Uncovering informative content in metabolomics data:

From pre-processing of ^{1}H NMR spectra to biomarkers discovery in multifactorial designs

PhD thesis

Manon Martin

CBIO, DDUV, UCLouvain

Thesis supervisor: Prof. Bernadette Govaerts

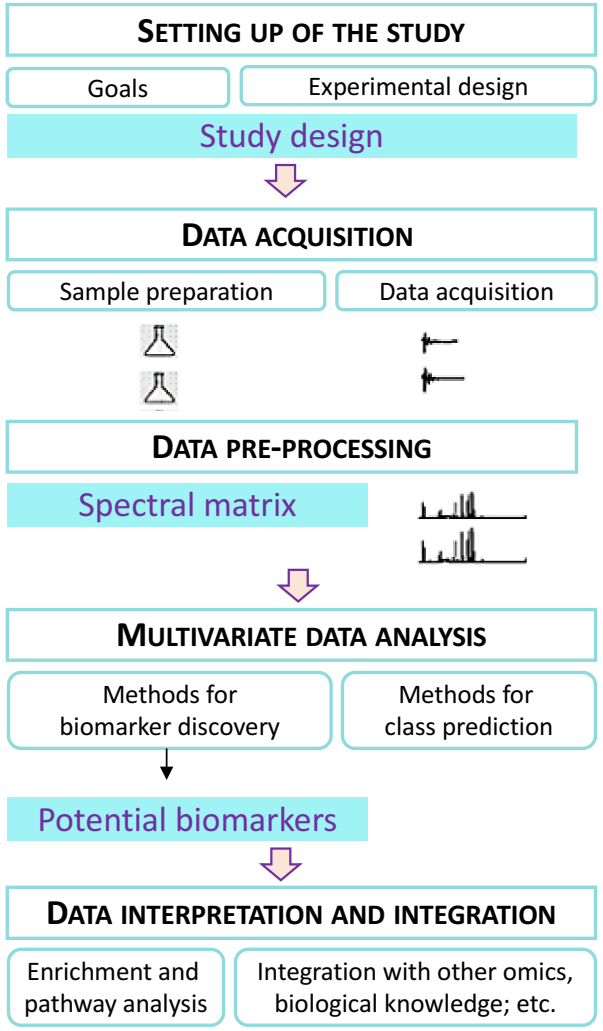
UCLouvain

Introduction

[Chapter 1]

Complexity of metabolomics studies

A typical metabolomics study



Challenging data treatment and analysis:

- Expensive data acquisition → reduced Nobs
- Nvar ≫ Nobs: High dimensional but short dataset
- Biological variability and instrumental noise/artifacts
- Need of heavy pre-processing
- Complex correlation structure: 1 metabolite can have ≠ peaks and descriptors, and correlations between metabolites → multicollinearity

Where statisticians do help?

- Everywhere...
- DOE, data pre-processing, biomarker search, predictive modelling, data fusion, etc.

Doctoral project

= 3 different & complementary research axes in metabolomics

Datasets

- Mostly raw and pre-processed metabolomics ^{1}H NMR spectra from blood and urine samples (human, rats)
- Generated from simple or multifactorial experimental designs

End products of this thesis :

Deliver practical solutions to non-statistician researchers

All Chapter's R codes available on GitHub (https://github.com/ManonMartin/thesisMaterial)

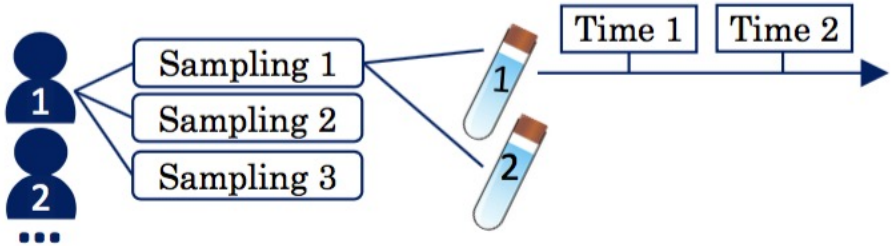
Measure the spectral repeatability

- How to quantify the spectral repeatability/reproducibility and the quality of samples analysed by: ≠ acquisition techniques (1D vs 2D),≠ analytical platforms (MS, NMR), with ≠ pre-processings,≠ sample preparations, etc.?
- Statistical perspective of spectral reproducibility/repeatability and quality control not yet well studied in metabolomics

Applied methodologies:

- Metabolomic Informative Content (MIC) concept from Féraud et al. (2015) with different quantitative quality indices for predictive power and clusters homogeneity applied to different pre-processing strategies [Chapter 2]
- ASCA framework to quantify and compare the sources of variability in the data with fixed and mixed effects models [Chapters 4 and 5]

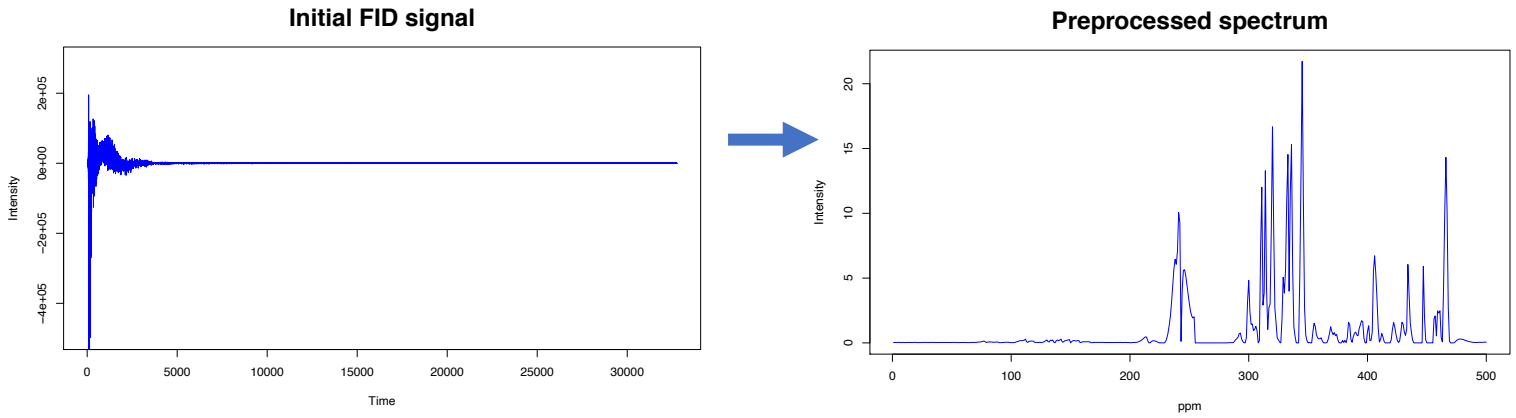
Nested design representation



PepsNMR for ^{1}H NMR data pre-processing

[Chapter 2]

1D ^{1}H NMR data pre-processing



→ These analytical platforms need a dedicated and advanced pre-processing

Current pre-processing (Engel et al., 2013):

- Manual & time-consuming
- Can be incomplete
- Proprietary software and black box
- No clear application strategy: Which method? In which order?

→ Incentive for the package creation

Input: Free Induction Decay (FID) = discretised complex signal in the time domain with > 30.000 points

Output: (n × m) spectral matrix with n observations and m variables (= spectral intensities) in the frequency domain

Pre-processing = mathematical & statistical algorithms designed for:

