Multiple Nested Reductions of Single Data Modes as a Tool to Deal with Large Data Sets

Iven Van Mechelen and Katrijn Van Deun

K.U.Leuven
Psychology Department
and
Center for Computational Systems Biology

Invited IFCS session at COMPSTAT 2010

Overview:

- introduction
- principles
- example 1: existing model
- example 2: novel model
- discussion

Overview:

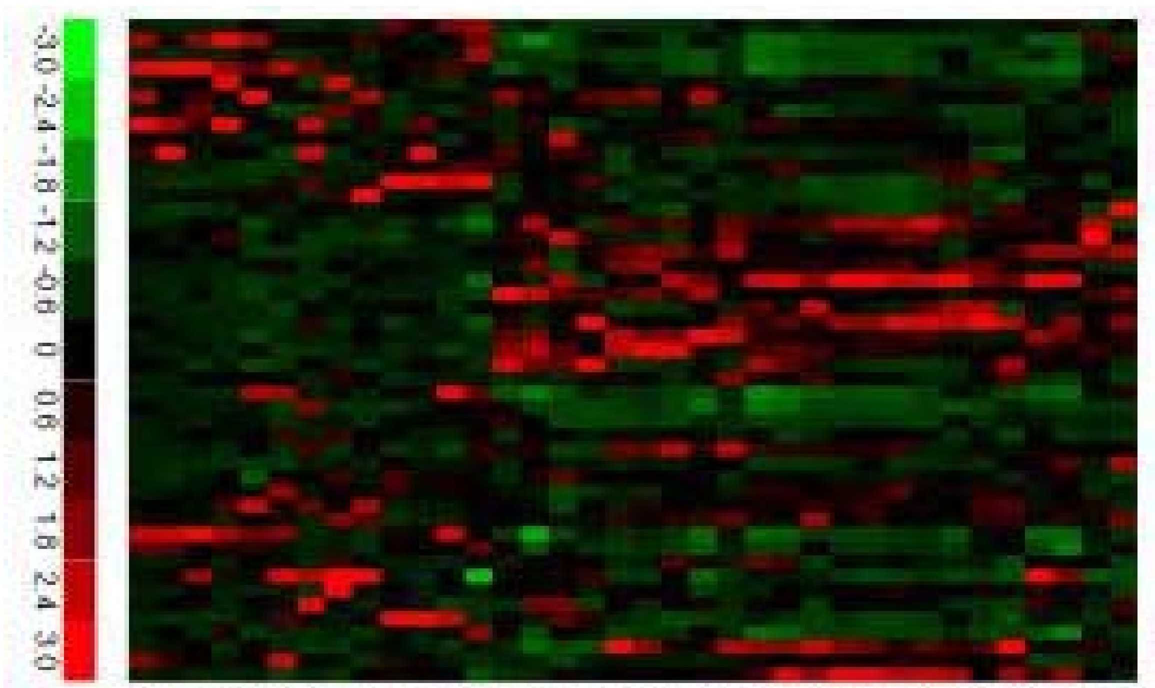
- introduction
- principles
- example 1: existing model
- example 2: novel model
- discussion

Introduction

- in many research areas:
  - accessibility of novel measurement technologies
  - data tsunami: highdimensional data sets
  - example: various types of ‘omics’ data

Introduction

- in many research areas:
  - accessibility of novel measurement technologies
  - data tsunami: highdimensional data sets
  - example: various types of ‘omics’ data



Introduction

- in many research areas:
  - accessibility of novel measurement technologies
  - data tsunami: highdimensional data sets
  - example: various types of ‘omics’ data

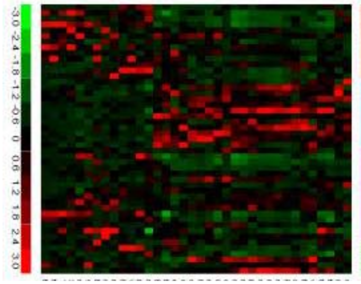
- concerted use of technologies in many settings
  - data sets with large number of experimental units

Introduction (ctd)

- problems:

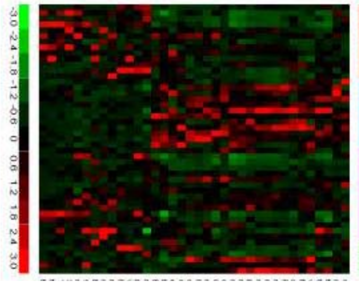
Introduction (ctd)

- problems:
  - redundancies, dependencies, ill-conditioned optimization problems



Introduction (ctd)

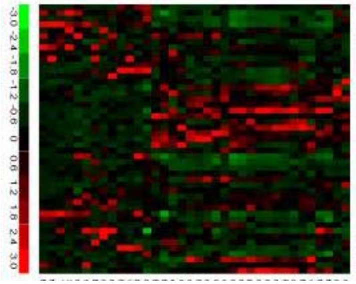
- problems:
  - redundancies, dependencies, ill-conditioned optimization problems



  - computational bottlenecks

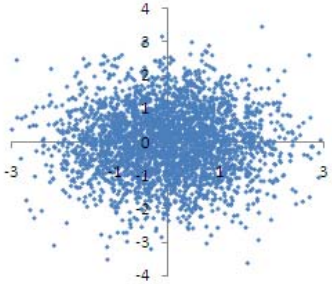
<u>Introduction</u> (ctd)

- problems:
  - redundancies, dependencies, ill-conditioned optimization problems



  - computational bottlenecks

  - displaying output prohibitive



Introduction (ctd)

- possible solution: classical reduction methods (categorical: clustering; continuous: dimension reduction)

Introduction (ctd)

- possible solution: classical reduction methods (categorical: clustering; continuous: dimension reduction)

- however: often breakdown of such methods …

Introduction (ctd)

- possible solution: classical reduction methods (categorical: clustering; continuous: dimension reduction)

- however: often breakdown of such methods …

- possible rescue missions: variable selection, sparseness penalty or constraints, …

Introduction (ctd)

- possible solution: classical reduction methods (categorical: clustering; continuous: dimension reduction)

