Analyzing unbalanced multifactorial experiments with ASCA and APCA: Application in metabolomics

Bernadette Govaerts: UCL
Michel Thiel: Janssen Pharmaceutica

UCL
Université catholique de Louvain

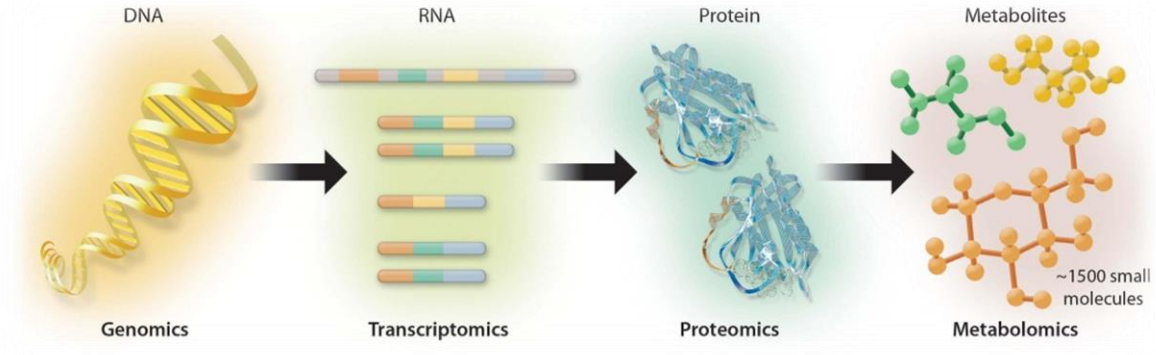
Janssen
PHARMACEUTICAL COMPANIES OF Johnson & Johnson

Context of this project

Master thesis in biostatistics at UCL

Metabolomics

- Study of metabolites
- Newest omics science
- Complex DoE's



ASCA and APCA

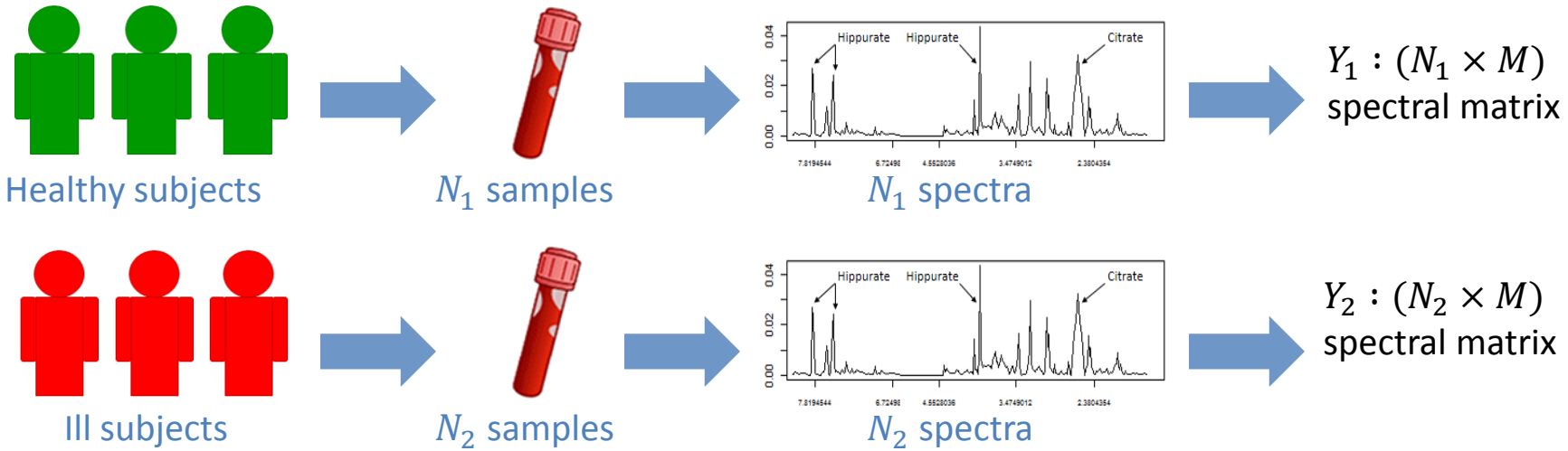
- Combination of ANOVA and PCA
- Analysis of sources of variability on spectral data
- Not well adapted to unbalanced data

