

# Data analysis for metabolomics

EMBL-EBI Introduction to metabolomics analysis course

Gonçalo dos Santos Correia, PhD 30 March 2023

March of Dimes Prematurity Research Center

Imperial College London, IRDB, Department of Metabolism, Digestion, and Reproduction

1. [Introduction](#_bookmark0)

Table of contents

1. [Data pre-treatment](#_bookmark1)
2. [Exploratory data analysis](#_bookmark2)
3. [Statistical hypothesis testing and regression models](#_bookmark3)
4. [Multivariate data analysis](#_bookmark4)
5. [Conclusion](#_bookmark5)

1

# [Introduction](#_bookmark0)

## Data is peak-picked,

Metabolomic data analysis

batch-corrected, and filtered... What next?

2

What is the main purpose of the analysis?

Choosing the right method

* Visualisation and exploratory data analysis
* Differential analysis, biological interpretation
* Prediction, biomarker studies
* Stratification and identification of molecular phenotypes.
* Support metabolite identification
* ...

3

Mass spectrometry data has specific characteristics that require particular at- tention when choosing statistical modeling strategy:

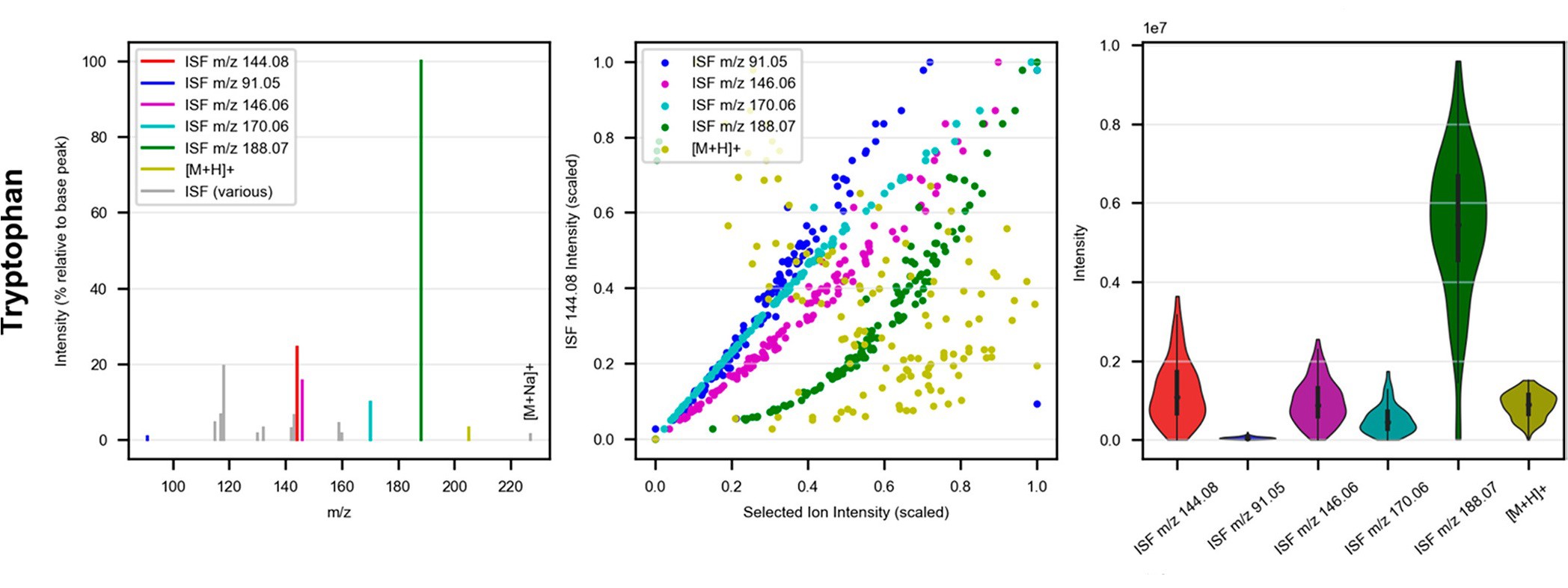
Characteristics of mass spectrometry data

* Large number of variables, multicollinearity and redundancy
* Complex noise structure with heteroskedasticity (unequal noise variance), etc...
* Type of assay (hyphenated, direct infusion, or imaging)
* Ionisation: 1 compound many MS features. Adducts, isotopes, dimmers, in source fragments, etc...

→

4

Characteristics of mass spectrometry data



Adapted from Sands et al[1](#_bookmark7)

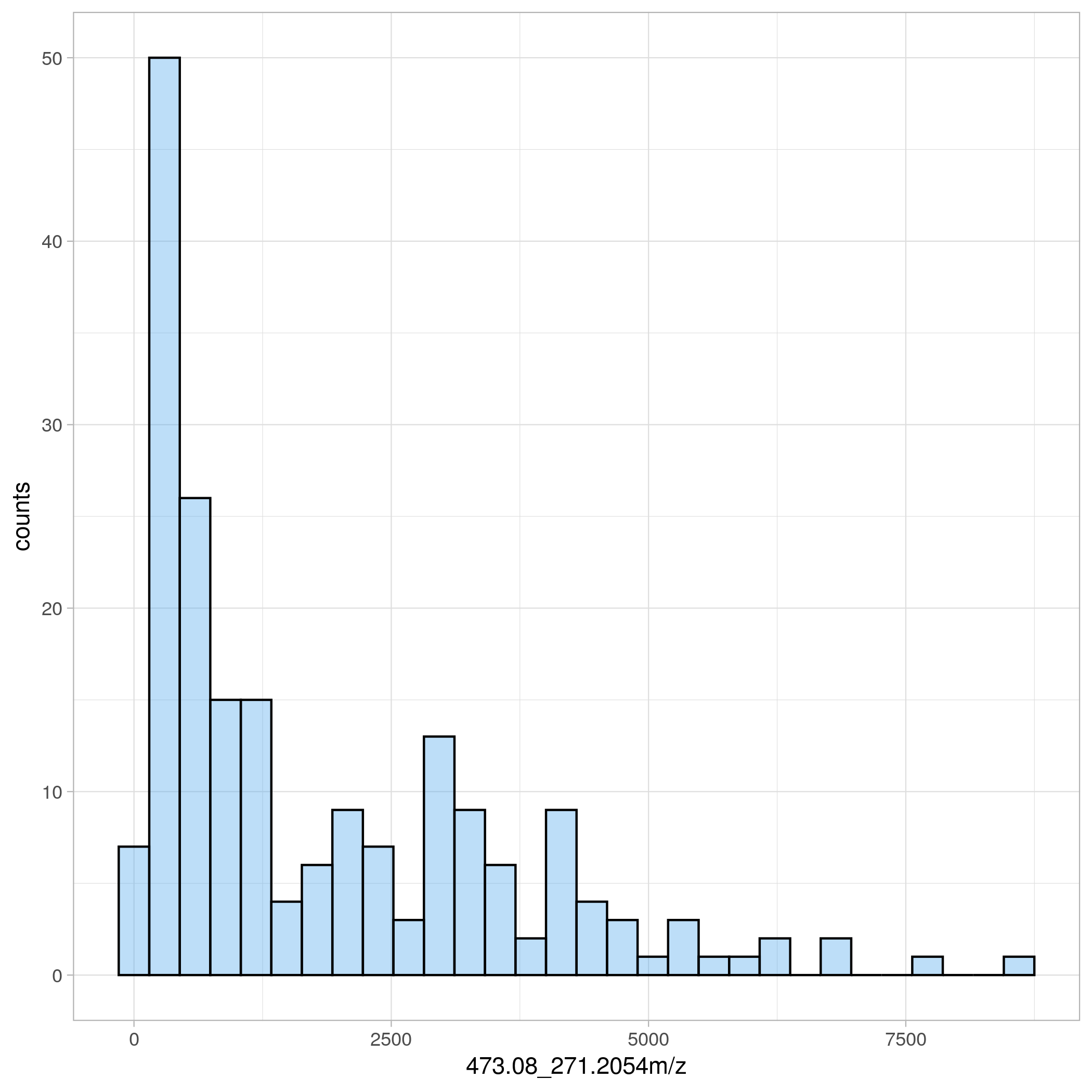
5

* Biological correlations from biochemical pathways and physiology

Characteristics of mass spectrometry data

* Spatial correlation in MS imaging
* ”Advantages” of metabolomics: everything can be a confounder!
* Data pre-processing workflow:
  + Spectra: Each individual peak is represented over multiple data point.
  + Peak-picked: Intensity values indexed by *m/z* and retention time
  + Grouping algorithms - features linked in groups (e.g. de-isotoping, pseudo-spectra)
  + Unknown (rt, m/z) or annotated (compound) features

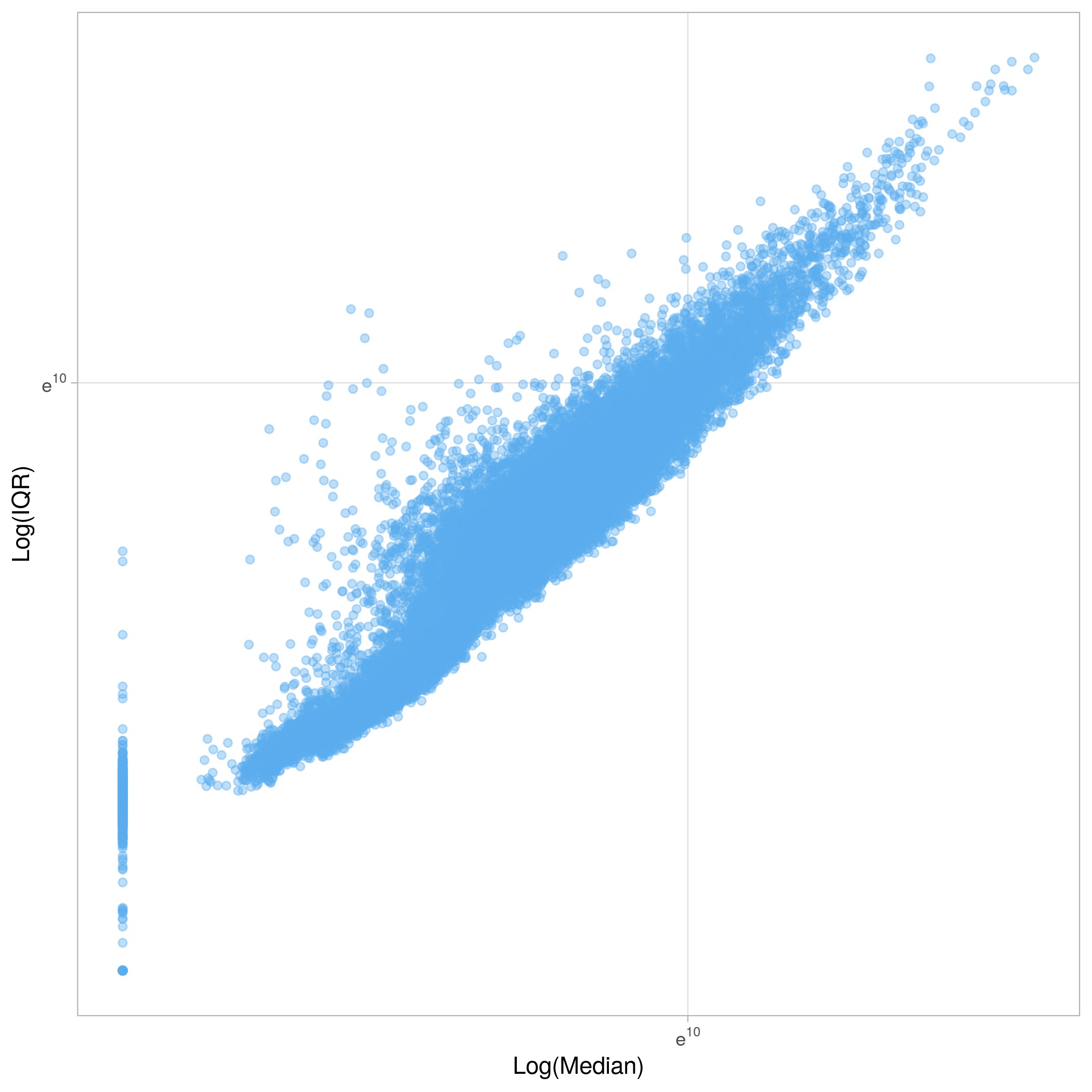
6



Characteristics of mass spectrometry data

Non-normally distributed, positive values only [0*,* ∞)

7



Characteristics of mass spectrometry data

Heteroskedasticity

8

Which methods are applicable metabolomics? No restrictions (usual caveats apply), but some popular choices are:

Data analysis methods for metabolomics

* Visualisation and exploratory data analysis - Descriptive statistics and plots, PCA, PLS, correlation
* Differential analysis - Hypothesis testing, regression models
* Prediction - Multivariate (PLS/PLS-DA) and machine learning models
* Stratification - Clustering/bi-clustering algorithms
* Metabolite identification - Correlation and clustering algorithms

