Combining ASCA and mixed models to analyse *high dimensional designed data*

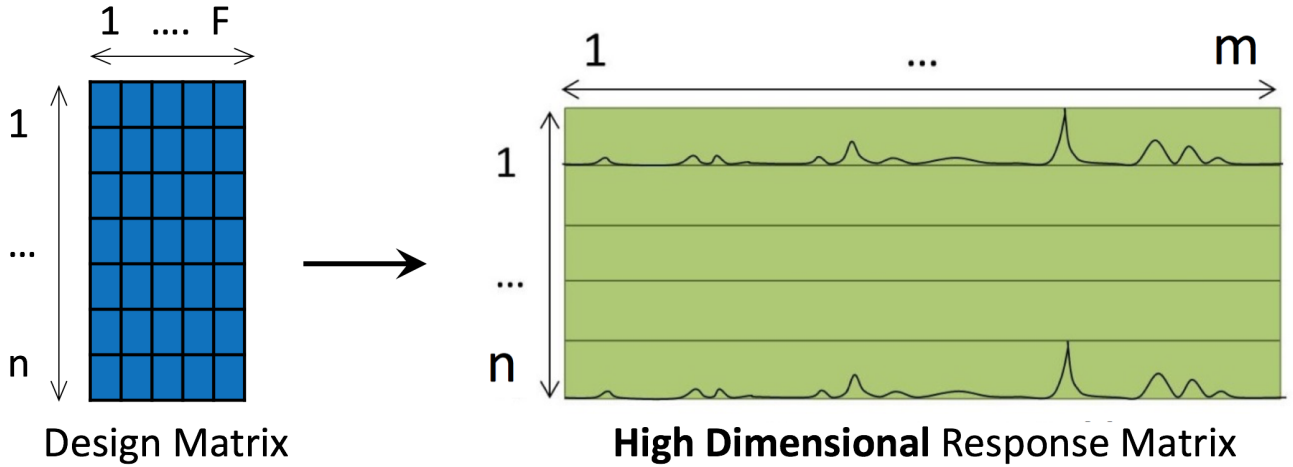
M. Martin [1] P. De Tullio [2] B. Govaerts [1]

[1] Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université catholique de Louvain (UCL), Belgium

[2] Laboratoire de Chimie Pharmaceutique, Université de Liège (ULg), Liège 1, Belgium

January 30, 2018

High Dimensional Experimental Design data



Challenging data analysis

Application to life science data:

- From -omics sciences: genomics, transcriptomics, metabolomics
- High Dimensional: multivariate; often m variables > n samples
- Biological variability, instrumental noise/artifacts
- Multicollinearity between variables

Nontrivial Design Of Experiments (DOE):

- ex.: longitudinal, multi-centre, cross-over studies, etc.
- Presence of random factors (day, lab variation, ...)

Table of content

- High Dimensional designed data analysis
- ASCA methodology
- Extension to mixed models: description & application

Methods needed to analyse HD designed data

Multivariate projection methods

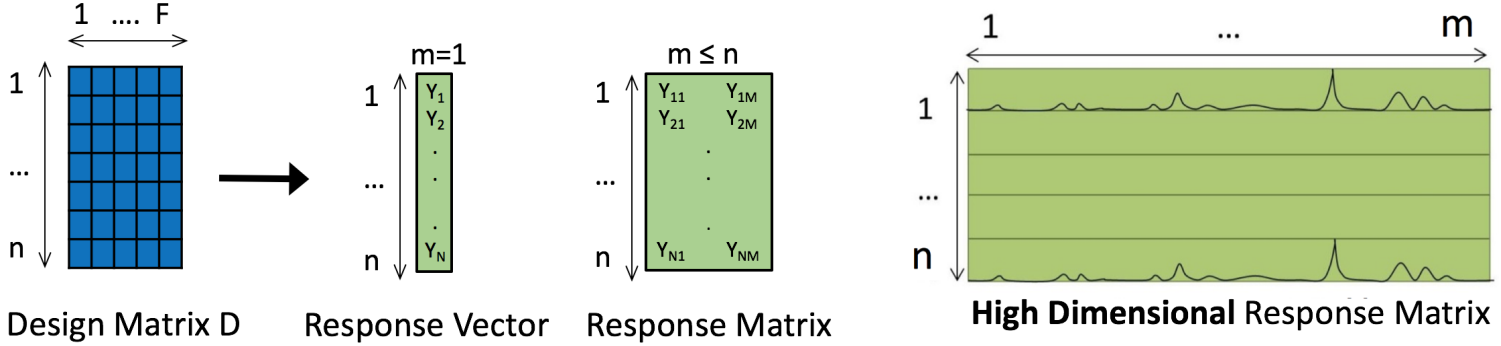
- PCA, ICA, (O)PLS, ...
- Disadvantages:
  - Simple exp. designs (*e.g.* 2 groups comparison)
  - Few statistical modelling and tests

Statistical regression methods

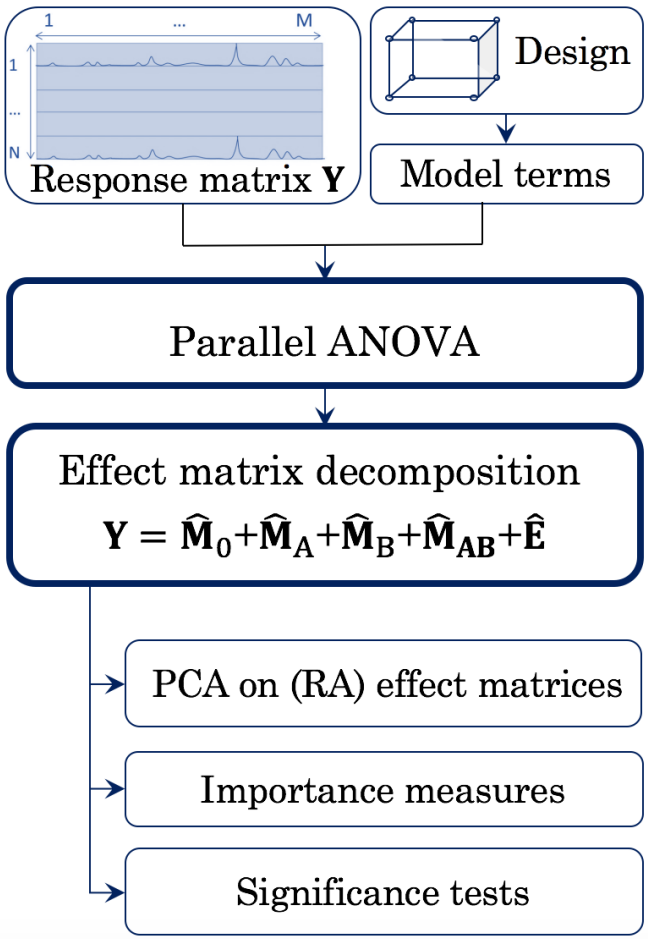
Depends on the **response dimension**:

