Multiway-SIR for Longitudinal Multi-table Data Integration

Nathalie Villa-Vialaneix, Valérie Sautron, Marie Chavent & Nathalie Viguerie

nathalie.villa@toulouse.inra.fr

http://www.nathalievilla.org

COMPSTAT - August 25th, 2016

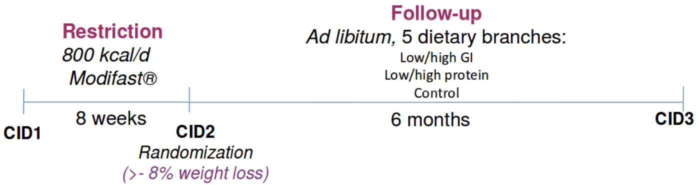
Sommaire

Sommaire

A longitudinal study on weight loss induced by a low calorie diet

Data collected during EU project, 8 centers, 450 families

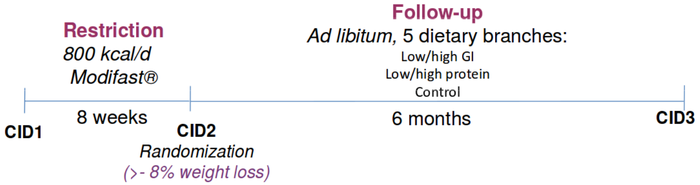
Purpose: effect of glycemic index, protein content... on weight maintenance after a diet for obese people



A longitudinal study on weight loss induced by a low calorie diet

Data collected during EU project, 8 centers, 450 families

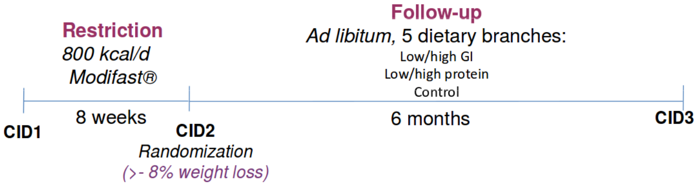
Purpose: effect of glycemic index, protein content... on weight maintenance after a diet for obese people



Data description: at each Clinical Intervention Day (CID), measure of 15 clinical variables (weight, HDL, ...) on 135 obese women

A longitudinal study on weight loss induced by a low calorie diet

Data collected during EU project, 8 centers, 450 families
Purpose: effect of glycemic index, protein content... on weight maintenance after a diet for obese people



Data description: at each Clinical Intervention Day (CID), measure of 15 clinical variables (weight, HDL, ...) on 135 obese women
Targeted problem: exploratory data analysis aimed at explaining the success/failure of the diet (in term of weight loss/regain)

Data description and notations

$\mathbf{X}_{..t}$

$n \times p$-matrix

$(n > p)$

for

$t = 1, \ldots, T$

exploratory variables

(clinical)

$\mathbf{y}$

numeric vector of length n

target variable

(evolution of weight: $\frac{\text{weight}_3 - \text{weight}_1}{\text{weight}_1}$)

Method presented in this talk

Features:

1. longitudinal data integration (with T small; here $T = 3$)

Method presented in this talk

Features:

1. longitudinal data integration (with T small; here $T = 3$)
2. multidimensional data analysis with simple graphical representations (similar to PCA)

Method presented in this talk

Features:

1. longitudinal data integration (with T small; here $T = 3$)
2. multidimensional data analysis with simple graphical representations (similar to PCA)
3. designed to highlight differences in the numeric target variable

Method presented in this talk

Features:

1. longitudinal data integration (with T small; here $T = 3$)
2. multidimensional data analysis with simple graphical representations (similar to PCA)
3. designed to highlight differences in the numeric target variable
4. designed to highlight differences/commonalities between variable structure (rather than individual structure) between the different time steps

Sommaire

Factor analysis methods for multi-tables data integration

Standard methods to analyze data such as $(\mathbf{X}_{..t})_{t=1,\ldots,T}$:

- Multiple Factor Analysis (MFA) [Escofier and Pagès, 2008]
- STATIS and DUAL STATIS [Lavit et al., 1994, Abdi et al., 2012]

Factor analysis methods for multi-tables data integration

Standard methods to analyze data such as $(\mathbf{X}_{..t})_{t=1,\ldots,T}$:

- Multiple Factor Analysis (MFA) [Escofier and Pagès, 2008]: weights for tables $\mathbf{X}_{..t}$ according to a measure of their variability (using first eigenvector of the PCA);
- STATIS and DUAL STATIS [Lavit et al., 1994, Abdi et al., 2012]

Factor analysis methods for multi-tables data integration

Standard methods to analyze data such as $(\mathbf{X}_{..t})_{t=1,\ldots,T}$:

- Multiple Factor Analysis (MFA) [Escofier and Pagès, 2008]: weights for tables $\mathbf{X}_{..t}$ according to a measure of their variability (using first eigenvector of the PCA);
- STATIS and DUAL STATIS [Lavit et al., 1994, Abdi et al., 2012]: larger importance to the most consensual tables ⇒ common representation which is the best overall compromise.

