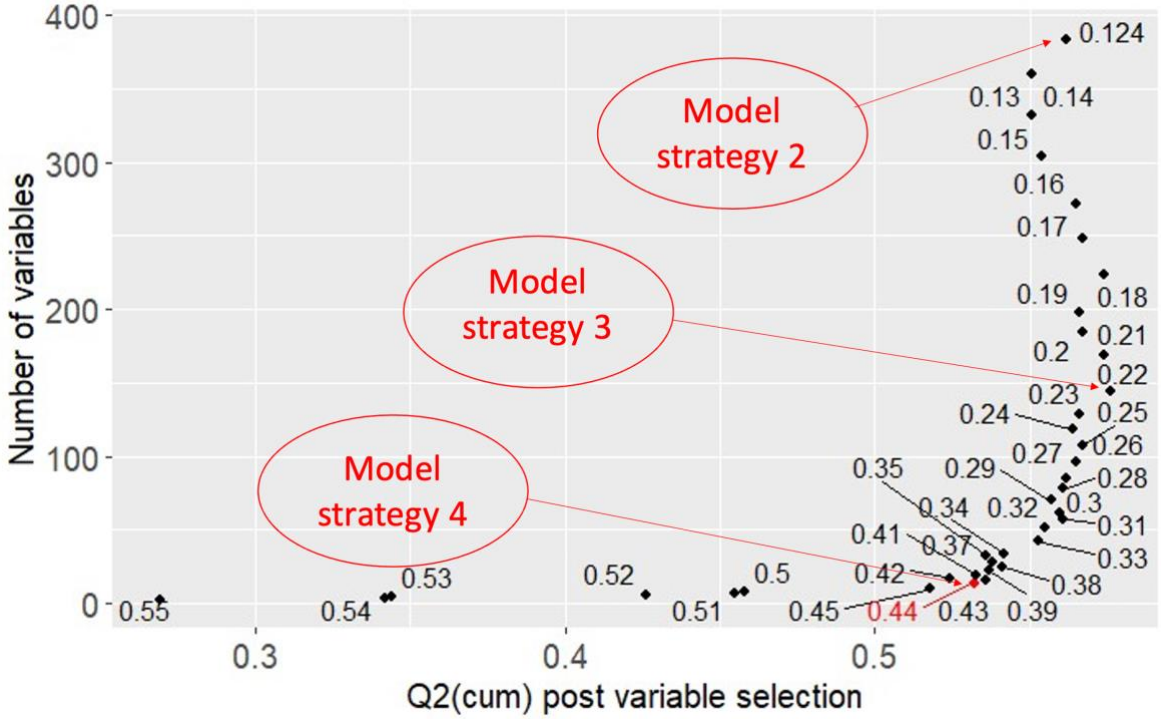

Figure 6. *Visualization of the optimization of p(corr) cutoffs for models comparing E+P versus Nonuser, 61-65 ages using Modeling strategy 2-4.*

The ropls-pvs package generates a plot illustrating the Q^2 post variable selection against the number of selected variables in the respective model to track iterative model optimization based on p(corr). This plot indicates the p(corr) cutoff for the model, as shown in Figure 6 for optimizing the contrast between E+P and non-users in the 61-65 age group across Modeling strategies 2 through 4. For Modeling strategy 2, we utilized a p(corr) cutoff of 0.124, selecting 384 variables with Q^2=0.56. Modeling strategy 3 employed a p(corr) cutoff of 0.22, selecting 144 variables, resulting in Q^2=0.58. Finally, Modeling strategy 4 with a p(corr) cutoff of 0.44 reduced the selected variables to 14, yielding Q^2=0.53.

Figure 7 presents the optimized OPLS model for the comparison between E+P users and hormone non-users in the 61-65 age group, employing Modeling strategy 4. The separation

between the two groups was highly significant based on the predictive component, as evidenced by the t-test ($p=10^{-47}$) (Figure 7A). Additionally, the corresponding permutation over variable selection demonstrated high significance for both R^2 and Q^2, as indicated by p[R2_perm_over_vs], p[Q2_perm_over_vs], and the correlation coefficient between Y and permuted Y (Figure 7C-D).