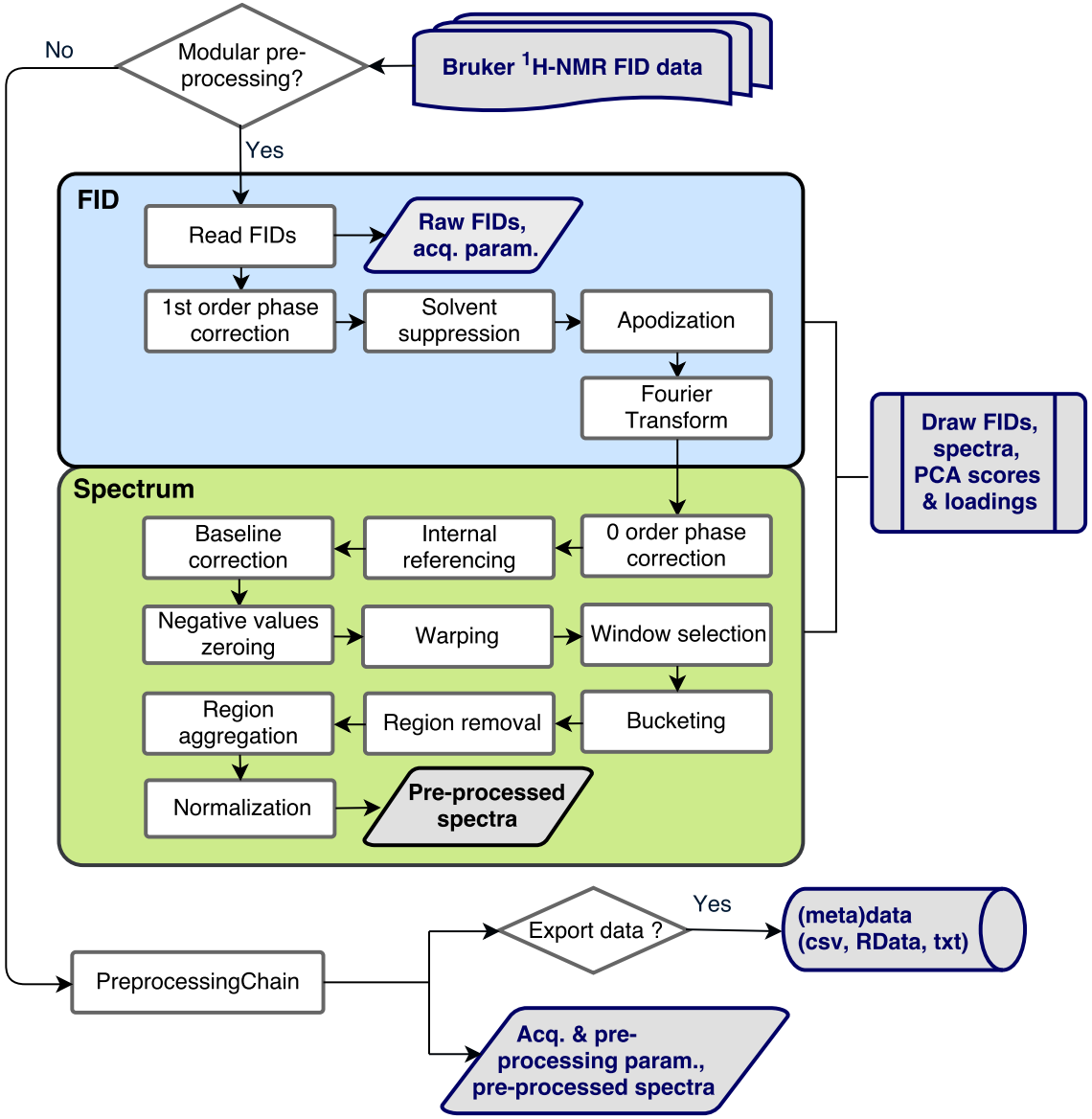
- Instrumental artefacts & biologically-related corrections
- SNR increase
- Data domain or scale transformation
- Data reduction

R package PepsNMR

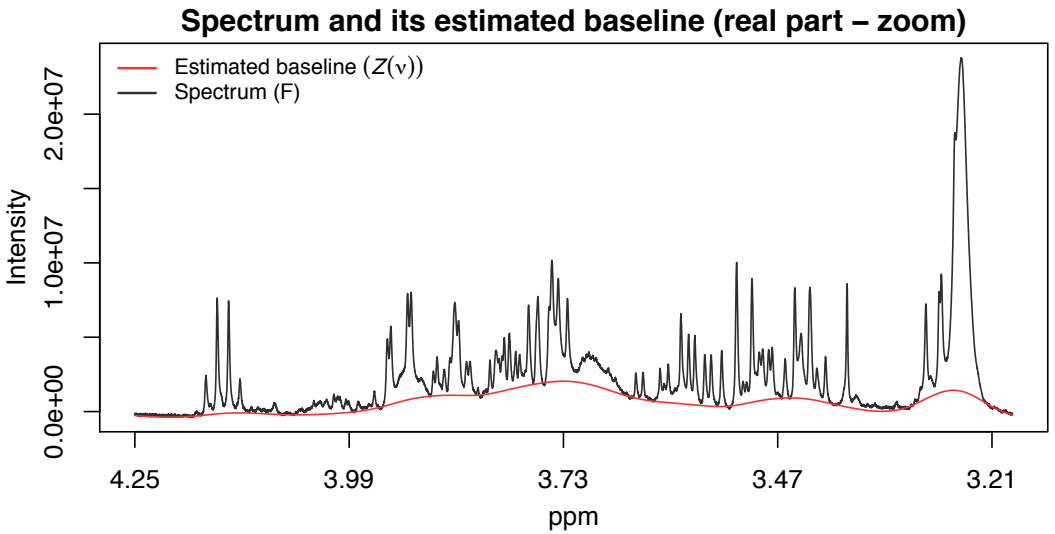
Chapter's objectives

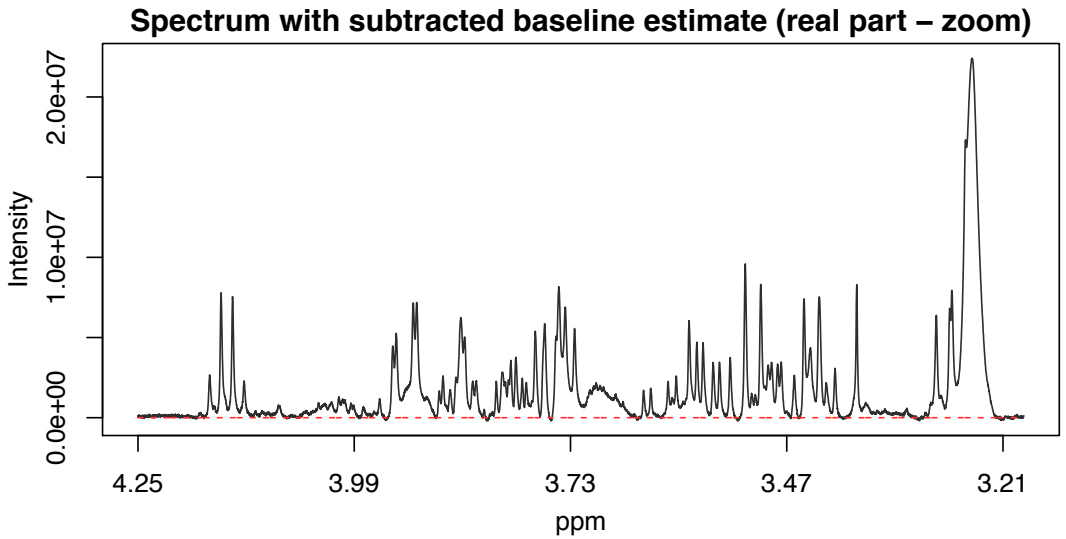
- Publish a complete and flexible software with innovative and well known algorithms for ^{1}H NMR pre-processing
- Package validation and consolidation
- Comparison with standard pre-processing based on quantitative quality criteria

- Former Matlab package (worked with a UI) mainly developed by Eli Lilly, P. H. C. Eilers and J. Vanwinsberghe (Vanwinsberghe, 2005)
- Includes gold-standard but also non- and semi-parametric methods



Baseline Correction





Issue: Presence of baseline artifacts (should be flat) in the spectrum *f*

Example of a non-parametric smoothing method:

- Asymmetric Least Squares smoothing (Eilers and Boelens, 2005) to estimate & remove the baseline *z*

$$\arg\min_{z} \; Q = \sum_{j=1}^{m} \psi_j (f_j - z_j)^2 + \lambda \sum_{j=3}^{m} (\Delta^2 z_j)^2$$

with weight ψ_j for the j^{th} frequency: if $f_j > z_j$, $\psi_j = p$,
else $\psi_j = (1-p)$ with $p \in [0, 1]$ close to 0.