9

## X matrix: The data matrix containing the metabolic profiles

Notation and conventions

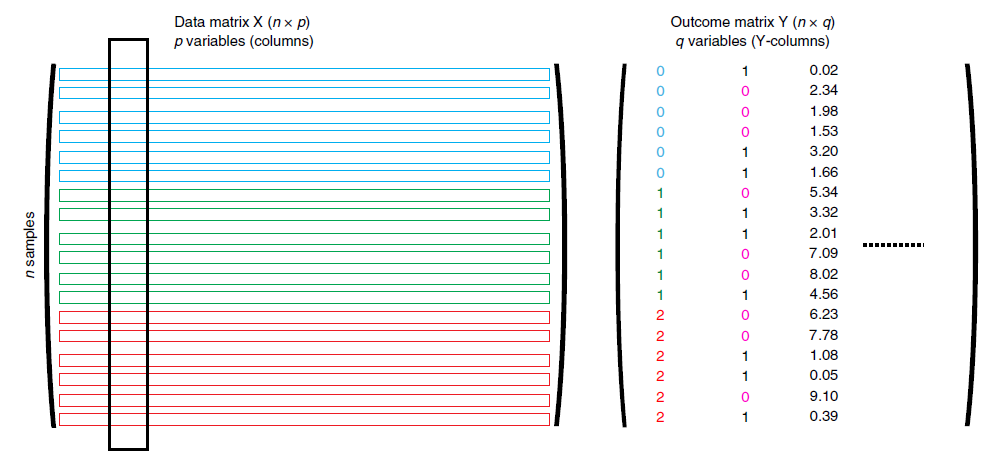
* + Observations (*n*) in rows (metabolic profiles)
  + Features (*p*) in columns (digitised point/peak/MS feature/compound abundance)

## Y matrix/Y vector: The data matrix containing the metabolic profiles

* + Observations (*n*) in rows
  + Study variables (experimental groups, demographic variables) (*q*) in columns

10

Notation and conventions, continued



## Adapted from Blaise et al[2](#_bookmark8)

11

Tutorial examples are available via GitHub: <https://github.com/Gscorreia89/metabolomics-course-ebi>

Tutorials code and examples

Click on the binder sticker on the repository readme page to run via the web browser 

In these examples, we will use data from human urine biofluid samples which were profiled by a reversed phase positive ionisation LC-MS assay[3](#_bookmark9).

The raw data is available in Metabolights [MTBLS719](https://www.ebi.ac.uk/metabolights/MTBLS719).

12

These urine samples originate from the AddNeuroMed and ART/DCR consortia studies on neurocognitive decline and Alzheimer’s disease. For more infor- mation see Lovestone et al[4](#_bookmark10) and the ANMERGE repository[5](#_bookmark11).

Example dataset

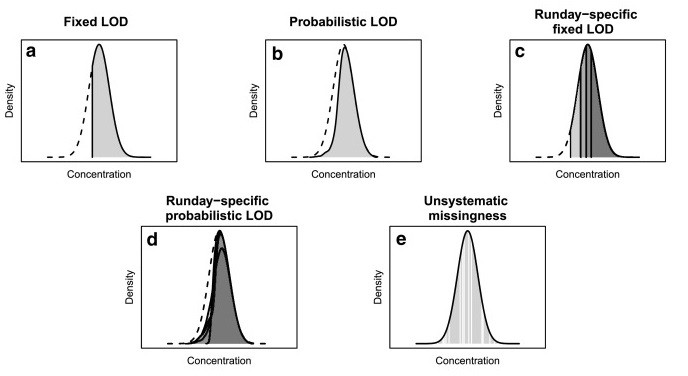
* *n* = 200 baseline sport urine samples (1st sample after recruitment to study)
* *p* = 27,870 LC-MS features, generated with xcms[6](#_bookmark12)
* 2 study variables, Age and Sex.

13

# [Data pre-treatment](#_bookmark1)

Missing observations (NA values) are common in MS, caused by lower or upper signal truncation (LOD, saturation) or by failures during peak detection.

Missing data



From Do et al[7](#_bookmark13)

14

Many statistical analysis methods cannot handle missing data and these must be imputed (replaced with a numerical value).

Missing data imputation - Types of missing data

Common strategies to impute missing data:

* Replace with 0
* Replace with noise integration (e.g. xcms fillPeaks).
* Replace with LOD/smallest non-0 value.
* Multivariate imputation methods (e.g. MICE, kNN, Random forests).

For a discussion on missing patterns in MS datasets a comparison of multiple imputation methods, see Hrydziuszko et al[8](#_bookmark14) and Do et al[7](#_bookmark13).

Warning

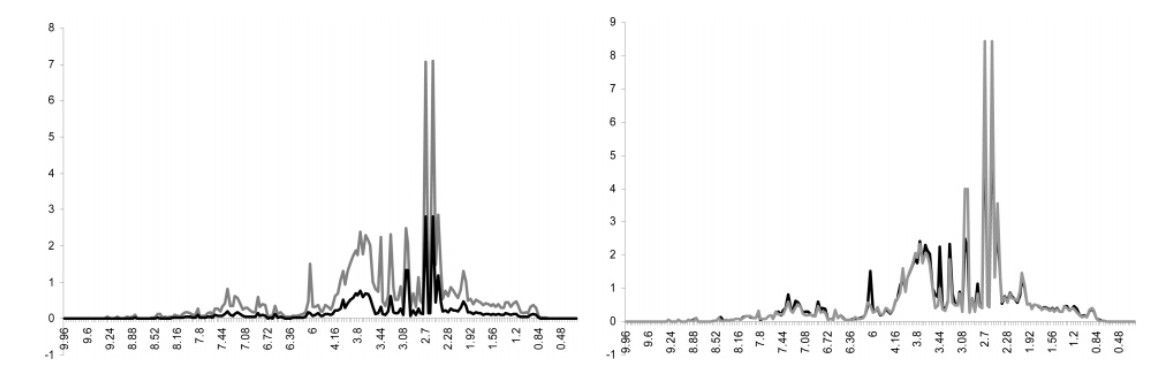
In untargeted profiling experiments it can be hard to detect ”what” is a

missing value (e.g. xcms fillPeaks) steps, and determine LODs

15

Used to adjust for changes in global profile due to varying sample amounts, dilution effects, or instrument intensity drift (e.g. varying tissue weight, urinary dilution, signal decay over run-order).

Normalisation



From Dieterle et al[9](#_bookmark15)

In normalisation, each row of *X* is divided by its corresponding normalisation

coefficient *Nc*:

*Xnormalised* =  *X*

*N*

*c*

16

Normalisation coefficients can be measured or estimated computationally us- ing various methods:

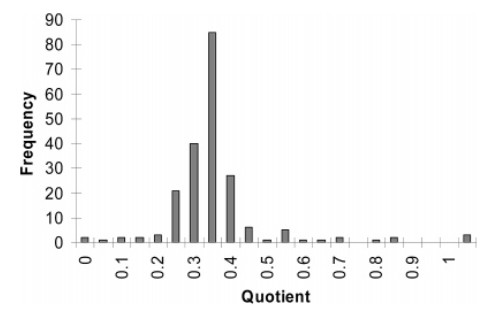
Normalisation methods

* Weight normalisation
* Creatinine, specific gravity, or osmolality (urine)
* Total area normalisation (sum of all signal/features)
* Probabilistic quotient normalisation (PQN)[9](#_bookmark15)
* Run order drift correction methods (e.g. LOWESS QC correction)[10](#_bookmark16) For more suggestions, see Kohl et al[11](#_bookmark17).

17

Introduced by Dieterle et al[9](#_bookmark15), PQN is robust to interference from high intensity signals (e.g. large contaminant or xenobiotic signals).

Probabilistic Quotient Normalisation (PQN)

How it works:

* Select a reference such as the median spectra
* Calculate a fold-change data

matrix *Fx* =

*X Reference*

* The PQN coefficient is the row-wise median

(across-variables)

*PQNc* = *median*(*Fx*) Democracy!

Warning

Normalisation procedures have key effects on downstream analyses[12](#_bookmark18), and

sometimes its worth checking normalisation coefficients for bias[13](#_bookmark19).

18

Scaling (column-wise operation) is used to account for in overall magnitude differences between variables. Differences in scale impact multivariate model performance and variable importance.

Scaling

*Xscaled* = *X* − *µx*

*s*

*σ*

Common scaling methods in metabolomic analysis:

* Mean centring (*s* = 0)
* Pareto (*s* = 1*/*2, or √*s*)
* Unit-Variance (UV) scaling, standardization or z-score (*s* = 1) Recommended reading: Berg et al[14](#_bookmark20)

19

Transform each column by applying a non-linear mathematical function. Transformations modify the relative distance between data-points and the data distribution.

Transformations

*Xt* = *f*(*X, c*)

Some examples of transformations:

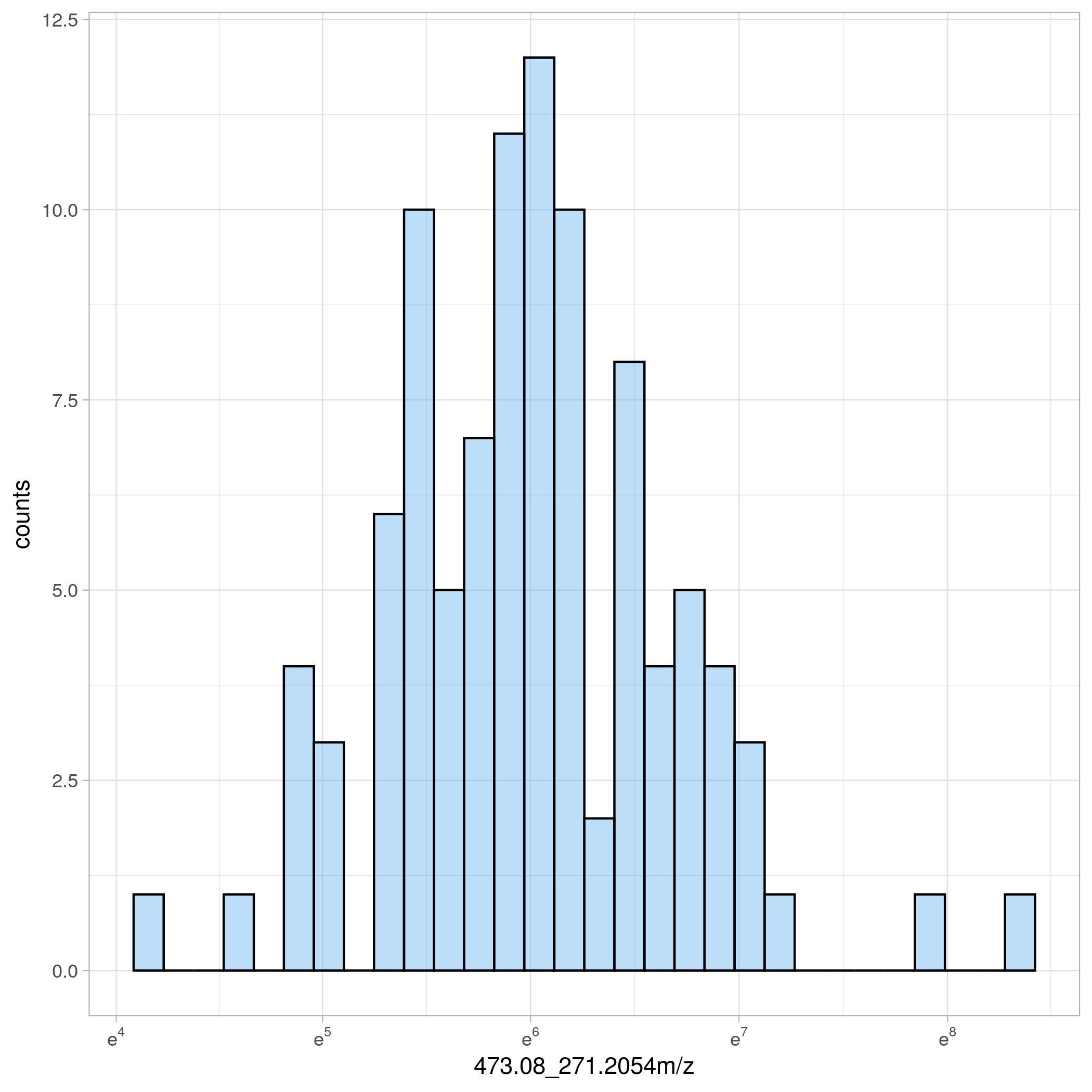
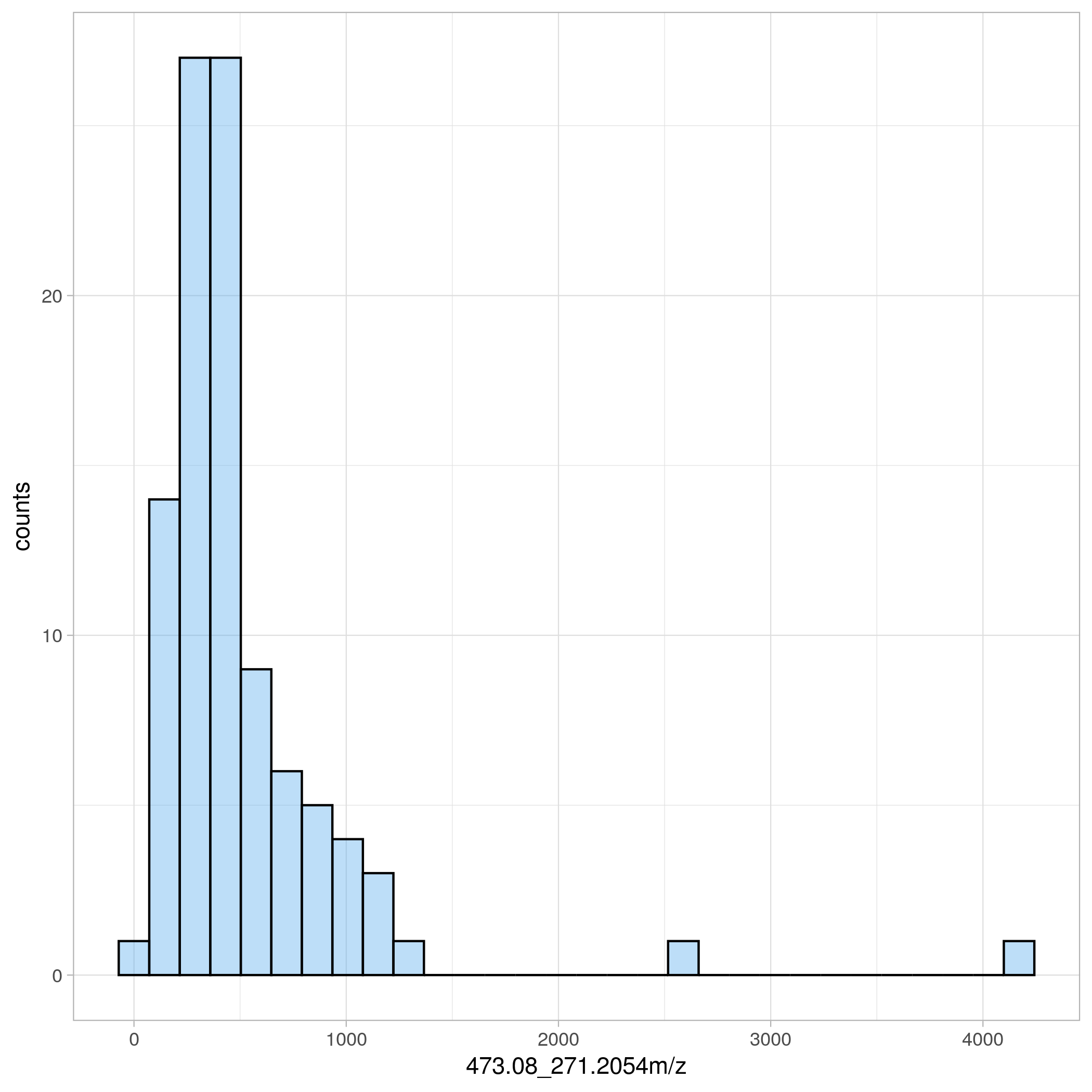
* Log transformation
* Square-root transformation
* Box-Cox transforms
* Variance stabilising transformation[15](#_bookmark21),[16](#_bookmark22) Recommended reading: Berg et al[14](#_bookmark20)

