



KATHOLIEKE UNIVERSITEIT
LEUVEN

A structured overview of simultaneous component methods

K. Van Deun I. Van Mechelen,
A. Smilde, H. Kiers, & M. Van der Werf





Outline

- Introduction
 - Coupled data
 - Simultaneous component methods
- Coupled two-way two mode data
- The general framework
- Specific methods
 - SUM-PCA
 - unr. PCovR
 - MFA
 - STATIS
 - SCA-P
- Comparison of the methods
- Illustrative application
- Discussion



Introduction: Coupled data

- Challenge
 - Represent coupled data such that shared and specific information is captured
 - E.g., retrieve modules of genes with same transcription factors and co-regulated under same conditions (common info)
 - E.g., highlight classes of compounds measured by both separation methods & by a specific method
 - Note: all targeted info pertains to variance within each of the matrices)!



Introduction: Simultaneous Component methods

- Extract components from all data matrices simultaneously
- SUM-PCA, unr. PCovR, SCA-P, MFA, STATIS
 - Use different terminologies / mathematical frame (stem from different disciplines) => not easy to compare
 - Yield different results



Introduction: Simultaneous Component methods

- Goal of this talk:

Provide a structured overview that relates all methods to the same mathematical framework

=> allows for a detailed comparison
(common core + particularities)

