Factorial Methods

K. Gibert [1,2]

[1] *Department of Statistics and Operation Research*

[2] *Knowledge Engineering and Machine Learning group*
Universitat Politècnica de Catalunya, Barcelona

Master Oficial en Enginyeria Informàtica
Universitat Politècnica de Catalunya

Factorial Methods

- Find the isomorph transformation from original space

  keeps the adjacency relationships among variables

- Results expressed in a ficticious space

- Might produce interpretation problems

- Methods
  - PCA (Principal components analysis)
  - Simple correspondence analysis
  - Multiple correspondence analysis

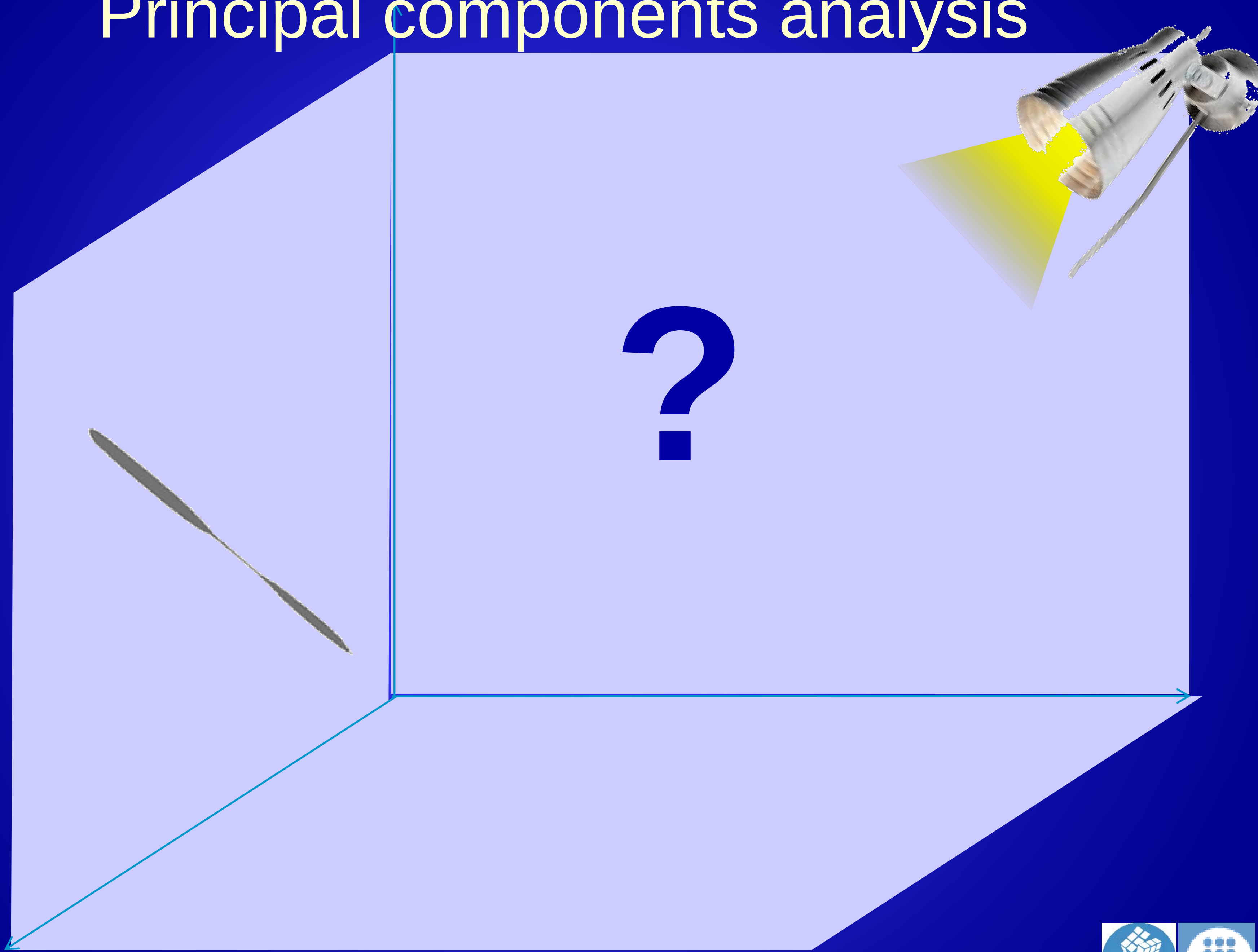
Factorial Methods

- Principal Components Analysis
  - Only numerical variables
  - Find the most informative projection planes

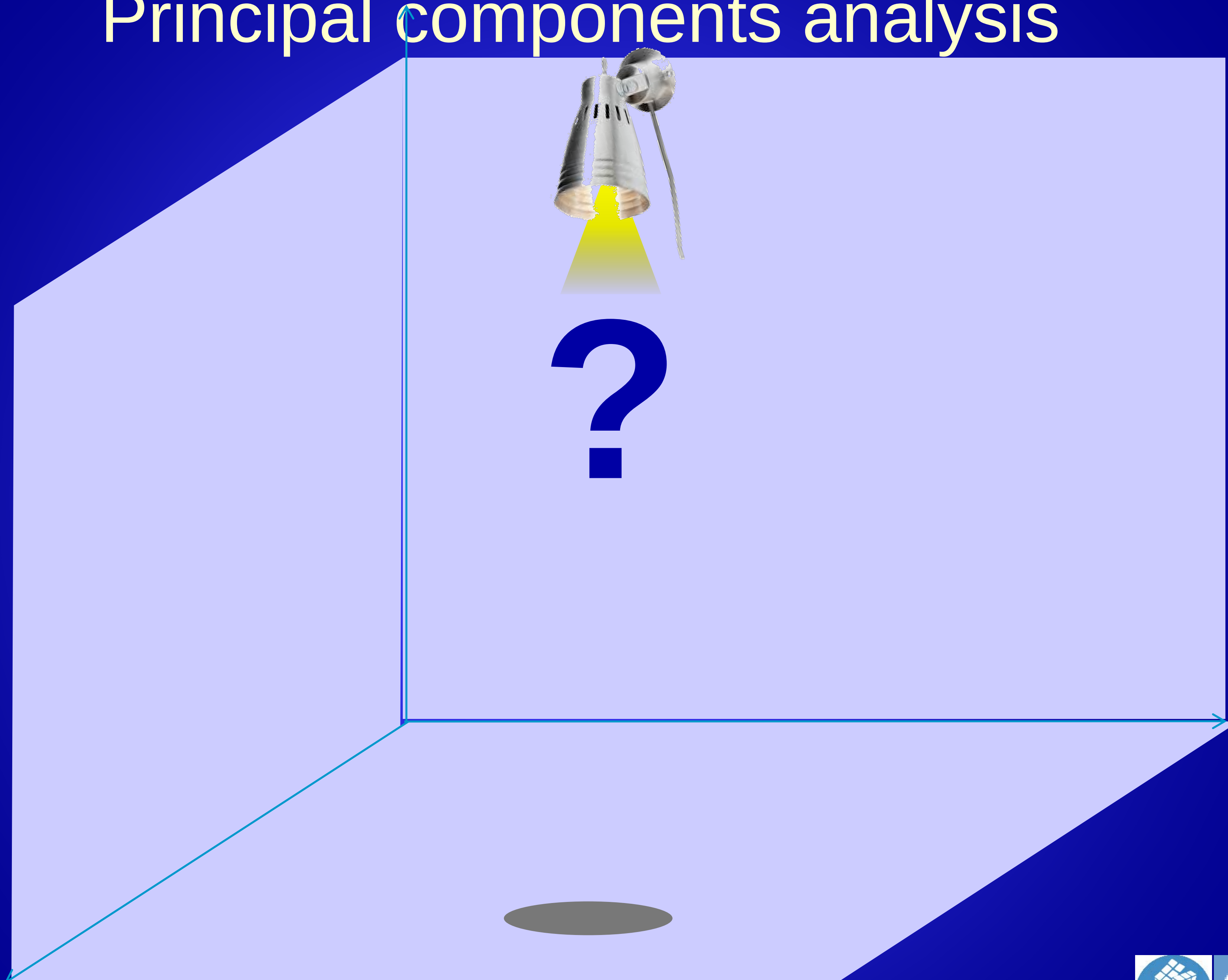
    (factorial planes)

Example “Copas”