20

*Xt* = *log*(*x* + *c*)

Log transformation

## A small offset (*c* = 1) is used to handle the 0 values.



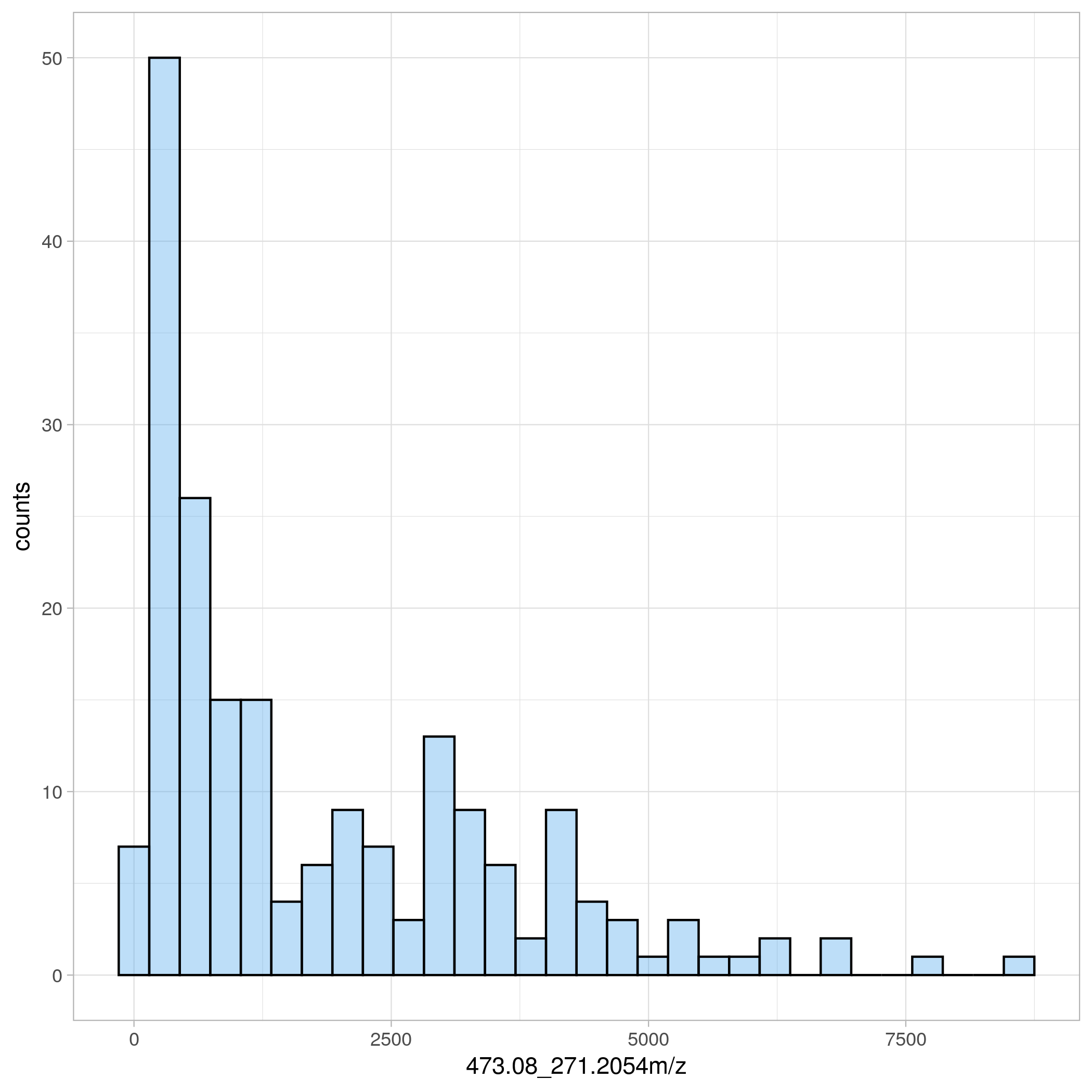
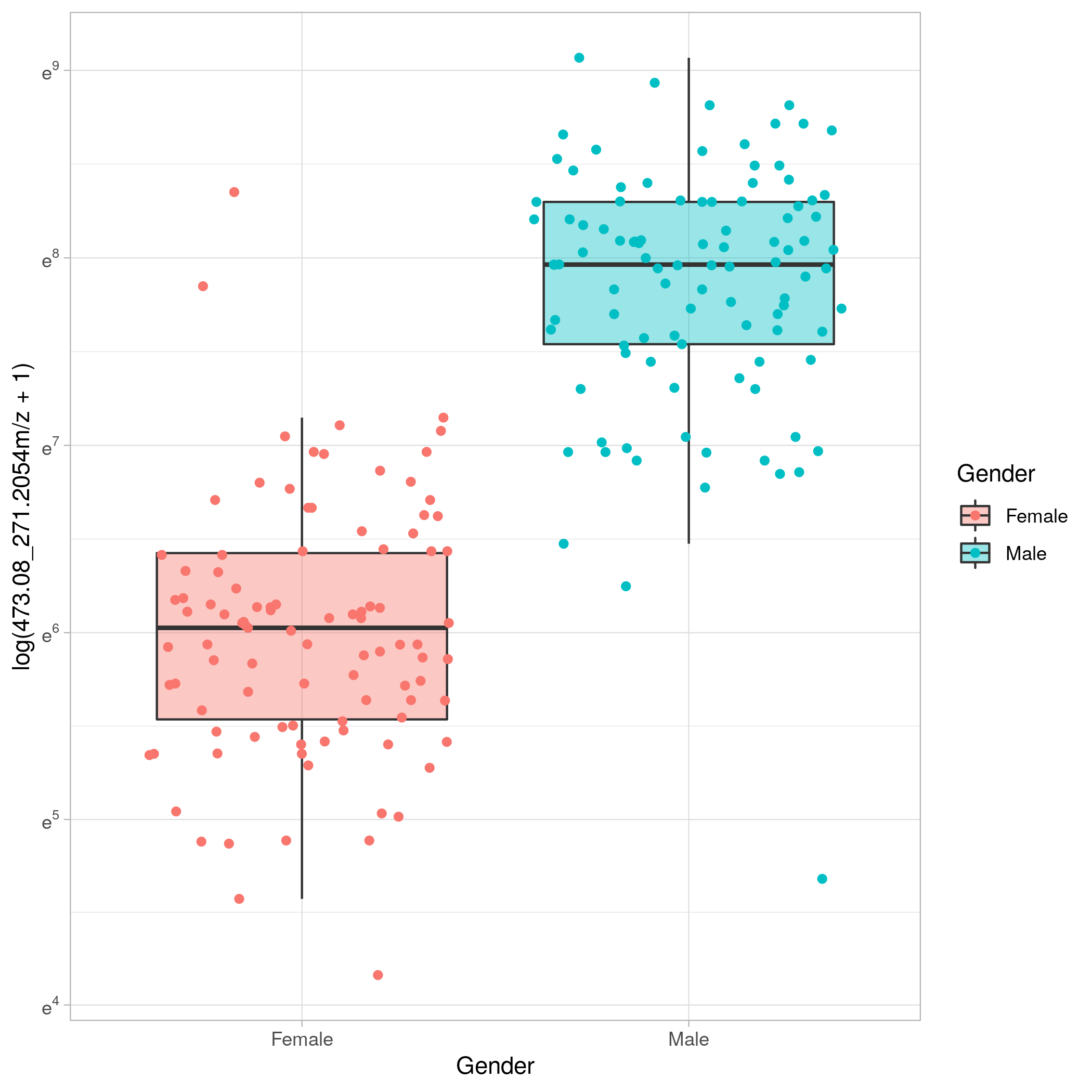
Before After log-transform

21

# [Exploratory data analysis](#_bookmark2)

Plotting is key...

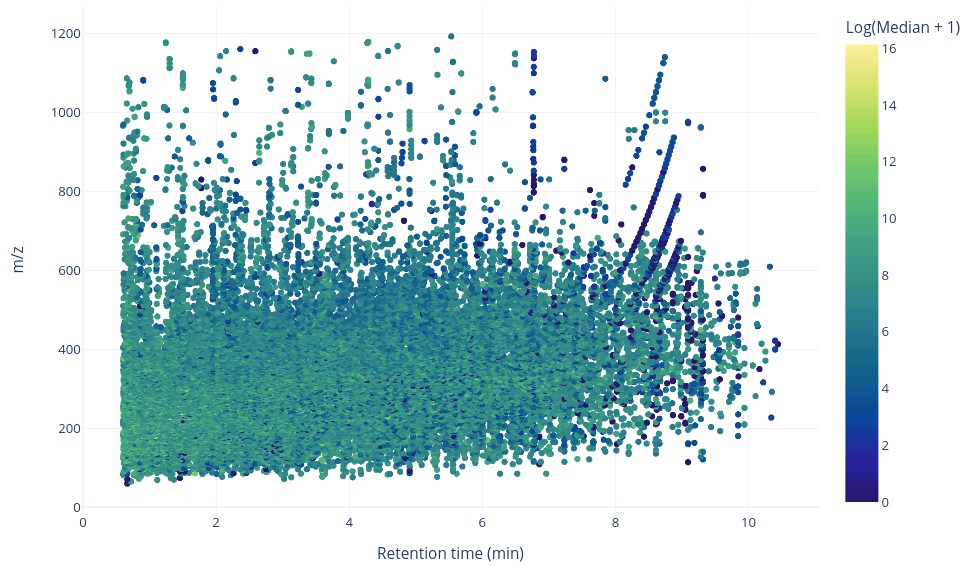
Data visualization



22

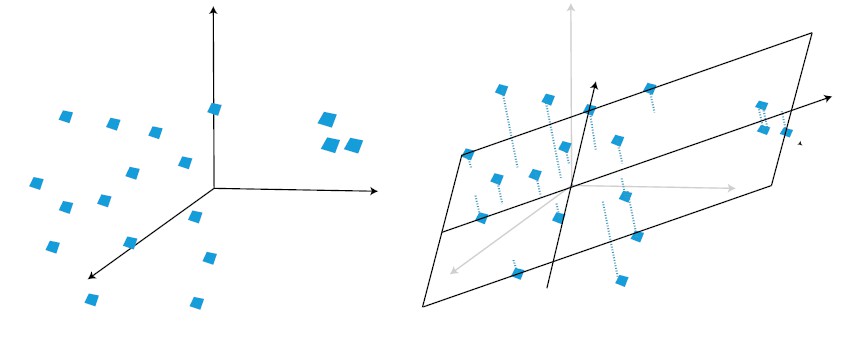
## Too many variables in *-omics*...