Spectra realignments

Issue: Variation in experimental conditions (*e.g.* pH, temperature or concentration) induce misalignments between identical features from different spectra → intensities cannot be compared

Example of a semi-parametric method: Semi-parametric Time Warping (SPTW)
SPTW is inspired by (Eilers, 2004) and (Van Nederkassel et al., 2006)

Principle: Build a warping function $w(\nu)$ such that the distance between a warped spectrum $F(w(\nu))$ and a reference spectrum G is minimised by Penalized Least Squares:

$$\underset{\beta_k,\alpha_l}{\arg\min}\ Q = \sum_{j=1}^{m}(g_j - f_j(w(\nu)))^2 + \lambda\sum_{l=3}^{L}(\Delta^2\alpha_l)^2 + \kappa\sum_{l=1}^{L}\alpha_l^2$$

$$\text{with } w(\nu) = \sum_{k=0}^{K}\beta_k\nu^k + \sum_{l=1}^{L}\alpha_l B_l(\nu)$$

Soft Bucketing Used for dimension reduction & soft realignment (either trapezoidal or rectangular integration)

Comparison with classic pre-processing

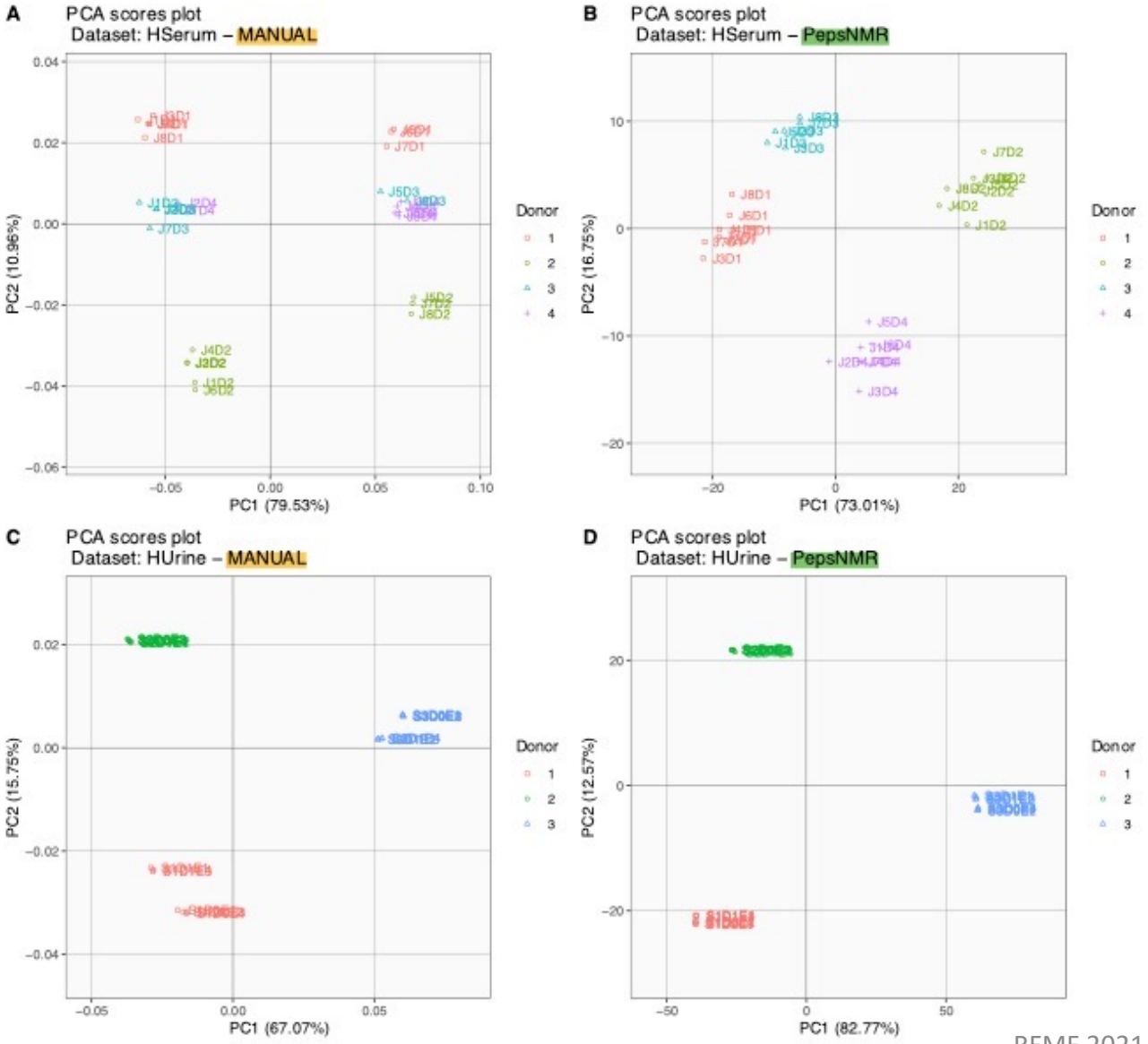
- Compare PepsNMR to “classic” pre-processing (Topspin+AMIX)
- 2 experimental datasets (HUrine & HSerum): ≠ donors (class), noisy factors (dilution, day, etc.)
- **MIC Quality criteria** measured from: inertia, PCA, clustering & PLS-DA indices

Max
Min

Dataset	Pre-processing	Inertia (%)		Ward hierarchical clustering				PLS-DA
		Between	Within	Rand	Adj. Rand	Dunn	D-B	Q2
HSerum	manual	21.44	78.56	0.64	0.14	0.62	0.56	-0.53
HSerum	PepsNMR	88.86	11.14	0.88	0.72	0.87	0.68	0.79
HUrine	manual	82.14	17.86	1	1	0.83	0.69	0.98
HUrine	PepsNMR	95.32	4.68	1	1	1.38	0.32	1

Table: Repeatability is partly assessed by variance decomposition methods: inertia decomposition (% of **between** and **within** inertia), Clustering/classification results for repeatability assessment: indexes derived from unsupervised Hierarchical clustering ((**Adjusted**) **Rand**, **Dunn** and Davies-Bouldin (**D-B**) indexes) and a supervised PLS-DA validation criterion (the cross-validated coefficient of determination **Q2**

Comparison with classic pre-processing



Main results for the comparison study

- Increased spectral repeatability
- Overall better information recovery
- Enhanced predictive power in a classification context

Built-in assets

- Modularity: can be combined/compared to other processing tools & software
- Open-source, semi-automated, structured documentation, reproducible results and reporting

End products

Publication This chapter has been published as a **journal article**:

M. Martin, B. Legat, J. Leenders, J. Vanwinsberghe, R. Rousseau, B. Boulanger, P. H.C. Eilers, P. De Tullio, and B. Govaerts. PepsNMR for 1H NMR metabolomic data pre-processing. *Analytica Chimica Acta*, 1019:1 - 13, 2018

R package The development version of **PepsNMR** is available on **GitHub** (https://github.com/ManonMartin/PepsNMR). The release version is published on **Bioconductor**. The package has *help* pages for functions and datasets, as well as a *vignette*, a long format case study. The *ad hoc* package, **PepsNMRData**, is also available on Github and Bioconductor.

The 9 first pre-processing steps of PepsNMR are implemented in **Workflow4Metabolomics** (Guitton et al., 2017) for NMR pre-processing on the Galaxy web platform.

Biomarker search in simple designs

[Chapter 3]

Feature selection with multivariate projection methods

Discrimination problem: dummy-coded response **y**

Tested methods:

- t-tests
- PLS (Varmuza and Filzmoser, 2016),
- OPLS (Trygg and Wold, 2002): PLS with an OSC filter and 1 predictive latent variable
- SPLS (Chun and Keleş, 2010): Addition of a L1 sparsity constraint (lasso) to the SIMPLS maximisation problem
- LSOPLS (Féraud et al., 2017): Take advantage of OPLS and SPLS for biomarkers discovery: sparsity and orthogonal variation filtering

Chapter's objectives

- Evaluate and compare different common biomarker selection techniques derived from multivariate projection methods in metabolomics based on a dataset with known true biomarkers' location.
- Compare several evaluation criteria

Meta-parameters:

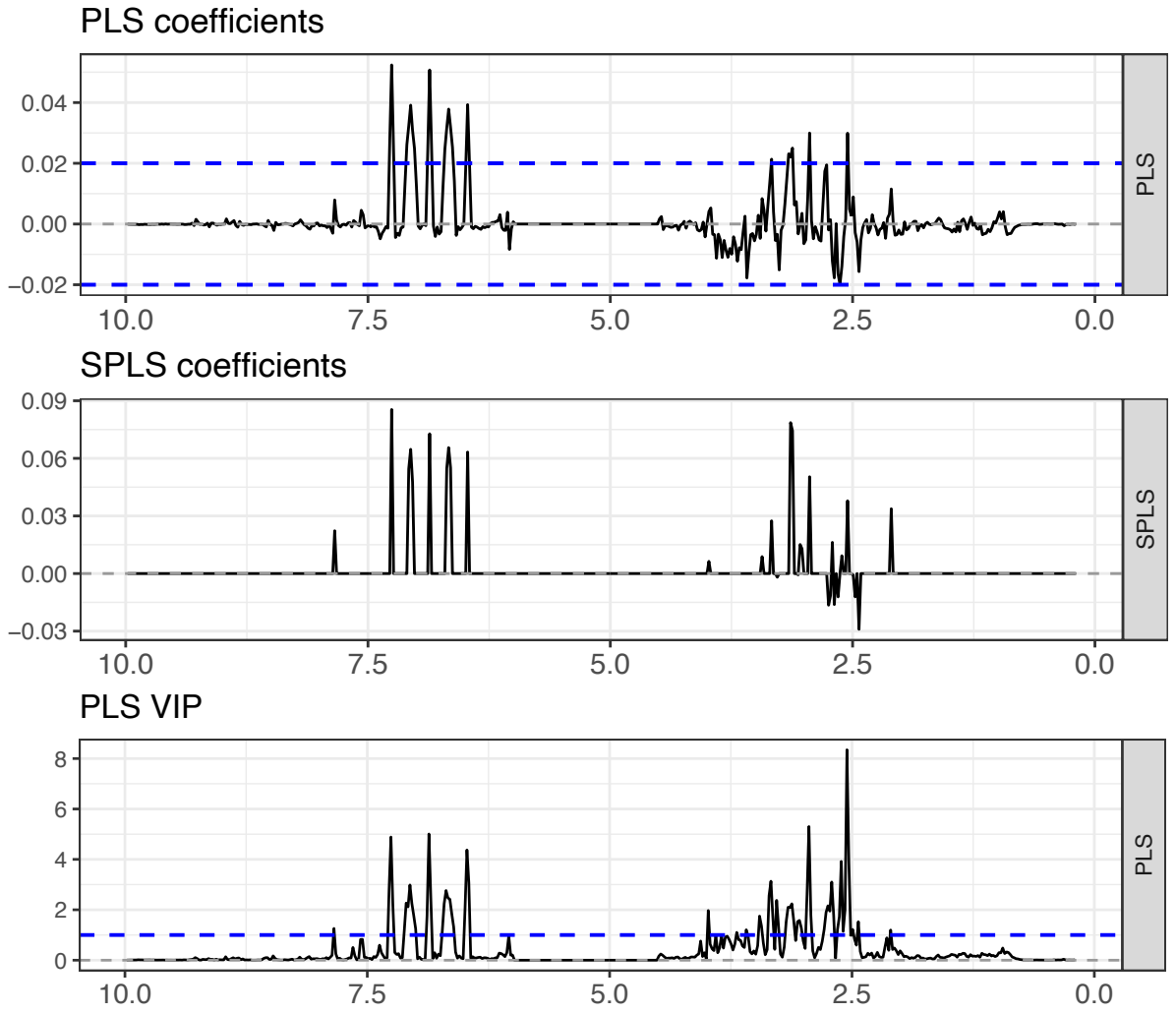
	PrC	OC	η
PLS	?	-	-
OPLS	1	?	-
SPLS	?	-	?
LSOPLS	1	?	?

Incentives for sparse methods

- Consistency of PLS estimators breaks down when m >> n
- Embedded: FS is integrated to the classifier training
- Sparse models are easier to interpret by clinicians

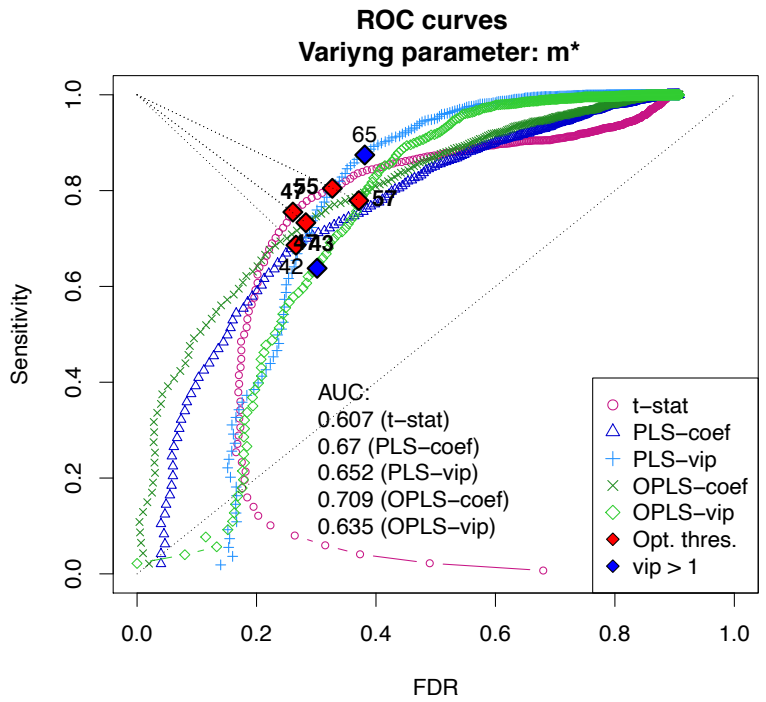
Comparison study

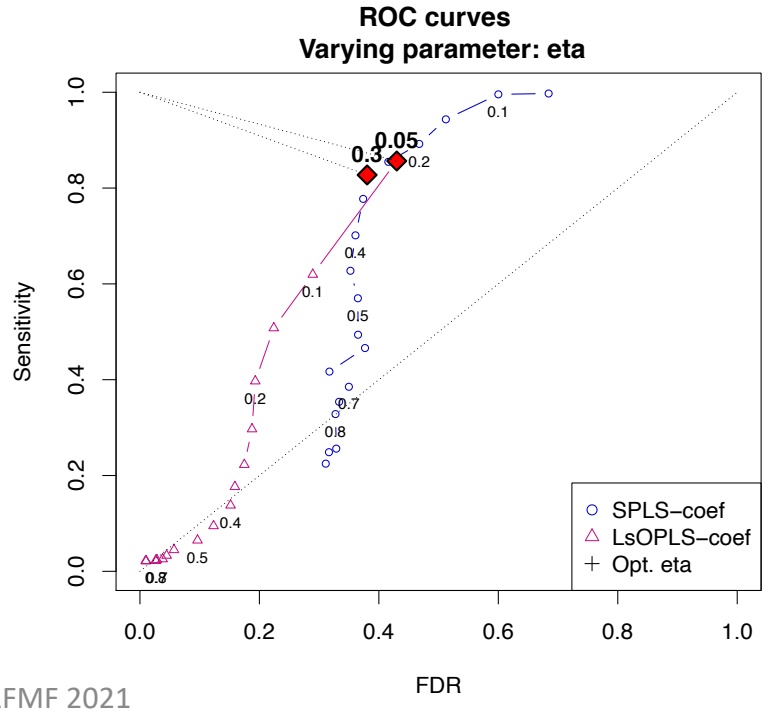
- **Semi-artificial NMR spectra**
  Rat urine spectra, half are artificially altered with artificial biomarkers
  ≠ Sample sizes (n = 60, 200)
  ≠ Signal-to-Noise-Ratio
  ≠ Biomarker intensities
- **Feature** *ranking or selection* based on **r**egression coefficients or VIP values
- **Comparison study:** double CV loop based on the AUC criterion
- **Evaluation criteria** of the final models:
  - ROC curves
  - Discriminant power
  - Stability
  - Graphical outputs



Main results

- Success of biomarker discovery heavily dependent on the multivariate method(s) used to extract the features but also on the data characteristics (different sample sizes, different levels of noise, etc.)
- Importance of the MP choice, especially for sparse embedded methods
- T-test: lots of false discoveries, sensitive to a smaller sample size, recovers correlated true features
- VIP ≥ 1 threshold: adapted for PLS but too high for OPLS
- Orthogonal filter: better for recovery in noisy area but not for lower intensities (and inversely without)
- Sparse classifiers: still a lot of FP, do not detect all correlated features
- PLS-VIP1: best overall performances on this dataset





End products

Publication This chapter has been published as a **Discussion Paper** at ISBA, UCLouvain:

> Martin, M. and Govaerts, B. (2019). Feature selection in metabolomics with PLS-derived methods. Discussion Paper DP 2019/20, Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium

The implemented (LS)OPLS methods were also employed in Féraud et al. (2017).