- $\boldsymbol{y}_{(1\times m)}$: linear or logistic regression, ANOVA, mixed models
- $\boldsymbol{Y}_{(n\times m)}$ with $m < n$: MANOVA, multivariate-GLM

But often $m \ggg n \Rightarrow$ need to combine dimension reduction and statistical modelling.



ASCA methodology



$\Rightarrow$ Combine the dimension reduction methods and statistical modelling [Jansen et al., 2005]:

ANOVA-Simultaneous Components Analysis (ASCA)

Example: **crossed ANOVA II model**:

$$y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

STEP 1: **Parallel ANOVA decomposition** of $\boldsymbol{Y}$

$\boldsymbol{Y}_{(n\times m)}$ is decomposed into **effect matrices**:

$$\boldsymbol{Y} = \hat{\boldsymbol{M}}_0 + \hat{\boldsymbol{M}}_A + \hat{\boldsymbol{M}}_B + \hat{\boldsymbol{M}}_{AB} + \hat{\boldsymbol{E}}$$

STEP 2: (Residual-Augmented) **effect matrices visualisation**

ASCA: $\boldsymbol{M}_f = \boldsymbol{T}_f \boldsymbol{P}_f'$ APCA: $\boldsymbol{M}_f + \boldsymbol{E} = \tilde{\boldsymbol{T}}_f \tilde{\boldsymbol{P}}_f'$

STEP 3: **Effect importance** quantification based on Frobenius norms $||\boldsymbol{M}_f||^2$

STEP 4: Global measure of **effect significance** based on permutation tests

Limitations with classic ASCA and new developments

1. **Balanced** ANOVA designs with **fixed factors**

   $\Rightarrow$ Generalisation to unbalanced data with fixed and random factors

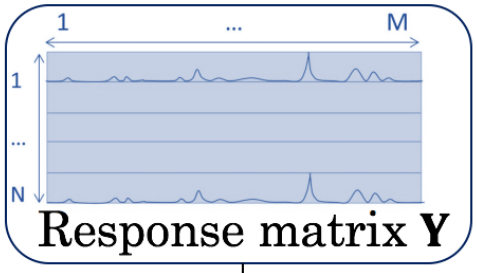
2. // ANOVA does not take into account the correlation between the responses

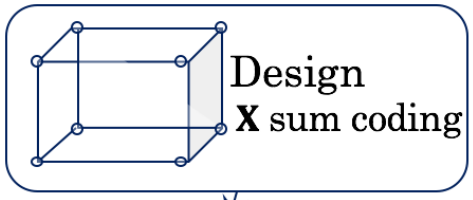
   $\Rightarrow$ Prior PCA reduction & back-transformations

3. **Permutation tests** implementation is challenging for advanced DOE [Anderson and Braak, 2003]

   $\Rightarrow$ Use of alternative test strategies: Likelihood ratio tests & bootstrap

APCA+ for unbalanced designs (Thiel et al. [2017], Guisset et al. [Submitted])





Model terms $\boldsymbol{\theta}$

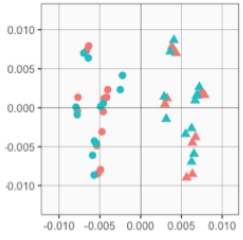
General Linear Model

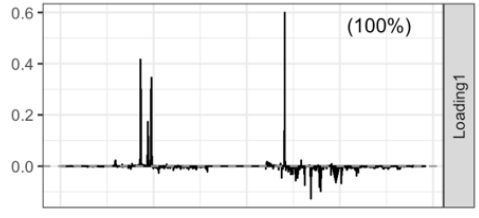
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{E}$$

Effect matrix decomposition

$$\mathbf{Y} = \hat{\mathbf{M}}_0 + \hat{\mathbf{M}}_A + \hat{\mathbf{M}}_B + \hat{\mathbf{M}}_{AB} + \hat{\mathbf{E}}$$

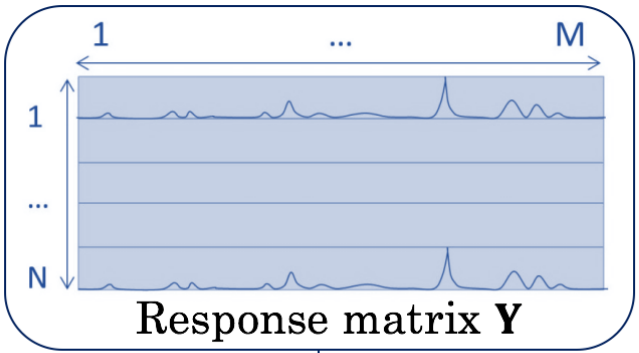
- PCA on (RA) effect matrices
- Importance measures
- Significance tests

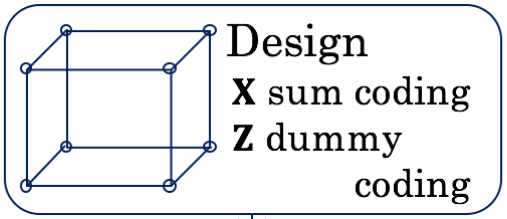




	% Var	P-value
Factor A	18.1	0.000
Factor B	12.1	0.004
Interaction AB	1.0	0.950
Error	53.9	

Mixed Models for High Dimensional Designed Data (MiMoHD^3)





Legend:

