Descriptive and Exploratory Methods

Léon Bottou

largely copied from Mireille Summa-Gettler lectures (in french)

COS 424 – 3/23/2010

Agenda

Goals	Classification, clustering, regression, other.
Representation	Parametric vs. kernels vs. nonparametric Probabilistic vs. nonprobabilistic Linear vs. nonlinear Deep vs. shallow
Capacity Control	Explicit: architecture, feature selection Explicit: regularization, priors Implicit: approximate optimization Implicit: bayesian averaging, ensembles
Operational Considerations	Loss functions Budget constraints Online vs. offline
Computational Considerations	Exact algorithms for small datasets. Stochastic algorithms for big datasets. Parallel algorithms.

Today's topic fits poorly in this picture.

Introduction

Predictive methods

– Construct models using examples (the training set).

– Hope that it works well for future situations (e.g. on a testing set.)

Descriptive methods

– Describe the distribution of examples.

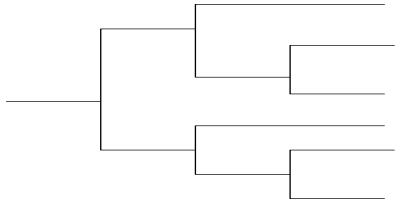
– Investigate the geometry of the data.

– Hope to acquire insights about the underlying phenomenon.

A catalog of descriptive methods

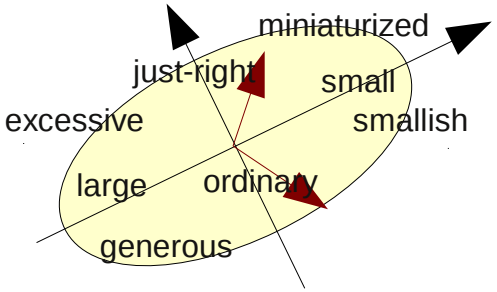
Clustering methods

– K-means, K-medioids, Gaussian mixtures. . .
– Hierarchical clustering. . .



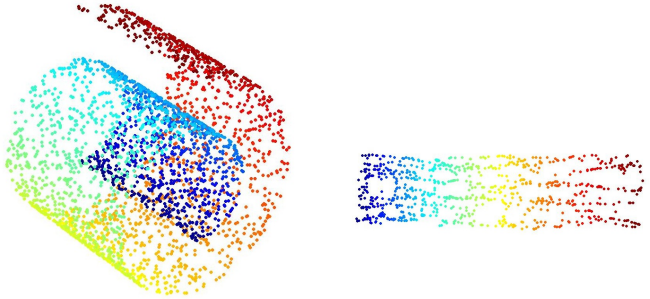
Projection methods

– Principal component analysis (PCA) [Hotelling, 30s]
– Correspondence analysis (CA) [Benzecri, 60s]
– Multiple correspondence analysis (MCA)
– Canonical correlation analysis (CCA), . . .



Embedding methods

– Kernel PCA
– Locally linear embbedding (LLE)
– ISOMAP



I. Principal Component Analysis

Sparkling water springs

Observations

– 21 sparkling water springs in France.

Continuous variables

– 8 ion concentrations (calcium, magnesium, . . .)
– price per liter.

Categorical variables

– Total minerality (low, medium, high)
– Compliance with regulations (yes, no)
– Region (Alps, Auvergne, Languedoc, . . .)

Sparkling water springs

Mineral water	Calcium	Magnésium	Potassium	Bicarbonates	Sulfates	Fluorures	Sodium	Nitrates	Price	Minerality	Compliance	Region
Arcens	14.5	24	10.7	1213	11	1.3	439	0.14	0.34	medium	no	Rhône-Alpes
Arvie	170	92	130	2195	31	0.9	650	0	0.44	high	no	Auvergne
Badoit	190	85	10	1300	40	1	150	5.8	0.64	medium	no	Rhône-Alpes
Beckerich	83.4	32	7.6	353	124	0.6	34	1	0.17	low	yes	Other
Châteauneuf	152	36	40	1799	195	3	651	0.05	0.58	medium	no	Auvergne
Eau de Perrier	149	7	1.4	420	42	0.05	11.5	5.2	0.72	low	yes	Languedoc-Roussillon
Faustine	170	50	26	1200	8	2	230	0.05	0.2	medium	no	Rhône-Alpes
La Salvetat	253	11	3	820	25	0.25	7	0.05	0.38	medium	yes	Other
Perrier	149	7	1.4	420	42	0.05	11.5	5.6	0.94	low	yes	Languedoc-Roussillon
Puits St-Georges	46	34	18.5	1373	10	0.5	434	8	0.35	medium	no	Rhône-Alpes
Pyrénées	48	12	1	183	18	0.05	31	5	0.3	low	yes	Other
Quézac	241	95	49.7	1685	143	2.1	255	0.05	0.52	high	no	Languedoc-Roussillon
San Pellegrino	185	53	2.5	237.9	444	0.6	35	2	0.65	medium	no	Other
St-Diéry	85	80	65	1350	25	0.3	385	1.9	0.32	high	no	Auvergne
St-Jean	76	25	36	908	52	1.1	228	1.4	0.4	medium	no	Rhône-Alpes
St-Pierre	35	20	36	1180	35	1.7	383	0.05	0.32	medium	no	Rhône-Alpes
St-Yorre	90	11	132	4368	174	9	1708	2.5	0.53	high	no	Auvergne
Vernet	29	17	22	470	7	1.3	120	0.05	0.36	low	no	Other
Vernière	190	72	49	1170	158	0.05	154	1.2	0.39	medium	no	Languedoc-Roussillon
Vichy-Célestins	103	10	66	2989	138	5	1172	1.5	0.59	high	no	Auvergne
Wattwiller	135	15.4	1.9	172	247	1.6	3	0	0.77	medium	no	Other