- however: often breakdown of such methods …

- possible rescue missions: variable selection, sparseness penalty or constraints, …

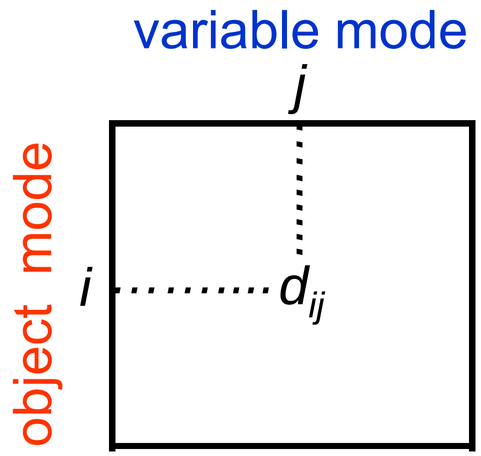
- alternative solution: multiple nested reductions of single data modes (within framework of global model for data, fitted with a simultaneous optimization procedure)

Overview:

- introduction
- principles
- example 1: existing model
- example 2: novel model
- discussion

Principles

- data: $I \times J$ object by variable (e.g., tissue by gene) data matrix **D**



Principles (ctd)

- (deterministic core of) generic decomposition model (Van Mechelen & Schepers, 2007):
  - reduction of object (tissue) mode by means of (binary or real-valued) $I \times P$ quantification matrix **A**

    examples:

Principles (ctd)

- (deterministic core of) generic decomposition model (Van Mechelen & Schepers, 2007):
  - reduction of object (tissue) mode by means of (binary or real-valued) $I \times P$ quantification matrix **A**

  examples:

Tissue_1	1	0	0
Tissue_2	1	0	0
Tissue_3	0	0	1
Tissue_4	0	0	1
Tissue_5	0	1	0
...			

Principles (ctd)

- (deterministic core of) generic decomposition model (Van Mechelen & Schepers, 2007):
  - reduction of object (tissue) mode by means of (binary or real-valued) $I \times P$ quantification matrix **A**

  examples:

$Tissue_1$	1	1	0
$Tissue_2$	1	1	0
$Tissue_3$	1	0	1
$Tissue_4$	1	0	1
$Tissue_5$	1	0	1
...			

Principles (ctd)

- (deterministic core of) generic decomposition model (Van Mechelen & Schepers, 2007):
  - reduction of object (tissue) mode by means of (binary or real-valued) $I \times P$ quantification matrix **A**

examples:

Tissue_1	3.2	5.2	5.1
Tissue_2	4.1	-6.7	3.4
Tissue_3	5.8	3.9	1.9
Tissue_4	1.0	-2.1	0.5
Tissue_5	-2.3	8.0	-1.7
...			

Principles (ctd)

- (deterministic core of) generic decomposition model (Van Mechelen & Schepers, 2007):
  - reduction of object (tissue) mode by means of (binary or real-valued) $I \times P$ quantification matrix **A**
  - reduction of variable (gene) mode by means of (binary or real-valued) $J \times Q$ quantification matrix **B**

Principles (ctd)

- (deterministic core of) generic decomposition model (Van Mechelen & Schepers, 2007):
  - reduction of object (tissue) mode by means of (binary or real-valued) $I \times P$ quantification matrix **A**
  - reduction of variable (gene) mode by means of (binary or real-valued) $J \times Q$ quantification matrix **B**
  - $P \times Q$ core matrix **W**

Principles (ctd)

- (deterministic core of) generic decomposition model (Van Mechelen & Schepers, 2007):
  - reduction of object (tissue) mode by means of (binary or real-valued) $I \times P$ quantification matrix **A**
  - reduction of variable (gene) mode by means of (binary or real-valued) $J \times Q$ quantification matrix **B**
  - $P \times Q$ core matrix **W**
  - decomposition operator *f*, which is such that:

$$\mathbf{D} = f\left(\mathbf{A}, \mathbf{B}, \mathbf{W}\right) + \mathbf{E}$$

with $f(\mathbf{A},\mathbf{B},\mathbf{W})_{ij}$ only depending on $\mathbf{A}_{i\cdot}$ and $\mathbf{B}_{j\cdot}$

Principles (ctd)

$$\mathbf{D} = f\left(\mathbf{A}, \mathbf{B}, \mathbf{W}\right) + \mathbf{E}$$

- special cases:

Principles (ctd)

$$\mathbf{D} = f(\mathbf{A}, \mathbf{B}, \mathbf{W}) + \mathbf{E}$$

- special cases:
  - **A** and **B** binary, *f* additive operator:

$$f(\mathbf{A}, \mathbf{B}, \mathbf{W}) = \mathbf{A}\,\mathbf{W}\,\mathbf{B}^t$$

$$f(\mathbf{A}, \mathbf{B}, \mathbf{W})_{ij} = \sum_{p=1}^{P} \sum_{q=1}^{Q} a_{ip}\, b_{jq}\, w_{pq}$$

(general additive two-mode clustering model)

$$f(\mathbf{A},\mathbf{B},\mathbf{W})_{ij} = \sum_{p=1}^{P}\sum_{q=1}^{Q} a_{ip}\, b_{jq}\, w_{pq}$$

A

	$A_{\bullet 1}$	$A_{\bullet 2}$
O_1	0	0
O_2	1	0
O_3	1	0
O_4	1	1
O_5	0	1
O_6	0	0

V_1	V_2	V_3	V_4	V_5	V_6	V_7	
0	0	0	0	0	0	0	O_1
0	2	2	2	0	0	0	O_2
0	2	2	2	0	0	0	O_3
0	2	2	5	3	3	0	O_4
0	0	0	3	3	3	0	O_5
0	0	0	0	0	0	0	O_6

W

$B_{\bullet 1}$	2	0
$B_{\bullet 2}$	0	3
	$A_{\bullet 1}$	$A_{\bullet 2}$

B

0	1	1	1	0	0	0	$B_{\bullet 1}$
0	0	0	1	1	1	0	$B_{\bullet 2}$
V_1	V_2	V_3	V_4	V_5	V_6	V_7	