Extensions

Model terms $\boldsymbol{\theta}$; $\boldsymbol{\Gamma}$

PCA dimension reduction

$\mathbf{Y} = \mathbf{T_C P_C}' + \mathbf{F}$

Parallel linear mixed models

$\mathbf{T_C} = \mathbf{X\theta} + \mathbf{Z\Gamma} + \mathbf{E}$

Effect matrix decomposition

$\hat{\mathbf{T}}_\mathbf{C} = \hat{\mathbf{M}}_{f1} + \hat{\mathbf{M}}_{f2} + \ldots + \hat{\mathbf{M}}_{r1} + \hat{\mathbf{M}}_{r2} + \cdots + \hat{\mathbf{E}}$

Back transformation → PCA on (RA) effect matrices

Importance measures

Significance measures

The Metabiose repeatability datasets

Context:

- Urine and Serum ^{1}H-NMR spectral data from control and endometriosis patients
- Statistical perspective of spectral reproducibility/repeatability and quality control: not yet well studied in metabolomics
- Unbalanced design for the Serum dataset

3 factors:

- 1 fixed: **groups** (**G**; 2)
- 2 random: **patients** (**P**; 7/group) & **repetitions** over weeks (**W**; 3)

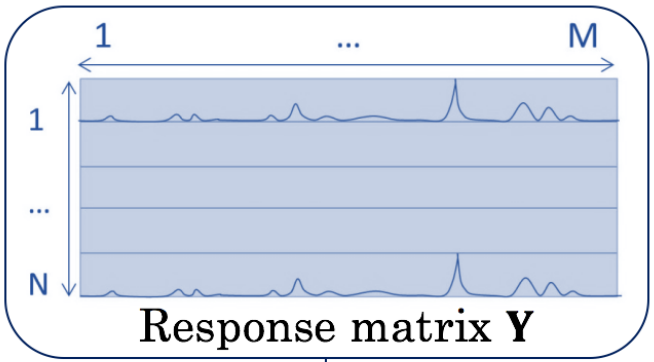
Main research questions:

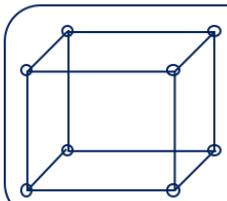
- Compare the groups
- Quantify the variability of the repetitions and the patients
- Test the significance of these random/fixed effects

Experimental design

		W1	W2	W3
Endometriosis	P1			
	P2			
	...	...	...	...
Control	P8			
	P9			
	...	...	...	...

Step 0: Prior PCA dimension reduction (1)





Design
X sum coding
Z dummy coding

Model terms $\boldsymbol{\theta}$; $\boldsymbol{\Gamma}$

PCA dimension reduction
$\mathbf{Y} = \mathbf{T_C P_C}' + \mathbf{F}$

Parallel linear mixed models
$\mathbf{T_C} = \mathbf{X\theta} + \mathbf{Z\Gamma} + \mathbf{E}$

Effect matrix decomposition
$\hat{\mathbf{T}}_\mathbf{C} = \hat{\mathbf{M}}_{f1} + \hat{\mathbf{M}}_{f2} + \ldots + \hat{\mathbf{M}}_{r1} + \hat{\mathbf{M}}_{r2} + \cdots + \hat{\mathbf{E}}$

Back transformation → PCA on (RA) effect matrices

Importance measures

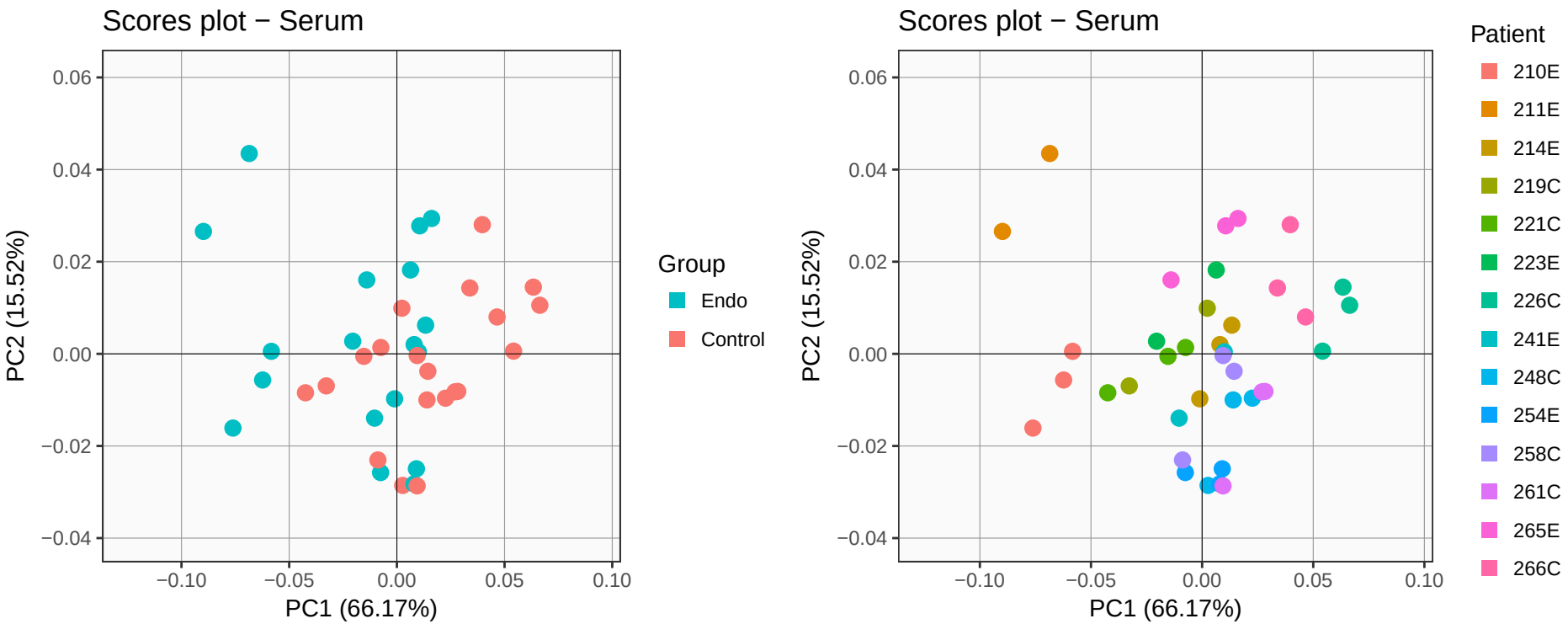
Significance measures

Step 0: Prior PCA dimension reduction (2)