Factor analysis methods for multi-tables data integration

Standard methods to analyze data such as $(\mathbf{X}_{..t})_{t=1,\ldots,T}$:

- Multiple Factor Analysis (MFA) [Escofier and Pagès, 2008]: weights for tables $\mathbf{X}_{..t}$ according to a measure of their variability (using first eigenvector of the PCA);
- STATIS and DUAL STATIS [Lavit et al., 1994, Abdi et al., 2012]: larger importance to the most consensual tables ⇒ common representation which is the best overall compromise. STATIS: compromise according to individuals and DUAL STATIS: compromise according to correlations between variables.

Overall description of DUAL STATIS

Data: $\mathbf{X} = \begin{pmatrix} \mathbf{X}_{..1} \\ \vdots \\ \mathbf{X}_{..T} \end{pmatrix}$

same number of variables, potentially different number of individuals
data are supposed centered and scaled for all t

Overall description of DUAL STATIS

Data: $\mathbf{X} = \begin{pmatrix} \mathbf{X}_{..1} \\ \vdots \\ \mathbf{X}_{..T} \end{pmatrix}$

same number of variables, potentially different number of individuals
data are supposed centered and scaled for all t

$\longrightarrow$

inter-structure analysis
analysis of $\widetilde{\mathbf{C}}$ ($T \times T$ similarity matrix)

Overall description of DUAL STATIS

Data: $\mathbf{X} = \begin{pmatrix} \mathbf{X}_{..1} \\ \vdots \\ \mathbf{X}_{..T} \end{pmatrix}$

same number of variables, potentially different number of individuals
data are supposed centered and scaled for all t

$\longrightarrow$

inter-structure analysis
analysis of $\widetilde{\mathbf{C}}$ ($T \times T$ similarity matrix)

table weights
$(\alpha_1, \ldots, \alpha_T)$

compromise matrix
$\Gamma^c = \sum_{t=1}^T \alpha_t \Gamma_t$ in which Γ_t is the covariance matrix of $\mathbf{X}_{..t}$

Overall description of DUAL STATIS

Data: $\mathbf{X} = \begin{pmatrix} \mathbf{X}_{..1} \\ \vdots \\ \mathbf{X}_{..T} \end{pmatrix}$

same number of variables, potentially different number of individuals
data are supposed centered and scaled for all t

$\longrightarrow$

inter-structure analysis
analysis of $\widetilde{\mathbf{C}}$ ($T \times T$ similarity matrix)

table weights
$(\alpha_1, \ldots, \alpha_T)$

compromise matrix
$\Gamma^c = \sum_{t=1}^T \alpha_t \Gamma_t$ in which Γ_t is the covariance matrix of $\mathbf{X}_{..t}$

GSVD

eig. dec.

compromise analysis
representations of variables and of individuals
compromise positions or time-dependant positions

Inter-structure analysis

Notations: $\forall\, t = 1, \ldots, T$, $\Gamma_t = \frac{1}{n}\mathbf{X}_{..t}^\top \mathbf{X}_{..t}$ and $\widetilde{\Gamma}_t = \frac{\Gamma_t}{\|\Gamma_t\|_F}$

Inter-structure analysis

Notations: $\forall\, t = 1, \ldots, T$, $\Gamma_t = \frac{1}{n}\mathbf{X}_{..t}^\top \mathbf{X}_{..t}$ and $\widetilde{\Gamma}_t = \frac{\Gamma_t}{\|\Gamma_t\|_F}$

Similarity between time steps: $\widetilde{\mathbf{C}} = (c_{tt'})_{t,t'=1,\ldots,T}$ with

$$c_{tt'} = \langle \widetilde{\Gamma}_t, \widetilde{\Gamma}_{t'} \rangle_F$$

(cosinus between Γ_t and $\Gamma_{t'}$).

Inter-structure analysis

Notations: $\forall\, t = 1, \ldots, T$, $\Gamma_t = \frac{1}{n}\mathbf{X}_{..t}^\top \mathbf{X}_{..t}$ and $\widetilde{\Gamma}_t = \frac{\Gamma_t}{\|\Gamma_t\|_F}$

Similarity between time steps: $\widetilde{\mathbf{C}} = (c_{tt'})_{t,t'=1,\ldots,T}$ with

$$c_{tt'} = \langle \widetilde{\Gamma}_t, \widetilde{\Gamma}_{t'} \rangle_F$$

(cosinus between Γ_t and $\Gamma_{t'}$).

