european conference on quality in official statistics 2014

SAPIENZA UNIVERSITÀ DI ROMA

New challenges for modelling Big Data

Maurizio Vichi

Department of Statistics
SapienzaUniversity of Rome

em: maurizio.vichi@uniroma1.it

Position of the problem: Big Data and Methodological solutions

The information acquires a central position as a consequence of three cultural and technological revolutions, transforming in the last 40 years the ***industrial society*** into the ***information society***.

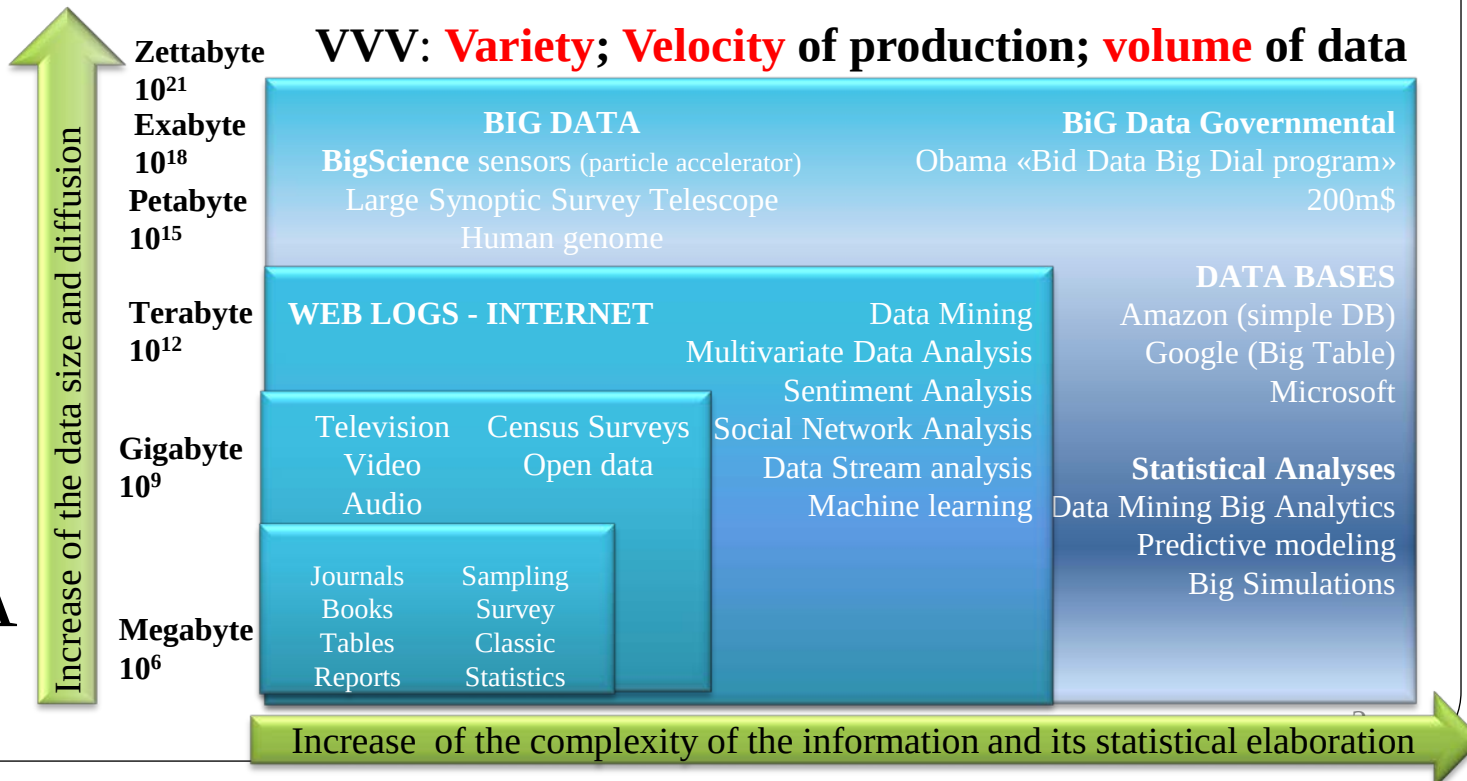
First revolution : ***The industrial Press.*** In the 20th century *Journals* and *Books* are largely reproduced increasing the volume of the *information* in the society; read, write and counting is necessary for job market inclusion;

Second revolution: ***Radio and TV*** produce a big massification of culture and increase of *information*; sources of information are still domain of a restricted number of media;

Third revolution: ***computers*** and ***Internet*** induce the *globalization of* information and economies; there are infinite sources of production of information. New technologies and automation reduce manual skills

A new paradigm of the information *velocity – variety* and *volume*

the data Deluge BIG DATA



Typical complete data for official statistics: Data Cube: Three-way Data array X

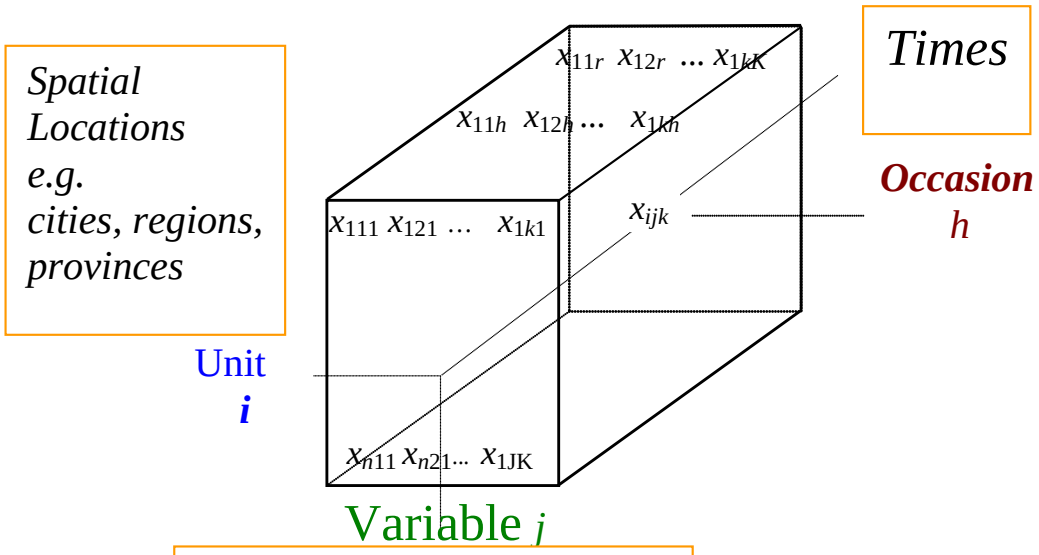
a set **X** of $n \times J \times T$ values related to:

J **variables** **measured** (**observed**, **estimated**) on

n **objects** (**individuals or aggregates with a spatial location**) at

T **occasions** (**times**, **locations**, different sources of data, etc.)

The set **X** is organized as a 3-Way Array



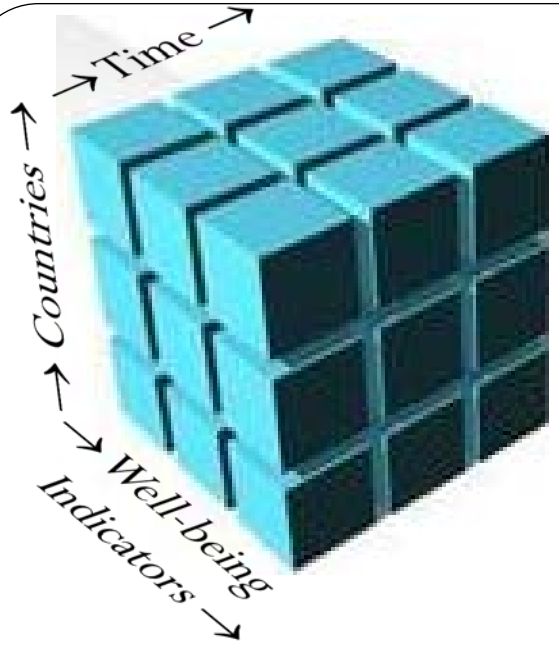
Variables (indicators) specifying a Multidimensional Phenomenon e.g., well-being indicators