Table 1

Group 1		Group 2			Model sans v.s.			Perm. sans v.s.			Model post v.s.					Perm. post v.s.		Perm. over v.s.			Correlation perm. over v.s.						
Treatm.	n	Treatm.	n	age	Ort	R^2Y	Q^2	mean Q^2	pR^2Y	pQ^2	p(corr) cutoff	K	Ort	R^2Y	Q^2	pR^2Y	pQ^2	mean Q^2	pR2Y	pQ^2	y-axis intercept	$k(R^2Y)$	$k(Q^2)$	p-value R^2Y	p-value Q^2	overfit perm. Q^2	Adj. Q^2
E-only	20	Nonuser	41	56-60	1	0.91	0.42	-0.21	0.008	≤0.001	0.45	22	0	0.71	0.66	≤0.001	≤0.001	0.09	≤0.001	≤0.001	0.06	0.14	0.12	6E-06	8E-05	0.29	0.37
E-only	91	Nonuser	121	61-65	1	0.74	0.44	-0.30	≤0.001	≤0.001	0.44	20	1	0.51	0.46	≤0.001	≤0.001	0.02	≤0.001	≤0.001	0.01	0.35	0.39	5E-30	2E-37	0.33	0.14
E-only	96	Nonuser	202	66-70	2	0.81	0.62	-0.35	≤0.001	≤0.001	0.44	24	1	0.56	0.53	≤0.001	≤0.001	0.01	≤0.001	≤0.001	0.00	0.55	0.56	6E-81	1E-84	0.37	0.17
E-only	90	Nonuser	185	71-75	2	0.84	0.65	-0.34	≤0.001	≤0.001	0.52	7	1	0.61	0.60	≤0.001	≤0.001	0.01	≤0.001	≤0.001	0.00	0.53	0.58	1E-73	2E-92	0.35	0.24
E-only	32	Nonuser	103	76-80	1	0.82	0.47	-0.30	≤0.001	≤0.001	0.37	22	1	0.56	0.51	≤0.001	≤0.001	0.05	≤0.001	≤0.001	0.04	0.28	0.30	4E-19	4E-22	0.35	0.16
E-only	332	Nonuser	667	joint	2	0.69	0.61	-0.25	≤0.001	≤0.001	0.45	23	1	0.52	0.51	≤0.001	≤0.001	0.00	≤0.001	≤0.001	-0.01	0.84	0.85	2E-268	8E-274	0.25	0.26
E+P	40	E-only	20	56-60	0	0.60	-0.11	-0.19	0.72	0.33	0.57	6	0	0.41	0.38	≤0.001	≤0.001	0.12	0.002	≤0.001	0.04	0.07	0.07	0.04	0.02	0.32	0.06
E+P	129	E-only	91	61-65	0	0.32	0.05	-0.21	0.44	≤0.001	0.53	23	1	0.14	0.10	≤0.001	≤0.001	0.02	≤0.001	≤0.001	0.01	0.19	0.19	3E-09	2E-09	0.23	-0.13
E+P	91	E-only	96	66-70	1	0.67	0.09	-0.30	0.05	≤0.001	0.18	144	0	0.40	0.30	≤0.001	≤0.001	0.01	≤0.001	≤0.001	0.00	0.18	0.19	9E-09	1E-09	0.31	-0.01
E+P	59	E-only	90	71-75	0	0.44	0.06	-0.24	0.21	≤0.001	0.43	32	2	0.29	0.18	≤0.001	≤0.001	0.04	≤0.001	≤0.001	0.03	0.13	0.09	4E-05	0.004	0.27	-0.10
E+P	14	E-only	32	76-80	0	0.76	-0.21	-0.17	0.13	0.63	0.45	8	0	0.40	0.34	≤0.001	≤0.001	0.23	0.25	0.11	0.22	0.05	0.05	0.09	0.11	0.39	-0.05
E+P	337	E-only	332	joint	3	0.61	0.21	-0.36	≤0.001	≤0.001	0.14	163	2	0.37	0.23	≤0.001	≤0.001	-0.02	≤0.001	≤0.001	-0.02	0.62	0.44	4E-108	9E-50	0.34	-0.11
E+P	40	Nonuser	41	56-60	3	0.96	0.50	-0.24	0.13	≤0.001	0.38	22	1	0.72	0.62	≤0.001	≤0.001	0.06	≤0.001	≤0.001	0.05	0.18	0.13	1E-08	3E-05	0.30	0.31
E+P	129	Nonuser	121	61-65	2	0.81	0.58	-0.34	≤0.001	≤0.001	0.44	14	1	0.57	0.53	≤0.001	≤0.001	0.02	≤0.001	≤0.001	0.00	0.49	0.50	1E-60	6E-65	0.36	0.17
E+P	91	Nonuser	202	66-70	2	0.81	0.59	-0.35	≤0.001	≤0.001	0.41	19	1	0.63	0.60	≤0.001	≤0.001	0.01	≤0.001	≤0.001	0.00	0.51	0.55	1E-66	2E-81	0.36	0.23
E+P	59	Nonuser	185	71-75	2	0.80	0.48	-0.33	≤0.001	≤0.001	0.31	32	0	0.47	0.44	≤0.001	≤0.001	0.02	≤0.001	≤0.001	0.01	0.39	0.43	1E-38	9E-47	0.35	0.09
E+P	14	Nonuser	103	76-80	4	0.98	0.40	-0.31	0.006	≤0.001	0.26	37	1	0.68	0.53	≤0.001	≤0.001	0.04	≤0.001	≤0.001	0.03	0.14	0.17	1E-05	3E-08	0.35	0.18
E+P	337	Nonuser	667	joint	3	0.75	0.65	-0.31	≤0.001	≤0.001	0.31	39	2	0.61	0.59	≤0.001	≤0.001	0.00	≤0.001	≤0.001	-0.01	0.84	0.84	2E-264	7E-264	0.31	0.28