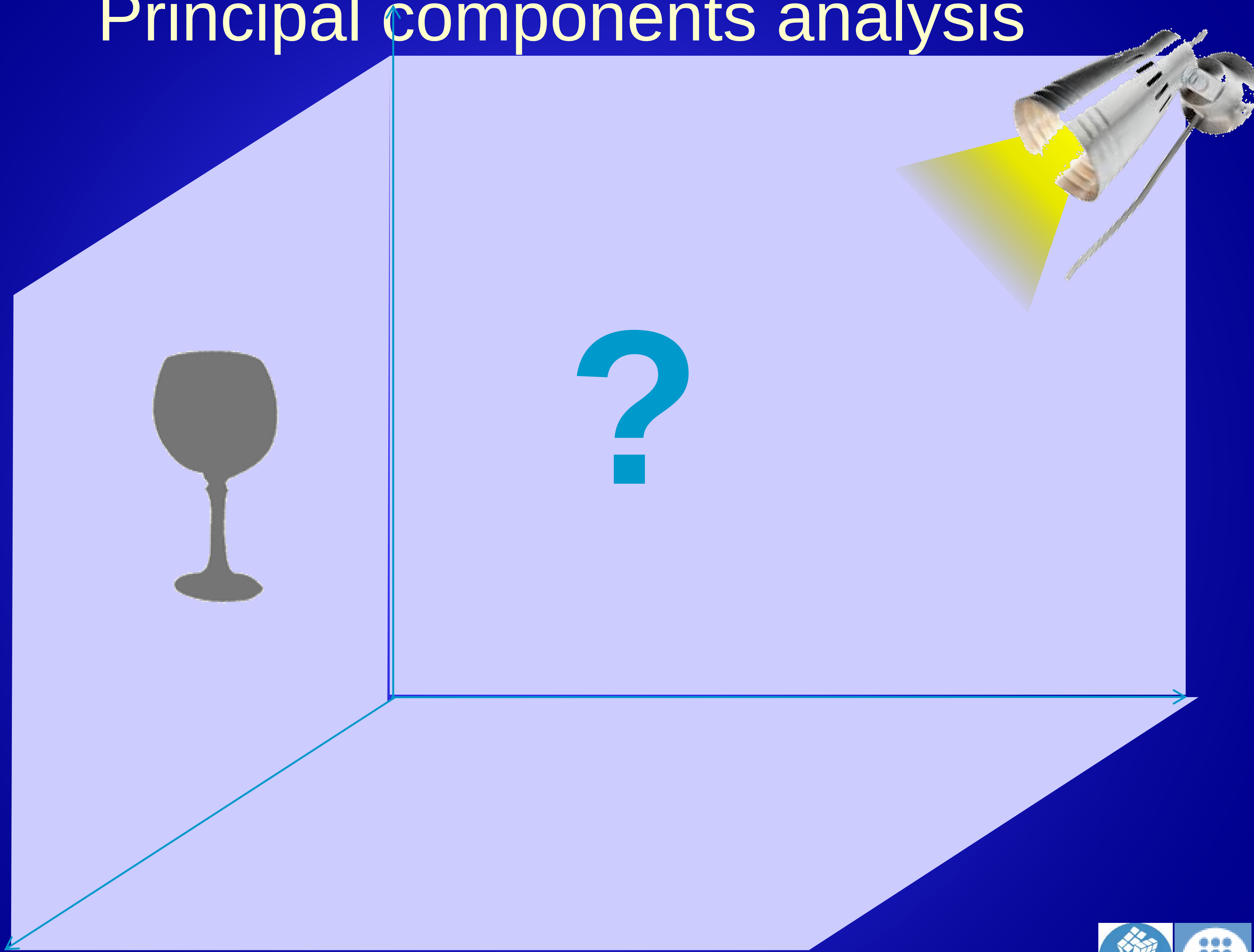
Principal components analysis



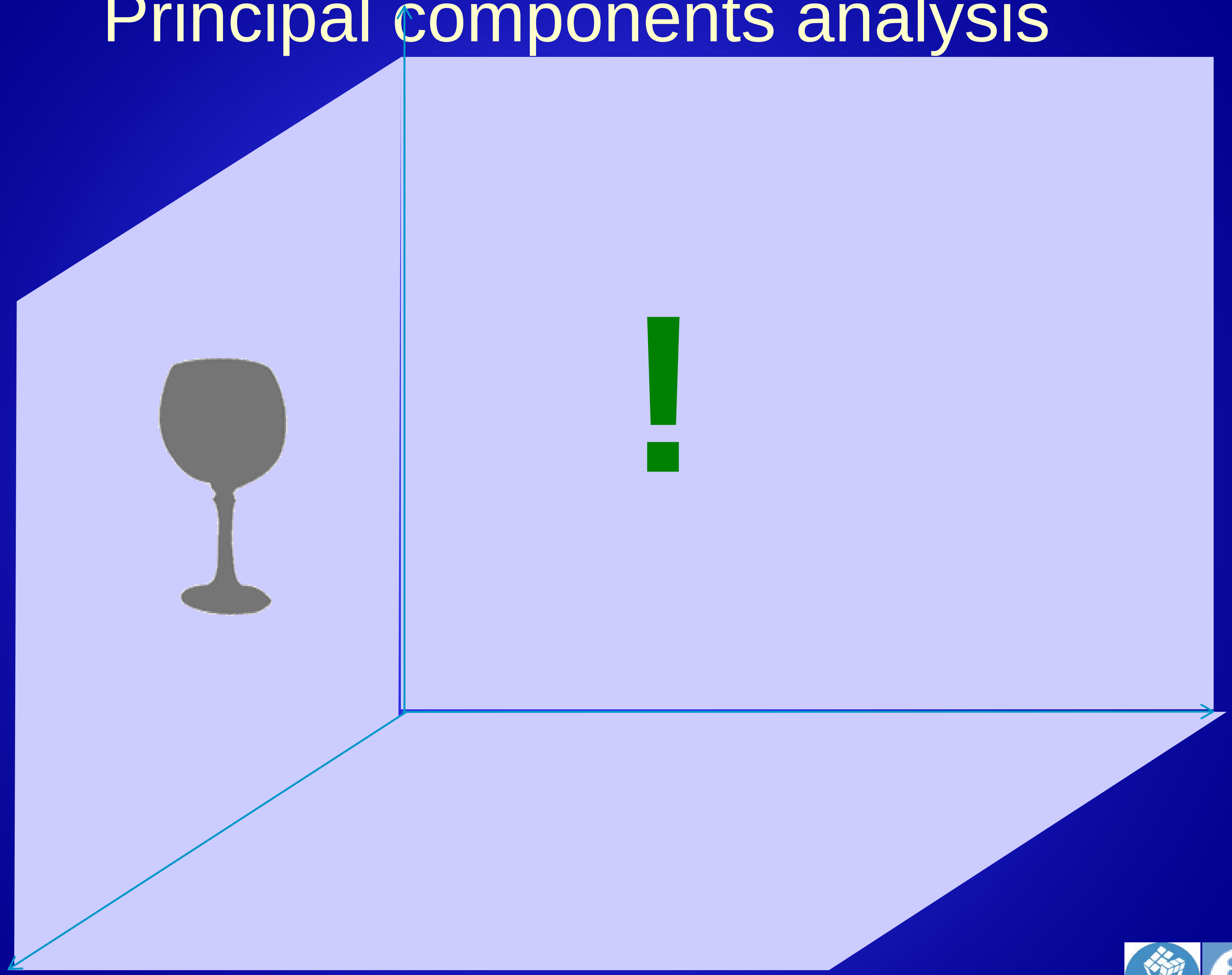
Principal components analysis



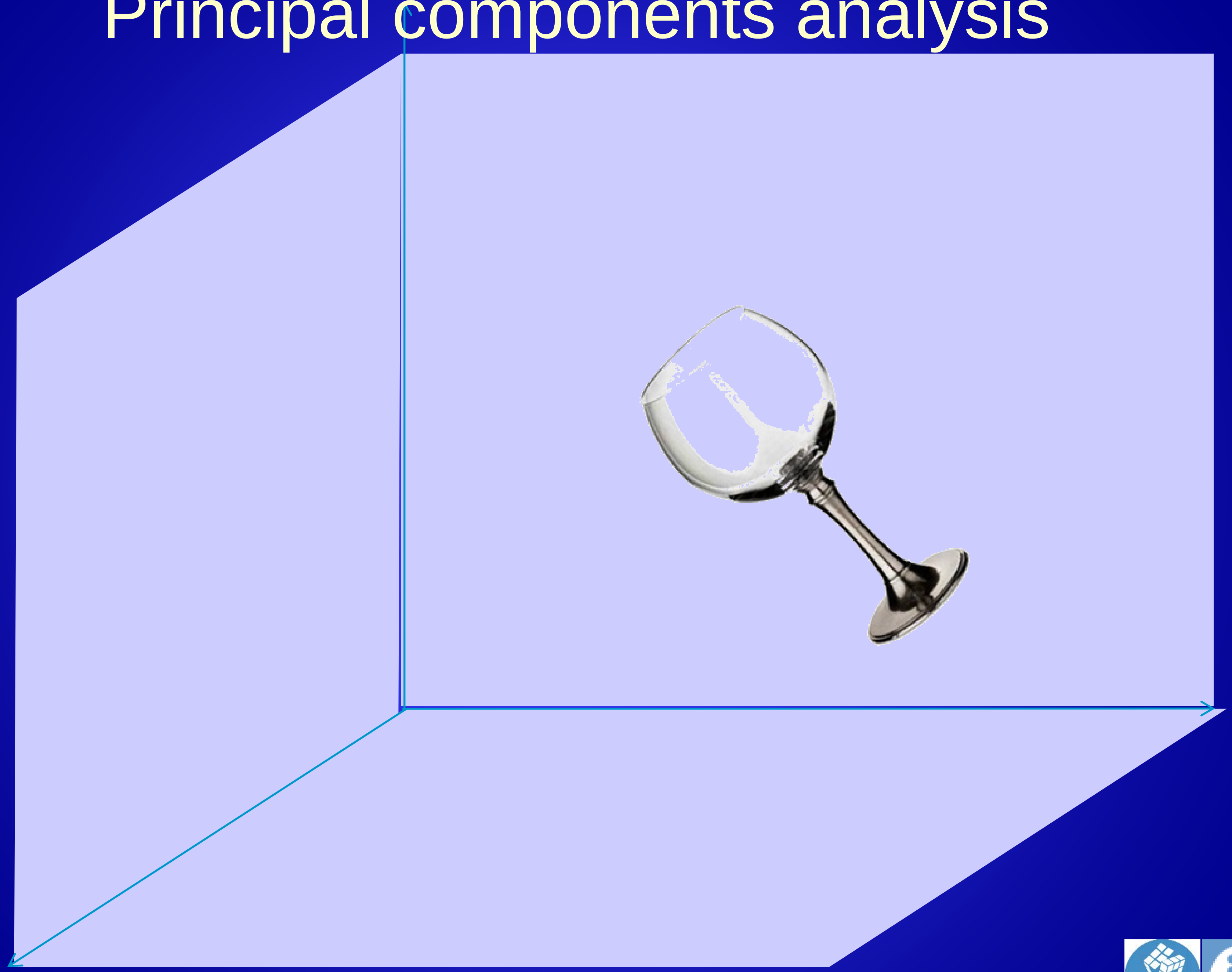
Principal components analysis



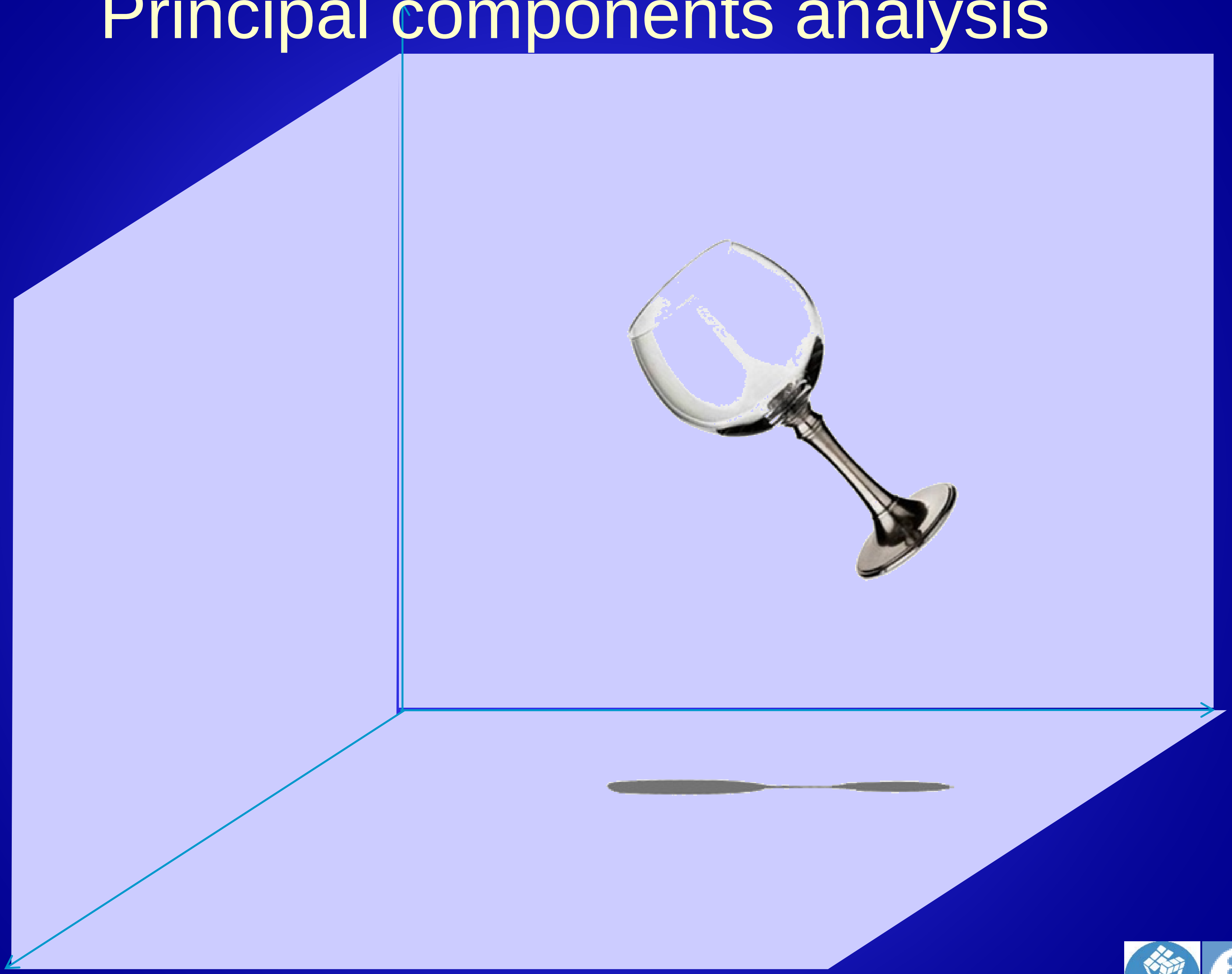
Principal components analysis



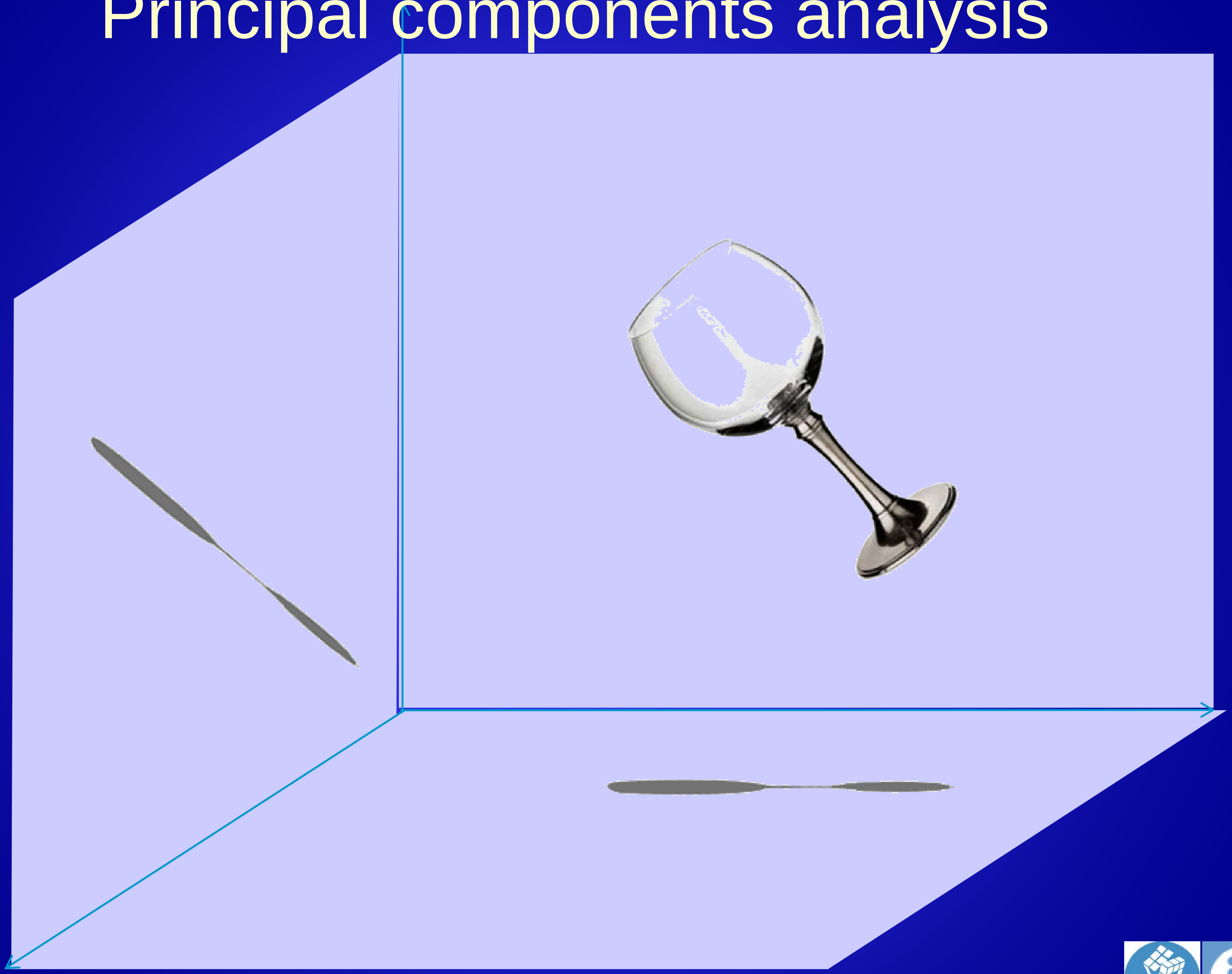
Principal components analysis



Principal components analysis

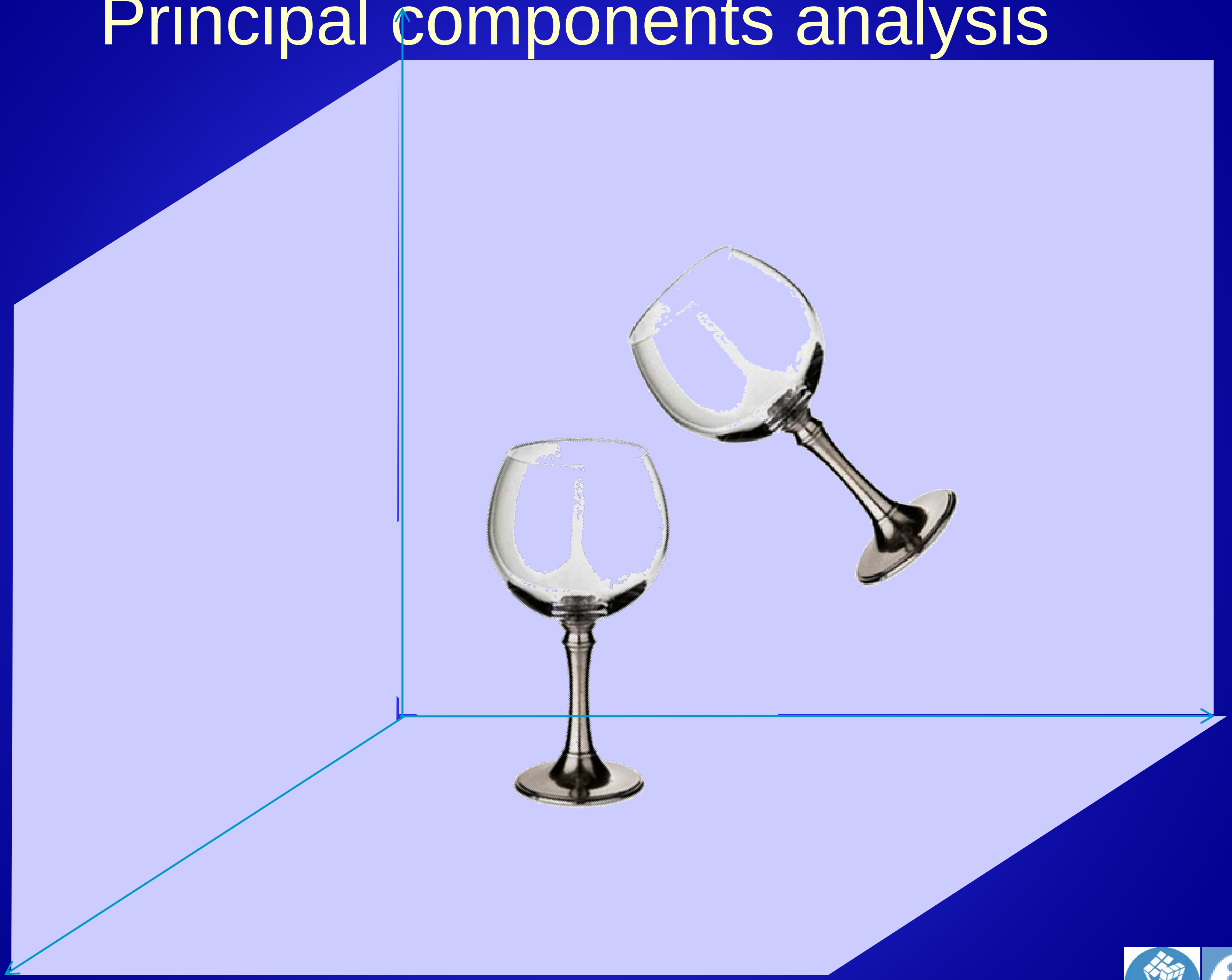


Principal components analysis



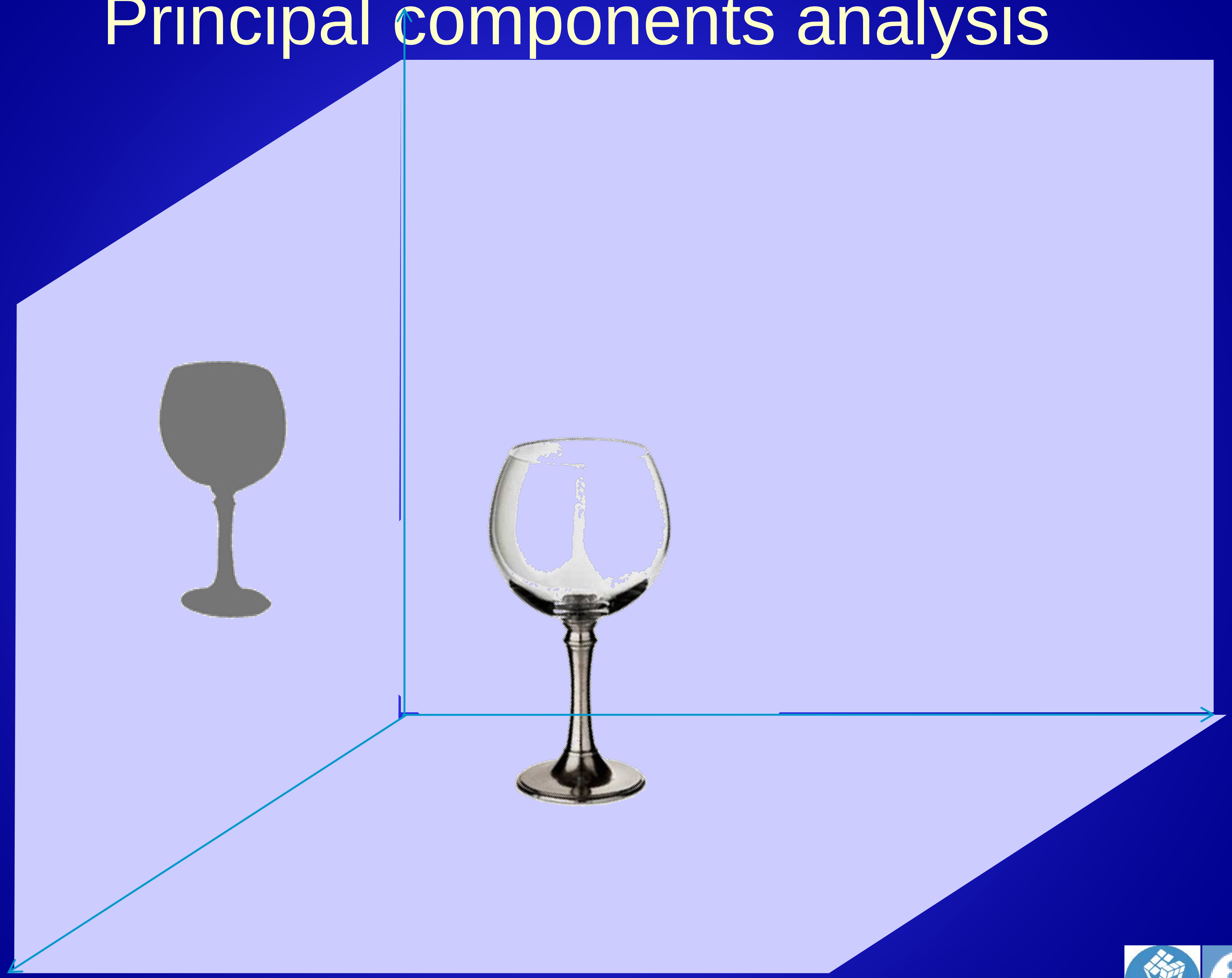
©K. Gibert

Principal components analysis

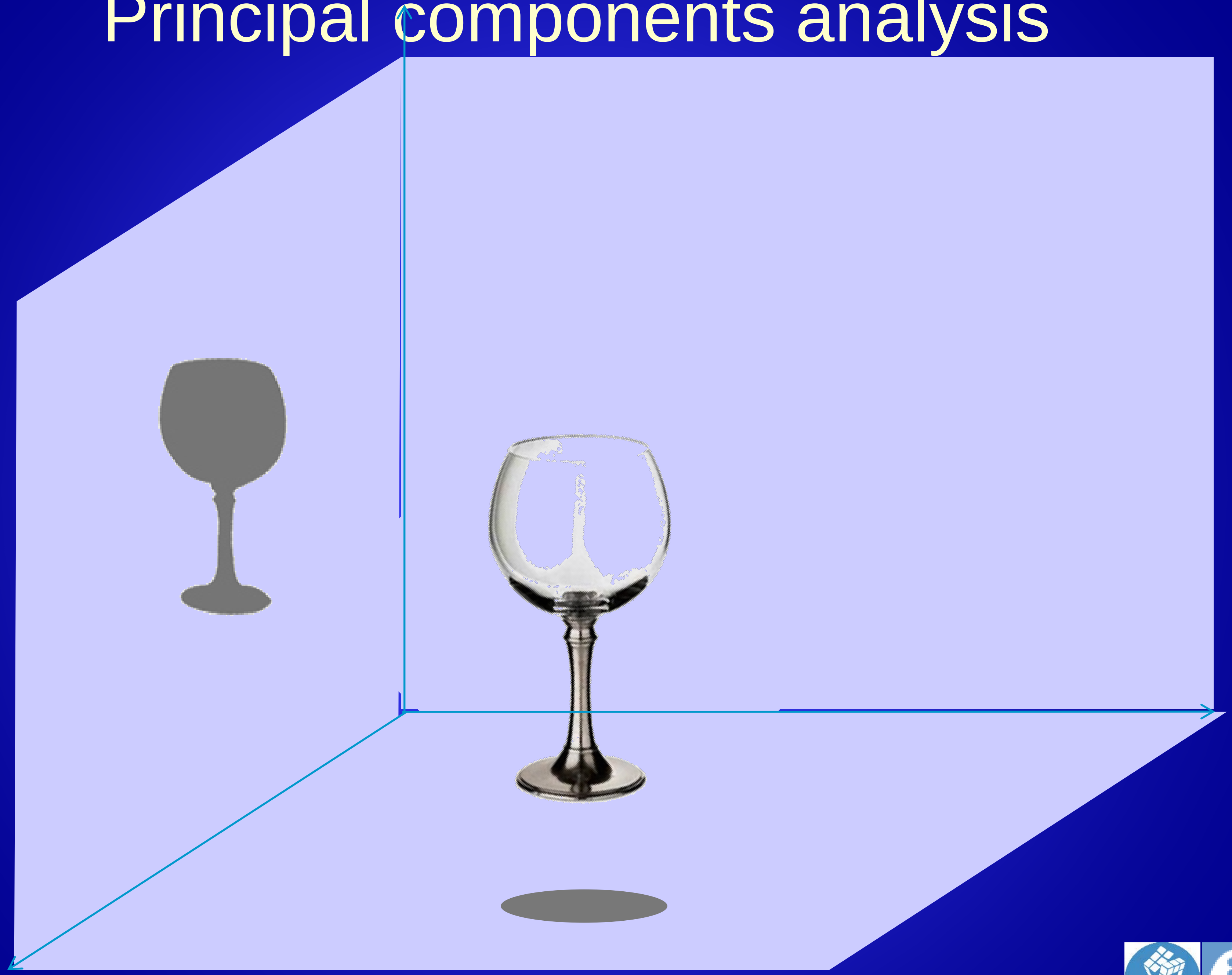


©K. Gibert

Principal components analysis

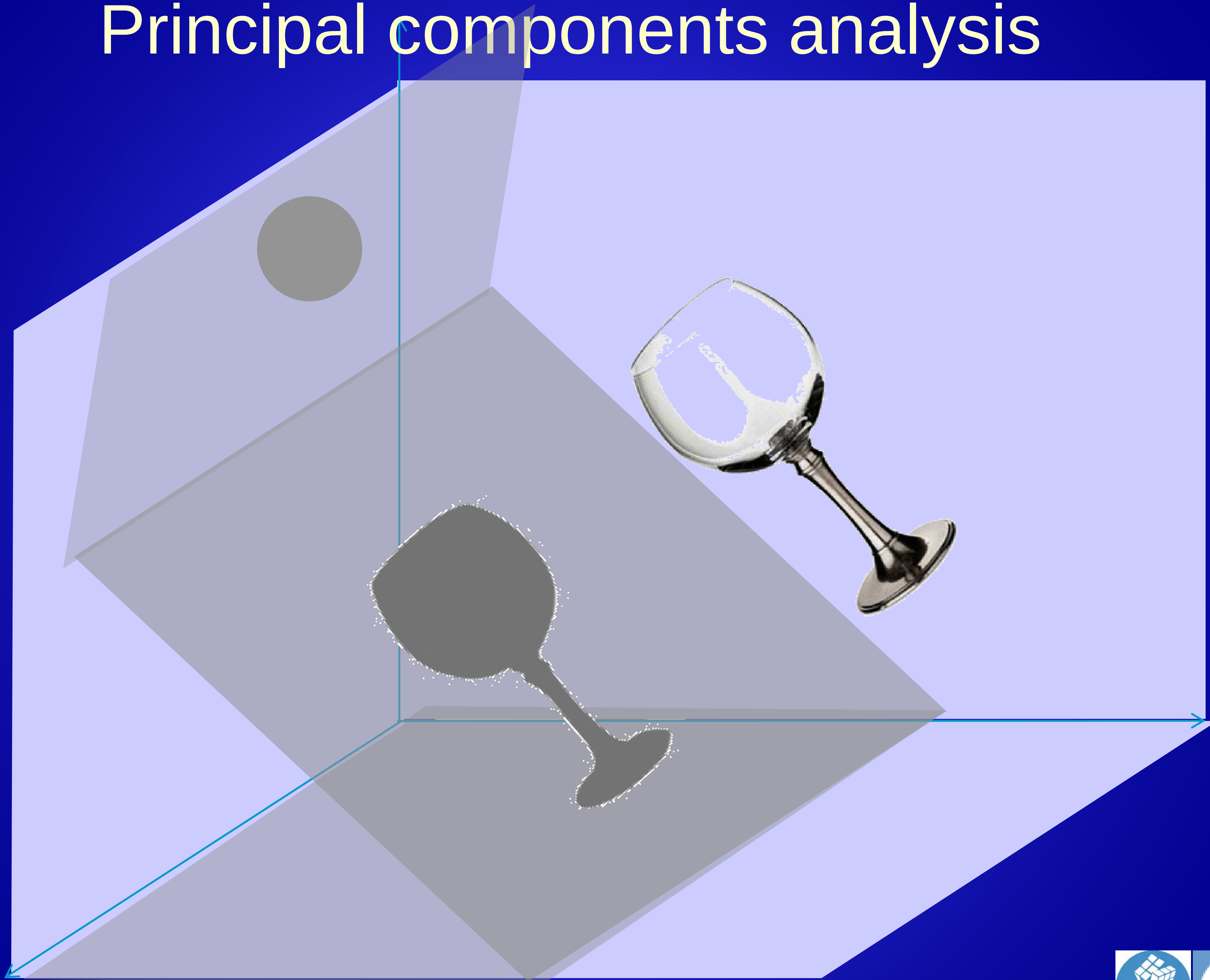


Principal components analysis



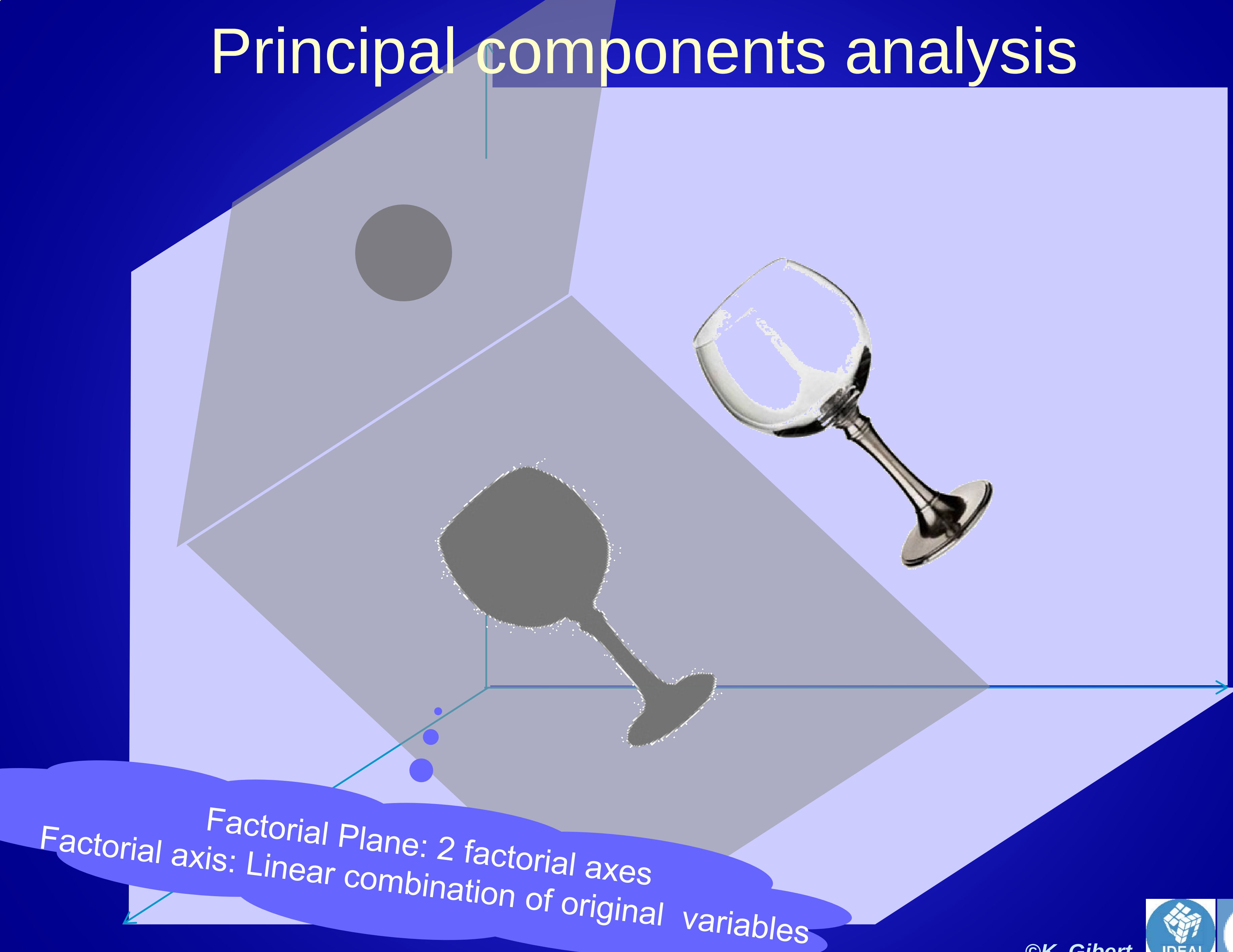
©K. Gibert

Principal components analysis

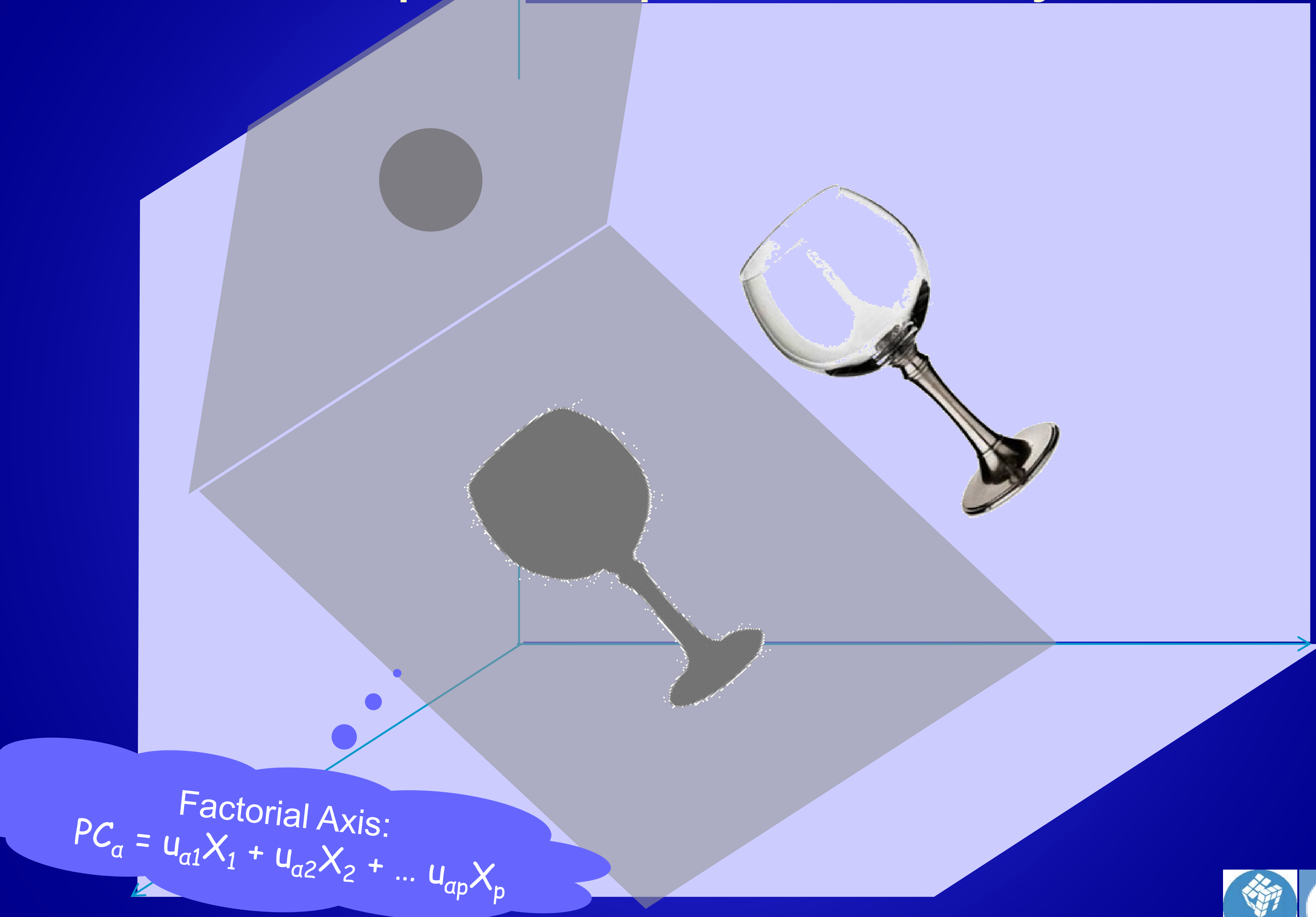


©K. Gibert

Principal components analysis



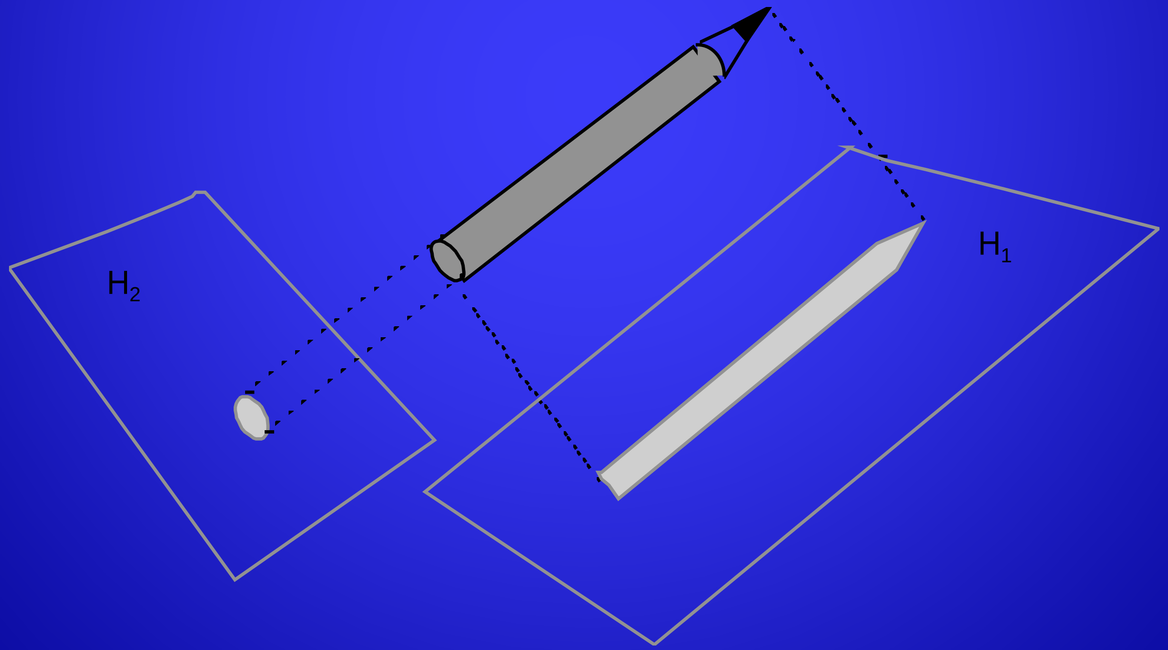
Principal components analysis



Principal components analysis

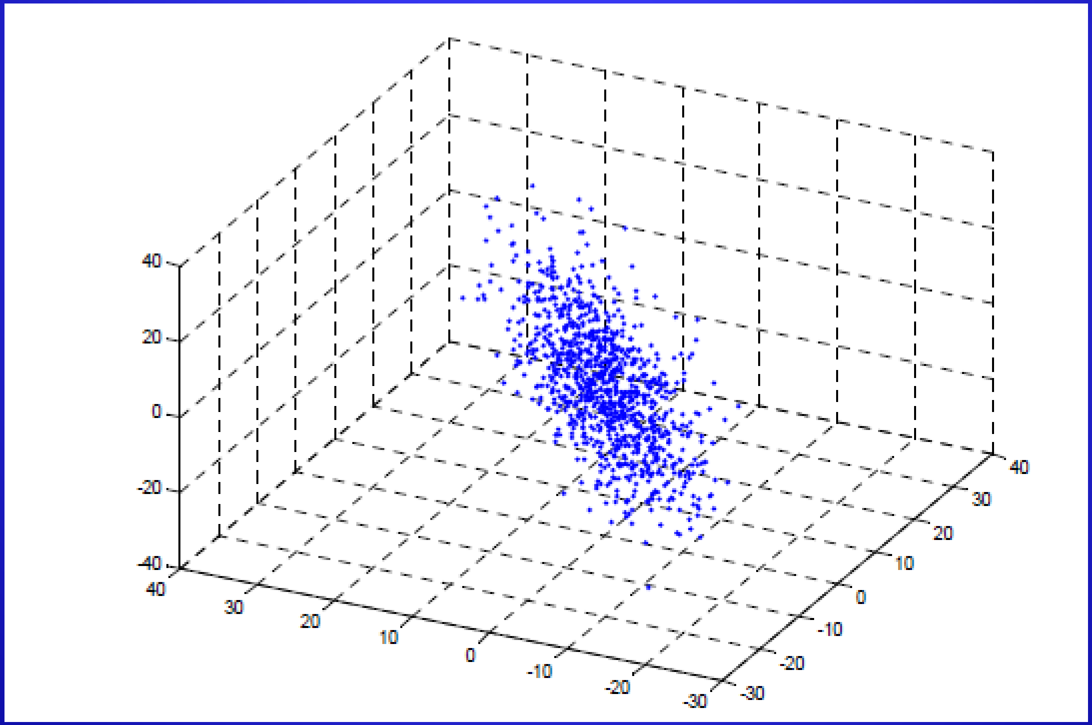
- Purpose:
  - To project the cloud of points upon a subspace (plane) retaining as much original cloud information.

    (see video)



Principal components analysis

- Find the most informative projection planes of data cloud *(factorial planes)*



Factorial Methods

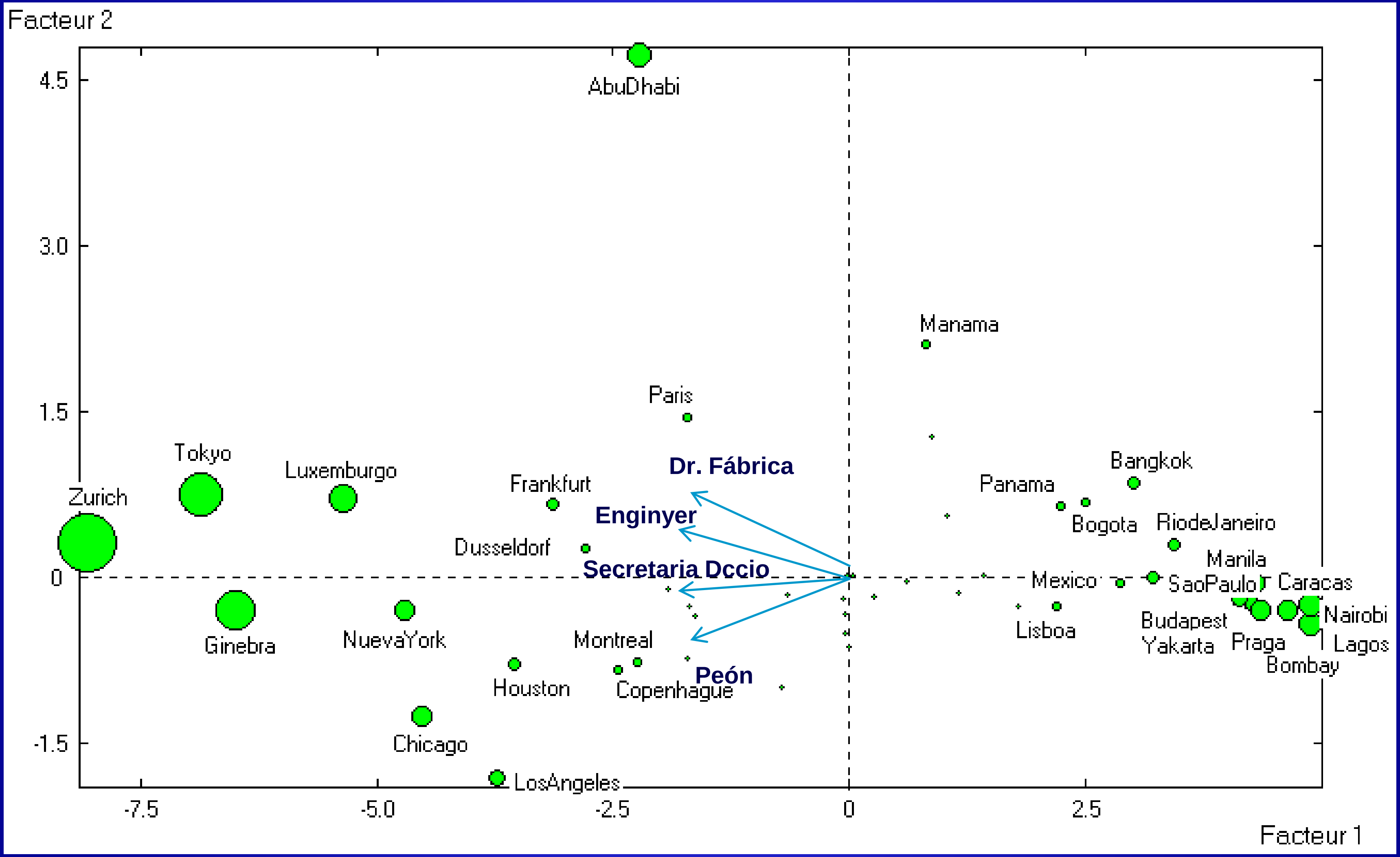
- Output: K factors rotating original X variables
- Factors: Linear combinations of original variables

Several uses:

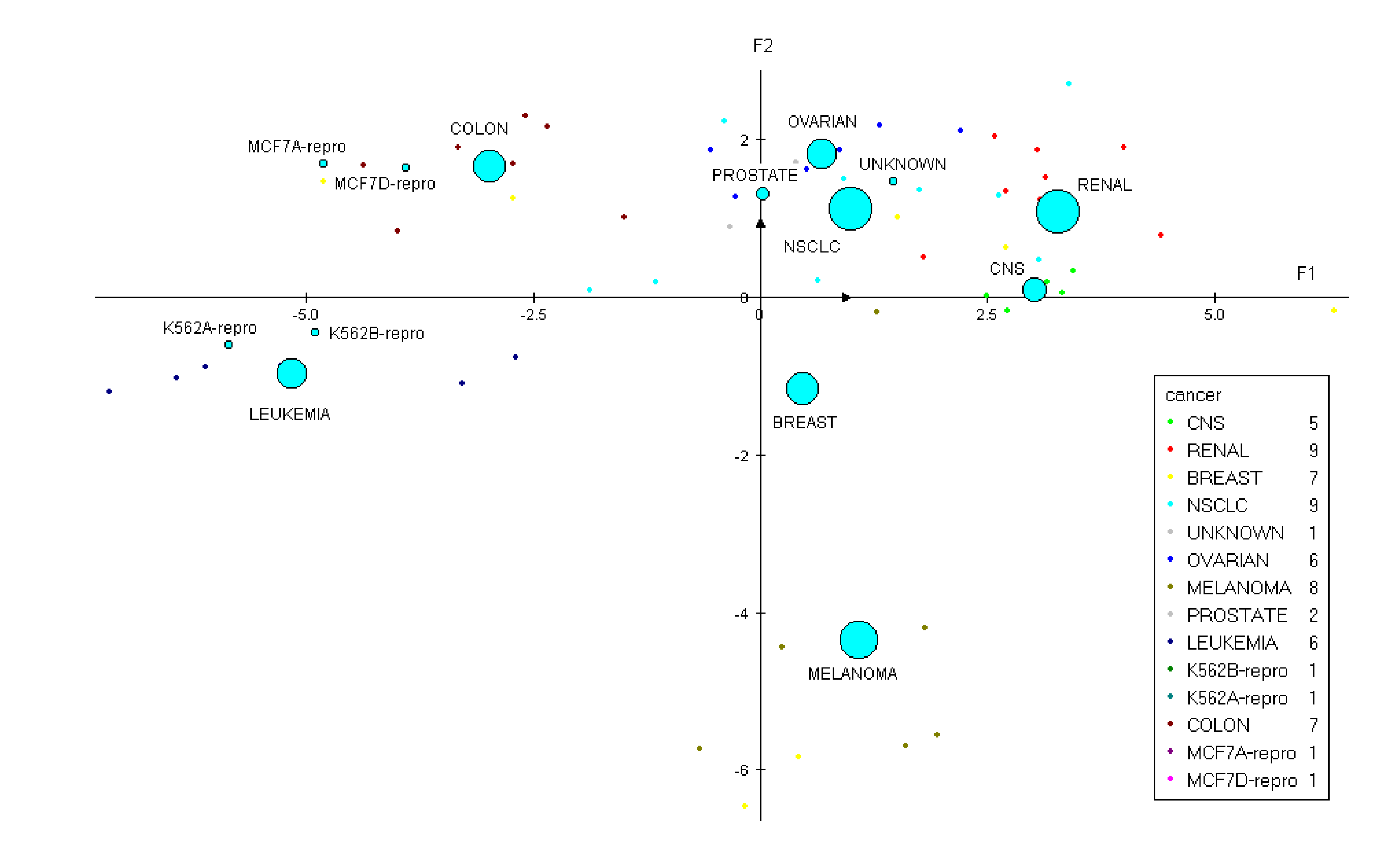
- As an associative data mining method:
  analyze relationships among variables
  Project variables and modalities and find associations

Visualisation of international cities according their salaries. USB 1994.

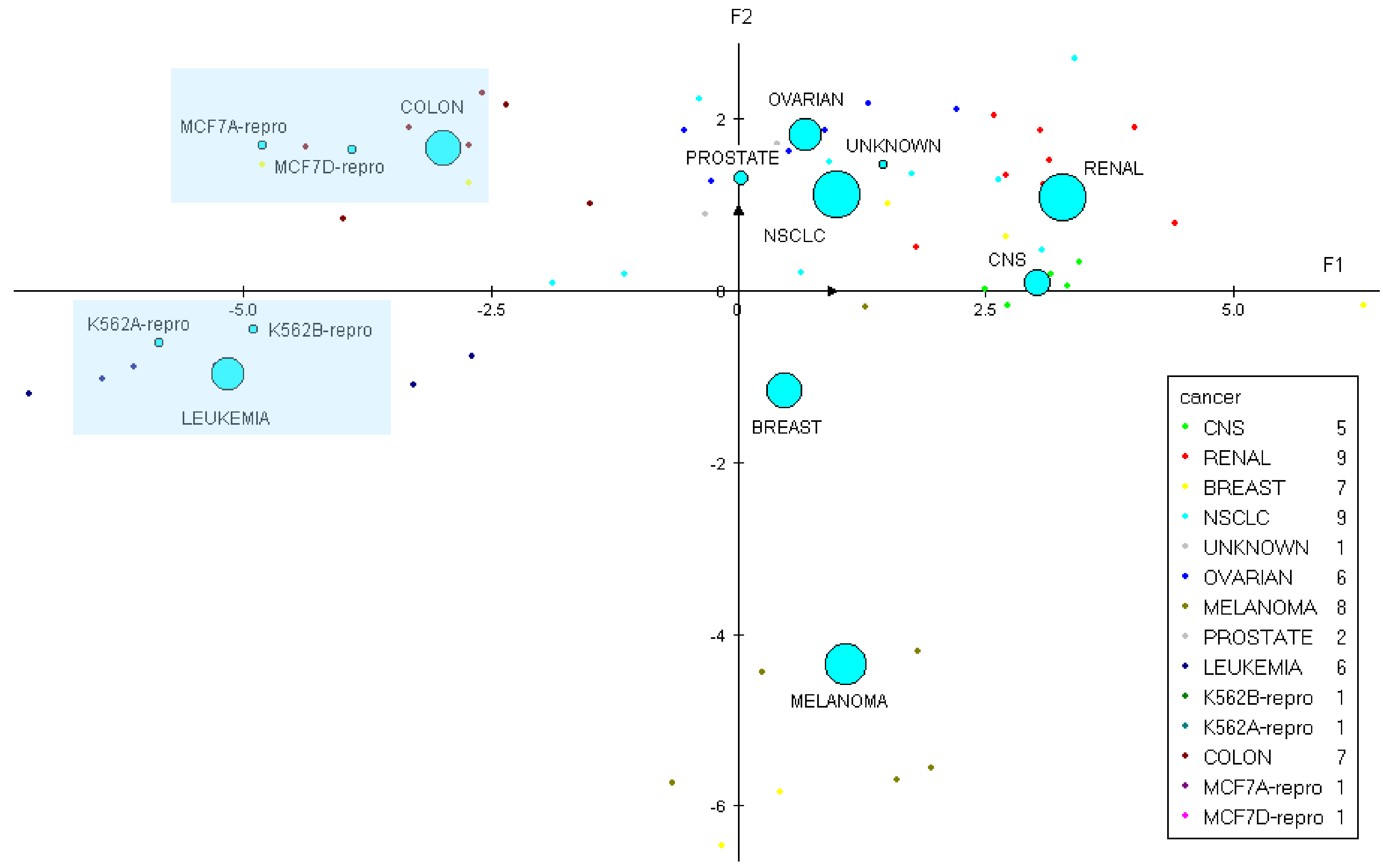
Visualisation of international cities according their salaries. USB 1994.