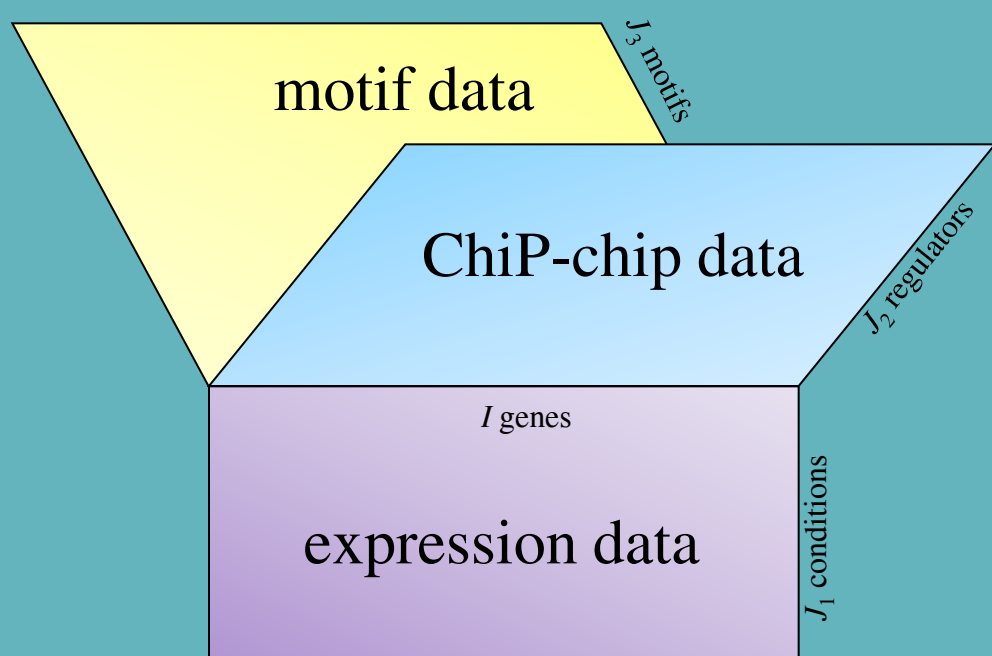


Coupled two-way two-mode data

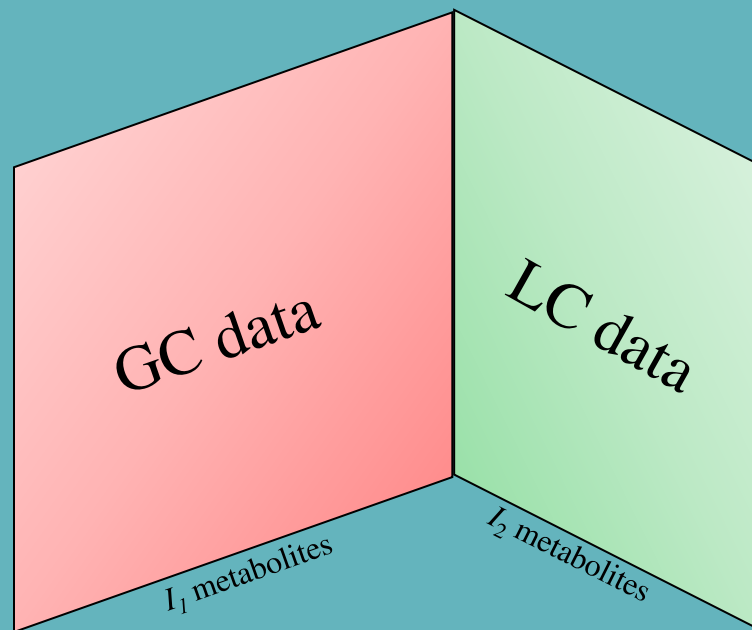
- SC methods apply to data
 - consisting of at least two two-way two-mode data blocks (multi-block / -set data)
 - at least one *mode* is common
 - indicates the sets of units that underlie the data
 - E.g., gene expression data: condition & gene mode
 - Convention: rows-> object mode; columns -> variable mode



Coupled two-way two-mode data



Coupled object / row mode



Coupled variable / column mode



Coupled two-way two-mode data

- Note: defining data as coupled two-way two-mode data might be subjective to some extent
 - Splitting a data matrix into several data matrices
 - Decoupling three-way three-mode data



The general framework

- Elements:
 1. The data
 2. Pre-processing
 3. Model
 4. Objective function
 5. Identification constraints
 6. Model estimation



The general framework

1. The data

- Multiple two-way two-mode data matrices that have one mode in common
- Common mode:
 - Object or variable mode



The general framework

2. Pre-processing

- Correct for irrelevant differences
 - Measurement scale differences between variables
 - Centering
 - Scaling
 - Standardizing / Auto-scaling
 - Size / variation differences between matrices
 - Scale to sum of squares one



The general framework

3. The model

- Principal components structure for one $I_k \times J_k$ data matrix \mathbf{X}_k ($k=1, \dots, K$)

$$w_k \mathbf{X}_k = \mathbf{T}_k \mathbf{P}_k' + \mathbf{E}_k \quad (1)$$

- w_k ($w_k \geq 0$): pre-specified matrix-specific weight
- \mathbf{T}_k : components scores ($I_k \times R$)
- \mathbf{P}_k : loadings ($J_k \times R$)
- \mathbf{E}_k : residuals ($I_k \times J_k$)



The general framework

3. The model

$$w_k \mathbf{X}_k = \mathbf{T}_k \mathbf{P}'_k + \mathbf{E}_k \quad (1)$$

- Simultaneous components structure for coupled row data: **Constraint: $\mathbf{T}_k = \mathbf{T}_1 = \dots = \mathbf{T}_K = \mathbf{T}$** (common component scores \mathbf{T})

$$w_k \mathbf{X}_k = \mathbf{T} \mathbf{P}'_k + \mathbf{E}_k, \quad \forall k \quad (2)$$

- Simultaneous components structure for coupled column data: **Constraint: $\mathbf{P}_k = \mathbf{P}_1 = \dots = \mathbf{P}_K = \mathbf{P}$** (common loadings \mathbf{P})

$$w_k \mathbf{X}_k = \mathbf{T}_k \mathbf{P}' + \mathbf{E}_k, \quad \forall k \quad (3)$$



The general framework

4. Objective function

- Common object mode

$$\min_{\mathbf{T}, \mathbf{P}_k} \sum_k \|\mathbf{X}_k - \mathbf{T}\mathbf{P}'_k\|^2 \quad (4)$$

- Common variable mode

$$\min_{\mathbf{T}_k, \mathbf{P}} \sum_k \|\mathbf{X}_k - \mathbf{T}_k\mathbf{P}'_k\|^2 \quad (5)$$