v.s.: variable selection; Perm.: permutation analysis; Treatm.: treatment group; n: number of subjects in group; Ortho: number of orthongonal components in the respective OPLS model; pR^2Y: significance of nonparametric permutation analysis for R^2Y; pQ^2: significance of nonparametric permutation analysis for Q^2; K: number of selected variables in optimized model; mean Q^2: mean Q^2 for 1000 permutation analyses over v.s.; $k(R^2Y)$: Pearson correlation coefficient between R^2Y and correlation between permutated and unpermutated group belonging; $k(Q^2)$: Pearson correlation coefficient between Q^2 and correlation between permutated and unpermutated group belonging; p-value R^2Y: significance of Pearson correlation coefficient between R^2Y and correlation between permutated and unpermutated group belonging; p-value Q^2: significance of Pearson correlation coefficient between Q^2 and correlation between permutated and unpermutated group belonging; overfit perm Q^2: mean (Q^2 permutated models over v.s.)-mean (Q^2 permutated models sans v.s.), represents an estimation of the overinflation occurring due to variable selection; Adj. Q^2: Q^2 post v.s. - overfit perm. Q^2, represents an estimation of the Q^2 corrected for overinflation by variable selection; E-only: cohort group receiving estrogen hormone replacement therapy; Nonuser: cohort group not receiving hormone replacement therapy; E+P: cohort group receiving combined estrogen and progestin hormone replacement therapy.