Microarray data: 64 cancers 6830 gen cromotografy

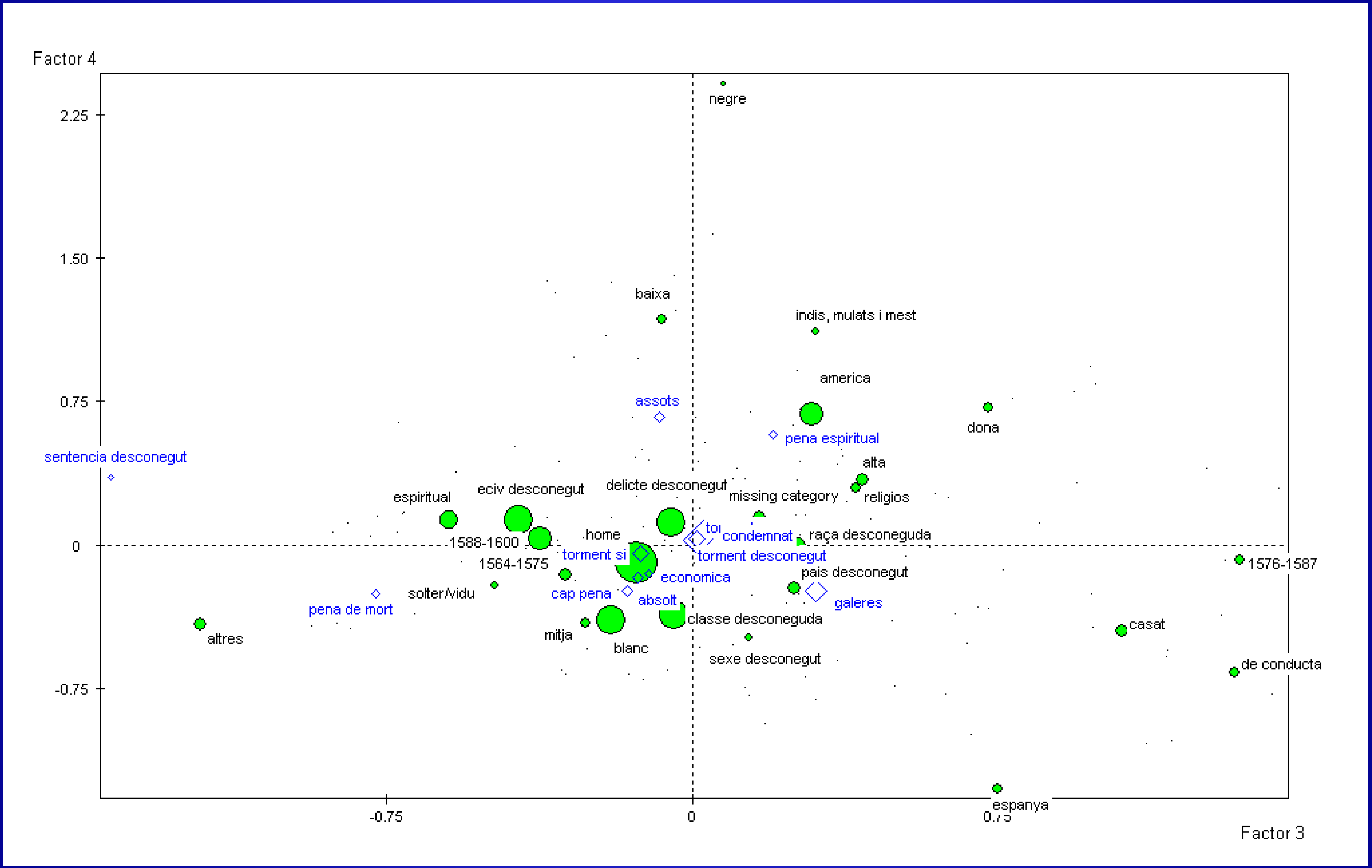


Microarray data: 64 cancers 6830 gen cromotografy



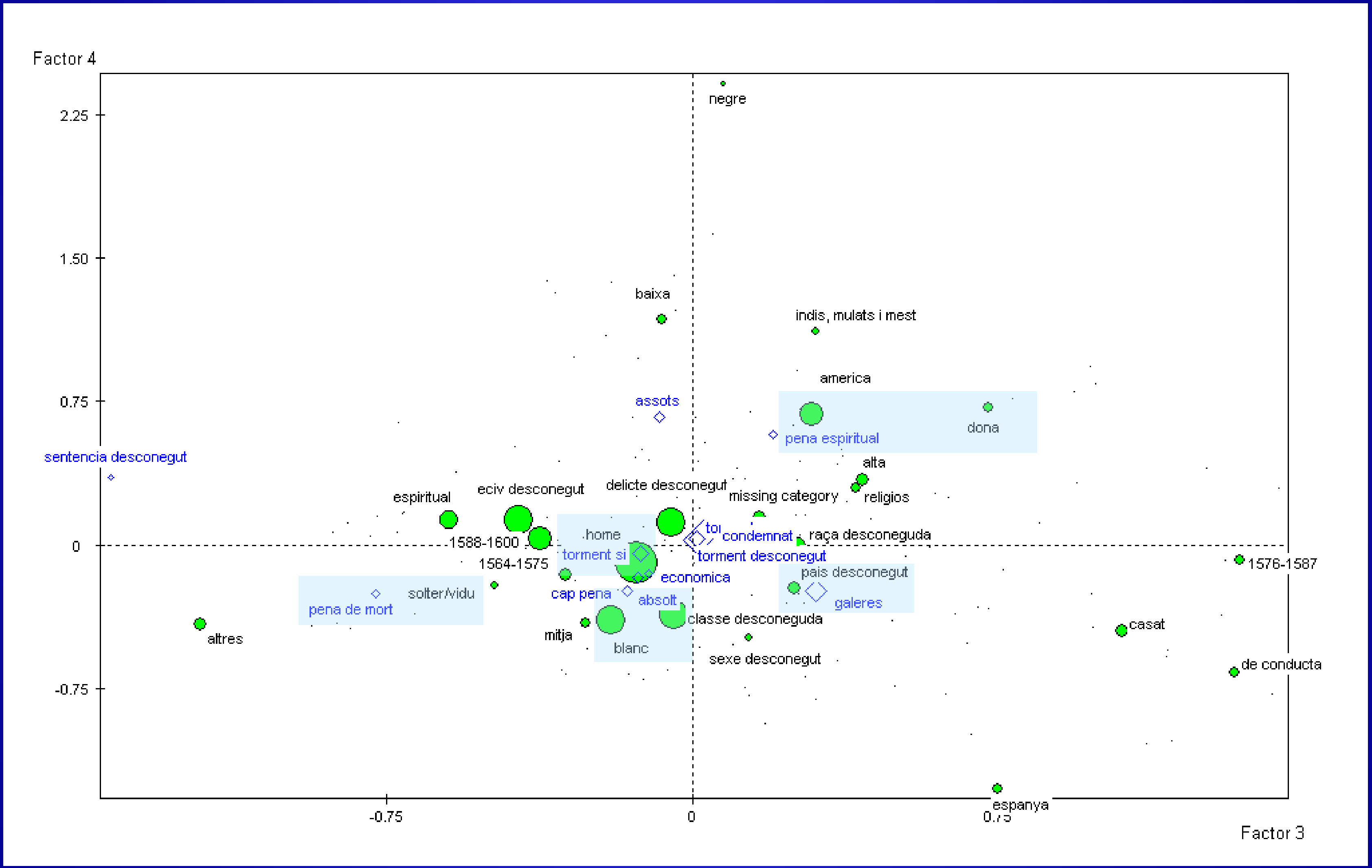
Spanish inquisition 1567-1600

sentences & crimes



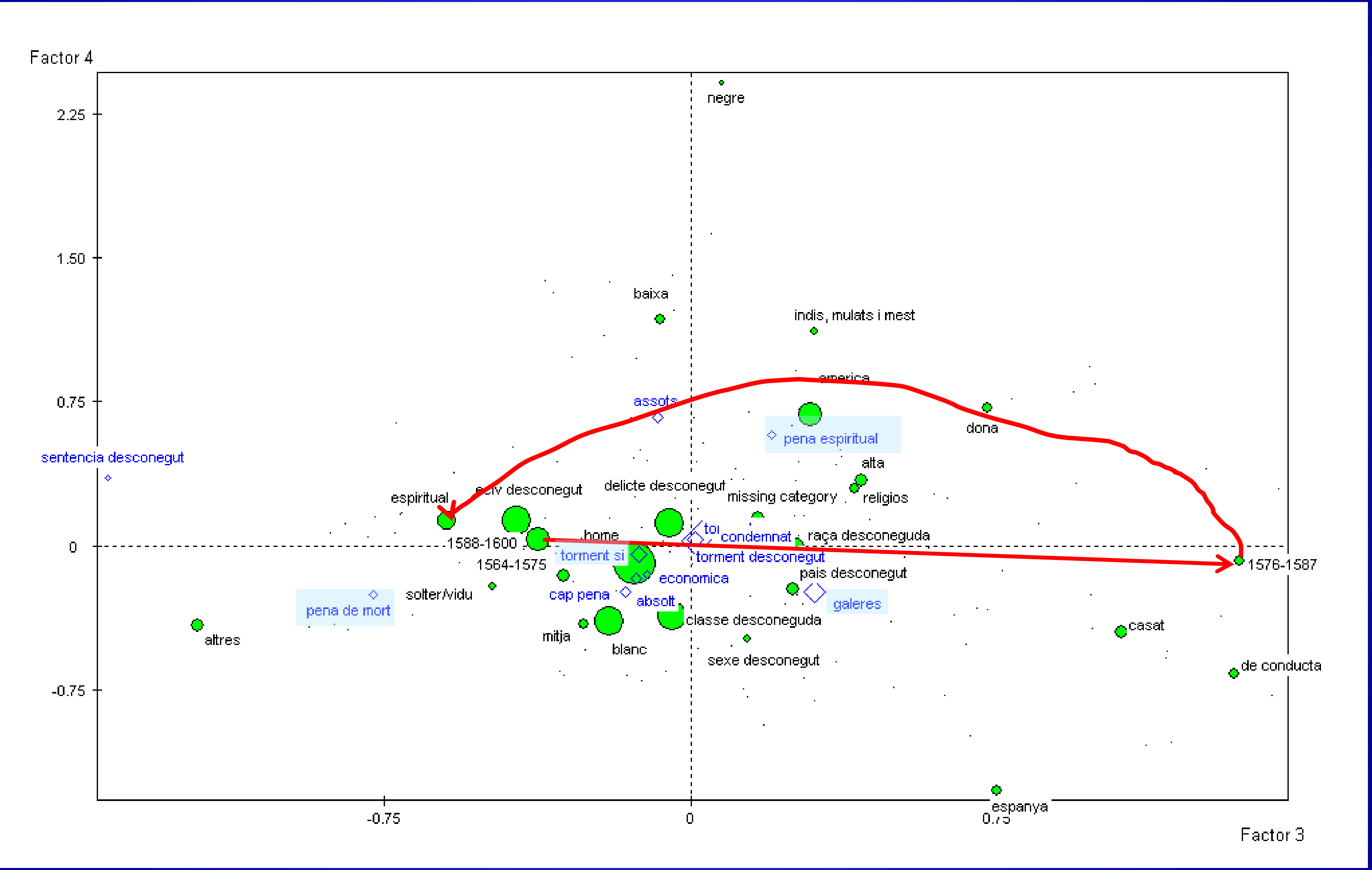
Spanish inquisition 1567-1600

sentences & crimes

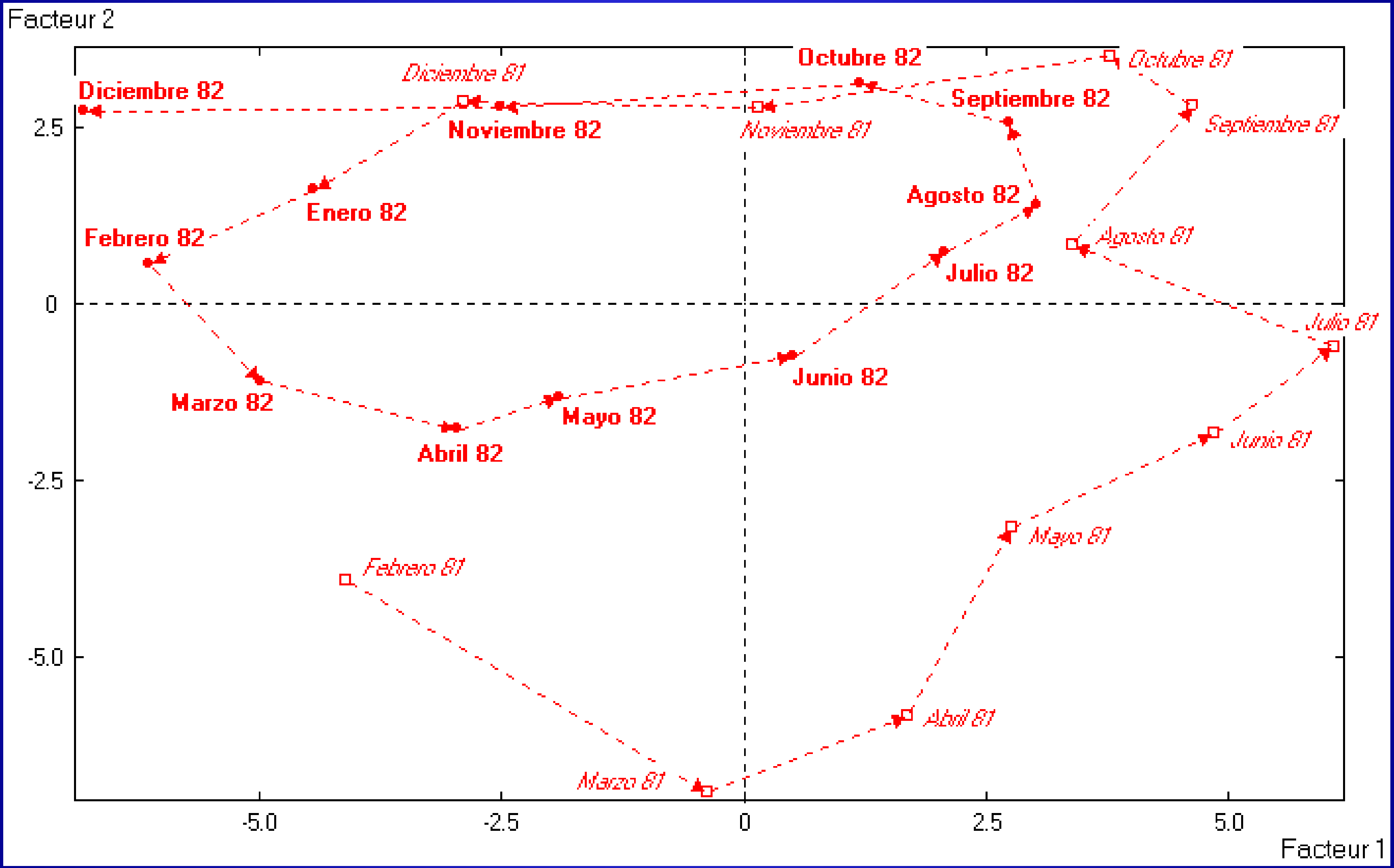


Spanish inquisition 1564-1600

sentences & crimes



Monitoring of the inner temperatures of Lascaux cave (France)



Factorial Methods

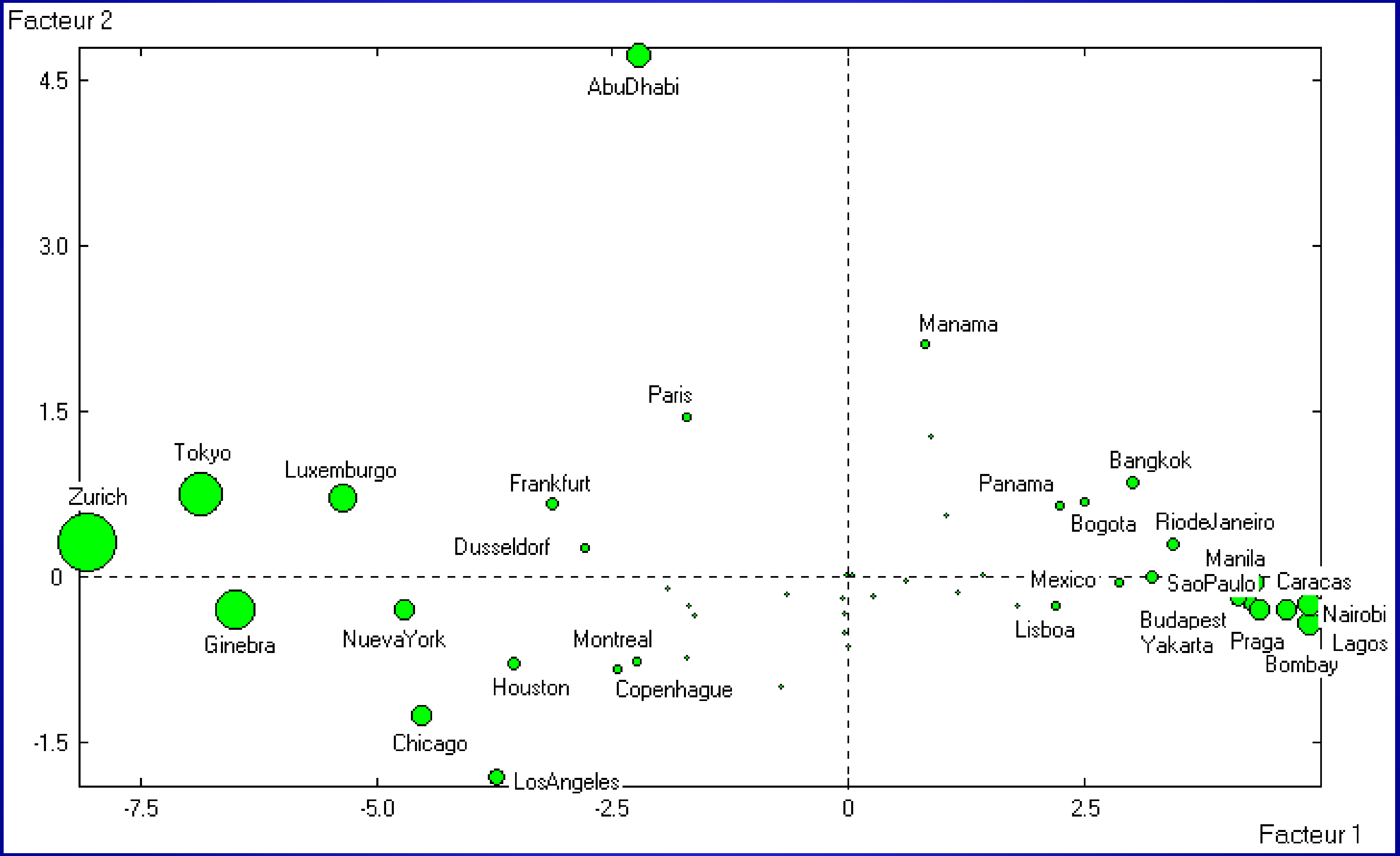
– Output: K factors rotating original X variables
– Factors: Linear combinations of original variables

Several uses:

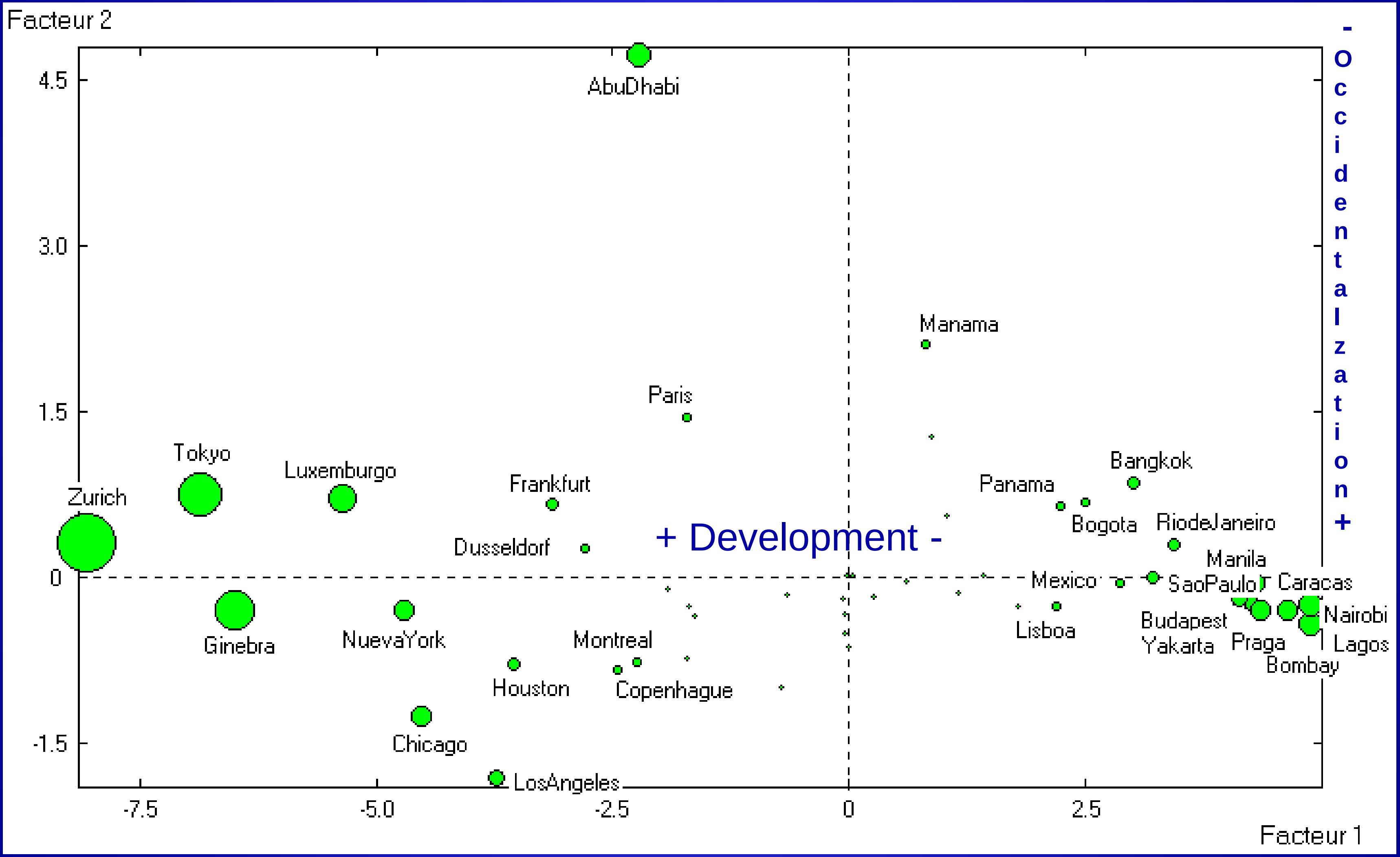
– As an associative data mining method to analyze relationships among variables
  Project variables and modalities and find associations

– As a preprocessing method for elicitation of latent variables
  Project active and illustrative variables/individuals on first/second factorial plane and interpret factors (find latent variables)

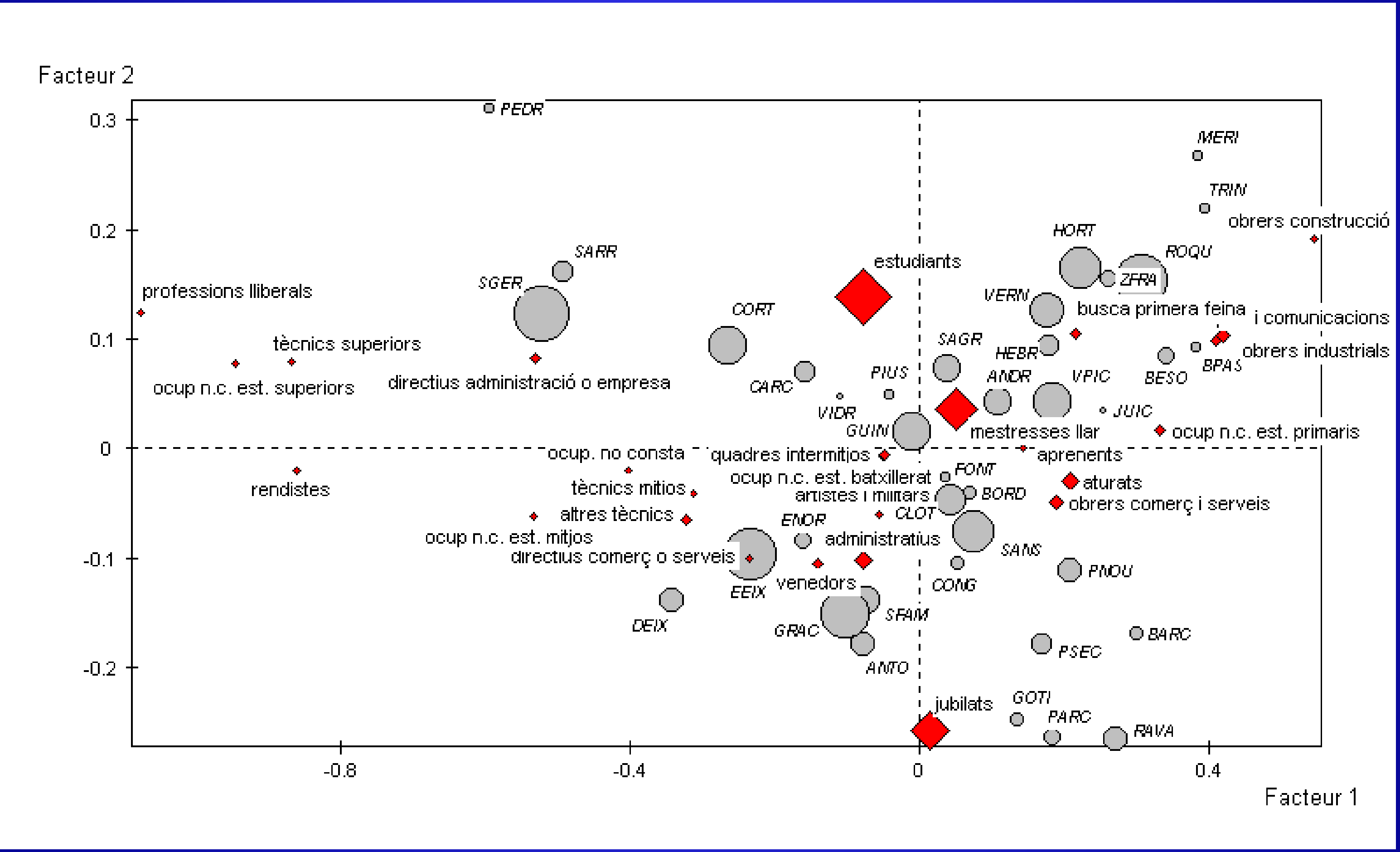
Visualisation of international cities according their salaries. USB 1994.



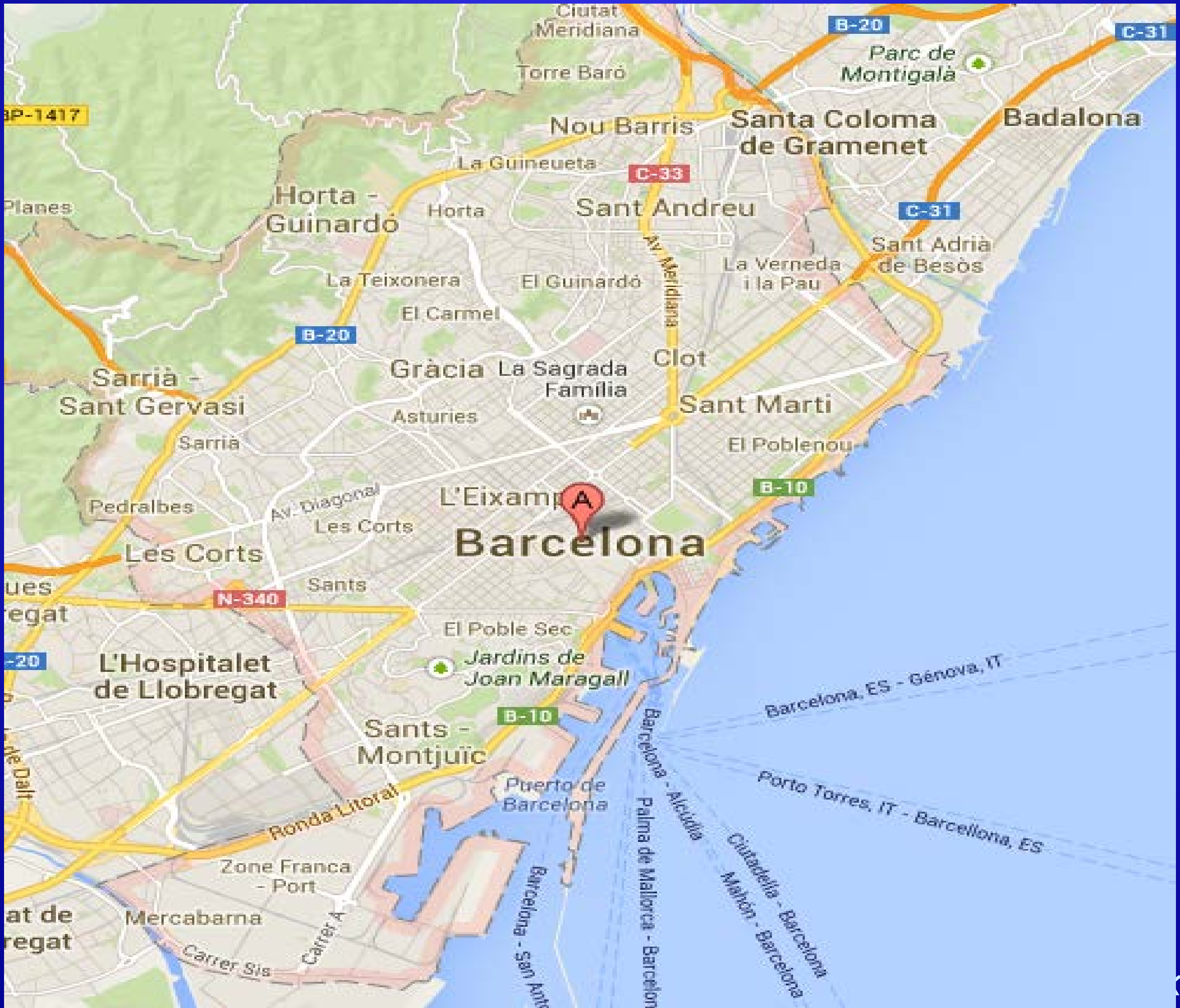
Visualisation of international cities according their salaries. USB 1994.



Visualization of the table *BCN* Quarters x *Profession of inhabitants*

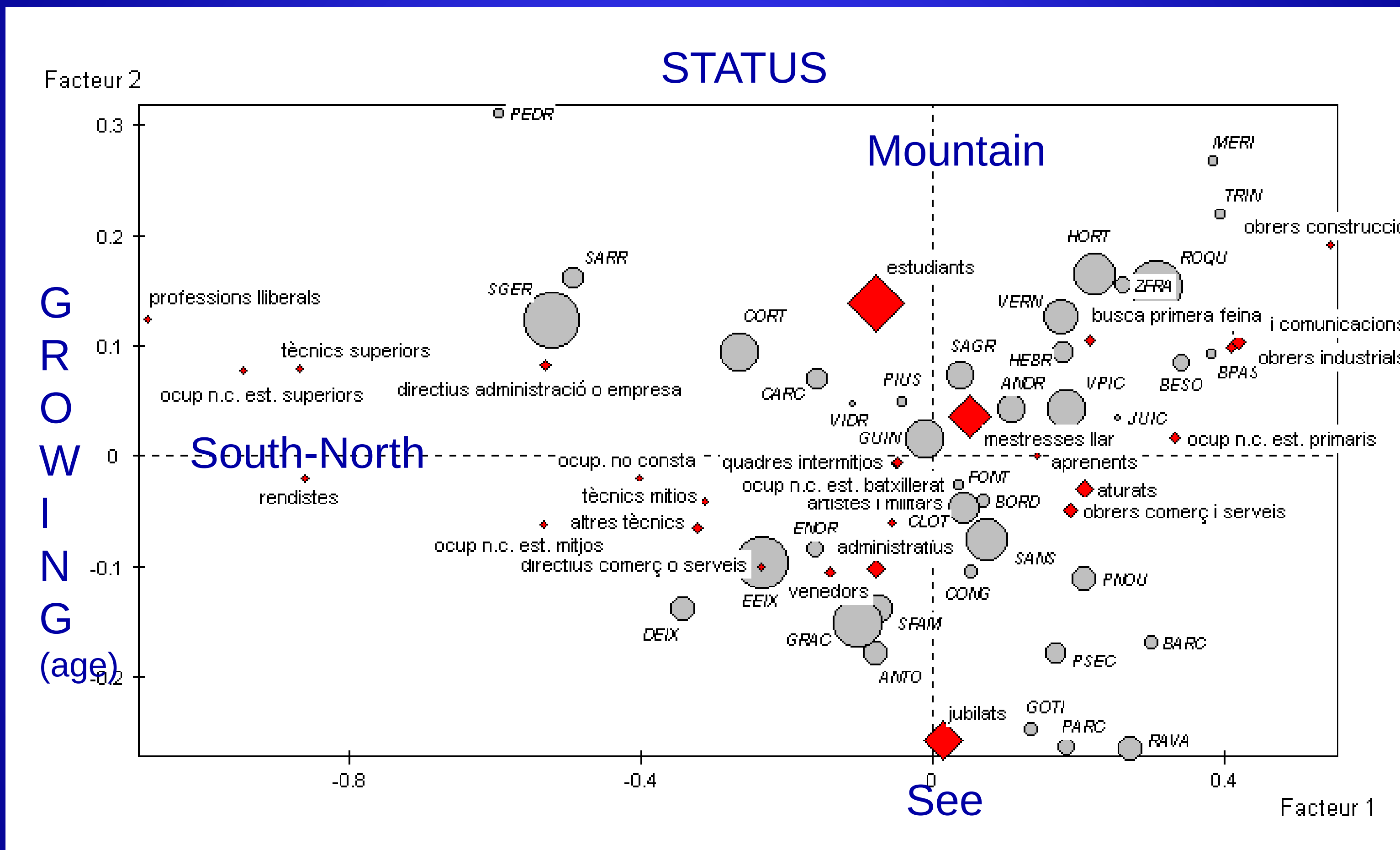


Visualization of the table
BCN Quarters x *Profession of inhabitants*



K. Gibert

Visualization of the table
BCN Quarters x *Profession of inhabitants*



Factorial Methods

- Output: K factors rotating original X variables
- Factors: Linear combinations of original variables

Several uses:

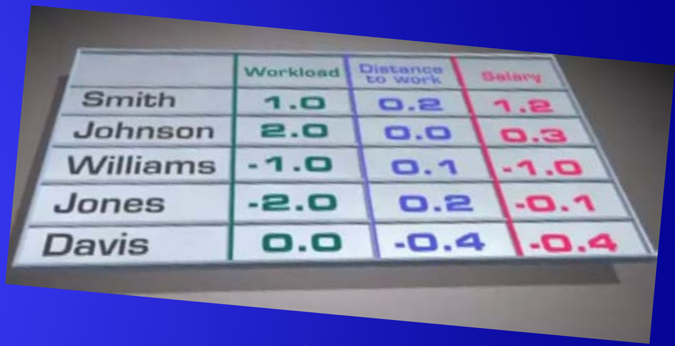
- As an associative data mining method to analyze relationships among variables
  Project variables and modalities and find associations
- As a preprocessing method for elicitation of latent variables
  Project active and illustrative variables/individuals on first/second factorial plan and interpret factors (find latent variables)
- As a preprocessing method for multidimensionality reduction

Factorial Methods

Data	Factorial Method
Continuous variables	Principal Component Analysis PCA
Contingency table	(Simple) Correspondence Analysis CA
Categorical variables	Multiple Correspondence Analysis MCA

Factorial Methods

- Principal Components Analysis
  - Only numerical variables
  - Find the most informative projection planes (factorial planes, maximize projected inertia)



	Workload	Distance to work	Salary
Smith	1.0	0.2	1.2
Johnson	2.0	0.0	0.3
Williams	-1.0	0.1	-1.0
Jones	-2.0	0.2	-0.1
Davis	0.0	-0.4	-0.4

Given <X,M,D>

- A data matrix X (nxp) centered
- A matrix of individuals weights D (nxn)
- Assume euclidean metrics to compare individuals (M= $\mathbb{I}_p$)

Si les dades estan centrades l'angle entre dues variables projectades coincideix amb la correlació entre elles

Matrix $M^{1/2} X'DXM^{1/2}$

- Product of data with the two metrics
- Simetric,
- Semidefinite
- Catches relationships and opositions of data

Factorial Methods

Given triplet <X,M,D>, diagonalize $M^{1/2} X'DXM^{1/2}$

Data	Factorial Method	X	M	D
Continuous variables	PCA	Centered data matrix	$\mathbb{I}_p$	$\mathbb{I}_n$
Contingency table (n_{ij})	CA	$F=(n_{ij}/n_i)$	$diag(1/f_j)$	$diag\ (f_i)$
		$G=(n_{ij}/n_j)$	$diag(1/f_i)$	$diag\ (f_j)$
Categorical variables	MCA	$F=(f_{ij}/(f_i/\sqrt{f_j}))$	$\mathbb{I}_p$	$diag\ (f_i)$
		Burt table	$\mathbb{I}_{n+p}$	$diag(n_{ij})$

Factorial Methods

- Principal Components Analysis

$M^{½}X'DXM^{½}$ catches well the data structures

Rang($M^{½}X'DXM^{½}$) =r, r=rang(X) r positive vaps and p-r null vaps

Trace($M^{½}X'DXM^{½}$)=$\sum_{\alpha=1}^{r} \lambda_\alpha$ (λ_α, the r non null vaps)

M= $\mathbb{I}_p$: $M^{½}X'DXM^{½}$ = X'DX

X centered and D diagonal : X'DX= Cov(X)

X standardized and D diagonal : X'DX= Corr(X)

(preferred, big variabilities do not dominate analysis)

Build variances and covariances matrix: X'DX

Diagonalize X'DX (i.e. solving the equation) $X'DXu= \lambda u$

provides eigen values λ_α and

eigenvectors $u_\alpha = (u_{\alpha 1} .... u_{\alpha p})$

Factorial Methods

- Principal Components Analysis

Diagonalize X'DX (i.e. solving the equation) $X'DXu = \lambda u$ (1)

det(X'DX-λ)=0 (find roots of characteristic polynomial)

provides eigen values λ_α *(*α *=1:r, r=rang(X))*

substituting in (1) provides eigenvectors $u_\alpha = (u_{\alpha 1} \ldots u_{\alpha p})$

$u^{-1}X'DXu = \lambda$ *is a diagonal matrix*

(X'DX becomes diagonal when pre/post multiplied by u)

$u^{-1}=u'$ *in orthonormal basis:* $u'X'DXu = \lambda$

X'DX decompose in a product by a diagonal matrix $X'DX = u\lambda u'$

$X'DX = u\lambda u' = u\lambda^{1/2}\lambda^{1/2} u' = u\lambda^{1/2}\mathbb{I}\lambda^{1/2} u' = u\lambda^{1/2} u'u\ \lambda^{1/2} u' = A^{1/2} A^{1/2}$

X'DX decompose in a product of something by itself (A square root)

Trace(X'DX)=Trace(λ*)* (property of diagonalization)

Factorial Methods

- Given <X,M,D>

Diagonalize correlations matrix (with normalized data X'DX)

Get r eigen values λ_α and sort decreasingly

$\{\lambda_\alpha\}_{\alpha=1:r} \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots\ldots \geq \lambda_r$

Corresponding eigenvectors $u_\alpha = (u_{\alpha 1} \ldots . \, u_{\alpha p})$

$|u_\alpha|=1$

$u_\alpha u_{\alpha'}=0$

$\{u_\alpha\}_{\alpha=1:r}$ *orthonormal base for individuals*

The subspace generated by $\{u_\alpha\}_{\alpha=1:r}$ is the same as the subspace generated by the rows of X

Factorial Methods

- Given <X,M,D>

Diagonalize Correlations matrix X'DX

Get r eigen values λ_α and sort decreasingly

$\{\lambda_\alpha\}_{\alpha=1:r} \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots\ldots \geq \lambda_r$

Corresponding eigenvectors $u_\alpha = (u_{\alpha 1} \ldots\ldots u_{\alpha p})$

for $M = \mathbb{I}_p$: $u^*_\alpha = u_\alpha$; *for* $M \neq \mathbb{I}_p$: $u^*_\alpha = M^{-½} u_\alpha$

$\{u^*_\alpha\}_{\alpha=1:r}$ *orthonormal base for individuals*

u^*_α *are the principal factors of X : good rotation directions*

$U^* = ([u^*_1]\ [u^*_2] \ldots\ldots [u^*_r])$ *is the basis for the projection space*

Factorial Methods

- Given <X,M,D>

In general *Diagonalize* $M^{½}X'DXM^{½}$

Get r eigen values λ_α *and sort decreasingly (vaps are conserved!!!!)*

$$\{\lambda_\alpha\}_{\alpha=1:r} \qquad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots\ldots \geq \lambda_r$$