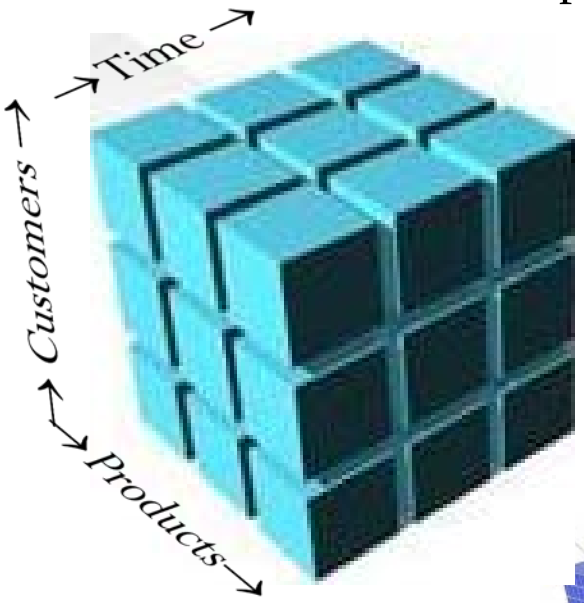
Examples of Data Cubes

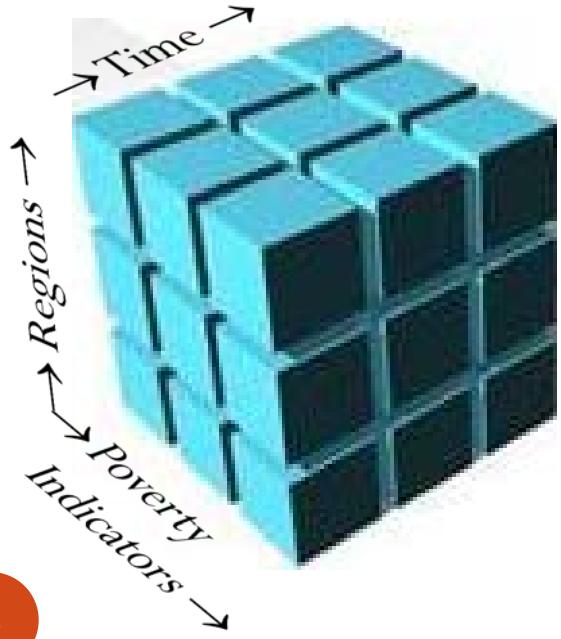
Countries × Well-being indicators × Time

Nuts2 regions × poverty indicators × Time

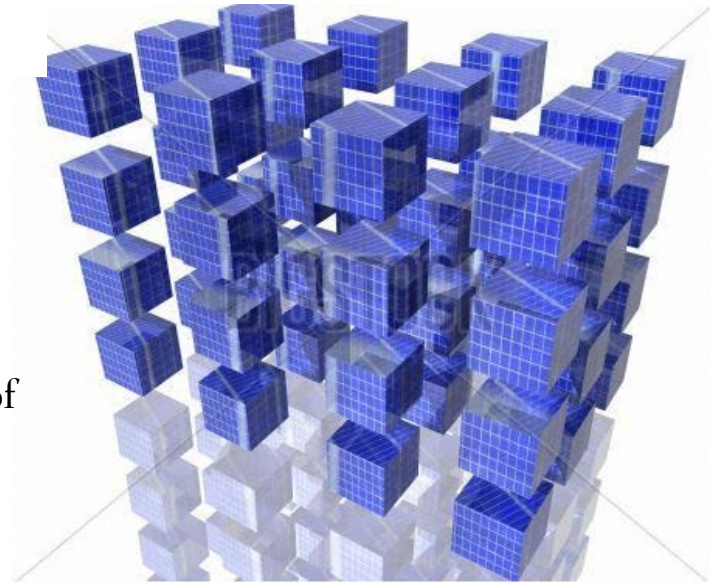
Customer × products × time







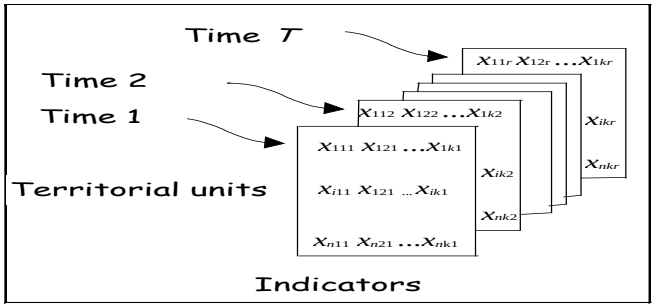
NSIs are Databases of Data Cubes



Cross-sectional, Spatial, and time series comparisons

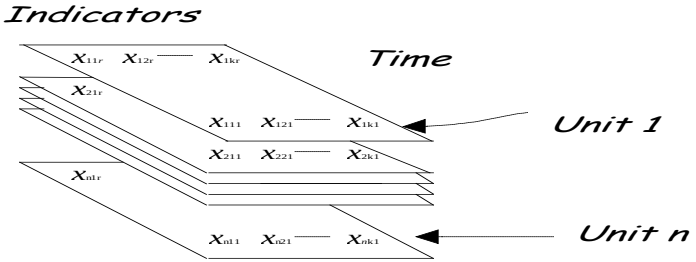
Spatial Comparison among multivariate territorial units in *T* times
(slicing the cube vertically)

First view of the Data Cube
T Cross-section Territorial comparisons



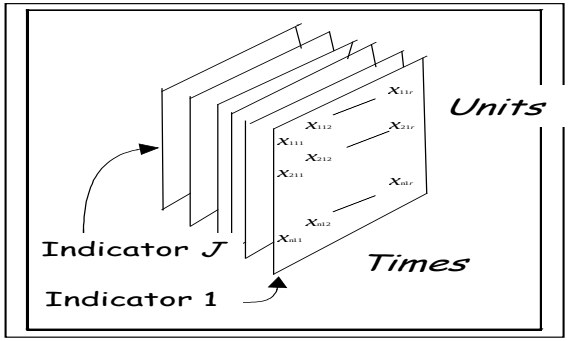
Time series comparison and statistical forecasting
(slicing the cube horizontly) *n* **Multiple time series**

Second view of the Data Cube
Time Series comparisons



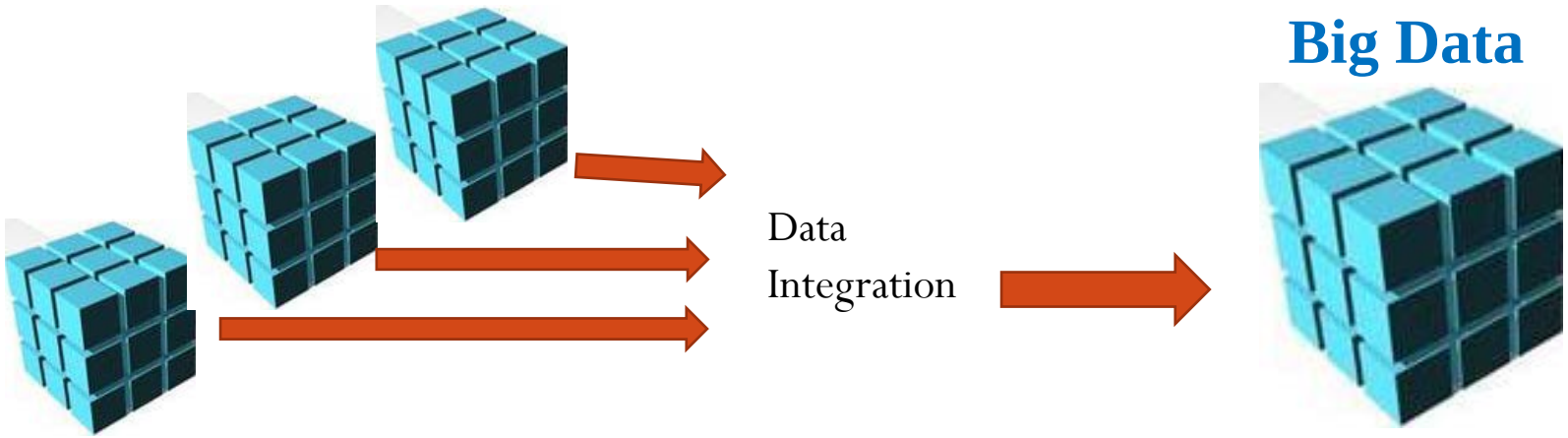
Indicators comparison
(slicing latelarly) A set of *K* **Multiple Time Series**

Third view of the Data Cube
Indicators comparisons



Large Data Cubes are Big Data for Official Statistics

How Big data are formed? Data coming from different sources, frequently automatically collected and reused to form a Data Cube



Scheveningen Memorandum (adopted by ESSC)