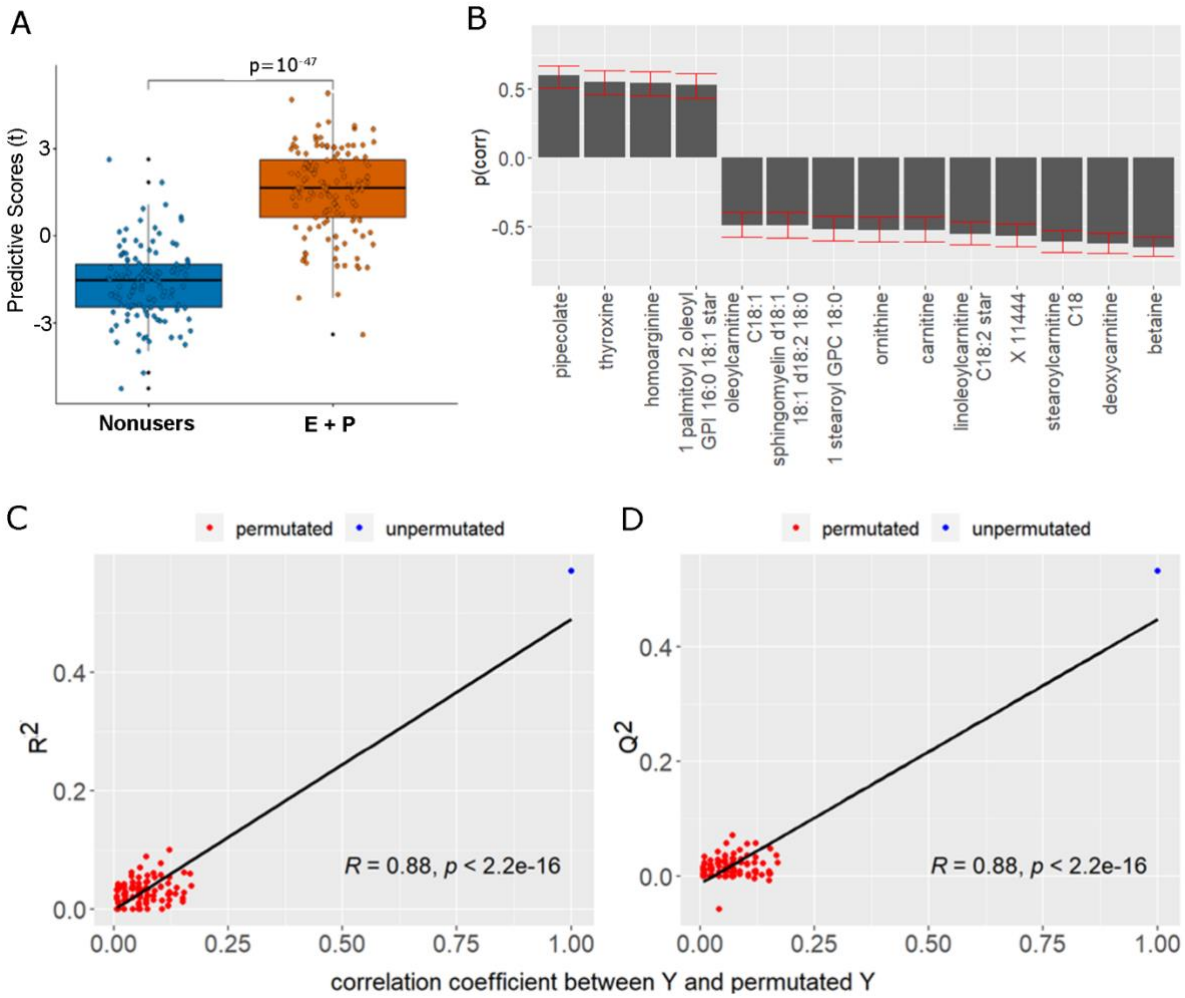

Figure 7. *Scores plot (A) and the scaled loadings (B) of E+P users versus hormone Nonusers age group 61-65 years using Modelling strategy 4. Error bars in B indicate the 95% confidence interval of p(corr). Permutation over variable selection for the model is shown for R^2 (C) and Q^2 (D), with permutated models displayed in red and unpermutated models in blue. p: Panel A: t-test of predictive scores, panels C-D: Pearson's correlation.*

Effect of Small Sample Sizes on R^2 and Q^2 and the Significance of Using Permutations

The study examined how decreasing sample size affects R^2 and Q^2, as well as its implications for significance using permutations without, post, and over variable selection. Specifically, we evaluated the robustness of a strong model comparing group E+P versus non-users against a weak model of E-only versus non-users and a comparison expected to yield no significant models of non-users versus non-users, all within the 61-65 age group. We drew equal sample sizes from each group, ranging from n=4 to n=80, and constructed OPLS models using Modeling strategy 3.

