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

Marika Ström[a] and Åsa M. Wheelock[a, b,*]

[a] Respiratory Medicine Unit, Department of Medicine and Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden, [b] Department of Respiratory Medicine and Allergy, Karolinska University Hospital, Stockholm, Sweden

[*] Correspondence to be addressed to

Åsa Wheelock, E-mail: asa.wheelock@ki.se

Wheelock lab, BioClinicum J7:30, 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

Many inflammatory diseases are polygenic with a multitude of genetic, environmental, and life style factors influencing the development of a wide range of phenotypes. To investigate these complex diseases, it can be advantageous to investigate a large number of variables from multiple anatomical locations and molecular levels in a multi-omics fashion. However, in clinical cohorts where invasive sampling is utilized, such as investigations on respiratory inflammatory diseases such as is performed in our own laboratory, The use of multivariate statistical modelling in omics research can be a powerful tool for biomarker discovery, as well as for extraction of relevant variables prior to more mechanistic investigations such as pathway enrichment analysis. In pulmonary research the local inflammation and injury is studied by samples collected from the lungs using bronchoscopy. The invasiveness of this procedure results in the number of subjects studied often being low. The resulting small sample sizes combined with a large number of analytes in omics investigations pose a risk for a false discovery rate [1]. One way to increase the power but keep the rate of false discoveries low is to use multivariate analysis. Principle component analysis (PCA) is an unsupervised method widely used to uncover structures in the dataset and find outliers. Partial least squares (PLS) is a multivariate modelling method [2] widely 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 in order to facilitate efficient supervised modeling of group separation.

OPLS has proven to be an efficient tool for analysis of omics data sets [4]. The model statistics provided by the original algorithms in the SIMCA software 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). However, the resulting OPLS model can also be used to narrow the number of variables from the hundreds-to-thousands of variables typically constituting an omics data set, to a subset of features driving the group separation that are of interest as putative biomarkers or molecular targets, a process referred to as variable selection or feature selection. Variable selection is frequently used using univariate methods, e.g. using random forest, support vector machine learning, or simply an alfa-cutoff based on t-statistics corrected for multiple testing. In OPLS, primarily two different variable selection metrics have been proposed; Variable Importance in the Projection (VIP), and the scaled loadings p(corr). VIP is simply a ranking of the variables where 1.0 represents the average contribution to group separation. The relative nature and high dependence on the number of variables included in the respective model makes the VIP metric difficult to use in variable selection, particularly when comparing across models. p(corr) is a more robust metric as it is related to the loadings of the predictive variable, and the fact that it is scaled as a correlation coefficient always ranging between -1 and 1 makes it easier to compare between models. Here, we propose to use 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 and describes the variation of the dataset at hand 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 model performance [5].

It is well known that OPLS models are prone to overfitting [6]. Adding additional principal components to a model increases R^2 and can always create a complete separation between the groups but on adding principal components the predictability Q^2 decreases indicating over-fitting of the model. Especially, decreasing the ratio between number of samples and features result in increasing separation resulting in perfect separation of a dataset when the ratio is low and has been estimated to 1:2 or below [7]. Also, during variable selection R^2 and Q^2 increase and the extent of the increase is partly due to overinflation of the model statistics. Due to the large risk of overfitting OPLS models, it is necessary to test the models for significance. Common methods to test OPLS models for significance is using cross validated analysis of variance (CV-ANOVA) and accuracy, accuracy in this context being defined as the proportion of the subject group belonging being 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 of the data set are scrambled prior to modeling the dataset. The procedure is iterated, and the resulting model statistics are representative of models that can occur by chance. Permutation tests has been implemented in tools for multivariate analysis such as the widely used SIMCA software (Sartorius) [9] and in the R package ropls. This procedure is commonly performed either prior to variable selection, i.e., by including all variables in the full dataset to calculate the significance of R^2 and Q^2, or post-variable selection, i.e., by including only the final selected variable of interest in the permutation test. Whereas the latter assures that the selected variables cannot produce a significant separation of groups of subjects randomly picked from cohort at hand. However, the limitation of this commonly used approach is that is only controls for the variables of interest and does not assess if the alterations represent random variation.

A more rigorous test of significance is to perform permutation testing including the variable selection step, sometimes referred to as permutation test over variable selection. During this procedure permutations are initially performed in the same way as for sans-variable selection. Subsequently, variable selection is performed on every permutated dataset before fitting a model to the permutated data. Permutations including variable selection has been investigated on OPLS models [10] and has been implemented in the MUVR package for PLS models [11].

By comparing model statistics of the permutated models to the model statistics of the unpermutated model to be tested, the significance of the model can be estimated by calculating the ratio between permutated models above and below the unpermutated reference model (nonparametric p-value). This method is used in the ropls package yielding p-values for R^2 and Q^2.

The aim of this study was to develop an R script roplspvs (**R o**rthogonal **p**rojections of **l**atent **s**tructures with **p**ermutation over **v**ariable **s**election) that uses the Bioconductor package ropls for the OPLS modeling and adds variable selection as well as permutations both sans, post and over variable selection. We also aimed at establishing a significance level for the predictive power of Q^2 of models post variable selection by comparing to the estimated overinflation of the models caused by reducing the number of variables (variable selection).

This estimated overinflation represents a threshold for Q^2, with models obtaining a Q^2 above this threshold considered not to have occurred by chance. Subsequently, an adjusted Q^2 was calculated by subtracting the estimated threshold, resulting in a cutoff for significance of 0. In this paper the script is used to study how R^2, Q^2, overinflation and the proposed adjusted Q^2 is affected by decreasing sample sizes, using a publicly available metabolomics dataset. We also investigate whether permutation sans, post and over variable selection can distinguish between a significant model versus a model produced from random data at different sample sizes. As OPLS models are prone to overfitting, particularly post variable selection, a tool that can contribute to avoiding publishing insignificant models is desperately needed, and may be counteract the well-established issue of irreproducibility in research.

Methods

The roplspvs package

Roplspvs is an R package and workflow developed to facilitate semi-automated pairwise group comparisons in a dataset using OPLS modeling based on the primary group belonging (e.g., treatment vs placebo groups), with the option to stratify the analysis based on additional variables of interest (e.g., sex, smoking status). In order to provide robust, reliable model optimization, the purpose was to include thorough test for significance by permutations including variable selection.

OPLS analysis is performed using the ropls R package. For producing html reports with plots, it is also dependent on rmarkdown, tools, ggplot2, ggrepel, kableExtra, gridExtra, ggpubr, matrixStats, stringr, tryCatchLog, devtools, DescTools, precrec, pROC and rstatix.

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

Workflow

The workflow of roplspvs consists of a run fil (Run.R) that runs the oplspvs function, (Figure 1). A subset of user-defined settings can be modified through two different configuration files: The configure file Configure_Get_Started.R consists of basic parameter settings that contains information about the data matrix such as variable names, preprocessing information, as well as information about the observations such as sample ID, primary group belonging, and metadata to be used for stratification in the OPLS analyses. In addition, the