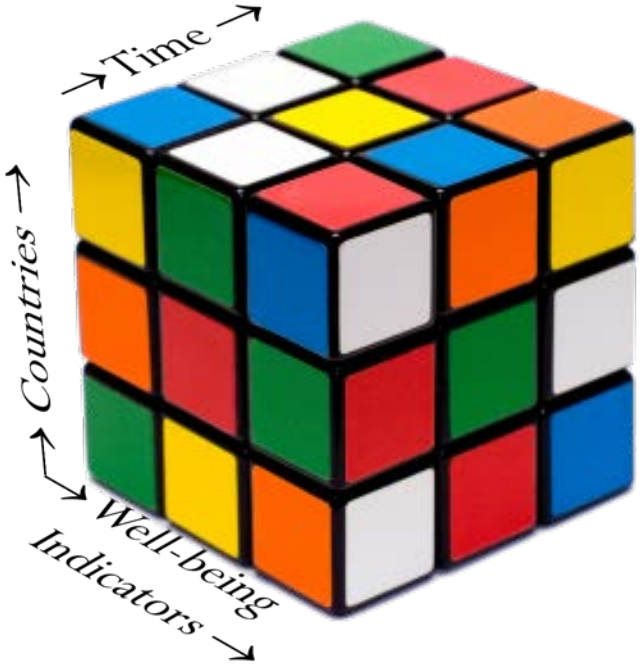
Get the Official Statistics on board on the BIG Data Era

- Big Data as new opportunities
- examine the potential of Big Data sources
- relevant action plan on
  - new methodology developments
  - skills and new competences
  - legislative action regarding the use

New methodologies for BIG DATA

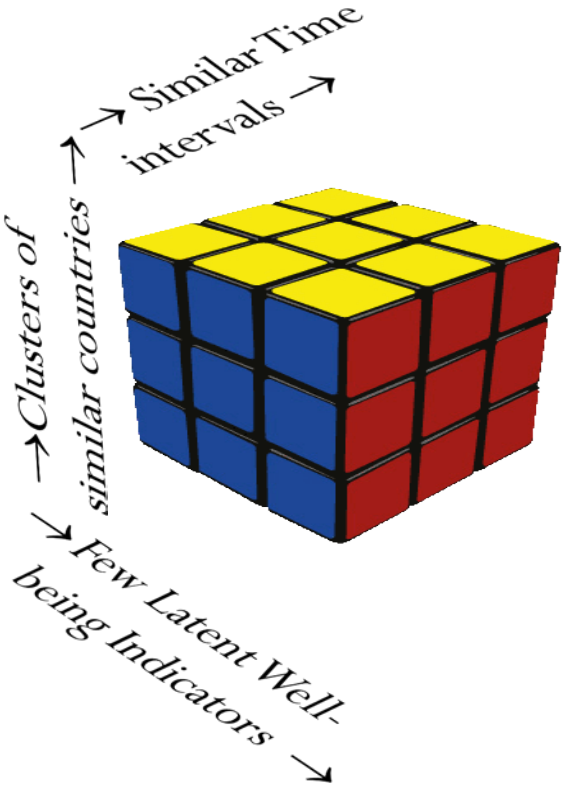
Data Reduction

From Big Data **to** Knowledge



From **Big Data**

Statistical Model



To the relevant **Knowledge**

Data Reduction

- Symmetrical & Asymmetrical mode reduction

Two approaches

1. Symmetric treatment of Units (Rows), Variables (Columns), Times (Tubes)

Clustering for Units
Clustering for Vars.
Clustering for times
Result: reduced set of *K* mean profiles for Units (Rows)
reduced set of *Q* mean profiles of Variables (Columns);
reduced set of *R* mean profiles of Occasions (Tubes);

OR

2. Asymmetric treatment of Units (Rows), Variables (Columns), occasions (Tubes)

Clustering for Units and times (periods)
Factorial methods for Variables
Result: reduced set of *K* mean profiles for Units and time periods
reduced set of Q components (factors) comp. indicators for variables

General model for three-way data

$$\mathbf{X} = \mathbf{AU}\overline{\mathbf{Y}}(\mathbf{W}'\mathbf{C} \otimes \mathbf{V}'\mathbf{B}) + \mathbf{E}$$

where

$\mathbf{X} = \mathbf{X}_{n,JH}$ $(n \times JH)$, $[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_H]$, three-way data matrix obtained by placing side by side $\mathbf{X}_h$;

$\mathbf{E} = \mathbf{E}_{n,JH} = [e_{ijk}] (n \times JH)$, $[\mathbf{E}_1, \ldots, \mathbf{E}_H]$; three-way error matrix

$\overline{\mathbf{Y}} = \overline{\mathbf{Y}}_{K,QR} = [\bar{x}_{kq}] (K \times QR)$, $[\overline{\mathbf{X}}_1, \ldots, \overline{\mathbf{X}}_R]$ three-way centroid matrix

$\mathbf{U} = [u_{ik}]$ $(n \times K)$ binary and row stochastic matrix for the partition of objects into K clusters, with $u_{ik}=1$ if the i^{th} object belongs to cluster k, $u_{ik}=0$ otherwise;

$\mathbf{V} = [v_{jq}]$ $(J \times Q)$ binary and row stochastic matrix for the partition of variables into Q clusters, with $v_{jq}=1$ if the j^{th} variable belongs to q^{th} cluster, $v_{jq}=0$, otherwise;

$\mathbf{W} = [w_{hr}]$ $(H \times R)$ binary matrix defining a partition of occasions into R clusters, with $w_{hr}=1$ if the h^{th} occasions belongs to r^{th} cluster, $w_{hr}=0$, otherwise;

Matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are diagonal matrices for objects, variables and occasions;

$\mathbf{A}=\mathrm{dg}(a_1,\ldots,a_n)$ $(n \times n)$ diagonal matrix for objects, $\sum_{i=1}^{n} u_{ik}\, a_i^2 = 1; \sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}\, a_{ik}^2 = K;$

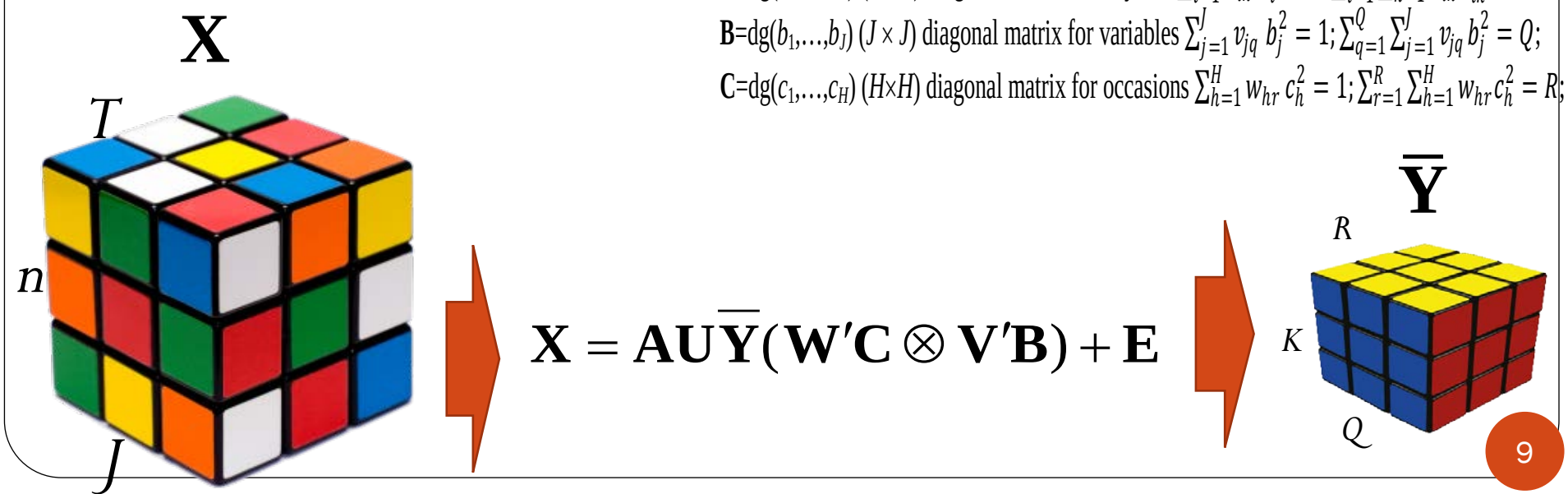
$\mathbf{B}=\mathrm{dg}(b_1,\ldots,b_J)$ $(J \times J)$ diagonal matrix for variables $\sum_{j=1}^{J} v_{jq}\, b_j^2 = 1; \sum_{q=1}^{Q}\sum_{j=1}^{J} v_{jq}\, b_j^2 = Q;$

$\mathbf{C}=\mathrm{dg}(c_1,\ldots,c_H)$ $(H \times H)$ diagonal matrix for occasions $\sum_{h=1}^{H} w_{hr}\, c_h^2 = 1; \sum_{r=1}^{R}\sum_{h=1}^{H} w_{hr}\, c_h^2 = R;$

subject to

U, **V**, **W** binary and row stochastic, membership matrices UNITS, VARS, TIMES; **A**, **B**, **C** diagonal and s.t.