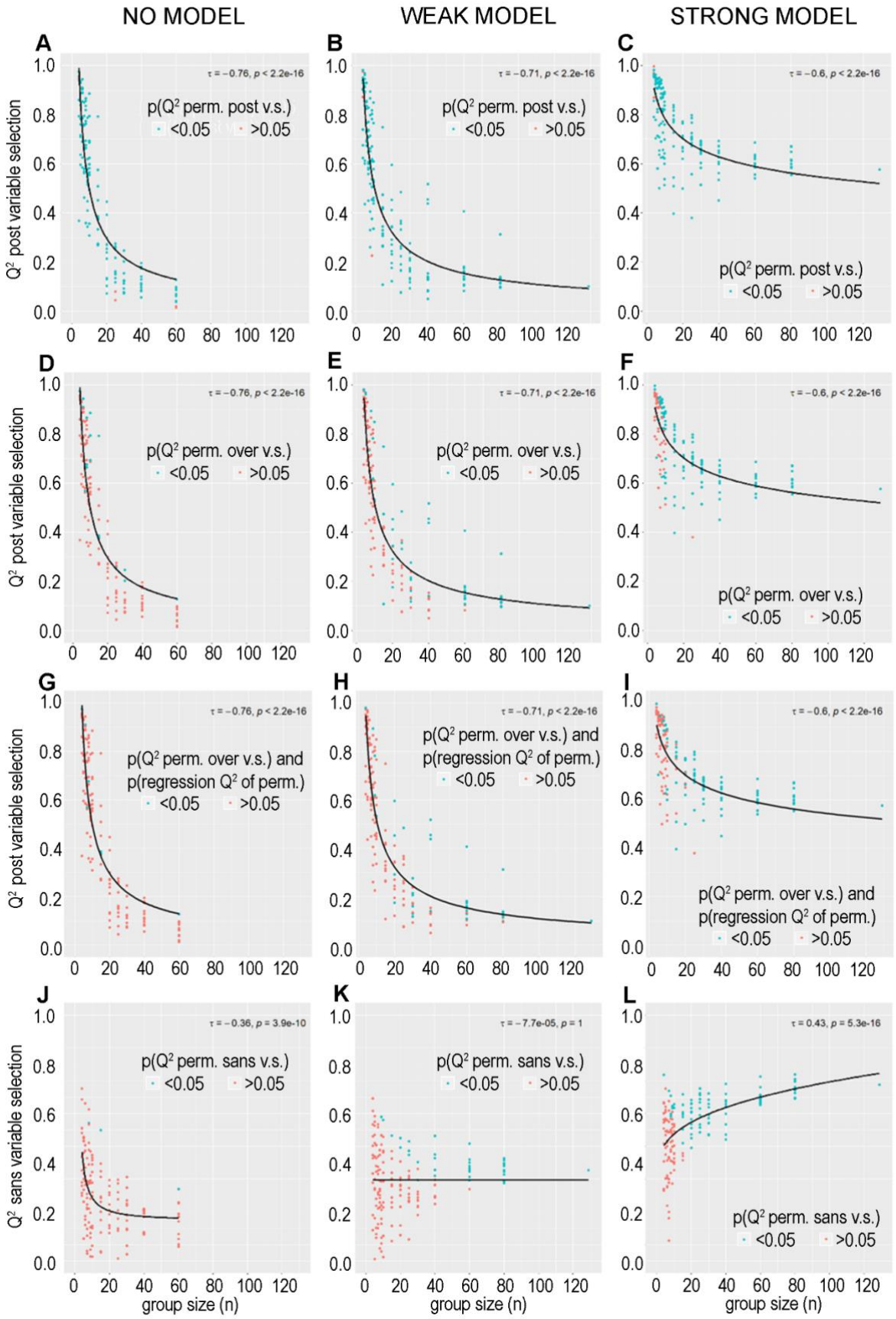

Figure 8: *Q^2 of models post variable selection using Modelling strategy 3 was significant across all samples sizes and models (A-C, colored by p[Q^2 permutated post variable*

selection]<0.05). However, the majority of the models in the weak- and non-models did not pass permutation over variable selection (D-E), whereas the majority of the strong models passed permutation over variable selection at a cutoff of n=8 (panel F). The same trend held true when also including the significance of the correlation of the permutation, but reduced the false discovery rate to 6%(panels G-I). Permutation sans variable selection displayed the opposite trend, with a decreasing Q^2 with decreasing group size in the strong model (panel L). This trend may either imply that this approach is too stringent, particularly at group sizes below n=10, or simply reflect the decreasing statistical power resulting from the smaller n. As expected, the none models did not pass the sans v.s. permutations tests (J)- The weak model required a sample size of n=40-60 to identify the models as significant using permutation test (K).

With decreasing sample sizes, Q^2 inflated, approaching 1.0 in models post variable selection for both the strong and weak models, as well as when comparing non-users to non-users (Figure 8). The trend was similar for R^2 (data not shown). In models without variable selection, decreasing sample size resulted in decreasing Q^2 in the strong model (Figure 8L) but was unaffected in the weak model (Figure 8K) and increased when modeling non-users versus non-users (Figure 8J).

When we investigated model significance using permutation post variable selection, Q^2 was significant for all models, including non-users versus non-users, regardless of sample size (Figure 8D-F).

Examining the significance of Q^2 using permutations over variable selection, the majority of the strong models remained significant with decreasing sample size (Figure 8I), while half of the weak models were significant using Modeling strategy 3, with more significance at larger sample sizes (Figure 8H). Thirteen percent of models comparing non-users to non-users were significant (Figure 8G). When considering the regression coefficient (>0.1) and p-value for the regression coefficient over the permutations (<0.05), 6% of the models comparing non-users versus non-users were significant (Figure 8J).

Assessing the significance of Q^2 by permutating the models without variable selection, the strong models remained significant as long as more than 10 subjects were used in each group. As expected, almost all models comparing non-users versus non-users (i.e., no model) remained insignificant (Figure 8J).

Establishing a Statistically Sound Cutoff for Q^2

To establish a significance level for Q^2, we estimated the overinflation of the models by comparing the Q^2 of models created from permutated data, including variable selection, to the Q^2 of models without variable selection. Additionally, we estimated the difference between the Q^2 of unpermutated models without and post variable selection. These two estimates of overinflation overlapped each other at sample sizes below 20 in this dataset (Figure 9).

We then calculated an adjusted Q^2 by subtracting the overinflation derived from permutated data from the Q^2 of the models post variable selection. This adjusted Q^2 was independent of sample size and served as a more valid measure of how well a model performs than Q^2 post variable selection. A Q^2 greater than zero indicates true variation larger than random explained by the model.

The number of orthogonal components was either decreased or not affected by decreasing sample sizes, indicating that the increase in R^2 and Q^2 was not due to the number of orthogonals (Figure 10).