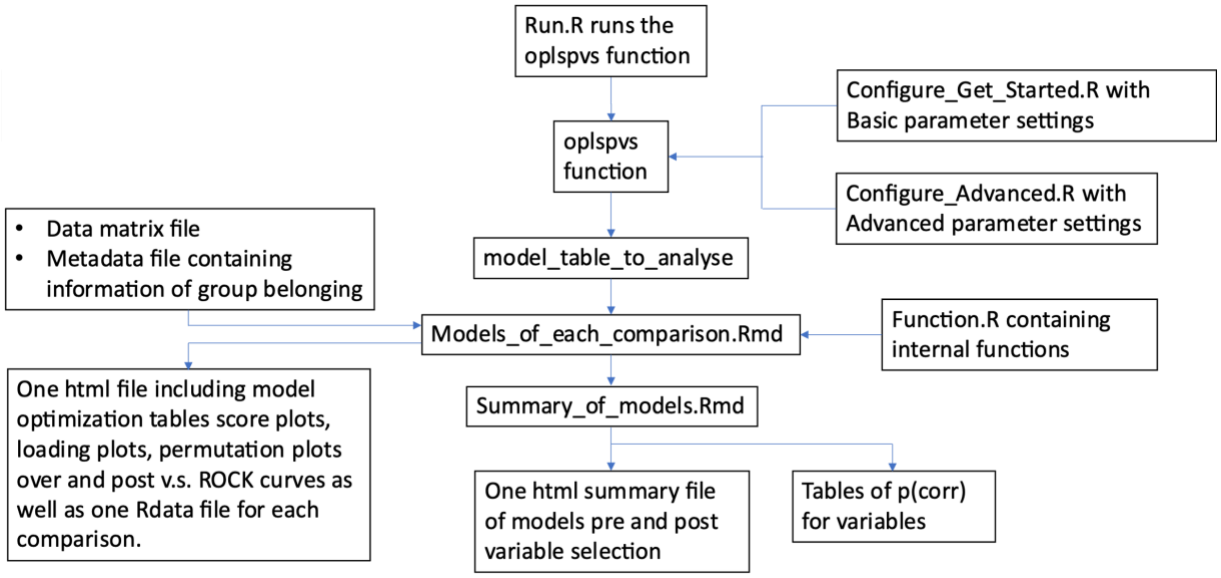

Figure 1. *Schematic of the roplspvs workflow and package*

Configure_Get_Started.R file also contains information regarding the number of permutations desired, the p(pearson) significance level to be used for the p(corr) cutoff, and the order of groups to define the directionality in the OPLS score plots (i.e., healthy vs. diagnosis or diagnosis vs healthy), as well as the ability to restrict what pairwise comparisons to run (the default being pairwise comparisons of all groups in the sampleID file). The Configure_Advanced.R file contains information about the default settings of directories and folder names for output files, which can be adjusted if an alternate folder structure is desired. This Configure_Advanced.R file also provides the option for the user to alter the variable selection settings, set the maximum amount of orthogonals allowed, to alter the settings for selecting best performing model, and settings to adjust length and part of variable names to be displayed in the OPLS loadings plots. Finally, it is also possible to choose which modelling strategies to apply (the default being 0-5), and whether both comparisons and summary output files are to be generated.

In a first step, the roplspvs script generates a table (model_table_to_analyse) containing information about each pairwise comparison to be performed. Subsequently, parameters from model_table_to_analyse and from the two configure files are sent to the Rmarkdown files. The data matrix, metadata file and functions from roplspvs_Functions.R are sourced into the Rmarkdown files to perform the modeling to produce an Rdata file and an html output file for each comparison containing score plots, loading plots, permutation plots over and post variable selection as well as Receiver-operator-characterstics (ROC) curves. Finally, html files are created summarizing the model statistics for all pairwise comparisons, and displaying variable lists with associated p(corr) values for variables selected by the respective modelling strategies. An overview of the input- and output files of roplspvs is provided in Figure 1.

Preprocessing of dataset

The dataset is pre-processed according to the schematic in Figure 2. The pre-process is performed iteratively for each pairwise comparison in order to optimize the variables considered for each comparison. It includes the options of replacing zero values with NA or the lower limit of detection (LLD), followed by an optional replacement of NA with LLD. The LLD can either be set by the user, or be estimated as 1/3 of the minimum value in the dataset. The data filtering step removes variables according to a user-defined missing value tolerance level for each group in the comparison.

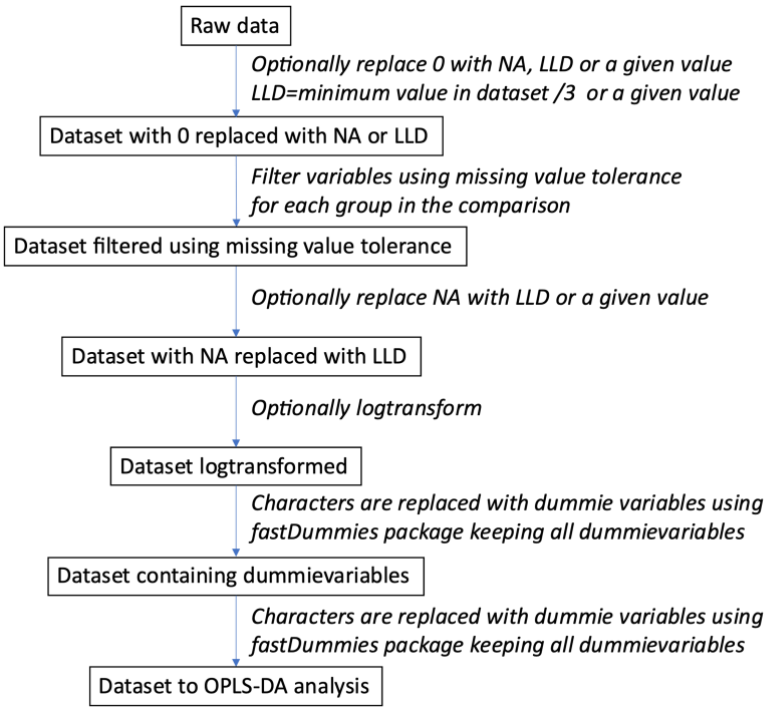

Figure 2. *Schematic of data pre-processsing options in the ropls package.*

OPLS and PCA modeling

The Bioconductor R package ropls (https://bioconductor.org/packages/release/bioc/html/ropls.html) is used to perform the initial PCA and OPLS modeling {Thevenot, 2015 #28}. The default setting of OPLS is applied, performing NIPALS (Nonlinear Iterative Partial Least Squares) for missing data if NAs are still present in the dataset. Mean centering and scaling to unit variance (UV) also allows variables with low abundance or amplitude to contribute to the model.

PCA models are created prior to variable selection for each comparison, generating both PCA models including all observations in the comparison, as well as PCA models for each group individually to facilitate strong outlier identification. OPLS models are created for each comparison both sans variable selection by including all variables that pass the preprocessing QC criteria, as well as post variable selection by including only the variables selected by the respective Modelling strategy (see section below). The models are evaluated reporting the model statistics R^2, Q^2 and RMSE as well as significance of the models established using permutations.

Variable selection

Feature selection, a.k.a variable selection, is applied using five different modelling strategies and/or stringency levels.

The default variable selection procedures in all model strategies are based on p(corr), with the option of adding VIP as an additional variable selection metric. Different p(corr) cutoffs for variable selection are used in the 5 strategies. The p(corr) cutoff is mainly determined by

iteratively optimizing the predictive power, Q^2, of the model post variable selection. This process ensures that overfitting is avoided, thereby maintaining the maximum predictive power of the model post variable selection. An overview of the available modelling strategies is presented in Figure 3.

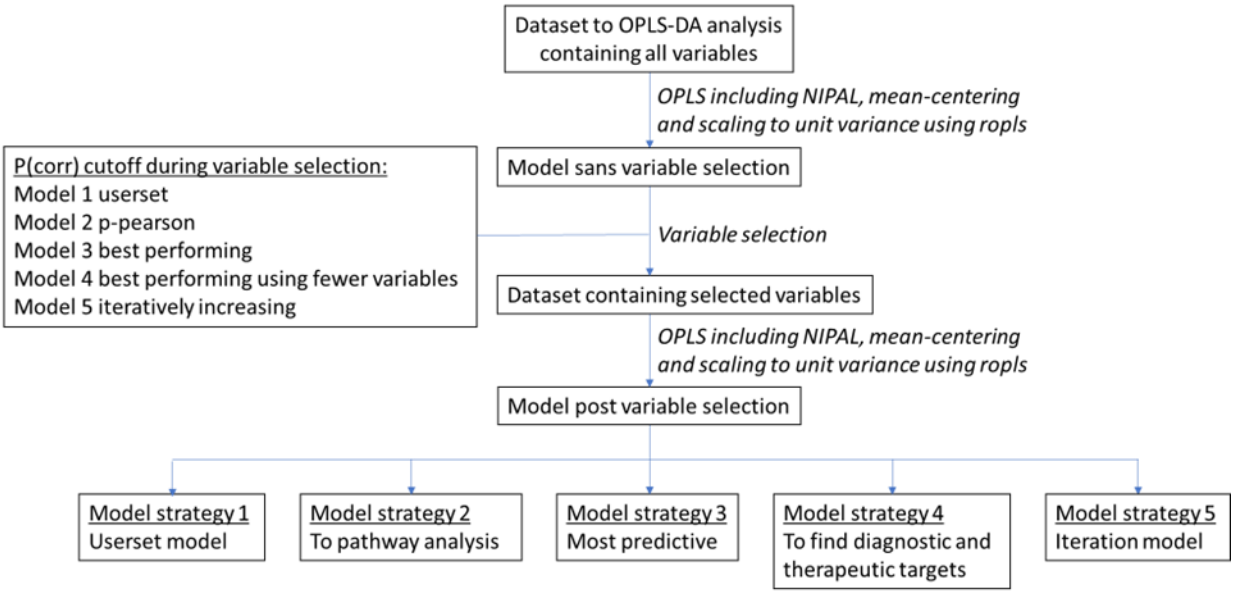

Figure 3*. Overview of the five different Modelling strategies available in the roplspvs package. The p(corr) cutoff used in the respective modelling strategies are indicated in the text box to the left.*