$$\mathbf{u}_k'\mathbf{A}'\mathbf{A}\mathbf{u}_k = 1,\ (k=1,\ldots,K);$$
$$\mathbf{v}_q'\mathbf{B}'\mathbf{B}\mathbf{v}_q = 1,\ (q=1,\ldots,Q);$$
$$\mathbf{w}_r'\mathbf{C}'\mathbf{C}\mathbf{w}_r = 1,\ (r=1,\ldots,R);$$



- **Some special cases for two-mode data**
  - **Least-Square estimation**
    - **Generalized Double *K*-Means (GDKM)**
    - **Clustering & Disjoint PCA (CDPCA)**
    - **Clustering & Orthogonal Disjoint PCA (CDPCA)**
  - **Maximum Likelihood estimation**
    - **ML DKM**

(*T*=1, C=1) *Generalized Double K-Means (GDKM)*

Symmetric reduction of rows and columns of the data matrix **X**
Classes of objects and variables are summarized by components (latent objects and latent variables)

$$\mathbf{X} = \mathbf{A}\mathbf{U}\overline{\mathbf{Y}}(1 \otimes \mathbf{V}'\mathbf{B}) + \mathbf{E}$$

i.e., the *Generalized Double K-means* (GDKM) model for asymmetrical single partitioning

$$\mathbf{X} = \mathbf{A}\mathbf{U}\overline{\mathbf{Y}}\mathbf{V}'\mathbf{B} + \mathbf{E}$$

Subject to

U, **V** binary and row stochastic

$\mathbf{u}_k'\mathbf{A}'\mathbf{A}\mathbf{u}_k = 1$, for k=1,…,K;

$\mathbf{v}_q'\mathbf{B}'\mathbf{B}\mathbf{v}_q = 1$, for q=1,…,Q.

Clustering and Disjoint Principal Component Analysis

(Vichi, Saporta, 2009 CSDA)

Disjoint PCA (Vichi, Saporta 2009)

$\mathbf{X} = \mathbf{YA}' + \mathbf{E}_1$ with $\mathbf{Y}=\mathbf{XA}$

subject to

(1) $\sum_{j=1}^{J} a_{jk}^2 = 1$ $(k=1, \ldots, K)$

(2) $\sum_{j=1}^{J} (a_{jk} a_{jr})^2 = 0$ $(k=1,\ldots,K\text{-}1;\ r=k+1,\ldots,K)$

(3) $\sum_{k=1}^{K} a_{jk}^2 > 0$, $(j = 1,\ldots,n)$

LS estimation Y=XA

$min \|\mathbf{X} - \mathbf{YA}'\|^2 = \max tr(\mathbf{\Sigma_Y})$
subject to (1), (2) and (3)

reparameterizing **A=BV**

$\max tr(\mathbf{\Sigma_Y}) = \max tr(\mathbf{V'B\Sigma_X BV})$
subject to

$\mathbf{V} = [v_{jk} \in \{0,1\}]$	(binary)
$\mathbf{V1}_K = \mathbf{1}_J$	(row stochastic)
$\mathbf{B} = diag(b_1, \ldots, b_J)$	(diagonal)
$\mathbf{V'BBV} = \mathbf{I}_K$	(orthogonal)

K-means

$\mathbf{Y} = \mathbf{U\bar{Y}} + \mathbf{E}_2$

subject to

$\mathbf{U} = [u_{ip} \in \{0,1\}]$	(binary)
$\mathbf{U1}_P = \mathbf{1}_n$	(row stochastic)

Clustering & Disjoint PCA

$\mathbf{X} = (\mathbf{U\bar{Y}} + \mathbf{E}_2)\mathbf{A}' + \mathbf{E}_1$

rewriting $\mathbf{E} = \mathbf{E}_2\mathbf{A}' + \mathbf{E}_1$ and being $\mathbf{A}=\mathbf{BV}$

$\mathbf{X} = \mathbf{U\bar{Y}V'B} + \mathbf{E}$

subject to

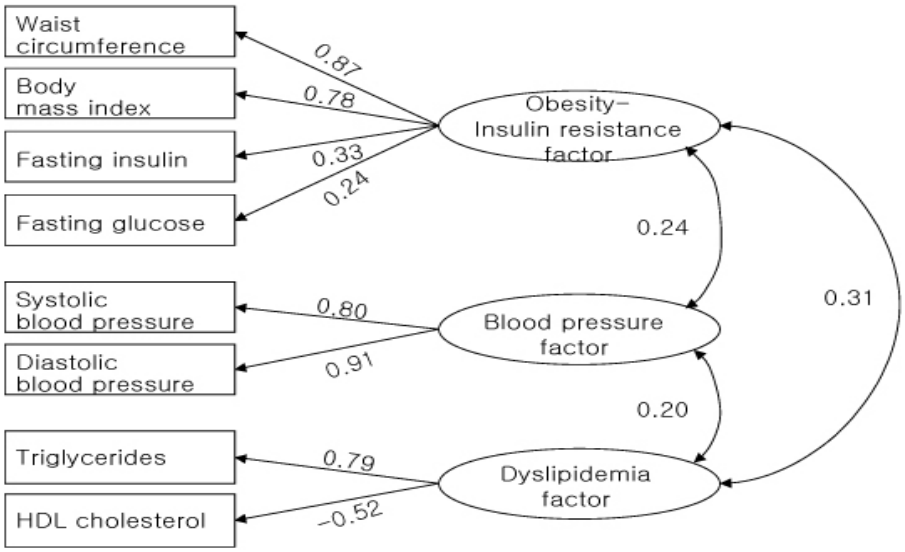
$\mathbf{U} = [u_{ip} \in \{0,1\}]$	(binary)
$\mathbf{U1}_P = \mathbf{1}_n$	(row stochastic)
$\mathbf{V} = [v_{jk} \in \{0,1\}]$	(binary)
$\mathbf{V1}_K = \mathbf{1}_J$	(row stochastic)
$\mathbf{B} = diag(b_1, \ldots, b_J)$	(diagonal)
$\mathbf{V'BBV} = \mathbf{I}_K$	(orthogonal)

Simplest and most sparse loading matrix A

- The loading matrix has only *J* non null values of the *JK*;
- Each manifest variable loads only to a single factor;
- Variables are partitioned into K classes;
- Each factor is linear combination of some manifest variables

A

	F1 Obesity	F2 Blod press	F3 Dyslipi-demia
Waist circumference	0.87	0	0
Body mass index	0.78	0	0
Fasting insulin	0.33	0	0
Fasting glucose	0.24	0	0
Systolic blood pressure	0	0.80	0
Dyastolic blood pressure	0	0.91	0
Triglycerides	0	0	0.79
HDL cholesterol	0	0	-0.52



As we have seen this is the most used sparse matrix in Confirmatory Factor Analysis

Short Term Indicators and Economic Performance Indicators ($\mathbf{X}$: 20 × 6)

20 Countries: Australia (A-lia), Canada (Can), Finland (Fin), France (Fra), Spain (Spa), Sweden (Swe), United States (USA), Netherlands (Net), Greece (Gre), Mexico (Mex), Portugal (Por), Austria (A-tria), Belgium (Bel), Denmark (Den), Germany (Ger), Italy (Ita), Japan (Jap), Norway (Nor), Switzerland (Swi), United Kingdom (UK)

6 Macro Eco. VARS: Gross Domestic Product (GDP), Leading Indicator (LI), Unemployment Rate (UR), Interest Rate (IR), Trade Balance (TB), Net National Savings (NNS)