Data visualization



23

Principal component analysis



Principal components maximise variance explained (covariance between scores T and X variables).

*X* = *t*1*p′* + *t*2*p′* + *...* + *trp′* + *E*

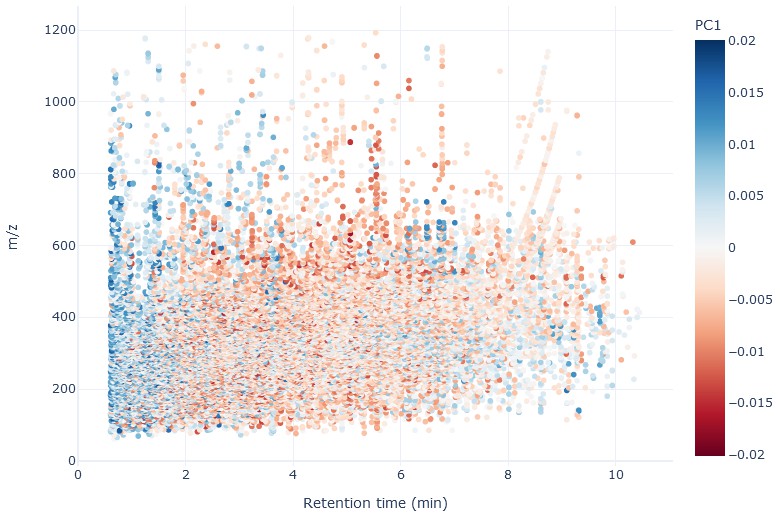
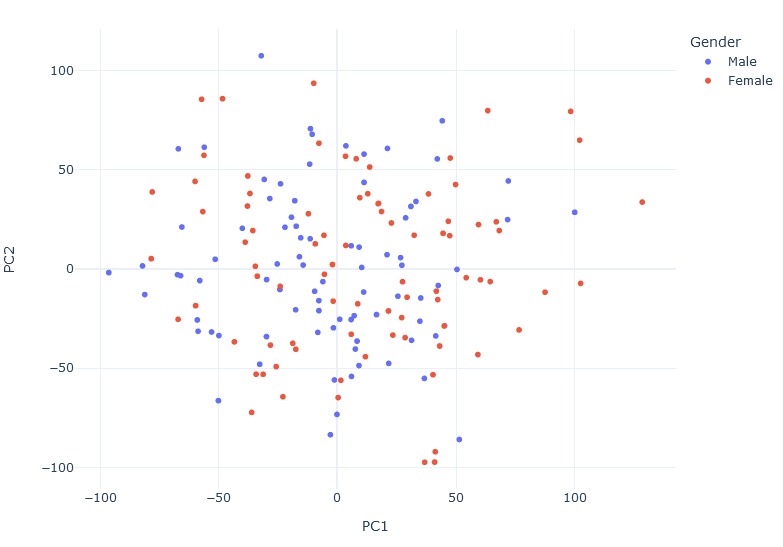
1 2 *r*

X = original matrix (n x p) T = score matrix (n x r) P = loading matrix (p x r)

24

How to interpret a PCA model?

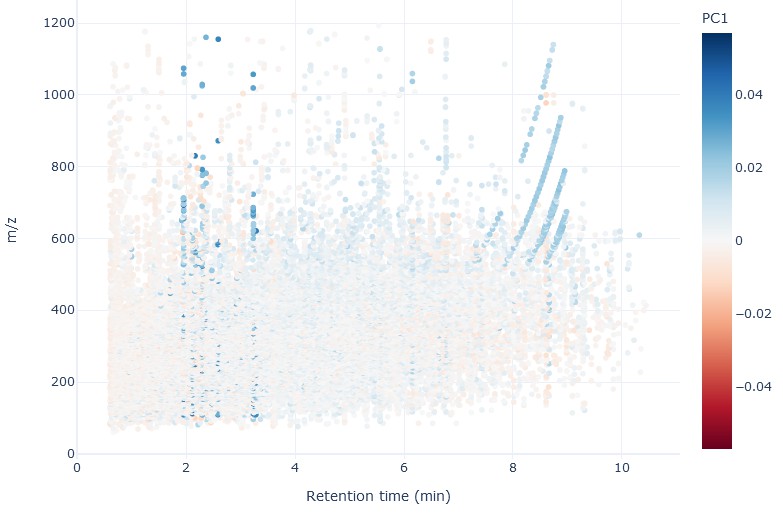
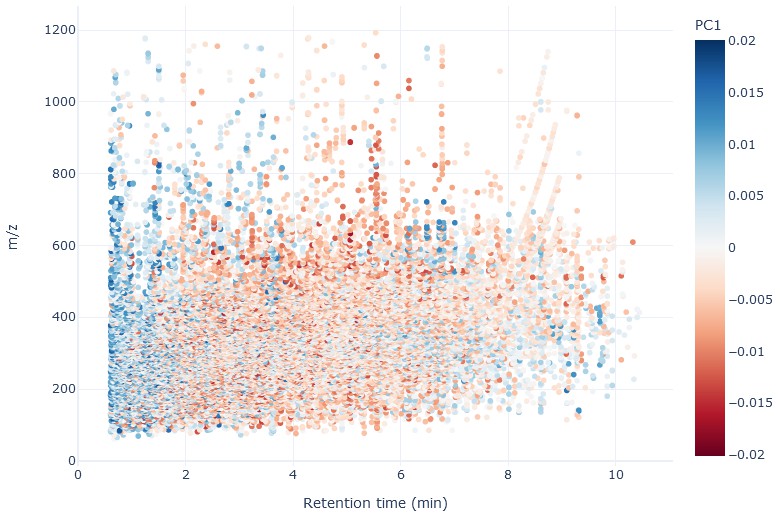
PCA - scores and loadings



Scores plot PC1 Loadings

25

Impact of scaling on multivariate models

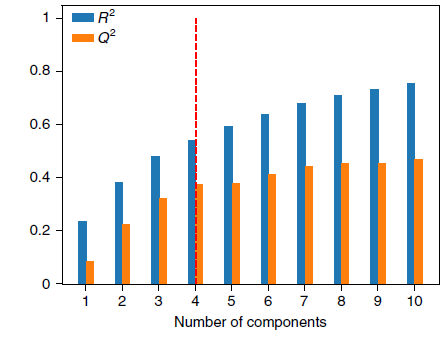
 

Mean centring (no scaling) UV scaling

26

How many principal components to use? And when does it matter?

Choosing the number of components

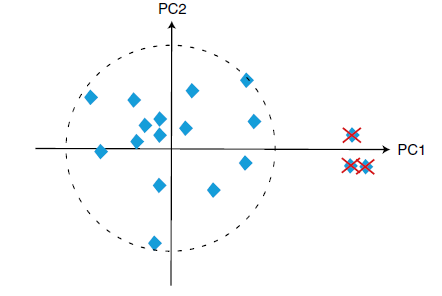


Scree plot display the variance explained per each component. Can also be combined with cross-validation (*Q*2 often reported in the literature).

27

PCA models can be used to screen for outlying observations using the Hotelling *T*2 or other metrics. The *T*2 is a multivariate generalization of the Student’s *t*- distribution.

Outlier detection with PCA



Warning

Scaling and data transformations affect outliers. The ”default” PCA method is

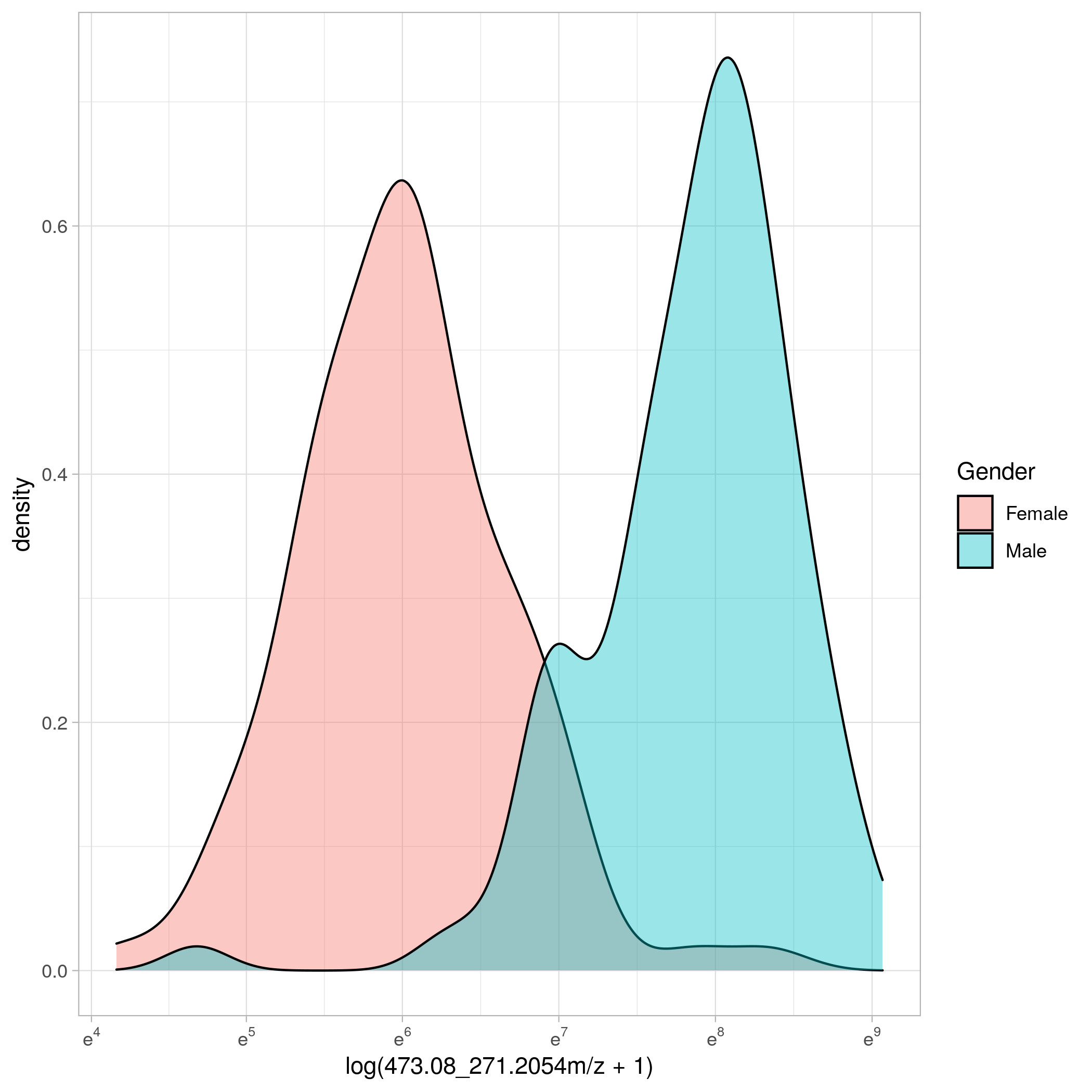
not robust to outliers (high leverage observations).

28

# [Statistical hypothesis testing and](#_bookmark3) [regression models](#_bookmark3)

Statistical hypothesis testing provides a straightforward way to identify fea- tures whose abundance is different between experimental groups.