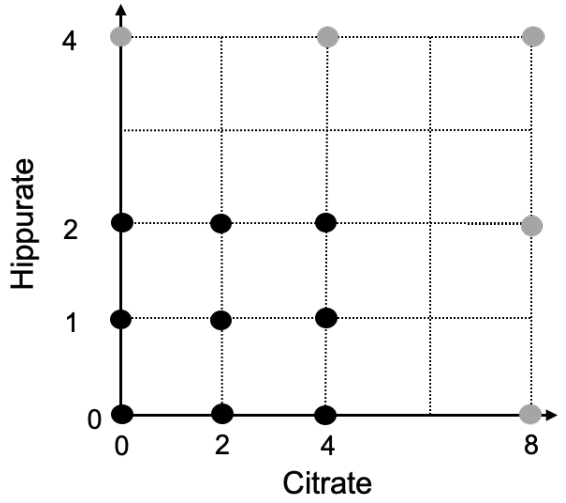
R package The R package **MBXUCL**, on GitHub (https://github.com/ManonMartin/MBXUCL) also contains all the applied methods, in addition to other common multivariate methods (PCA, clustering).

Extending ASCA^{+}

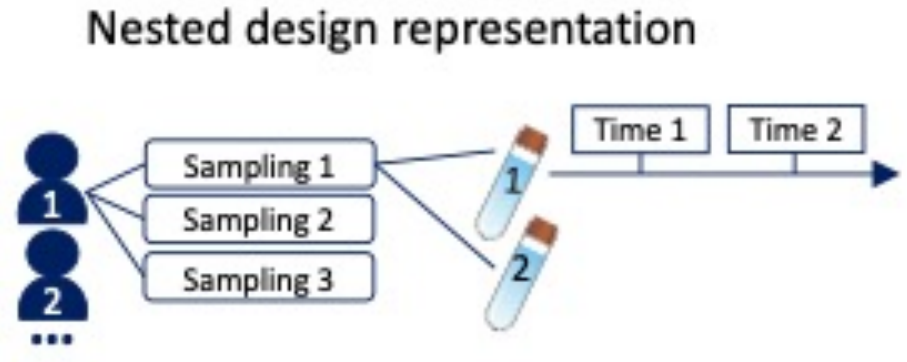
[Chapters 4 and 5]

Examples of multifactorial designs

- **Citrate-Hippurate Urine** dataset [Chapter 4]
  Urine spectra spiked with ≠ amounts of citrate and hippurate



- **Repeatability/reproducibility** dataset [Chapter 5]
  Human serum spectra affected by ≠ factors (nested design):
- biological (volunteer, sampling), analytical (tube, time replication) factors



Data frequency table

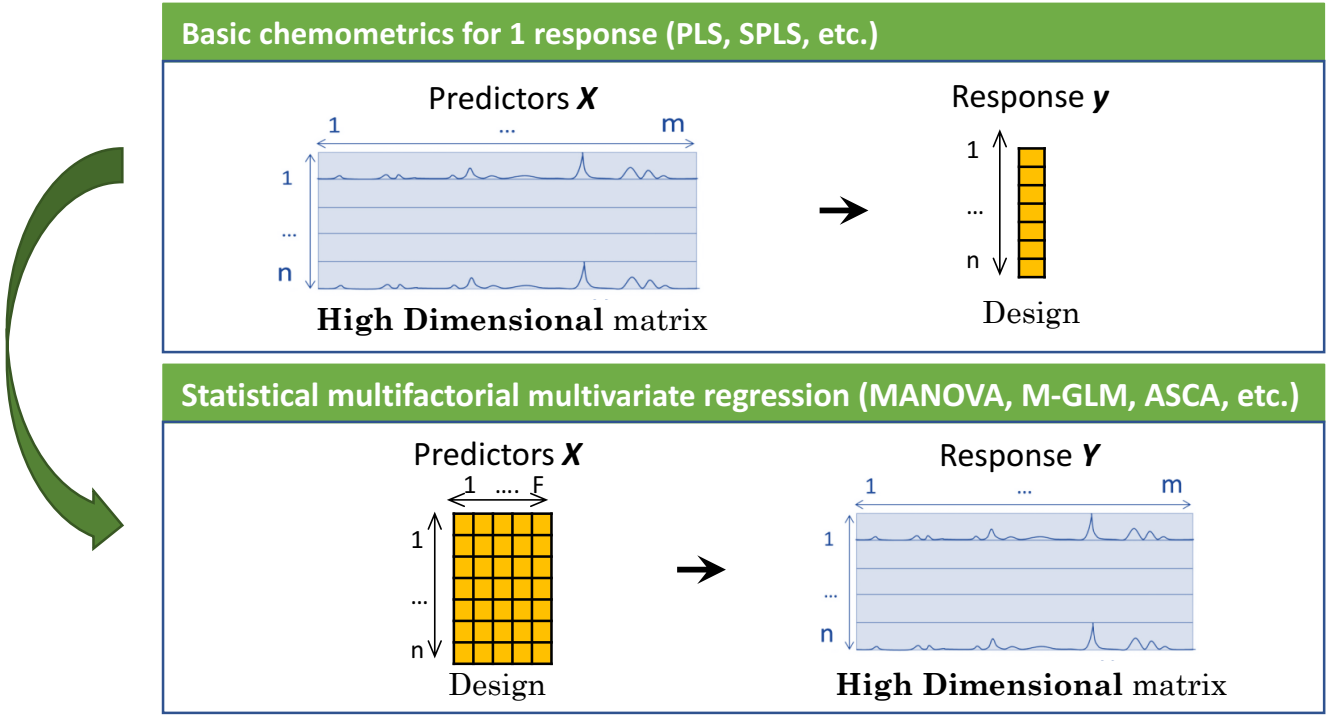
Tube	Time	Sampling		
		1	2	3
1	1	12	12	12
1	2	12	12	12
2	1	11	11	12
2	2	11	11	12

Multivariate projection methods

- **Disadvantages of multivariate projection methods (e.g. PLS):**
  - Rely on simple experimental designs (often only the class of the observation), without considering the full info. from the design
  - Few statistical tests
- **Classic statistical methods:**

Dealing with the dependent variable(s) dimension

- Linear or logistic regression, ANOVA, mixed models: **y** (1×m)
- MANOVA or multivariate-GLM: **Y** (n×m) with n < m
- **But** m ≫ n for metabolomic data !



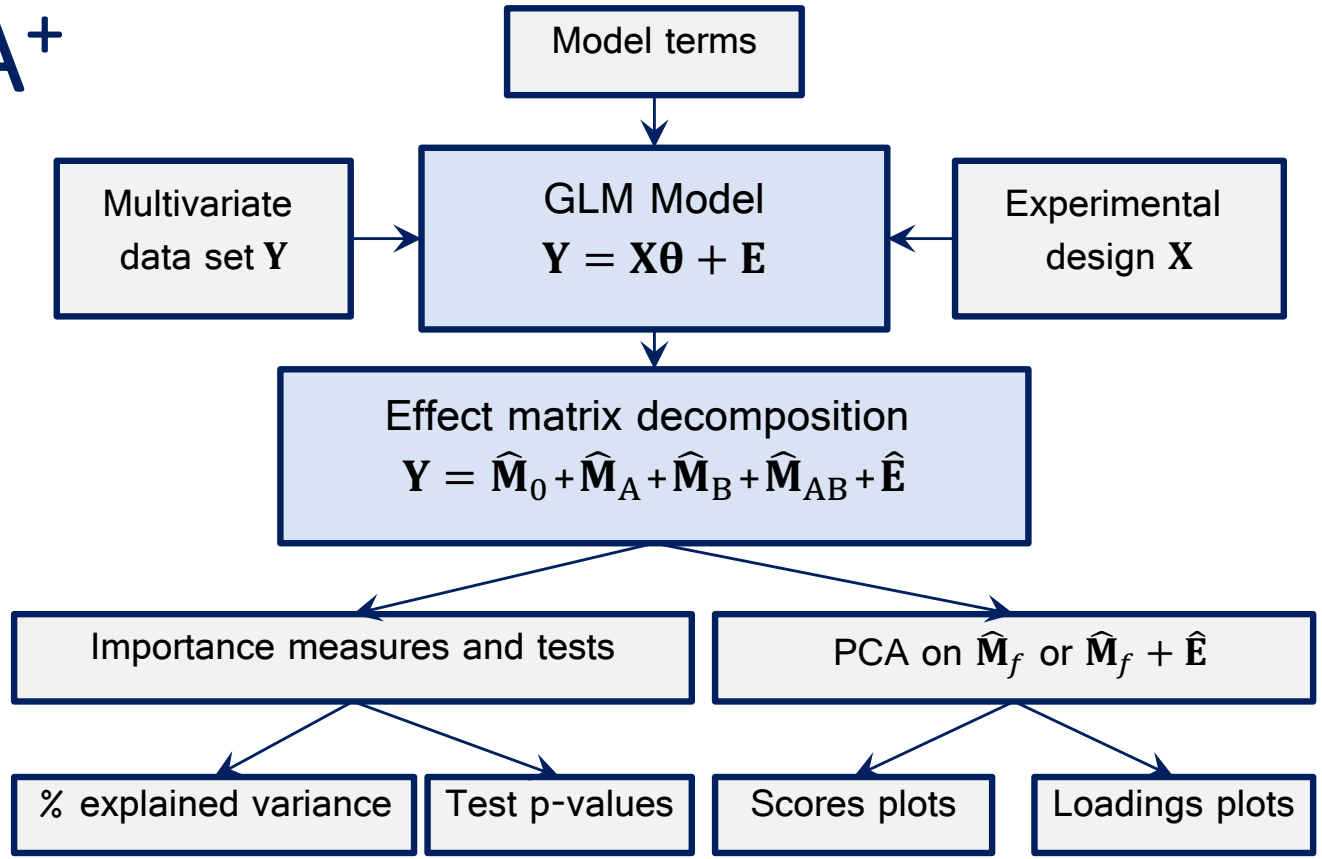
→ **One solution**: combine dimension reduction techniques and statistical modeling