Inter-structure analysis: Solve

$$\mathbf{u}_1 = \arg\max_{\mathbf{u}=(u_1,\ldots,u_T)^\top \in \mathbb{R}^T,\ \|\mathbf{u}\|=1} \left\| \sum_{t=1}^{T} u_t \widetilde{\Gamma}_t \right\|_F^2$$

Inter-structure analysis

Notations: $\forall\, t = 1, \ldots, T$, $\Gamma_t = \frac{1}{n}\mathbf{X}_{..t}^\top \mathbf{X}_{..t}$ and $\widetilde{\Gamma}_t = \frac{\Gamma_t}{\|\Gamma_t\|_F}$

Similarity between time steps: $\widetilde{\mathbf{C}} = (c_{tt'})_{t,t'=1,\ldots,T}$ with

$$c_{tt'} = \langle \widetilde{\Gamma}_t, \widetilde{\Gamma}_{t'} \rangle_F$$

(cosinus between Γ_t and $\Gamma_{t'}$).

Inter-structure analysis: Solve

$$\mathbf{u}_1 = \arg\max_{\mathbf{u}=(u_1,\ldots,u_T)^\top \in \mathbb{R}^T,\ \|\mathbf{u}\|=1} \left\| \sum_{t=1}^T u_t \widetilde{\Gamma}_t \right\|_F^2$$

Solution: u_1 is the first eigenvector of $\widetilde{\mathbf{C}}$ $\Rightarrow$ compromise weights:

- $\forall\, t = 1, \ldots, T$, $\alpha_t = \frac{u_{1t}}{\sum_{t'} u_{1t'}}$
- $\Gamma^c = \sum_{t=1}^T \alpha_t \widetilde{\Gamma}_t$

Compromise analysis

GPCA of $(\widetilde{\mathbf{X}}, \mathbb{I}_p, \mathbf{D})$ with $\widetilde{\mathbf{X}}_{..t} = \frac{\mathbf{X}_{..t}}{\sqrt{\|\Gamma_t\|_F}}$ and $\mathbf{D} = \frac{1}{n}\mathrm{Diag}(\alpha_1, \ldots, \alpha_T) \otimes \mathbb{I}_n$

Compromise analysis

GPCA of $(\widetilde{\mathbf{X}}, \mathbb{I}_p, \mathbf{D})$ with $\widetilde{\mathbf{X}}_{..t} = \frac{\mathbf{X}_{..t}}{\sqrt{\|\Gamma_t\|_F}}$ and $\mathbf{D} = \frac{1}{n}\text{Diag}(\alpha_1, \ldots, \alpha_T) \otimes \mathbb{I}_n$

$\Leftrightarrow$ eigendecomposition of Γ^c

Compromise analysis

GPCA of $(\widetilde{\mathbf{X}}, \mathbb{I}_p, \mathbf{D})$ with $\widetilde{\mathbf{X}}_{..t} = \frac{\mathbf{X}_{..t}}{\sqrt{\|\mathbf{\Gamma}_t\|_F}}$ and $\mathbf{D} = \frac{1}{n}\text{Diag}(\alpha_1, \ldots, \alpha_T) \otimes \mathbb{I}_n$

$\Leftrightarrow$ eigendecomposition of Γ^c

Solution writes:

$$\widetilde{\mathbf{X}} = \mathbf{P}\Lambda\mathbf{Q}^\top \qquad \text{with } \mathbf{P}^\top\mathbf{D}\mathbf{P} = \mathbf{Q}^\top\mathbf{Q} = \mathbb{I}_r \text{ and } \Lambda = \text{Diag}(\sqrt{\lambda_k})_{k=1,\ldots,r}$$

Compromise analysis

GPCA of $(\widetilde{\mathbf{X}}, \mathbb{I}_p, \mathbf{D})$ with $\widetilde{\mathbf{X}}_{..t} = \frac{\mathbf{X}_{..t}}{\sqrt{\|\Gamma_t\|_F}}$ and $\mathbf{D} = \frac{1}{n}\text{Diag}(\alpha_1, \ldots, \alpha_T) \otimes \mathbb{I}_n$

$\Leftrightarrow$ eigendecomposition of Γ^c

Solution writes:

$$\widetilde{\mathbf{X}} = \mathbf{P}\Lambda\mathbf{Q}^\top \qquad \text{with } \mathbf{P}^\top\mathbf{D}\mathbf{P} = \mathbf{Q}^\top\mathbf{Q} = \mathbb{I}_r \text{ and } \Lambda = \text{Diag}(\sqrt{\lambda_k})_{k=1,\ldots,r}$$

Representations of:

- observations: *time-specific positions* on the k-th principal component:

  $$k\text{-th column of } \mathbf{F} = \mathbf{P}\Lambda = \begin{bmatrix} \mathbf{F}_1 \\ \vdots \\ \mathbf{F}_T \end{bmatrix}$$

  compromise positions on the k-th principal component: k-th column of $\mathbf{F}^c = \sum_{t=1}^T \alpha_t \mathbf{F}_{..t}$