3 Prototype Countries × 2 Factors ($\bar{\mathbf{Y}}$: 3 × 2)

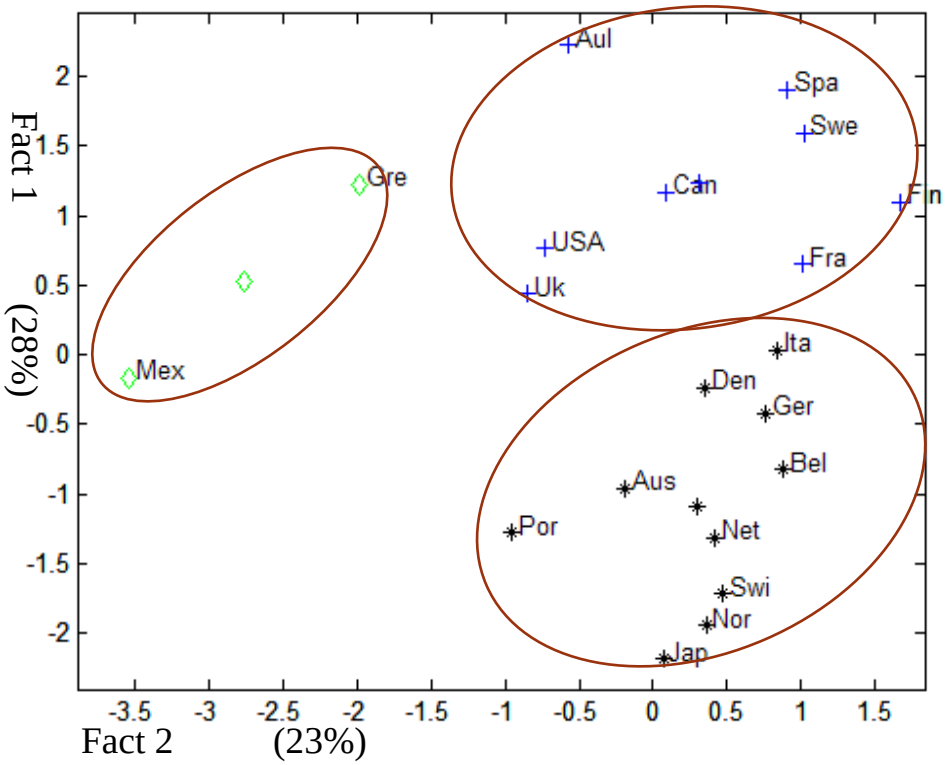
Component loadings of PCA

	GDP	IR	LI	UR	NNS	TB
Fact2	-0.065	**-0.696**	**-0.229**	**0.367**	-0.092	**0.563**
Fact1	**-0.567**	-0.175	**-0.192**	**-0.489**	**0.607**	0.059

Var(Dim1) =1.6531, Var(Dim2) = 1.3680

51% Variance explained

K-MEANS on the two components



Short Term Indicators and Economic Performance Indicators

20 Countries: Australia (A-lia), Canada (Can), Finland (Fin), France (Fra), Spain (Spa), Sweden (Swe), United States (USA), Netherlands (Net), Greece (Gre), Mexico (Mex), Portugal (Por), Austria (A-tria), Belgium (Bel), Denmark (Den), Germany (Ger), Italy (Ita), Japan (Jap), Norway (Nor), Switzerland (Swi), United Kingdom (UK)

6 Macro Eco. VARS: Gross Domestic Product (GDP), Leading Indicator (LI), Unemployment Rate (UR), Interest Rate (IR), Trade Balance (TB), Net National Savings (NNS)

Loadings of Clustering and Disjoint PCA

	GDP	IR	LI	UR	NNS	TB
Fact2	0	**-0.697**	**-0.229**	0	0	**0.679**
Fact1	**-0.383**	0	0	**-0.498**	**0.778**	0

Var(Fact1) =1.5601, Var(Fact2) = 1.2553

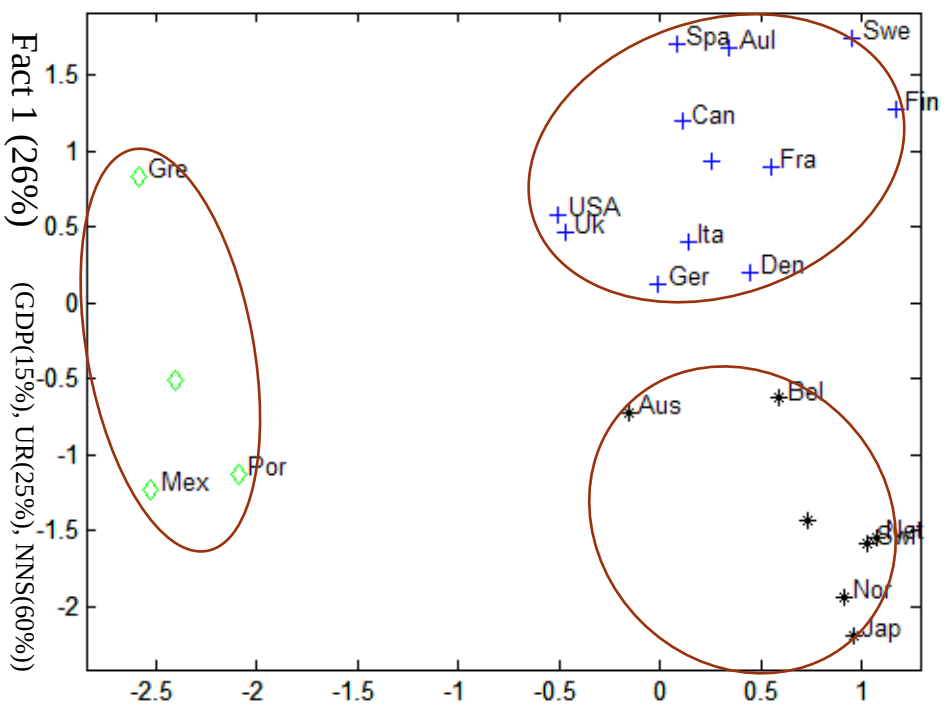
47% Variance explained

Component loadings of PCA

	GDP	IR	LI	UR	NNS	TB
Fact2	-0.065	**-0.696**	**-0.229**	**0.367**	-0.092	**0.563**
Fact1	**-0.567**	-0.175	**-0.192**	**-0.489**	**0.607**	0.059