The general framework

5. Identification constraints

- In general: If \mathbf{T} and \mathbf{P}_k is a solution, \mathbf{TB} and $\mathbf{P}_k\mathbf{B}^{-1}$ is a solution too (\mathbf{B} non-singular)
- Common identification constraints
 - Principal axes orientation
 - Orthonormality of component scores or loadings



The general framework

6. Model estimation

- Two methods:
 - SVD of concatenated data
 - Two step approach (ED of sum of cross product matrices + (block-specific) regression)



The general framework

- SVD of concatenated data:

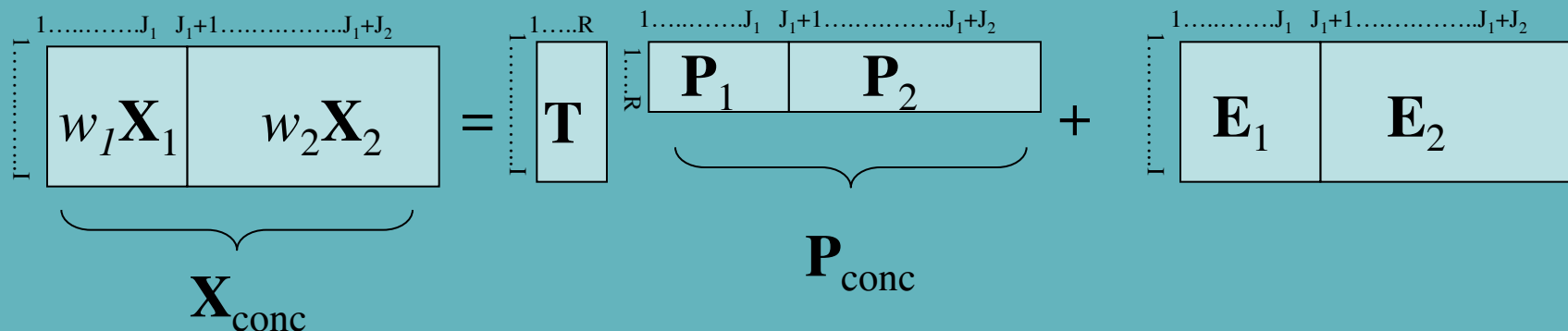
$$\min_{\mathbf{T}, \mathbf{P}_k} \sum_{k=1}^K \left\| w_k \mathbf{X}_k - \mathbf{T} \mathbf{P}'_k \right\|^2 = \min_{\mathbf{T}, \mathbf{P}_k} \left\| \mathbf{X}_{\text{conc}} - \mathbf{T} \mathbf{P}'_{\text{conc}} \right\|^2 \quad (6)$$

$$\mathbf{X}_{\text{conc}} = \mathbf{U} \mathbf{S} \mathbf{V}' = \mathbf{U} \mathbf{S}^p \mathbf{S}^{p-1} \mathbf{V}'$$

$$\mathbf{T} = \mathbf{U} \mathbf{S}^p$$

$$\mathbf{P}_{\text{conc}} = \mathbf{V} \mathbf{S}^{p-1}$$

(with $p=0$ if $\mathbf{T}'\mathbf{T}=\mathbf{I}$ and $p=1$ if $\mathbf{P}_{\text{conc}}'\mathbf{P}_{\text{conc}}=\mathbf{I}$)





The general framework

- Two step approach:
 1. Derive \mathbf{T} by ED of

$$\mathbf{X}_{conc} \mathbf{X}'_{conc} = w_k^2 \sum_k \mathbf{X}_k \mathbf{X}'_k \quad (7)$$

2. Derive $\mathbf{P}_k / \mathbf{P}_{conc}$ by regression

$$\mathbf{P}_k = \mathbf{X}'_k \mathbf{T} (\mathbf{T}' \mathbf{T})^{-1}$$

$$\mathbf{P}_{conc} = \mathbf{X}'_{conc} \mathbf{T} (\mathbf{T}' \mathbf{T})^{-1} \quad (8)$$