Principles (ctd)

$$\mathbf{D} = f\left(\mathbf{A}, \mathbf{B}, \mathbf{W}\right) + \mathbf{E}$$

- special cases (ctd):

  - **A** and **B** real-valued, **W** identity matrix, *f* additive operator:

$$f\left(\mathbf{A}, \mathbf{B}, \mathbf{W}\right) = \mathbf{A}\mathbf{B}^t$$

$$f\left(\mathbf{A}, \mathbf{B}, \mathbf{W}\right)_{ij} = \sum_{p=1}^{P} a_{ip} \, b_{jp}$$

(principal component analysis)

Principles (ctd)

$$\mathbf{D} = f\left(\mathbf{A}, \mathbf{B}, \mathbf{W}\right) + \mathbf{E}$$

- special cases (ctd):
  - **A** and **B** real-valued, **W** identity matrix, *f* Euclidean distance-based operator:

$$f\left(\mathbf{A}, \mathbf{B}, \mathbf{W}\right)_{ij} = \left[\sum_{p=1}^{P}\left(a_{ip} - b_{jp}\right)^2\right]^{\frac{1}{2}}$$

(multidimensional unfolding)

Principles (ctd)

$$\mathbf{D} = f\left(\mathbf{A},\mathbf{B},\mathbf{W}\right) + \mathbf{E}$$

- multiple nested reductions:
  - decomposition of core matrix **W**:

$$\mathbf{W} = f^{*}\left(\mathbf{A}^{*},\mathbf{B}^{*},\mathbf{W}^{*}\right)$$

and therefore:

$$\mathbf{D} = f\left(\mathbf{A},\mathbf{B},f^{*}\left(\mathbf{A}^{*},\mathbf{B}^{*},\mathbf{W}^{*}\right)\right) + \mathbf{E}$$

with **A*** denoting a $P \times P^*$ quantification matrix,
B* a $Q \times Q^*$ quantification matrix,
f^* a decomposition operator,
and with $f^*(\mathbf{A}^*,\mathbf{B}^*,\mathbf{W}^*)_{pq}$ only depending on $\mathbf{A}^*_{p\cdot}$ and $\mathbf{B}^*_{q\cdot}$

Principles (ctd)

$$\mathbf{D} = f\left(\mathbf{A},\mathbf{B},f^{*}\left(\mathbf{A}^{*},\mathbf{B}^{*},\mathbf{W}^{*}\right)\right)+\mathbf{E}$$

- remarks:
  - each of the quantification matrices (**A**, **A***, **B**, **B***) can be an identity matrix (no reduction), a binary matrix (categorical, cluster-based reduction), or a real-valued matrix (continuous, dimension reduction)
  - model is to be estimated as a whole, making use of one overall objective or loss function (unlike in '*tandem*' approaches)

Overview:

- introduction
- principles
- example 1: existing model
- example 2: novel model
- discussion

Example 1: Existing model

$$\mathbf{D} = f\left(\mathbf{A},\mathbf{B},f^{*}\left(\mathbf{A}^{*},\mathbf{B}^{*},\mathbf{W}^{*}\right)\right)+\mathbf{E}$$

- two-mode unfolding clustering:
  - **A** and **B** binary partition matrices, *f* additive operator (i.e., outer model = two-mode partitioning)
  - **A*** and **B*** real-valued matrices, **W*** identity matrix, *f* Euclidean-distance based operator (i.e., inner model = multidimensional unfolding)

$$d_{ij} = \left[\sum_{p=1}^{P}\sum_{q=1}^{Q} a_{ip}\, b_{jq}\left[\sum_{p^*=1}^{P^*}\left(a^{*}_{pp^*} - b^{*}_{qp^*}\right)^{2}\right]^{\frac{1}{2}}\right] + e_{ij}$$

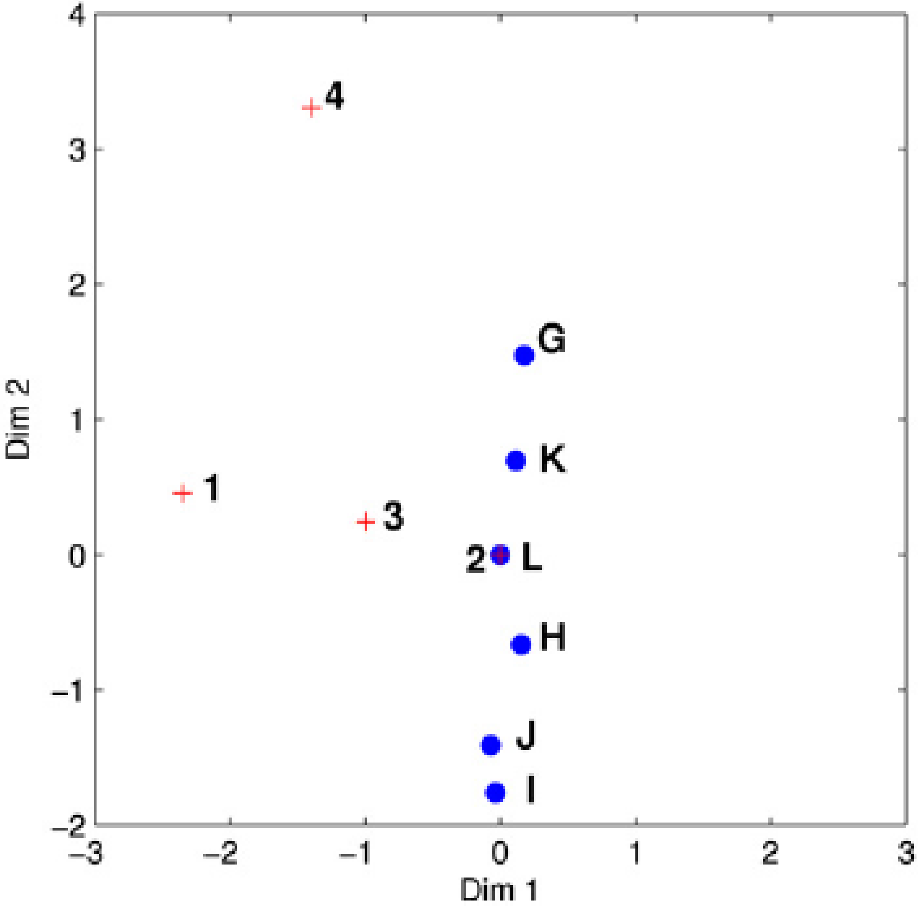
Example 1: Existing model (ctd)

$$d_{ij} = \left[\sum_{p=1}^{P} \sum_{q=1}^{Q} a_{ip} \, b_{jq} \left[\sum_{p^*=1}^{P^*} \left(a^*_{pp^*} - b^*_{qp^*} \right)^2 \right]^{\frac{1}{2}} \right] + e_{ij}$$