Var(Fact1) =1.6531, Var(Fact2) = 1.3680

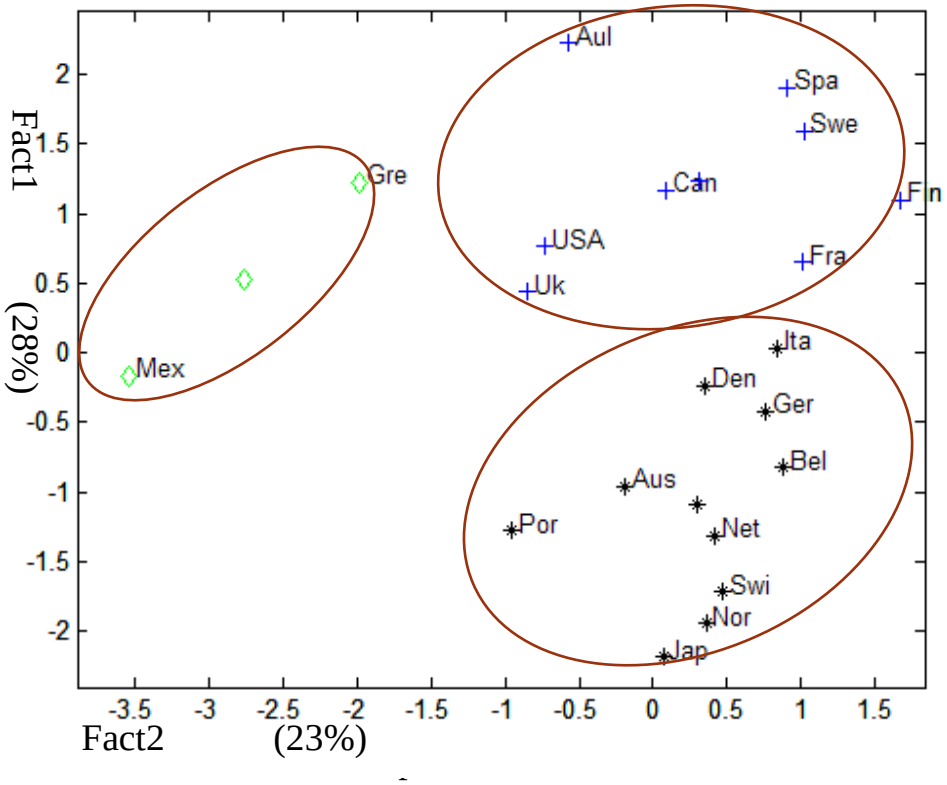
51% Variance explained

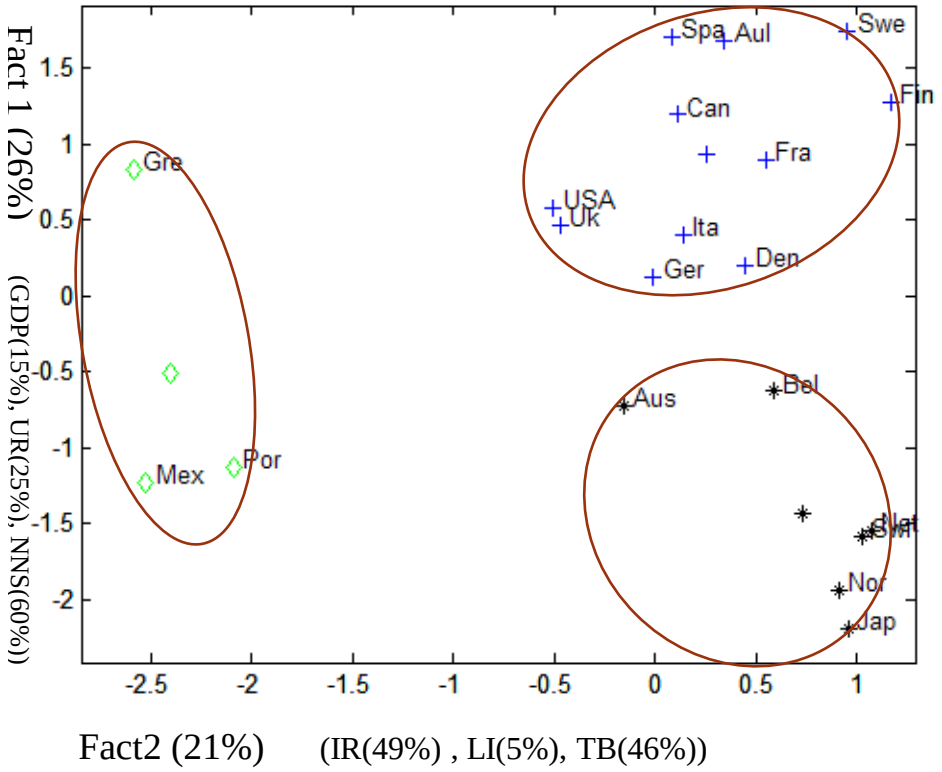


Short Term Indicators and Economic Performance Indicators

20 Countries: Australia (A-lia), Canada (Can), Finland (Fin), France (Fra), Spain (Spa), Sweden (Swe), United States (USA), Netherlands (Net), Greece (Gre), Mexico (Mex), Portugal (Por), Austria (A-tria), Belgium (Bel), Denmark (Den), Germany (Ger), Italy (Ita), Japan (Jap), Norway (Nor), Switzerland (Swi), United Kingdom (UK)

6 Macro Eco. VARS: Gross Domestic Product (GDP), Leading Indicator (LI), Unemployment Rate (UR), Interest Rate (IR), Trade Balance (TB), Net National Savings (NNS)





Mex, Por, Gre are in the same class also in k-means on the original variables

Ita, Ger Den have almost equal NNS, GDP;

Orthogonal CDPCA

$$\mathbf{Y} = \mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{V}\mathbf{C}$$

C is an orthogonal matrix necessary to obtain orthogonal Components.
It is a **rotation matrix** used to perform a **rotation in Euclidean Space** .
It includes *improper rotations* (det(**C**)=-1), i.e., *rotation+reflection*

$$\mathbf{\Sigma_X} = \frac{1}{n}\mathbf{X}'\mathbf{X} = \mathbf{\Sigma_{X,B}} + \mathbf{\Sigma_{X,W}} = \frac{1}{n}\bar{\mathbf{X}}'\mathbf{U}'\mathbf{U}\bar{\mathbf{X}} + \frac{1}{n}(\mathbf{X} - \mathbf{U}\bar{\mathbf{X}})'(\mathbf{X} - \mathbf{U}\bar{\mathbf{X}})$$

Cov.Matrix = Betw.Cov.Matr+Wit.Cov.Matr

Problem to solve

$$\max tr(\mathbf{\Sigma_Y}) = \max tr(\frac{1}{n}\mathbf{C}'\mathbf{V}'\mathbf{B}\bar{\mathbf{X}}'\mathbf{U}'\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{V}\mathbf{C}) = \max tr(\mathbf{C}'\mathbf{V}'\mathbf{B}\mathbf{\Sigma_{X,B}}\mathbf{B}\mathbf{V}\mathbf{C})$$

subject to

$\mathbf{V} = [v_{jk} \in \{0,1\}$	$(j=1,\ldots,J;\ k=1,\ldots,K)$
$\mathbf{V}\mathbf{1}_K = \mathbf{1}_J$	(row stochastic)
$\mathbf{U} = [u_{ip} \in \{0,1\}$	$(i=1,\ldots,n;\ p=1,\ldots,P)$
$\mathbf{U}\mathbf{1}_P = \mathbf{1}_n$	(row stochastic)
$\mathbf{V}'\mathbf{B}\mathbf{B}\mathbf{V} = \mathbf{I}_K$	(columns orthonormal)
$\mathbf{C}'\mathbf{C} = \mathbf{I}_K$	(columns orthonormal)
$\mathbf{\Sigma_Y} = \text{diag}(\lambda_1,\ldots,\lambda_k)$.	(orthogonal PCs).

Holzinger-Swineford (1939) ability tests (X: 301 students × 9 ability tests)

Mental ability test scores on 301 students, 9 variables are considered identifying three typologies of tests spatial tests - verbal tests - speed tests .

The spatial tests consist of visual, cubes, paper, flags, …, etc.
The verbal tests consist of general comprehension, paragraph, sentence, word meaning.
The speed tests consist of addition, code, counting, and straight.

PCA REDUCTION ($\bar{\mathbf{Y}}$: 301 × 3)

PCA loading matrix

Explained variance of the components

F1	F2	F3	Total expl	Total
3.2163	1.6387	1.3652	6.2202 (69%)	9

T1-S- VISUAL PERCEPTION TEST FROM SPEARMAN VPT,	**0.3671**	0.0984	0.3174	
T2-S- CUBES SIMPLIFICATION (BRIGHAM'S TEST)	0.2173	0.0677	**0.5318**	spatial tests
T4-S- LOZENGES FROM THORNDIKE—SHAPES	0.2660	0.2574	**0.4644**	
T6-V- PARAGRAPH COMPREHENSION TEST	**0.4270**	-0.3488	-0.1450	
T7-V- SENTENCE COMPLETION TEST	**0.4113**	-0.3778	-0.1856	verbal tests
T9-V- WORD MEANING TEST	**0.4306**	-0.3350	-0.0972	
T10- SPEEDED ADDITION TEST	0.1945	0.3912	**-0.5059**	
T12 SPEEDED COUNTING OF DOTS IN SHAPE	0.2533	**0.4796**	-0.2823	speed tests
T13 SPEEDED DISCRIM STRAIGHT AND CURVED CAPS	0.3294	**0.3998**	-0.0169	

DPCA loading matrix

Explained variance of the components

F1	F2	F3	Total expl	Total
2.4385	1.8538	1.7221	6.0144(67%)	9

T1-S- VISUAL PERCEPTION TEST FROM SPEARMAN VPT,	0	0	**0.5900**
T2-S- CUBES SIMPLIFICATION (BRIGHAM'S TEST)	0	0	**0.5297**
T4-S- LOZENGES FROM THORNDIKE—SHAPES	0	0	**0.6094**
T6-V- PARAGRAPH COMPREHENSION TEST	**0.5772**	0	0
T7-V- SENTENCE COMPLETION TEST	**0.5813**	0	0
T9-V- WORD MEANING TEST	**0.5736**	0	0
T10- SPEEDED ADDITION TEST	0	**0.5684**	0
T12 SPEEDED COUNTING OF DOTS IN SHAPE	0	**0.6128**	0
T13 SPEEDED DISCRIM STRAIGHT AND CURVED CAPS	0	**0.5490**	0

LossVariance(PCA vs DPCA) = (6.2202 - 6.0144)/9 *100 = 2.29%

Correlation between factors (factors are weakly correlated)

	F1	F2	F3
F1	1.0000	0.2238	0.3166
F2	0.2238	1.0000	0.2805
F3	0.3166	0.2805	1.0000

Orthogonal DPCA loading matrix

Explained variance of the components

F1	F2	F3	Total expl.	Total
3.1742	1.6222	1.2380	6.0344 (67%)	9

T1 VISUAL PERCEPTION TEST FROM SPEARMAN VPT,	0.2904	0.1367	**0.4950**
T2 CUBES SIMPLIFICATION (BRIGHAM'S TEST)	0.2608	0.1227	**0.4444**
T4 LOZENGES FROM THORNDIKE—SHAPES	0.3000	0.1412	**0.5113**
T6 PARAGRAPH COMPREHENSION TEST	**0.4274**	-0.3567	-0.1523
T7 SENTENCE COMPLETION TEST	**0.4304**	-0.3593	-0.1534
T9 WORD MEANING TEST	**0.4248**	-0.3545	-0.1513
T10 SPEEDED ADDITION TEST	0.2600	**0.4270**	-0.2705
T12 SPEEDED COUNTING OF DOTS IN SHAPE	0.2803	**0.4603**	-0.2916
T13 SPEEDED DISCRIM STRAIGHT AND CURVED CAPS	0.2512	**0.4124**	-0.2613