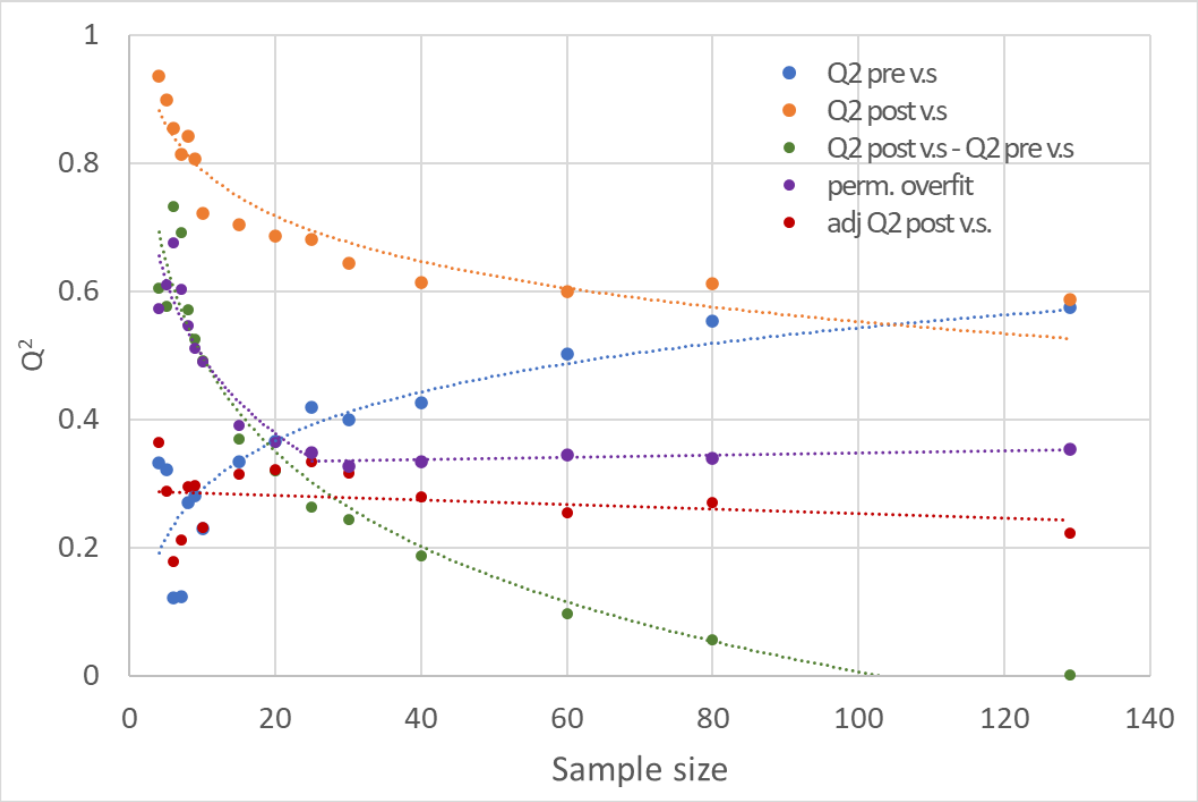

Figure 9. *The figure displays Q^2 versus group size of models comparing the groups of E+P versus Nonuser, age group 61-65. Q^2 sans v.s: Q^2 of models sans variable selection; Q^2 post v.s.: Q^2 of models post variable selection; Q^2 post v.s. – Q^2 sans v.s.: The difference between Q^2 sans and post variable selection; perm.overfit: Permutated overinflation is calculated by subtracting the average of the permutated models sans variable selection from the Q^2 of the permutated models over variable selection; adj Q^2 post v.s.: Adjusted Q^2 post variable selection is obtained by subtracting the overinflation from the Q^2 post variable selection.*

