Combining ASCA and mixed models to analyse high dimensional designed data

M. Martin\textsuperscript{1} P. De Tullio\textsuperscript{2} B. Govaerts\textsuperscript{1}

\textsuperscript{1}Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université catholique de Louvain (UCL), Belgium

\textsuperscript{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:

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

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

⇒ 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 \( Y \)

\( Y_{(n \times m)} \) is decomposed into effect matrices:

\[
Y = \hat{M}_0 + \hat{M}_A + \hat{M}_B + \hat{M}_{AB} + \hat{E}
\]

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

**ASCA**: \( M_f = T_f P'_f \)

**APCA**: \( M_f + E = \tilde{T}_f \tilde{P}'_f \)

**STEP 3**: Effect importance quantification based on Frobenius norms \( ||M_f||^2 \)

**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**
   ⇒ Generalisation to **unbalanced** data with fixed and **random** factors

2. ANOVA does not take into account the **correlation** between the responses
   ⇒ Prior **PCA** reduction & back-transformations

3. **Permutation tests** implementation is challenging for advanced DOE
   [Anderson and Braak, 2003]
   ⇒ Use of alternative test strategies: **Likelihood ratio tests & bootstrap**
APCA+ for unbalanced designs (Thiel et al. [2017], Guisset et al. [Submitted])

Response matrix $Y$

$\text{Design X sum coding}$

Model terms $\theta$

General Linear Model

$Y = X\theta + E$

Effect matrix decomposition

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

PCA on (RA) effect matrices

Importance measures

Significance tests

<table>
<thead>
<tr>
<th></th>
<th>% Var</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor A</td>
<td>18.1</td>
<td>0.000</td>
</tr>
<tr>
<td>Factor B</td>
<td>12.1</td>
<td>0.004</td>
</tr>
<tr>
<td>Interaction AB</td>
<td>1.0</td>
<td>0.950</td>
</tr>
<tr>
<td>Error</td>
<td>53.9</td>
<td></td>
</tr>
</tbody>
</table>
Mixed Models for High Dimensional Designed Data (MiMoHD$^3$)

Response matrix $Y$

PCA dimension reduction

$Y = T_CP_C' + F$

Design

$X$ sum coding

$Z$ dummy coding

Model terms $\theta; \Gamma$

Parallel linear mixed models

$T_C = X\theta + Z\Gamma + E$

Effect matrix decomposition

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

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
Step 0: Prior PCA dimension reduction (1)

Response matrix $Y$

PCA dimension reduction

$Y = T_c P_c' + F$

Design

$X$ sum coding

$Z$ dummy coding

Model terms $\theta ; \Gamma$

Parallel linear mixed models

$T_c = X\theta + Z\Gamma + E$

Effect matrix decomposition

$\hat{T}_c = \hat{M}_{f1} + \hat{M}_{f2} + ... + \hat{M}_{r1} + \hat{M}_{r2} + ... + \hat{E}$

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 $Y = T_C P_C^t + F$

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

![Scores plot – Serum](image1)

![Scores plot – Serum](image2)

![Scores plot – Serum](image3)
Step 1: Parallel mixed models on $T_n \times C$

Response matrix $Y$

PCA dimension reduction $Y = T_c P_c' + F$

Design
- $X$ sum coding
- $Z$ dummy coding

Model terms $\theta ; \Gamma$

Parallel linear mixed models
$T_c = X\theta + Z\Gamma + E$

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

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:

$$t_c = X\beta + Z\gamma + \epsilon$$

- Fixed Group effect
- Random Patient effect
- Random Repetition 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

Response matrix $Y$

$$Y = T_c P_c' + F$$

PCA dimension reduction

Design
- $X$ sum coding
- $Z$ dummy coding

Model terms $\theta ; \Gamma$

Parallel linear mixed models

$$T_c = X\theta + Z\Gamma + E$$

Effect matrix decomposition

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

Back transformation

PCA on (RA) effect matrices

Importance measures

Significance measures
PCA on fixed/random pure effect matrices

Scores plot – Group effect – Serum

Scores plot – Repetition effect – Serum

Scores plot – Patient effect – Serum

Loadings plot – Group effect – Serum

Loadings plot – Repetition effect – Serum

Loadings plot – Patient effect – Serum
PCA on fixed/random Residual-Augmented effect matrices

In APCA: \( E \) added to \( M_f \)

But which variance components should be added for mixed models?

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

\[
\tilde{M}_G = M_G + M_P + E \\
\tilde{M}_P = M_P + E \\
\tilde{M}_R = M_R + E
\]

Scores plot – RA Group effect – Serum

Scores plot – RA Repetition effect – Serum

Scores plot – RA Patient effect – Serum
Steps 3 & 4: Global measure of factor importance/significance

**Response matrix** $\mathbf{Y}$

PCA dimension reduction

$$\mathbf{Y} = \mathbf{T}_c \mathbf{P}_c' + \mathbf{F}$$

**Design**
- $\mathbf{X}$ sum coding
- $\mathbf{Z}$ dummy coding

**Model terms** $\mathbf{\theta} ; \mathbf{\Gamma}$

**Parallel linear mixed models**

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

**Effect matrix decomposition**

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

- **Back transformation**
- **PCA on (RA) effect matrices**
- **Importance measures**
- **Significance measures**
Step 3: Quantification of effect importance

For each response $t_c, c = 1, \ldots, C$:

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:

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

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

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

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

- Assess the significance of an effect based on:
  - The $\chi^2$ distribution with known $df$ (fixed effects)
  - **bootstrap simulations** (fixed/random effects)
Histograms of bootstrapped GLRT ($N_{sim} = 2000$)

**Fixed Group effect – Serum**
- True GLRT: 8.93
- $p$-val: 0.72

**Fixed Group effect – Urine**
- True GLRT: 15.61
- $p$-val: 0.365
Histograms of bootstrapped GLRT ($N_{sim} = 2000$)

Random Patient effect – Serum
- True GLRT: 112.44
- Kernel density
- p-val: 0***

Random Repetition effect – Serum
- True GLRT: 2.01
- Kernel density
- p-val: 0.84

Random Patient effect – Urine
- True GLRT: 850.15
- Kernel density
- p-val: 0***

Random Repetition effect – Urine
- True GLRT: 0
- Kernel density
- p-val: 1
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/reproducibility study & longitudinal data
References