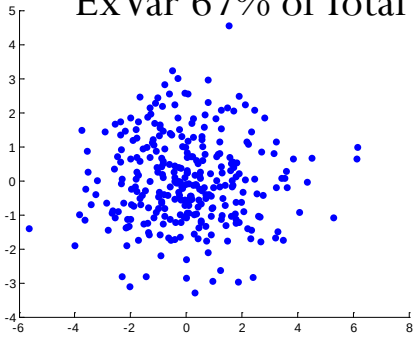
and the orthogonal rotation matrix C

0.7405	-0.6181	-0.2638
0.4923	0.2317	0.8390
0.4574	0.7512	-0.4758

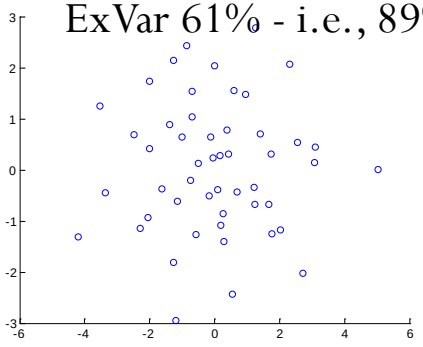
Clustering and DPCA

Different reductions of the original matrix Matrix (301 × 9)

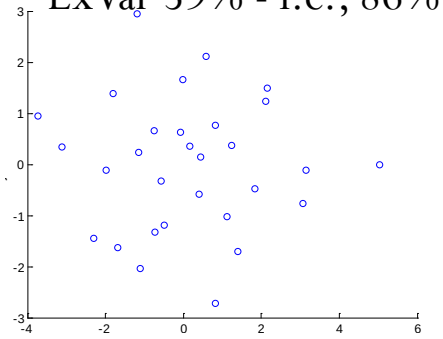
Matrix ($\overline{\mathbf{Y}}$: 301 × 3)
ExVar 67% of Total



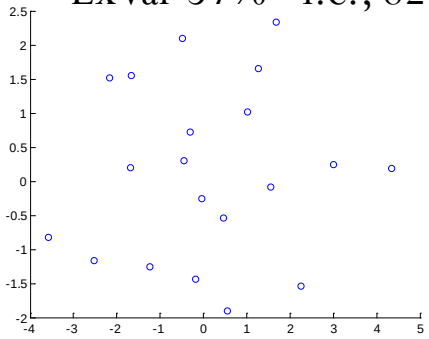
Matrix ($\overline{\mathbf{Y}}$: 50 × 3)
ExVar 61% - i.e., 89%



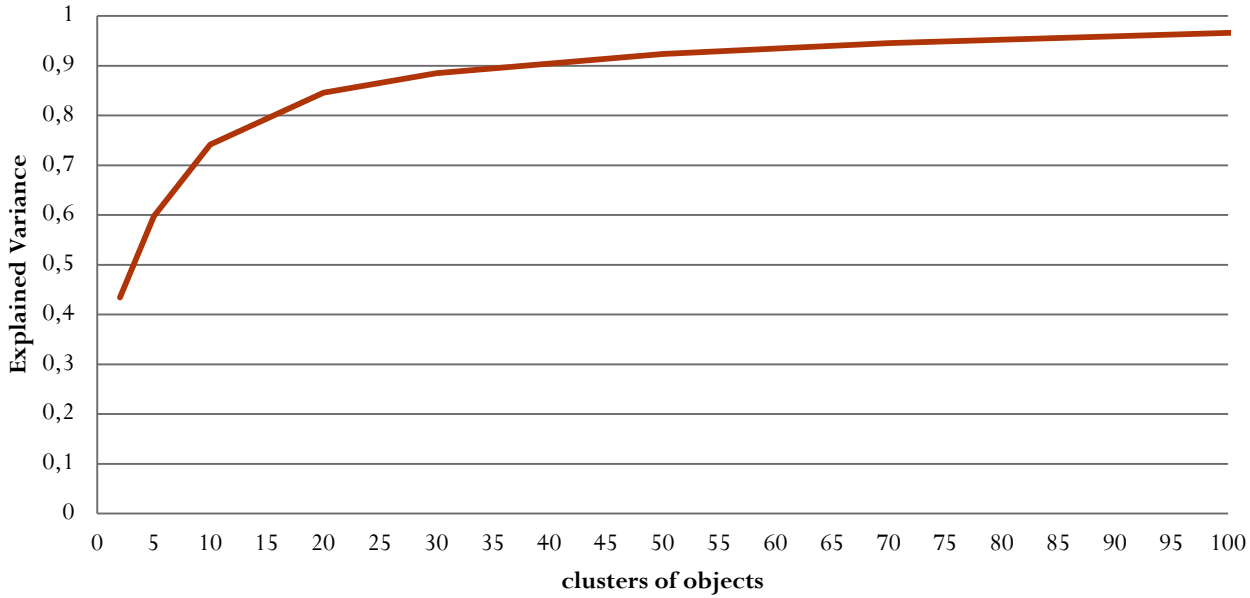
Matrix ($\overline{\mathbf{Y}}$: 30 × 3)
ExVar 59% - i.e., 86%



Matrix ($\overline{\mathbf{Y}}$: 20 × 3)
ExVar 57% - i.e., 82%



Variance of components explained by clusters



Maximum likelihood estimation of the DKM

Let us consider the double *K*-means model specified in row form

$$\mathbf{x}_i = \mathbf{V}\overline{\mathbf{X}}'\mathbf{u}_i + \mathbf{e}_i \qquad i = 1, \ldots, n,$$

Assumptions: **Homoscedasticity** var($\mathbf{x}_i$)=$\boldsymbol{\Sigma}$ (for parsimony reasons)
Multinormality given **U** and **V**

$$f\left(\mathbf{x}_i, \mathbf{U}, \mathbf{V}, \overline{\mathbf{X}}, \boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{(2\pi)^J}} \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}_i - \mathbf{V}\overline{\mathbf{X}}'\mathbf{u}_i\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i - \mathbf{V}\overline{\mathbf{X}}'\mathbf{u}_i\right)\right\}$$

Let ($\mathbf{x}_1$, $\mathbf{x}_2$, . . . , $\mathbf{x}_n$) be a sample of i.i.d. *J*-dimensional observations drawn from the density $f\left(\mathbf{x}_i, \mathbf{U}, \mathbf{V}, \overline{\mathbf{X}}, \boldsymbol{\Sigma}\right)$ the corresponding likelihood function for fixed **U** and **V** matrices is given by

$$L\left(\mathbf{U}, \mathbf{V}, \overline{\mathbf{X}}, \boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{(2\pi)^{nJ}}} \left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{n}{2}} \exp\left\{-\frac{1}{2} tr\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{X} - \mathbf{U}\overline{\mathbf{X}}\mathbf{V}'\right)'\left(\mathbf{X} - \mathbf{U}\overline{\mathbf{X}}\mathbf{V}'\right)\right]\right\},$$

Thus

$$\ln L\left(\mathbf{U}, \mathbf{V}, \overline{\mathbf{X}}, \boldsymbol{\Sigma}\right) = -nJ \ln\sqrt{2\pi} + \frac{n}{2}\ln\left|\boldsymbol{\Sigma}^{-1}\right| - \frac{1}{2} tr\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{X} - \mathbf{U}\overline{\mathbf{X}}\mathbf{V}'\right)'\left(\mathbf{X} - \mathbf{U}\overline{\mathbf{X}}\mathbf{V}'\right)\right] \rightarrow \max_{\mathbf{U}, \mathbf{V}, \overline{\mathbf{X}}, \boldsymbol{\Sigma}}$$

Subject to

$\mathbf{V} = [v_{jk} \in \{0,1\} (j=1,\ldots,J;\ k=1,\ldots,K)$
$\mathbf{V}\mathbf{1}_K = \mathbf{1}_J$ (row stochastic)
$\mathbf{U} = [u_{ik} \in \{0,1\}$ $(i=1,\ldots,n;\ k=1,\ldots,K)$
$\mathbf{U}\mathbf{1}_K = \mathbf{1}_n$ (row stochastic)

Maximum likelihood estimation of the DKM

$$\ln L(\mathbf{U},\mathbf{V},\bar{\mathbf{X}},\boldsymbol{\Sigma}) = -nJ\ln\sqrt{2\pi} + \frac{n}{2}\ln\left|\boldsymbol{\Sigma}^{-1}\right| - \frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{U}\bar{\mathbf{X}}\mathbf{V}')'(\mathbf{X}-\mathbf{U}\bar{\mathbf{X}}\mathbf{V}')\right] \rightarrow \max_{\mathbf{U},\mathbf{V},\bar{\mathbf{X}},\boldsymbol{\Sigma}}$$