In Modelling strategy 1, the p(corr) cutoff may be set by the user. This enables the user to rerun the script using other settings for the model than the script has selected, or to evaluate the significance of a model optimized in another software by using the permutation analyses provided by the roplspvs script. The p(corr) cutoff used in the variable selection for Modelling strategy 2 is set to include all variables with significant correlation to the score. This p(corr) cutoff is set by a user set p-value for the p(corr) cut off. The p-value corresponding to the Pearson's correlation coefficient for the corresponding group size. This p(corr) cutoff is also used as the default p(corr) cutoff for Modelling strategy 1 if no other cutoff is defined. If, instead of including many variables, a model that is best performing is desired Modelling strategy 3 should be used. It results in the best performing model post variable selection here defined as models that after variable selection give high Q^2, low difference between R^2 and Q^2 and also low p[R2_perm_post_vs] and low p[Q2_perm_post_vs]. The highest Q^2 corresponds to the most predictive model. Low difference between R^2 and Q^2 is a means of avoiding overfitting, as overfitting a model increases R^2 without increasing Q^2 to the same extent. Finally, low p[R2_perm_post_vs] and low p[Q2_perm_post_vs] avoids selecting a model that is not significantly better than random. The

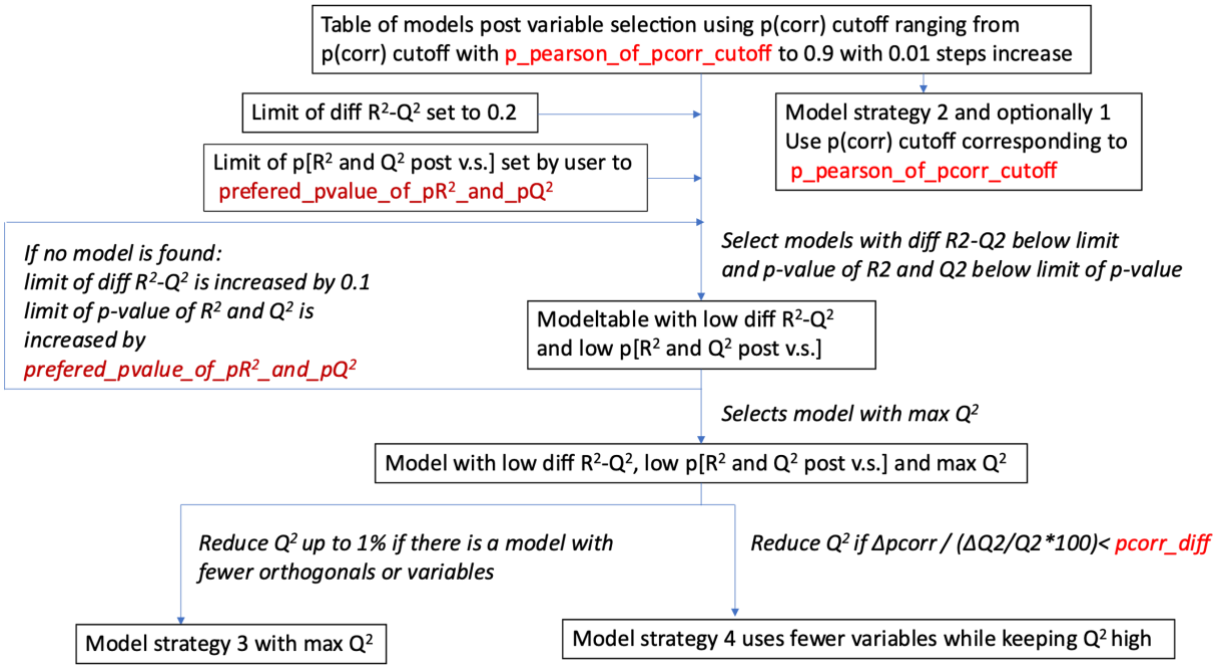

***Figure 4**. Details for selecting p(corr) cutoff of Modelling strategies 3 and 4. Parameters that are user-defined in roplspvs are marked in red.*

weight between low difference between R^2 and Q^2 and low p[Q2_perm_post_vs] may be set by the user through the variable prefered_pR2_and_pQ2_permutated_post_vs with lower value giving more weight to having low p[R2_perm_post_vs] and p[Q2_perm_post_vs]. Decreasing the number of variables selected even further is useful in the search for diagnostics and potential therapeutics. Modelling strategy 4 keeps the number of selected variables to a minimum by extracting the variables most contributing to the model without losing too much predictability. P(corr) cutoff is selected to give the best performing model post variable selection increasing p(corr) cutoff as long as Q^2 is not decreased more than 1% in the model post variable selection. This procedure result in a model that includes fewer variables. How much the number of variables is reduced is set by how much p(corr) is allowed to increase when Q^2 is reduced by 1% and may be set by user in the variable named pcorr_diff. If fewer variables are desired or if there are many variables in the original model, pcorr_diff should be larger. The details of the selection of p(corr) cutoff are shown in Figure 4.

To enable the user to evaluate the optimization of p(corr) cutoff a plot of Q^2 versus number of variables using different p(corr) cutoff is created, showing the selected p(corr) cutoff for Modelling strategies 3 and 4. This together with a model table enables the user to easily verify that the script has selected the best performing model.

The traditional way of performing variable selection is to use an iterative procedure either by selecting the most predictive variables and adding additional variables if the model is improved or by deselecting variables if the model is improved. After adding or deselection variables the model is refitted and the procedure is repeated until the model cannot be improved further. Modelling strategy 5 represent an iterative model using increasing p(corr) cutoff and refitting the model each time p(corr) cutoff is increased by 0.01 units resulting in

removing the variables least correlated and as such least driving the model. P(corr) cutoff is increased as long as Q^2 of the model post variable selection is not decreased more than 1%.

Selection of number of orthogonal variables

The amount of orthogonals is set by the user for Modelling strategy 1 with default set to 0 and for model strategies 2-5 the number of orthogonals is set by using the ropls default method adding an orthogonal as long as Q^2 is increased by 1 % in the model currently evaluated. The maximum amount of orthogonals for both model sans and post variable selection may be set by user. The script also check that no strategy produces a better performing model post variable selection using fewer number of orthogonals.

Permutations

Significance of models are evaluated using permutations sans, post and over variable selection Figure 5.