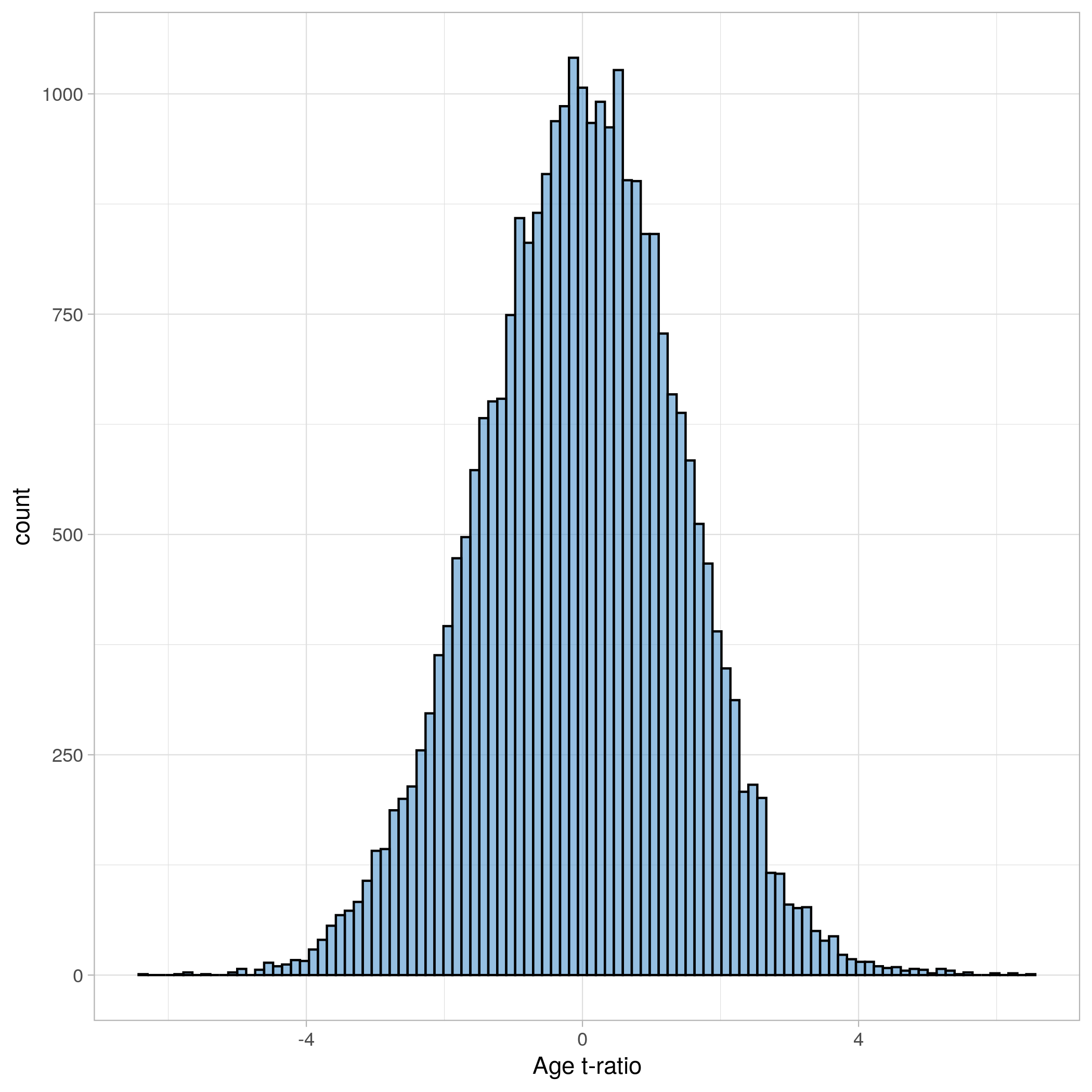
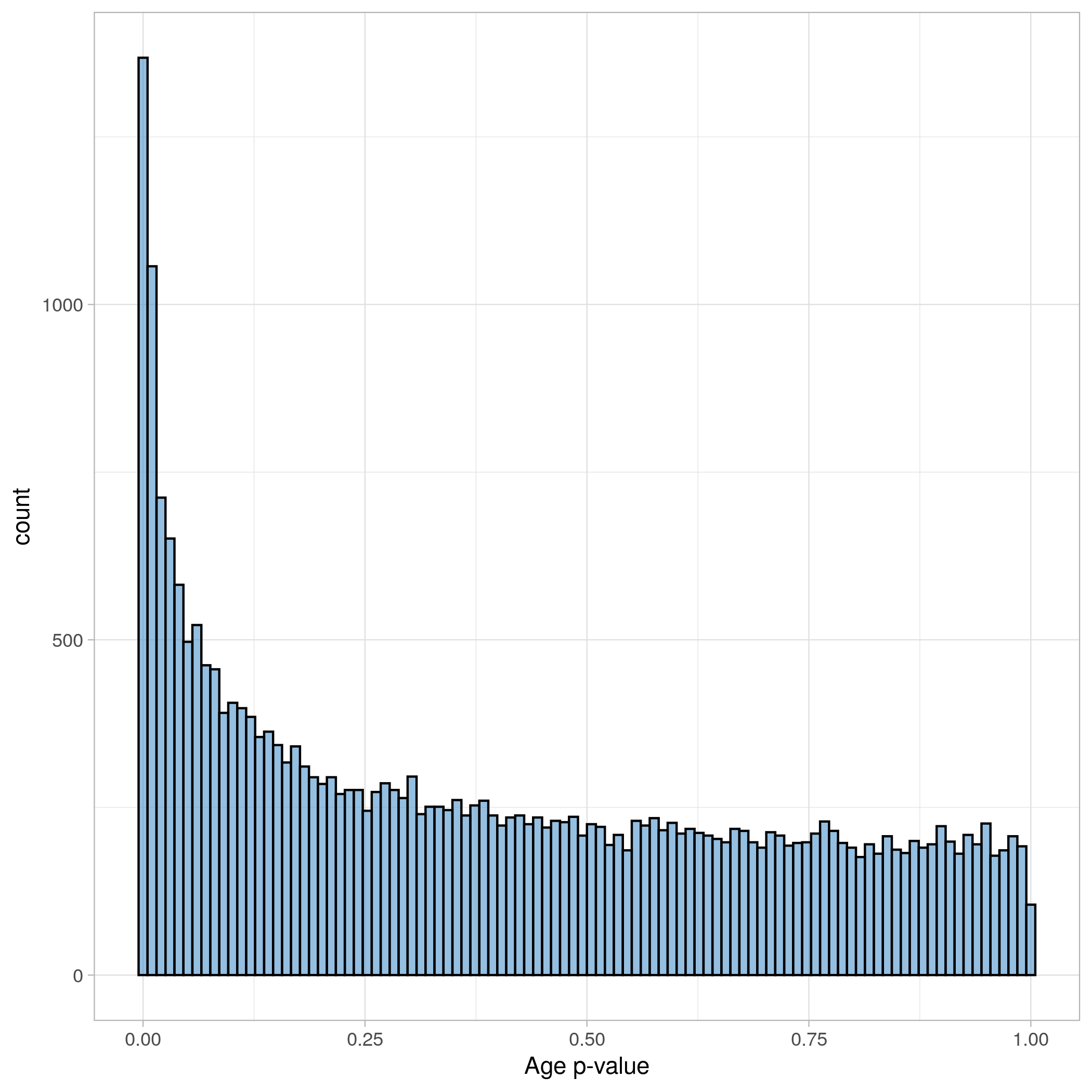
Statistical hypothesis testing



Test each variable with parametric (e.g, Student’s *t-test*, ANOVA) or non-parametric tests (e.g. Wilcoxon-Man-Whitney, Kruskal-Wallis). Usual recommendations

from statistical textbooks apply: check assumptions of tests and pretreat the data as required. 29

Univariate analysis

*t-ratio* distribution *p-value* distribution A mixture of distributions = null is true + alternative hypothesis is true.

30

Hypothesis testing outcomes

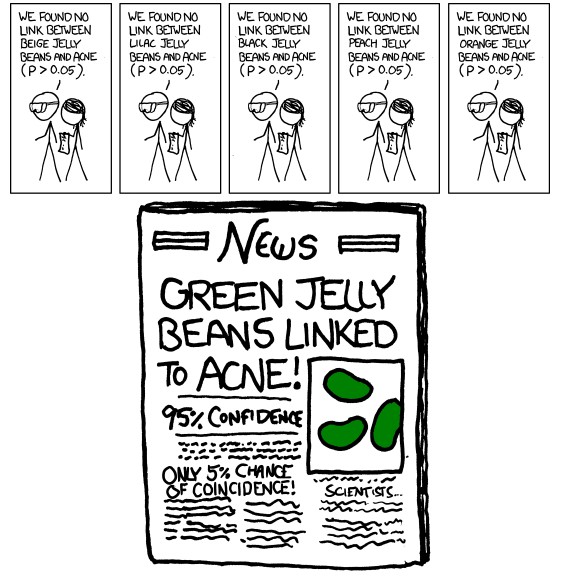
|  |  |  |  |
| --- | --- | --- | --- |
|  | | *H*0 is (Ground Truth) | |
| True | False |
| Rejection of *H*0? | Yes  No | False Positive (Type I Error) | True Positive (Type I Error) |
| True Negative | False Negative (Type II Error) |

31

When doing *n* tests, at a significance level of *α*, *n* x *α* false positives are ex- pected just by chance. The family wise error rate (FWER) is the probability of having one or more false positives in a multiple testing procedure:

Multiple testing problem

*FWER* = 1 − (1 − *α*)*n*

[Obligatory XKCD](https://xkcd.com/882/)

32

Bonferroni correction adjusts the significance threshold (*α*) to control FWER:

FWER adjustment

*αadjusted* = *α*

*n*

*tests*

*padjusted* = *p*

*n*

*tests*

Bonferroni correction assumes the tests are independent *ntests* = *p*, which is too stringent in the presence of dependency between tests. There are more suitable methods to estimate the effective number of tests: Metabolome wide significance level (MWSL)[17](#_bookmark23),[18](#_bookmark24)

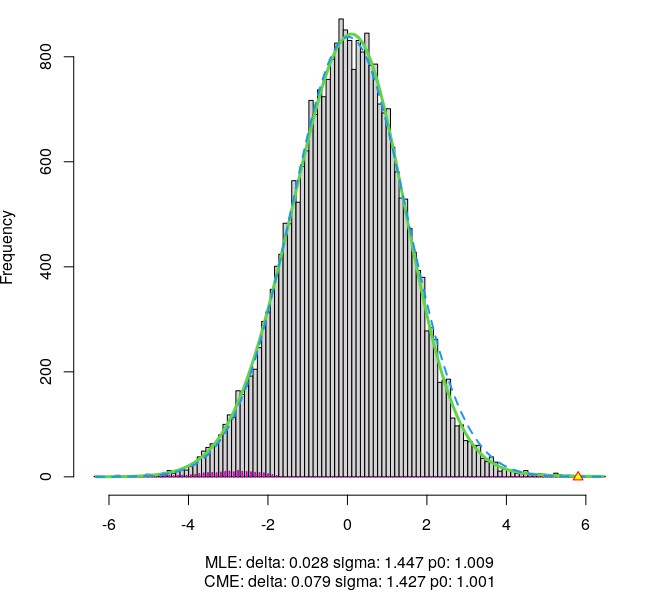
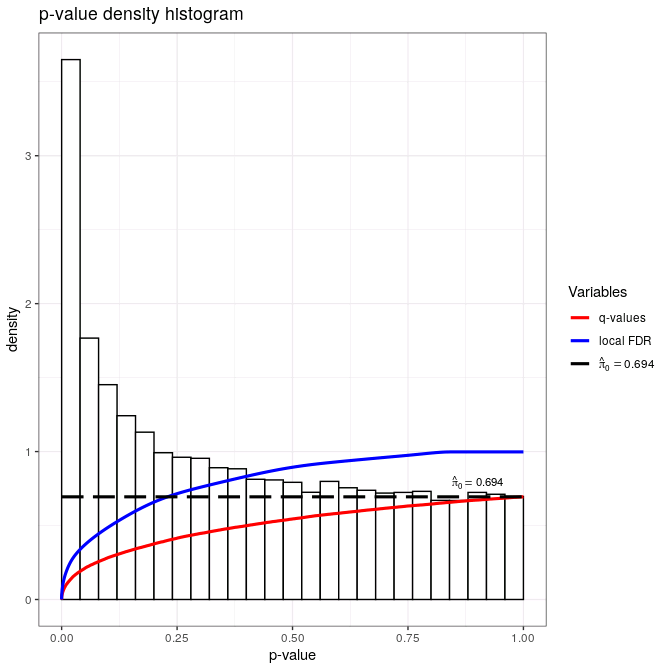
33

False discovery rate based methods control the proportion of false positives in the signature (all positive hits) instead.