State of the art: ASCA+

2 main steps:

1. **Parallel GLM** on each column of the response matrix **Y** and effect matrix decomposition with LS estimators according to the experimental design.
2. **Effect matrix interpretation**: Importance measure and effects test, PCA on each pure ($\hat{\boldsymbol{M}}_f = \boldsymbol{T}_f \boldsymbol{P}'$) or augmented ($\hat{\boldsymbol{T}}_f^a = (\hat{\boldsymbol{M}}_f + \boldsymbol{E}) \times \hat{\boldsymbol{P}}_f$) effect matrix.

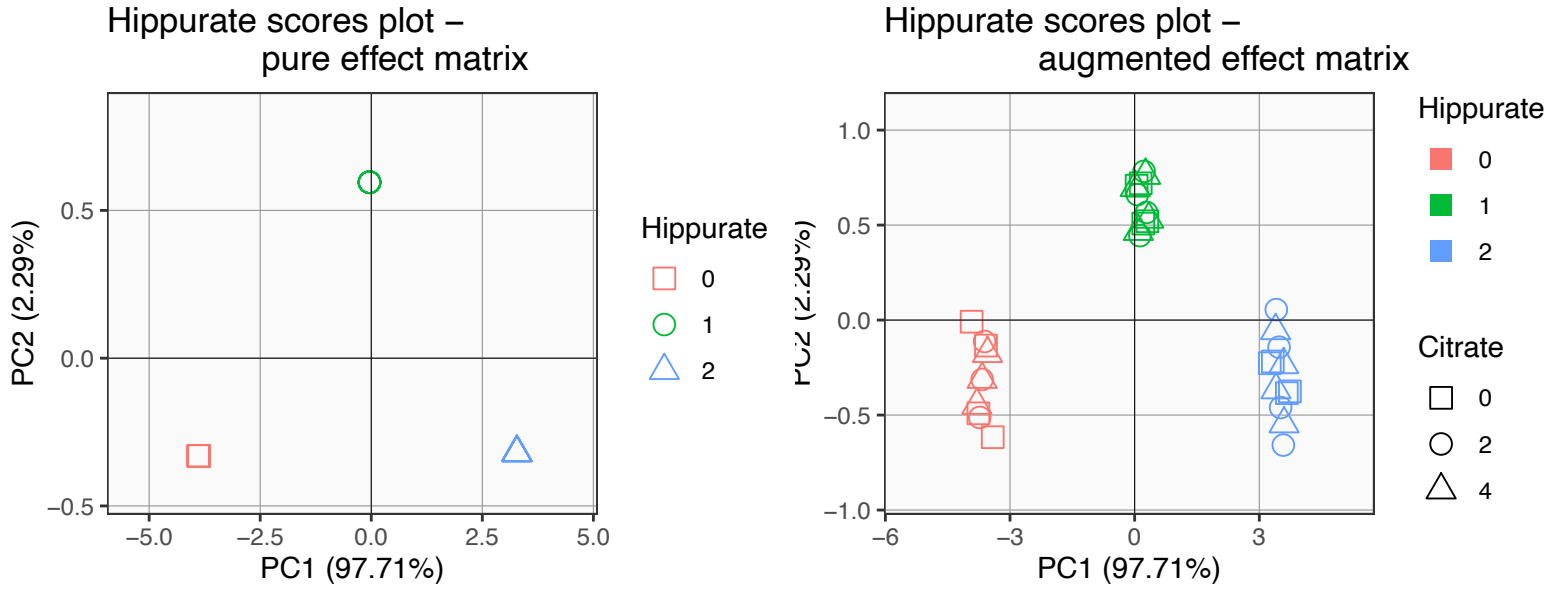
- = General Linear Model (GLM) version of ASCA
- For (un)balanced designs with fixed categorical variables
- Factors *sum* coding in the model matrix



Effect matrix augmentation

Table: Two way fixed ANOVA table with the $E(MS)$ and the F-test values

Effect	SS	df	MS	$E(MS)$	F-test
A	SSA	$a-1$	MSA	$\sigma^2 + \frac{nb\sum \alpha_i^2}{(a-1)}$	$\frac{MSA}{MSE}$
B	SSB	$b-1$	MSB	$\sigma^2 + \frac{na\sum \beta_i^2}{(b-1)}$	$\frac{MSB}{MSE}$
AB	$SSAB$	$(a-1)(b-1)$	$MSAB$	$\sigma^2 + \frac{n\sum\sum (\alpha\beta)_{ij}^2}{(a-1)(b-1)}$	$\frac{MSAB}{MSE}$
$Error$	SSE	$ab(n-1)$	MSE	σ^2	



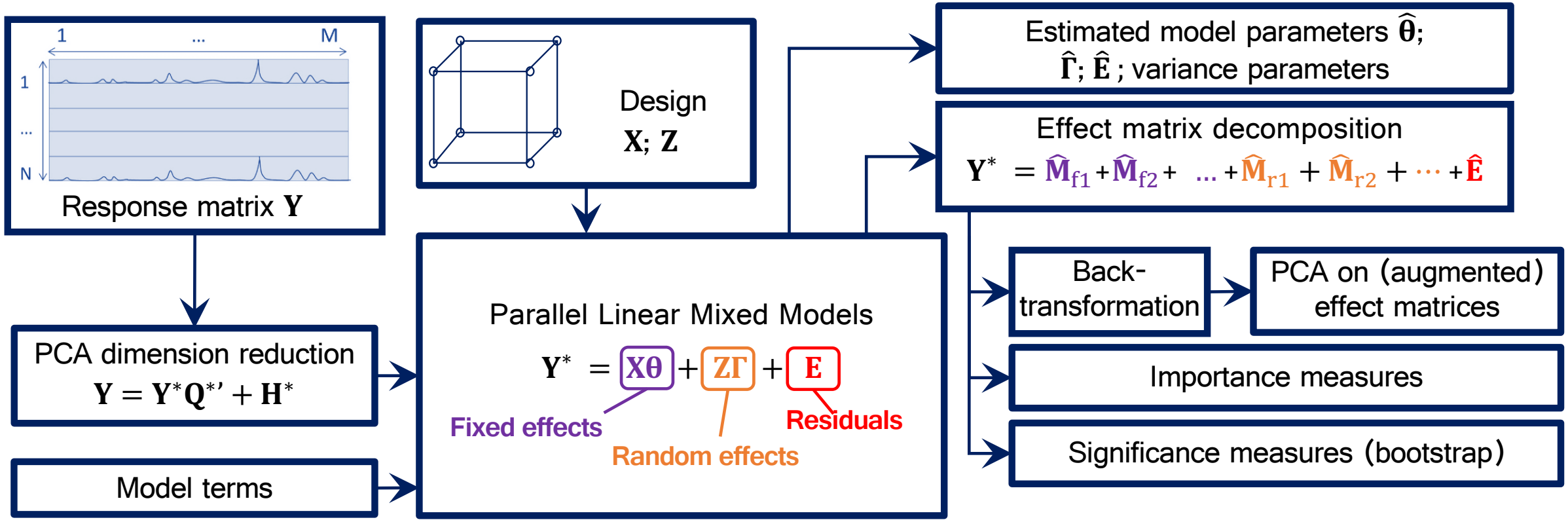
New developments

- **A. Modification of step 2** [Chapter 4]: Extension to other dimension reduction methods
  - **Incentive**: Applying other dimension reduction methods than PCA could provide supplementary information or other visualisation tools
- **B. Modification of step 1** [Chapter 5]: Extension of GLM to Linear Mixed Model (LMM) and bootstrap tests
  - **Incentives**:
    - ASCA+ not adapted to model random effects
    - Permutation tests implementation is challenging for advanced DOE [Anderson and Braak, 2003]