Corresponding eigenvectors $u^*_\alpha = (u^*_{\alpha 1} \ldots u^*_{\alpha p})$

by algebraic properties, u^*_α *can be found from* u^*

$$u^*_\alpha = M^{-½} u_\alpha$$

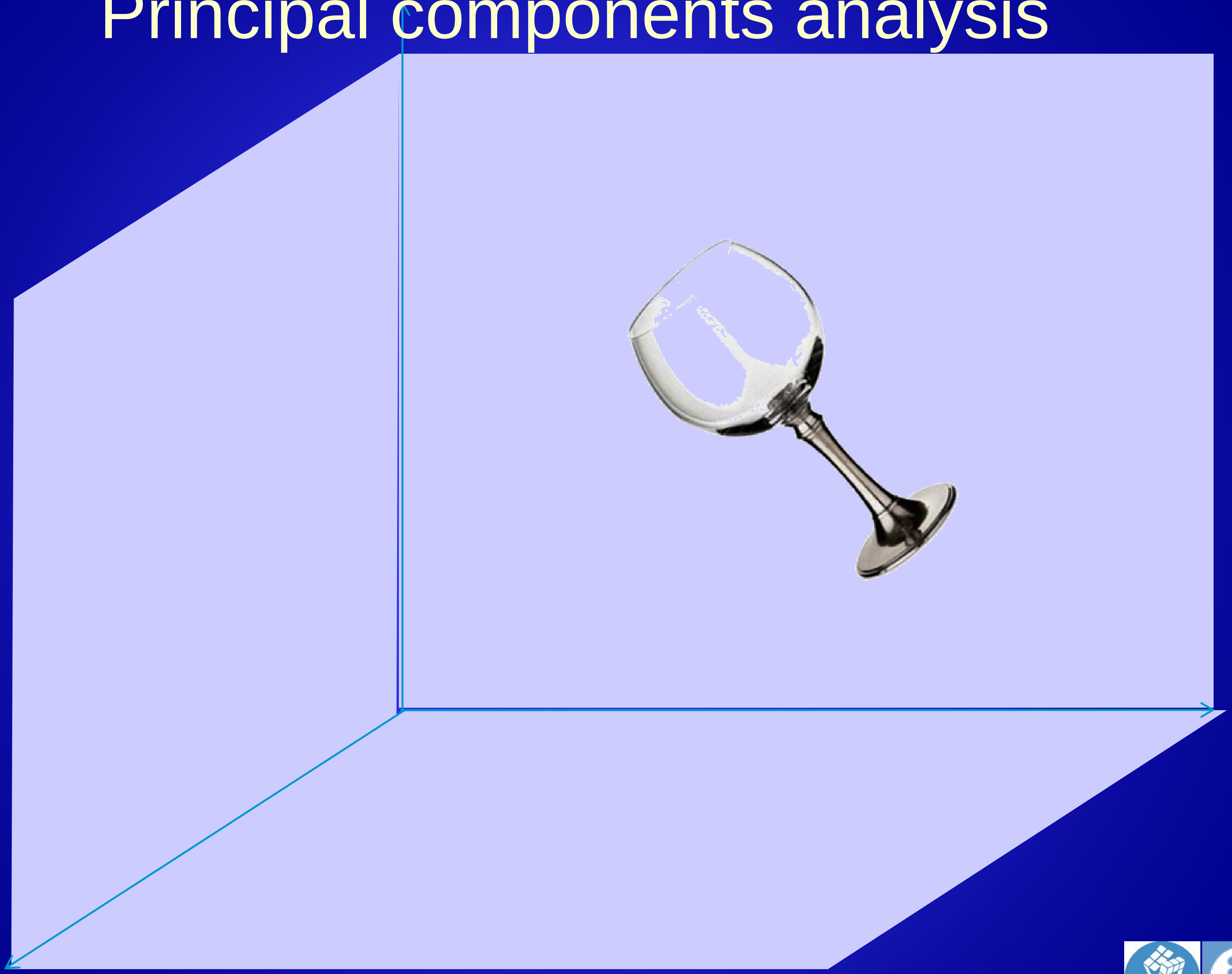
$\{u^*_\alpha\}_{\alpha=1:r}$ *orthonormal base for individuals*

$$|u^*_\alpha|_M = 1 : \quad u^{*\prime}_\alpha M u^*_\alpha = u'_\alpha M^{-½} M M^{-½} u_\alpha = 1$$

$$u^*_\alpha M u^*_{\alpha'} = 0: \quad u^*_\alpha M u^*_{\alpha'} = u'_\alpha M^{-½} M M^{-½} u_{\alpha'} = 0$$

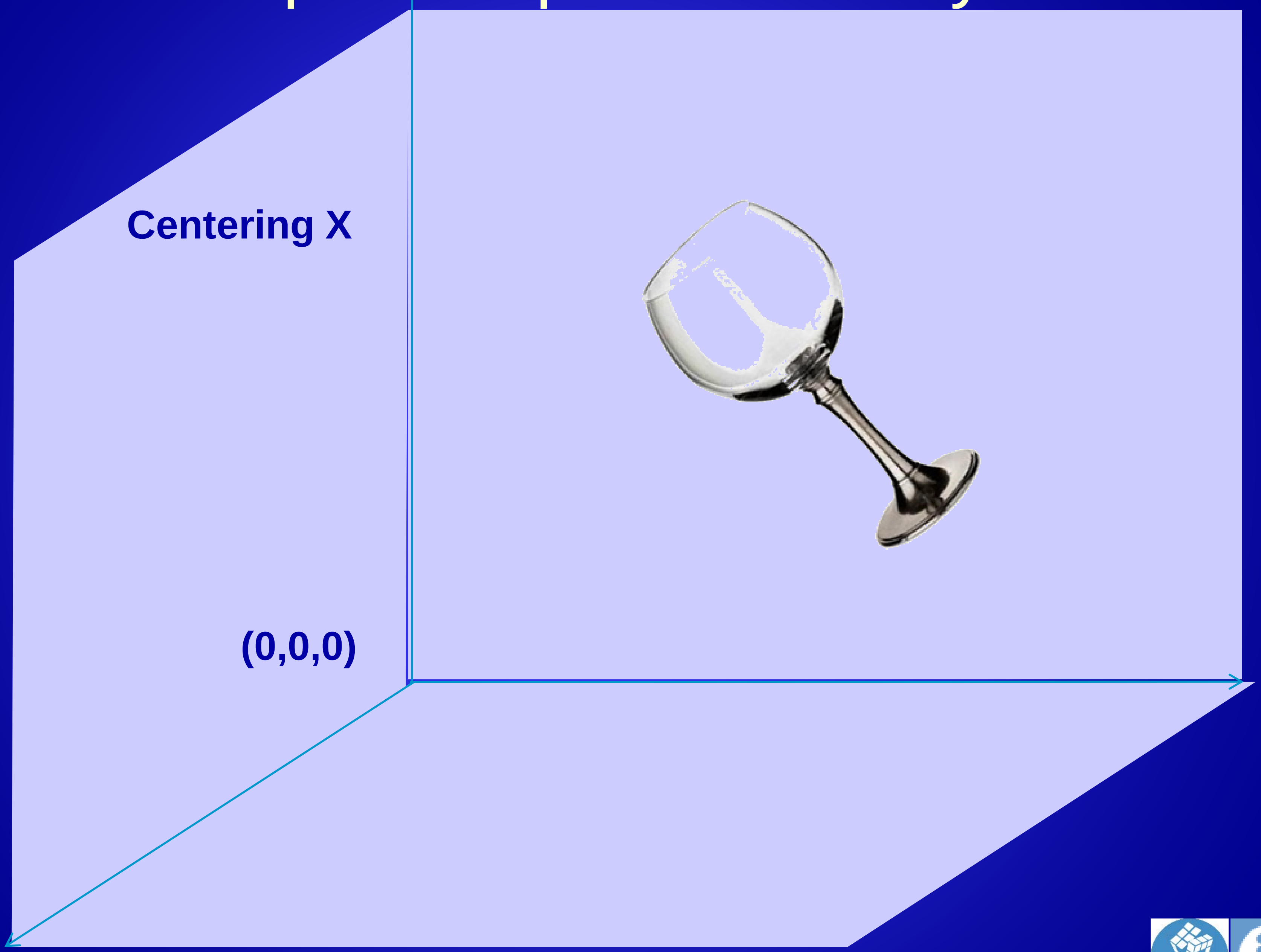
Subspace generated by $\{u^*_\alpha\}_{\alpha=1:r}$ = *Subspace generated by X rows*

Principal components analysis

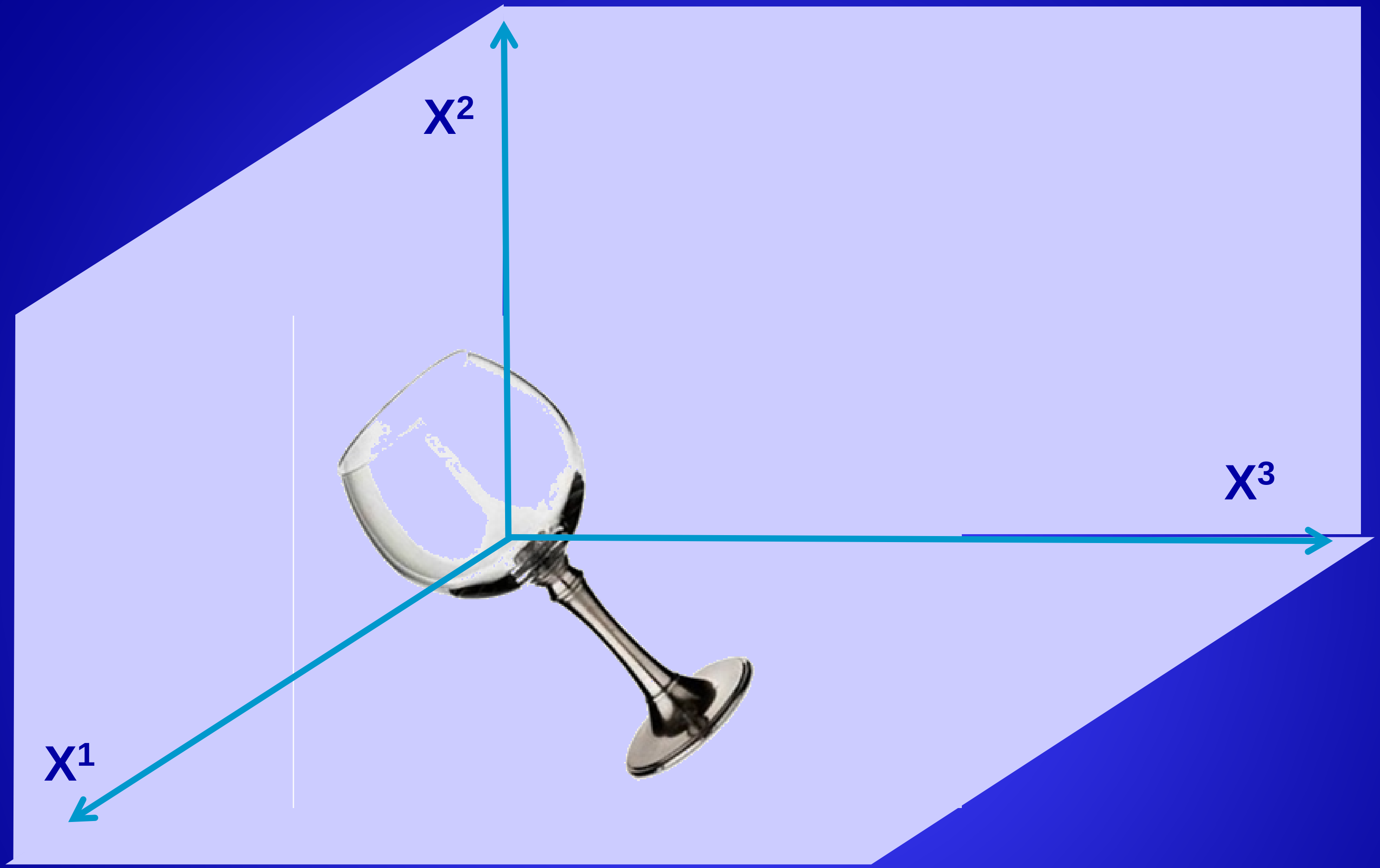


©K. Gibert

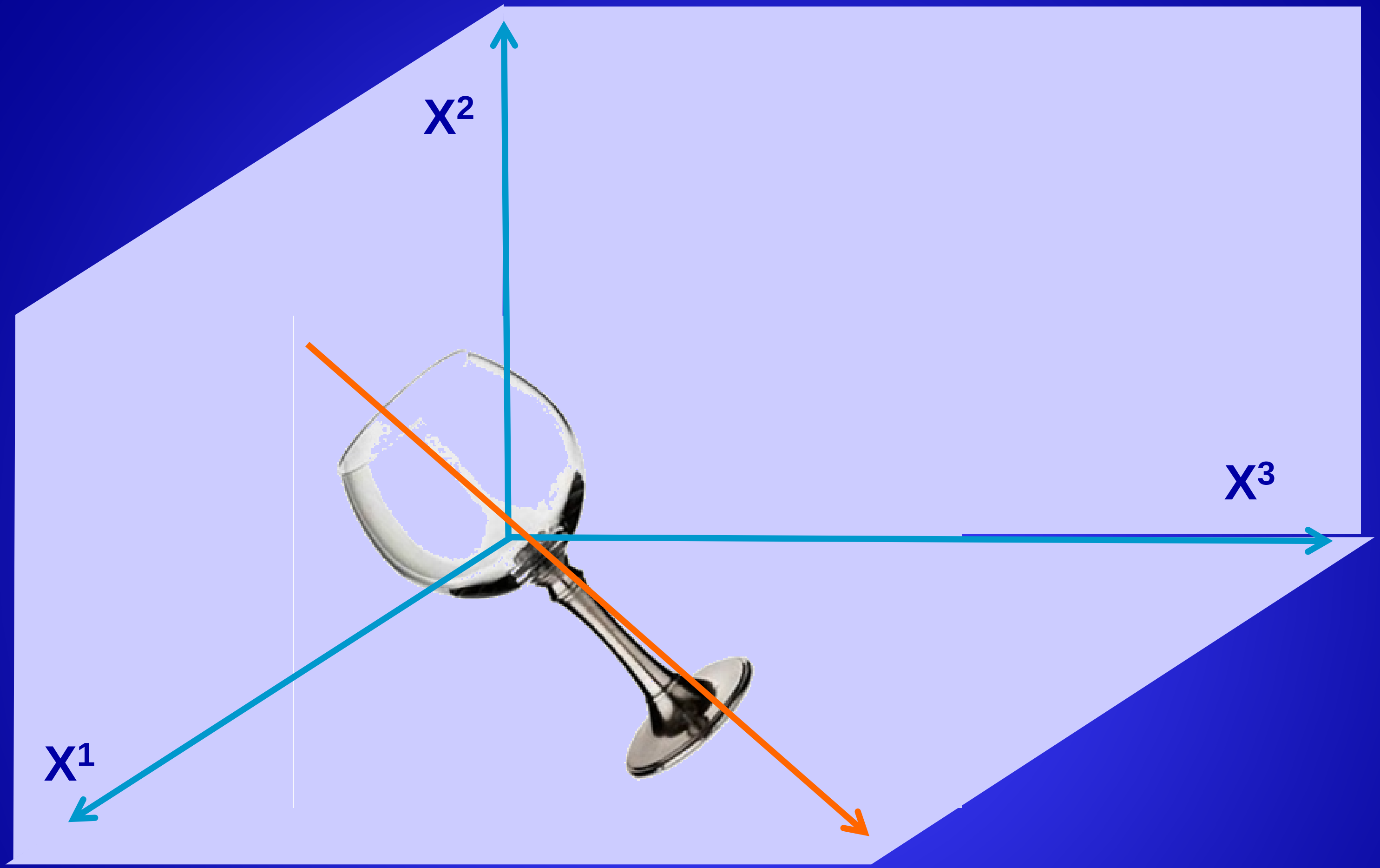
Principal components analysis



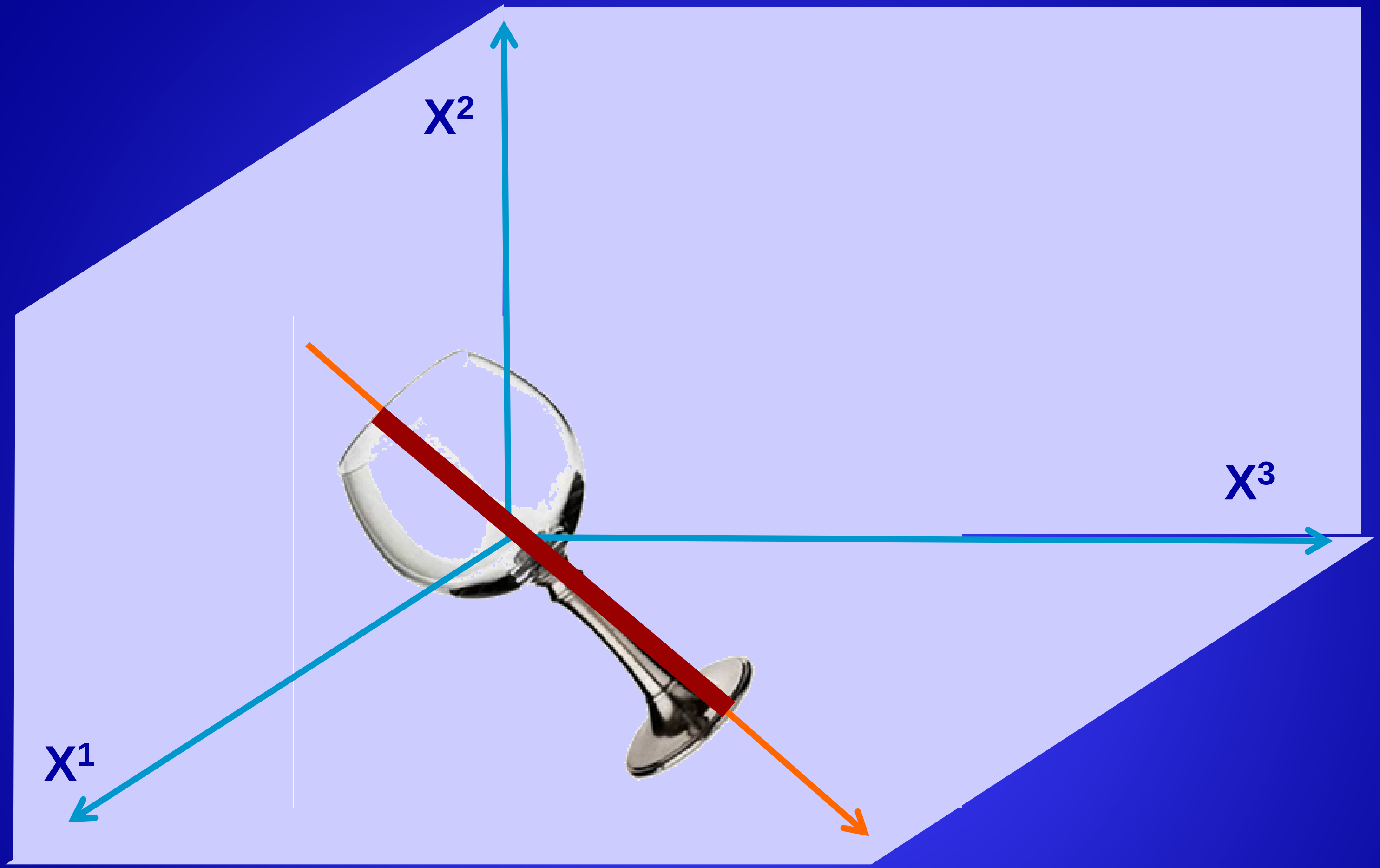
Principal components analysis



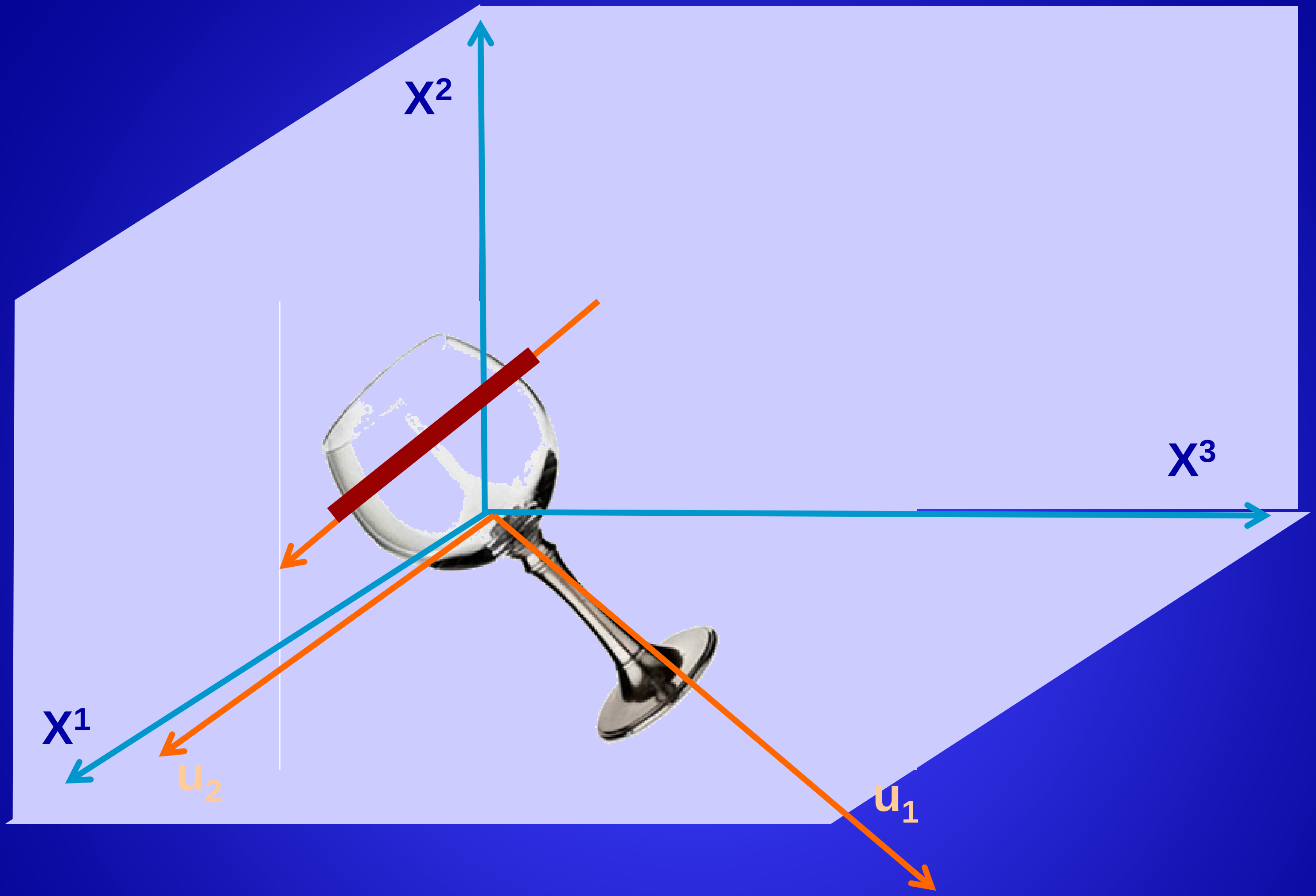
Principal components analysis



Principal components analysis

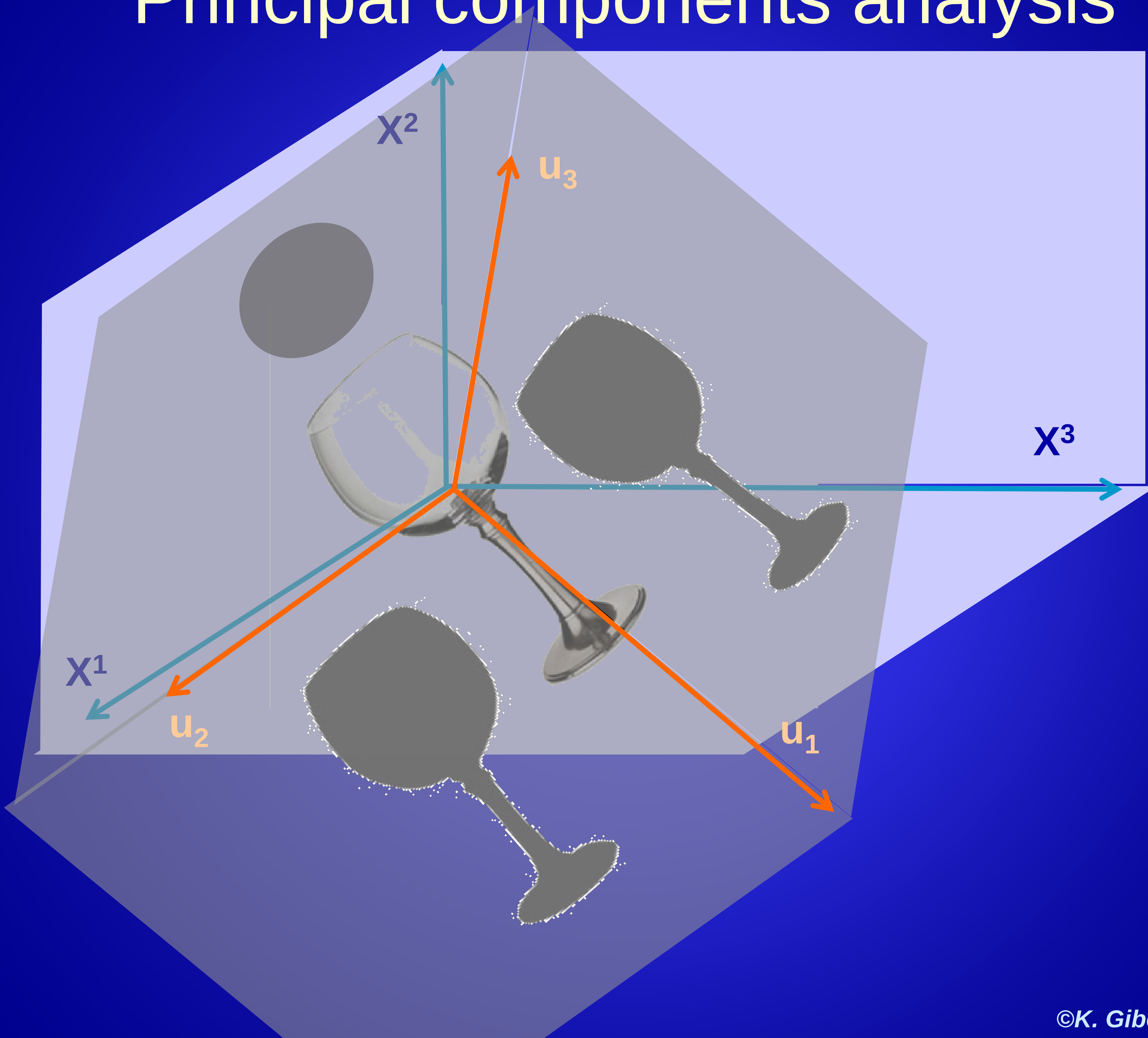


Principal components analysis

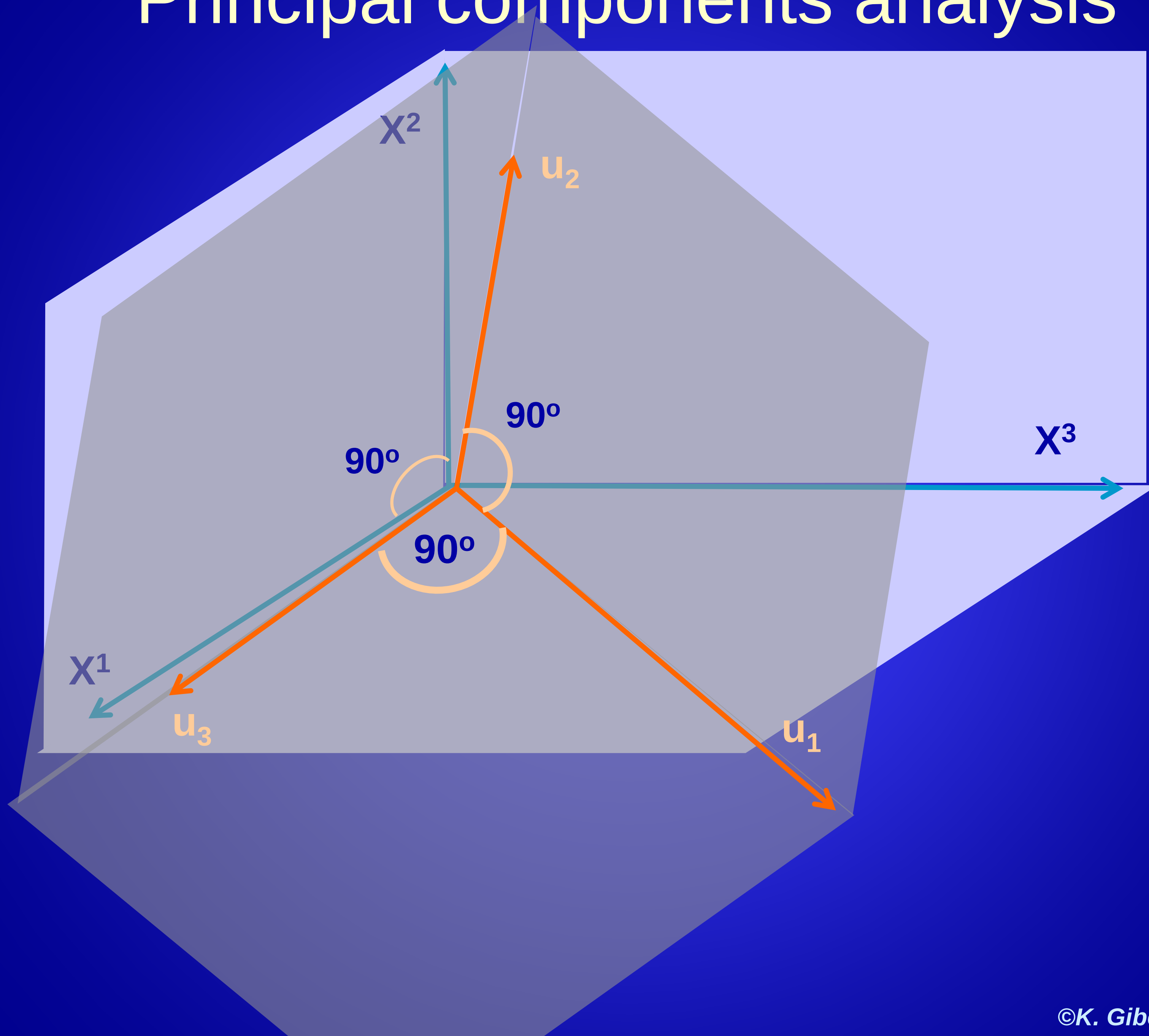


©K. Gibert

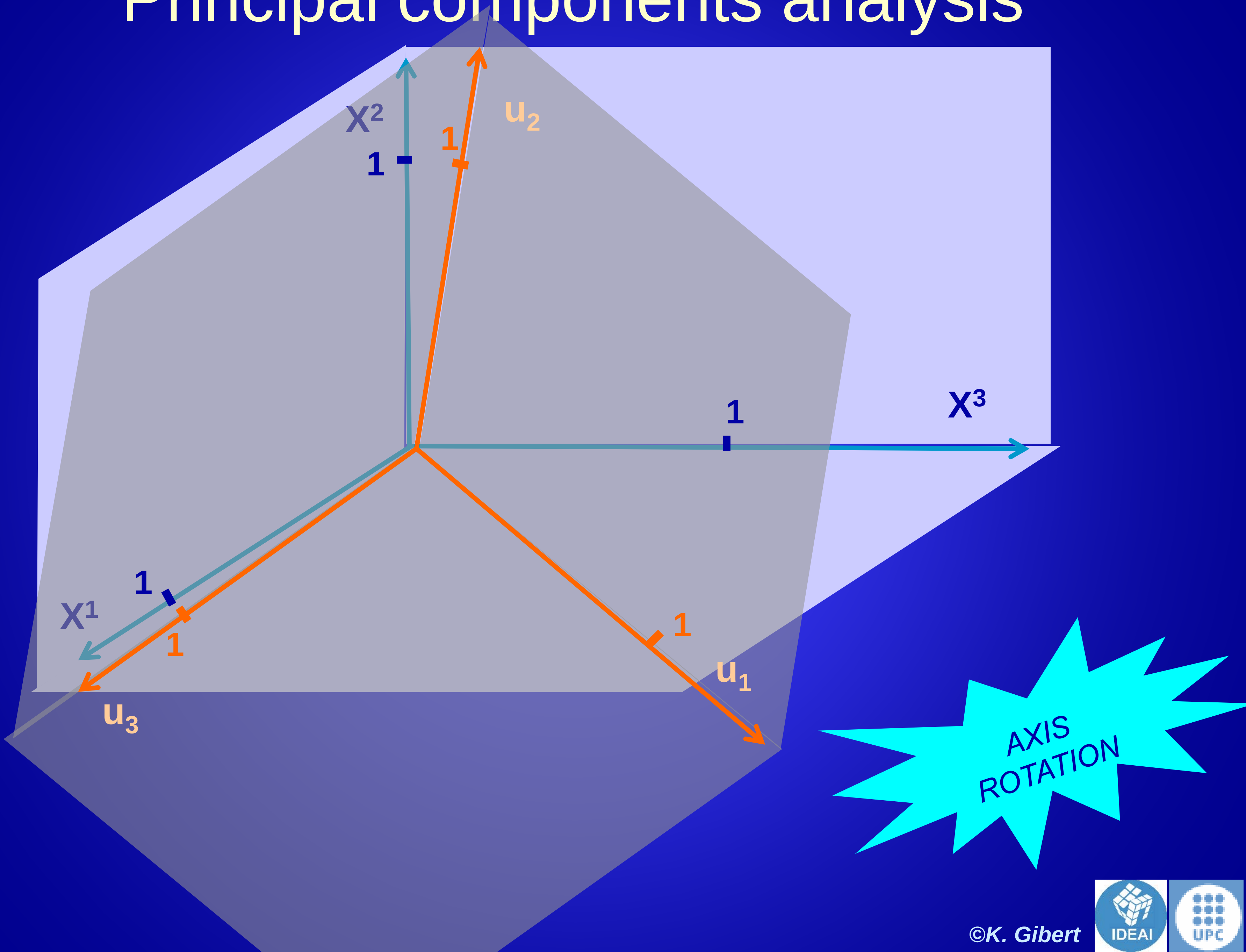
Principal components analysis



Principal components analysis



Principal components analysis



Factorial Methods

- Given <X,M,D>

Diagonalize Correlations matrix X'DX

Get r eigen values λ_α and sort decreasingly

$\{\lambda_\alpha\}_{\alpha=1:r} \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots\ldots \geq \lambda_r$

Corresponding eigenvectors $u_\alpha = (u_{\alpha 1} \ldots\ldots u_{\alpha p})$

for $M = \mathbb{I}_p$: $u^*_\alpha = u_\alpha$; *for* $M \neq \mathbb{I}_p$: $u^*_\alpha = M^{-½} u_\alpha$