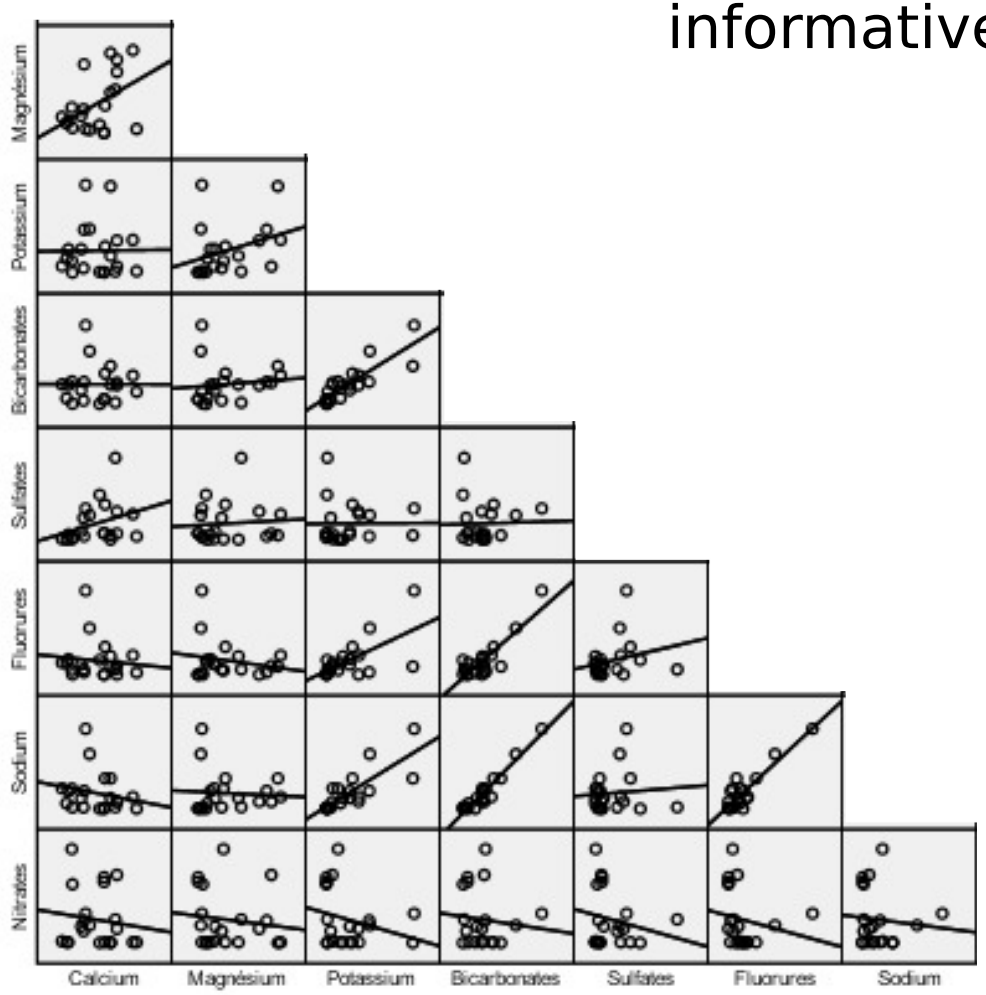
Mean	*123.5*	*38*	*34*	*1229*	*94*	*1.5*	*338*	*2.0*	*0.47*
Sdev	*68.0*	*29*	*38*	*990*	*105*	*2.0*	*417*	*2.4*	*0.19*
Minimum	*14.5*	*7*	*1*	*172*	*7*	*0.05*	*3*	*0*	*0.17*
Maximum	*253*	*95*	*132*	*4368*	*444*	*9*	*1708*	*8*	*0.94*