- Gives the same results as SVD of \mathbf{X}_{conc}



Specific methods: 1. SUM-PCA

- Smilde et al. 2003
- This is not SUMPCA (Kiers, 1989)
- Data: common object mode
- Pre-processing
 - Variables: standardized (within each matrix)
 - Matrices: scaled to sum of squares one
- Model: $\mathbf{X}_k = \mathbf{T}\mathbf{P}_k' + \mathbf{E}_k, (w_k=1)$ (9)
- Id. constr.: principal axes + $\mathbf{T}'\mathbf{T}=\mathbf{I}$



Specific methods: 2. unr.PCovR

- de Jong & Kiers (1992), Gurden (unpubl.)
- Data: common object mode, 2 blocks
- Pre-processing
 - Variables: different options (centering, scaling)

- Model:
$$\begin{aligned}\beta\mathbf{X}_1 &= \mathbf{TP}'_1 \\ (1-\beta)\mathbf{X}_2 &= \mathbf{TP}'_2\end{aligned}\quad (10)$$

with β , $0 \leq \beta \leq 1$, obtained by cross-validation
(asymmetry + unsatisfying results)



Specific methods: 2. unr.PCovR

- Model:
$$\begin{aligned}\beta \mathbf{X}_1 &= \mathbf{TP}'_1 \\ (1 - \beta) \mathbf{X}_2 &= \mathbf{TP}'_2\end{aligned}\quad (10)$$

with β , $0 \leq \beta \leq 1$, obtained by cross-validation

Note: Introduces asymmetry and gave unsatisfying results

- Id.constraints: princ.axes $+\mathbf{T}'\mathbf{T}=\mathbf{I}$



Specific methods: 3. MFA

- Escofier & Pagès (1983, 1998)
- Data: common object mode
- Pre-processing
 - Variables: standardized (within each matrix)
- Model:

$$\sigma_{k1}^{-1} \mathbf{X}_k = \mathbf{TP}'_k + \mathbf{E}_k, \quad (11)$$

with σ_{k1} the largest singular value of \mathbf{X}_k



Specific methods: 3. MFA

- Model:

$$\sigma_{k1}^{-1} \mathbf{X}_k = \mathbf{TP}'_k + \mathbf{E}_k, \quad (11)$$

Corrects for size and redundancy:

$$\frac{1}{\sqrt{(\sum_j \sigma_{kj}^2)}} \frac{1}{\sqrt{\left(\frac{\sigma_{k1}^2}{\sum_j \sigma_{kj}^2}\right)}} \mathbf{X}_k = \frac{1}{\sigma_{k1}} \mathbf{X}_k \quad (12)$$

- Id. constraints: princ.axes + $\mathbf{T}'\mathbf{T}=\mathbf{I}$



Specific methods: 4. STATIS

- L'hermier des Plantes & Thiébaud (1977)
- Data: common object mode (3-w)
- Pre-processing
 - Original publ.: not discussed
 - Lavit et al. (1994): centering and scaling;
Stanimirova et al. (2004): correction for block size



Specific methods: 4. STATIS

- Model:

$$a_k \mathbf{X}_k = \mathbf{TP}'_k + \mathbf{E}_k, \quad (13)$$

with a_k derived as follows:

- Derive CP matrices $\mathbf{S}_k = \mathbf{X}_k \mathbf{X}_k'$
- Construct \mathbf{F} , of size $p \times K$, with $\text{vec}(\mathbf{S}_k)$ in the k^{th} column
- a_k are loadings on first PC of \mathbf{F}



Specific methods: 4. STATIS

- larger weights can be expected for
 - matrices with larger values,
 - larger matrices,
 - matrices with more covariation between objects,
 - matrices with more similar CP matrices ($K > 2$)
- Id. constraints: princ. axes + $\mathbf{P}'_{\text{conc}} \mathbf{P}_{\text{conc}} = \mathbf{I}$