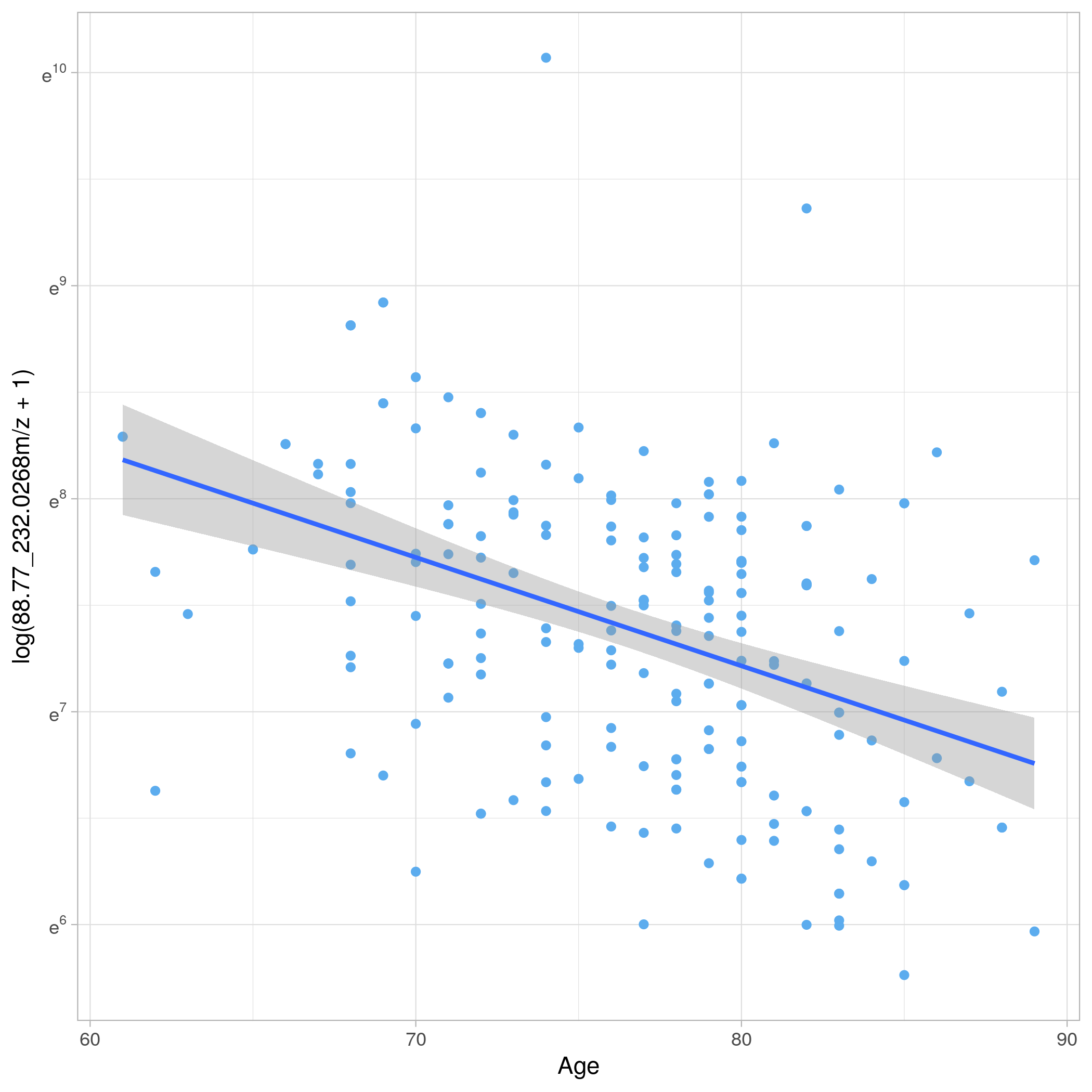
False discovery rate correction

Some exemplar methods:

* Benjamini-Hochberg and Benjamini-Yekutieli[19](#_bookmark25),[20](#_bookmark26)
* Storey’s FDR[21](#_bookmark27)
* Efron’s locfdr[22](#_bookmark28)



Storey’s FDR Efron’s locfdr 34

Linear regression models the conditional expectation of the dependent vari- able on the predictors E(Y|X). Linear regression is extremely flexible, and al- lows for explicit confounder adjustment.

Linear regression

*y* = *xβ* + *α* + *ϵ ϵ* ∼ *Normal*(0*, σ*)

[Many statistical tests can be implemented via regression!](https://lindeloev.github.io/tests-as-linear/linear_tests_cheat_sheet.pdf)

35

Step 1: Build a model structure:

Linear regression example

*log*(*MSFeature*) = *xgenotypeβgenotype* + *xageβage* + *xsexβsex* + *α* + *ϵ*

How do each of the independent variables affect the mean value of the MS feature?

Step 2: Extract the quantities of interest from each model - *β*, *t-ratio*, *p-values*. Step 3: Apply a multiple testing correction method to the vector of *p-values*

obtained to restrict the signature.

36

Used to model other probability distributions with linear regression. Logistic regression is a GLM where the the dependent variable Y follows a Binomial distribution.

Generalized Linear models (GLM)

*P*(*case*) = *xMSFeatureβMSFeature* + *xageβage* + *xsexβsex* + *α* + *ϵ*

Relates the abundance of the ”MS Feature” and other covariates with the prob- ability of case - how does a marker relate to risk of disease/outcome.

37

# [Multivariate data analysis](#_bookmark4)

With linear regression we can model a variable as a function of the entire metabolic profile (multiple), or the metabolic profile as a function of other covariates (multivariate). However, in case of multicollinearity and *p >> n* the OLS estimator is ill-conditioned and produces unstable models.

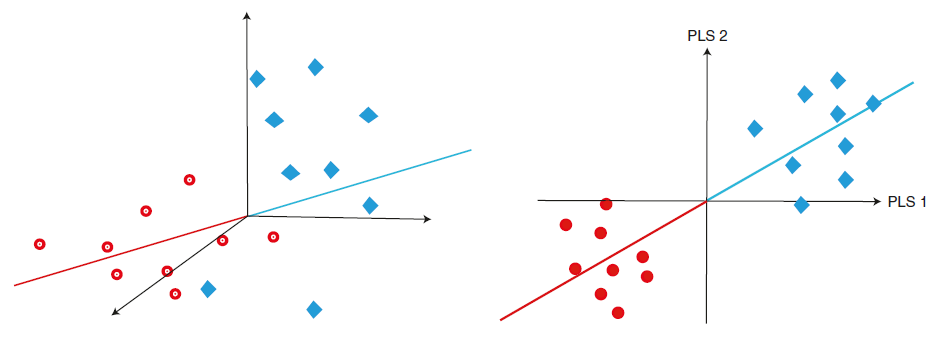
Multivariate regression

*Regularised*[23](#_bookmark29) regression methods are required in this setting:

* Ridge regression
* LASSO
* Elastic Net regression
* Partial Least Squares (PLS)[24](#_bookmark30)

38

Partial Least Squares (PLS)



## PLS[25](#_bookmark31),[26](#_bookmark32) combines dimensionality reduction with regression. PLS components are chosen so to maximise covariance between the projection axis and the target (y).

39

The ”classic” (NIPALS/Wold PLS) model[24](#_bookmark30),[26](#_bookmark32)–[28](#_bookmark33).

The PLS Model

*u*1 = *b*1*t*1*, ..., ur* = *brtr*

”Inner relation”: a linear regression at latent variable level

*X* = *t*1*p′* + *t*2*p′* + *EX Y* = *u*1*q′* + *u*2*q′* + *EY*

1 2 1 2

X = original matrix (n x p) T = score matrix (n x r) P = loading matrix (p x r)

E*X* = X residual matrix (n x p)

Y = original matrix (n x q) U = score matrix (n x r) Q = loading matrix (q x r)

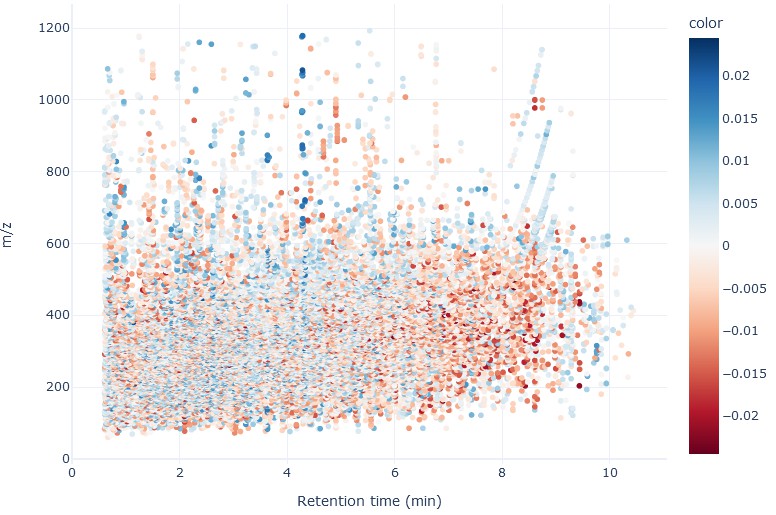
E*Y* = Y residual matrix (n x q)

The ”default” PLS algorithm can be used to explore covariance between 2 data matrices, akin to Canonical Correlation Analysis.

40

How to interpret a PLS model?

PLS - scores and loadings...

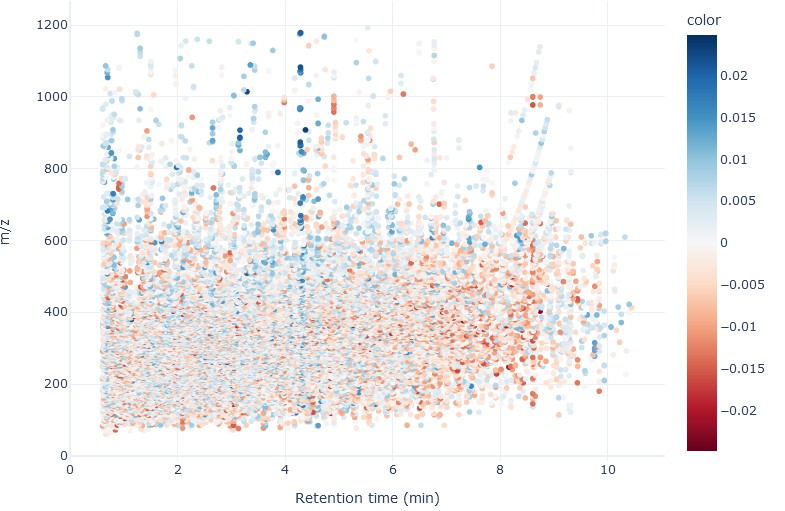
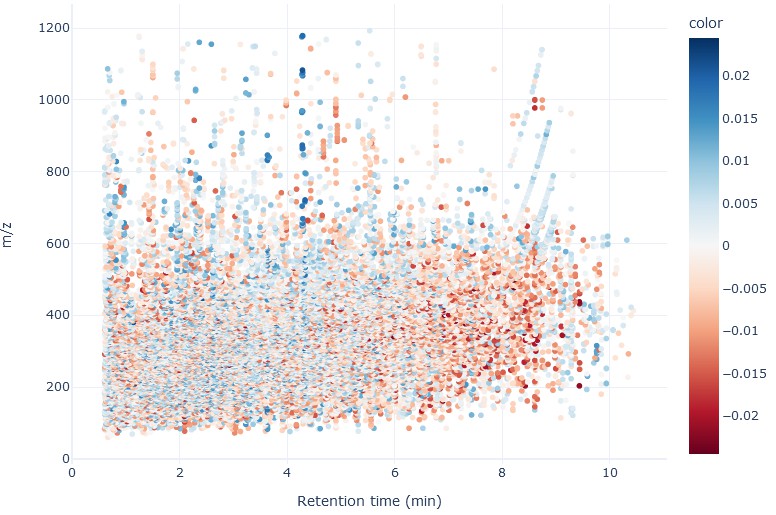


Scores plot PLS 1 Loadings *p*

41

How to interpret a PLS model?

PLS - loadings and weights...



PLS1 Loadings *p* PLS 1 Weights *w*

42

How to examine variable importance in PLS?

PLS model interpretation and variable importance

* Weights *w*
* Loadings *p*
* Regression coefficients *β*
* VIPs: Weighted sum (by the variance explained) of *w*2

Depends on the purpose: Use loadings *p* to interpret scores plots, and *w* for assessing which X variables have more covariance with y.

Warning

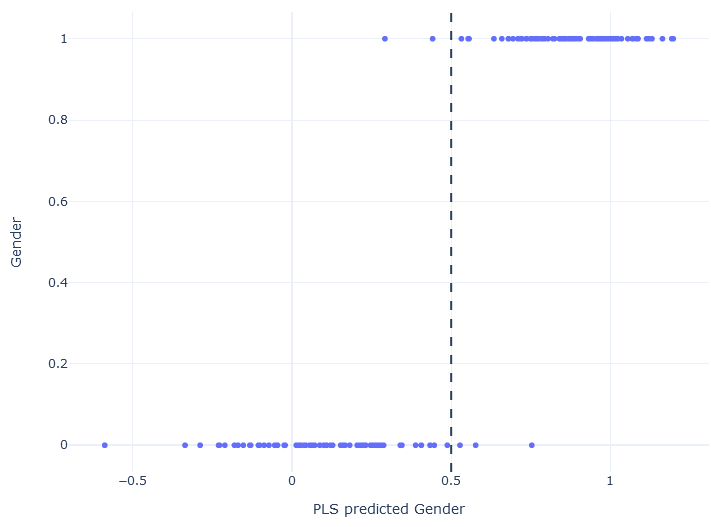
There are no formal procedures for ”statistical significance” of PLS models.

Don’t overthink multivariate model interpretation...

43

PLS-DA = PLS[29](#_bookmark34). Encode class membership as a Y dummy matrix (0’s and 1’s) and fit a PLS model.