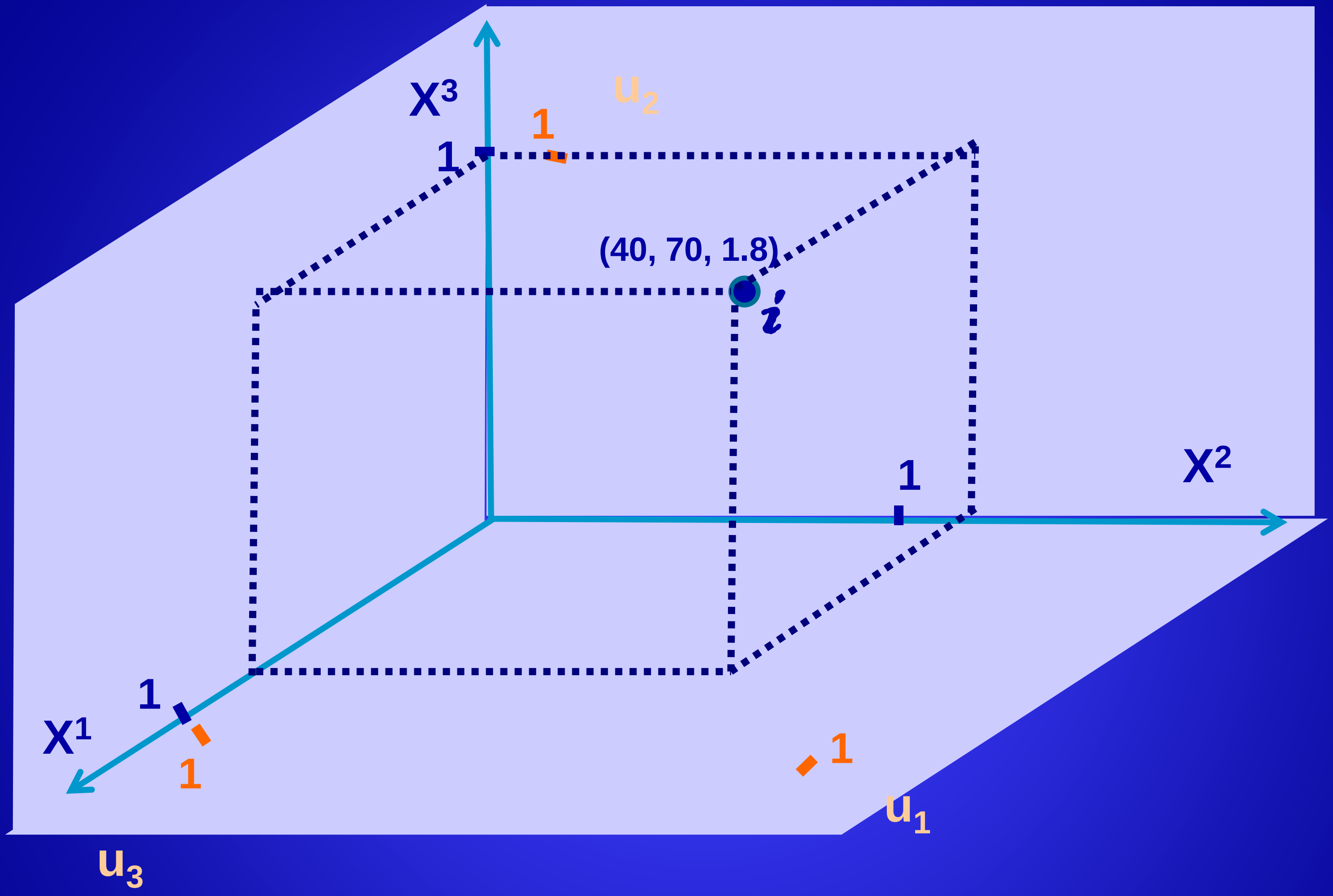
$\{u^*_\alpha\}_{\alpha=1:r}$ *orthonormal base for individuals*

u^*_α *are the principal factors of X : good rotation directions*

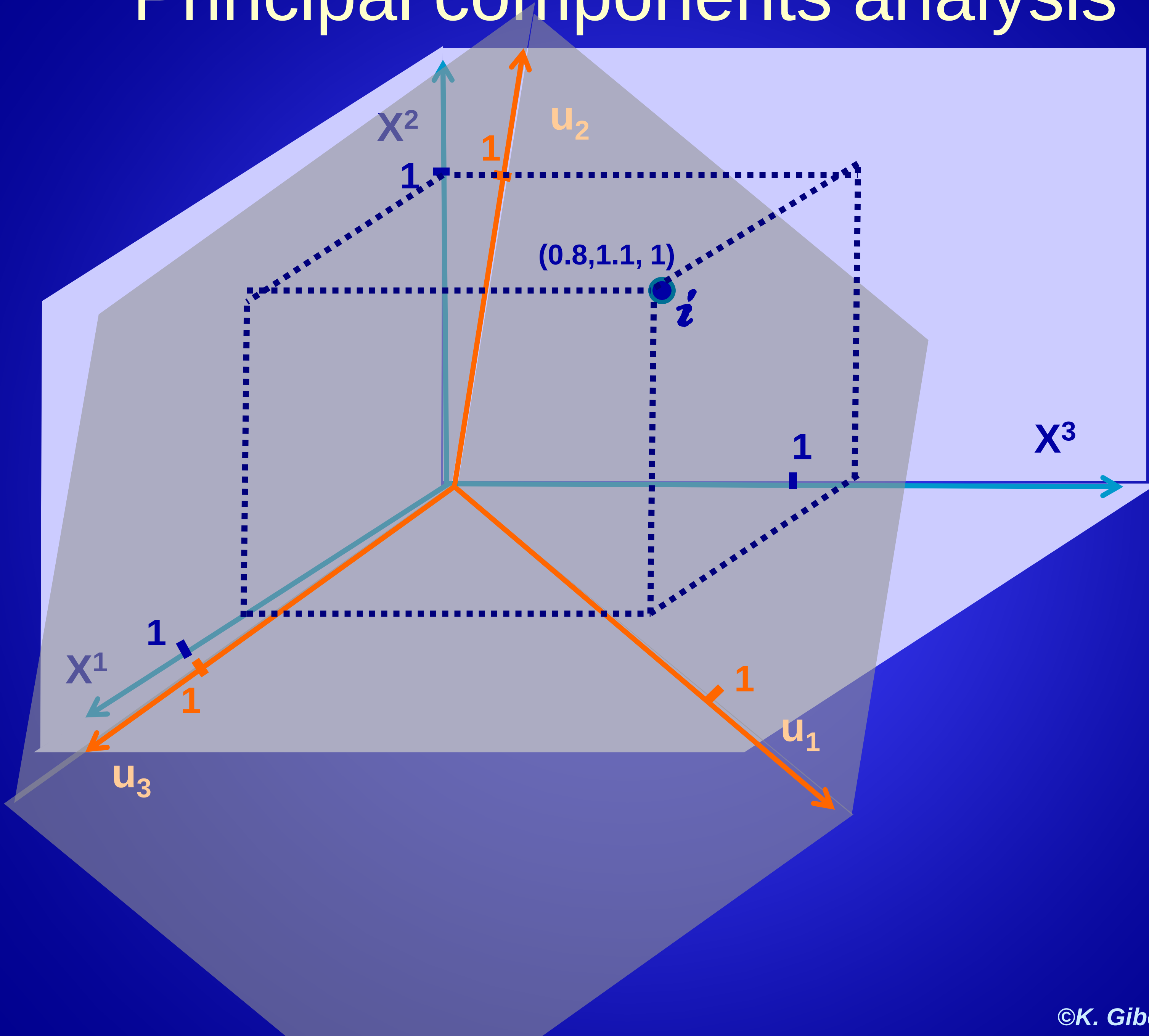
$U^* = ([u^*_1]\ [u^*_2] \ldots\ldots [u^*_r])$ *is the basis for the projection space*

How is i expressed in rotated space?

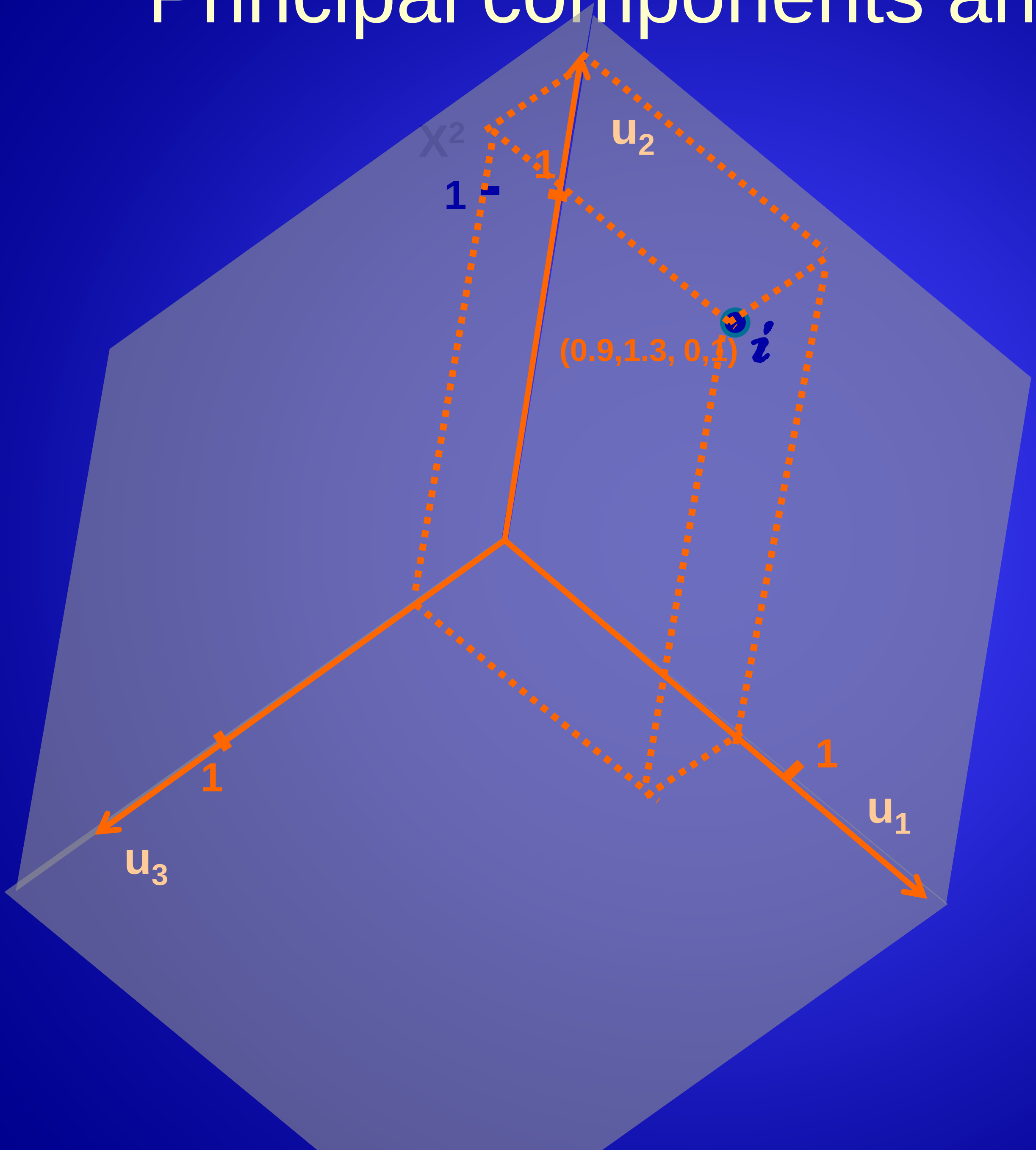
Principal components analysis



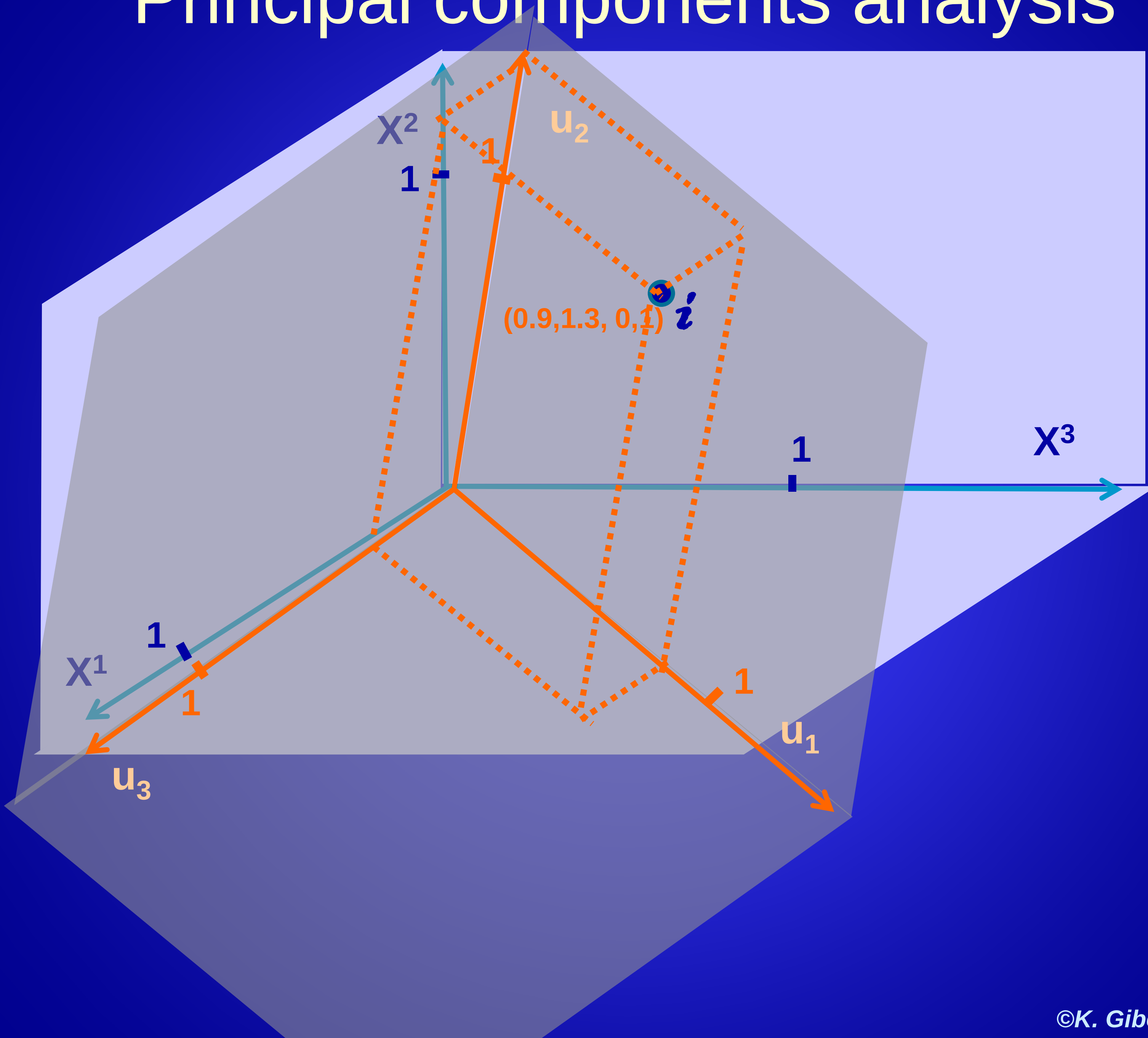
Principal components analysis



Principal components analysis

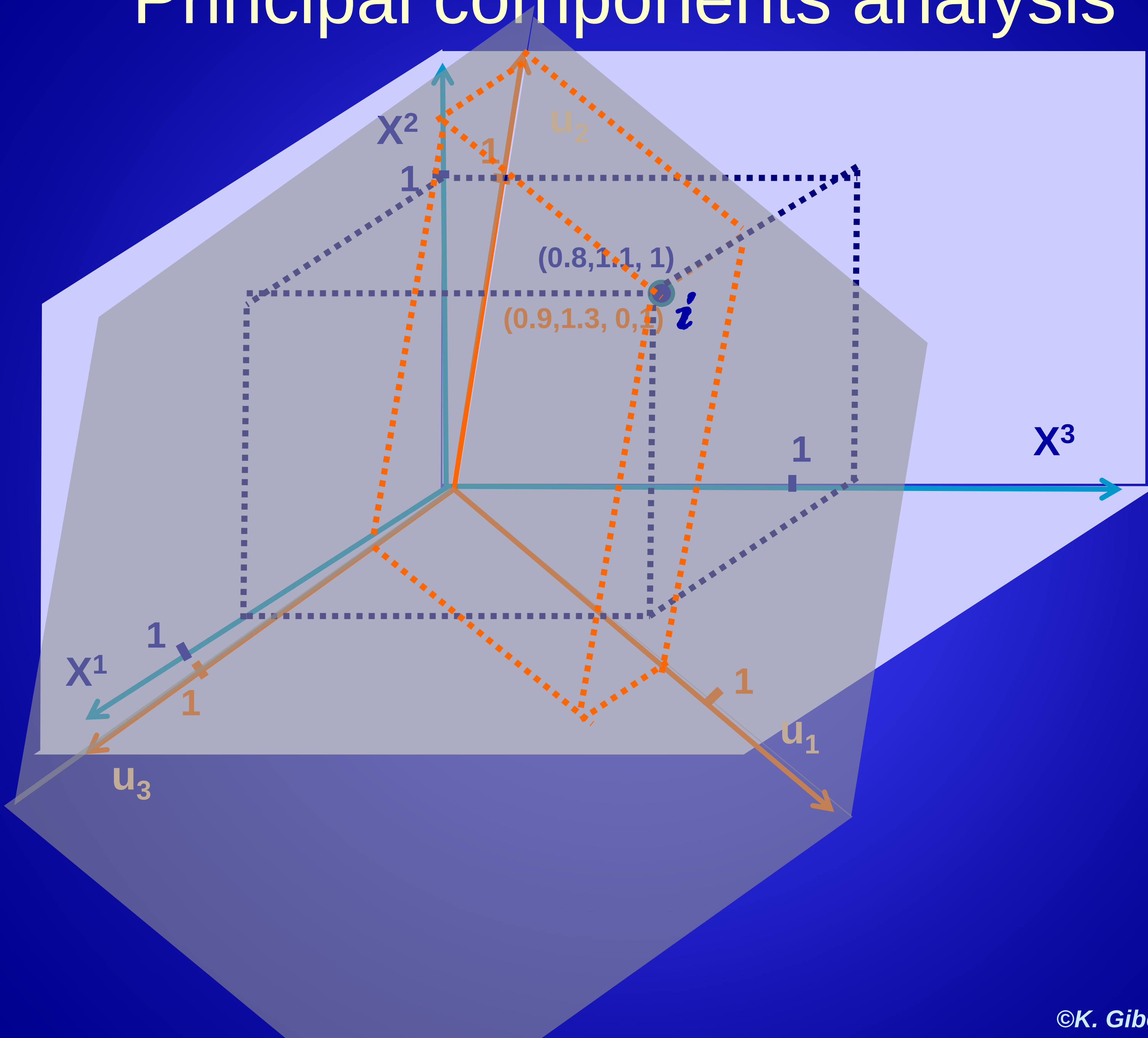


Principal components analysis



IDEAI UPC

Principal components analysis



Factorial Methods

- Given <X,M,D>

Can we find coordinates in rotated space from original ones?

The projection matrix $P = U^*_k U^{*\prime}_k M$

Projection of a single individual: $Pr(i)= U^*_k U^{*\prime}_k M x_i$

Projection of all individuals: $Pr(X)= U^*_k U^{*\prime}_k M X'$

Get a matrix with projections in ROWS: $Pr(X)'= XMU^*_k U^{*\prime}_k$

Projections expressed in original vectorial space

The best possible projection over k dimensions

Factorial Methods

- Given <X,M,D>

Matrix $XMU^*_k U^{*\prime}_k$ provides the best possible k-projection of X

Silver-Smidth norm: $||X||^2_{MD} = \sum_{\alpha=1}^{r} \lambda_\alpha$

Measures variability, information contained in X

Property: $||XMU^*_k U^{*\prime}_k||^2_{MD} = ||X||^2_{MD}$

Any other k-projection of X

- Provides smallest values of Silver-Smidth norm
- Has less variability
- Keeps smallest information from X

Factorial Methods

- Given <X,M,D>

Diagonalize correlations matrix (with normalized data)

eigenvectors $u_\alpha = (u_{\alpha 1} \ldots u_{\alpha p})$ *(direction of factor α, α =1:p)*

$u_{\alpha p}$: contribution of variable p to the factor α

$(u_1 \ldots\ldots\ldots u_k)$ *ortonormal*

eigen values λ_k *(quantity of information converved by factor k)*

(Projected inertia)

$\{\lambda_\alpha\}_{\alpha=1:r} \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r$

$\Sigma_{\forall\alpha}\lambda_\alpha$ = Total inertia of X (information in data)

Close objects project close
proximity linked with association

Factorial Methods

- Given <X,M,D>

  eigenvectors $u_\alpha = (u_{\alpha 1} \dots u_{\alpha p})$ *(direction of factor k)*

  $u_{\alpha p}$: contribution of variable p to the factor α

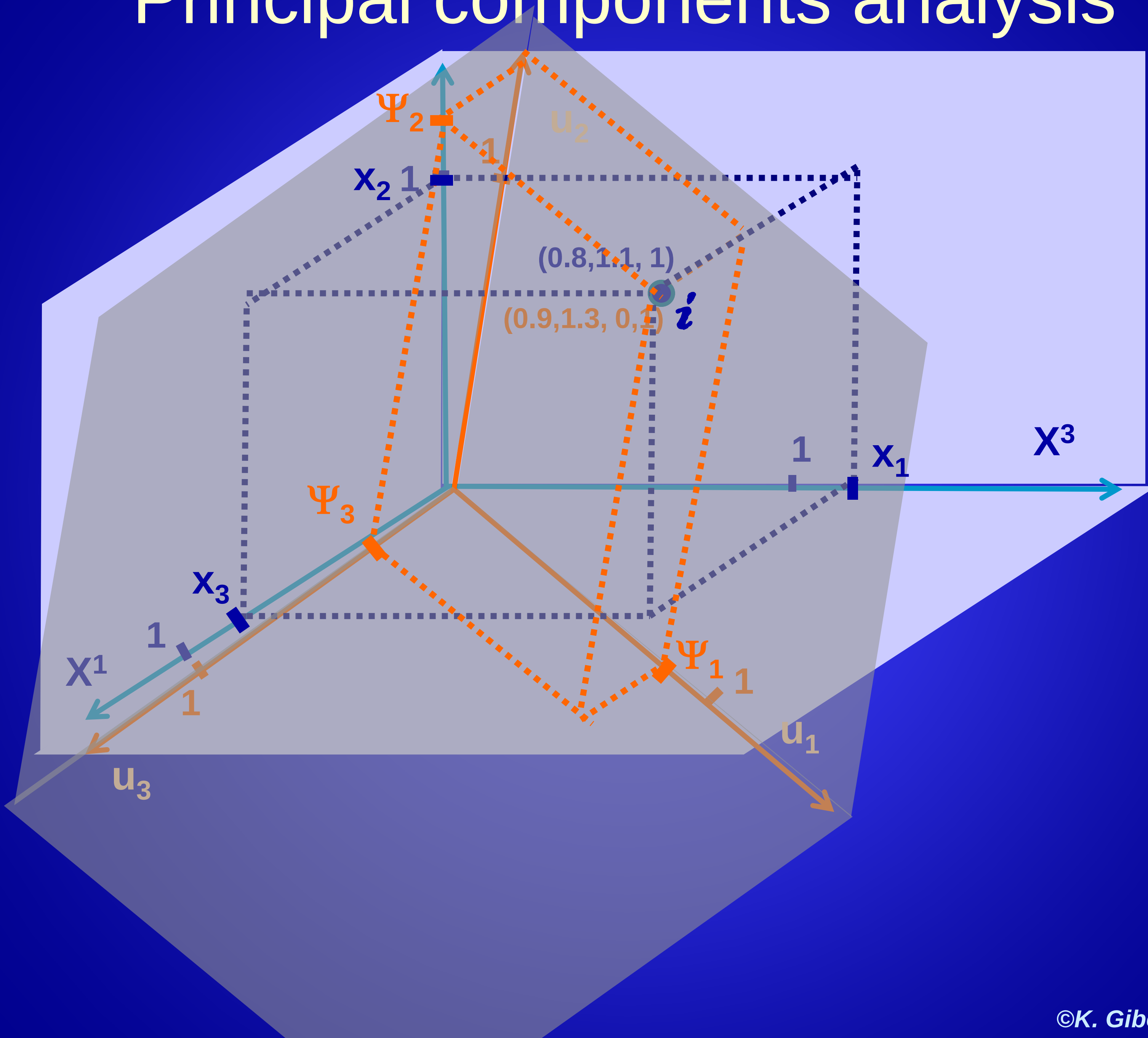
  eigen values λ_α *(quantity of information conserved by factor* α*)*

  (Projected inertia)

  $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_r$

  $\Sigma_{\forall\alpha}\ \lambda_\alpha$ = Total inertia of X (information of data)

Principal components analysis



Factorial Methods

- Given <X,M,D>

$$i = (x_{1i}, \dots, x_{pi})$$

Points in projected space: $i = (\Psi_{1i}, \dots, \Psi_{\alpha i}, \dots, \Psi_{ri})$ (often r=p)

$$\Psi_{\alpha i} = x_{1i}u_{\alpha 1} + x_{2i}u_{\alpha 2} + \cdots + x_{pi}u_{\alpha p} \qquad \psi_{\alpha} = Xu_{\alpha}$$

Then $\Psi_{\alpha}^{'} D\Psi_{\alpha} = \lambda_{\alpha}$

Illustrative points z also projectable

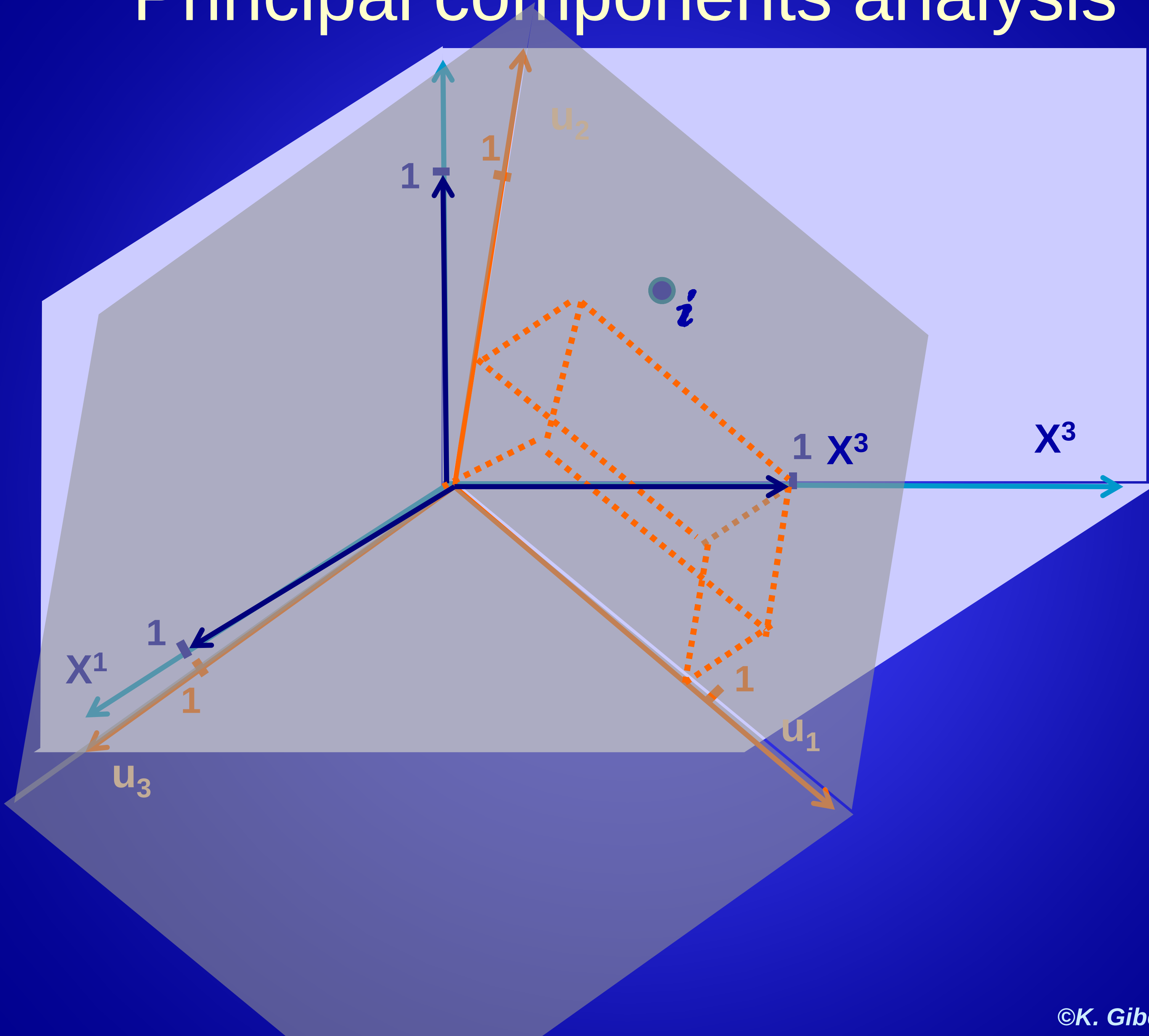
$$\Psi_{\alpha z} = x_{1z}u_{\alpha 1} + x_{2z}u_{\alpha 2} + \cdots + x_{pz}u_{\alpha p}$$