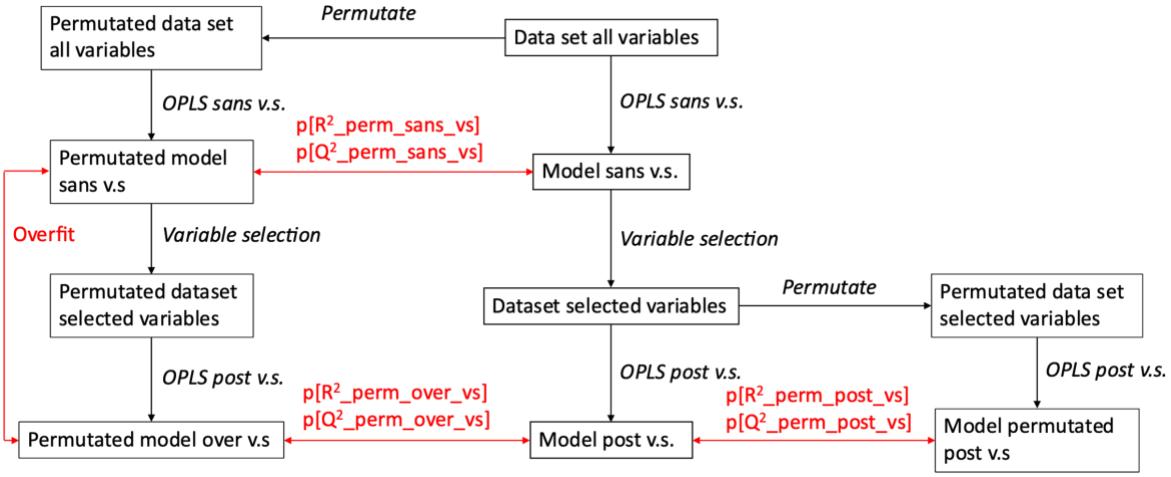

Figure 5. *Estimations of the significance level of each model statistics is provided based on permutation test at three levels; prior to variable selection (sans v.s.), after variable selection (post v.s.), as well as over variable selection (over v.s.) p[R2 and Q2 for permutations sans, post and over 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.*

Permutations sans and post variable selection are performed using the ropls package and includes randomization of group belonging of subjects. The permutations are here performed allowing for a by random correct set of group belongings. During permutations sans variable selection, the dataset including all variables resulting from the preprocessing procedure are modeled against permutated group belonging while permutations post variable selection is performed by modeling the dataset only including variables selected during variable selection against the permutated group belonging. The model statistics R2 and Q2 of the permutated models are compared to the R2 and Q2 of the original unpermutated models resulting in p-values of R^2 and Q^2 sans variable selection and post variable selection, here called p[R2_perm_sans_vs] and p[Q2_perm_sans_vs] respectively p[R2_perm_post_vs] and p[Q2_perm_post_vs].

As performed in permutations sans variable selection, the permutations over variable selection are performed by modeling the dataset including all variables derived from the preprocessing

procedure. Each permutated model undergoes variable selection using the same p(corr) cutoff as the unpermutated model under investigation. The model statistics R^2 and Q^2 derived from the permutated models post variable selection are compared to model statistics of the unpermutated models post variable selection. The proportion of the permutated model compared to the unpermutated model that has higher R2 and Q2 result in the p-value for R^2 and Q^2 over the variable selection p[R2_perm_over_vs] and p[Q2_perm_over_vs].

The permutation procedure over variable selection determines the probability that using the variable selection procedure would render a significant model by chance. One limitation of the permutations over variable selection is that the settings of the unpermutated model is used for the permutated models. This means that the p(corr) cutoff is optimized for the unpermutated model but not for the permutated models. This may result in an increased significance for p[R2_perm_over_vs] and p[Q2_perm_over_vs] compared to if p(corr) was optimized also for the permutated models. To avoid this the regression over R^2 and Q^2 of permutated models may be taken into account. The R^2 and Q^2 of the permutated models are plotted versus correlation of permutated and unpermutated group belonging. For a significant model the regression should result in a significant positive correlation coefficient since the higher correlation to original group belonging should result in higher R^2 and Q^2. Here we evaluate a limit for the significance of 0.1 for the coefficient and 0.05 for the p-value of the regression. Sometimes the use of the optimized p(corr) cutoff result in no variables selected in the permutated dataset and thus no model is created. When a permutation renders no model, R^2 and Q^2 is set to 0 for this model when including it in the regression. No model is also deemed poorer than the unpermutated model during calculation of p[perm. over v.s.].

The number of permutations performed on the models during model optimization and on optimized model may be set by user, with the default set to 20 permutations representing the lowest number that may result in a p[perm.] of 0.05 in order to decrease the computational time for initial analysis. The amount of permutations should however be increased for the final analysis.

Establishing a statistically sound cutoff for Q^2

The overinflation of the OPLS models due to variable selection was established by the difference in Q^2 of the permutated models over variable selection to the Q^2 of the permutated models sans variable selection. The overinflation of Q^2 represents a robustness threshold for Q^2, and models above this Q^2 cutoff can be considered not to have occurred by chance. An *adjusted Q^2* was calculated by subtracting the estimated overinflation of Q^2 from the Q^2 of the model post variable selection. Accordingly, models with an adjusted Q^2 above zero can be considered not to have occurred by random.

Figures

The ropls package is used to produce scores plots as well as loading-, prediction-, diagnostic- and outlier-plots as well as plots of permutations sans and post variable selection.

Plots produced by roplspvs includes score plots of predictive component shown as boxplots and p(corr) plots of variables driving the separation of model showing all variables and if the number of variables exceeds 50, the 50 most driving variables are shown. The optimization of the models is illustrated by comparing the number of variables to Q^2 post variable selection. Plots of permutations over variable selection described in section above is produced by roplspvs.

After the models and html files have been created, they may be compared in shared and unique structures (SUS) plots. SUS plots are correlation plots of p(corr) for each variable of two models [12]. To facilitate inclusion of the variables that are unique for one model, new models including all variables from both models to be compared are produced. However, p(corr) from the optimized models are displayed in the resulting SUS plot. The SUS plot enables identification of variables that have similar or different effect when comparing the groups and may thus extract information about possible interactions or confounders in the comparison.

Qualitative variables

Qualitative data in the dataset are converted into dummy variables, which is the same procedure used in SIMCA. It is common to exclude one of the dummy variables in an analysis as using all contains redundant information. Here all dummy variables are used in the modeling which has the advantage of showing how all settings affect the outcome but may render a slight overfit. Being able to have qualitative variables in the analysis is useful when creating models using clinical data.

Cohort and dataset

The untargeted metabolomics dataset MTBLS136 [13] from the Cancer Prevention II Nutrition Cohort [14] studying metabolites in postmenopausal women was used to demonstrate the roplspvs package. The data set is publicly available at Metabolights (www.ebi.ac.uk/metabolights/MTBLS136/files). The dataset consists of blood metabolome profiled generated on the Metabolon platform to investigate post-menopausal hormone treatments in women. In the cohort of 1336 women, 332 received estrogen therapy alone (E-only), 337 received a combined estrogen and progestin treatment (E+P), and 667 women received no hormone replacement therapy (nonusers). In the analyses, the women were stratified into age groups less or equal to 55, 56-60, 61-65, 66-70, 71-75, 76-80 and 81-85. Values of 0 in the dataset were replaced by NA, followed by filtering allowing for 25% missing values and log-transformation.

Comparing hormone users to nonusers using roplspvs

OPLS modeling using roplspvs was demonstrated using on the metabolomic dataset MTBLS136 comparing progestin plus estrogen users (E+P), estrogen only users (E-only) and hormone nonusers (nonusers) versus each other and stratifying by age groups. A maximum of 5 orthogonals in the models both sans and post variable selection was set reducing risk of overfitting due to too many orthogonals. The variable prefered_pR2_and_pQ2_permutated_post_vs was set to 0.05 balancing between keeping p[R2 and Q2 post v.s.] and difference between R^2 and Q^2 low while maximizing Q^2. Pcorr_diff used for restricting the number of variables in Modelling strategy 4 was set to 0.01.