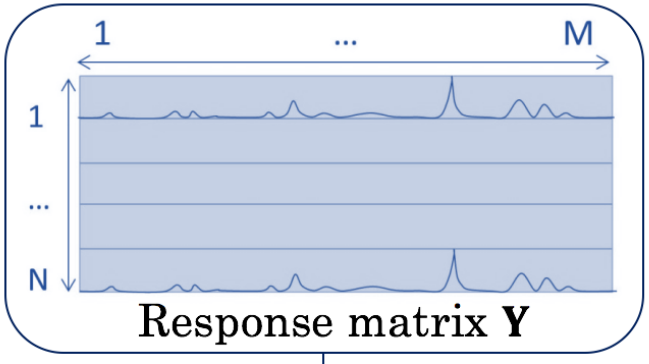
Transform the highly correlated response matrix into a reduced number of orthogonal components without information loss

PCA dimension reduction on the response matrix $\boldsymbol{Y} = \boldsymbol{T}_C \boldsymbol{P}_C' + \boldsymbol{F}$

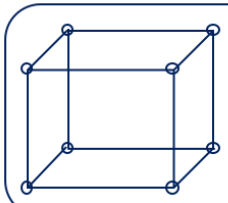
Keep $C = 11$ first Principal Components (PC) with $\sum_C \text{var}(\text{PC}) \geq 99\%$



Step 1: Parallel mixed models on $T_{n \times C}$



Response matrix **Y**



Design
X sum coding
Z dummy coding

Model terms **θ** ; **Γ**

PCA dimension reduction
$\mathbf{Y} = \mathbf{T_C P_C}' + \mathbf{F}$

Parallel linear mixed models
$\mathbf{T_C} = \mathbf{X\theta} + \mathbf{Z\Gamma} + \mathbf{E}$

Effect matrix decomposition
$\hat{\mathbf{T}}_\mathbf{C} = \hat{\mathbf{M}}_{f1} + \hat{\mathbf{M}}_{f2} + \ldots + \hat{\mathbf{M}}_{r1} + \hat{\mathbf{M}}_{r2} + \cdots + \hat{\mathbf{E}}$

Back transformation → PCA on (RA) effect matrices

Importance measures

Significance measures

General framework for mixed models

Fixed + **random** factors *vs* linear regression (only 1 source of random variation)

The **mixed model** for one response t_c can be written as:

$$t_c = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\epsilon}$$

$$= \boldsymbol{X_G\beta_G} + \boldsymbol{Z_P\gamma_P} + \boldsymbol{Z_R\gamma_R} + \boldsymbol{\epsilon}$$

Random Repetition effect

Fixed Group effect

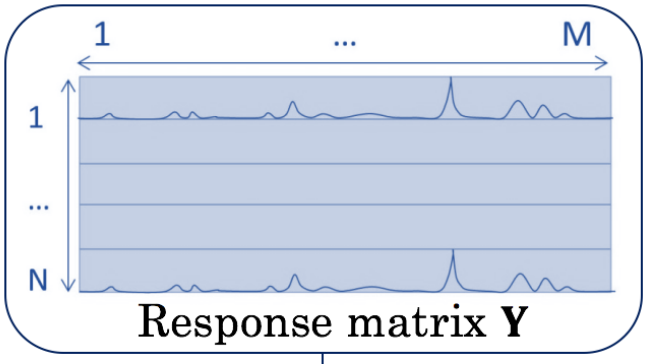
Random Patient effect

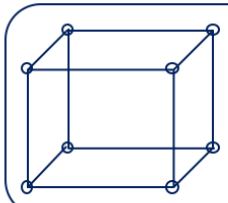
Random residuals

Variance components $= \sigma_P^2 + \sigma_R^2 + \sigma^2$

- Drop the hypothesis of independence between the samples
  $\Rightarrow$ model advanced designs
- Coding: Sum coding for fixed and dummy coding for random effects
- Typical applications: Multicentre study, Multilevel data, Repeated data, etc.
- Parameters are estimated with the Restricted Maximum Likelihood (REML) method

Step 2: PCA decomposition of (RA) effect matrices





Model terms $\boldsymbol{\theta}$; $\boldsymbol{\Gamma}$

PCA dimension reduction

$$\mathbf{Y} = \mathbf{T_C P_C}' + \mathbf{F}$$

Parallel linear mixed models

$$\mathbf{T_C} = \mathbf{X\theta} + \mathbf{Z\Gamma} + \mathbf{E}$$

Effect matrix decomposition

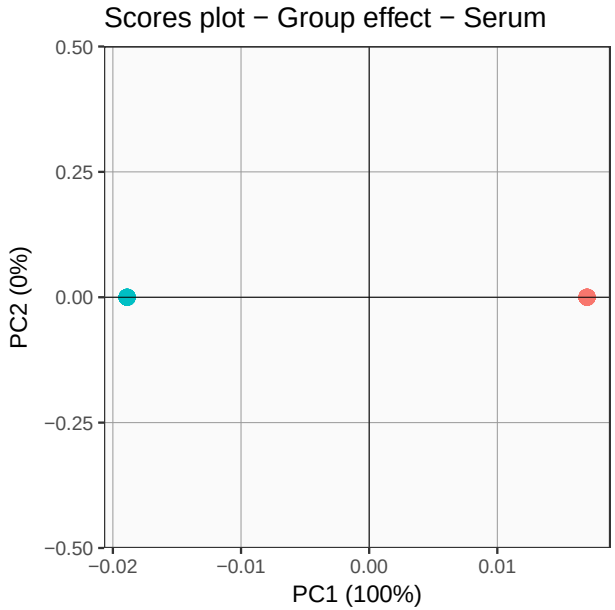
$$\hat{\mathbf{T}}_\mathbf{C} = \hat{\mathbf{M}}_{f1} + \hat{\mathbf{M}}_{f2} + \ldots + \hat{\mathbf{M}}_{r1} + \hat{\mathbf{M}}_{r2} + \cdots + \hat{\mathbf{E}}$$