Specific methods: 5. SCA-P

- Kiers & Ten Berge (1994)
- Data: common variable mode
- Pre-processing
 - Variables: standardized within each block
 - Note: Timmerman & Kiers (2003) center each variable within each matrix and scale to $ss=1$ over the matrices
- Model

$$\mathbf{X}_k = \mathbf{T}_k \mathbf{P}' + \mathbf{E}_k, \quad (14)$$



Specific methods: 5. SCA-P

- Model

$$\mathbf{X}_k = \mathbf{T}_k \mathbf{P}' + \mathbf{E}_k, \quad (14)$$

so, no block-specific weighting but note that

$$\sum_k \mathbf{X}'_k \mathbf{X}_k = \sum_k \mathbf{R}_k, \quad (15)$$

- Id. constraints: princ.axes+ $\mathbf{T}_{\text{conc}}' \mathbf{T}_{\text{conc}} = \mathbf{I}$



Comparison of the methods

	Common	Pre-processing	Matrix-specific	Identification
	mode		weights	constraint
SUM-PCA	object	Variables: auto-scaling Matrices: scaled to sum of squares one	All $w_k=1$	Principal axes $\mathbf{T}^T \mathbf{T} = \mathbf{I}$
unrestricted PCovR	object	Variables: auto-scaling	Minimize cross-validation error	Principal axes $\mathbf{T}^T \mathbf{T} = \mathbf{I}$
MFA	object	Variables: auto-scaling	Inverse of largest singular value	Principal axes $\mathbf{T}^T \mathbf{T} = \mathbf{I}$
STATIS	object		Compromise weights	Principal axes $\mathbf{P}_{\text{conc}} \mathbf{P}_{\text{conc}} = \mathbf{I}$
SCA-P	variable	Variables: auto-scaling	All $w_k=1$	Principal axes $\mathbf{T}_{\text{conc}}^T \mathbf{T}_{\text{conc}} = \mathbf{I}$



Comparison of the methods

- Aspects that are consequential for the obtained results:
 - Pre-processing
 - Only STATIS does not standardize the variables
=> solutions may be dominated by a few variables
 - Matrix-specific weights
 - Extreme: all weight put on a single matrix
 - If high matrix-correlation (e.g., RV coeff.) weighting is of little impact



Comparison of the methods

- Aspects that are consequential for the obtained results:
 - Labeling of common mode as object or variable in combination with autoscaling

$$\sum_k \mathbf{X}_k \mathbf{X}'_k \left\{ \begin{array}{l} \rightarrow \sum_k \mathbf{R}_k : \text{common variable mode} \\ \rightarrow \sum_k \mathbf{S}_k : \text{common object mode} \end{array} \right.$$

Size of correlation: does not depend on sample size.
This is not so for cross-products.

=> Larger matrices do not necessarily contribute more! (do not naively correct for size)



A 'fair' integration

<i>Principle</i>	<i>Methods aiming at this principle</i>
Same weight for all matrices (naive approach)	SCA-P
More weight for smaller matrices	SUM-PCA, MFA
More weight for less redundant matrices	MFA
More weight for matrices with more stable prediction information	PCovR
More weight for matrices that share more information with other matrices ($K > 2$)	STATIS



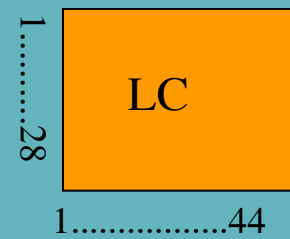
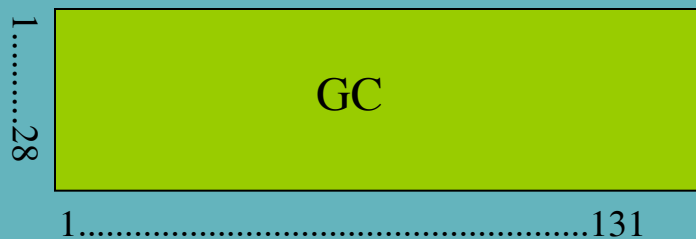
Discussion

- Combining principles (multiplying weights)
 - MFA
 - Scaled STATIS (Qanari et al.)
- MFA in presence of a very noisy data block
 - Noisy data block \approx least redundant thus largest weight



Illustrative application

- Metabolome of *E. coli* screened under various exper.cond. & fermentation times
 - Both GC/MS and LC/MS were used for the same 28 samples of *E. coli*
- => data consisting of two data matrices that are coupled by the object mode