DKM estimated with a Mahalanobis metrics

$$||\mathbf{X}-\mathbf{U}\bar{\mathbf{X}}\mathbf{V}'||^2_{\boldsymbol{\Sigma}^{-1}} = tr\left[\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{U}\bar{\mathbf{X}}\mathbf{V}')'(\mathbf{X}-\mathbf{U}\bar{\mathbf{X}}\mathbf{V}')\right] \rightarrow \min_{\mathbf{U},\bar{\mathbf{X}},\mathbf{V},\boldsymbol{\Sigma}}$$

or

$$||\mathbf{U}\bar{\mathbf{X}}\mathbf{V}'||^2_{\boldsymbol{\Sigma}^{-1}} = tr\left[\boldsymbol{\Sigma}^{-1}(\mathbf{U}\bar{\mathbf{X}}\mathbf{V}')'(\mathbf{U}\bar{\mathbf{X}}\mathbf{V}')\right] \rightarrow \max_{\mathbf{U},\bar{\mathbf{X}},\mathbf{V},\boldsymbol{\Sigma}}$$

Subject to

$\mathbf{V} = [v_{jk} \in \{0,1\} (j=1,\ldots,J;\ k=1,\ldots,K)$
$\mathbf{V}\mathbf{1}_K = \mathbf{1}_J$ (row stochastic)
$\mathbf{U} = [u_{ik} \in \{0,1\}$ $(i=1,\ldots,n;\ k=1,\ldots,K)$
$\mathbf{U}\mathbf{1}_K = \mathbf{1}_n$ (row stochastic)

Coordinate Ascent Algorithm

Updating $\bar{\mathbf{X}}$:

When **U**, **V** and **Σ** are fixed, the ML estimate of $\bar{\mathbf{X}}$ is given by solving the likelihood equations

$$\frac{\partial \ln L}{\partial \bar{\mathbf{X}}} \propto \mathbf{U}'\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{V} - \mathbf{U}'\mathbf{U}\bar{\mathbf{X}}\mathbf{V}'\boldsymbol{\Sigma}^{-1}\mathbf{V} = 0,$$

which leads to the ML estimator of $\bar{\mathbf{X}}$ is given by

$$\hat{\bar{\mathbf{X}}} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{V}(\mathbf{V}'\boldsymbol{\Sigma}^{-1}\mathbf{V})^{-1},$$

Updating Σ:

When **U**, **V** and **Σ** are fixed, the ML estimate of $\bar{\mathbf{X}}$ is given by solving the likelihood equations

$$\frac{\partial \ln L}{\partial \boldsymbol{\Sigma}^{-1}} \propto \frac{n}{2}\boldsymbol{\Sigma} - \frac{1}{2}\left(\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{V}'\right)'\left(\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{V}'\right) = 0$$

from which we have

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\left(\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{V}'\right)'\left(\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{V}'\right).$$

Updating U:

When **V**, $\bar{\mathbf{X}}$ and **Σ** are fixed, for each $i = 1, \ldots, n$, let

$$u_{ik} = \begin{cases} 1 \; if \; lnL(\cdot, u_{ik} = 1) = \max\{lnL(\cdot, u_{ih} = 1) : h = 1, \ldots, K\} \\ 0 \quad otherwise \end{cases}$$

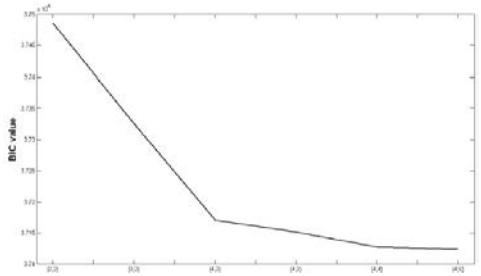
Updating V:

When **U**, $\bar{\mathbf{X}}$ and **Σ** are fixed, for each $j = 1, \ldots, J$, let

$$v_{jq} = \begin{cases} 1 \; if \; lnL(\cdot, v_{jq} = 1) = \max\{lnL(\cdot, v_{jp} = 1) : p = 1, \ldots, J\} \\ 0 \quad otherwise \end{cases}$$

Application cutaneous melanoma Bittner *et al.*, 2000

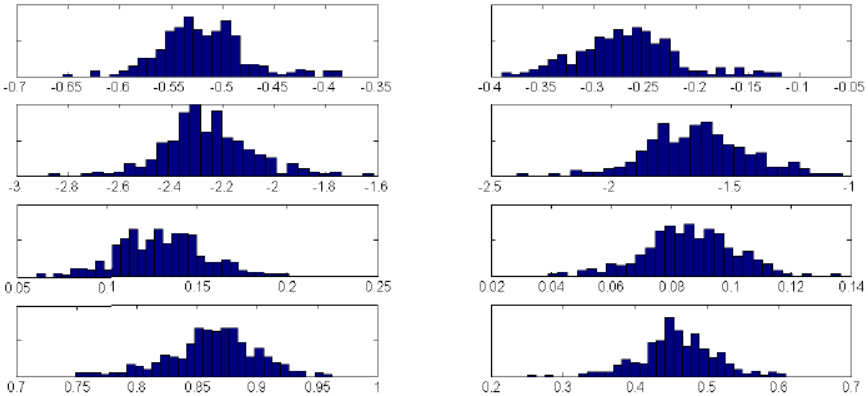
31 samples of cutaneous melanomas x 3,613 genes. Bitner *et al.* obtained two clusters of 10 and 21 samples,
It was originally analyzed to determine whether or not molecular gene profiles could be used to identify distinct subtypes of cutaneous melanoma.

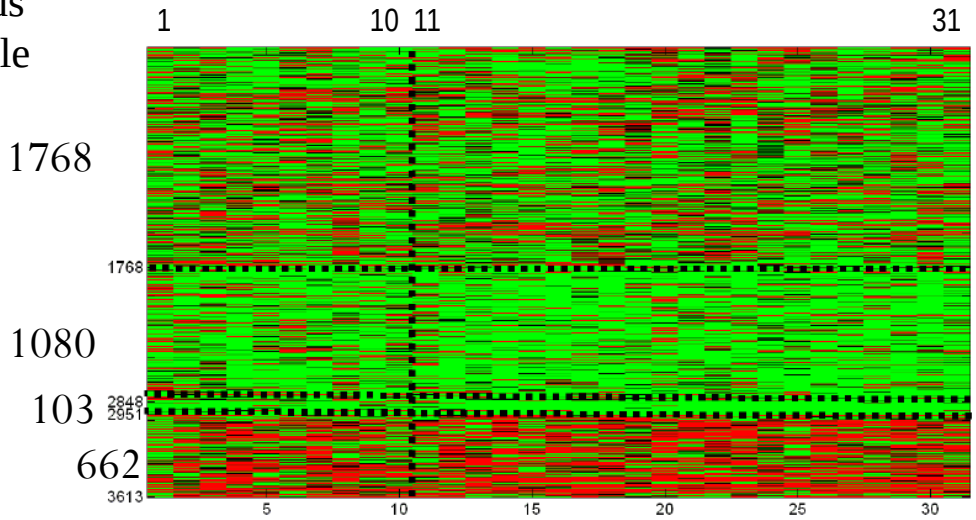


The partition K=4, Q=2 identifies the elbow in the curve of BIC

4 × 2 *blocks partitioning* of *Double K-Means*

Empirical sampling distributions of block centroids generated by 300 samples of the Stratified resample procedure to compute confidence intervals





Gene cluster size	Mean tissue cluster 1	Mean tissue cluster 2	Mean F	95% CI mean tissue cluster 1	95% CI mean tissue cluster 2
1768	0.12	0.08	10.06	(0.053, 0.175)	(0.053, 0.112)
1080	-0.52	-0.27	20.65	(-0.593, -0.416)	(-0.328, -0.152)
103	-2.20	-1.57	105.09	(-2.582, -1.885)	(-2.063, -1.223)
662	0.87	0.47	40.10	(0.778, 0.933)	(0.348, 0.571)

References

- Martella F., Alfò M., Vichi M. (2010). Hierarchical mixture models for biclustering in microarray, *Statistical Modelling*.
- Rocci R., Gattone A., Vichi M. (2011). A new Dimension Reduction Method: Factor Discriminant *K*-means, *Journal of Classification*.
- Vichi M., Rocci R. (2008). Two-mode Multi-partitioning. *Computational Statistics & Data Analysis*, vol. 52, pp. 1984-2003 ISSN: 0167-9473.
- Vicari D., Vichi M. (2012). On Multivariate Linear Regression for Heterogeneous Data, Submitted.
- Vichi M., Saporta G. (2009). Clustering and Disjoint Principal Component Analysis. *Computational Statistics & Data Analysis* vol. 53; p. 3194-3208, ISSN: 0167-9473, doi: 10.1016/j.csda.2008.05.028
- Vichi M. (2008) Fitting Semiparametric Clustering Models to Dissimilarity Data, *Advances in Data Analysis and Classification*, vol. 2, 2, 121-161.
- Vichi M. (2012) Fitting Hierarchical Clustering Models to Dissimilarity Data, Submitted.