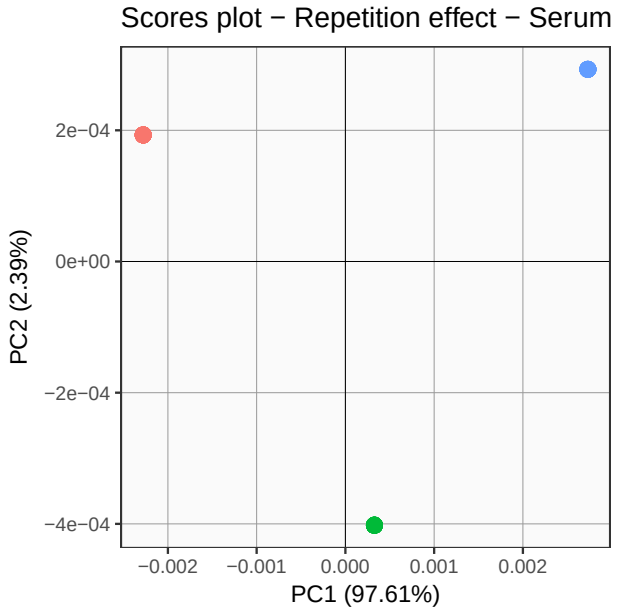
Back transformation → PCA on (RA) effect matrices

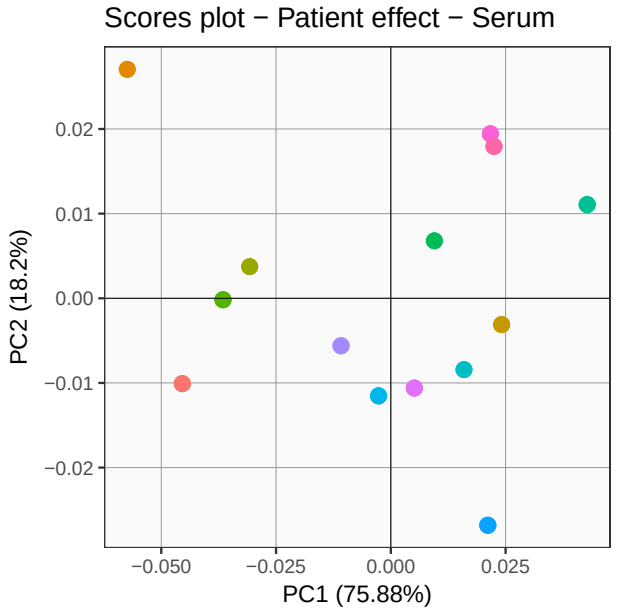
Importance measures

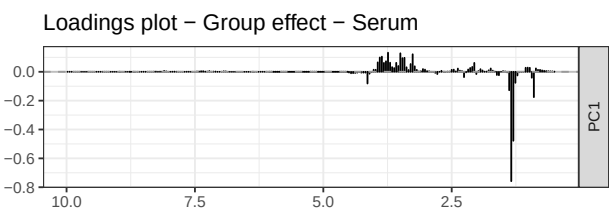
Significance measures

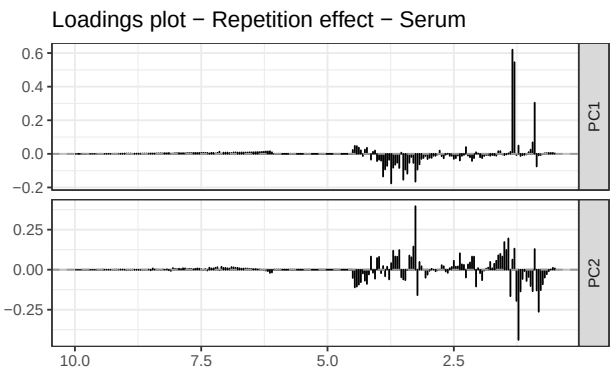
PCA on fixed/random pure effect matrices

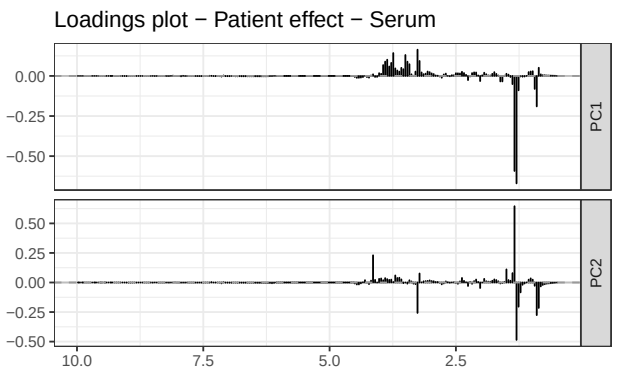












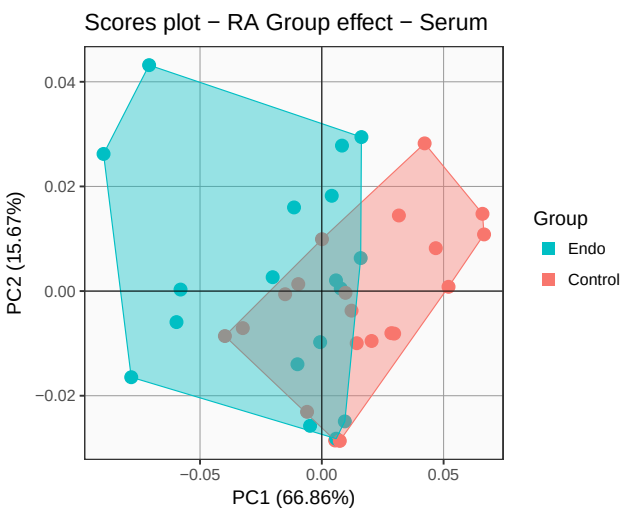
PCA on fixed/random Residual-Augmented effect matrices

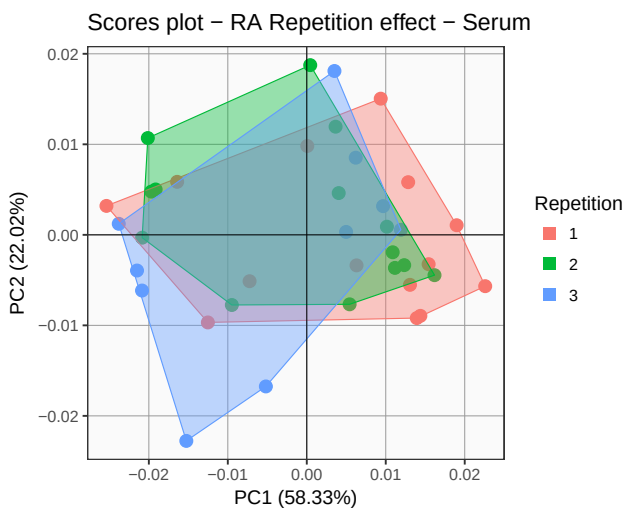
In APCA: $\boldsymbol{E}$ added to $\boldsymbol{M}_f$