Effect of small sample sizes on OPLS models

To study the effect of sample sizes on OPLS models the roplspvs package was used. MTBLS136 dataset was used avoiding any unrelated variability depending on gender and age in the models. The age groups of 61-65 years was selected due to the relatively large n (n_{E+P} = 129, $n_{E\text{-only}}$= 91 and $n_{nonusers}$=121), as well as the homogeneous nature of the groups as evaluated by PCA. A strong model comparing Estrogen plus progestin users to postmenopausal hormones nonusers with Q^2=0.53, significant by permutation tests at all levels (p[Q2_perm_ post_vs] ≤0.01, (p[Q2_perm_over_vs] ≤0.01), and one very weak model

with Q^2=0.10, but still passing the significance level by permutation test at all levels (p[Q2_perm_sans_vs] ≤0.01, p[Q2_perm_ post_vs] ≤0.01, p[Q2_perm_over_vs] ≤0.01). Finally, also hormone nonusers were compared to nonusers to verify that models were considered insignificant when there should be no difference between the groups. A range of subsets of subjects from the three models were drawn using the same sample size for each group. The number of subjects included in the subsets spanned from 4 to 80 for each group (4 to 60 for each group of nonusers) and were repeated 12 times for each subset. Modelling strategy 3 and 4 were applied on each subset using roplspvs. R^2 and Q^2 for models both sans and post variable selection was plotted against sample sizes and the models were colored according to p[R2 and Q2 perm. sans, post and over v.s.]. The number of significant models from the three comparisons were compared to investigate whether the permutations could be used to discriminate between a significant models and models produced from random data using decreasing sample sizes.

The average overinflation and adjusted Q^2 was determined for each sample size for the strong model.

Results

Comparing hormone users to nonusers using roplspvs

The utility of the roplspvs package was demonstrated using a publicly available data set investigating the blood metabolite composition in relation to hormone replacement therapy in post-menopausal women. Initially, pairwise comparisons were performed for all contrasts of the three study groups of estrogen-and-progestin users (E+P), estrogen-only users (E only), and subjects reporting no hormone replacement use (nonusers), with analyses performed both for all age groups jointly (joint), as well as stratified by age groups using 5-year bins. Age groups age 81 and above were removed due to the small sample sizes.

A summary of the model statistics for all pairwise comparisons using Modelling strategy 4 is shown in Table 1. All models using all Modelling strategies were significant with regards to permutation post-variable selection p[Q2_perm_post_vs]. When applying the more stringent approach of permutation testing over variable selection p[Q2_perm_over_vs], the majority of the comparisons were still significant for all Modelling strategies, with the exception of E+P versus E-only age group 56-60 which was insignificant using Modelling strategy 1 and 2 as well as age group 76-80 using Modelling strategy 1 through 4. Also, p[Q2_perm_sans_vs] indicated that all but the same two comparisons E+P versus E-only age group 56-60 and 76-80 were significant. The two insignificant comparisons (age group 56-60 and 76-80, E+P versus E-only) were also insignificant regarding p-value for correlation of permutations using Modelling strategy 1 through 4. Both regarding models sans and post variable selection for Modelling strategy 2 through 4, all models over all age groups of E+P versus E-only had low Q^2 (<0.4) while models E+P and E-only versus nonusers had high Q2 (>0.4). Adjusted Q2 was low (<0.1) for all models of all age groups comparing E+P versus E-only.

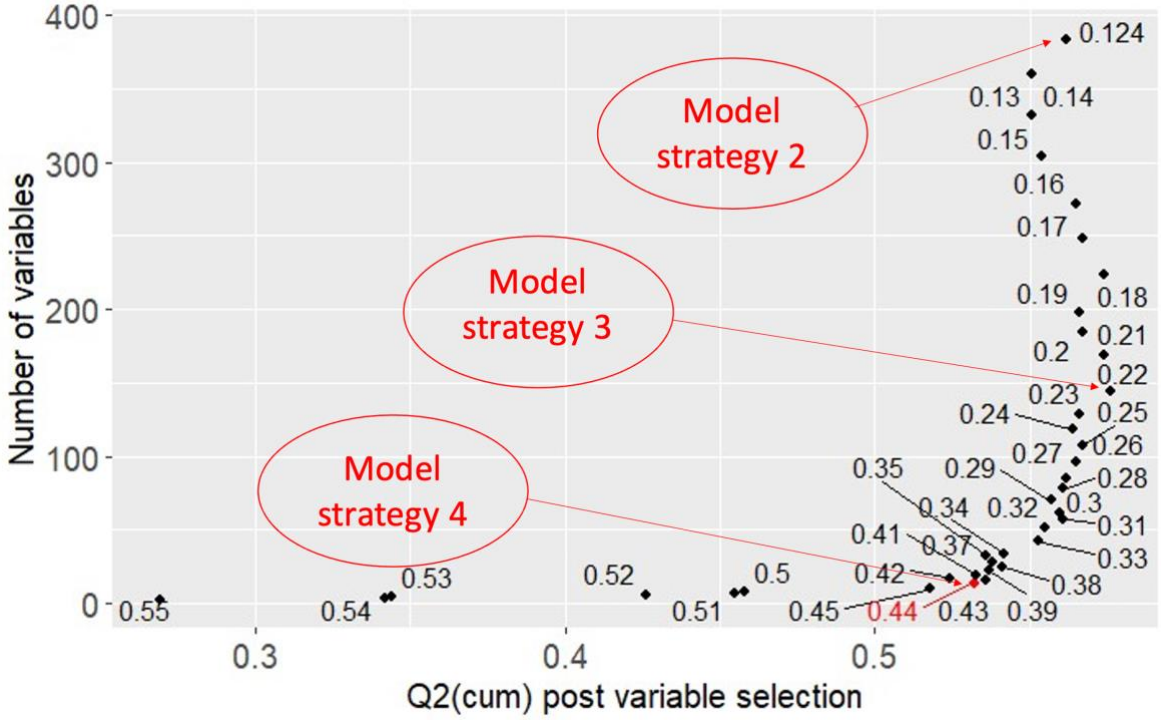

Figure 6. *Visualization of the optimization of p(corr) cutoffs for models comparing E+P versus Nonuser, 61-65 ages using Modelling strategy 2-4.*

To be able to follow the iterative model optimization step based on p(corr), the roplspvs package outputs a plot visualizing the Q^2 post v.s. versus the number of selected variables in the respective model, with the p(corr) cutoff for the model indicated by the data point, as shown in Figure 6 for optimization of the contrast E+P versus Nonuser, age group 61-65 using Modelling strategy 2 through 4. Modelling strategy 2 used p(corr) cutoff 0.124 which selected 384 variables with Q^2=0.56. P(corr) cutoff 0.22 for Modelling strategy 3 selected 144 variables resulting in Q^2=0.58 and finally p(corr) cutoff 0.44 using Modelling strategy 4 decreased the number of selected variables to 14 and resulted in Q^2=0.53.

The optimized OPLS model for comparison of E+P users versus hormone Nonusers, age group 61-65, when using Modelling strategy 4 is displayed in Figure 7. The separation of the two groups was highly significant using t-test ($p=10^{-47}$) based on the predictive component (Figure 7A). The corresponding permutation over variable selection was highly significant both for R^2 and Q^2 both in terms of p[R2_perm_over_vs], p[Q2_perm_over_vs] as well as for the correlation to the correlation coefficient between Y and permutated 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.