Factors are linear combinations of original variables

Original variables project as VECTORS over factorial space

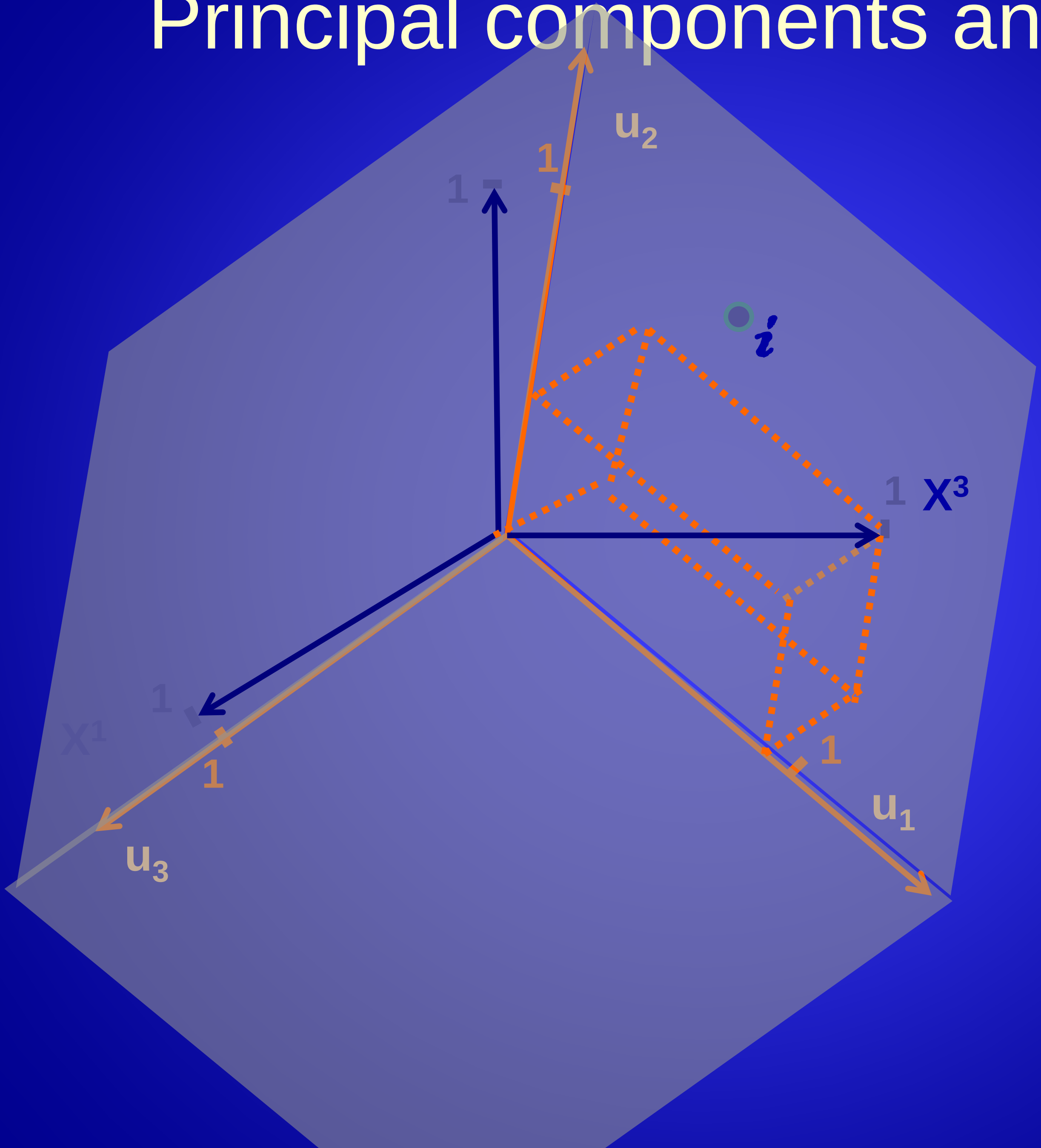
angle and lenght important

Principal components analysis

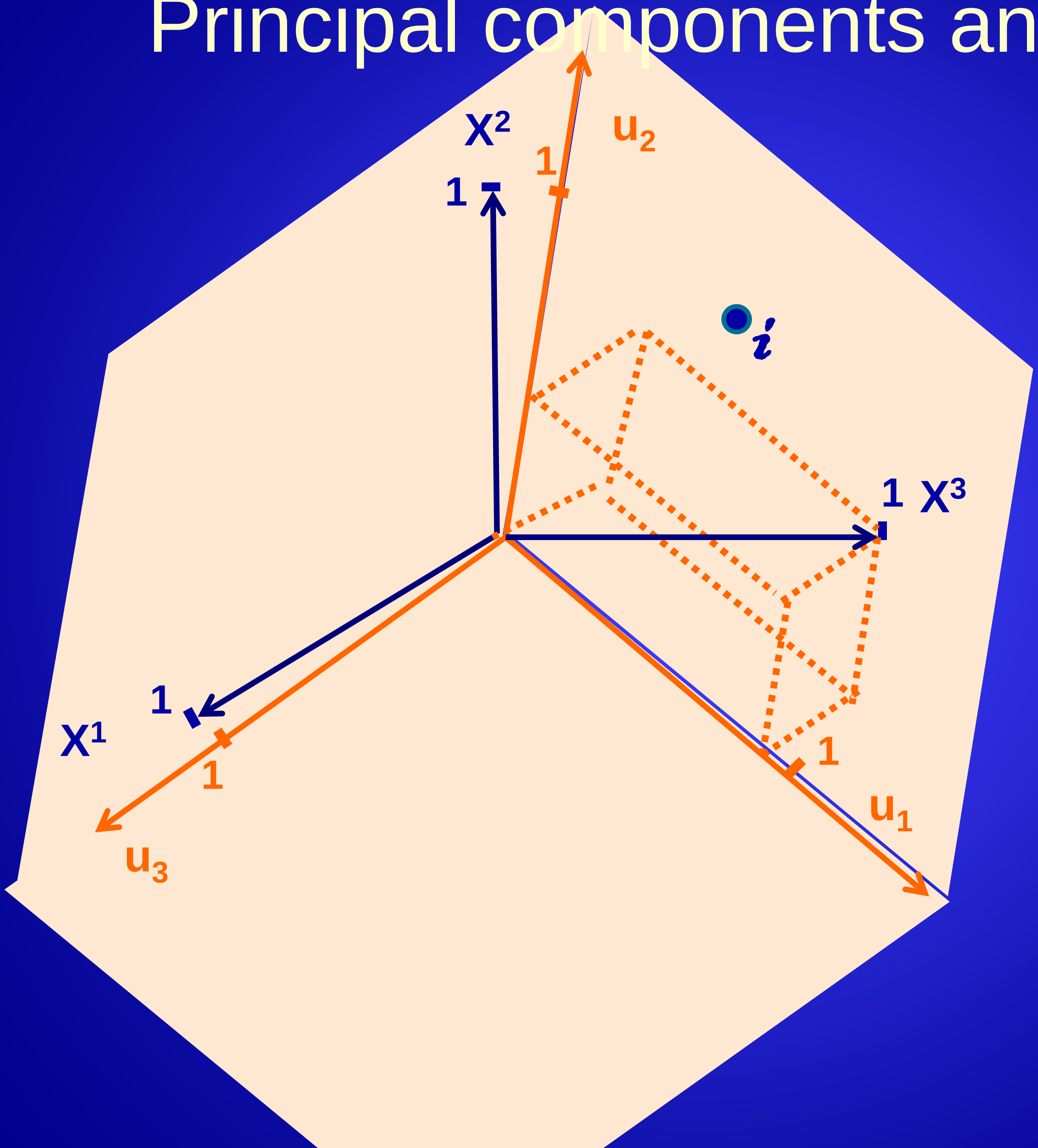


©K. Gibert

Principal components analysis

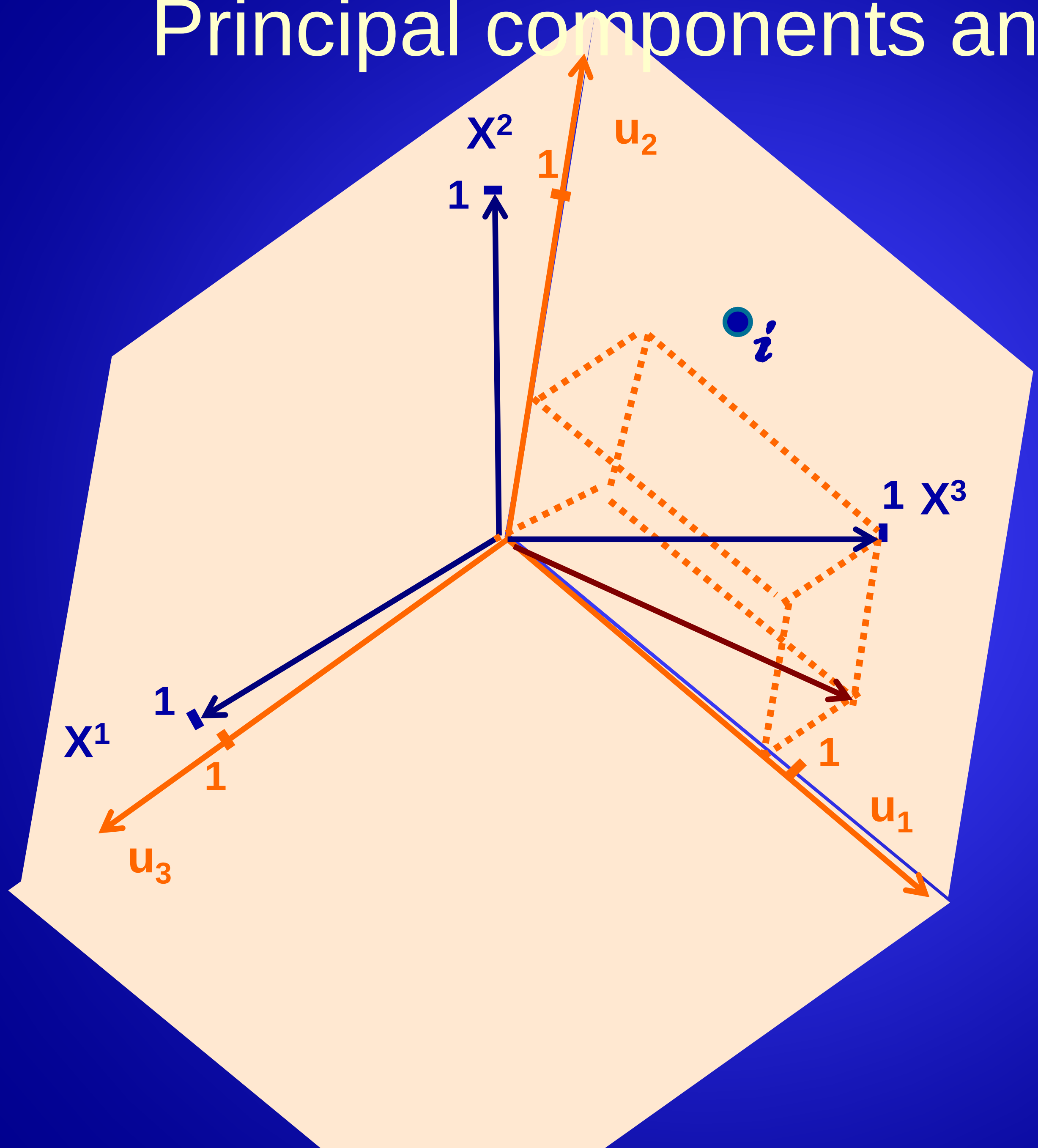


Principal components analysis



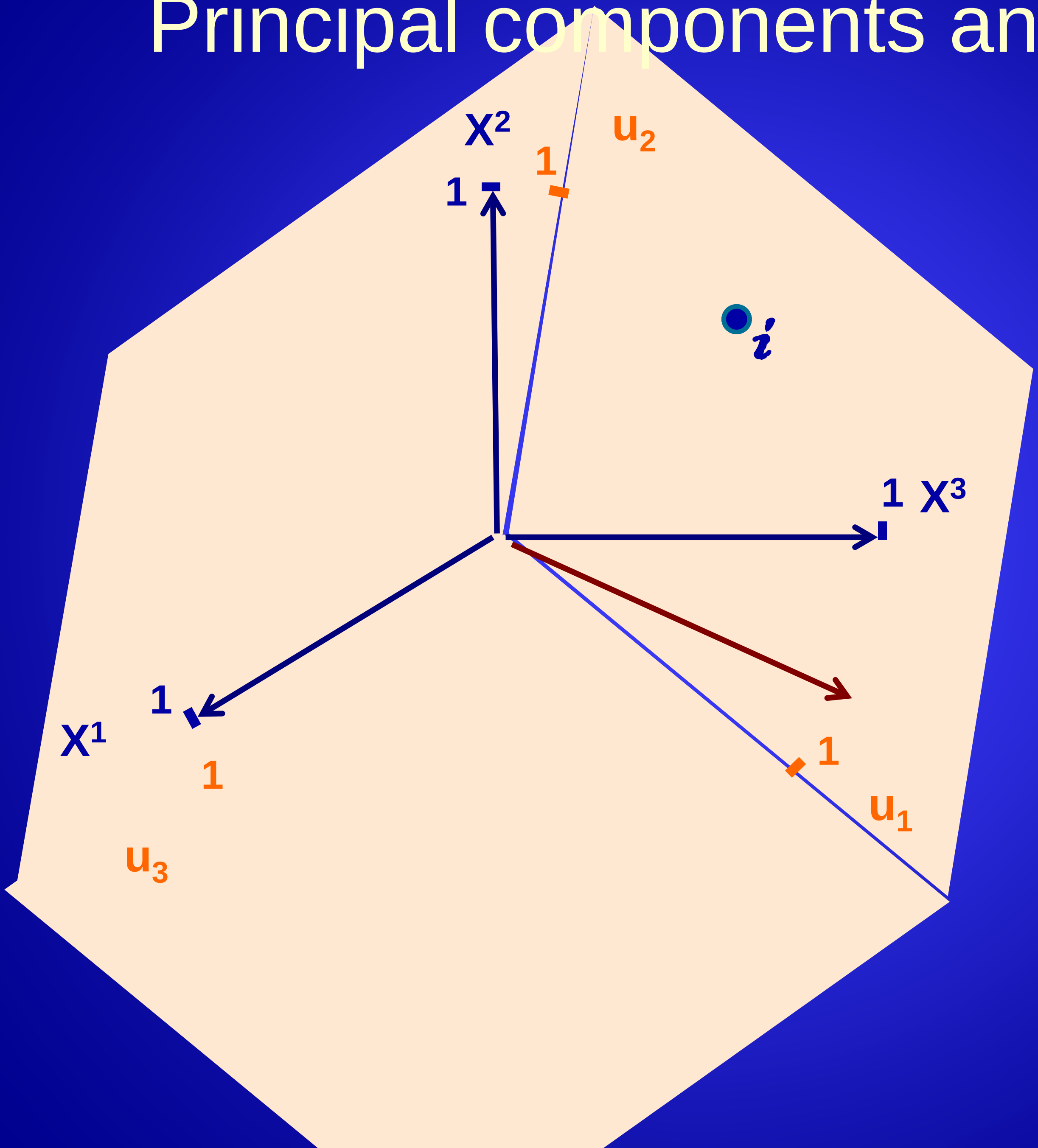
©K. Gibert

Principal components analysis

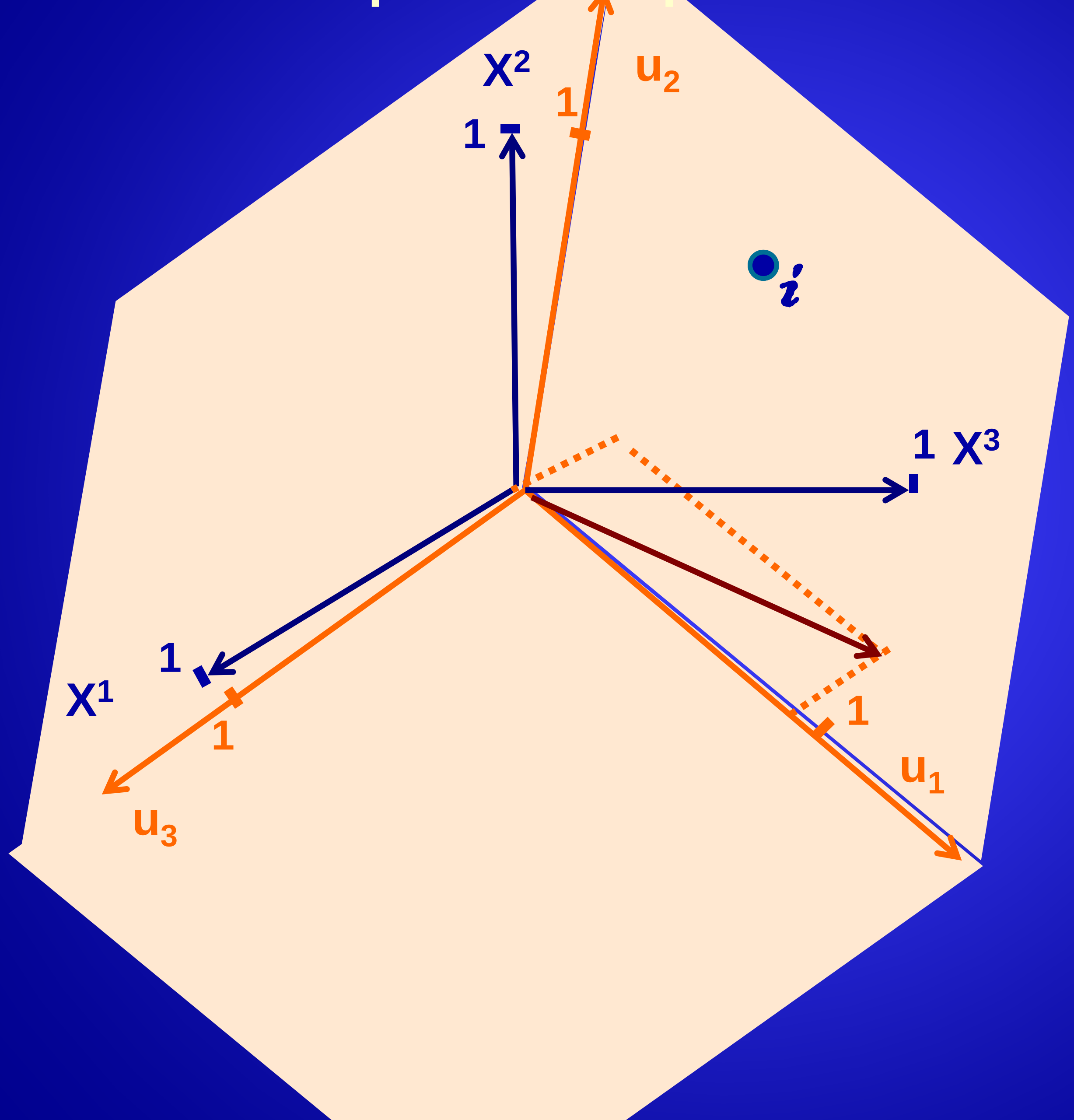


IDEAI UPC

Principal components analysis

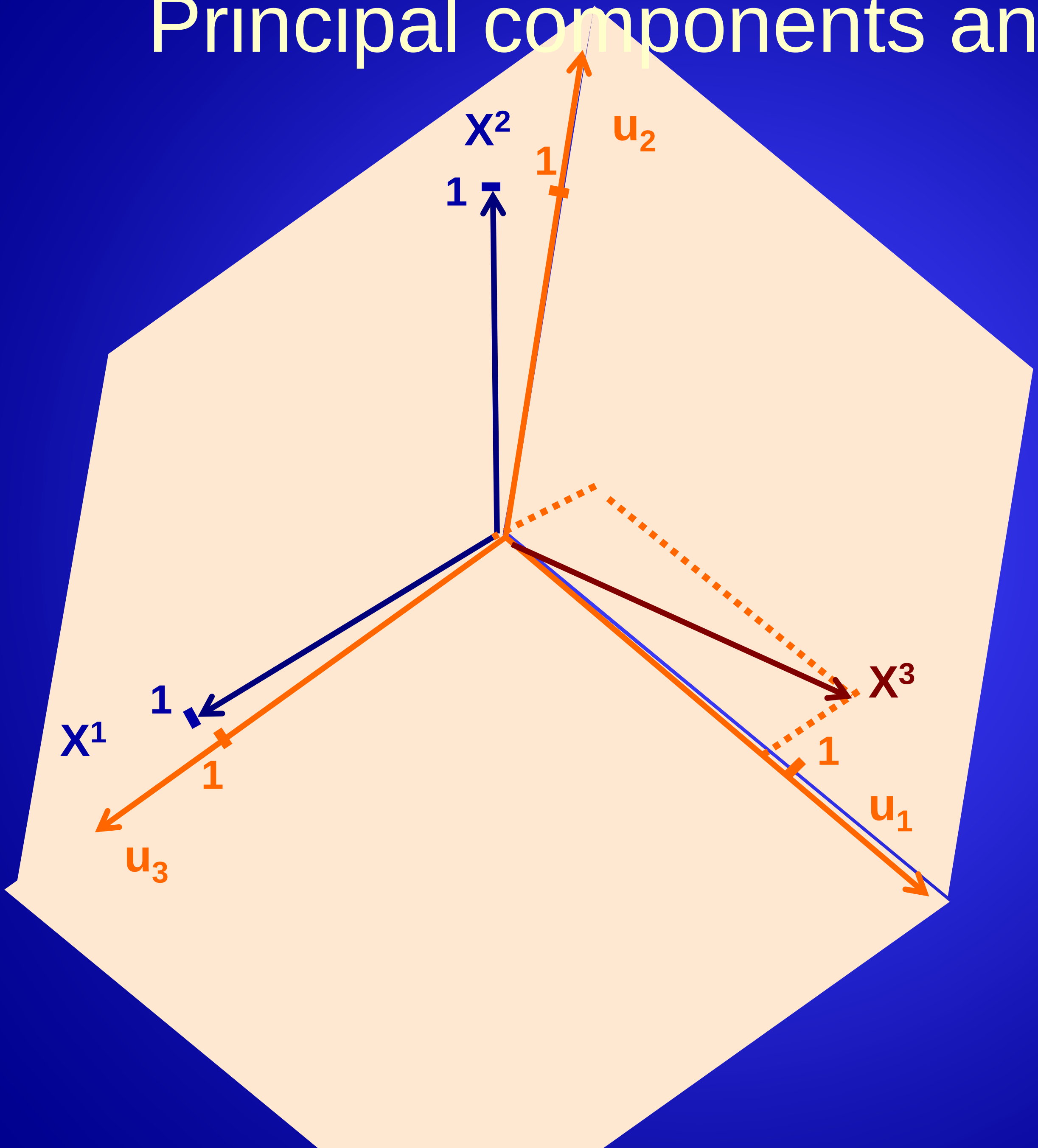


Principal components analysis

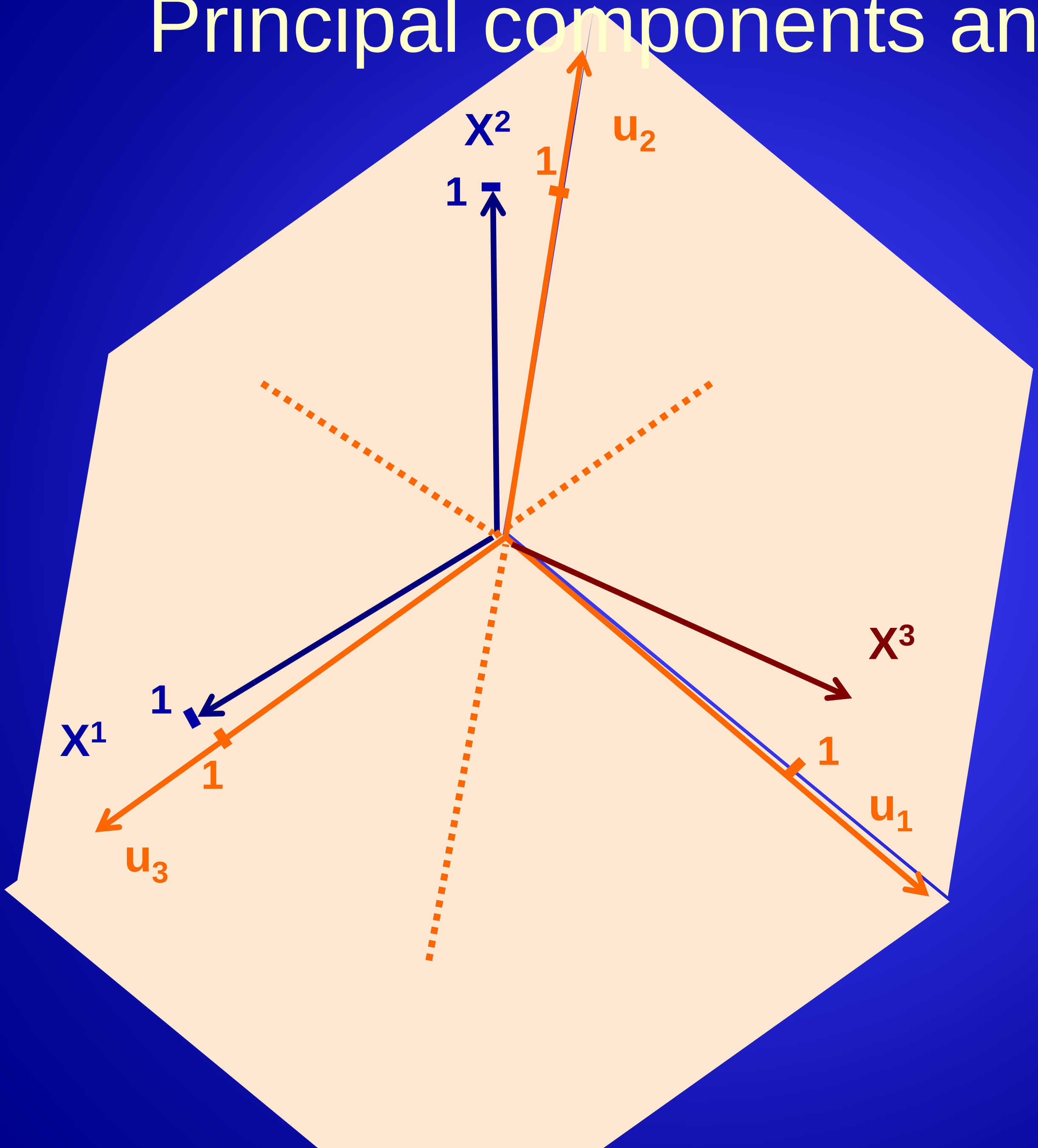


©K. Gibert

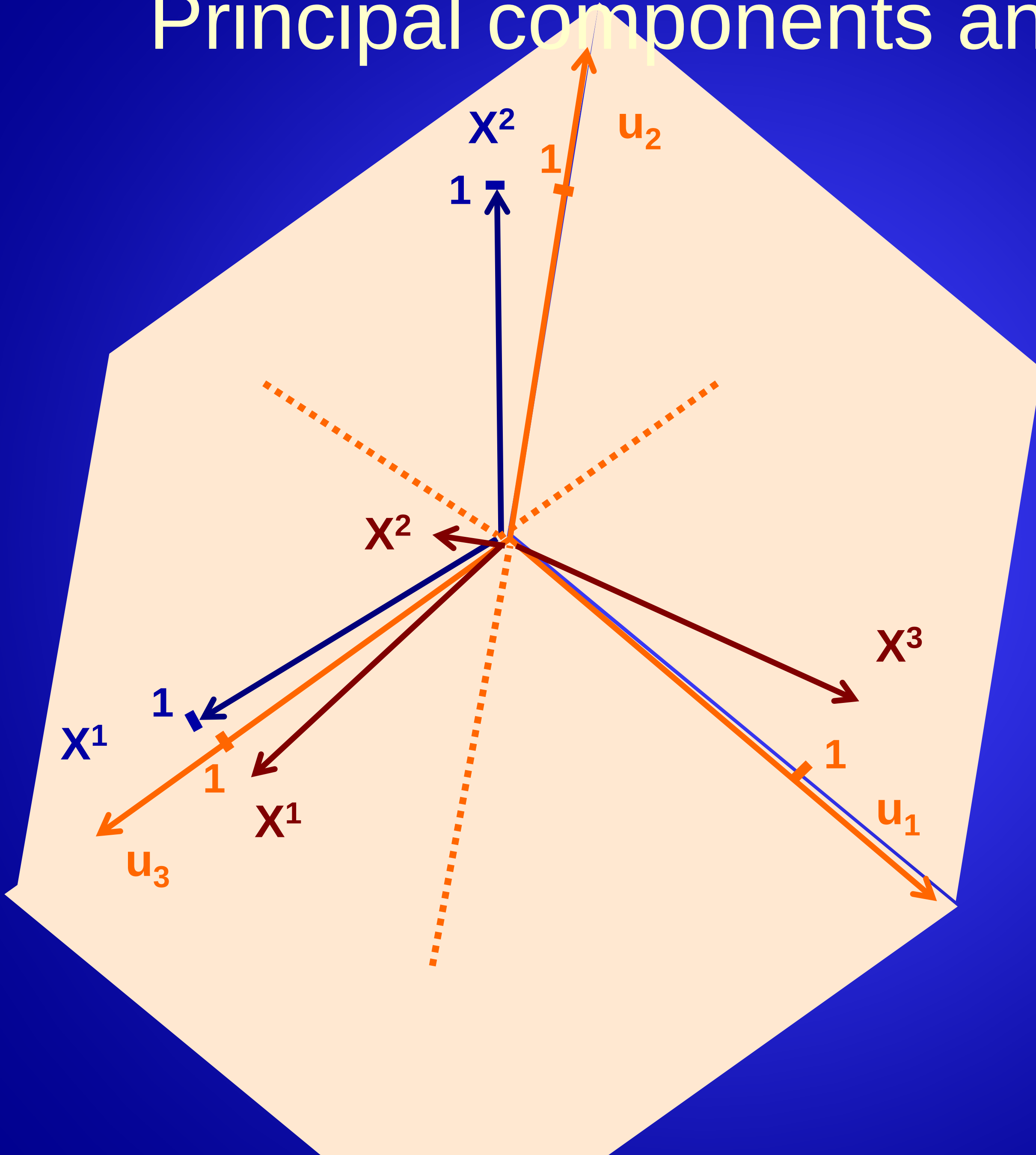
Principal components analysis



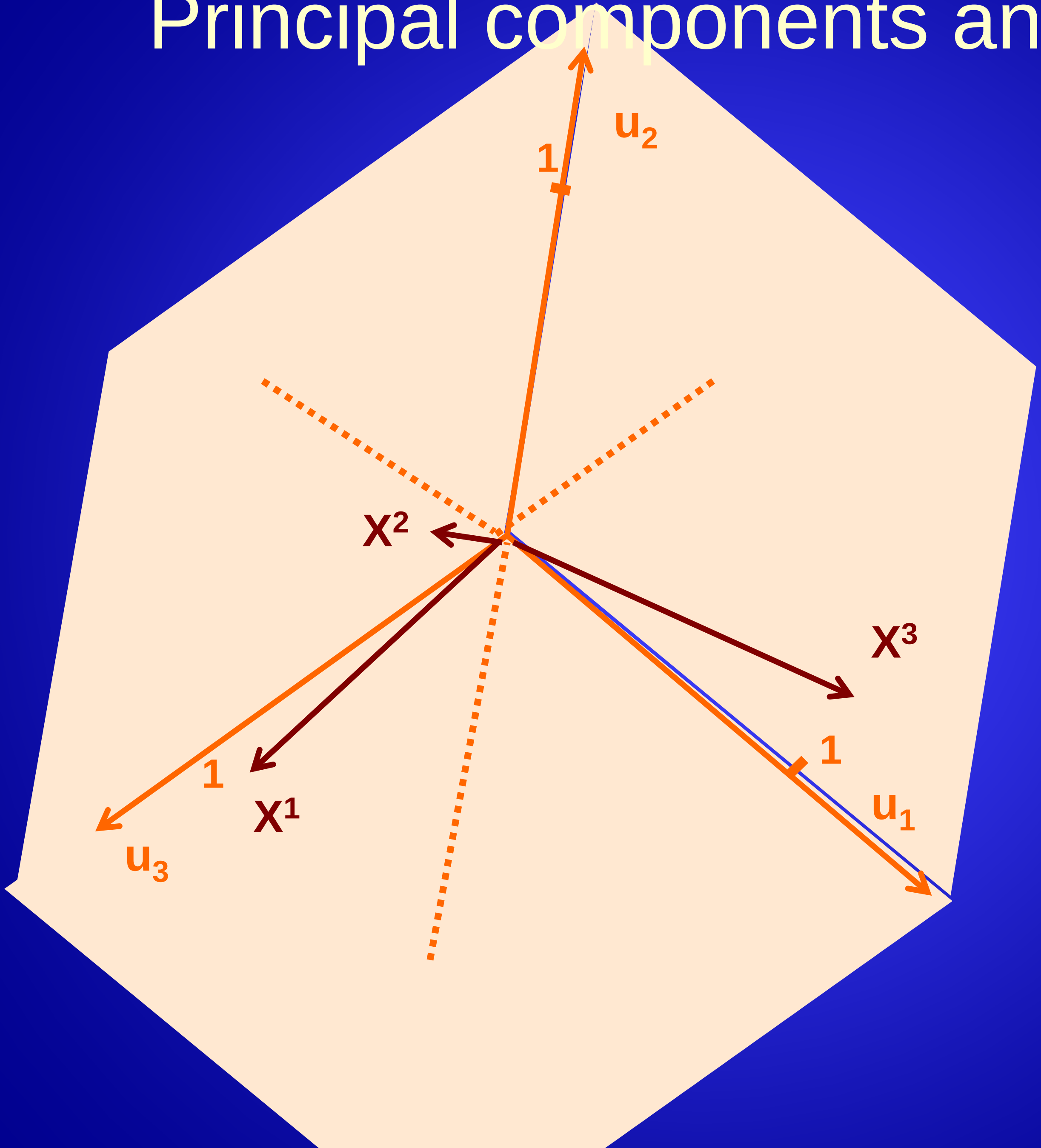
Principal components analysis



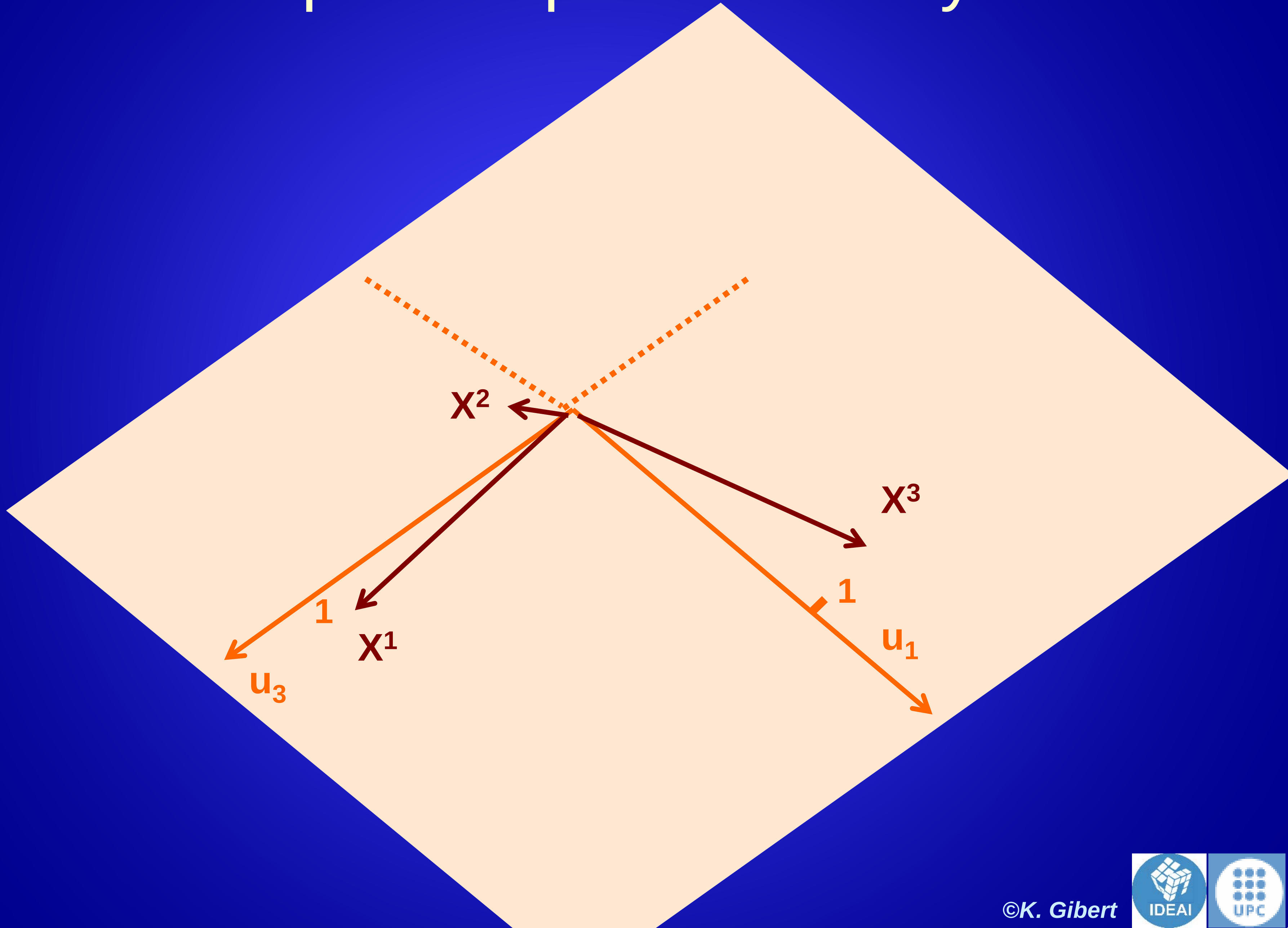
Principal components analysis



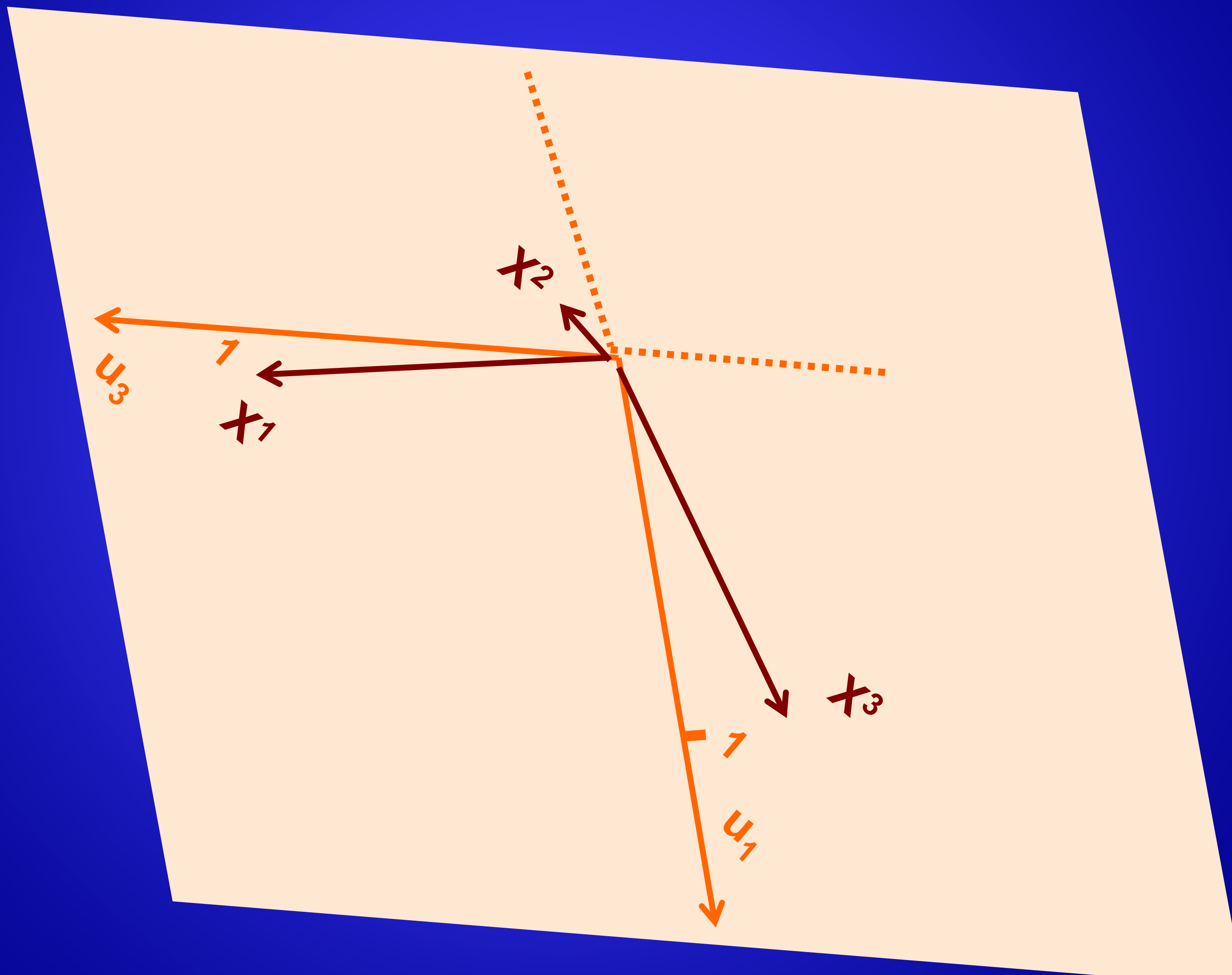
Principal components analysis



Principal components analysis

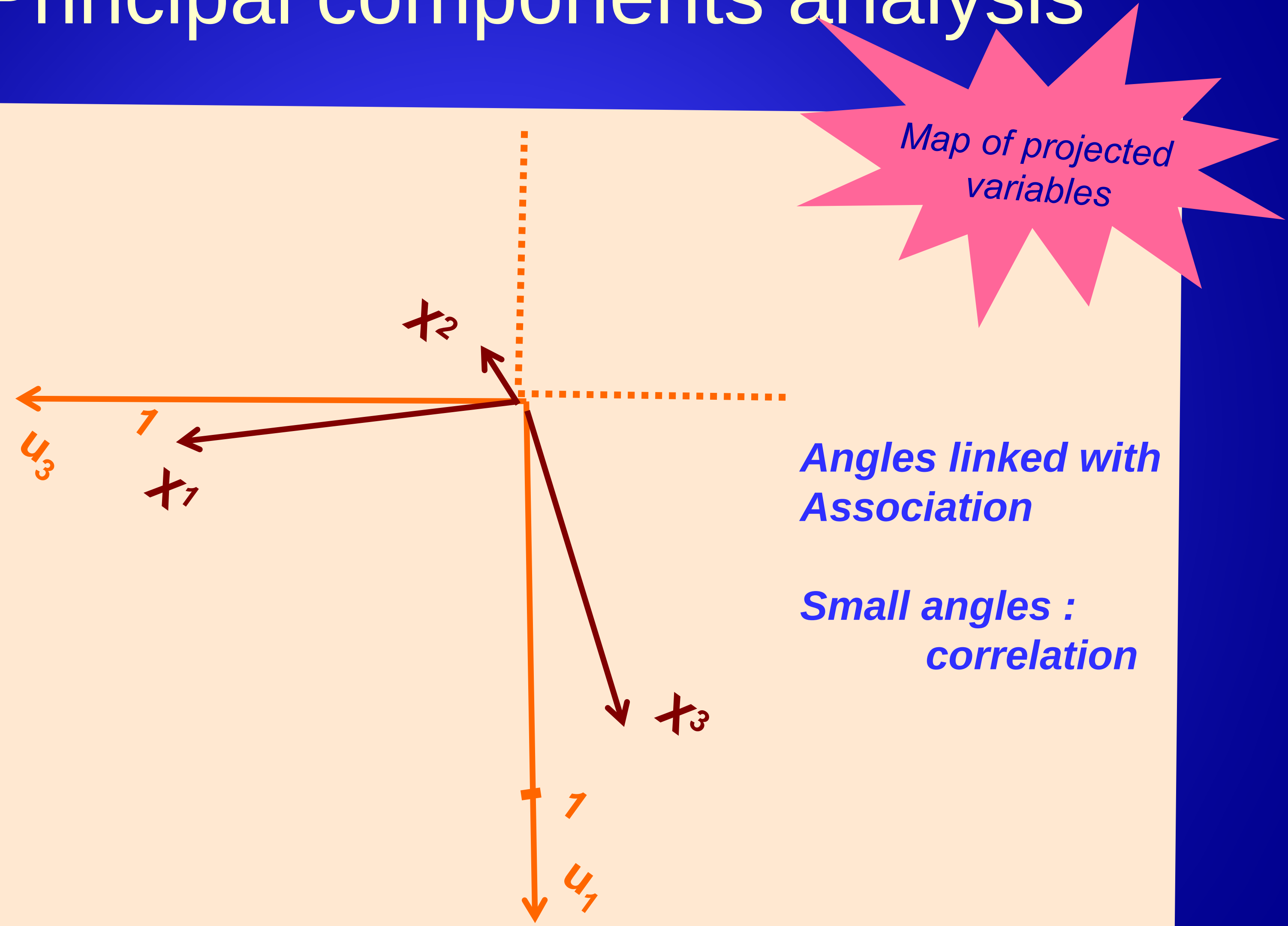


Principal components analysis



©K. Gibert

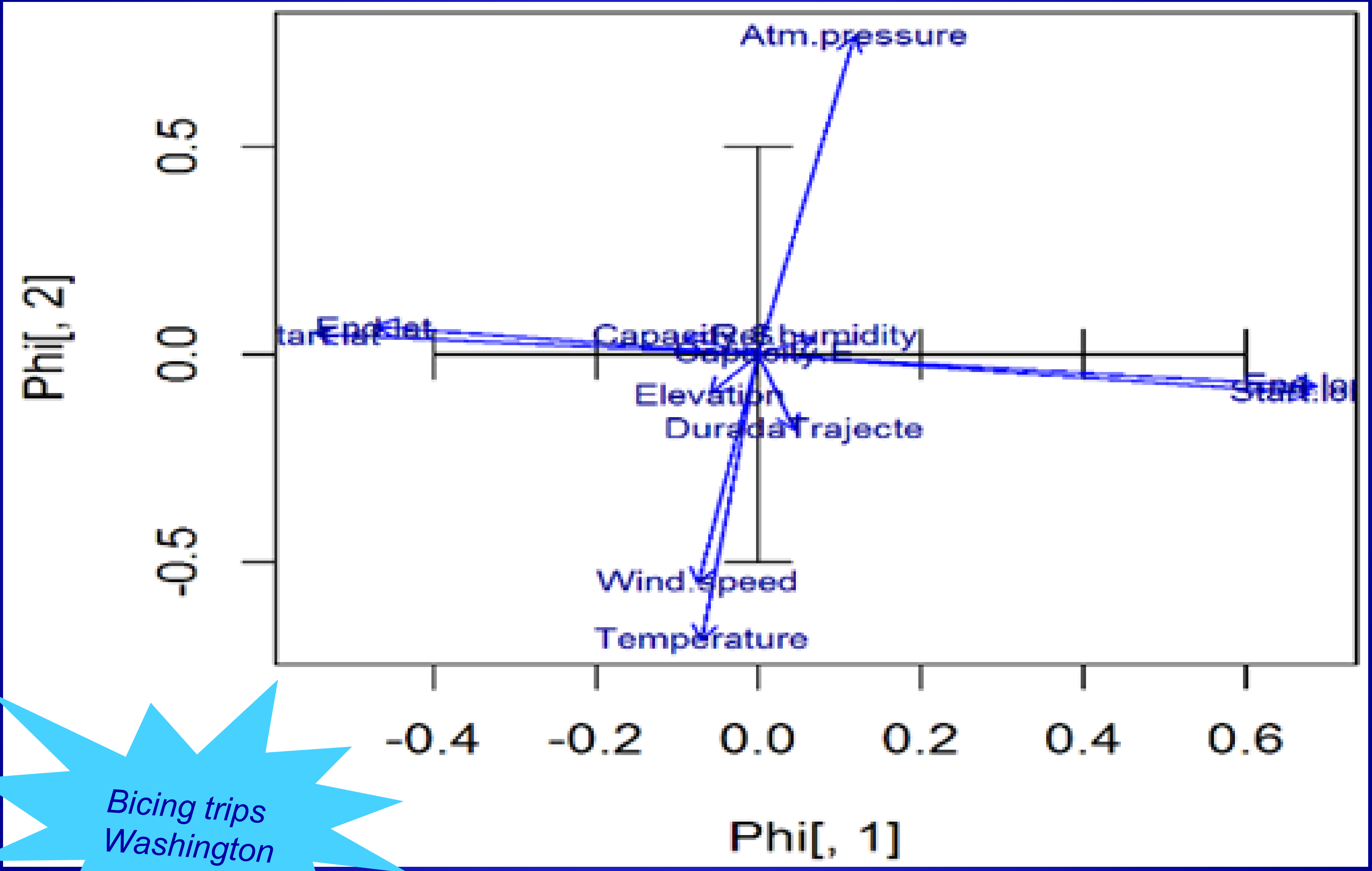
Principal components analysis



Angles linked with Association

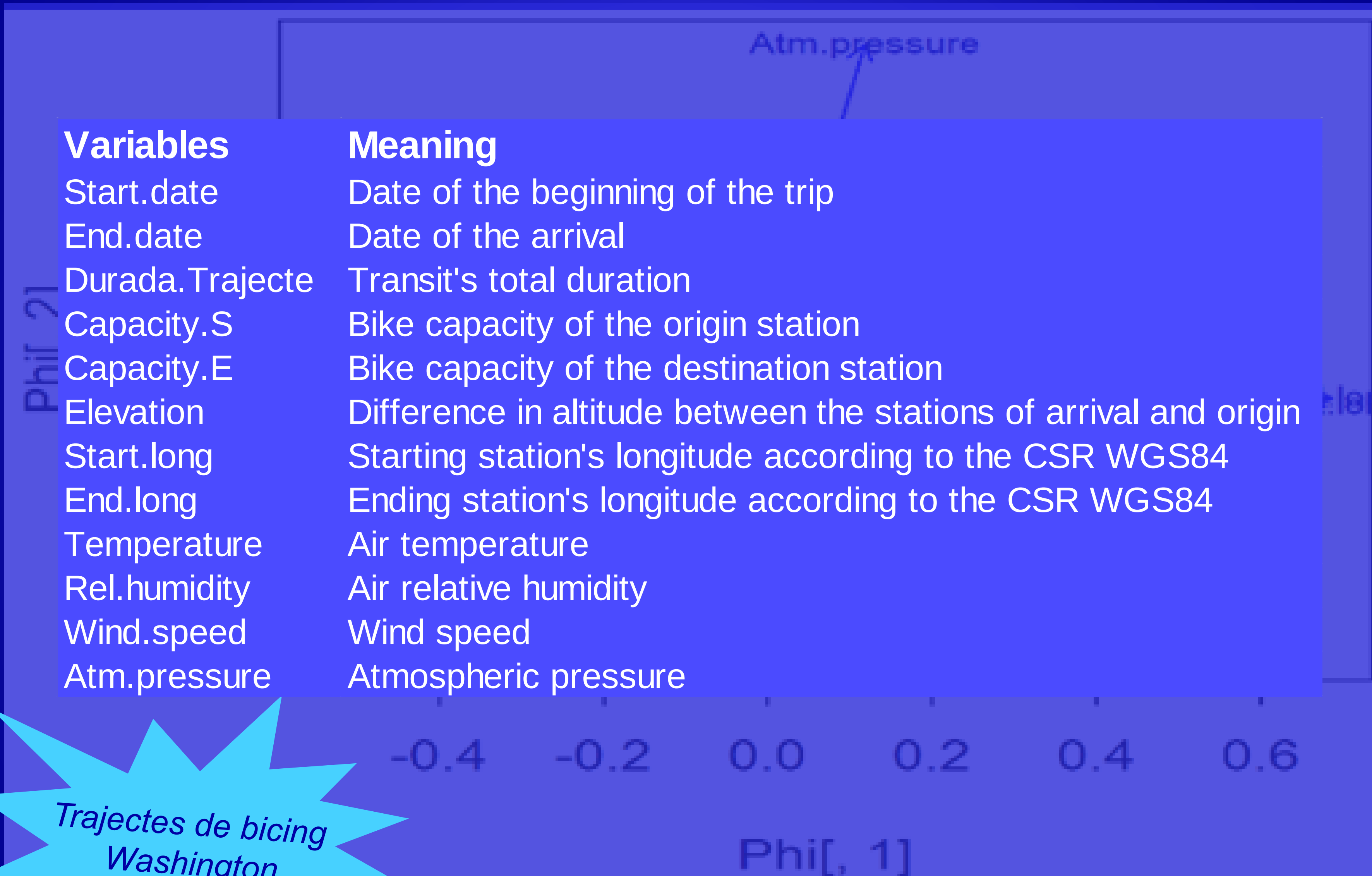
Small angles : correlation

Principal components analysis



Bicing trips Washington

Principal components analysis



Variables	Meaning
Start.date	Date of the beginning of the trip
End.date	Date of the arrival
Durada.Trajecte	Transit's total duration
Capacity.S	Bike capacity of the origin station
Capacity.E	Bike capacity of the destination station
Elevation	Difference in altitude between the stations of arrival and origin
Start.long	Starting station's longitude according to the CSR WGS84
End.long	Ending station's longitude according to the CSR WGS84
Temperature	Air temperature
Rel.humidity	Air relative humidity
Wind.speed	Wind speed
Atm.pressure	Atmospheric pressure

Trajectes de bicing Washington

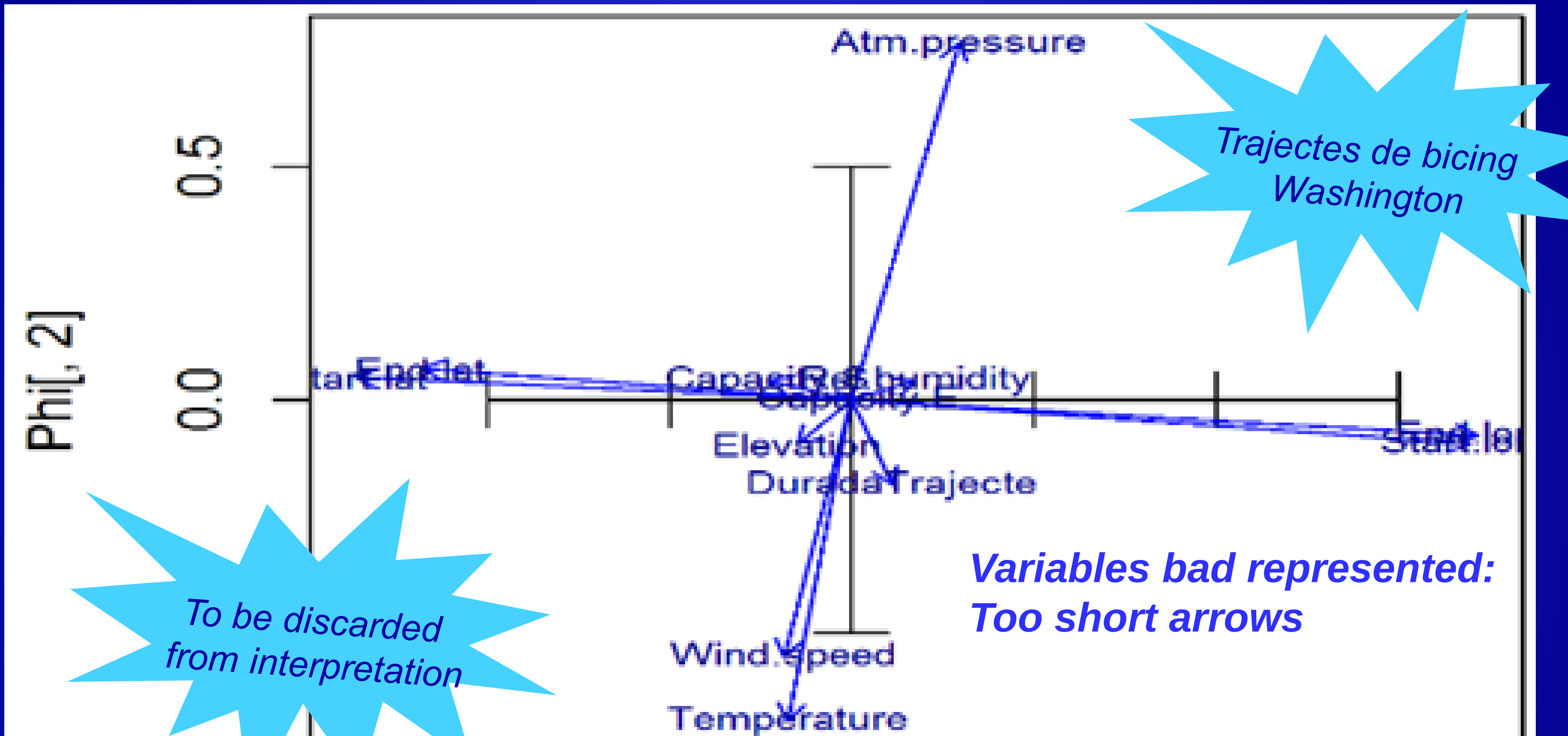
Principal components analysis

Process to interpret a factorial map

- Forget about variables bad represented in the factorial plan
- Which are the variables with relevant direct contribution to Factor in Axis X (eg. PCA1)?
- Which are the variables with relevant inverse contribution to Factor in Axis X (eg. PCA1)
- (later introduce info on qualitative variables as well)
- Analyze profiles opposed in two extremes of Axis X
- Induce a label for the Factor that represents the concept

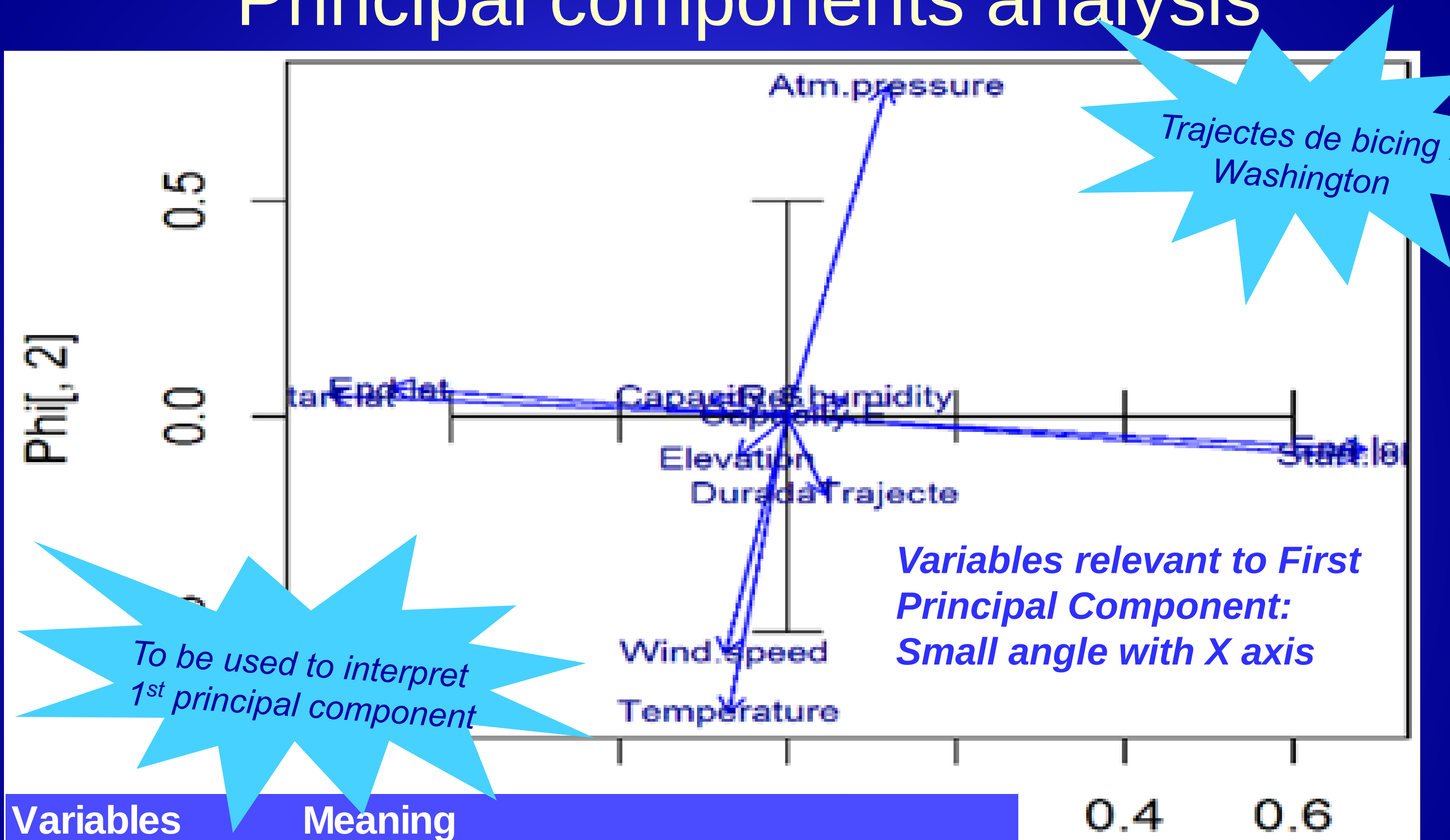
- Repeat with Factor in Axis Y

Principal components analysis



Variables	**Meaning**
Durada.Trajecte	Transit's total duration
Capacity.S	Bike capacity of the origin station
Capacity.E	Bike capacity of the destination station
Elevation	Difference in altitude between the stations of arrival and origin