Development of two new methods: ASCA+ and APCA+

Table of contents

Typical metabolomic studies

Linking a biological reaction and changes in metabolites



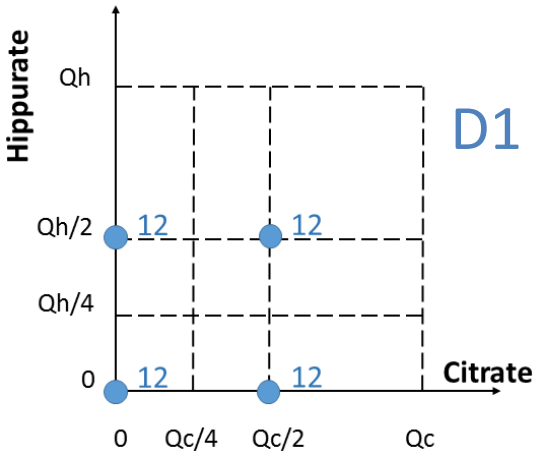
Multivariate databases: *M* variables >>> $N = N_1 + N_2$ observations

Complex designs with multiple factors: subject, time, treatment,…

PCA: doesn't consider the design of experiment

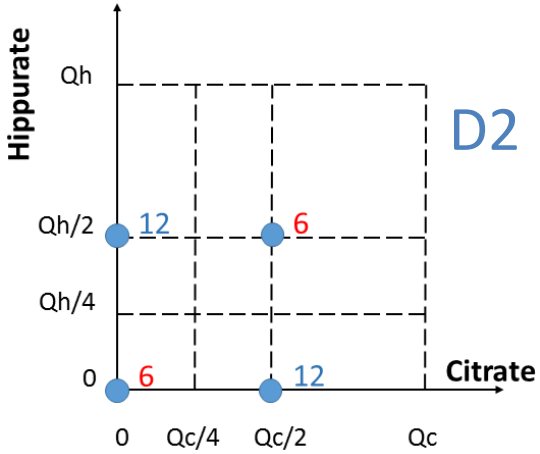
ASCA and APCA: extension of ANOVA in multivariate situations

^{1}H-NMR spectra



D1

balanced



D2

unbalanced

Urine samples spiked with 2 chemicals

Datasets used here:

- D1: balanced (N=48)
- D2: unbalanced (N=36)

Factors of interest

- Hippurate: 2 doses
- Citrate: 2 doses
- Crossed design with 2 factors

Advantages of ASCA+/APCA+ w.r.t. ASCA/APCA will be showed on D2

Table of contents

From ANOVA to ASCA and APCA

ANOVA

Univariate results following a design of experiment

$y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$ ⟶ ANOVA 2 crossed model

ASCA and APCA

Extension of ANOVA to multivariate cases

$\mathbf{Y} = \mathbf{M_0} + \mathbf{M_A} + \mathbf{M_B} + \mathbf{M_{AB}} + \boldsymbol{\varepsilon}$

ASCA (ANOVA-simultaneous component analysis)