Chapters' objectives

- **Chapter 4**
  - Present and compare within a unified framework with common notations ASCA$^+$ and three global techniques (PARAFASCA, AComDim and AMOPLS) and extend them with the ASCA$^+$ GLM approach
  - advanced graphical results and permutations test to test the effects significance
- **Chapter 5**
  - Suggest a generic framework, LiMM-PCA, to extend the ASCA$^+$ methodology to deal with random effects and allow more flexible modelling to analyse more advanced designs, identify, quantify and compare the mixed (fixed or random) sources of variability
  - Adapt the workflow (effect matrix augmentation, effect importance, bootstrap test)
- Application of the methodologies to experimental datasets

LiMM-PCA methodology description



LiMM-PCA methodology description

- **Effect matrix augmentation**
  - Based on the E(MS) and F-tests in ANOVA tables for mixed effects + correction factor (involving the effective model dimension and the corresponding F-distribution quantile)
- **Quantification of effects importance**
  - Compare the fixed and random sources of variation
  - Nakagawa and Schielzeth (2013): General definition of marginal and conditional R2 in the univariate LMM for independent random effects.
  - Random effects: variance components $\widehat{var(\boldsymbol{y}_j^*|\boldsymbol{X})} = \sum_{r=1}^{R} \hat{\sigma}_{rj}^2 + \hat{\sigma}_{\epsilon j}^2$
  - Fixed effects: population variance of the estimated effect matrices $\hat{\sigma}_{fj}^2 = var(\hat{\boldsymbol{M}}_{fj})$
  - Global variance $\widehat{var(\boldsymbol{Y}^*)} = \sum_{j=1}^{m^*} \widehat{var(\boldsymbol{y}_j^*)} = \sum_{j=1}^{m^*} \left(\sum_{f=1}^{F} \hat{\sigma}_{fl}^2 + \sum_{r=1}^{R} \hat{\sigma}_{rl}^2 + \hat{\sigma}_{\epsilon j}^2 \right)$
- **Parametric bootstrap test**
  - Multivariate generalisation of the LLR test under the hypothesis of near independence between the individual transformed responses
  - Global Log Likelihood Ratio test statistics for (a group of) effect(s):
    $\Lambda_{\tilde{g}}^{obs} = 2[\sum_{j=1}^{m^*}(\log(L_{H_1,j\tilde{g}}) - \log(L_{H_0,j\tilde{g}}))]$
  - Parametric random effects and residuals bootstrap: simulate $\boldsymbol{Y}^*$ under H_0, measure $\Lambda_{\tilde{g}}^{b}$
  - $p_{\tilde{g}}^{\text{boot}} = \frac{\sum_{b=1}^{B} I(\Lambda_{\tilde{g}}^{b} \geq \Lambda_{\tilde{g}}^{\text{obs}}) + 1}{B+1}$

Results on the repeatability/reproducibility dataset

Unbalanced multifactorial design with 4 effects:

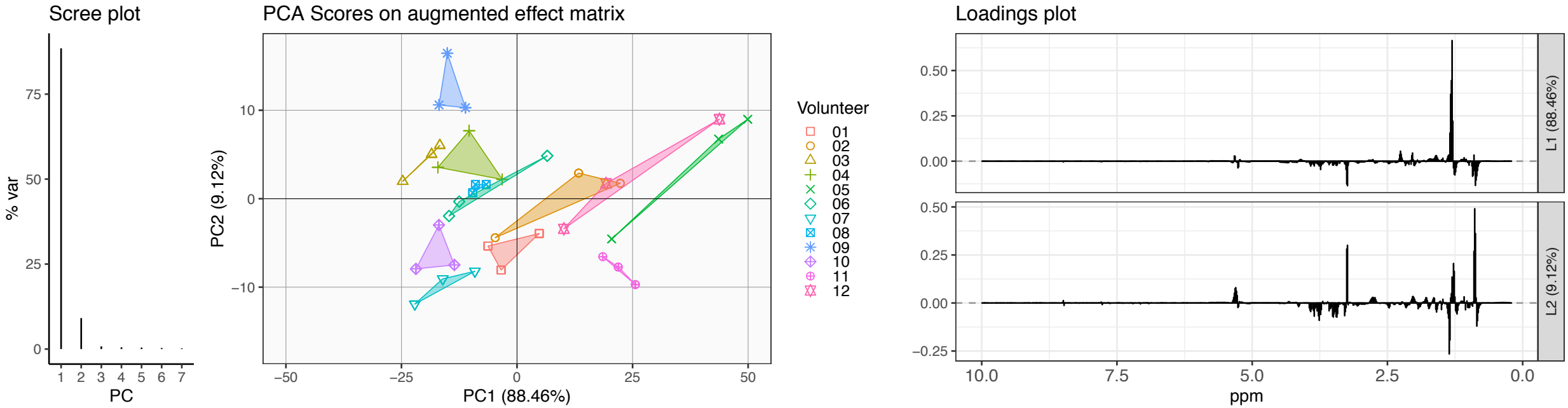
- 2 fixed (Tube, Time)
- 2 random (Volunteer, Sampling)

Main research questions:

- Compare the groups
- Quantify the variability of the repetitions and the patients
- Test the significance of these random/fixed effects

D

Volunteer effect matrix



Results on the repeatability/reproducibility dataset

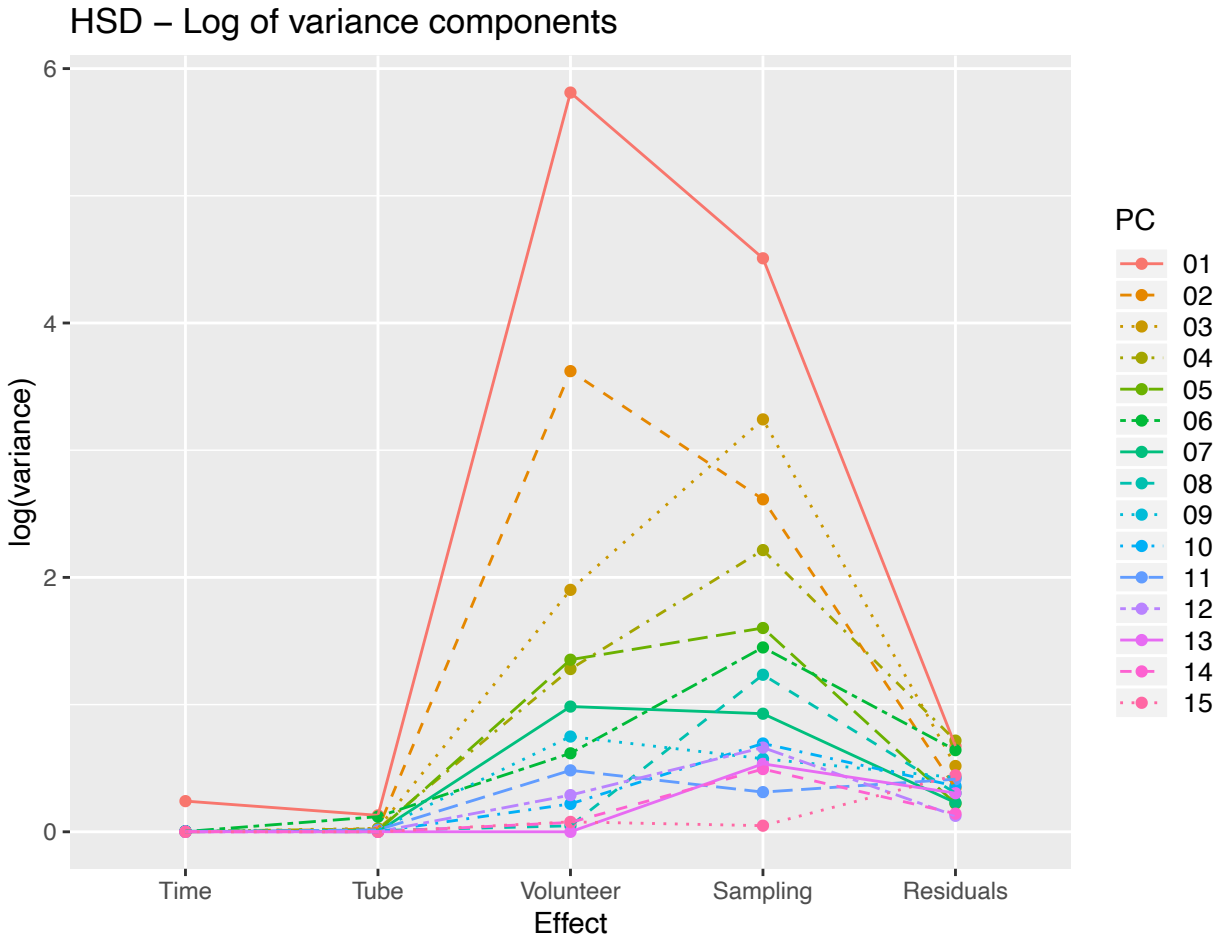
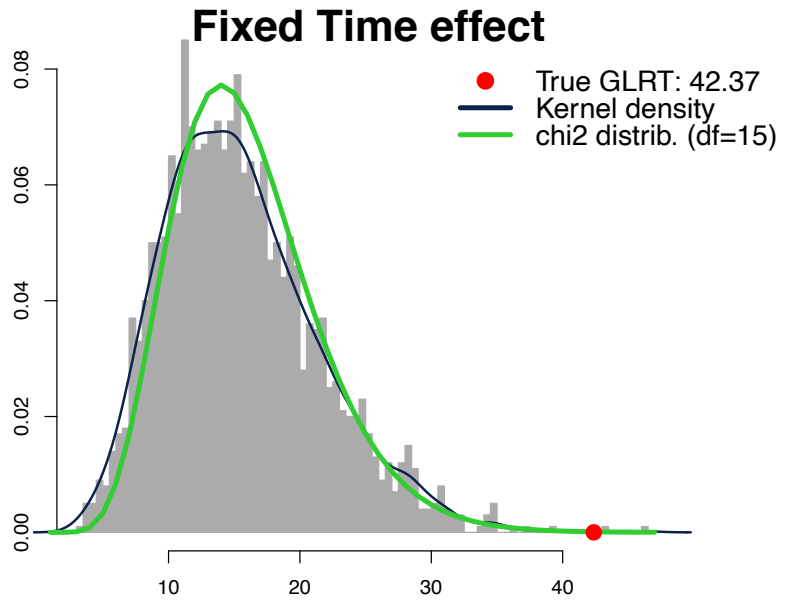