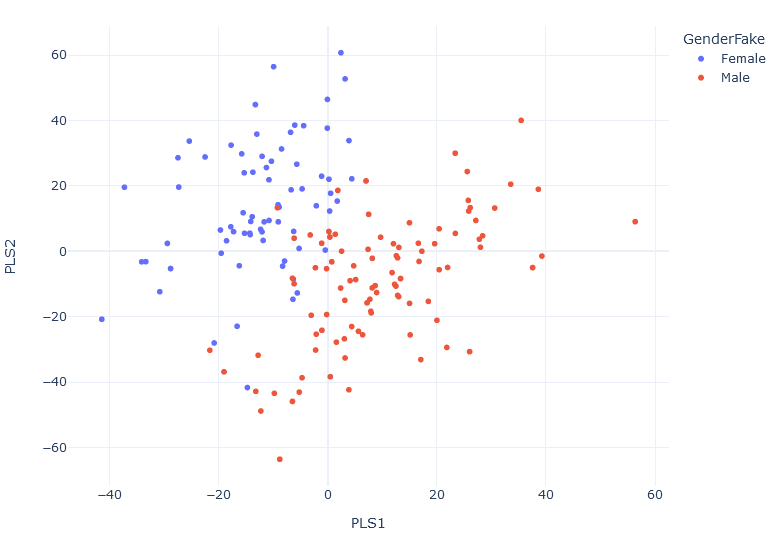
Partial Least Squares - Discriminant Analysis (PLS-DA)



Classification rules can be devised based on model prediction or scores[29](#_bookmark34)–[32](#_bookmark35).

44

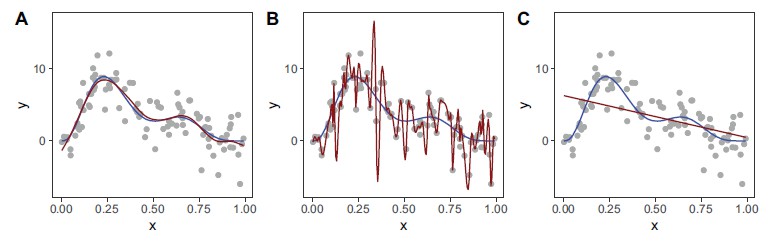
Partial Least Squares - Discriminant Analysis (PLS-DA)



*Good separation...*

45

Overfitting

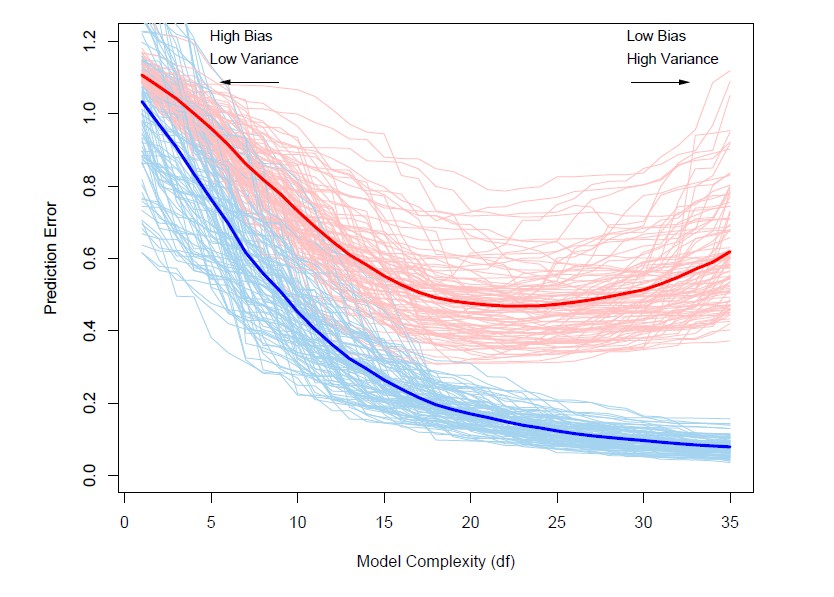


From Pedersen et al[33](#_bookmark36)

Balance between enough complexity and risk of overfitting... Supervised mod- els can overfit and PLS is no exception[34](#_bookmark37). Increasing the model complexity (number of PLS components) increases the likelihood of overfitting.

46

Overfitting



From Hastie et al[35](#_bookmark38)

Take home message: model performance must be assessed in held-out data.

47

Resampling and cross-validation methods allow us to obtain a better estimate of model error with a limited number of observations. Models are trained on the training sets and benchmarked on the test sets.

Model (cross)-validation

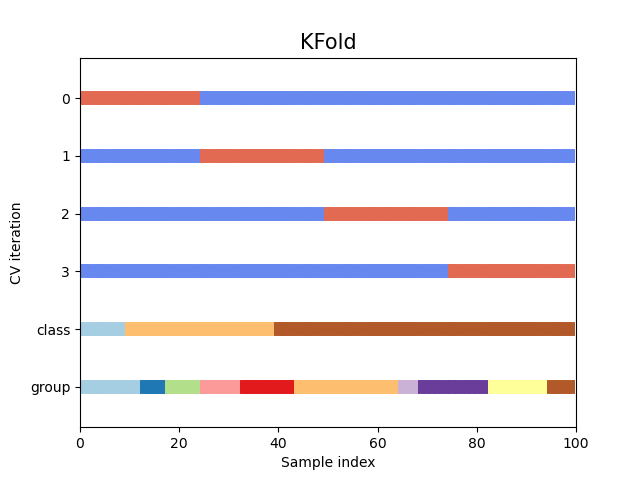
* Single data split: Keep % of the data as a held-out test set
* K-Fold: Split the data into K disjoint test sets with size *n* . Use remaining samples as training set. Leave one out cross validation (LOOCV) = K-Fold when *K* = *n*.

*K*

* Monte Carlo CV: Choose a % train-test split ratio, and sample randomly the members of the train and test set. Repeat this procedure *r* times (e.g. 1000).

48

K-fold Cross-validation



[From scikit-learn](https://scikit-learn.org/stable/auto_examples/model_selection/plot_cv_indices.html)

49

* Adapt K to your sample-size

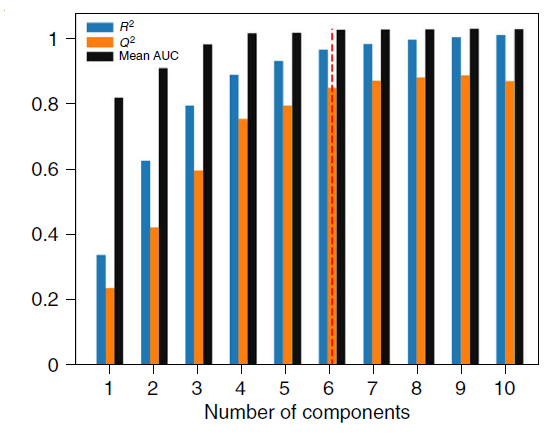
Cross-validation - recommendations

* Use stratified cross-validation - preserves the outcome distribution in folds
* Shuffle observations before cross-validation[36](#_bookmark39)
* Any parameters (e.g. scaling) or procedures (variable selection) should be part of CV procedure
* Whenever possible make use of repeated and double cross-validation schemes[37](#_bookmark40)
* Use open-source tools for flexibility: scikit-learn[38](#_bookmark41) and caret[39](#_bookmark42)

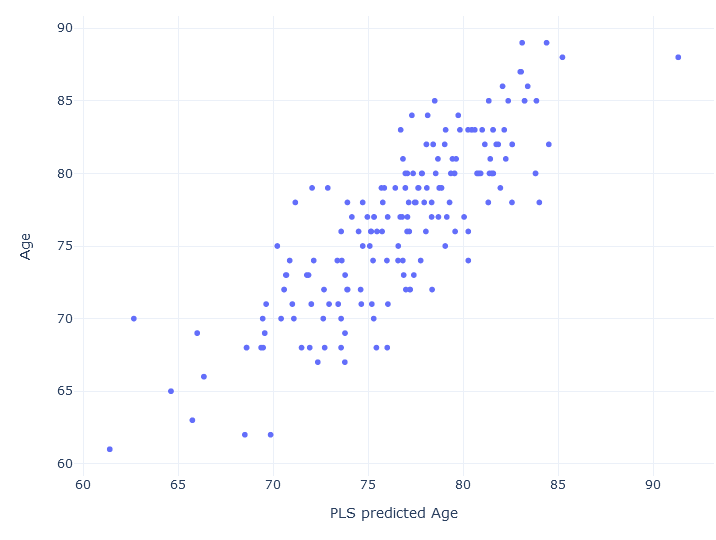
50

The choice of PLS components is critical. Use cross-validation and be parsi- monious (less is better)!

PLS - Choosing the number of components



51

Metrics for regression models:

Model performance metrics: Regression

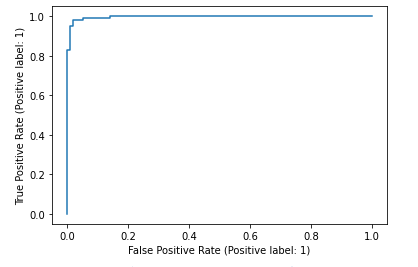
* *R*2: Variance of Y explained by the model. Should be estimated on external test sets (i.e. using CV)

*Y*

* *Q*2: Cross-validated version of the R-squared measure, calculated with K-Fold or leave one out CV

*Y*

52

Metrics for classification:

Model performance metrics: Classification

* Accuracy, Precision, Recall, Sensitivity, Specificity
* ROC curves and ROC AUC.
* f1-score, kappa
* Confusion Matrices
* *Q*2 in PLS-DA models - not recommended

*Y*

Warning

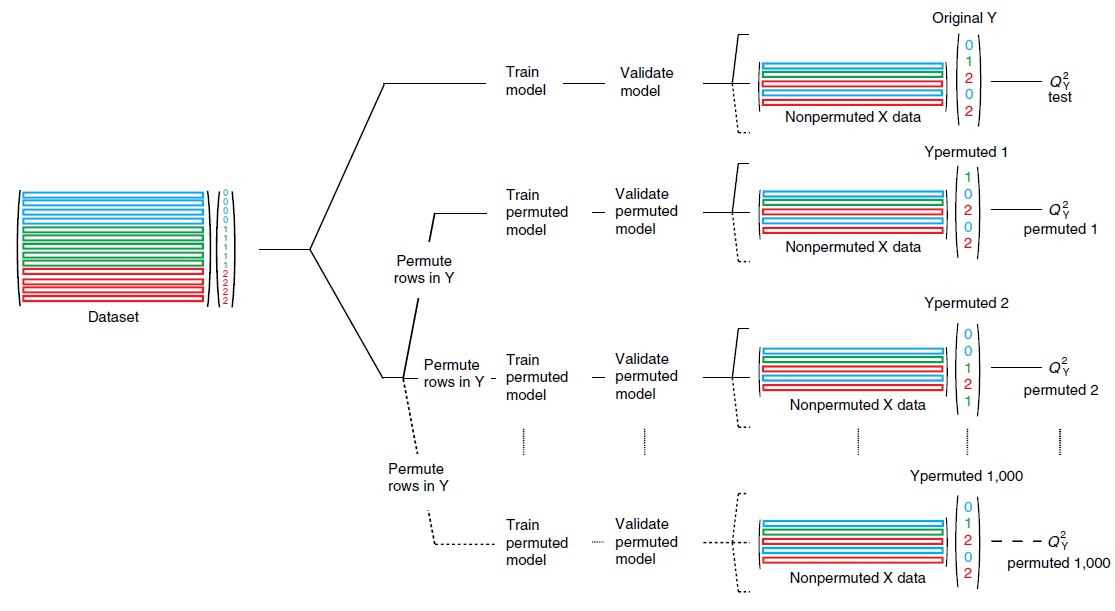
When the classes are unbalanced, make sure to use adequate metrics (e.g.

balanced accuracy).

53

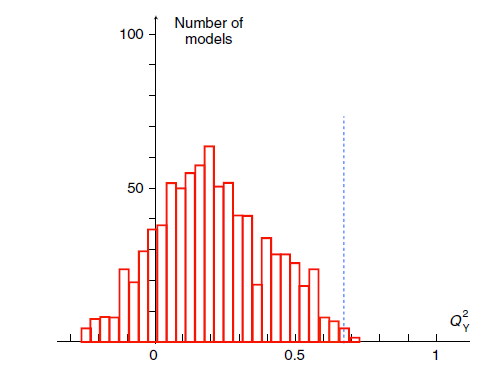
One step beyond: can we obtain equally good CV metrics by chance?

Permutation randomisation tests



54

Permutation randomisation tests



The permuted null distribution conveys the range of goodness-of-fit or clas- sification metrics that can be obtained from chance alone assuming a simi- lar dataset (sample size, correlation structure, number of variables, etc) and modeling strategy.

55

# [Conclusion](#_bookmark5)

There are no restrictions on what can be used, but choose statistical model based on what your aim is and characteristics of the data.

Conclusion and questions

Multivariate (PLS) based methods are common in the field, but have their own limitations (always cross-validate!!).

Visualise hits - including raw data/chromatograms! Univariate or multivariate analysis? Use both[40](#_bookmark43).

Questions?