Compromise analysis

GPCA of $(\widetilde{\mathbf{X}}, \mathbb{I}_p, \mathbf{D})$ with $\widetilde{\mathbf{X}}_{..t} = \frac{\mathbf{X}_{..t}}{\sqrt{\|\Gamma_t\|_F}}$ and $\mathbf{D} = \frac{1}{n}\text{Diag}(\alpha_1, \ldots, \alpha_T) \otimes \mathbb{I}_n$

$\Leftrightarrow$ eigendecomposition of Γ^c

Solution writes:

$$\widetilde{\mathbf{X}} = \mathbf{P}\Lambda\mathbf{Q}^\top \qquad \text{with } \mathbf{P}^\top\mathbf{D}\mathbf{P} = \mathbf{Q}^\top\mathbf{Q} = \mathbb{I}_r \text{ and } \Lambda = \text{Diag}(\sqrt{\lambda_k})_{k=1,\ldots,r}$$

Representations of:

- observations: *time-specific positions* on the k-th principal component:

  $$k\text{-th column of } \mathbf{F} = \mathbf{P}\Lambda = \begin{bmatrix} \mathbf{F}_1 \\ \vdots \\ \mathbf{F}_T \end{bmatrix}$$

  compromise positions on the k-th principal component: k-th column of $\mathbf{F}^c = \sum_{t=1}^T \alpha_t \mathbf{F}_{..t}$
- variables: *compromise positions on circle of correlations* for variable j on k-th axis: $\frac{\sqrt{\lambda_k}}{\hat{\sigma}_j} Q_{jk}$ (*time-specific positions* can also be derived)

SIR Regression framework

[Li, 1991]

$$Y = f(X^\top \mathbf{a}_1, \ldots, X^\top \mathbf{a}_d, \epsilon)$$

for $d < p$ and $f : \mathbb{R}^{d+1} \to \mathbb{R}$, an arbitrary (non linear) function.

$\mathcal{S}_{Y|X} = \text{Span}\{\mathbf{a}_1, \ldots, \mathbf{a}_d\}$ is: Effective Dimension Reduction (EDR) space.

SIR Regression framework

[Li, 1991]

$$Y = f(X^\top \mathbf{a}_1, \ldots, X^\top \mathbf{a}_d, \epsilon)$$

for $d < p$ and $f : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$, an arbitrary (non linear) function.

$\mathcal{S}_{Y|X} = \text{Span}\{\mathbf{a}_1, \ldots, \mathbf{a}_d\}$ is: Effective Dimension Reduction (EDR) space.

Equivalence between SIR and eigendecomposition

$\mathcal{S}_{Y|X}$ is included in the space spanned by the first d Γ-orthogonal eigenvectors of the Γ^e, with $\Gamma = \mathbb{E}\left[(X - \mathbb{E}(X))^T X\right]$ and $\Gamma^e = \mathbb{E}\left(\mathbb{E}(Z|Y)^T \mathbb{E}(Z|Y)\right)$ for $Z = \Gamma^{-1/2}(X - \mathbb{E}(X))$.

SIR in practice

Estimation (when $n > p$)

- compute $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and $\Gamma = \frac{1}{n}\mathbf{X}^T(\mathbf{X} - \bar{\mathbf{x}})$

SIR in practice

Estimation (when $n > p$)

- compute $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and $\Gamma = \frac{1}{n}\mathbf{X}^T(\mathbf{X} - \bar{\mathbf{x}})$
- split the range of Y into H different slices: τ_1, ... τ_H and estimate

$$\mathbf{G} = \left(\frac{1}{n_h}\sum_{i:\ y_i\in\tau_h}\mathbf{z}_i\right)_{h=1,\ldots,H}$$

with $n_h = |\{i : \ y_i \in \tau_h\}|$ and

$$\Gamma^e = \mathbf{G}^\top\mathbf{M}\mathbf{G}$$

with $\mathbf{M} = \text{Diag}\left(\frac{n_1}{n}, \ldots, \frac{n_H}{n}\right)$

SIR in practice

Estimation (when $n > p$)

- compute $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and $\Gamma = \frac{1}{n}\mathbf{X}^T(\mathbf{X} - \bar{\mathbf{x}})$
- split the range of Y into H different slices: τ_1, ... τ_H and estimate

$$\mathbf{G} = \left(\frac{1}{n_h} \sum_{i:\ y_i \in \tau_h} \mathbf{z}_i\right)_{h=1,\ldots,H}$$

  with $n_h = |\{i :\ y_i \in \tau_h\}|$ and

$$\Gamma^e = \mathbf{G}^\top \mathbf{M} \mathbf{G}$$

  with $\mathbf{M} = \text{Diag}\left(\frac{n_1}{n}, \ldots, \frac{n_H}{n}\right)$
- generalized eigendecomposition of Γ^e with norm Γ