Active variables

Supplementary variables

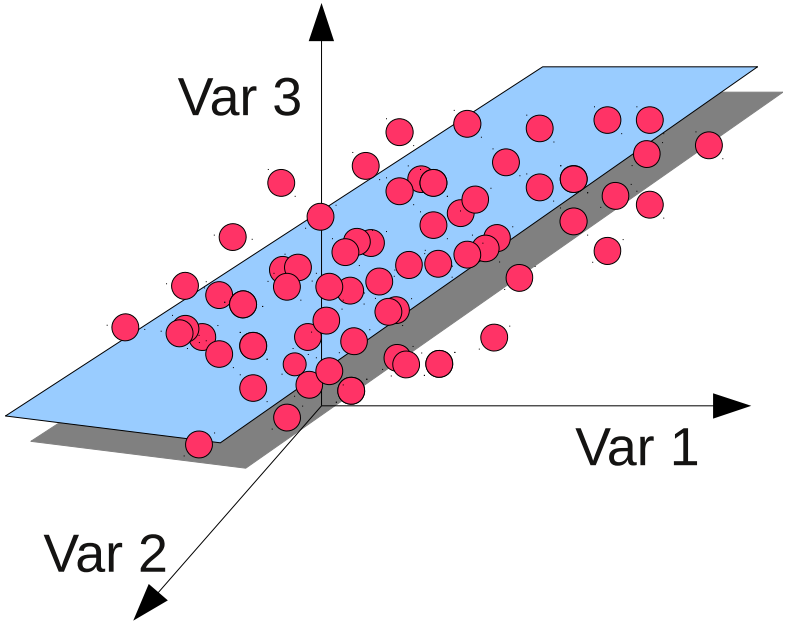
Elementary planes

Pairwise graphs are not informative

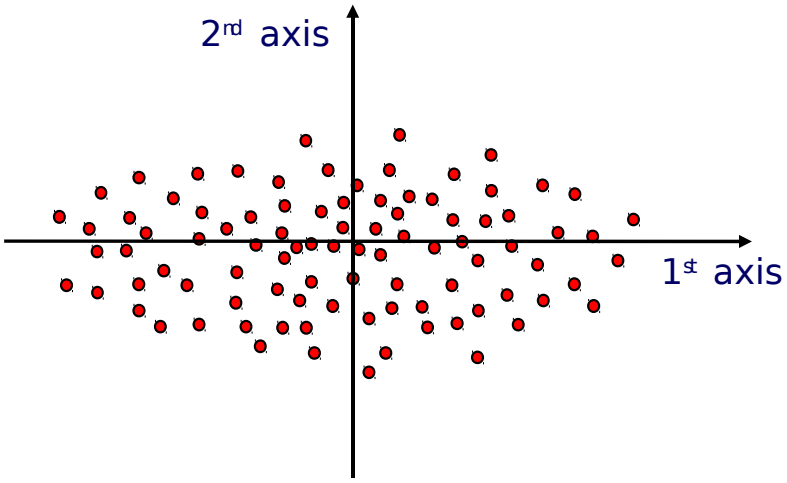


Approximate a data cloud by its projection

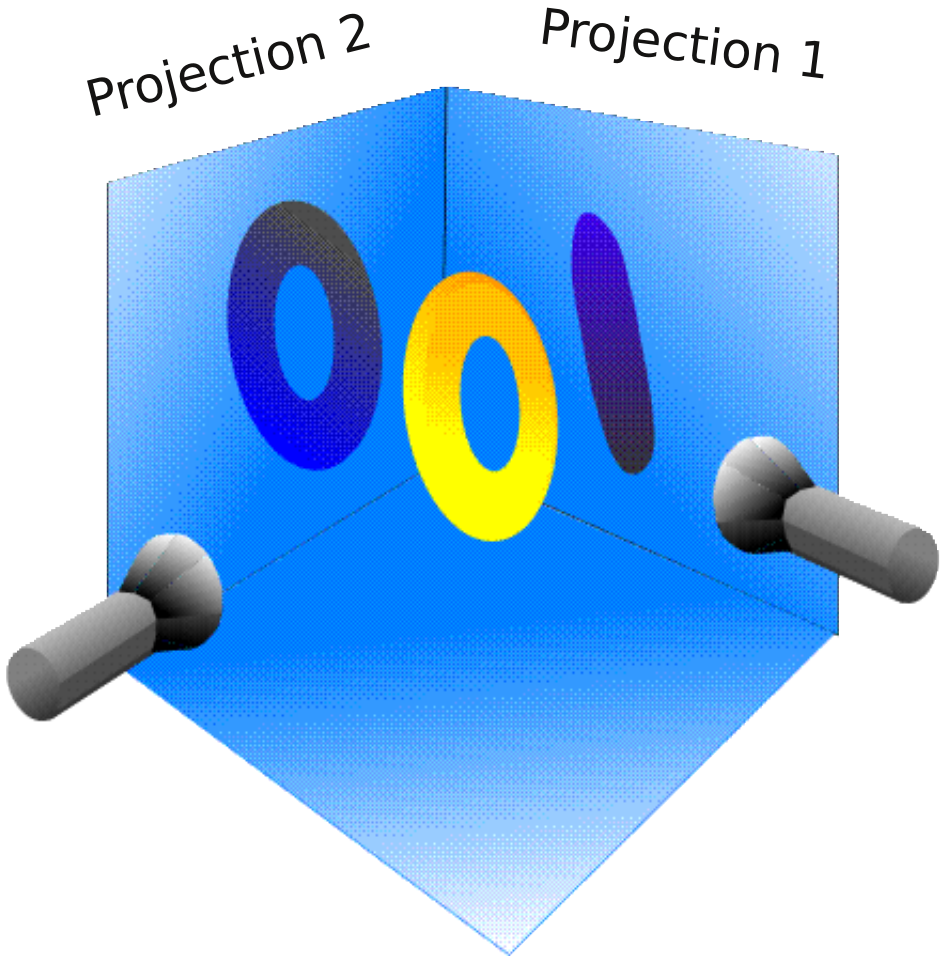
High dimensional cloud.



Low dimensional projection.