- two-mode unfolding clustering: (ctd)
  - originally proposed (in deterministic form) by Van Mechelen & Schepers (2007)
  - stochastic variant (making use of double mixture approach) proposed by Vera, Macías & Heiser (2009) under the name dual latent class unfolding
  - special case: **A** or **B** identity matrix (outer categorical reduction of one mode only): latent class unfolding as proposed by De Soete & Heiser (1993)

Example 1: Existing model (ctd)

- application (Vera et al.): respondent by statement on internet use



Overview:

- introduction
- principles
- example 1: existing model
- example 2: novel model
- discussion

Example 2: Novel model

$$\mathbf{D} = f\left(\mathbf{A},\mathbf{B},f^{*}\left(\mathbf{A}^{*},\mathbf{B}^{*},\mathbf{W}^{*}\right)\right) + \mathbf{E}$$

- two-mode principal component clustering:
  - data centered or standardized variablewise
  - **A** and **B** binary partition matrices, *f* additive operator (i.e., outer model = two-mode partitioning)
  - **A*** and **B*** real-valued matrices, **W*** identity matrix, *f* additive operator (i.e., inner model = principal component analysis)

$$d_{ij} = \left[\sum_{p=1}^{P}\sum_{q=1}^{Q} a_{ip}\, b_{jq} \left(\sum_{p^*=1}^{P^*} a^{*}_{pp^*}\, b^{*}_{qp^*}\right)\right] + e_{ij}$$

Example 2: Novel model (ctd)

$$d_{ij} = \left[\sum_{p=1}^{P} \sum_{q=1}^{Q} a_{ip}\, b_{jq} \left(\sum_{p^*=1}^{P^*} a^*_{pp^*}\, b^*_{qp^*} \right) \right] + e_{ij}$$

- two-mode principal component clustering: (ctd)
  - in matrix notation:

$$\mathbf{D} = \mathbf{A}\left(\mathbf{A}^* \mathbf{B}^{*t}\right)\mathbf{B}^t + \mathbf{E}$$

  - special case: **B** identity matrix (no reduction)
    → *k*-means clustering in a low-dimensional Euclidean space (De Soete & Carroll, 1994)

- in deterministic scenario: least squares loss function

$$\min_{\mathbf{A},\mathbf{B},\mathbf{A}^*,\mathbf{B}^*} \left\| \mathbf{D} - \mathbf{A}\left(\mathbf{A}^* \mathbf{B}^{*t}\right)\mathbf{B}^t \right\|^2$$

Example 2: Novel model (ctd)

$$\min_{\mathbf{A},\mathbf{B},\mathbf{A}^*,\mathbf{B}^*} \left\| \mathbf{D} - \mathbf{A}\left(\mathbf{A}^* \mathbf{B}^{*t}\right)\mathbf{B}^t \right\|^2$$

- algorithmic solution (ALS type):

1. initialize **A** and **B**, e.g., through randomly started *k*-means analyses on rows and column of **D**

2. estimate/update **A*** and **B*** through generalized SVD in the metrics $\left[\mathrm{diag}\left(\mathbf{A}^t\mathbf{A}\right)\right]^{-1}$ and $\left[\mathrm{diag}\left(\mathbf{B}^t\mathbf{B}\right)\right]^{-1}$ of the matrix of the two-mode centroids, $\left[\mathrm{diag}\left(\mathbf{A}^t\mathbf{A}\right)\right]^{-1}\mathbf{A}^t\mathbf{D}\,\mathbf{B}\left[\mathrm{diag}\left(\mathbf{B}^t\mathbf{B}\right)\right]^{-1}$

3. update **A** and **B** through rowwise exhaustive search

Repeat 2 and 3 until convergence.