Back to our problem: multiway SIR

Basic ideas:

- perform DUAL STATIS analysis on center of gravity of the slices ($\mathbf{G}_{..t}$) instead of the original variables
- compromise analysis is similar to finding a compromise EDR space
- using slices make the method similar to FDA but other estimates of Γ_t^e (not based on slices) could be used

Overview of multiway SIR

Data: $\widetilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_{..1} \\ \vdots \\ \mathbf{X}_{..T} \end{pmatrix}$ and

$\mathbf{y} = (y_1, \ldots, y_n)^\top$

same number of variables, potentially different number of individuals

Overview of multiway SIR

Data: $\widetilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_{..1} \\ \vdots \\ \mathbf{X}_{..T} \end{pmatrix}$ and

$\mathbf{y} = (y_1, \ldots, y_n)^\top$

same number of variables, potentially different number of individuals

$\longrightarrow$

inter-structure analysis

analysis of $\widetilde{\mathbf{C}}$ ($T \times T$ similarity matrix between the covariance matrix of $\mathbf{G}_{..t}$, Γ_t^e)

Overview of multiway SIR

Data: $\widetilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_{..1} \\ \vdots \\ \mathbf{X}_{..T} \end{pmatrix}$ and

$\mathbf{y} = (y_1, \ldots, y_n)^\top$

same number of variables, potentially different number of individuals

→

inter-structure analysis

analysis of $\widetilde{\mathbf{C}}$ ($T \times T$ similarity matrix between the covariance matrix of $\mathbf{G}_{..t}$, Γ_t^e)

table weights

$(\alpha_1, \ldots, \alpha_T)$

compromise matrix

$\Gamma^{e,c} = \sum_{t=1}^T \alpha_t \Gamma_t^e$

Overview of multiway SIR

Data: $\widetilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_{..1} \\ \vdots \\ \mathbf{X}_{..T} \end{pmatrix}$ and

$\mathbf{y} = (y_1, \ldots, y_n)^\top$

same number of variables, potentially different number of individuals

→

inter-structure analysis

analysis of $\widetilde{\mathbf{C}}$ ($T \times T$ similarity matrix between the covariance matrix of $\mathbf{G}_{..t}$, Γ_t^e)

table weights

$(\alpha_1, \ldots, \alpha_T)$

compromise matrix

$\Gamma^{e,c} = \sum_{t=1}^{T} \alpha_t \Gamma_t^e$

GSVD

eig. dec.

compromise analysis

representations of variables and of individuals

compromise positions or time-dependant positions

Inter-structure analysis

Notations: $\forall\, t = 1, \ldots, T$, $\Gamma_t = \frac{1}{n}\mathbf{X}_{..t}^\top(\mathbf{X}_{..t} - \mathbf{1}_n\overline{\mathbf{x}}_t^\top)$, $\mathbf{Z}_{..t} = \left(\mathbf{X}_{..t} - \mathbf{1}_n\overline{\mathbf{x}}_t^\top\right)\Gamma_t^{-1/2}$, $\mathbf{G}_{..t} = \left(\frac{1}{n_h}\sum_{i:\, y_i\in\tau_h}\mathbf{z}_i\right)_{h=1,\ldots,H}$ and $\Gamma_t^e = \mathbf{G}_{..t}^\top\mathbf{M}\,\mathbf{G}_{..t}$ and $\widetilde{\Gamma}_t^e = \frac{\Gamma_t^e}{\|\Gamma_t^e\|_F}$.

Inter-structure analysis

Notations: $\forall\, t = 1, \ldots, T$, $\Gamma_t = \frac{1}{n}\mathbf{X}_{..t}^\top(\mathbf{X}_{..t} - \mathbf{1}_n\overline{\mathbf{x}}_t^\top)$, $\mathbf{Z}_{..t} = \left(\mathbf{X}_{..t} - \mathbf{1}_n\overline{\mathbf{x}}_t^\top\right)\Gamma_t^{-1/2}$,
$\mathbf{G}_{..t} = \left(\frac{1}{n_h}\sum_{i:\, y_i\in\tau_h}\mathbf{z}_i\right)_{h=1,\ldots,H}$ and $\Gamma_t^e = \mathbf{G}_{..t}^\top\mathbf{M}\,\mathbf{G}_{..t}$ and $\widetilde{\Gamma}_t^e = \frac{\Gamma_t^e}{\|\Gamma_t^e\|_F}$.

Similarity between time steps: $\widetilde{\mathbf{C}} = (c_{tt'})_{t,t'=1,\ldots,T}$ with

$$c_{tt'} = \langle\widetilde{\Gamma}_t^e, \widetilde{\Gamma}_{t'}^e\rangle_F$$

(cosinus between Γ_t^e and $\Gamma_{t'}^e$).