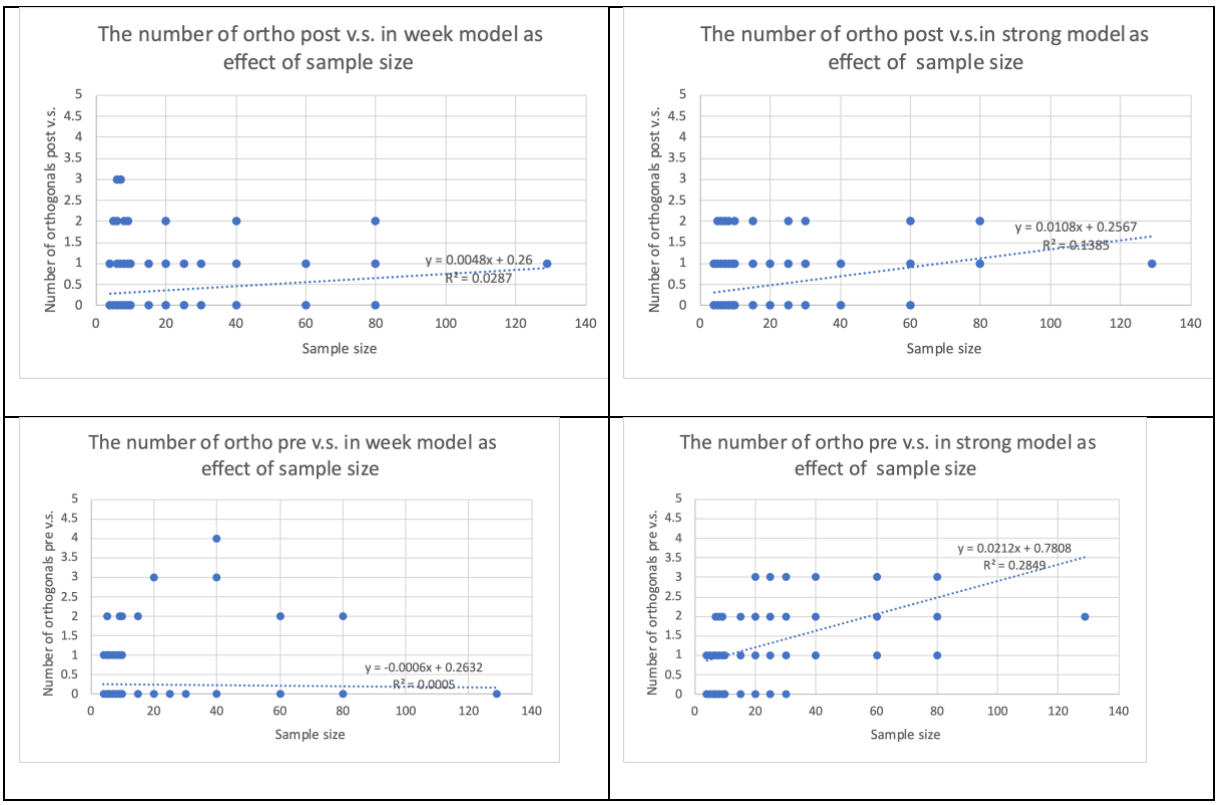

Figure 10. *The number of orthogonals post variable selection (top panels) and sans variable section (lower panels is not increased with reduced sample size in the significant models. This holds true both for the strong (right panels) and weak (left panels) models.*

Discussion

Here, we underscore the widespread utilization of OPLS modeling in various omics fields for discerning group separation significance and identifying variables driving such separation. Despite its advantages over univariate statistics, OPLS modeling faces challenges, notably overfitting, particularly in small cohorts. The determination of significance thresholds for model statistics lacks clear guidelines, posing complexity and uncertainty, especially for novices. Cross-validation (CV), particularly Q^2, serves as a gold standard for validation in small cohorts, but the applicable significance threshold remains elusive.

Lindgren et al. (1996) demonstrated the utility of permutations for OPLS models, showing that permutations over variable selection and prior to variable selection could distinguish a model from a randomly created one [10]. They also proposed estimating model overinflation by comparing Q^2 of the model fitted on permuted data without variable selection to Q^2 of the model fitted after variable selection on the permuted data. Here, we investigate the effectiveness of permutation procedures with small sample sizes, revealing that permutations before and over variable selection efficiently establish significance levels, even with smaller sample sizes.

In this study, we examined OPLS modeling performance on small sample sizes using a publicly available dataset and evaluated various permutation testing methods for assessing associated risks of overfitting or overinflation of model statistics. Our findings reveal a tendency of R^2 and Q^2 to increase with decreasing sample sizes. Notably, permutations without and over variable selection largely guard against the identification of random models as significant, while permutations post variable selection exhibit limited efficacy, particularly with small samples. Additionally, we estimated the magnitude of Q^2 overestimation resulting from variable selection, leading to an adjusted Q^2 more resilient against group size differences, thus serving as a dataset-specific significance threshold.

Furthermore, we introduced ropls-pvs, an R package and workflow incorporating variable selection and permutations without, post, and over variable selection, along with estimating overinflation and adjusted Q^2. We demonstrated its utility on the aforementioned publicly available metabolomic dataset, showcasing its automated comparison of pairwise cohort comparisons and stratified analyses. While the dataset's lack of smoking status information limits its variance assessment, the script's automation streamlines analysis, reduces subjectivity, and facilitates efficient reruns with new data or settings.

Another method suggested for testing overfitting or overinflation is the use of cross-validated score plots [19]. While it provides a more predictive picture of the score plot, it lacks a significance threshold. Future improvements to ropls-pvs could incorporate non-cross-validated score plots to address potential over-optimism.

Although ROC analysis is implemented in ropls-pvs using precrec and pROC packages, there's potential for further enhancement by integrating AUROC into significance testing through permutations to establish a threshold for significance [20].