Smilde *and al.* (2005): urine samples from guinea pigs with osteoarthritis

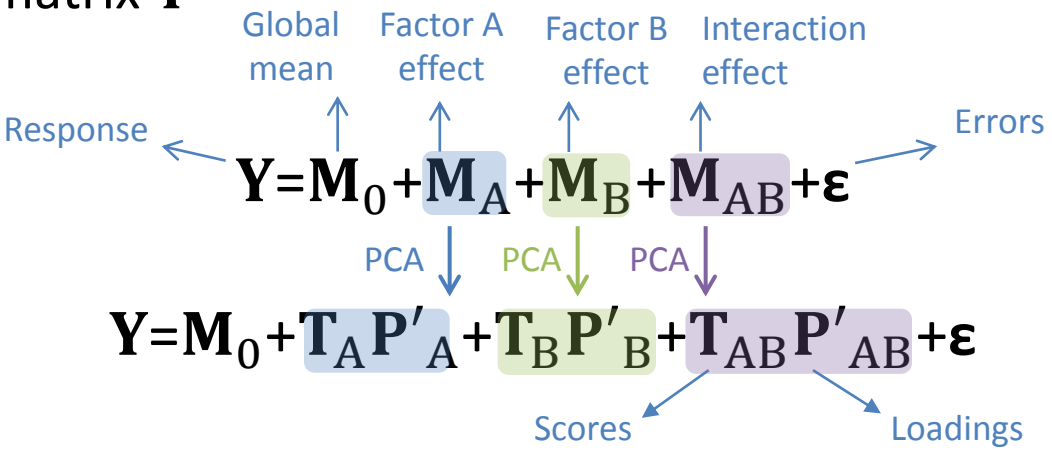
APCA (ANOVA-principal component analysis)

Harrington *and al.* (2004): amniotic fluid samples in premature infants

Steps of ASCA and APCA

1. Decomposition of a spectral matrix $\mathbf{Y}$

   $\mathbf{Y}=\mathbf{M}_0+\mathbf{M}_A+\mathbf{M}_B+\mathbf{M}_{AB}+\boldsymbol{\varepsilon}$



2. PCA on each effect matrix

   ASCA: $\mathbf{M}_A$, $\mathbf{M}_B$, $\mathbf{M}_{AB}$

   APCA: $\mathbf{M}_A+\boldsymbol{\varepsilon}$, $\mathbf{M}_B+\boldsymbol{\varepsilon}$, $\mathbf{M}_{AB}+\boldsymbol{\varepsilon}$

3. Graphical representation

   PCA scores help to visualize separation of spectra for each effect

   PCA loadings help to find biomarkers for each effect

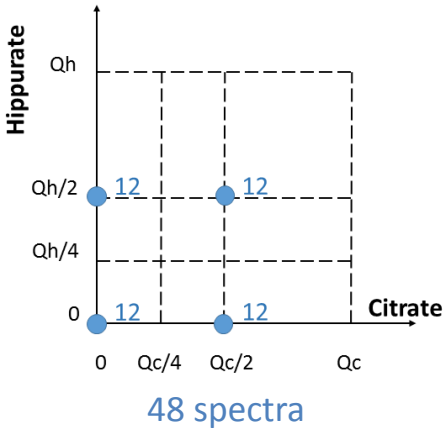
4. Calculate percentages of variance for each component

   $\|\mathbf{Y}-\mathbf{M}_0\|^2=\|\mathbf{M}_A\|^2+\|\mathbf{M}_B\|^2+\|\mathbf{M}_{AB}\|^2+\|\boldsymbol{\varepsilon}\|^2$ (Frobenius norm, balanced)