Inter-structure analysis

Notations: $\forall\, t = 1, \ldots, T$, $\Gamma_t = \frac{1}{n}\mathbf{X}_{..t}^\top(\mathbf{X}_{..t} - \mathbf{1}_n\overline{\mathbf{x}}_t^\top)$, $\mathbf{Z}_{..t} = \left(\mathbf{X}_{..t} - \mathbf{1}_n\overline{\mathbf{x}}_t^\top\right)\Gamma_t^{-1/2}$, $\mathbf{G}_{..t} = \left(\frac{1}{n_h}\sum_{i:\, y_i\in\tau_h}\mathbf{z}_i\right)_{h=1,\ldots,H}$ and $\Gamma_t^e = \mathbf{G}_{..t}^\top\mathbf{M}\,\mathbf{G}_{..t}$ and $\widetilde{\Gamma}_t^e = \frac{\Gamma_t^e}{\|\Gamma_t^e\|_F}$.

Similarity between time steps: $\widetilde{\mathbf{C}} = (c_{tt'})_{t,t'=1,\ldots,T}$ with

$$c_{tt'} = \langle\widetilde{\Gamma}_t^e, \widetilde{\Gamma}_{t'}^e\rangle_F$$

(cosinus between Γ_t^e and $\Gamma_{t'}^e$).

Inter-structure analysis: Solve

$$\mathbf{u}_1 = \arg\max_{\mathbf{u}=(u_1,\ldots,u_T)^\top\in\mathbb{R}^T,\ \|\mathbf{u}\|=1} \left\|\sum_{t=1}^T u_t\widetilde{\Gamma}_t^e\right\|_F^2$$

Inter-structure analysis

Notations: $\forall\, t = 1, \ldots, T$, $\Gamma_t = \frac{1}{n}\mathbf{X}_{..t}^\top(\mathbf{X}_{..t} - \mathbf{1}_n\bar{\mathbf{x}}_t^\top)$, $\mathbf{Z}_{..t} = \left(\mathbf{X}_{..t} - \mathbf{1}_n\bar{\mathbf{x}}_t^\top\right)\Gamma_t^{-1/2}$,
$\mathbf{G}_{..t} = \left(\frac{1}{n_h}\sum_{i:\, y_i\in\tau_h}\mathbf{z}_i\right)_{h=1,\ldots,H}$ and $\Gamma_t^e = \mathbf{G}_{..t}^\top \mathbf{M}\,\mathbf{G}_{..t}$ and $\widetilde{\Gamma}_t^e = \frac{\Gamma_t^e}{\|\Gamma_t^e\|_F}$.

Similarity between time steps: $\widetilde{\mathbf{C}} = (c_{tt'})_{t,t'=1,\ldots,T}$ with

$$c_{tt'} = \langle \widetilde{\Gamma}_t^e, \widetilde{\Gamma}_{t'}^e \rangle_F$$

(cosinus between Γ_t^e and $\Gamma_{t'}^e$).

Inter-structure analysis: Solve

$$\mathbf{u}_1 = \arg\max_{\mathbf{u}=(u_1,\ldots,u_T)^\top\in\mathbb{R}^T,\ \|\mathbf{u}\|=1} \left\| \sum_{t=1}^T u_t \widetilde{\Gamma}_t^e \right\|_F^2$$

Solution: u_1 is the first eigenvector of $\widetilde{\mathbf{C}}$ $\Rightarrow$ compromise weights:

- $\forall\, t = 1, \ldots, T$, $\alpha_t = \frac{u_{1t}}{\sum_{t'} u_{1t'}}$
- $\Gamma^{e,c} = \sum_{t=1}^T \alpha_t \widetilde{\Gamma}_t^e$

Compromise analysis

GPCA of $(\widetilde{\mathbf{G}}, \mathbb{I}_p, \mathbf{D})$ with $\widetilde{\mathbf{G}}_{..t} = \frac{\mathbf{G}_{..t}}{\sqrt{\|\Gamma_t^e\|_F}}$ and $\mathbf{D} = \mathrm{Diag}(\alpha_1, \ldots, \alpha_T) \otimes \mathbf{M}$

Compromise analysis

GPCA of $(\widetilde{\mathbf{G}}, \mathbb{I}_p, \mathbf{D})$ with $\widetilde{\mathbf{G}}_{..t} = \frac{\mathbf{G}_{..t}}{\sqrt{\|\Gamma_t^e\|_F}}$ and $\mathbf{D} = \text{Diag}(\alpha_1, \ldots, \alpha_T) \otimes \mathbf{M}$