The ropls-pvs package is designed for ease in modeling numerous group comparisons and assessing significance using stringent permutation testing, particularly beneficial for screening many groups for differences. It offers flexibility and automation, especially suited for studying ropls model performance, particularly in scenarios with small sample sizes prone to overinflated model statistics. By rigorously testing for significance, the package contributes to reducing the high rate of false findings in scientific research, thereby enhancing the robustness of OPLS models.

Modelling strategies provided by the script offer versatility for various analysis goals, from pathway enrichment to biomarker identification. The flexibility and user-friendliness of ropls-pvs, alongside its potential for parallel computing, ensure applicability across omics platforms and clinical data analysis. Overall, our study highlights the importance of robust statistical approaches and automated workflows in enhancing the reliability and efficiency of OPLS modeling in metabolomics and related fields.

Conclusion

The ropls-pvs package actively performs OPLS-DA modeling by semi-automatically optimizing models through variable selection. It also conducts permutation testing without, after, and over variable selection to establish a significance level for the predictive model statistics Q^2 of models post variable selection.

Through extensive quality control, permutation testing is conducted prior to variable selection and over variable selection to mitigate the risk of false positives. Permutation prior to variable selection ensures significant differences between groups, while permutations over variable selection ensure that models maintain significance after variable selection.

Demonstrating the utility of the package, we apply it to a publicly available metabolomic dataset. Our investigation into the effects of model on small group sizes reveals that R^2 and Q^2 increase with decreasing sample sizes. We show that both permutations without variable selection and permutations over variable selection significantly mitigate the identification of random models as significant, whereas permutations after variable selection do not.

The proposed adjusted Q^2, derived by subtracting the overinflation from the model post variable selection, remains stable across sample sizes and may serve as a threshold for the significance of OPLS models. This is essential as traditional statistical evaluation alone may not adequately discern whether a model is significantly better than a random model post variable selection.

References

1. Ioannidis JP. Why most published research findings are false, PLoS Med 2005;2:e124.
2. Ståhle L, Wold S. Partial least squares analysis with cross validation for the two class problem: A Monte Carlo study, Journal of Chemometrics 1987;1.
3. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS), Journal of Chemometrics 2002;16:119-128.
4. Wheelock AM, Wheelock CE. Trials and tribulations of 'omics data analysis: assessing quality of SIMCA-based multivariate models using examples from pulmonary medicine, Mol Biosyst 2013;9:2589-2596.
5. Triba MN, Le Moyec L, Amathieu R et al. PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters, Mol Biosyst 2015;11:13-19.
6. Westerhuis JA, Hoefsloot HCJ, Smit S et al. Assessment of PLSDA cross validation, Metabolomics 2008;4:81-89.
7. Ruiz-Perez D, Guan H, Madhivanan P et al. So you think you can PLS-DA?, BMC Bioinformatics 2020;21:2.
8. Pitman EJG. Significance Tests Which May be Applied to Samples From any Populations, Supplement to the Journal of the Royal Statistical Society 1937;4:119-130.
9. Eriksson L, Kettaneh-Wold N, Trygg J et al. Multi- and Megavariate Data Analysis : Part I: Basic Principles and Applications. Umetrics Inc, 2006.
10. Lindgren F, Hansen B, Karcher W et al. Model validation by permutation tests: Applications to variable selection, Journal of Chemometrics 1996;10:521-532.
11. Shi L, Westerhuis JA, Rosen J et al. Variable selection and validation in multivariate modelling, Bioinformatics 2019;35:972-980.
12. Wiklund S, Johansson E, Sjostrom L et al. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models, Anal Chem 2008;80:115-122.
13. Stevens VL, Wang Y, Carter BD et al. Serum metabolomic profiles associated with postmenopausal hormone use, Metabolomics 2018;14:97.
14. Calle EE, Rodriguez C, Jacobs EJ et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics, Cancer 2002;94:2490-2501.
15. Levanen B, Wheelock AM, Eklund A et al. Increased pulmonary Wnt (wingless/integrated)-signaling in patients with sarcoidosis, Respir Med 2011;105:282-291.
16. Lundstrom SL, Levanen B, Nording M et al. Asthmatics exhibit altered oxylipin profiles compared to healthy individuals after subway air exposure, PLoS One 2011;6:e23864.
17. Eriksson L, Trygg J, Wold S. CV-ANOVA for significance testing of PLS and OPLS (R) models, Journal of Chemometrics 2008;22:594-600.
18. Wakeling IN, Morris JJ. A test of significance for partial least squares regression, Journal of Chemometrics 1993;7:291-304.
19. Bevilacqua M, Bro R. Can We Trust Score Plots?, Metabolites 2020;10.
20. Szymanska E, Saccenti E, Smilde AK et al. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics 2012;8:3-16.