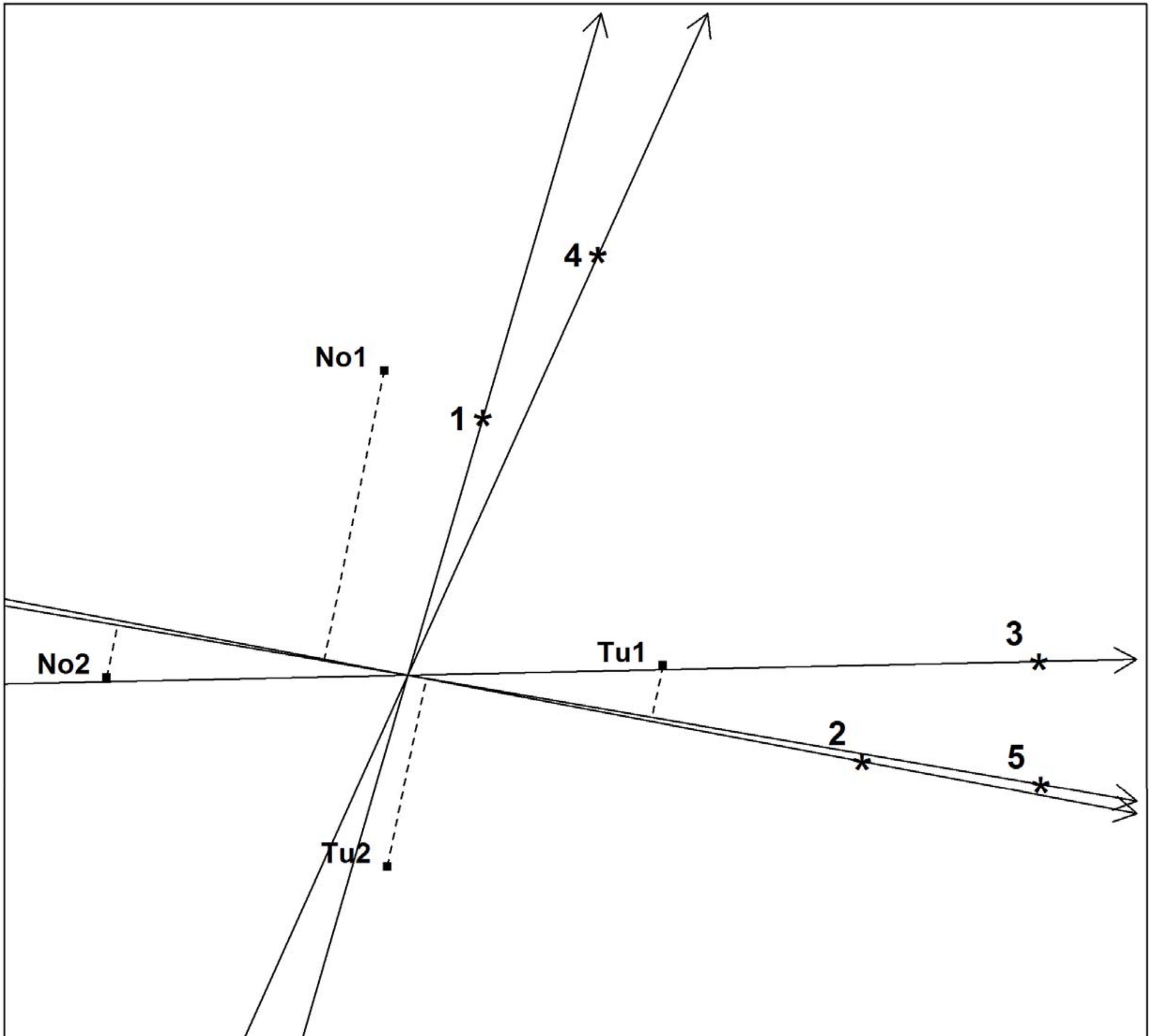
Example 2: Novel model (ctd)

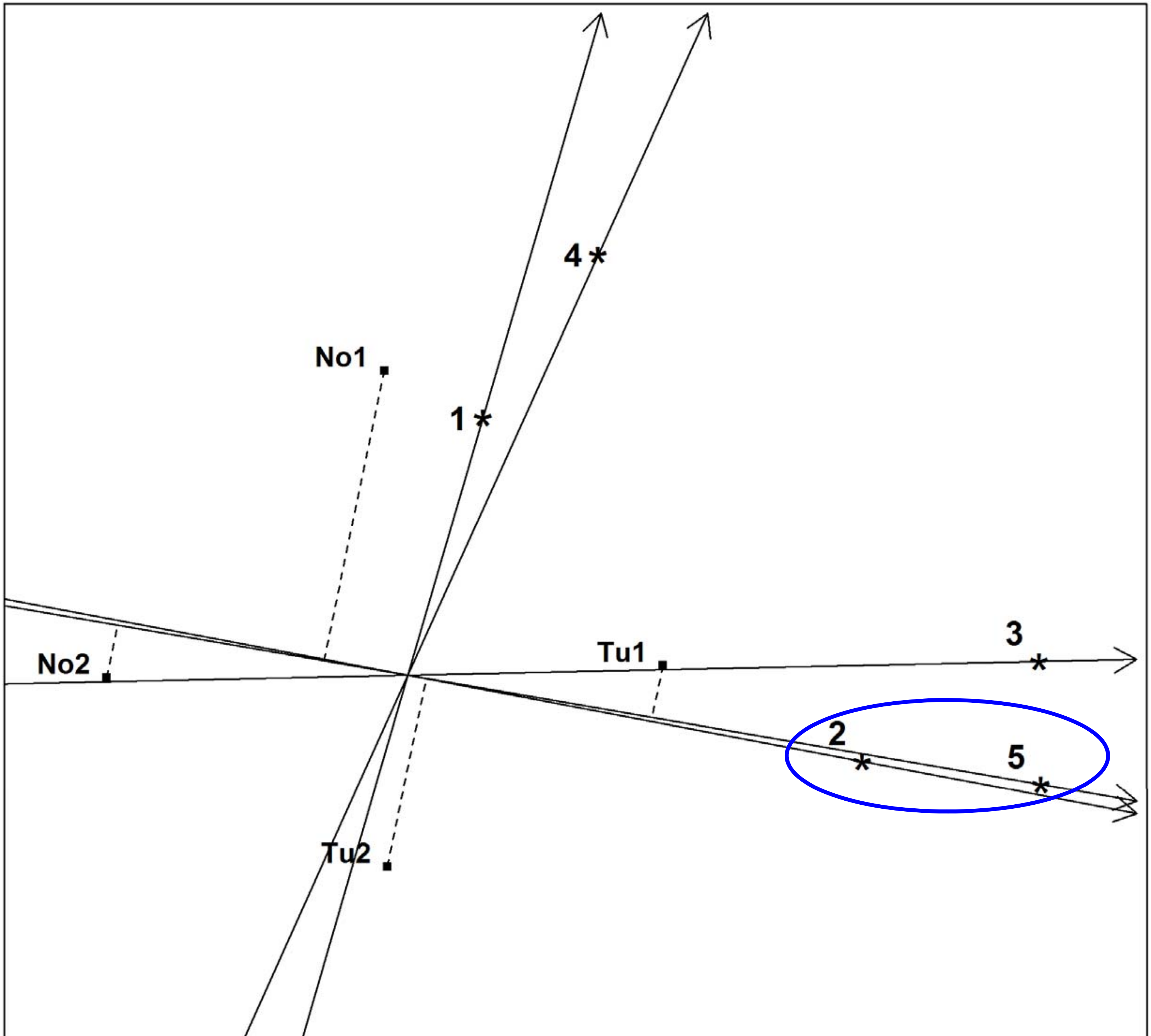
$$\min_{\mathbf{A},\mathbf{B},\mathbf{A}^*,\mathbf{B}^*} \left\| \mathbf{D} - \mathbf{A}\left(\mathbf{A}^* \mathbf{B}^{*t}\right)\mathbf{B}^t \right\|^2$$