$\Leftrightarrow$ eigendecomposition of $\Gamma^{e,c}$

Compromise analysis

GPCA of $(\widetilde{\mathbf{G}}, \mathbb{I}_p, \mathbf{D})$ with $\widetilde{\mathbf{G}}_{..t} = \frac{\mathbf{G}_{..t}}{\sqrt{\|\Gamma_t^e\|_F}}$ and $\mathbf{D} = \text{Diag}(\alpha_1, \ldots, \alpha_T) \otimes \mathbf{M}$

$\Leftrightarrow$ eigendecomposition of $\Gamma^{e,c}$

Solution writes:

$$\widetilde{\mathbf{G}} = \mathbf{P}\Lambda\mathbf{Q}^\top \qquad \text{with } \mathbf{P}^\top\mathbf{D}\mathbf{P} = \mathbf{Q}^\top\mathbf{Q} = \mathbb{I}_r \text{ and } \Lambda = \text{Diag}(\sqrt{\lambda_k})_{k=1,\ldots,r}$$

Compromise analysis

GPCA of $(\widetilde{\mathbf{G}}, \mathbb{I}_p, \mathbf{D})$ with $\widetilde{\mathbf{G}}_{..t} = \frac{\mathbf{G}_{..t}}{\sqrt{\|\Gamma^e_t\|_F}}$ and $\mathbf{D} = \text{Diag}(\alpha_1, \ldots, \alpha_T) \otimes \mathbf{M}$

$\Leftrightarrow$ eigendecomposition of $\Gamma^{e,c}$

Solution writes:

$$\widetilde{\mathbf{G}} = \mathbf{P}\Lambda\mathbf{Q}^\top \qquad \text{with } \mathbf{P}^\top\mathbf{D}\mathbf{P} = \mathbf{Q}^\top\mathbf{Q} = \mathbb{I}_r \text{ and } \Lambda = \text{Diag}(\sqrt{\lambda_k})_{k=1,\ldots,r}$$

Representations of:

- slices: *time-specific positions* on the k-th principal component: k-th column of $\mathbf{F} = \mathbf{P}\Lambda = \begin{bmatrix} \mathbf{F}_1 \\ \vdots \\ \mathbf{F}_T \end{bmatrix}$

  compromise positions on the k-th principal component: k-th column of $\mathbf{F}^c = \sum_{t=1}^T \alpha_t \mathbf{F}_{..t}$

Compromise analysis

GPCA of $(\widetilde{\mathbf{G}}, \mathbb{I}_p, \mathbf{D})$ with $\widetilde{\mathbf{G}}_{..t} = \frac{\mathbf{G}_{..t}}{\sqrt{\|\Gamma_t^e\|_F}}$ and $\mathbf{D} = \text{Diag}(\alpha_1, \ldots, \alpha_T) \otimes \mathbf{M}$

$\Leftrightarrow$ eigendecomposition of $\Gamma^{e,c}$

Solution writes:

$$\widetilde{\mathbf{G}} = \mathbf{P}\Lambda\mathbf{Q}^\top \qquad \text{with } \mathbf{P}^\top\mathbf{D}\mathbf{P} = \mathbf{Q}^\top\mathbf{Q} = \mathbb{I}_r \text{ and } \Lambda = \text{Diag}(\sqrt{\lambda_k})_{k=1,\ldots,r}$$

Representations of:

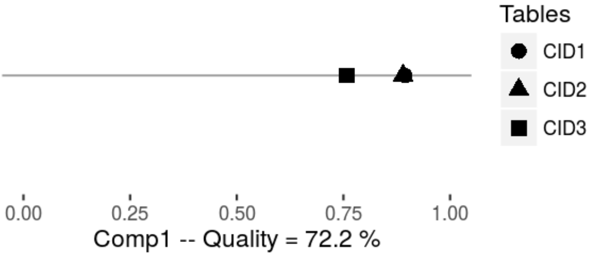
- slices: *time-specific positions* on the the k-th principal component: k-th column of $\mathbf{F} = \mathbf{P}\Lambda = \begin{bmatrix} \mathbf{F}_1 \\ \vdots \\ \mathbf{F}_T \end{bmatrix}$

  compromise positions on the k-th principal component: k-th column of $\mathbf{F}^c = \sum_{t=1}^T \alpha_t \mathbf{F}_{..t}$
- variables: *compromise positions on circle of correlations* for variable j on k-th axis: $\frac{\sqrt{\lambda_k}}{\hat{\sigma}_j} Q_{jk}$ (*time-specific positions* can also be derived)

Sommaire

Application of multiway SIR on Diogenes dataset I

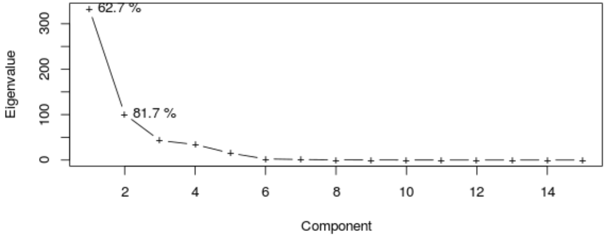
$H = 5$, interstructure analysis:



CID1 and CID2 have more similar Γ_t^e than CID3.