   $\%A=\frac{\|\mathbf{M}_A\|^2}{\|\mathbf{Y}-\mathbf{M}_0\|^2}\times 100$

5. Apply permutation tests to measure significance of each factor effect

ASCA and APCA results for a balanced design



48 spectra

PCA

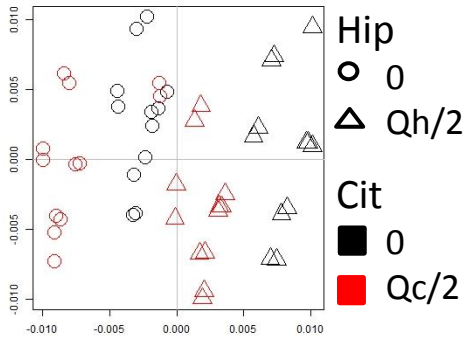
APCA

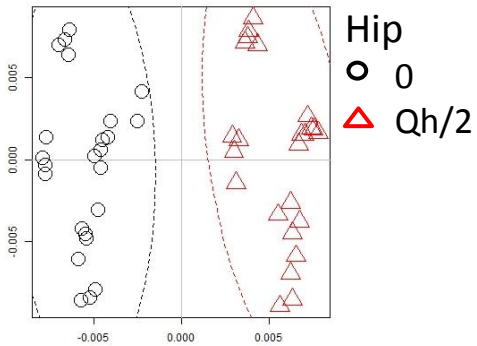
Hip

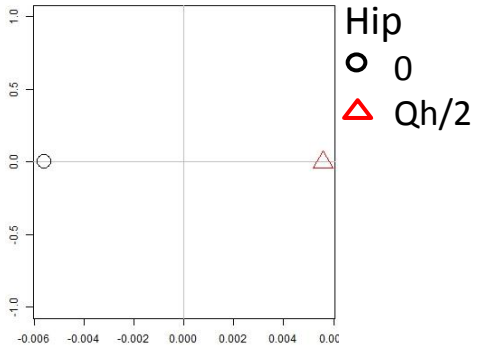
ASCA

Hip

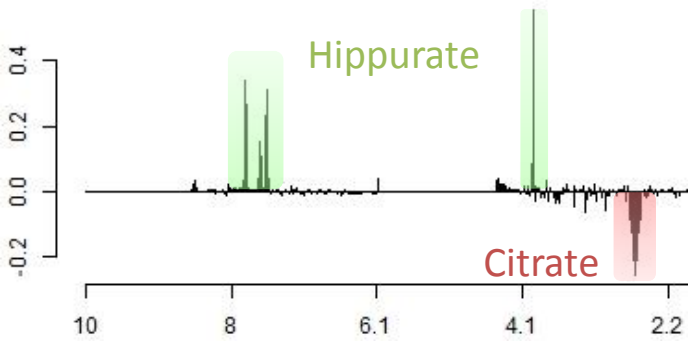
Score plot

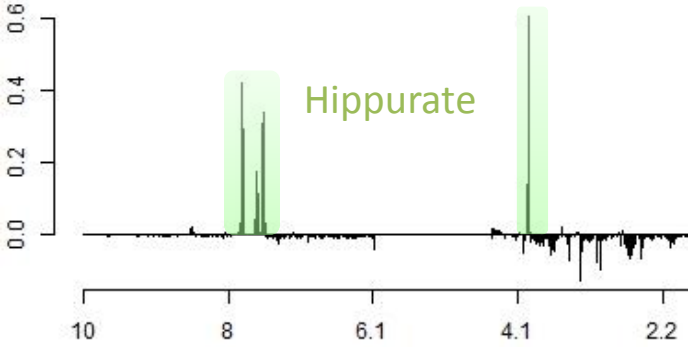






Loading plot, PC1





	% Var	P-value
Hippurate	27.55	0.000
Citrate	18.01	0.000
Interaction	0.51	0.984
Error	53.93	

M_x effect matrices calculation

ANOVA 2 model

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

response — global mean — factor A — factor B — interaction — errors

ANOVA effect decomposition (balanced)

$$y_{ijk} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})$$

$$y_{ijk} = \hat{\mu}_{..} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij} + e_{ijk}$$

Spectral matrix decomposition in ASCA and APCA

$$\mathbf{Y} = \mathbf{M}_0 + \mathbf{M}_A + \mathbf{M}_B + \mathbf{M}_{AB} + \boldsymbol{\varepsilon}$$

ANOVA decomposition applied to each column of the spectral matrix $\mathbf{Y}$

Problems with unbalanced designs:

Factor effect estimators are biased

Effect matrices $\mathbf{M}_A$, $\mathbf{M}_B$ … are not orthogonal → $\mathbf{M}_A \times \mathbf{M}_B \neq \mathbf{0}$

$$\|\mathbf{Y} - \mathbf{M}_0\|^2 \neq \|\mathbf{M}_A\|^2 + \|\mathbf{M}_B\|^2 + \|\mathbf{M}_{AB}\|^2 + \|\boldsymbol{\varepsilon}\|^2$$

Methodology used in ASCA+ and APCA+

Use of general linear models (GLM) to estimate parameters

- Least squares estimators in GLM ≠ simple differences of means in ANOVA
- Unbiased estimators with unbalanced designs
- GLM estimators and ANOVA estimators are identical in balanced designs

Multivariate GLM model:

$$\mathbf{Y}_{N\times M} = \mathbf{X}_{N\times P}\,\boldsymbol{\theta}_{P\times M} + \boldsymbol{\varepsilon}_{N\times M}$$

Spectral matrix — Model matrix — Parameter matrix — Error matrix

Parameters estimation:

$$\widehat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Predicted values:

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\theta}}$$

Errors:

$$\widehat{\mathbf{E}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\theta}}$$

ASCA+ and APCA+: details

1. Effect matrix estimation

Example with a balanced design

6 observations, 2 levels for factor A, 3 levels for factor B and M variables

$$\mathbf{X} = \left(\begin{array}{c|c|cc|cc} \mathbf{X}_0 & \mathbf{X}_A & \mathbf{X}_B & & \mathbf{X}_{AB} & \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{array}\right), \quad \hat{\theta} = \begin{pmatrix} \hat{\mu}_{..1} & \hat{\mu}_{..2} & \cdots & \hat{\mu}_{..M} \\ \hat{\alpha}_{11} & \hat{\alpha}_{12} & \cdots & \hat{\alpha}_{1M} \\ \hat{\beta}_{11} & \hat{\beta}_{12} & \cdots & \hat{\beta}_{1M} \\ \hat{\beta}_{21} & \hat{\beta}_{22} & \cdots & \hat{\beta}_{2M} \\ (\widehat{\alpha\beta})_{111} & (\widehat{\alpha\beta})_{112} & \cdots & (\widehat{\alpha\beta})_{11M} \\ (\widehat{\alpha\beta})_{121} & (\widehat{\alpha\beta})_{122} & \cdots & (\widehat{\alpha\beta})_{12M} \end{pmatrix} \rightarrow \hat{\boldsymbol{\theta}}_{\boldsymbol{A}}$$

$$\Rightarrow \quad \hat{\mathbf{M}}_A = \mathbf{X}_A \times \hat{\boldsymbol{\theta}}_{\boldsymbol{A}}$$

Model matrix ($N \times P$) Parameter matrix ($P \times M$)

$\hat{\mathbf{M}}_0$, $\hat{\mathbf{M}}_B$ et $\hat{\mathbf{M}}_{AB}$ estimated in a same way

$\mathbf{Y} = \hat{\mathbf{M}}_0 + \hat{\mathbf{M}}_A + \hat{\mathbf{M}}_B + \hat{\mathbf{M}}_{AB} + \mathbf{E}$ when the design is balanced

$\mathbf{Y} \neq \hat{\mathbf{M}}_0 + \hat{\mathbf{M}}_A + \hat{\mathbf{M}}_B + \hat{\mathbf{M}}_{AB} + \mathbf{E}$ when the design is unbalanced

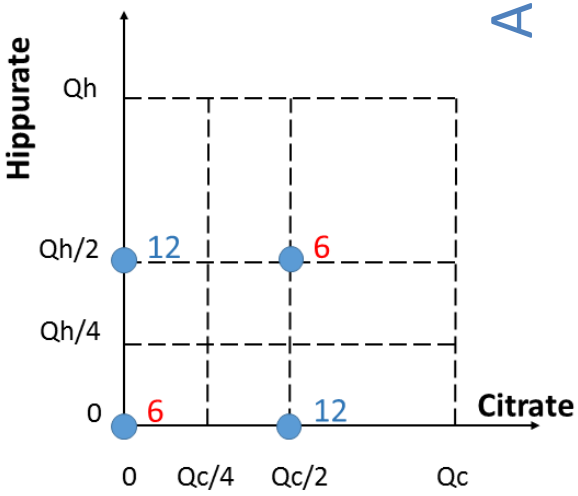
2. Pourcentage of variance calculation (measure of importance)