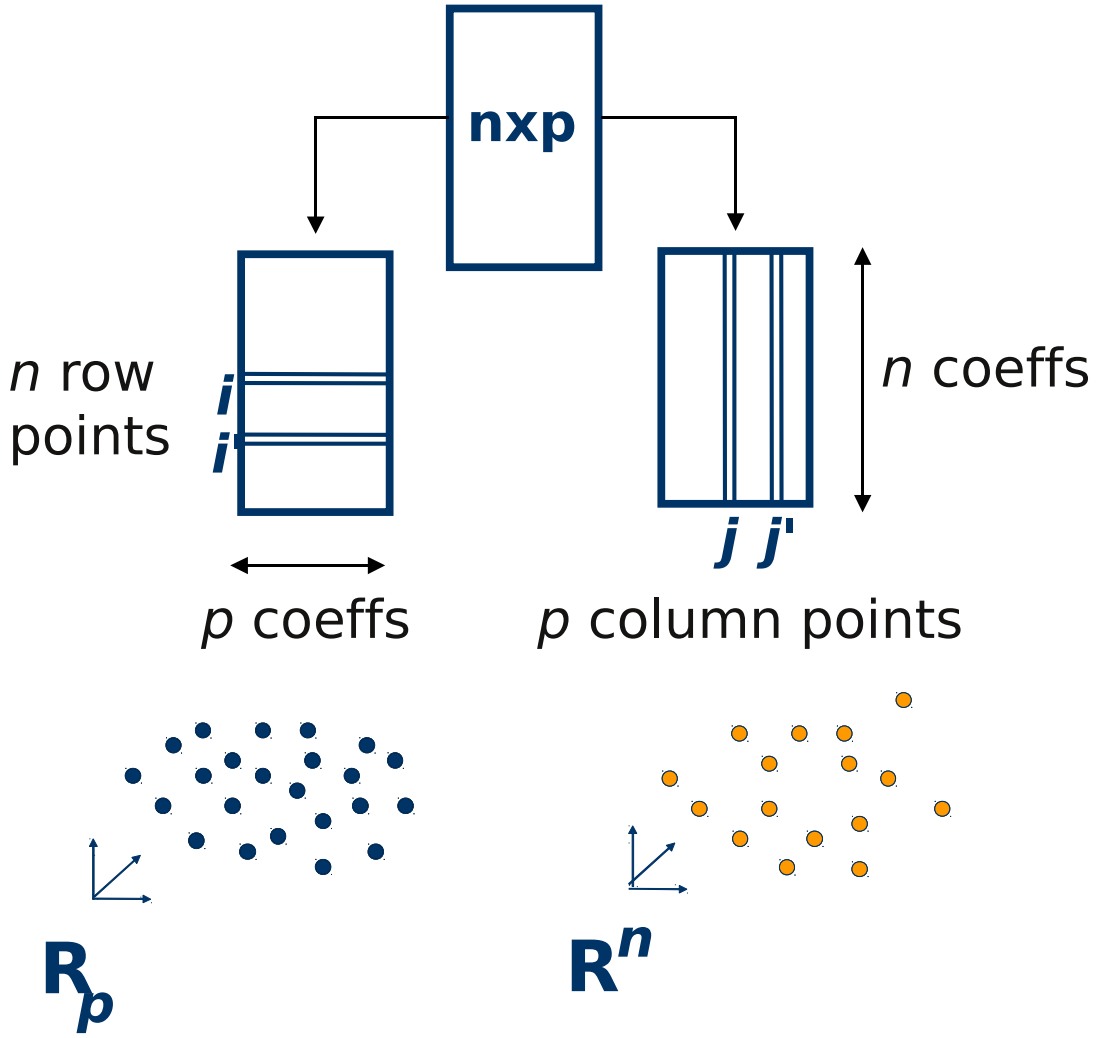


Some projections are more informative

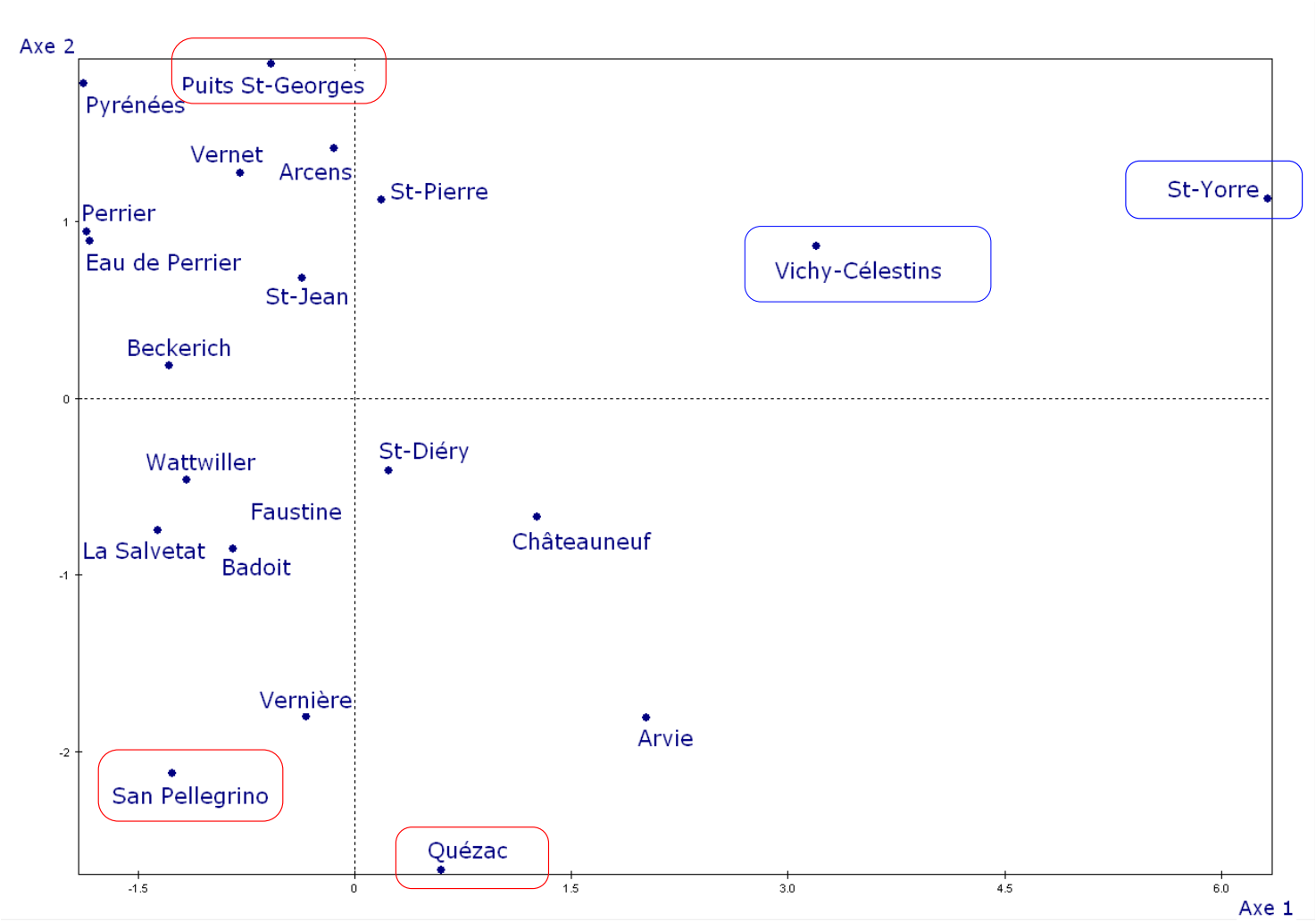


The main idea of PCA is the determination of a good projection.

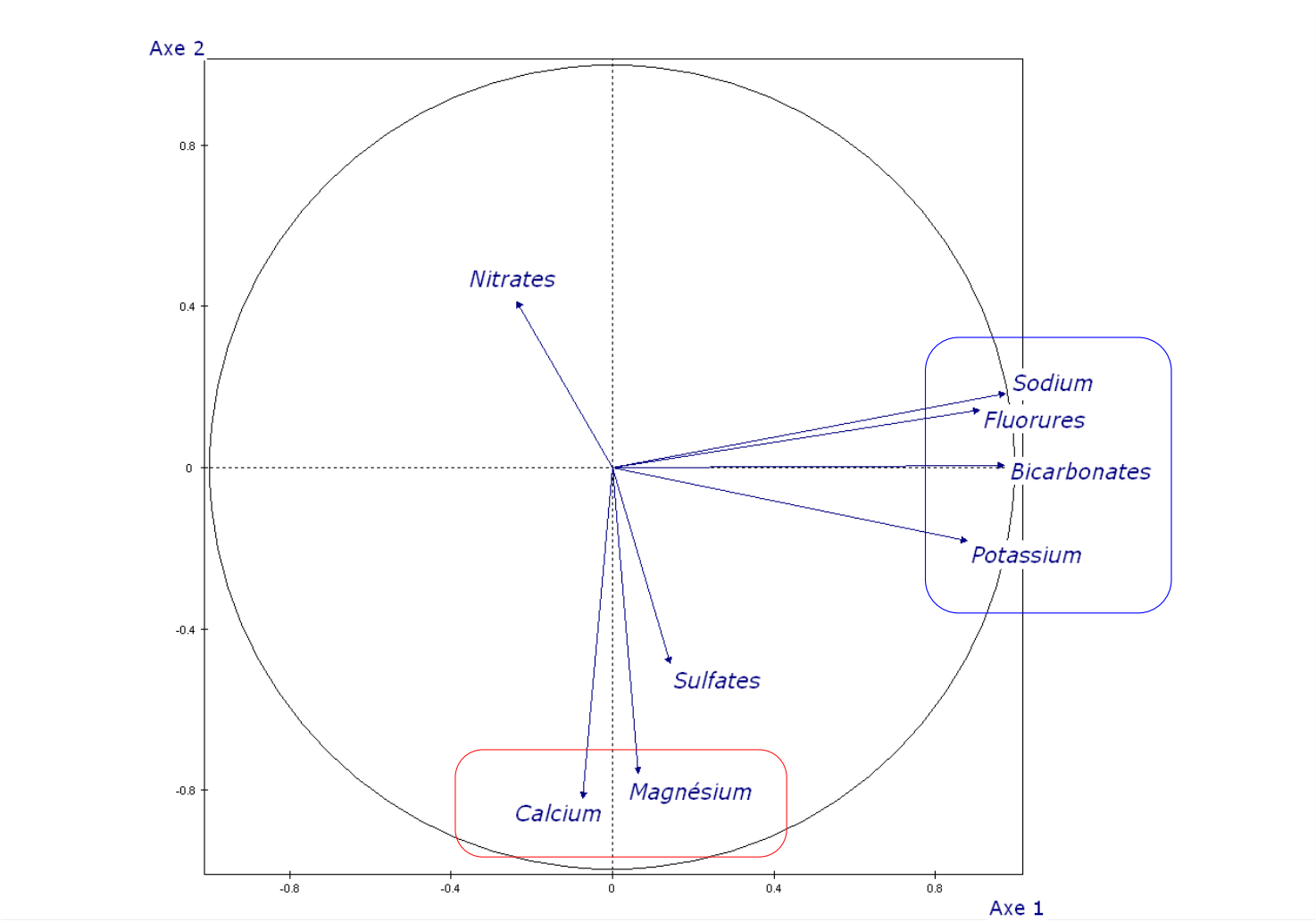
One data table, two data clouds



PCA projection of the 21 rows



PCA projection of the 8 columns



Summary

Principal component analysis

– Table of n observations represented by p continuous variables.
– Cloud of n row-points (observations) in dimension p.
– Cloud of p column-points (variables) in dimension n.
– Search the "best" projection for each cloud.

Interpretation

– Identify similar observations.
– Identify similar variables.

Best projection ?