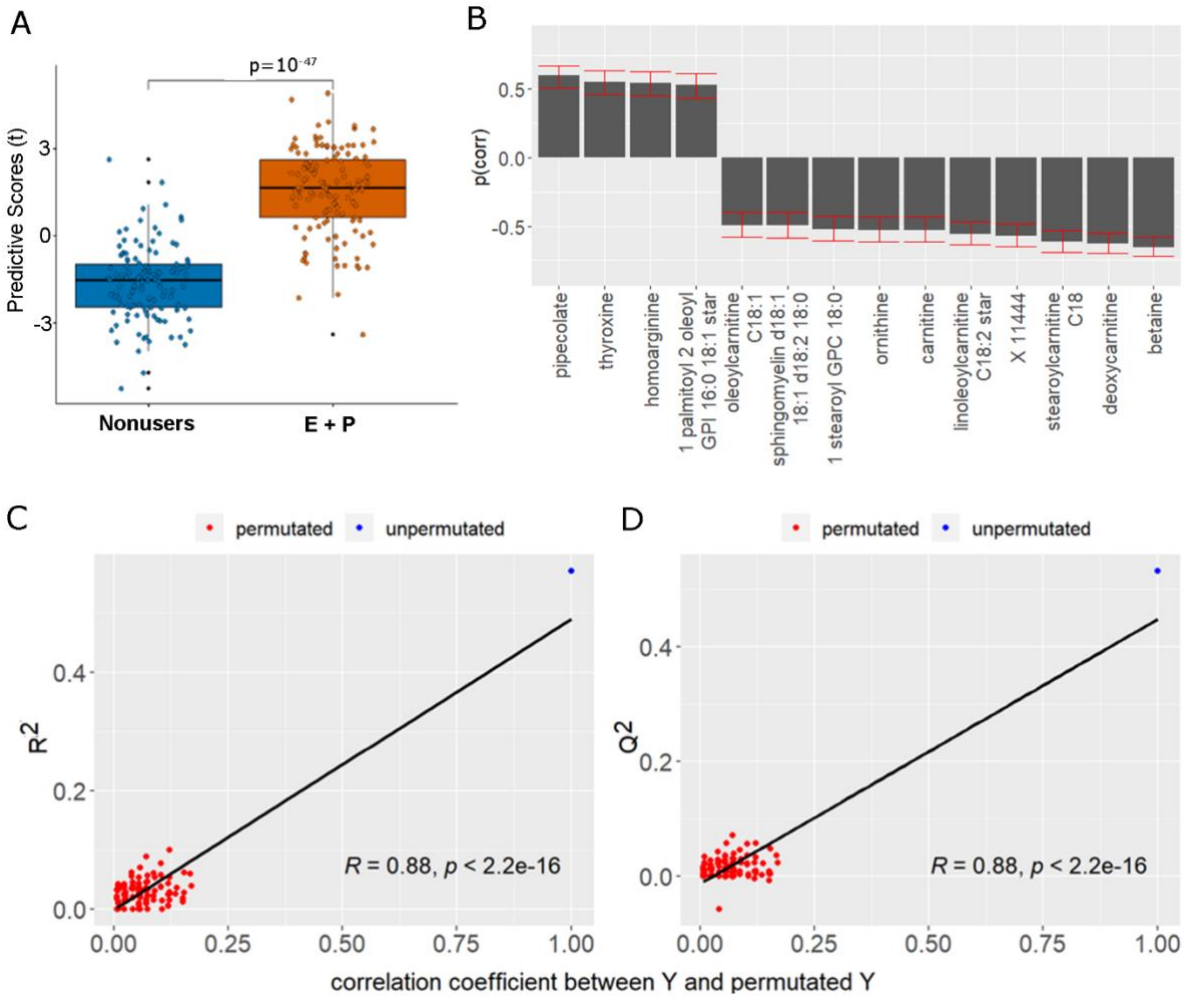

Figure 7. *Scores plot (A) and the scaled loadings (B) of E+P users versus hormone Nonusers age group 61-65 years using Modelling strategy 4. Error bars in B indicate the 95% confidence interval of p(corr). Permutation over variable selection for the model is shown for R^2 (C) and Q^2 (D), with permutated 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 using permutations

The effect of decreasing sample size on R^2 and Q^2, as well as the corresponding effects on the significance using permutations sans, post and over variable selection was investigated. The strong model of group E+P versus Nonuser was compared to a week model of E-only versus nonusers as well as a comparison that should render no significant models of nonusers versus nonusers, all at age groups 61-65. An equal number of samples were drawn from each group with sample sizes ranged from n=4 to n=80, and OPLS models were created using Modelling strategy 3.

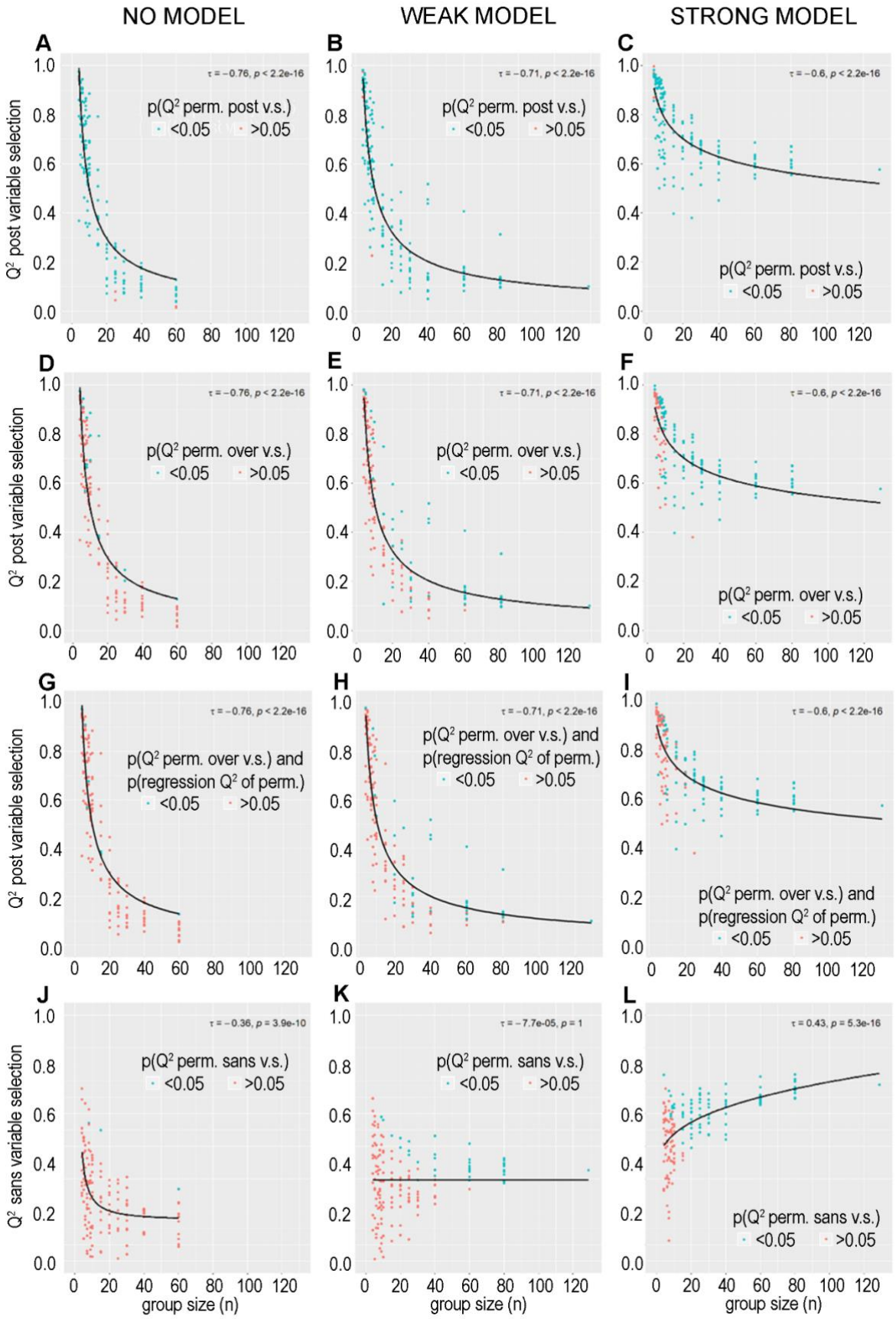

Figure 8: *Q^2 of models post variable selection using Modelling strategy 3 was significant across all samples sizes and models (A-C, colored by p[Q^2 permutated post variable*

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 was inflated, approaching 1.0 in models post variable selection for both the strong and the weak models as well as when comparing nonusers to nonusers (Figure 8). The trend was the same for R^2 (data not shown). In the model sans variable selection, using decreasing sample size resulted in decreasing Q^2 in the strong model (Figure 8L), but was not affected in the weak model (Figure 8K), and was increasing when nonusers versus nonusers were modeled (Figure 8J).

If model significance was investigated using permutation post variable selection, Q^2 was significant for all models also comparing nonusers versus nonusers, independently of sample size (Figure 8D-F).

Investigating the significance of Q^2 using permutations over variable selection, the vast majority of the strong models stayed significant when decreasing sample size (Figure 8I), while half of the weak models were significant using Modelling strategy 3, with more significance at larger sample sizes (Figure 8H). 13% of models comparing nonusers to nonusers were significant (Figure 8G). When also the regression coefficient (>0.1) and p-value for the regression coefficient over the permutations (<0.05) were considered, 6% of the models comparing nonusers versus nonusers were significant (Figure 8J).