For more tutorials, see Blaise et al[2](#_bookmark8) and these GitHub repositories: <https://github.com/Gscorreia89/chemometrics-tutorials>

<https://github.com/Gscorreia89/chemometrics-tutorials-urineLCMS>

56

# [References](#_bookmark6)

1. Sands, C. J. *et al.* Representing the Metabolome with High Fidelity: Range and Response as Quality Control Factors in LC-MS-Based Global Profiling. *Analytical Chemistry* 93, 1924–1933. ISSN: 15206882 (2021).
2. Blaise, B. J. *et al. Statistical analysis in metabolic phenotyping.* 2021.

[*https://www.nature.com/articles/s41596-021-00579-1*](https://www.nature.com/articles/s41596-021-00579-1).

1. Lewis, M. R. *et al.* Development and Application of Ultra-Performance Liquid Chromatography-TOF MS for Precision Large Scale Urinary Metabolic Phenotyping. *Analytical Chemistry* 88, 9004–9013. ISSN: 15206882 (2016).
2. Lovestone, S. *et al. AddNeuroMed - The european collaboration for the discovery of novel biomarkers for alzheimer’s disease.* in *Annals of the New York Academy of Sciences* 1180 (2009), 36–46. ISBN: 9781573317726.

56

1. Birkenbihl, C. *et al.* ANMerge: A Comprehensive and Accessible Alzheimer’s Disease Patient-Level Dataset. *Journal of Alzheimer’s Disease* 79, 423–431. ISSN: 1875-8908. [*https://www.medrxiv.org/*](https://www.medrxiv.org/content/early/2020/08/06/2020.08.04.20168229)

[*content/early/2020/08/06/2020.08.04.20168229*](https://www.medrxiv.org/content/early/2020/08/06/2020.08.04.20168229) (2021).

1. Smith, C. A., Want, E. J., O’Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry* 78, 779–787. ISSN: 00032700.

[*http://www.ncbi.nlm.nih.gov/pubmed/16448051*](http://www.ncbi.nlm.nih.gov/pubmed/16448051) (2006).

1. Do, K. T. *et al.* Characterization of missing values in untargeted

MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* 14, 1–18. ISSN: 15733890. [*https://link.springer.com/article/10.1007/s11306-*](https://link.springer.com/article/10.1007/s11306-018-1420-2)[*018-1420-2*](https://link.springer.com/article/10.1007/s11306-018-1420-2) (2018).

56

1. Hrydziuszko, O. & Viant, M. R. Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline. *Metabolomics* 8, 161–174. ISSN: 15733882. [*https://link.springer.com/article/10.1007/s11306-*](https://link.springer.com/article/10.1007/s11306-011-0366-4)[*011-0366-4*](https://link.springer.com/article/10.1007/s11306-011-0366-4) (2012).
2. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in1H NMR metabonomics. *Analytical Chemistry* 78, 4281–4290. ISSN: 00032700 (2006).
3. Dunn, W. B. *et al.* Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols* 6, 1060–1083. ISSN: 1754-2189.

[*https://www.nature.com/articles/nprot.2011.335*](https://www.nature.com/articles/nprot.2011.335) (2011).

56

1. Kohl, S. M. *et al.* State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics* 8, 146–160. ISSN: 15733882. */pmc/articles/PMC3337420//pmc/articles/ PMC3337420/?report=abstracthttps:*

[*//www.ncbi.nlm.nih.gov/pmc/articles/PMC3337420/*](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3337420/) (2012).

1. Saccenti, E. Correlation Patterns in Experimental Data Are Affected by Normalization Procedures: Consequences for Data Analysis and Network Inference. *Journal of Proteome Research* 16, 619–634. ISSN: 15353907. [*https://pubs.acs.org/sharingguidelines*](https://pubs.acs.org/sharingguidelines) (2017).
2. Correia, G. D. S. *et al.* 1H NMR Signals from Urine Excreted Protein Are a Source of Bias in Probabilistic Quotient Normalization. *Analytical Chemistry* 94, 6919–6923. ISSN: 0003-2700. [*https:*](https://pubs.acs.org/doi/full/10.1021/acs.analchem.2c00466)

[*//pubs.acs.org/doi/full/10.1021/acs.analchem.2c00466*](https://pubs.acs.org/doi/full/10.1021/acs.analchem.2c00466)

(2022).

56

1. Van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K. & van der Werf, M. J. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics* 7, 142. ISSN: 14712164. arXiv: [*806598*](https://arxiv.org/abs/806598).

[*http://bmcgenomics.biomedcentral.com/articles/10.*](http://bmcgenomics.biomedcentral.com/articles/10) *1186/1471-2164-7-142http:*

[*//www.ncbi.nlm.nih.gov/pubmed/16762068*](http://www.ncbi.nlm.nih.gov/pubmed/16762068) (2006).

1. Purohit, P. V., Rocke, D. M., Viant, M. R. & Woodruff, D. L. Discrimination models using variance-stabilizing transformation of metabolomic NMR data. *OMICS A Journal of Integrative Biology* 8, 118–130. ISSN: 15362310. [*www.liebertpub.com*](http://www.liebertpub.com/) (2004).
2. Veselkov, K. A. *et al.* Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry* 83, 5864–5872. ISSN: 00032700.

[*https://pubs.acs.org/doi/full/10.1021/ac201065j*](https://pubs.acs.org/doi/full/10.1021/ac201065j) (2011).

56

1. Chadeau-Hyam, M. *et al.* Metabolic profiling and the metabolome-wide association study: Significance level for biomarker identification. *Journal of Proteome Research* 9, 4620–4627. ISSN: 15353893 (2010).
2. Peluso, A., Glen, R. & Ebbels, T. M. Multiple-testing correction in metabolome-wide association studies. *BMC Bioinformatics* 22. ISSN: 14712105 (2021).
3. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 289–300. ISSN: 2517-6161. [*https:*](https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1995.tb02031.x)

[*//onlinelibrary.wiley.com/doi/full/10.1111/j.2517-*](https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1995.tb02031.x)[*6161.1995.tb02031.x*](https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1995.tb02031.x) (1995).

1. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29, 1165–1188. ISSN: 0090-5364. arXiv: [*0801.1095*](https://arxiv.org/abs/0801.1095).

[*http://projecteuclid.org/euclid.aos/1013699998*](http://projecteuclid.org/euclid.aos/1013699998) (2001).

56

1. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* 100, 9440–9445. ISSN: 00278424.

[*www.pnas.orgcgidoi10.1073pnas.1530509100*](http://www.pnas.orgcgidoi10.1073pnas.1530509100/) (2003).

1. Efron, B. Size, power and false discovery rates. *Annals of Statistics* 35,

1351–1377. ISSN: 0090-5364 (2007).

1. Frank, L. E. & Friedman, J. H. A statistical view of some chemometrics regression tools. *Technometrics* 35, 109–135. ISSN: 15372723. [*https://www.tandfonline.com/action/*](https://www.tandfonline.com/action/journalInformation?journalCode=utch20)[*journalInformation?journalCode=utch20*](https://www.tandfonline.com/action/journalInformation?journalCode=utch20) (1993).
2. Wold, S., Ruhe, A., Wold, H. & Dunn, III, W. J. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. en. *SIAM Journal on Scientific and Statistical Computing* 5, 735–743. ISSN: 0196-5204.

[*http://epubs.siam.org/doi/abs/10.1137/0905052*](http://epubs.siam.org/doi/abs/10.1137/0905052) (1984).

56

1. Wold, S., Sjöström, M. & Eriksson, L. *PLS-regression: A basic tool of chemometrics.* in *Chemometrics and Intelligent Laboratory Systems* 58 (2001), 109–130. ISBN: 0169-7439. arXiv: [*S0169-74390100155-1*](https://arxiv.org/abs/S0169-7439%2001%2000155-1).
2. Geladi, P. & Kowalski, B. R. Partial least-squares regression: a tutorial.

*Analytica Chimica Acta* 185, 1–17. ISSN: 00032670. arXiv:

*arXiv:1011.1669v3* (1986).

1. Wold, H. Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach. *Journal of Applied Probability* 12, 117–142. ISSN: 0021-9002.

[*https://www.cambridge.org/core/journals/journal-of-*](https://www.cambridge.org/core/journals/journal-of-applied-probability/article/abs/soft-modelling-by-latent-variables-the-nonlinear-iterative-partial-least-squares-nipals-approach/3DEC372F98373C30CB4DE1EE0C3DFEDF)[*applied-probability/article/abs/soft-modelling-by-*](https://www.cambridge.org/core/journals/journal-of-applied-probability/article/abs/soft-modelling-by-latent-variables-the-nonlinear-iterative-partial-least-squares-nipals-approach/3DEC372F98373C30CB4DE1EE0C3DFEDF)[*latent-variables-the-nonlinear-iterative-partial-*](https://www.cambridge.org/core/journals/journal-of-applied-probability/article/abs/soft-modelling-by-latent-variables-the-nonlinear-iterative-partial-least-squares-nipals-approach/3DEC372F98373C30CB4DE1EE0C3DFEDF)[*least-squares-nipals-*](https://www.cambridge.org/core/journals/journal-of-applied-probability/article/abs/soft-modelling-by-latent-variables-the-nonlinear-iterative-partial-least-squares-nipals-approach/3DEC372F98373C30CB4DE1EE0C3DFEDF)[*approach/3DEC372F98373C30CB4DE1EE0C3DFEDF*](https://www.cambridge.org/core/journals/journal-of-applied-probability/article/abs/soft-modelling-by-latent-variables-the-nonlinear-iterative-partial-least-squares-nipals-approach/3DEC372F98373C30CB4DE1EE0C3DFEDF) (1975).

1. Höskuldsson, A. PLS regression methods. *Journal of Chemometrics* 2,

211–228. ISSN: 0886-9383.

[*http://doi.wiley.com/10.1002/cem.1180020306*](http://doi.wiley.com/10.1002/cem.1180020306) (1988).

56

1. Brereton, R. G. & Lloyd, G. R. Partial least squares discriminant analysis: Taking the magic away. *Journal of Chemometrics* 28, 213–225. ISSN: 1099128X (2014).
2. Barker, M. & Rayens, W. Partial least squares for discrimination. *Journal of Chemometrics* 17, 166–173. ISSN: 08869383 (2003).
3. Gromski, P. S. *et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding.* 2015.
4. Indahl, U. G., Martens, H. & Næs, T. From dummy regression to prior probabilities in PLS-DA. *Journal of Chemometrics* 21, 529–536. ISSN: 08869383 (2007).
5. Pedersen, E. J., Miller, D. L., Simpson, G. L. & Ross, N. Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ* 2019, e6876. ISSN: 21678359.

[*https://peerj.com/articles/6876*](https://peerj.com/articles/6876) (2019).

56

1. Cloarec, O. Can we beat over-fitting? *Journal of Chemometrics* 28,

610–614. ISSN: 1099128X. [*https:*](https://onlinelibrary.wiley.com/doi/full/10.1002/cem.2602)

[*//onlinelibrary.wiley.com/doi/full/10.1002/cem.2602*](https://onlinelibrary.wiley.com/doi/full/10.1002/cem.2602)

(2014).

1. Hastie, T., Tibshirani, R. & Friedman, J. *Springer Series in Statistics.* 2,

83–85. ISBN: 9780387848570. arXiv: *arXiv:1011.1669v3*. [*http:*](http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf)

[*//www.springerlink.com/index/D7X7KX6772HQ2135.pdf*](http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf)

(2009).

1. Triba, M. N. *et al. PLS/OPLS models in metabolomics: The impact of permutation of dataset rows on the K-fold cross-validation quality parameters.* 2015. [*https://pubmed.ncbi.nlm.nih.gov/25382277/*](https://pubmed.ncbi.nlm.nih.gov/25382277/).

56

1. Szymańska, E., Saccenti, E., Smilde, A. K. & Westerhuis, J. A.

Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* 8, 3–16. ISSN: 1573-3882. [*http://www.pubmedcentral.nih.gov/articlerender.fcgi?*](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3337399&tool=pmcentrez&rendertype=abstract)[*artid=3337399&tool=pmcentrez&rendertype=abstract*](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3337399&tool=pmcentrez&rendertype=abstract)(2012).

1. Pedregosa, F., Grisel, O., Weiss, R., Passos, A. & Brucher, M. Scikit-learn : Machine Learning in Python. 12, 2825–2830. arXiv: *arXiv:1201.0490v2* (2011).
2. Kuhn, M. Building predictive models in R using the caret package. *Journal of Statistical Software* 28, 1–26. ISSN: 15487660. [*https://www.*](https://www.jstatsoft.org/index.php/jss/article/view/v028i05)[*jstatsoft.org/index.php/jss/article/view/v028i05*](https://www.jstatsoft.org/index.php/jss/article/view/v028i05)(2008).

56

1. Saccenti, E., Hoefsloot, H. C., Smilde, A. K., Westerhuis, J. A. & Hendriks, M. M. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics* 10, 361–374. ISSN: 15733890. [*http://link.springer.com/10.1007/s11306-013-0598-6*](http://link.springer.com/10.1007/s11306-013-0598-6)(2014).

56