Illustrative application

- Application of the different published methods
 - Pre-processing:
 - All values were log-transformed
 - Remaining pre-processing: according to the published method
 - Unr. PCovR: Two cases
 - PCovR GC: leave-one-out cross-validation using LC data to reproduce GC data
 - PCovR LC: use GC data to reproduce LC data



Illustrative application

	GC	LC
PCovR GC ¹	1.00	0
STATIS	.99	.01
SCA-P	.77	.23
MFA	.66	.34
SUM-PCA	.50	.50
PCovR LC ¹	0	1.00

Relative weights calculated as matrix-specific SS divided by SS of concatenated data



Illustrative application

- Modified RV coefficient between GC and LC: 0.20
 - Different weighting can be expected to give different results
 - Comparison of common component scores using Tucker's coefficient of congruence



Illustrative application

	PCov R GC	STATIS	SCA-P	MFA	SUM- PCA	PCovR LC	LC²
GC¹	1	.13	.91	.86	.81	.55	.55
PCovR GC³		.13	.91	.86	.81	.55	.55
STATIS			.12	.12	.11	.08	.08
SCA-P				.99	.96	.73	.73
MFA					.99	.79	.79
SUM-PCA						.84	.84
PCovR LC⁴							1

Tucker's coefficient of congruence between the component scores ($R=5$).