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

Establishing a statistically sound cutoff for Q^2

In order to establish a significance level of Q^2, we estimated the overinflation of the models by regarding the models created from permutated data including variable selection. This was performed by comparing Q^2 of permutated models post variable selection to Q^2 of permutated models sans variable selection. Another way to estimate the overinflation is to estimate the difference between Q^2 of unpermutated models sans- and post variable selection. The two estimates of overinflation overlapped each other at sample sizes below 20 in this dataset (Figure 9). An adjusted Q^2 was established 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 is a more valid measure of how well a model performs than

Q^2 post variable selection. If the adjusted Q^2 is >0 there is true variation larger than random explained by the model.

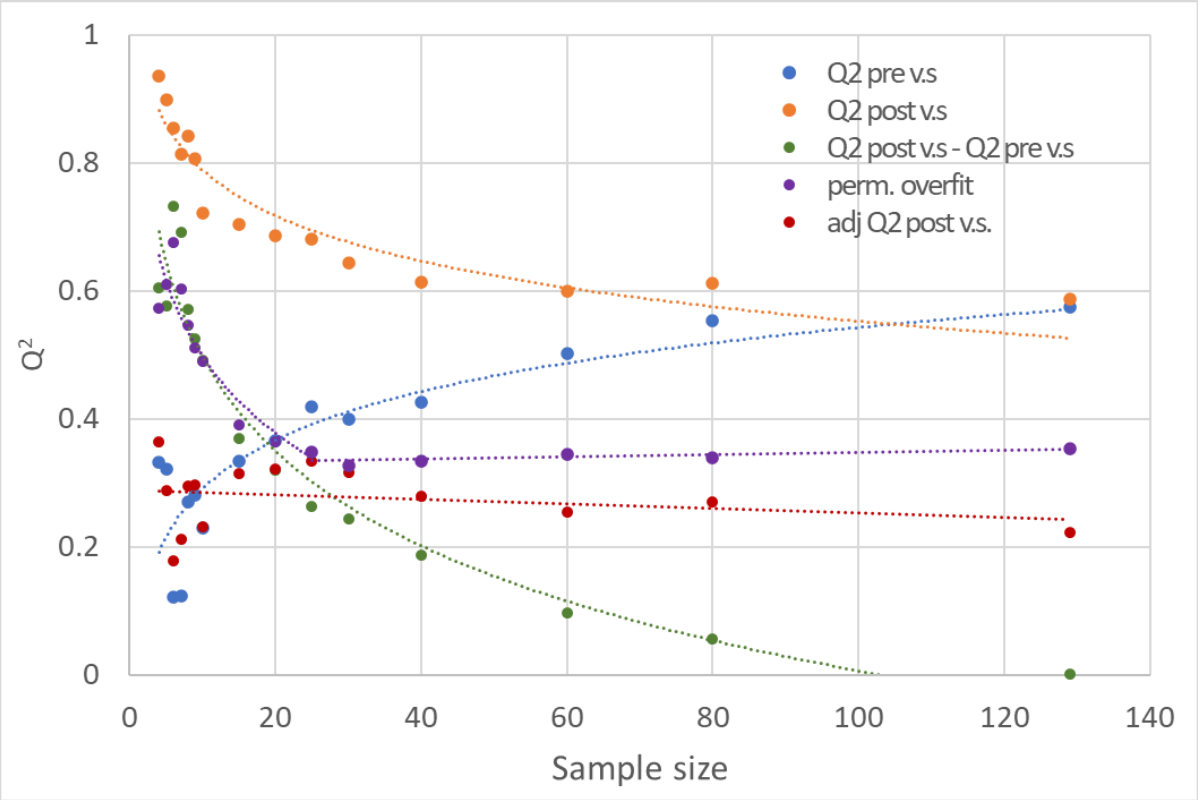

Figure 9. *The figure displays Q^2 versus group size of models comparing the groups of E+P versus Nonuser, age group 61-65. Q^2 sans v.s: Q^2 of models sans variable selection; Q^2 post v.s.: Q^2 of models post variable selection; Q^2 post v.s. – Q^2 sans v.s.: The difference between Q^2 sans and post 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.*

The number of orthogonal components were decreased or not affected with decreasing sample sizes indicating that it was not the number of orthogonals that gave increasing R^2 and Q^2, (Figure 10).

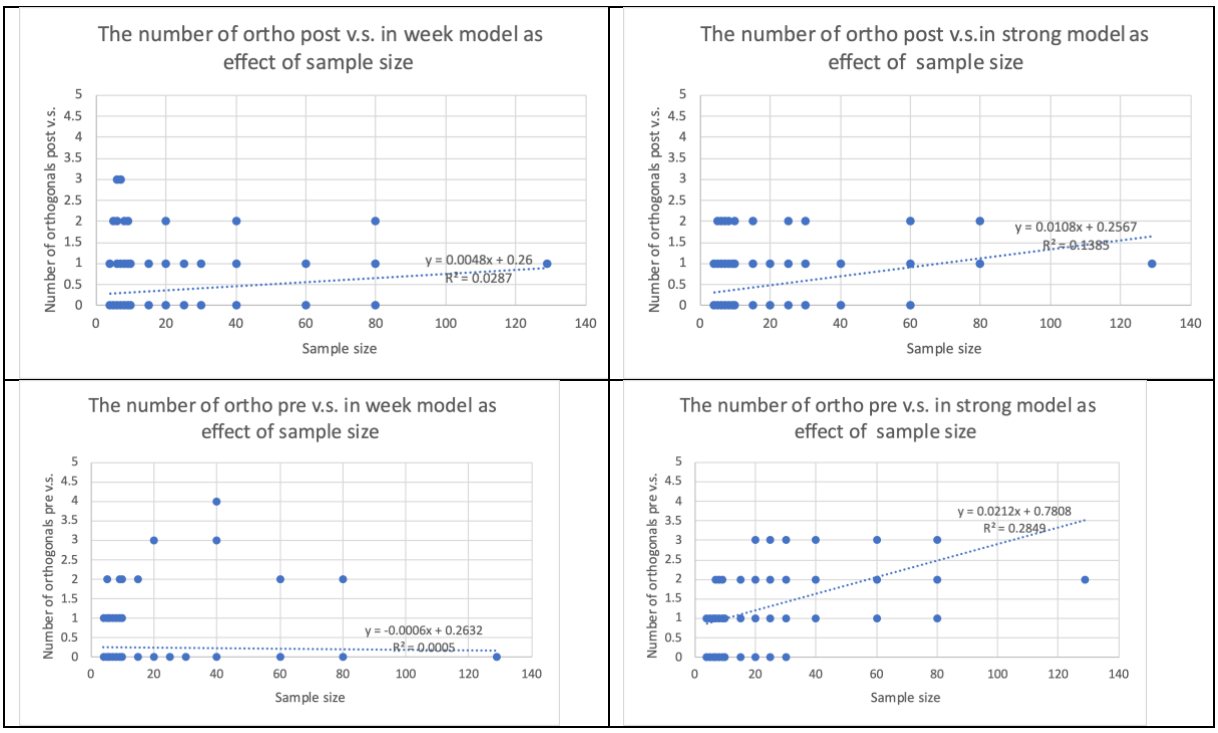

***Figure 10**. The number of orthogonals post variable selection (top panels) and sans variable section (lower panels is not increased with reduced sample size in the significant models. This holds true both for the strong (right panels) and weak (left panels) models.*

Discussion

OPLS modeling is widely used in various omics fields as a means to model both the significance of group separation, as well as for identification of variables driving the group separation. Although multivariate modeling provides many advantages over univariate statistics, e.g., avoiding high false discovery rates due to multiple hypothesis testing, the risk of overfitting is prominent particularly in small cohorts. The relevant threshold for significance in the utilized model statistics are highly dependent on the data set at hand, and the lack of clear guidelines adds an extra layer of complexity and uncertainty, especially for the inexperienced user. Cross validation (often termed Q^2) as a means of validation has become the gold standard when the cohort is too small to allow for division into a training- and test sets. Even so, the relevant significance threshold for the resulting model statistic is not clear cut, and Q^2 cutoffs ranging from 0.3 – 0.5 have been proposed [15, 16]. To test the statistical significance of the group separation, CV-ANOVA [17] can be used as a complementary metric. Particularly when Q^2 cutoffs below 0.5 are applied, permutation tests to assess the risk of erroneously rejecting the null hypothesis when the model in reality is produced by chance are highly recommended [18].