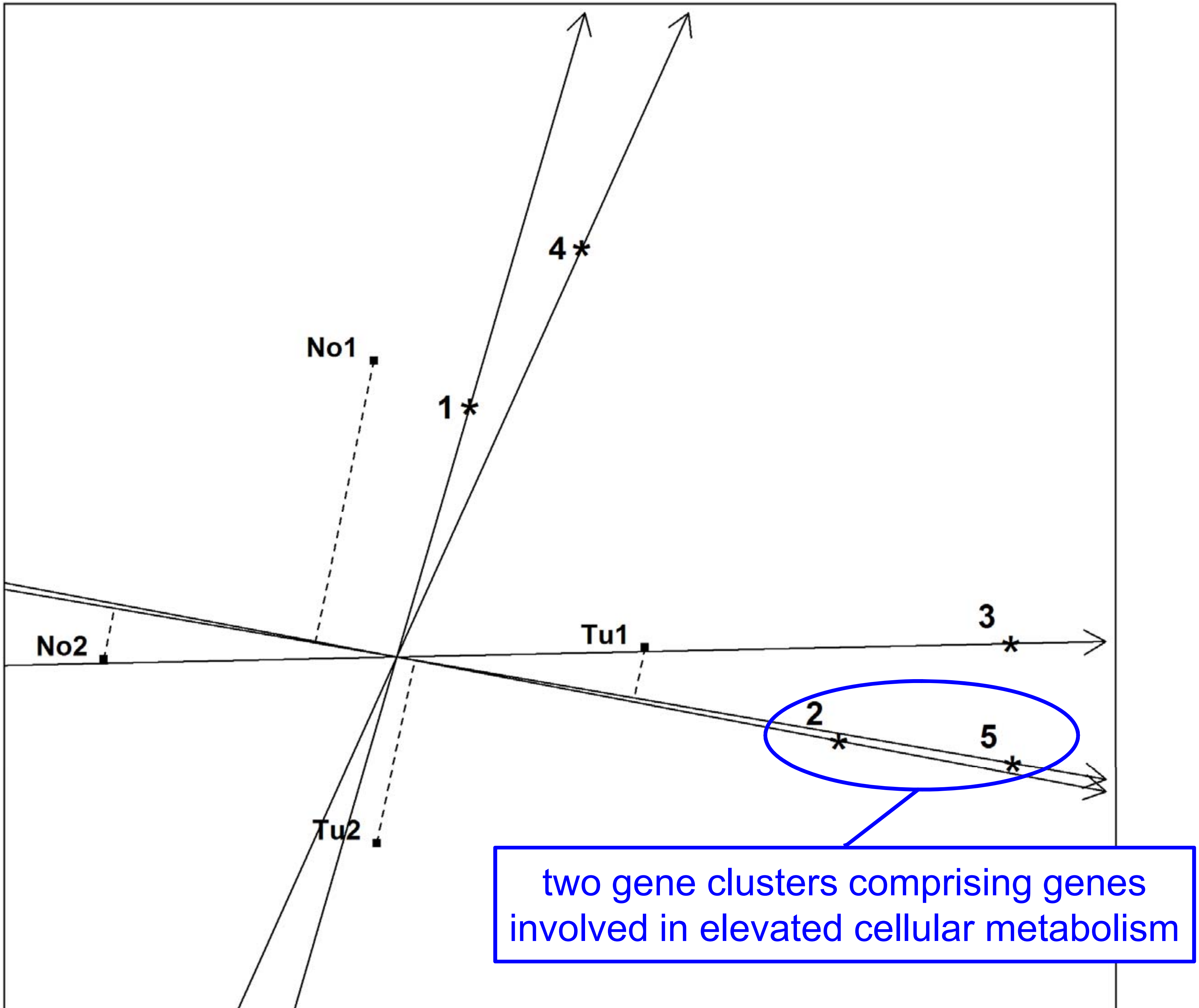
- algorithmic solution (ALS type): (ctd)
  - optional: postprocess final $\mathbf{A}^*$ by means of regular SVD to preserve columnwise orthonormality
  - possibility of convergence to local minimum → multistart strategy