Rel.humidity	Air relative humidity

Principal components analysis



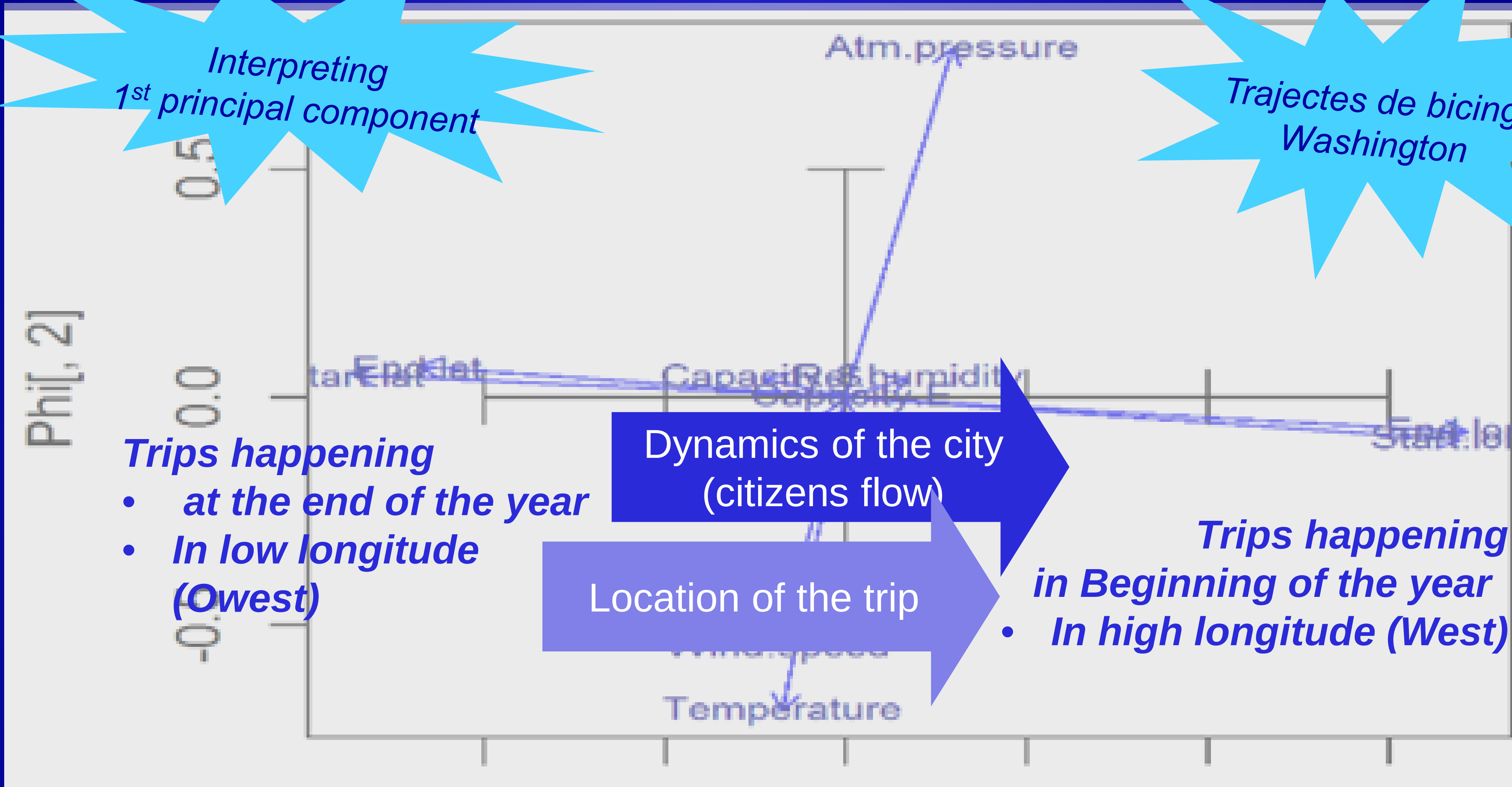
Variables relevant to First Principal Component: Small angle with X axis

Variables	Meaning	
Start.date	Date of the beginning of the trip	inverse
End.date	Date of the arrival	inverse
Start.long	Starting station's longitude	direct
End.long	Ending station's longitude	direct

Principal components analysis

Interpreting
1^{st} principal component

Trajectes de bicing
Washington



Trips happening
- **_at the end of the year_**
- **_In low longitude (Owest)_**

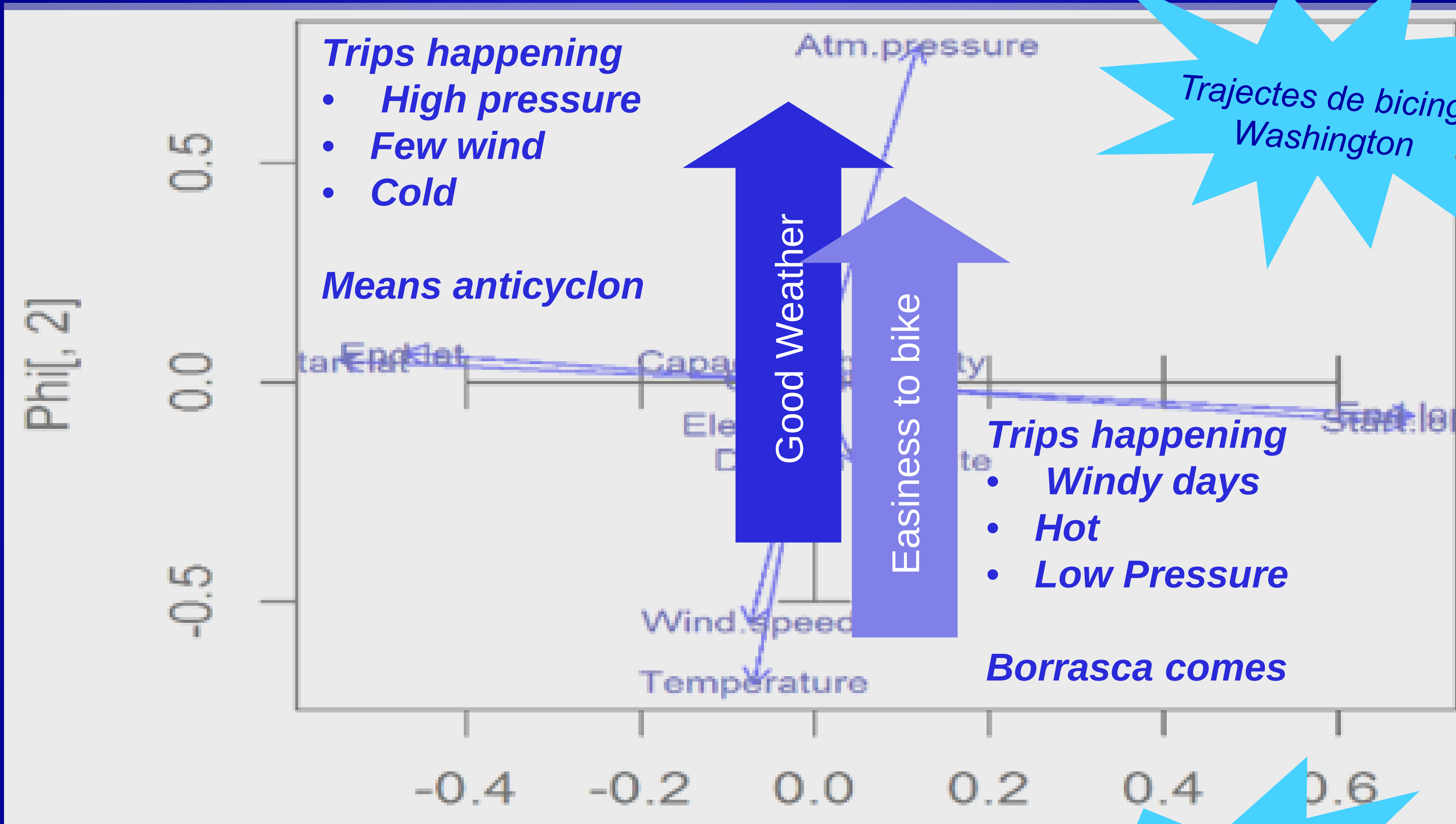
Dynamics of the city (citizens flow)

Location of the trip

Trips happening
in Beginning of the year
- **_In high longitude (West)_**

Variables	**Meaning**	
Start.date	Date of the beginning of the trip	inverse
End.date	Date of the arrival	inverse
Start.long	Starting station's longitude	direct
End.long	Ending station's longitude	direct

Principal components analysis



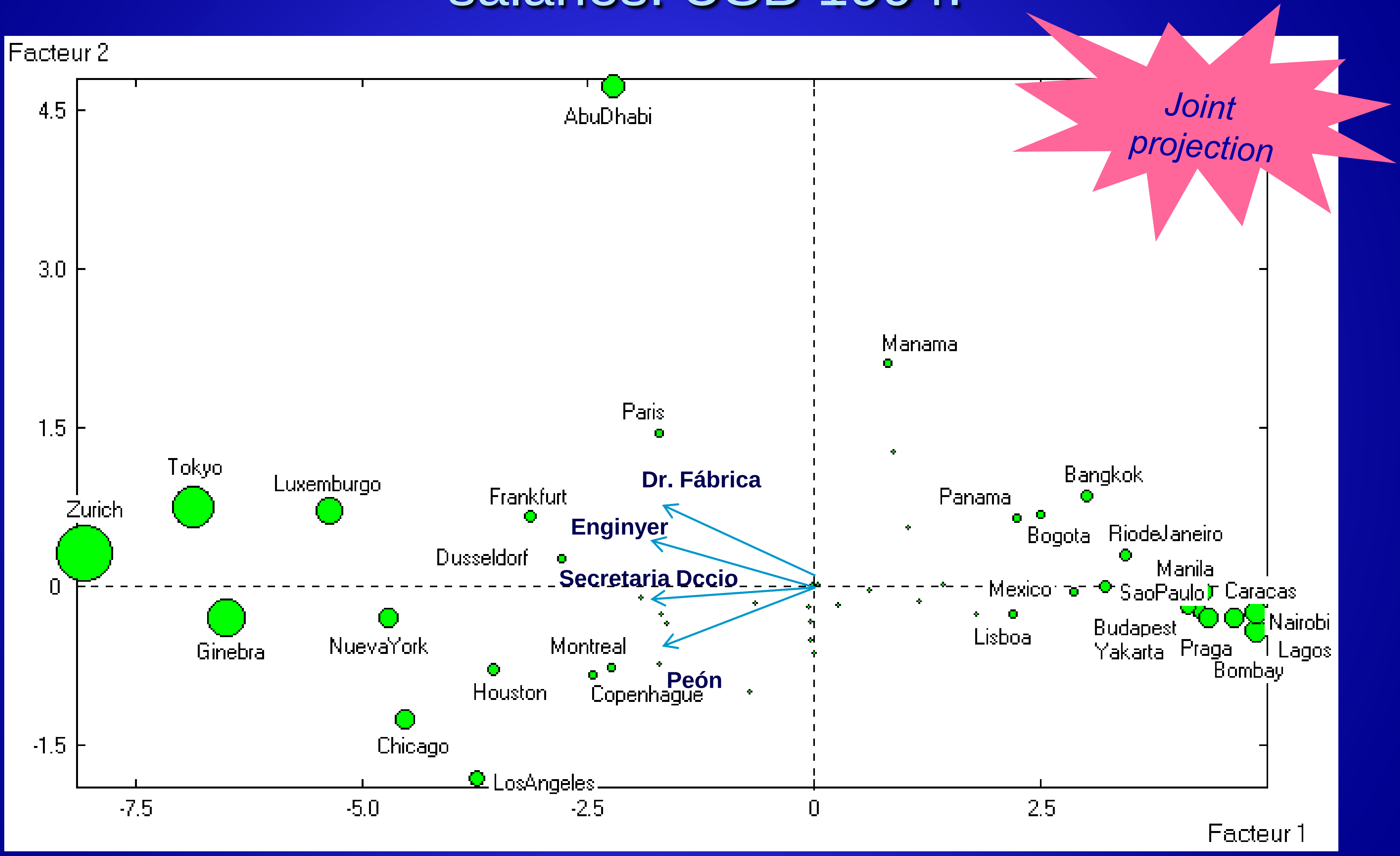
Trajectes de bicing Washington

Variables	Meaning	
Temperature	Air temperature	invers
Rel.humidity	Air relative humidity	invers
Atm.pressure	Atmospheric pressure	direct

Interpreting 2nd principal component

Visualisation of international cities according their salaries. USB 1994.

Visualisation of international cities according their salaries. USB 1994.



Factorial Methods

- Principal Components Analysis
  - Output: K factors rotating original X variables
  - Factors: Linear combinations of original variables

Several uses:

- As an associative data mining method to analyze relationships among variables

  Project variables and modalities and find associations

- As a preprocessing method for elicitation of latent variables

  Project active and illustrative variables/individuals on first/second factorial plane and interpret factors (find latent variables)

- As a preprocessing method for multidimensionality reduction

  Select more informative factors κ<<p *(accumulate 80% inertia)*
  Reduce data matrix to selected factors
  Alternative, keep variables mainly contributing to selected factors (smaller angles with factorial axis)

Factorial Methods

- Given <X,M,D>

Diagonalize Correlations matrix X'DX

Get r eigen values λ_α and sort decreasingly

$\{\lambda_\alpha\}_{\alpha=1:r} \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r$

Corresponding eigenvectors $u_\alpha = (u_{\alpha 1} \ldots u_{\alpha p})$

for $M = \mathbb{I}_p$: $u^*_\alpha = u_\alpha$; *for* $M \neq \mathbb{I}_p$: $u^*_\alpha = M^{-1/2} u_\alpha$

$\{u^*_\alpha\}_{\alpha=1:r}$ *orthonormal base for individuals*

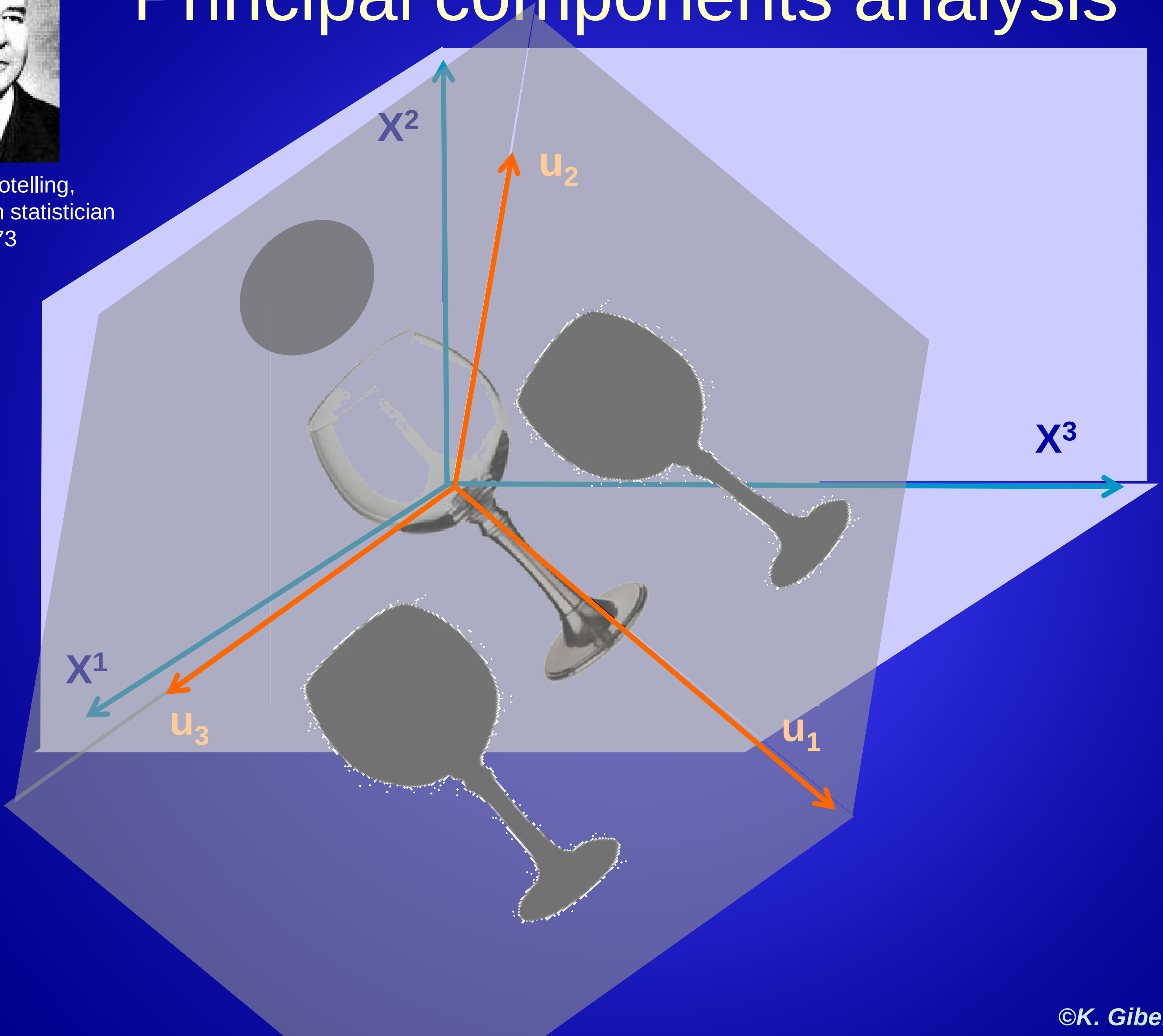
u^*_α are the principal factors of X : good rotation directions