Type III sum of squares used in GLM:

$$\%A = \frac{\left\|\hat{\mathbf{E}}_{/A}\right\|^2 - \left\|\hat{\mathbf{E}}_{Full}\right\|^2}{\left\|\mathbf{Y} - \hat{\mathbf{M}}_0\right\|^2} \times 100$$

Table of contents

Score plots in ASCA vs ASCA+

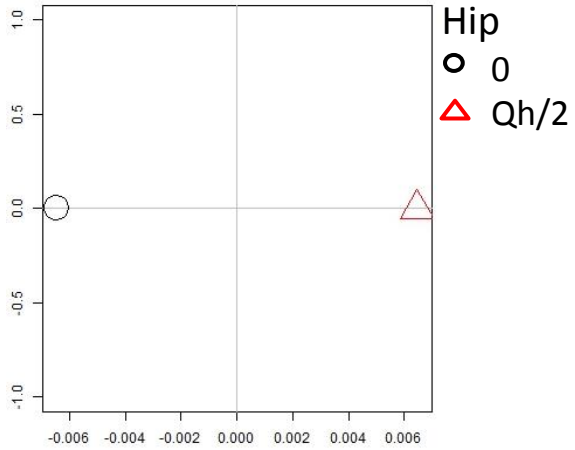


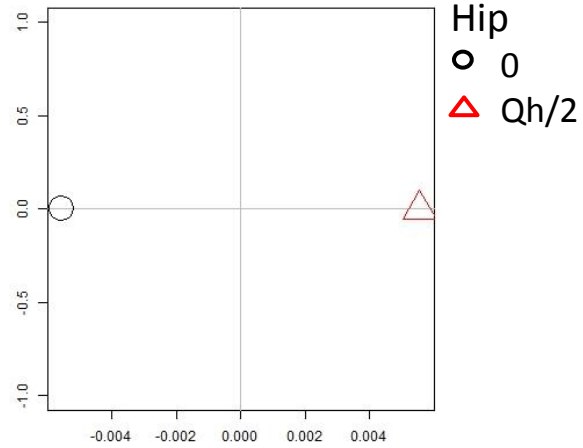
36 spectra

ASCA

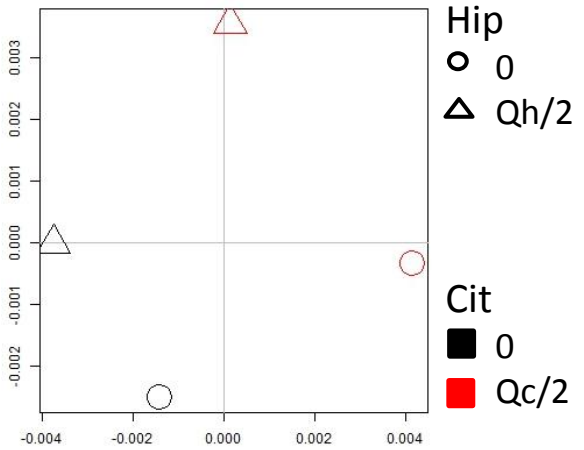
ASCA+

Hippurate

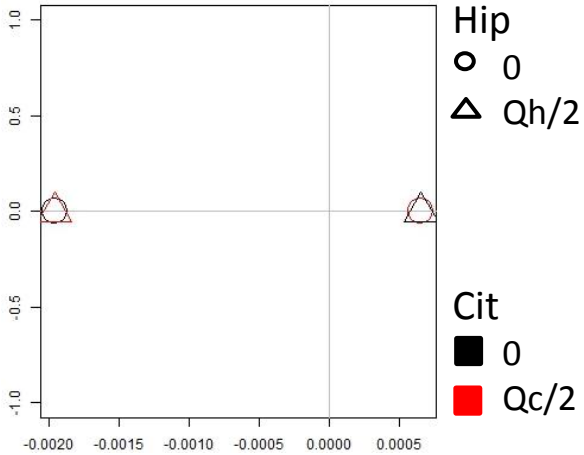




Interaction



Degrees of freedom created!