Here we investigate how OPLS modelling performs on small sample sizes using a publicly available dataset, and how various types of permutation testing can be used to assess the associated risk of overfitting or overinflation of model statistics. The results show that R^2 and Q^2 have a tendency to increase with decreasing samples sizes. We demonstrate that both permutations sans- and over variable selection to a large extent protects from identifying

random models as significant, while permutations post variable selection has a limited power in this aspect, particularly for small samples sizes. We also estimate the magnitude of overestimation of Q^2 resulting from variable selection by comparing the permutated models sans and over variable selection (Figure 9). Subtracting the calculated overestimation from Q^2 post variable selection result in a Q^2 that is adjusted for overinflation. The resulting adjusted Q^2 is more robust against differences in group sizes. As such, the adjusted Q^2 represents a threshold of significance for Q^2, and can be calculated as a threshold of model relevance for each specific dataset.

We also present an R package and workflow roplspvs (**R o**rthogonal **p**rojections of **l**atent **s**tructures with **p**ermutation over **v**ariable **s**election) that uses the Bioconductor package ropls for OPLS modeling and adds variable selection as well as permutations sans, post and over variable selection as well as provides an estimate of the overinflation and adjusted Q^2. An example of using the package on a publicly available metabolomic dataset to create OPLS models is demonstrated.

Permutations post variable selection resulted in all models regardless of sample size being significant (α=0.05) also when comparing the same group of subjects to each other. The model has thus found alterations in variables that are not larger than the average variation in the dataset. This shows that permutation post variable selection is not a sufficient tool to distinguish between significant and insignificant models.

Permutations of nonusers and fitting the permutated subjects to all variables sans variable selection and comparing permutated to unpermutated models, there is as expected no more than 5% significant models at α=0.05 identified in nonusers versus nonusers. In the strong models all models were significant when the number of subjects exceeded 15. This indicates that this is a good tool to evaluate if a model is significantly better than a random model. As permutations sans variable selection does not include any optimization of the model it results in a p-value that is not affected by the model under investigation being more optimized than the permutated models.

In addition to permutation sans variable selection, it is useful to perform variable selection on every permutated dataset with proceeding comparison of permutated and unpermutated models. This ensures that the model is significant also after variable selection and that the selected variables represent the main variation in the dataset and that this variation is larger than the variation present in the dataset by random.

Comparing nonusers to nonusers, permutations over variable selection result in 12 % significant models. One reason for the number of significant models being larger than the expected 5% may be that the model has optimized p(corr) cutoff while the permutated models use the same p(corr) cutoff as the unpermutated model. To account for the difference in optimization it is useful to take also the regression over permutations into account. The regression compares the Q2 to the correlation between the permutated group setting and the original group setting. A positive regression coefficient over the permutations indicates that the model is significant with higher Q2 in models with high correlation to unpermutated model under investigation. Taking also the regression into account when comparing nonusers to nonusers resulted in 6% significant models (p[Q2 perm over v.s.]<0.05 and regression coefficient>0.1 and p[regression over permutations]<0.05). This shows that it may be

important to also consider the regression over permutations in combination with permutation over variable selection when evaluating a model.

The utility and functions of the script is exemplified in a metabolomics dataset. The dataset contains only females, and they are stratified by age minimizing the variation in the dataset. One disadvantage in this public dataset is that the smoking status is not available. This is known to contribute to a large variation in the data and may, if the smokers are evenly distributed between the groups, result in higher variance with lower significance of the models than if analysis is stratified by current smoking status. Also, if the smokers are unevenly distributed between the groups the models may become either stronger or weaker than expected.

The script automates the comparison of pairwise comparisons of cohorts consisting of multiple sub-groups, with the possibility to stratify the analysis by various descriptors of interest. The automated workflow includes filtering, missing data imputation, and the option of automated replacement 0 or NA values with the lower limit of detection, PCAs, OPLS with scores and loading plots, significance test using permutations sans, post and over variable selection and ROC curves, as well as the possibility to compare the resulting models using shared and unique structure plots (SUS-plots). Automation not only increases speed of analysis compared to graphical user interfaces (GUI), but also reduces the subjectivity in the analysis, and facilitates efficient rerun with new data or new settings. The script is user friendly with parameter settings separated into parameters that are necessary to set, and parameter settings that may be of interest for the experienced user. The script is flexible with the possibility to set limit the comparisons and modelling strategies to run, as well as to include numerical and categorical data with possibility to implement a bash script for running parallel computing on clusters of high-performance computing systems for faster computation times. This ensures the utility for performing OPLS modelling of all different platforms of omics data analysis and clinical data.

Modelling strategy 2 includes all significant variables (i.e., above a p(corr) corresponding to a significant Pearson's correlation), is useful for pathway enrichment analysis where the goal is to identify all significantly altered pathways. The most predictive model is identified by Modelling strategy 3, and may be useful when the goal is simply to find groups that are significantly different. The fourth Modelling strategy extracts the minimum number of variables that produce a highly predictive model, useful for identifying putative biomarkers candidates. Finally, Modelling strategy 5 is produced with a conventional iterative procedure to enable comparison to a conventional optimization workflow.

The usefulness of permutations for OPLS models was demonstrated by Lindgren et al. already in 1996, where they showed that permutations over variable selection as well as permutations prior to variable selection could distinguish a model from a model created by random [10]. They also suggested to estimate the overinflation of a model by comparing Q^2 of the model fitted on permutated data sans variable selection compared to Q^2 of model fitted after performing variable selection on the permutated data. Here we investigate the usefulness of permutation procedures when small sample sizes are used. We show that permutations before and over variable selection are efficient tools to establish the significance level from considering a random model significant also using smaller sample sizes.

Other methods that have been suggested to test overfitting or overinflation is the use of cross-validated score plots [19]. It has the advantage of producing a more predictive picture of the score plot but does not provide a threshold for significance. This could be a future improvement of roplspvs as the non cross-validated score plots in roplspvs may be overoptimistic.

ROC analysis is implemented in roplspvs using the precrec and pROC packages. It has been suggested to implement AUROC in significance testing by permutations to get a threshold for significance also for AUROC [20]. This has not been implemented in roplspvs but may be a future improvement of the package.

The roplspvs package is designed to easily model a large number of group comparisons, as well as assessing the significance using stringent permutation testing. This is helpful if many groups are screened for differences between them. The roplspvs package is suited for studying the performance of ropls models as it is flexible and automated. This package is especially helpful for small sample sizes, since those are more prone to overinflation of model statistics. The thorough test for significance contributes to reducing the large rate of false findings in scientific research by increasing the robustness of OPLS models.

Conclusions

Roplspvs is a package that performs OPLS-DA modeling with semi-automated model optimization using variable selection, as well as permutation testing sans-, post-, and over variable selection aimed at establishing a significance level for the predictive model statistics Q^2 of models post variable selection.

The extensive quality control by permutation prior to variable selection together with the permutations over variable selection reduces the risk of false positives. Permutation prior to variable selection assures that the groups have significant difference, whereas permutations over variable selection assures that the models remain significant after variable selection. The utility of the package is demonstrated on a publicly available metabolomic dataset. Investigations of model effects on small group sizes showed that R^2 and Q^2 increases with decreasing samples sizes. We demonstrate that both permutations sans variable selection and permutations over variable selection to a large extent protects from identifying random models as significant, while permutations post variable selection does not. The proposed adjusted Q^2 resulting from subtracting the overinflation from the model post variable selection is stable over the sample sizes, and may be used as a threshold for the significance of OPLS models. This is well needed as it is difficult to evaluate if a model is significantly better than a random model simply by using traditional statistical evaluation of 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.