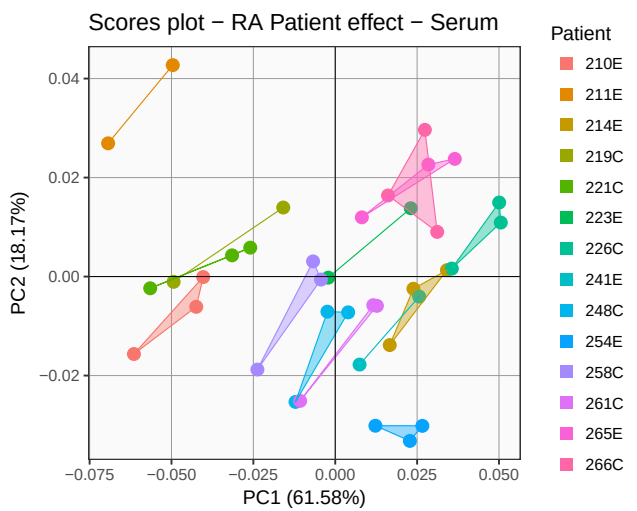
But which variance components should be added for mixed models?

Solution: RA effect matrices based on the ANOVA F-tests (Expected Mean Squares ratio)

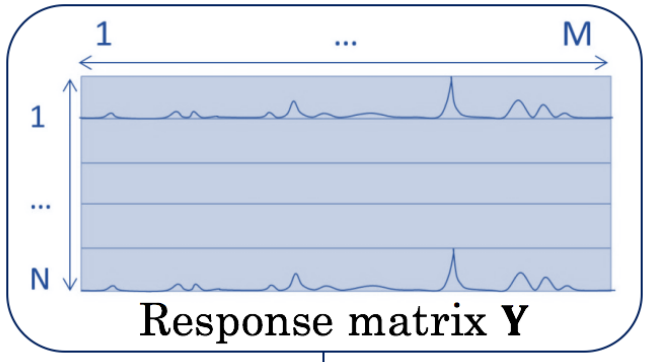
$$\tilde{\boldsymbol{M}}_G = \boldsymbol{M}_G + \boldsymbol{M}_P + \boldsymbol{E}$$
$$\tilde{\boldsymbol{M}}_P = \boldsymbol{M}_P + \boldsymbol{E}$$
$$\tilde{\boldsymbol{M}}_R = \boldsymbol{M}_R + \boldsymbol{E}$$

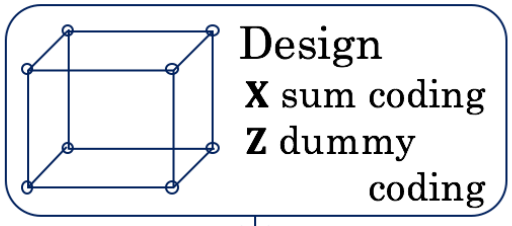






Steps 3 & 4: Global measure of factor importance/significance





Model terms $\boldsymbol{\theta}$; $\boldsymbol{\Gamma}$

PCA dimension reduction

$$\mathbf{Y} = \mathbf{T_C P_C}' + \mathbf{F}$$

Parallel linear mixed models

$$\mathbf{T_C} = \mathbf{X\theta} + \mathbf{Z\Gamma} + \mathbf{E}$$

Effect matrix decomposition

$$\hat{\mathbf{T}}_\mathbf{C} = \hat{\mathbf{M}}_{f1} + \hat{\mathbf{M}}_{f2} + \ldots + \hat{\mathbf{M}}_{r1} + \hat{\mathbf{M}}_{r2} + \cdots + \hat{\mathbf{E}}$$

Back transformation → PCA on (RA) effect matrices

Importance measures

Significance measures

Step 3: Quantification of effect importance

For each response t_c, $c = 1, .., C$:

Random effects: Variance components

$$\hat{\sigma}^2_{P,c}; \hat{\sigma}^2_{R,c}; \hat{\sigma}^2_c$$

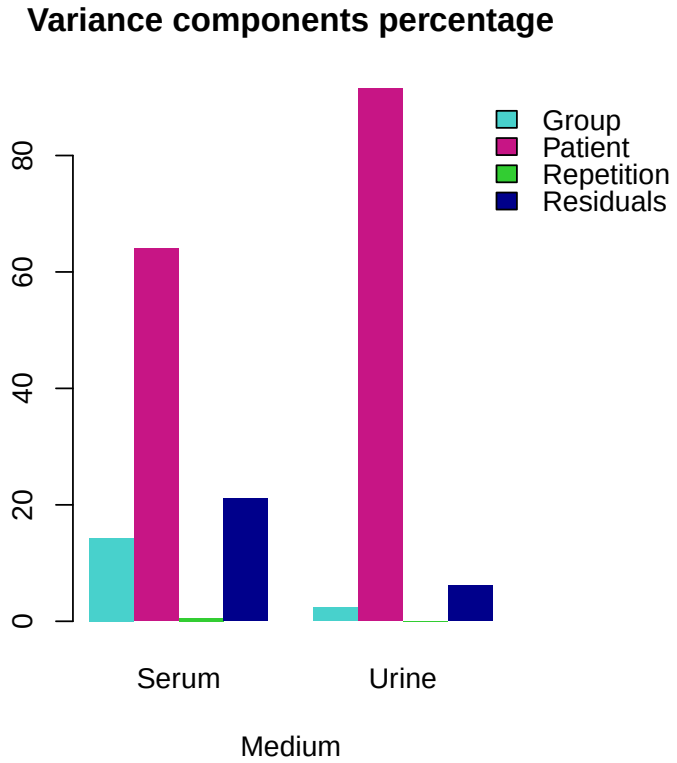
Fixed effects [Nakagawa and Schielzeth, 2013]:

$$\hat{\sigma}^2_{G,c} = \text{var}(\hat{\beta}_{G,c} x_G)$$

For all t_c responses:

Total variance:

$$\hat{\sigma}^2_{tot} = \sum_{c=1}^{C} (\hat{\sigma}^2_{G,c} + \hat{\sigma}^2_{P,c} + \hat{\sigma}^2_{R,c} + \hat{\sigma}^2_c)$$



	Serum	Urine
Group	14.32	2.4
Patient	64.03	91.48
Repetition	0.6	0
Residuals	21.05	6.12

Effect importance percentage

Step 4: Global measure of effect significance

- **Likelihood Ratio Test (LRT)**
  Test the significance of a fixed/random effect in a (mixed) linear model
  Compare the likelihoods L between nested models

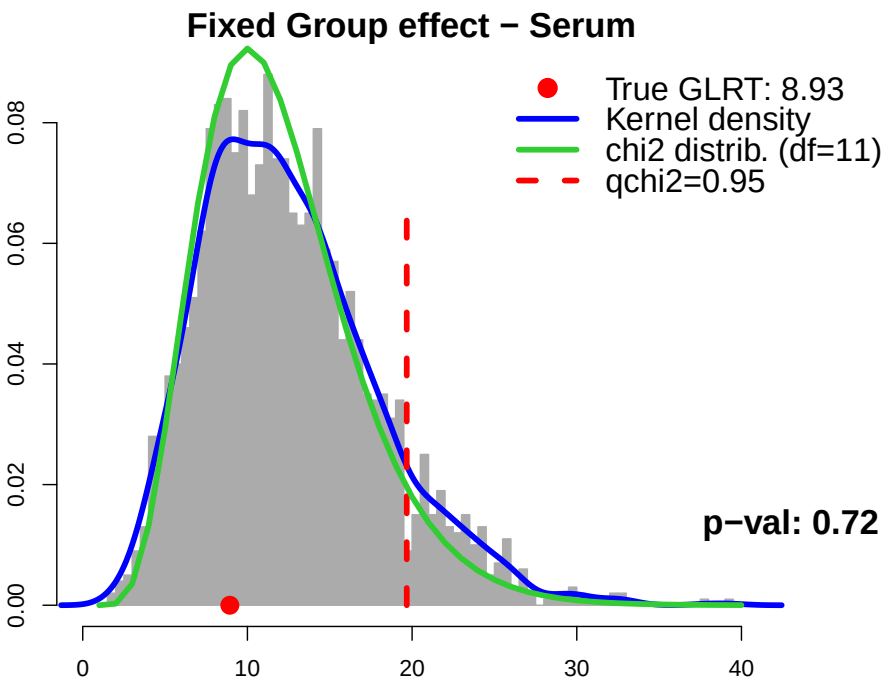
$$\text{LRT statistic: } 2[\log(L_{full}) - \log(L_{null})] \sim_{H_0} \chi^2_{df}$$

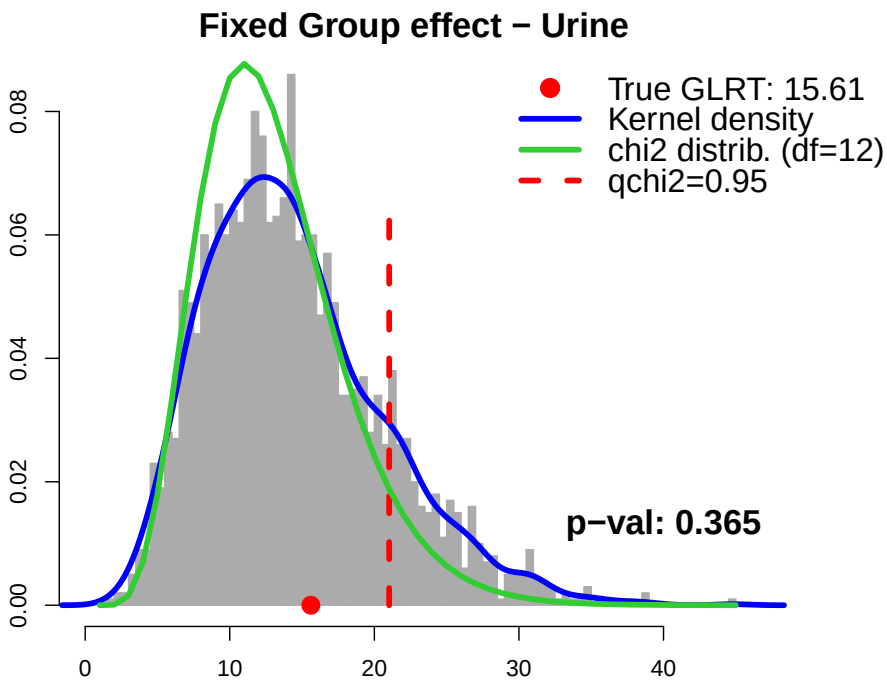
- **a Global LRT (GLRT)**
  The test statistic for a fixed/random effect matrix

$$\text{GLRT statistic: } 2[\sum_{c=1}^{C}(\log(L_{full,c}) - \log(L_{null,c}))] \sim_{H_0} \chi^2_{C\times df}$$

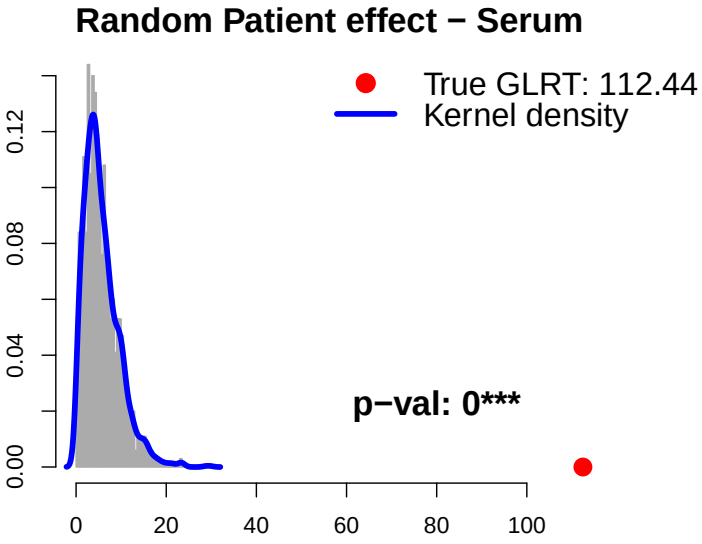
- Assess the significance of an effect based on:
  - The χ^2 distribution with known *df* (fixed effects)
  - **bootstrap simulations** (fixed/random effects)