Table: Global effect importance and tests of effect significance.

Effect	**Global variance** (%)	**Bootstrapped p-value**	χ^2 **test**
Time	0.05	0.0015	$<$ 5e-04
Tube	0.07	$<$ 5e-04	$<$ 5e-04
Volunteer	70.80	$<$ 5e-04	-
Sampling	27.69	$<$ 5e-04	-
Residuals	1.38	-	-



Main results

LiMM-PCA

- Innovative extension and generalisation of the ASCA^{+} framework
- Enable to model unbalanced designs with random factors
- Global test of effect significance
- Quantification and comparison of the mixed variability sources
- Targeted applications: repeatability/reproducibility study & longitudinal data

End products

Publication These chapters have been published as **journal articles**:

Chapter 4

Guisset, S., Martin, M., and Govaerts, B. (2019). Comparison of PARAFASCA, AComDim, and AMOPLS approaches in the multivariate GLM modelling of multi-factorial designs. *Chemometrics and Intelligent Laboratory Systems*, 184:44 - 63

Chapter 5

Martin, M., and Govaerts, B. (2020). LiMM-PCA: Combining ASCA$^+$ and linear mixed models to analyse high-dimensional designed data. *Journal of Chemometrics*, *34*(6), e3232.

R packages This work contributed to the creation of the R package **LMWiRe** for ASCA$^+$-related methodology and graphical outputs that is currently under active development on GitHub: https://github.com/bgovaerts/LMWiRe

General conclusions and perspectives

- **Field of metabolomics**
  - Recent research field
    Very informative generated data but heavy data (pre-)processing
  - Technology- and data-driven science → statistical and bioinformatics tools have to adapt
- **Doctoral research** = 4 different contributions to the field of metabolomics
  - Pre-processing: R package for ^{1}H NMR pre-processing published on Bioconductor
    Data analysis (can be generalised to datasets from other research fields)
  - Comparison of classic PLS approaches for biomarker discovery
  - Extension of ASCA^{+} to other global methods and to LMM
- **Perspectives**
  - **PepsNMR R package:** Continuous improvement of the package
    - Fix bugs, update codes, etc. ; Test and include alternative methods Read other raw data formats; Develop a strategy based on DOE methodology to assess the quality of the pre-processing steps for methods and parameters selection
  - **Feature selection**
    - Combine different FS methods
    - Take into account the correlation structure of the features eg: use sparse Group LASSO
    - Bagging principle applied in the features space
  - **ASCA^{+} and LiMM-PCA**
    - Theoretical extensions to include continuous coefficients in the model
    - Deal with data from longitudinal studies

Thank you for your attention!

- Thanks to the RFMF thesis prize and conference organising committees and to all the collaborators of my PhD thesis.
- Link to the slides of this presentation and the thesis codes: https://github.com/ManonMartin/thesisMaterial
- Any questions?

Manon Martin

Computational biology and bioinformatics (CBIO)

de Duve Institute - UCLouvain

manon.martin@uclouvain.be

References

- Boccard, J. and Rudaz, S. (2016). Exploring omics data from designed experiments using analysis of variance multiblock orthogonal partial least squares. *Analytica Chimica Acta*, 920:18 – 28.
- Bouveresse, D. J.-R., Pinto, R. C., Schmidtke, L., Locquet, N., and Rutledge, D. N. (2011). Identification of significant factors by an extension of anova–pca based on multi-block analysis. *Chemometrics and Intelligent Laboratory Systems*, 106(2):173–182.
- Chun, H. and Keles , S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25.
- Eilers, P. H. (2004). Parametric time warping. *Analytical Chemistry*, 76(2):404–411.
- Eilers, P. H. and Boelens, H. F. (2005). Baseline correction with asymmetric least squares smoothing. leiden university medical centre report.
- Engel, J., Gerretzen, J., Szymanska, E., Jansen, J. J., Downey, G., Blanchet, L., and Buydens, L. M. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, 50:96–106.
- Féraud, B., Govaerts, B., Verleysen, M., and De Tullio, P. (2015). Statistical treatment of 2d nmr cosy spectra in metabolomics: data preparation, clustering-based evaluation of the metabolomic informative content and comparison with 1h-nmr. *Metabolomics*, 11(6):1756–1768.
- Féraud, B., Munaut, C., Martin, M., Verleysen, M., and Govaerts, B. (2017). Combining strong sparsity and competitive predictive power with the l-sopls approach for biomarker discovery in metabolomics. *Metabolomics*, 13(11):130.
- Guitton Y., Tremblay-Franco M., Le Corguillé G., Martin J.F., Pétéra M., Roger-Mele P., Delabrière A., Goulitquer S., Monsoor M., Duperier C., Canlet C., Servien R., Tardivel P., Caron C., Giacomoni F., Thévenot E.A. (2017). Create, run, share, publish, and reference your LC–MS, FIA–MS, GC–MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics, *The International Journal of Biochemistry & Cell Biology*, ISSN 1357-2725, http://dx.doi.org/10.1016/j.biocel.2017.07.002.
- Jansen, J. J., Bro, R., Hoefsloot, H. C. J., van den Berg, F. W. J., Westerhuis, J. A., and Smilde, A. K. (2008). Parafasca: Asca combined with parafac for the analysis of metabolic fingerprinting data. *Journal of Chemometrics*, 22(2):114–121.
- Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.
- Trygg, J. and Wold, S. (2002). Orthogonal projections to latent structures (o-pls). *Journal of Chemometrics*, 16(3):119–128.
Van Nederkassel, A., Daszykowski, M., Eilers, P., and Vander Heyden, Y. (2006). A comparison of three algorithms for chromatograms alignment. *Journal of*
- *Chromatography A*, 1118(2):199–210.
- Vanwinsberghe, J. (2005). Bubble: development of a matlab tool for automated 1h-nmr data processing in metabonomics. Varmuza, K. and Filzmoser, P. (2016). *Introduction to multivariate statistical analysis in chemometrics*. CRC press.