Distance

Distances

– A good projection reveals whether two points were close or distant.
– We would like to use the convenient Euclidian distance.
– Variables often have very different numerical ranges.

	Calcium	Magnésium	Potassium	Bicarbonates	Sulfates	Fluorures	Sodium	Nitrates
Mean	*123.5*	*37.5*	*33.8*	*1228.9*	*93.8*	*1.5*	*337.7*	*2.0*
Sdev	*68.0*	*29.4*	*37.7*	*990.0*	*105.5*	*2.0*	*417.0*	*2.4*
Minimum	*14.5*	*7*	*1*	*172*	*7*	*0.05*	*3*	*0*
Maximum	*253*	*95*	*132*	*4368*	*444*	*9*	*1708*	*8*

Correlation PCA

– Normalize the mean and standard deviation of each variable, $x_{ij} = (z_{ij} - \bar{z}_j)/\sigma_j$.

– This is the default and this is what we discuss today.

Covariance PCA

– Normalize the mean of each variable, $x_{ij} = (z_{ij} - \bar{z}_j)$, but not the standard deviation.

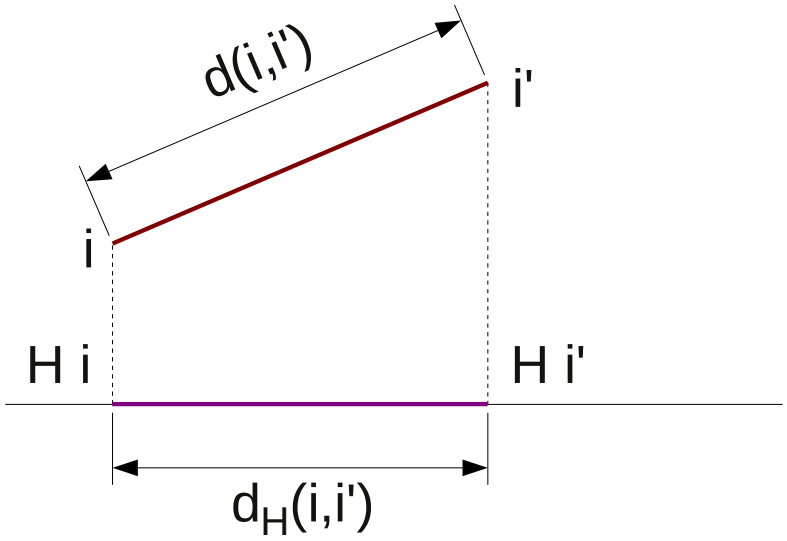
– This is sometimes useful.

Normed centered data

	Calcium	Magnesium	Potassium	Bicarbonates	Sulfates	Fluorures	Sodium	Nitrates	*Distance from origin*
Arcens	-1.60	-0.46	-0.61	-0.02	-0.78	-0.12	0.24	-0.77	*2.11*
Arvie	0.68	1.85	2.55	0.98	-0.60	-0.32	0.75	-0.83	*3.62*
Badoit	0.98	1.61	-0.63	0.07	-0.51	-0.27	-0.45	1.60	*2.66*
Beckerich	-0.59	-0.19	-0.70	-0.88	0.29	-0.47	-0.73	-0.41	*1.63*
Châteauneuf	0.42	-0.05	0.16	0.58	0.96	0.72	0.75	-0.81	*1.79*
Eau de Perrier	0.37	-1.04	-0.86	-0.82	-0.49	-0.74	-0.78	1.35	*2.42*
Faustine	0.68	0.42	-0.21	-0.03	-0.81	0.22	-0.26	-0.81	*1.46*
La Salvetat	1.91	-0.90	-0.82	-0.41	-0.65	-0.64	-0.79	-0.81	*2.72*
Perrier	0.37	-1.04	-0.86	-0.82	-0.49	-0.74	-0.78	1.52	*2.52*
Puits St-Georges	-1.14	-0.12	-0.41	0.15	-0.79	-0.52	0.23	2.53	*2.97*
Pyrénées	-1.11	-0.87	-0.87	-1.06	-0.72	-0.74	-0.74	1.27	*2.66*
Quézac	1.73	1.95	0.42	0.46	0.47	0.27	-0.20	-0.81	*2.86*
San Pellegrino	0.90	0.53	-0.83	-1.00	3.32	-0.47	-0.73	0.01	*3.82*
St-Diéry	-0.57	1.44	0.83	0.12	-0.65	-0.61	0.11	-0.03	*1.98*
St-Jean	-0.70	-0.43	0.06	-0.32	-0.40	-0.22	-0.26	-0.24	*1.05*
St-Pierre	-1.30	-0.60	0.06	-0.05	-0.56	0.08	0.11	-0.81	*1.74*
St-Yorre	-0.49	-0.90	2.61	3.17	0.76	3.68	3.29	0.22	*6.55*
Vernet	-1.39	-0.70	-0.31	-0.77	-0.82	-0.12	-0.52	-0.81	*2.17*
Vernière	0.98	1.17	0.40	-0.06	0.61	-0.74	-0.44	-0.33	*1.93*
Vichy-Célestins	-0.30	-0.94	0.86	1.78	0.42	1.70	2.00	-0.20	*3.46*
Wattwiller	0.17	-0.75	-0.85	-1.07	1.45	0.03	-0.80	-0.83	*2.43*
Mean	***0***	***0***	***0***	***0***	***0***	***0***	***0***	***0***	
Variance	***1***	***1***	***1***	***1***	***1***	***1***	***1***	***1***	

Preserve distances

Projection contracts the distances



PCA criterion: maximize sum of projected distances

$$\max_H \sum_{i=1}^{n} \sum_{i'=1}^{n} d_H^2(i,i')$$

Maximize dispersion

Observe $\sum_{i,i'} (\mathbf{x}_i - \mathbf{x}_{i'})^2 = \sum_{i,i'} \big((\mathbf{x}_i - \bar{\mathbf{x}}) - (\mathbf{x}_{i'} - \bar{\mathbf{x}})\big)^2 = \cdots = 2n \sum_i (\mathbf{x}_i - \bar{\mathbf{x}})^2 = 2n^2 \,\mathrm{Var}(\mathbf{x})$.