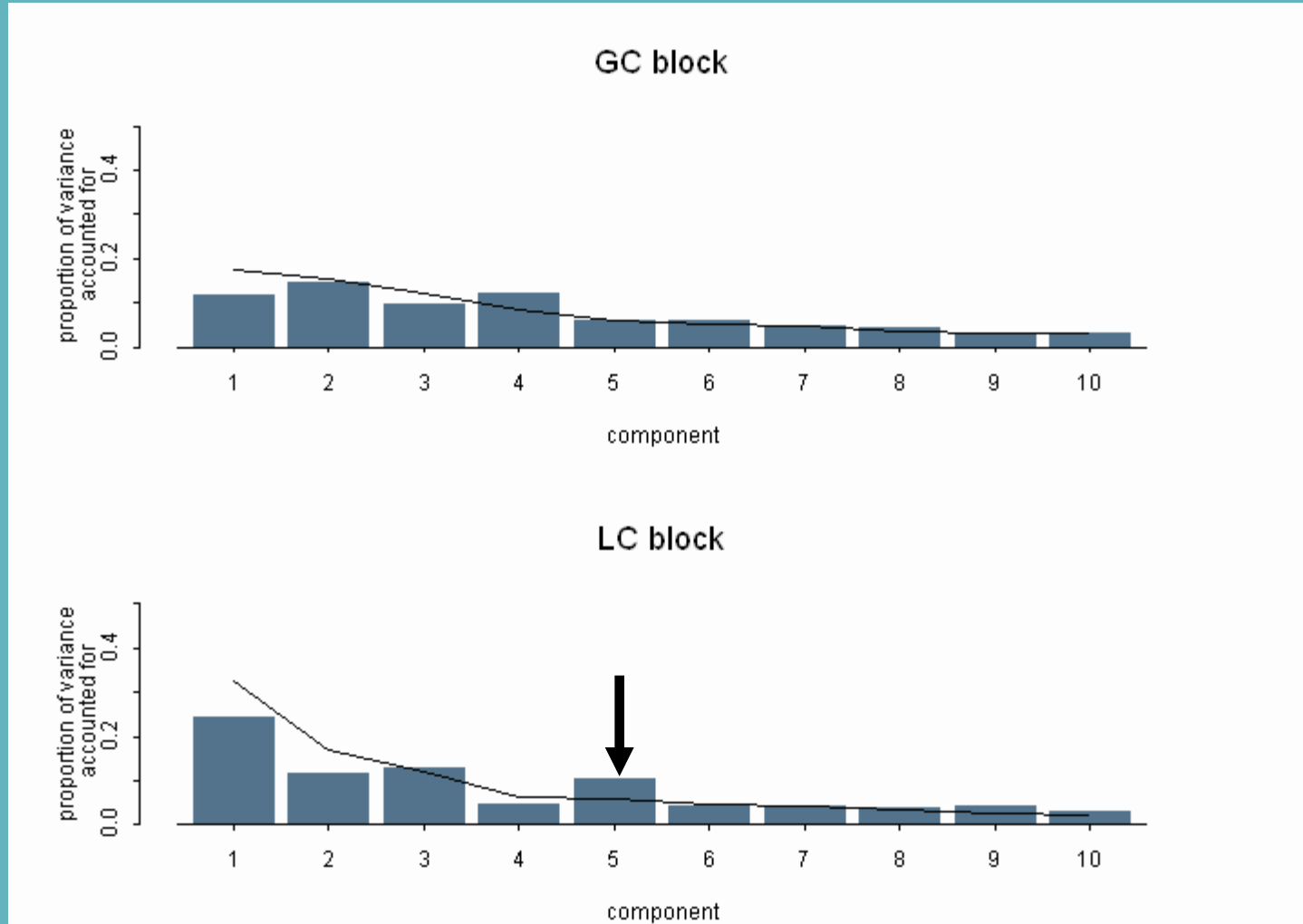
Illustrative application

- Which method?
 - GC larger than LC but biological processes are equally important => size correction
 - LC measures mainly nucleotides while GC measures a larger variety of metabolites; here interest in variety

⇒ MFA
- How many components?