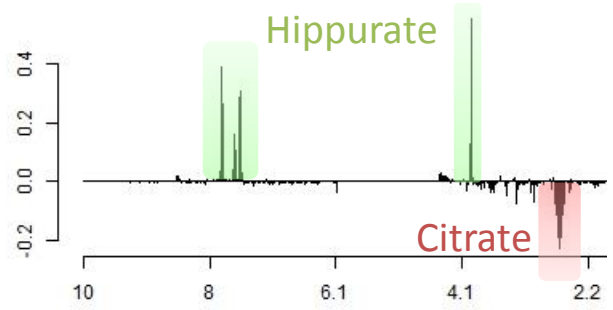


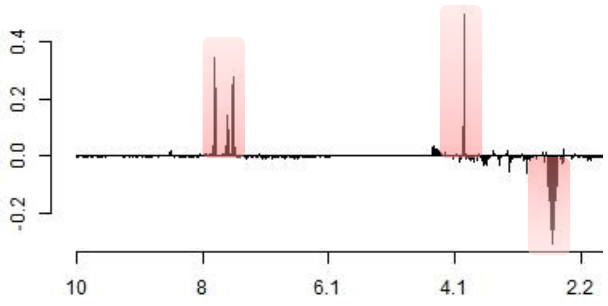
Loading plots in ASCA vs ASCA+

Hippurate

Interaction

ASCA





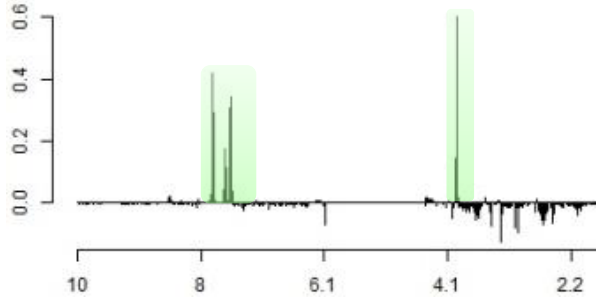
	% Var	P-value
Hippurate	32.7	0.000
Citrate	26.8	0.000
Interaction	12.2	0.016
Error	54.1	
Total	125.8	

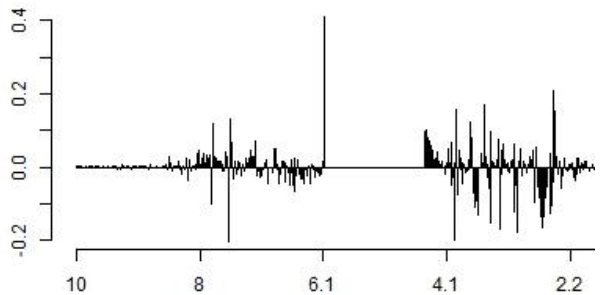
Citrate detected

Hippurate and citrate detected

Interaction significant

ASCA+





	% Var	P-value
Hippurate	18.1	0.000
Citrate	12.1	0.004
Interaction	1.0	0.950
Error	53.9	
Total	85.1	

Only hippurate

Only noise

No interaction

Conclusions and perspectives

ASCA and APCA

- Analysis of multivariate data with DoE's → metabolomics
- Limitation to balanced designs

ASCA+ and APCA+

- General linear models
- Analysis of unbalanced data
- Limitation to fixed factors

Generalize ASCA+ et APCA+ to MIXED linear models

- Fixed and random effects
- Necessary in biological experiments

Any questions?