Equivalent PCA criterion: maximize dispersion

– Maximize average distance to the cloud mean G.

$$\max_H \frac{1}{n} \sum_{i=1}^{n} d_H^2(i, G)$$

Equivalent PCA criterion: maximize variance

– Maximize variance of the projected points.

First factorial axis

– Pick unit vector $\mathbf{u}$.
– Project the $\mathbf{x}_i$ on the line of direction $\mathbf{u}$.
– Find $\mathbf{u}$ that maximizes the dispersion.

$$\max_{\mathbf{u}} \; C(\mathbf{u}) = \sum_{i=1}^{n} \left[\mathbf{u}^\top \mathbf{x}_i - \mathbf{u}^\top \bar{\mathbf{x}}\right]^2 \quad \text{subject to} \quad \mathbf{u}^\top \mathbf{u} = 1$$

The constraint means that $\mathbf{u}$ lives on the unit sphere. At the optimum $\mathbf{u}^*$, the gradient of the dispersion must be orthogonal to the surface of the sphere, otherwise we would be able to find a better solution by slightly moving $\mathbf{u}^*$ along the projection of the gradient. Therefore there exist a "Lagrange multiplier" λ such that $\frac{dC(\mathbf{u}^*)}{d\mathbf{u}} - \lambda \mathbf{u}^* = 0$.

This leads to the necessary condition $\left[\frac{1}{n}\sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top\right] \mathbf{u}^* = \lambda \mathbf{u}^*$.

Therefore $\mathbf{u}^*$ must be an eigenvector of the covariance matrix.
The best one is associated with the largest eigenvalue.

Orientation is arbitrary!

Successive factorial axes

First factorial axis

– Direction with maximal dispersion

– Eigenvector of the covariance matrix Σ with the highest eigenvalue.

Second factorial axis

– Direction with maximal dispersion orthogonal to the first axis

– Eigenvector of Σ with the second highest eigenvalue.

Third factorial axis

– Direction with maximal dispersion orthogonal to the first two axes

– Eigenvector of Σ with the third highest eigenvalue.

etc...

Basis change

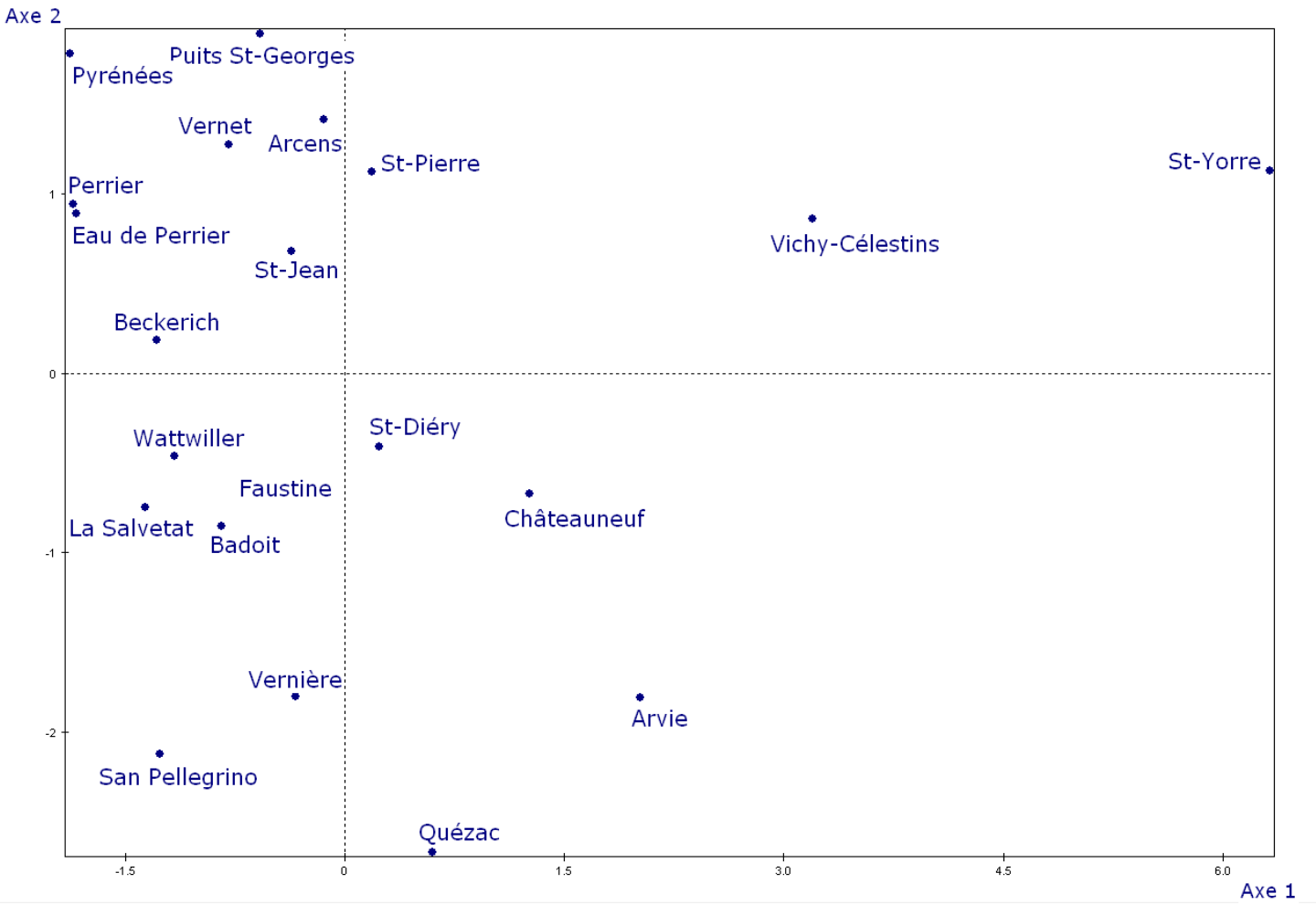
– The basis formed by the p variables is replaced
by the basis formed by the p principal axes.