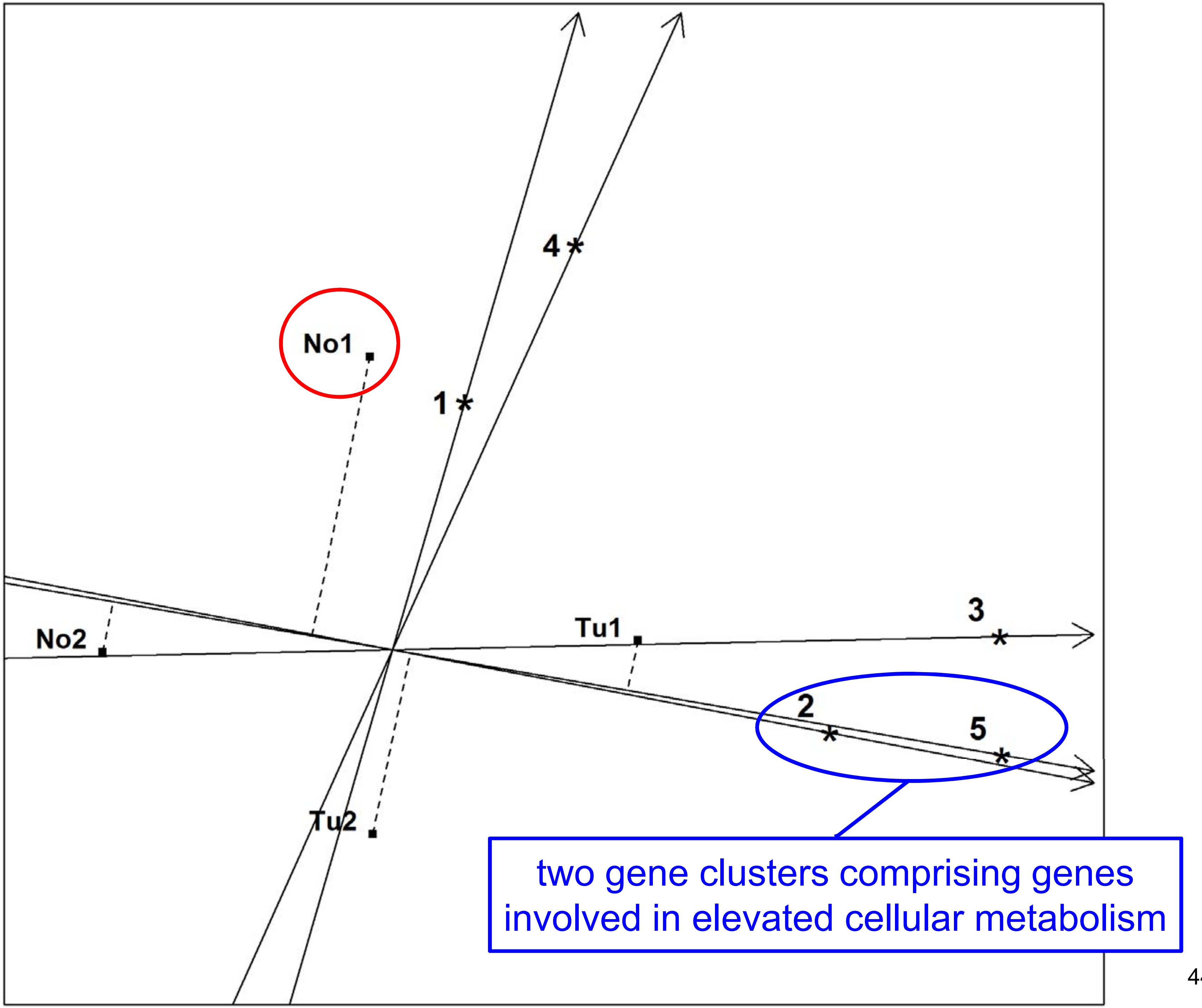
Example 2: Novel model (ctd)

- illustrative application:
  - data from study by Alon et al. (1999) on gene expression in 40 tumor and 22 normal tissues
  - here only data on 400 genes that maximally differentiated cancer from normal tissues
  - ALS algorithm with 500 starts
  - selection of model with 4 tissue clusters, 5 gene clusters and 2 components
  - two tissue clusters largely pertained to tumor tissues and the two other ones to normal tissues