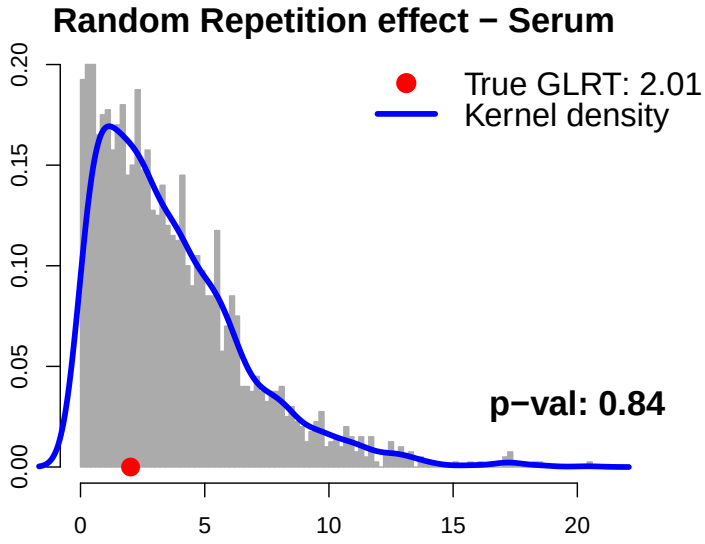
Histograms of bootstrapped GLRT (N_{sim} = 2000)

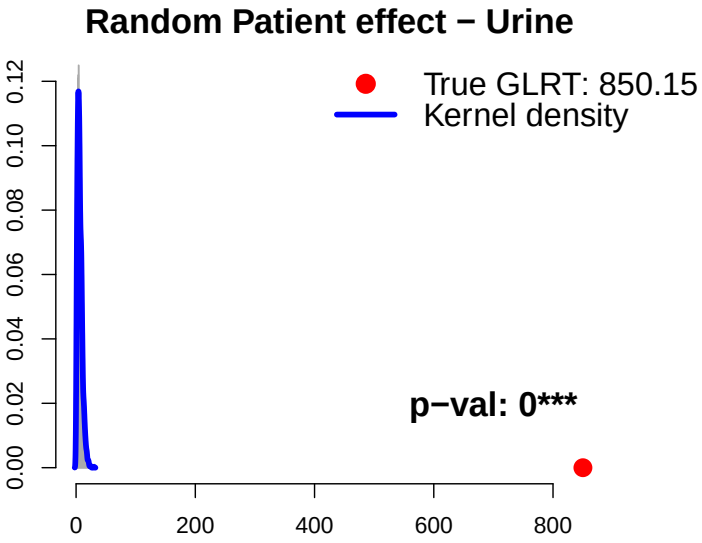


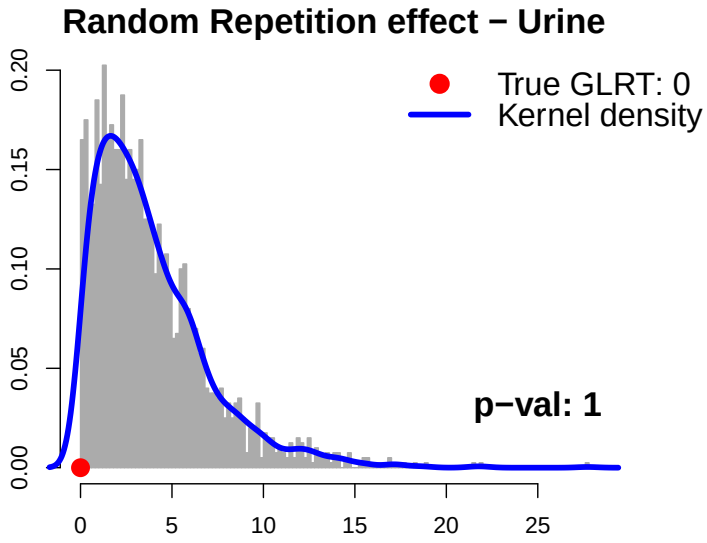


Histograms of boostraped GLRT (N_{sim} = 2000)









In summary

MiMoHD^3, a combination of mixed models & multivariate projection methods

- Innovative extension and generalisation in the ASCA(+) framework
- Enable to model unbalanced designs with random factors
- Take into account the correlation between the response variables
- Global test of effect significance
- Quantification and comparison of the mixed variability sources
- Targeted applications: repeatability/reproductibility study & longitudinal data

Acknowledgments

UCL Université catholique de Louvain

ISBA

RÉGION WALLONNE

LIÈGE université

CIRM

fnrs LA LIBERTÉ DE CHERCHER

References

Marti Anderson and Cajo Ter Braak. Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation*, 73(2):85–113, 2003. doi: 10.1080/00949650215733. URL http://dx.doi.org/10.1080/00949650215733.

Séverine Guisset, Bernadette Govaerts, and Manon Martin. Comparison of parafasca, acomdim, and amopls approaches in the multivariate glm modelling of multi-factorial designs. *Chemometrics and Intelligent Laboratory Systems*, Submitted.

Jeroen J. Jansen, Huub C. J. Hoefsloot, Jan van der Greef, Marieke E. Timmerman, Johan A. Westerhuis, and Age K. Smilde. Asca: analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics*, 19(9):469–481, 2005. ISSN 1099-128X. doi: 10.1002/cem.952. URL http://dx.doi.org/10.1002/cem.952.

Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142, 2013. ISSN 2041-210X. doi: 10.1111/j.2041-210x.2012.00261.x. URL http://dx.doi.org/10.1111/j.2041-210x.2012.00261.x.

Michel Thiel, Baptiste Féraud, and Bernadette Govaerts. Asca+ and apca+: Extensions of asca and apca in the analysis of unbalanced multifactorial designs. *Journal of Chemometrics*, 31(6): e2895–n/a, 2017. ISSN 1099-128X. doi: 10.1002/cem.2895. URL http://dx.doi.org/10.1002/cem.2895. e2895 cem.2895.