Illustrative application





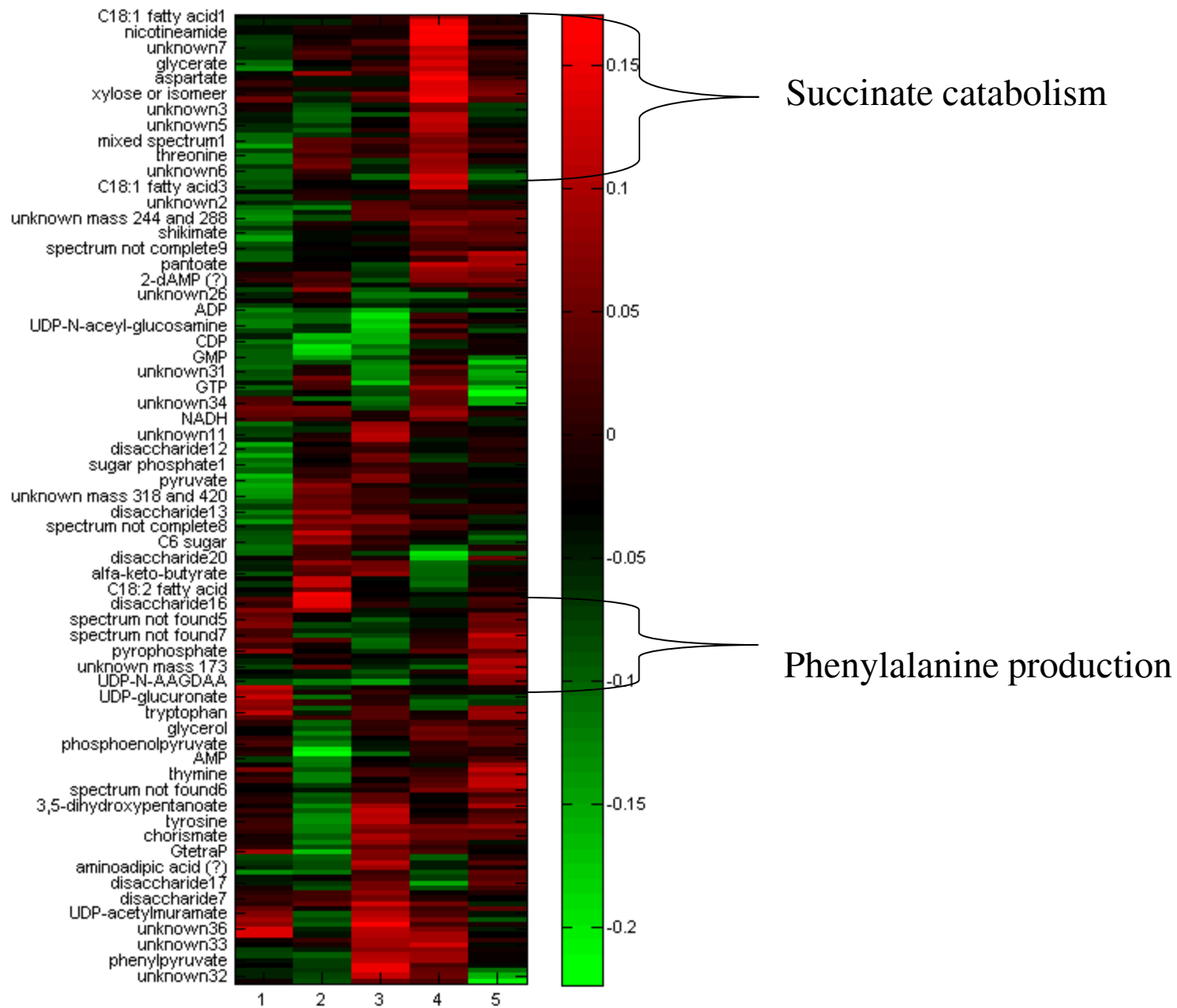
Illustrative application

	PC1	PC2	PC3	PC4	PC5
GC	0,14	0,12	0,08	0,14	0,06
LC	0,10	0,14	0,20	0,07	0,13
TOTAL	0,13	0,13	0,12	0,12	0,09

Proportion of VAF by MFA solution

		SC1	SC2	SC3	SC4	SC5
Reference	16	0,07	0,33	-0,04	0,01	-0,12
	24	-0,05	-0,02	-0,10	-0,09	-0,16
	32	-0,01	-0,30	-0,05	-0,07	-0,06
	40	0,10	-0,27	-0,01	-0,09	0,01
	48	0,18	-0,17	0,17	-0,03	0,10
pH+	16	0,18	-0,11	-0,29	-0,10	-0,61
	24	0,03	-0,07	0,36	-0,10	-0,03
	40	0,35	-0,17	0,26	0,03	-0,09
	48	0,22	-0,10	0,05	-0,10	0,03
Oxygen-	40	-0,24	0,04	0,02	0,00	-0,06
Oxygen?	16	0,04	0,34	0,01	-0,11	-0,06
	24	-0,26	-0,02	-0,02	-0,08	-0,16
	40	-0,45	-0,13	0,09	-0,05	0,08
	64	-0,37	-0,10	0,11	-0,02	0,07

		SC1	SC2	SC3	SC4	SC5
Phosphate +	16	0,01	0,38	-0,02	-0,04	0,03
	24	-0,09	0,43	0,04	-0,10	0,14
	40	-0,33	0,00	-0,04	-0,01	-0,01
	48	-0,04	-0,12	0,01	0,01	-0,15
Phosphate -	16	-0,03	0,06	0,00	-0,19	-0,09
	24	-0,03	-0,02	0,41	-0,12	0,07
	40	0,32	0,09	0,32	-0,01	0,14
Succinate	24	0,07	0,10	-0,02	0,55	-0,13
	40	-0,01	-0,02	-0,01	0,57	-0,03
	48	-0,05	-0,05	-0,05	0,46	0,11
Wild type	16	0,19	0,27	-0,25	-0,10	0,06
	24	0,02	-0,11	-0,38	-0,12	0,04
	40	0,08	-0,20	-0,32	-0,06	0,33
	48	0,11	-0,06	-0,26	-0,03	0,55





Concluding remarks

- We proposed a general framework for SC methods based on 6 aspects
 - Data
 - Pre-processing
 - Model
 - **Objective function**
 - Identification constraints
 - **Model estimation**



Concluding remarks

- The different published methods make choices wrt these aspects
- As shown, the choices wrt pre-processing and weighting are consequential;
- Labeling the common mode as object or variable is consequential when considered in combination with pre-processing.
- Posing the orthon.constr. on **T** or **P** is not consequential for the reproduced scores.



Concluding remarks

- The type of weighting is essential for SC methods and different methods link this up to different principles of a *fair* integration
- As illustrated, appropriate pre-processing and weighting depends on substantive issues.