Application of multiway SIR on Diogenes dataset II

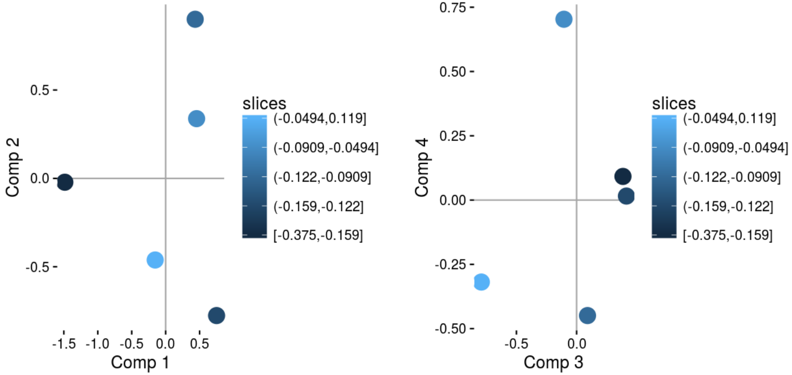
$H = 5$, compromise analysis: number of components



2 components are enough but we will analyze 4 for a deeper understanding of the data.

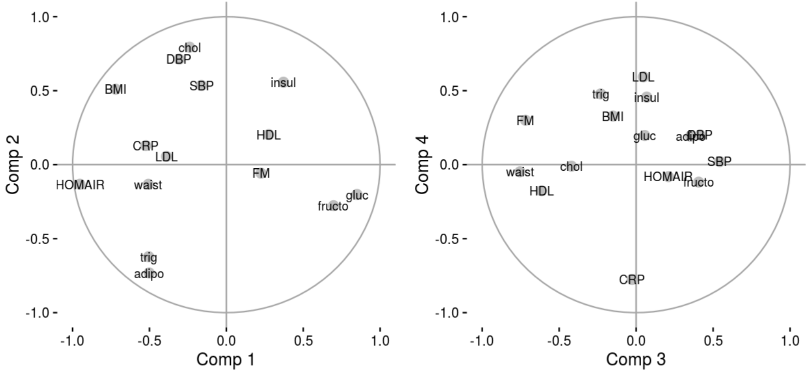
Application of multiway SIR on Diogenes dataset III

$H = 5$, compromise analysis: analysis of the slices (compromise)



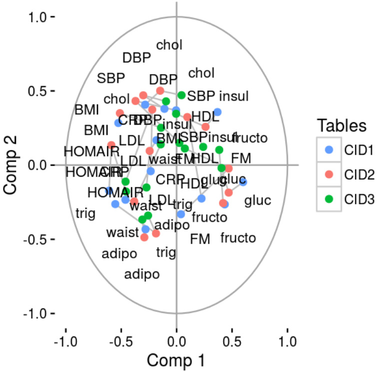
Application of multiway SIR on Diogenes dataset IV

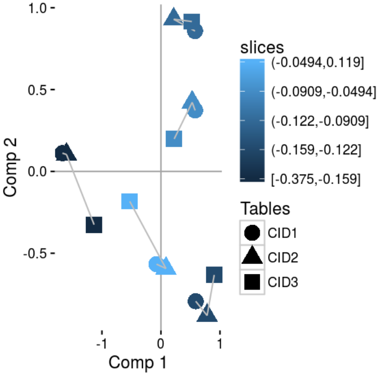
$H = 5$, compromise analysis: analysis of the variable (compromise)



Application of multiway SIR on Diogenes dataset V

$H = 5$, compromise analysis: longitudinal analysis





Conclusion

We have presented an exploratory analysis method

- based on DUAL STATIS
- able to focus on a numeric variable of interest similarly to SIR
- able to explain longitudinal evolutions when the number of time steps is small

Future work

- an R package is under development
- only valid for $n \geq p$: regularization approach is under study to allows for the analysis of $n < p$ cases

Abdi, H., Williams, L., Valentin, D., and Bennani-Dosse, M. (2012).
STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling.
Wiley Interdisciplinary Reviews: Computational Statistics, 4(2):124–167.

Escofier, B. and Pagès, J. (2008).
Analyses Factorielles Simples et Multiples.
Dunod.

Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P. (1994).
The ACT (STATIS method).
Computational Statistics and Data Analysis, 18(1):97–119.

Li, K. (1991).
Sliced inverse regression for dimension reduction.
Journal of the American Statistical Association, 86(414):316–342.