$U^* = ([u^*_1]\ [u^*_2] \ldots [u^*_r])$ is the basis for the projection space

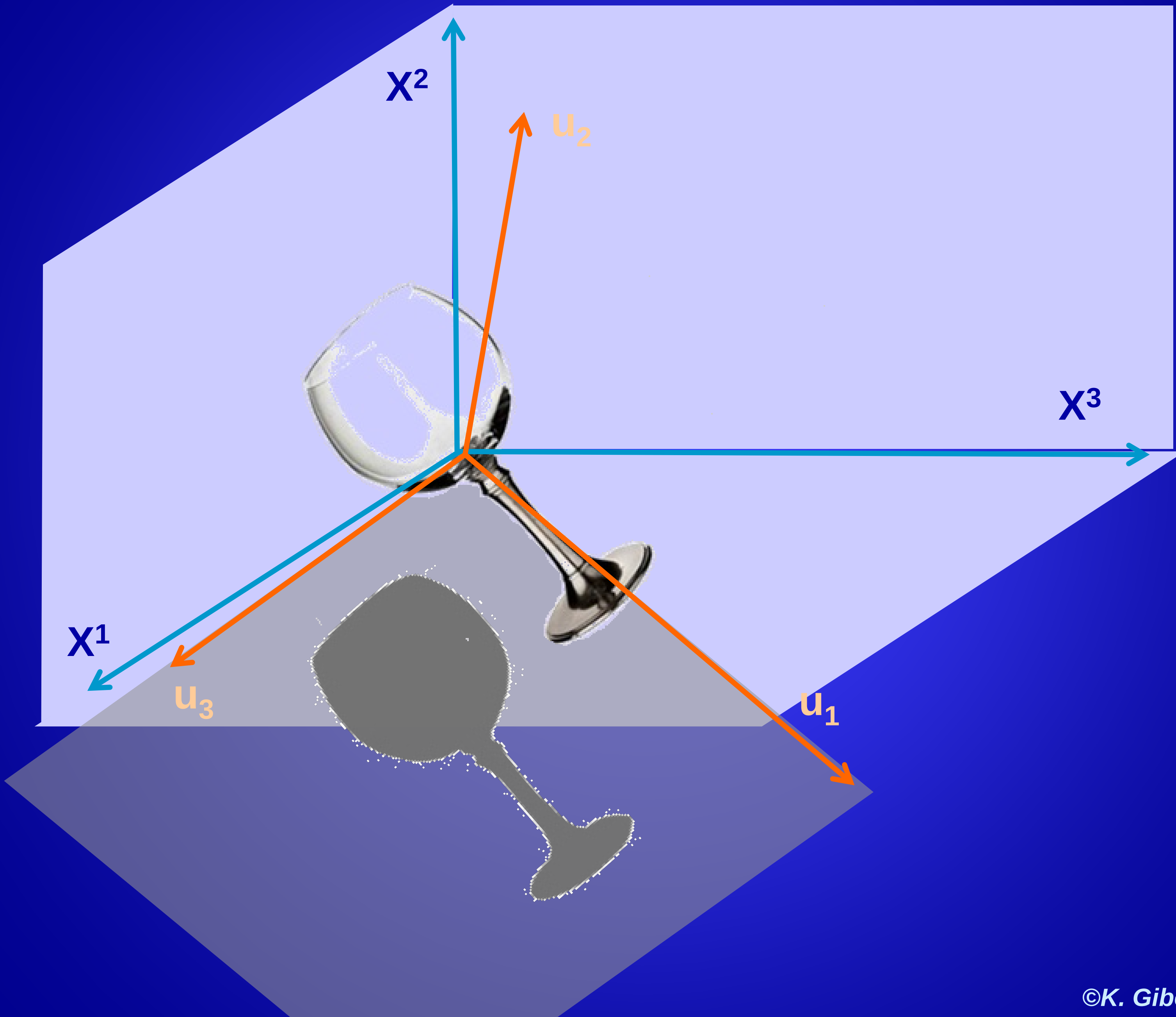
$U^*_k = ([u^*_1]\ [u^*_2] \ldots [u^*k])$ is the basis for projecting in first k dimensions(k<r)

Principal components analysis

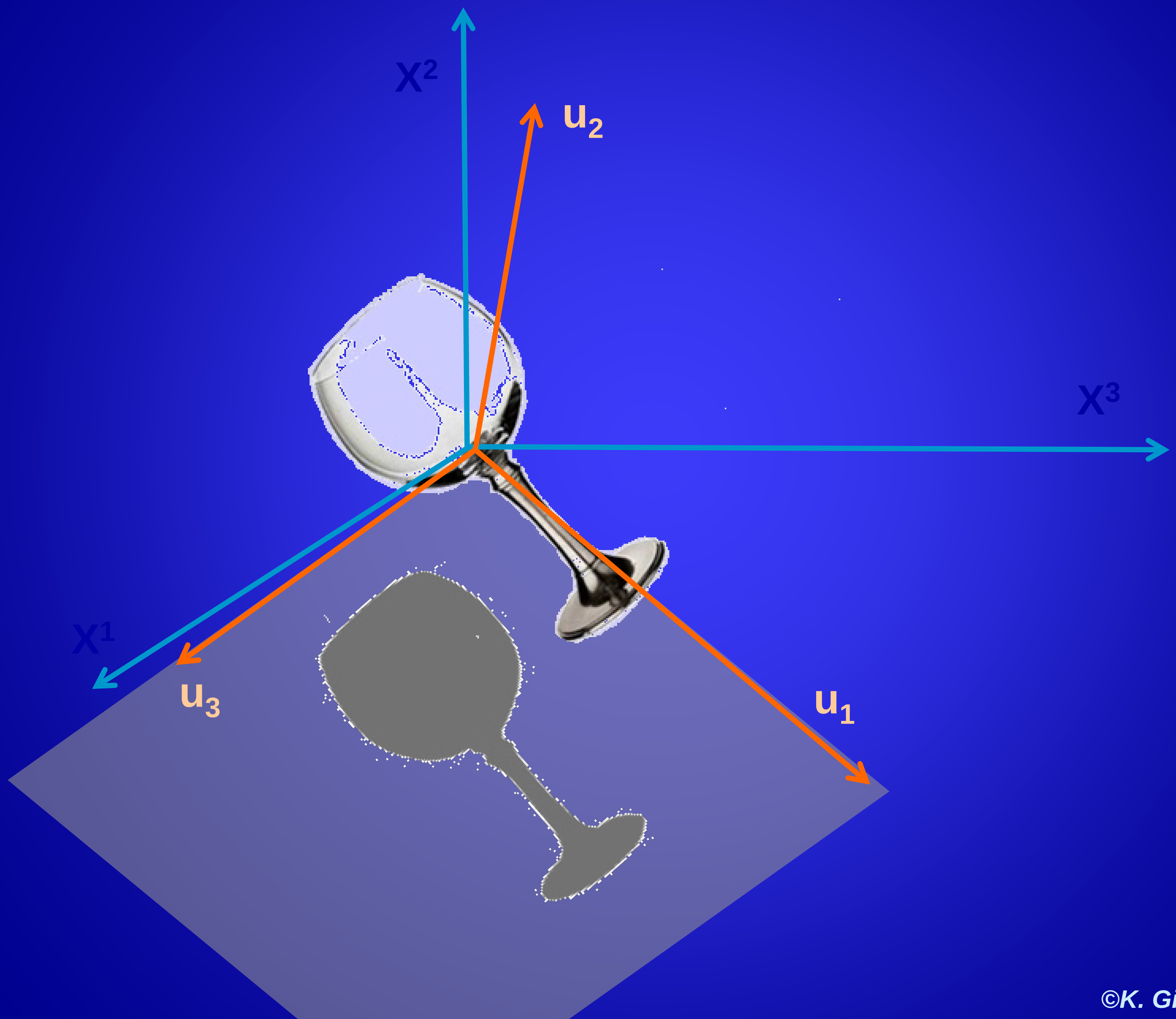
Harold Hotelling,
American statistician
1895-1973



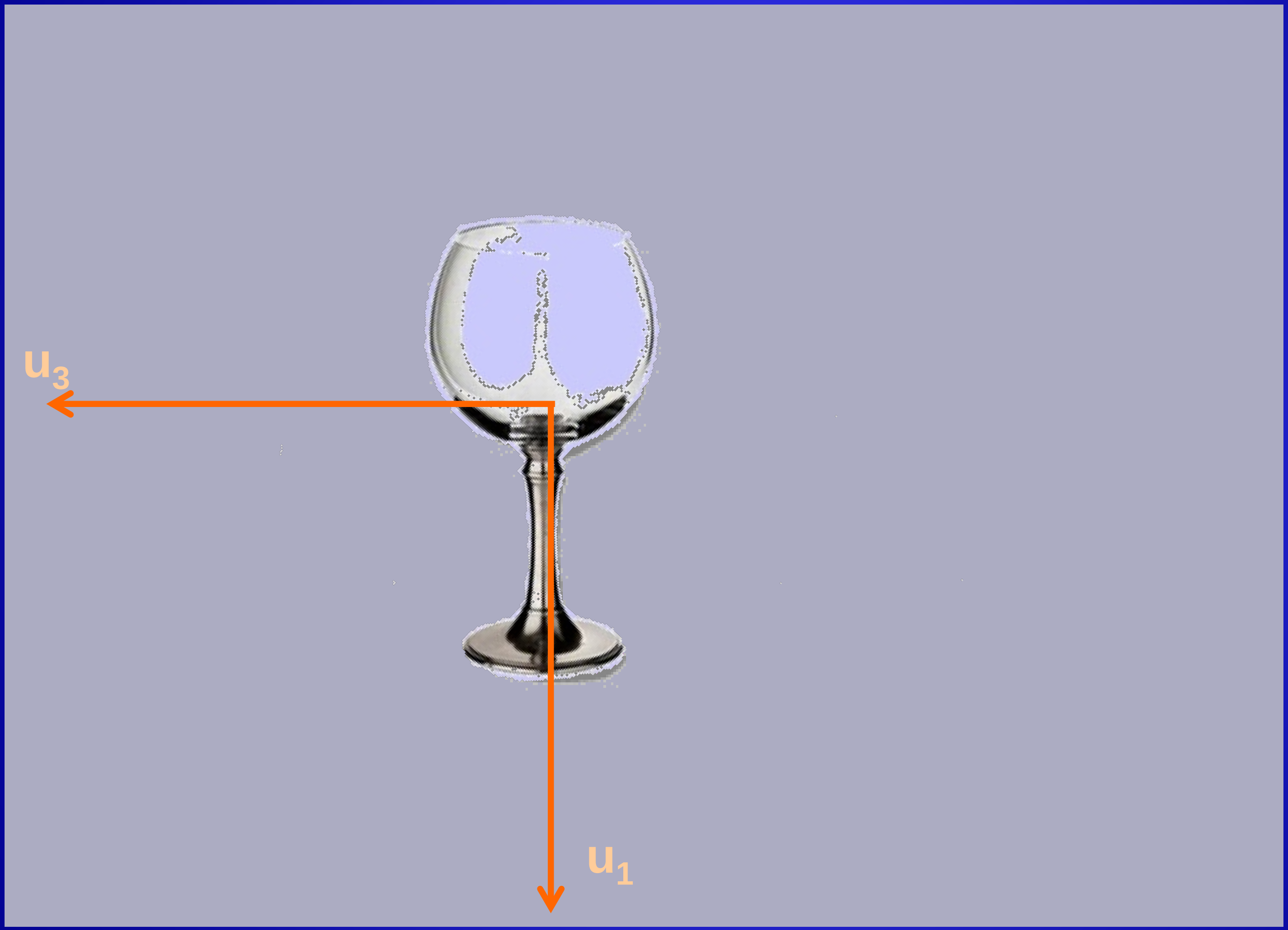
Principal components analysis



Principal components analysis



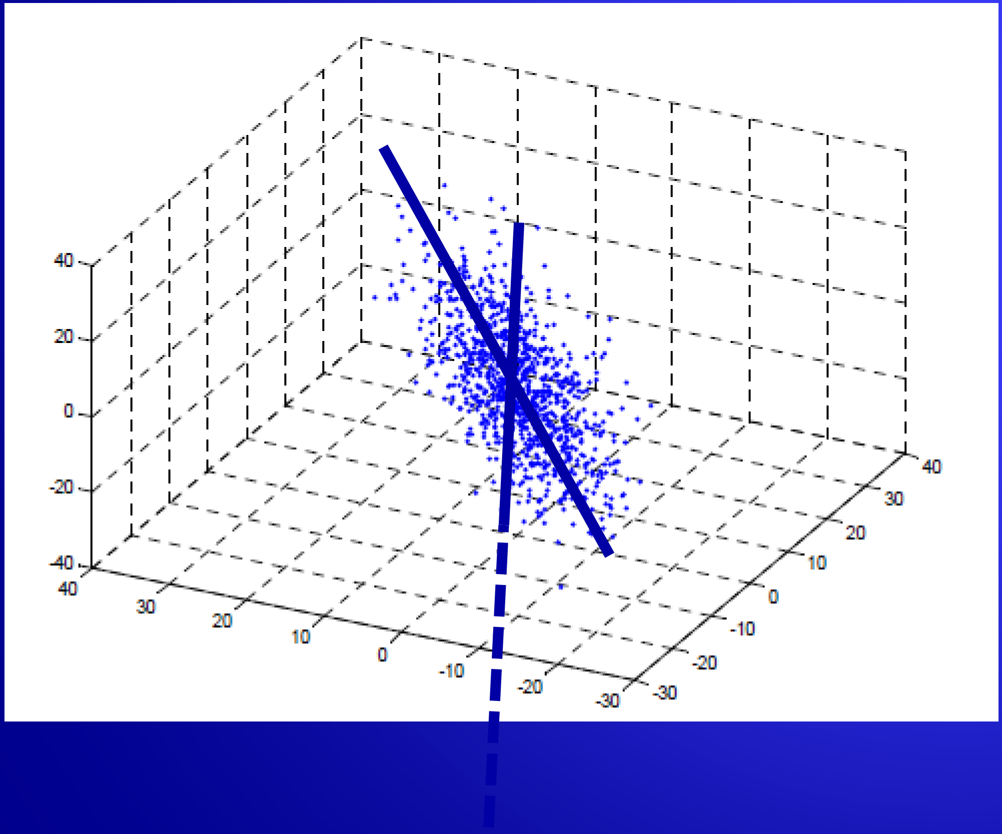
Principal components analysis

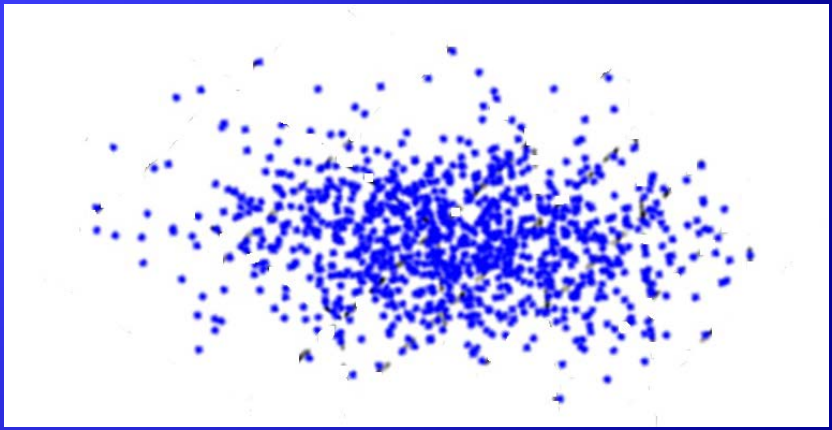


Principal components analysis

- Find the most informative projection planes of data cloud

  (factorial planes)

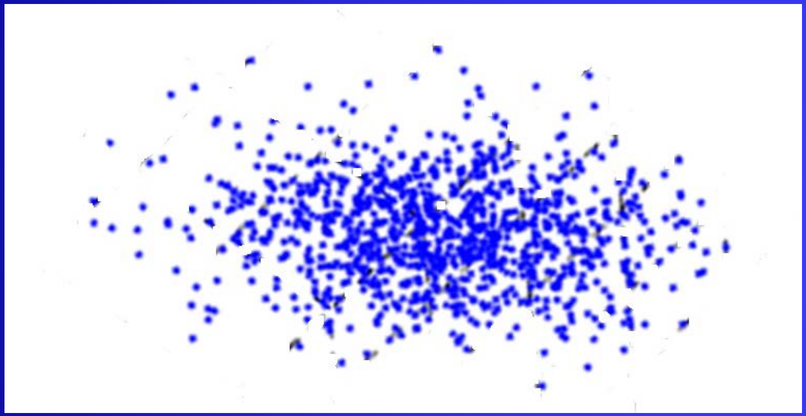


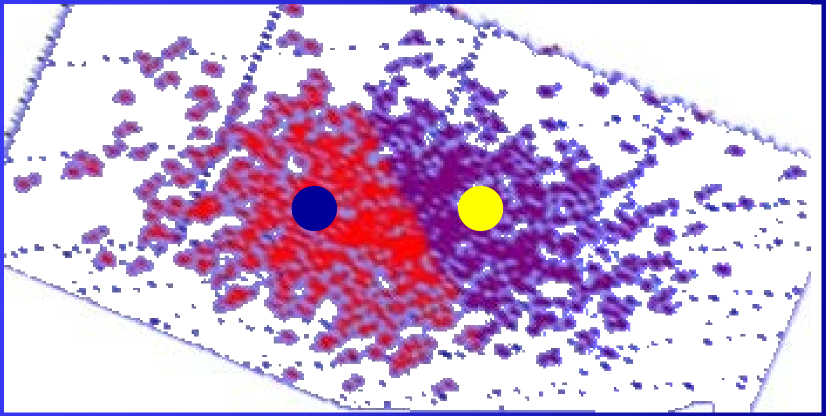


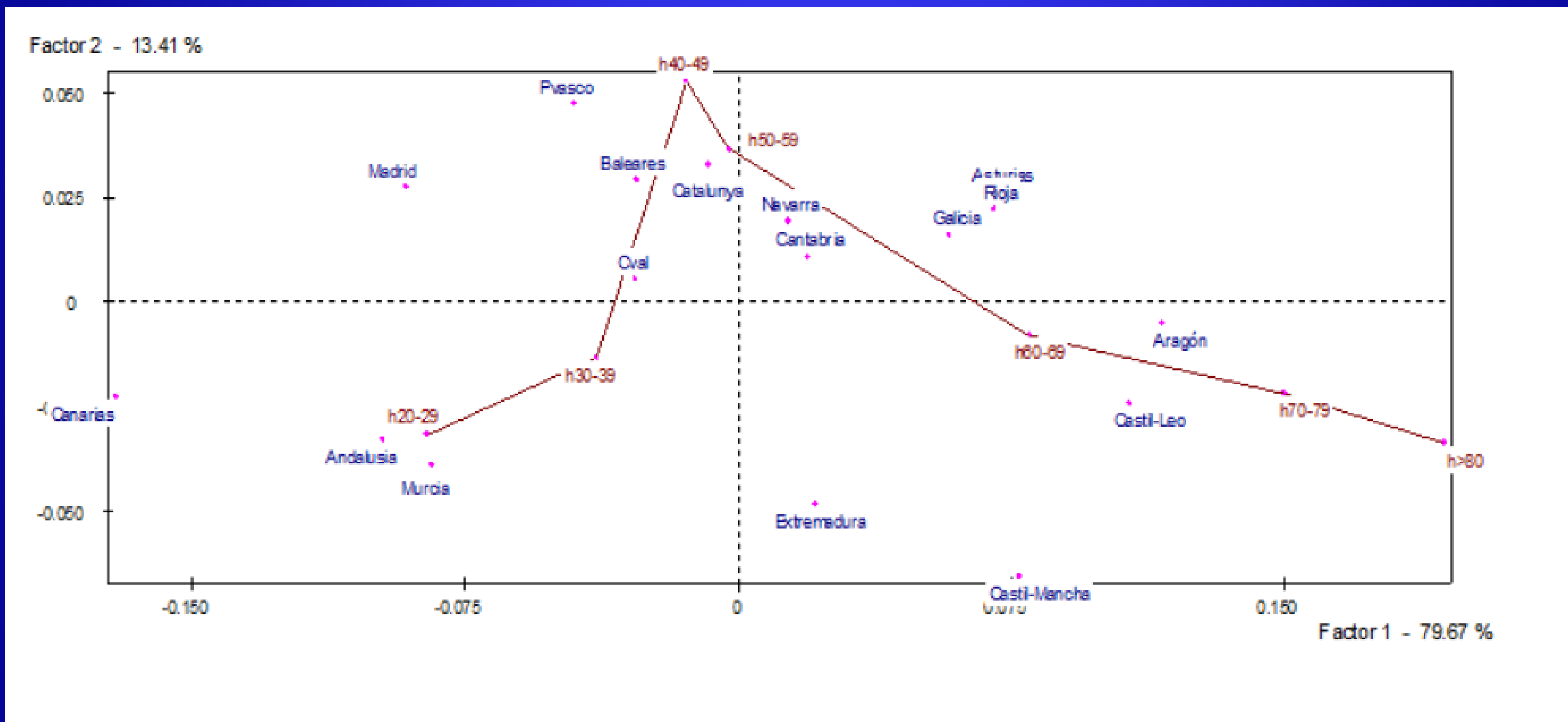
©K. Gibert

Principal components analysis

- Introduce qualitative information (projecting modalities)





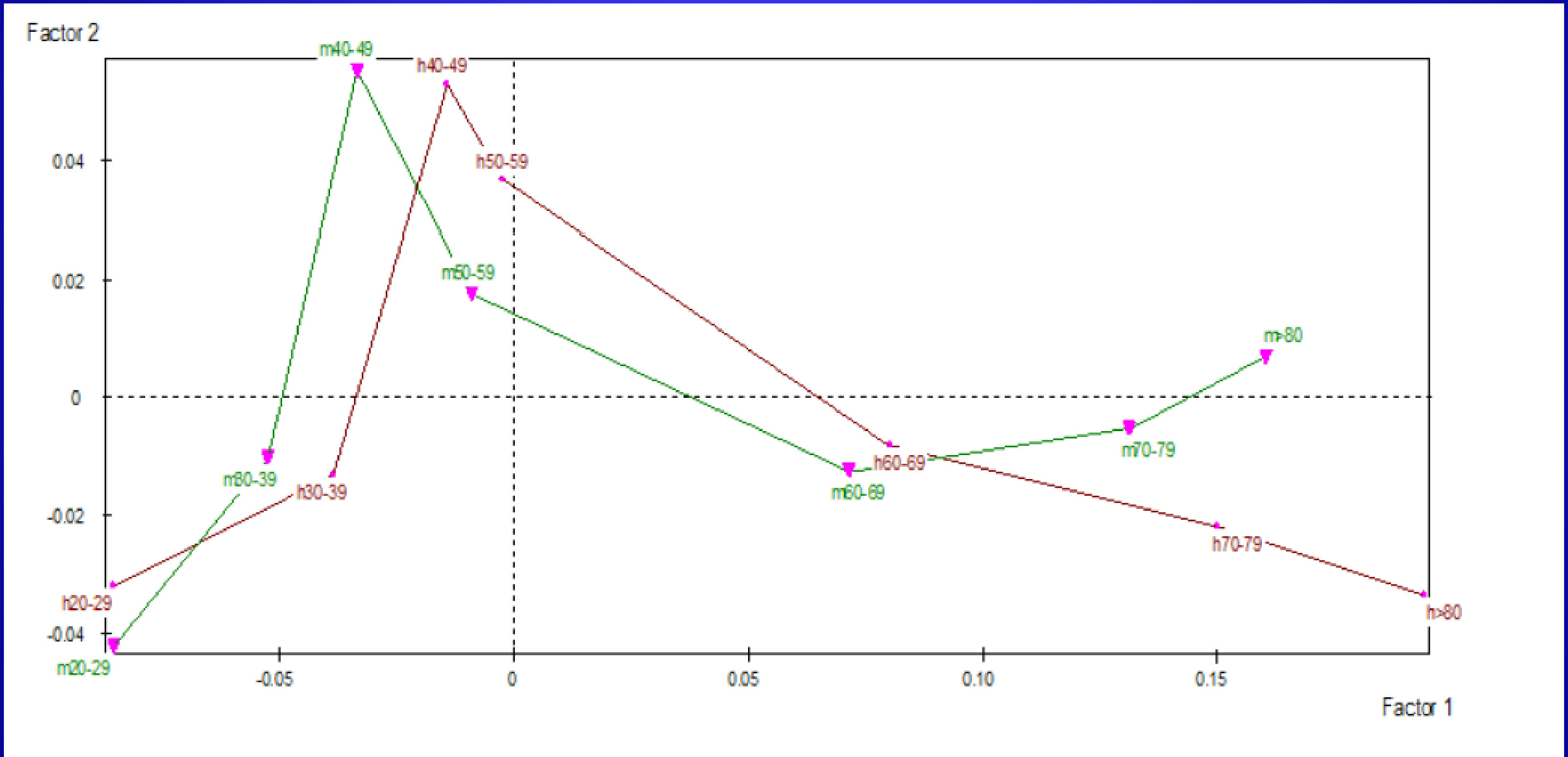


Factor 2 - 13.41 %
0.050
0.025
0
-0.050
-0.150
-0.075
0
0.150
Factor 1 - 79.67 %
h40-49
h50-59
h60-69
h70-79
h>80
h30-39
h20-29
Pvasco
Madrid
Balears
Catalunya
Navarra
Cantabria
Asturias
Rioja
Galicia
Cval
Aragón
Castil-Leo
Canarias
Andalusia
Murcia
Extremadura
Castil-Mancha

Projecting qualitative variables

Respect the following principles:

- Choose a diferent color for each qualitative variable
- Use the color of the variable for all centroids corresponding to the modalities of the variable
- Include a legent with the list of variables and associated color
- Ensure that legend to not hide any centroide in the factorial map
- For ordinal variables link modalities with rows in the right order and use the color of the variable for the arrows
- Manipulate the size of the font to guarantee the màximum visibility of the map



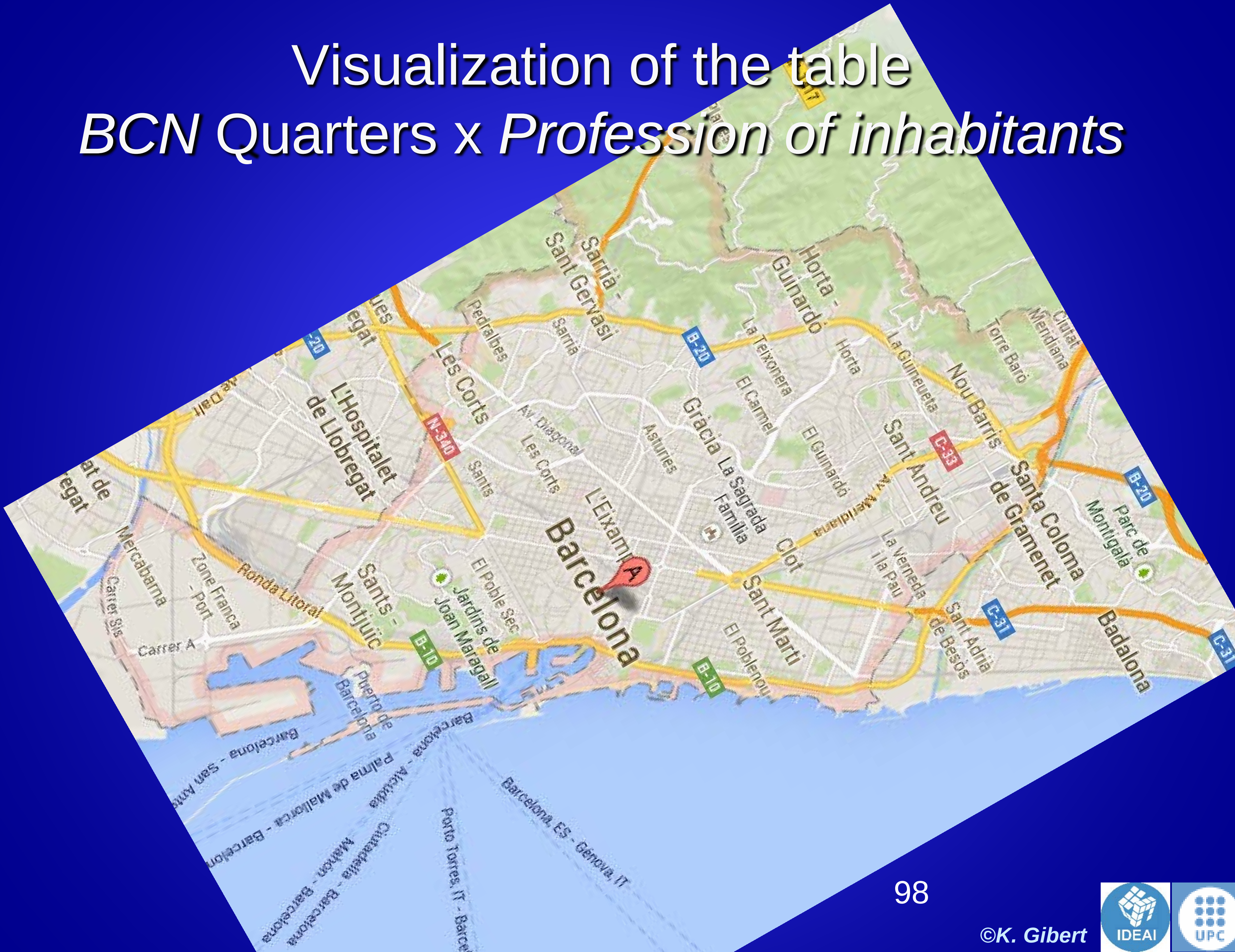
Factor 2
0.04
0.02
0
-0.02
-0.04
-0.05
0
0.05
0.10
0.15
Factor 1
m40-49
h40-49
h50-59
m50-59
m>80
m70-79
h60-69
m60-69
m30-39
h30-39
h70-79
h20-29
h>80
m20-29

©K. Gibert

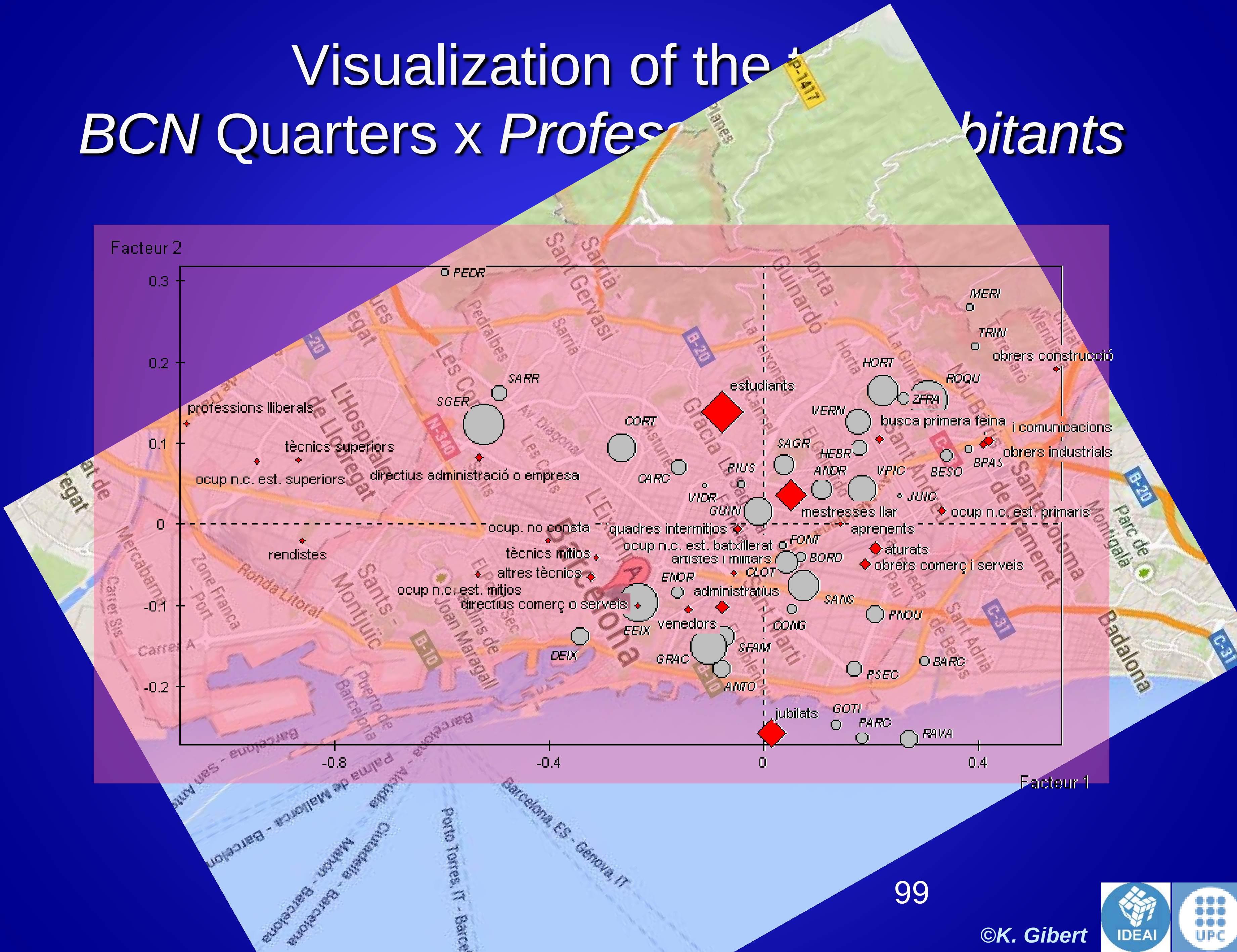
Efecte guttmann

http://www.ugr.es/~gallardo/

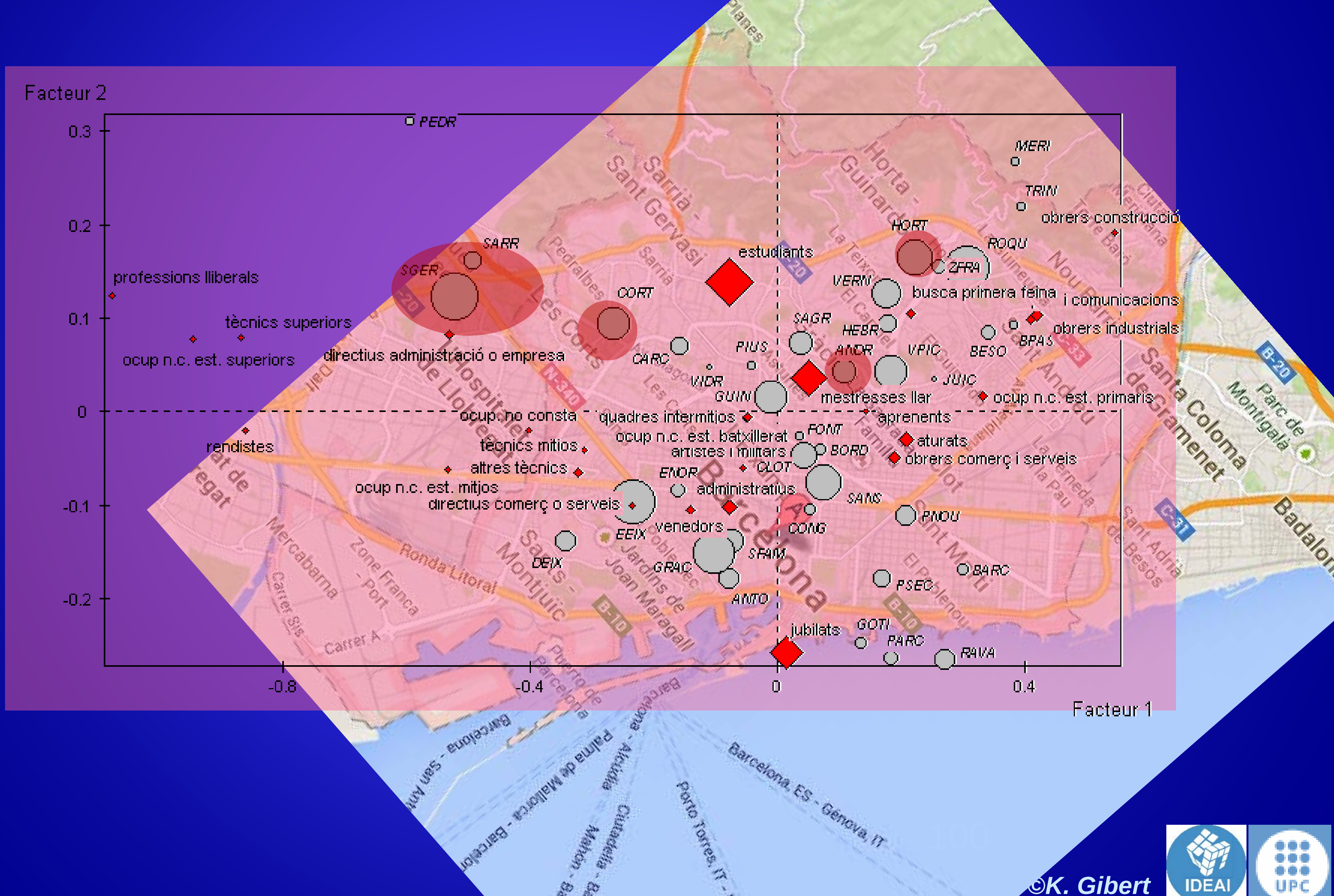
Visualization of the table *BCN* Quarters x *Profession of inhabitants*



Visualization of the t
BCN Quarters x *Profes* *bitants*



Visualization of the table
BCN Quarters x *Professions* of inhabitants



©K. Gibert