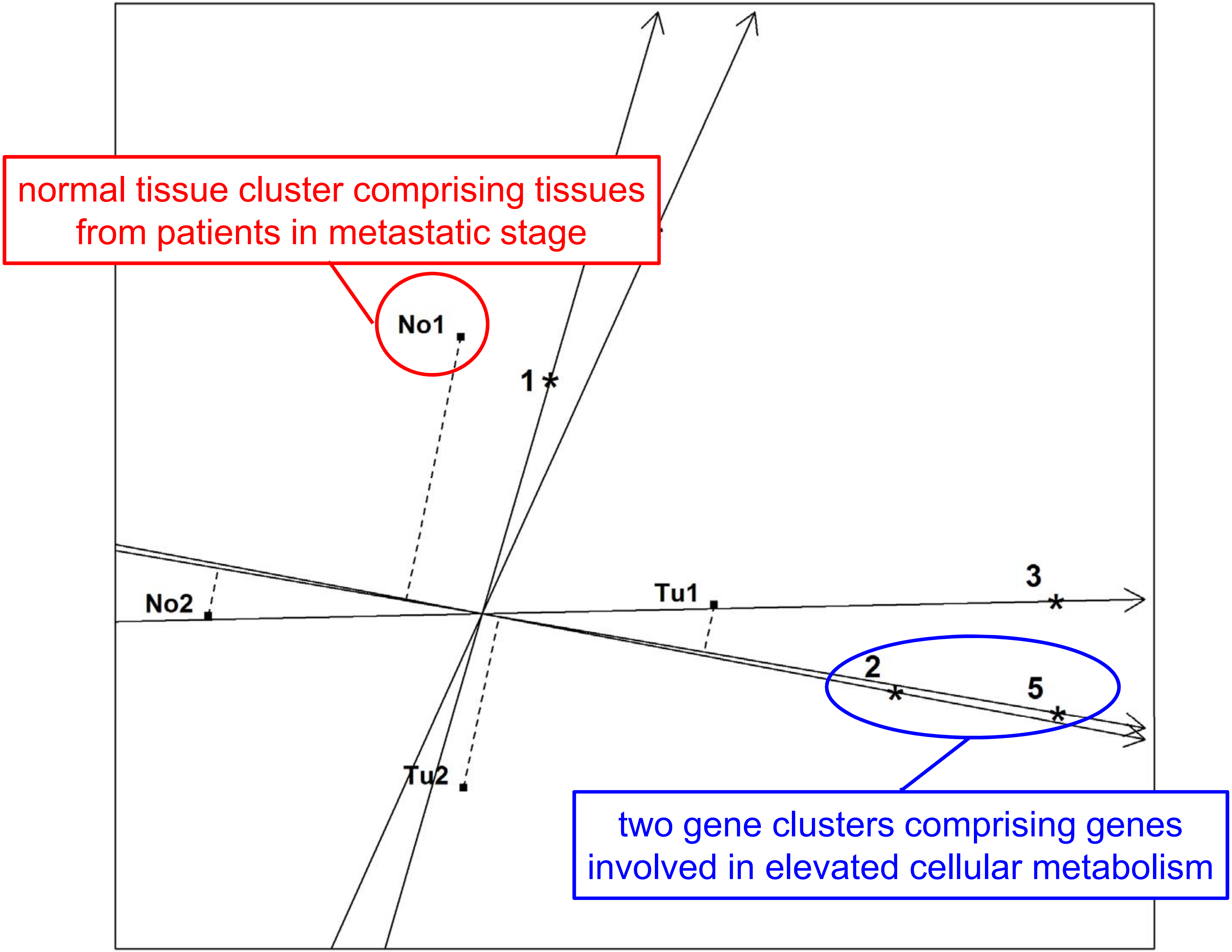

4
No1
1
3
No2
Tu1
2
5
Tu2


4
No1
1
No2
Tu1
3
2
5
Tu2


4
No1
1
3
Tu1
No2
2
5
Tu2
two gene clusters comprising genes involved in elevated cellular metabolism



4
No1
1
3
Tu1
No2
2
5
Tu2
two gene clusters comprising genes involved in elevated cellular metabolism


normal tissue cluster comprising tissues from patients in metastatic stage
No1
1
No2
Tu1
3
2
5
Tu2
two gene clusters comprising genes involved in elevated cellular metabolism

Overview:

- introduction
- principles
- example 1: existing model
- example 2: novel model
- discussion

Discussion

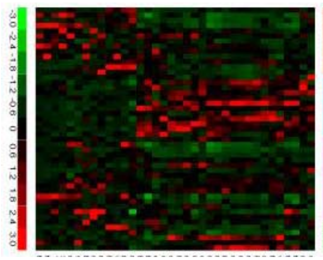
- principle of multiple nested reductions can be extended to:
  - three- and higher-mode data
  - more than two levels of reduction
- inner en outer reductions can fulfill different functions (e.g., outer ones may capture redundancies, and inner ones core substantive mechanisms)
- multiple nested reductions of a single data mode ≠ simultaneous single reductions of several modes (as in classical two-mode clustering techniques and in methods for multimode data analysis)
- multiple nested reductions of a single data mode ≠ inter-woven categorical/dimensional reductions as in 'clustering & disjoint principal component analyis' (Vichi & Saporta, 2009)

Discussion (ctd)

- approach addresses problems as outlined at the start:

Discussion (ctd)

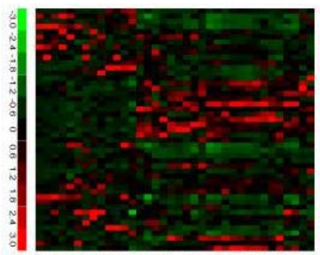
- approach addresses problems as outlined at the start:



  - redundancies, dependencies

Discussion (ctd)

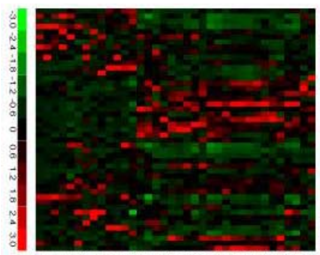
- approach addresses problems as outlined at the start:



  - redundancies, dependencies
    → through outer reduction (no need for discar-
    ding information or for arbitrary choices)

Discussion (ctd)

- approach addresses problems as outlined at the start:



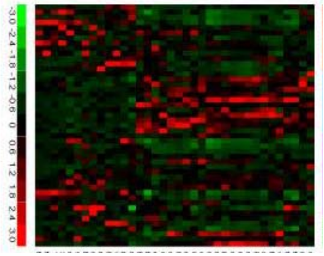
- redundancies, dependencies

  → through outer reduction (no need for discarding information or for arbitrary choices)

- computational bottlenecks

Discussion (ctd)

- approach addresses problems as outlined at the start:

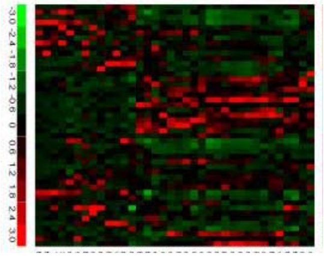


- redundancies, dependencies
  → through outer reduction (no need for discarding information or for arbitrary choices)

- computational bottlenecks
  → see, e.g., inner GSVD to be applied to small matrix with centroids

Discussion (ctd)

- approach addresses problems as outlined at the start:

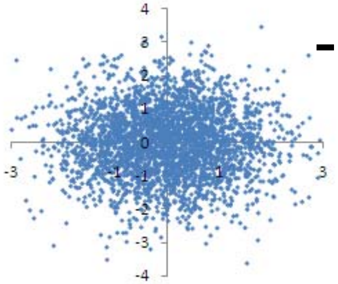


- redundancies, dependencies

  → through outer reduction (no need for discar-ding information or for arbitrary choices)

- computational bottlenecks

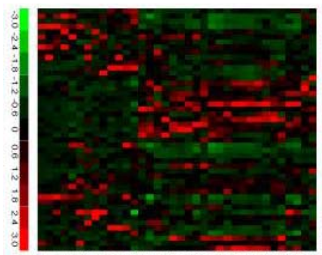
  → see, e.g., inner GSVD to be applied to small matrix with centroids



- displaying output prohibitive

Discussion (ctd)

- approach addresses problems as outlined at the start:

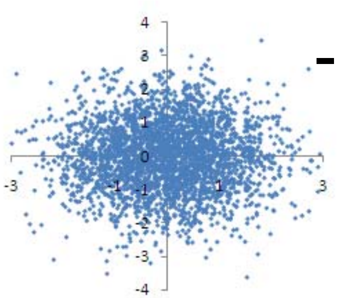


- redundancies, dependencies

  → through outer reduction (no need for discarding information or for arbitrary choices)

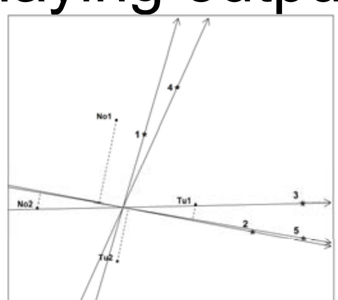
- computational bottlenecks

  → see, e.g., inner GSVD to be applied to small matrix with centroids



- displaying output prohibitive

  →



Iven.VanMechelen@psy.kuleuven.be

ppw.kuleuven.be/okp

thank you for your attention!