Factorial coordinates of the rows

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Axis 7	Axis 8	Distance from origin
Arcens	-0.14	1.41	-0.04	1.34	0.38	0.55	-0.44	0.01	*2.11*
Arvie	2.02	-1.81	2.02	0.84	0.25	-0.93	-0.02	-0.07	*3.62*
Badoit	-0.84	-0.85	1.55	-1.56	-0.08	0.87	0.12	-0.01	*2.66*
Beckerich	-1.29	0.18	-0.60	0.58	0.49	0.10	0.09	0.02	*1.63*
Châteauneuf	1.26	-0.67	-1.01	0.09	-0.08	0.18	-0.27	-0.10	*1.79*
Eau de Perrier	-1.83	0.89	0.06	-1.00	-0.72	-0.40	0.06	-0.02	*2.42*
Faustine	-0.11	-0.60	0.40	0.73	-0.85	0.54	0.20	-0.09	*1.46*
La Salvetat	-1.36	-0.75	-0.40	0.40	-2.13	-0.22	-0.29	0.04	*2.72*
Perrier	-1.86	0.94	0.11	-1.14	-0.70	-0.42	0.09	-0.02	*2.52*
Puits St-Georges	-0.58	1.89	1.32	-1.63	0.67	0.06	-0.26	0.04	*2.97*
Pyrénées	-1.88	1.78	0.27	-0.37	0.30	-0.16	0.18	-0.05	*2.66*
Quézac	0.60	-2.68	0.50	0.00	-0.25	0.55	0.19	0.07	*2.86*
San Pellegrino	-1.27	-2.12	-2.29	-1.20	1.31	-0.17	-0.12	-0.03	*3.82*
St-Diéry	0.24	-0.41	1.60	0.54	0.91	-0.07	-0.06	-0.06	*1.98*
St-Jean	-0.37	0.68	0.03	0.65	0.15	-0.23	0.13	0.07	*1.05*
St-Pierre	0.19	1.12	-0.15	1.28	0.27	0.02	-0.03	0.07	*1.74*
St-Yorre	**6.32**	1.13	-0.73	-0.99	-0.20	-0.07	0.35	0.04	***6.55***
Vernet	-0.79	1.27	-0.17	1.51	0.22	0.07	0.33	0.00	*2.17*
Vernière	-0.34	-1.80	0.36	-0.05	0.27	-0.31	-0.14	0.14	*1.93*
Vichy-Célestins	3.20	0.86	-0.79	-0.32	-0.35	0.09	-0.40	-0.03	*3.46*
Wattwiller	-1.16	-0.46	-2.04	0.31	0.16	-0.06	0.26	-0.02	*2.43*

Mean	***0***	***0***	***0***	***0***	***0***	***0***	***0***	***0***	
Variance	***3.57***	***1.74***	***1.12***	***0.87***	***0.49***	***0.15***	***0.05***	***0.00***	***8***
Percent	***44.6%***	***21.7%***	***14.0%***	***10.9%***	***6.1%***	***1.9%***	***0.6%***	***0.0%***	***100.0%***

First factorial plane



	Axis 1	Axis 2
Arcens	-0.14	1.41
Arvie	2.02	-1.81
Badoit	-0.84	-0.85
Beckerich	-1.29	0.18
Châteauneuf	1.26	-0.67
Eau de Perrier	-1.83	0.89
Faustine	-0.11	-0.60
La Salvetat	-1.36	-0.75
Perrier	-1.86	0.94
Puits St-Georges	-0.58	1.89
Pyrénées	-1.88	1.78
Quézac	0.60	-2.68
San Pellegrino	-1.27	-2.12
St-Diéry	0.24	-0.41
St-Jean	-0.37	0.68
St-Pierre	0.19	1.12
St-Yorre	**6.32**	1.13
Vernet	-0.79	1.27
Vernière	-0.34	-1.80
Vichy-Célestins	3.20	0.86
Wattwiller	-1.16	-0.46

Mean	***0***	***0***
Variance	***3.57***	***1.74***
Percent	***44.6%***	***21.7%***

Approximate reconstruction of distances

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Axis 7	Axis 8
Arcens	-0.14	1.41	-0.04	1.34	0.38	0.55	-0.44	0.01
Arvie	2.02	-1.81	2.02	0.84	0.25	-0.93	-0.02	-0.07

- Distance computed with the first two axes:
  $d^2(\text{Arcens}, \text{Arvie}) \approx (-0.14 2.02)^2 + (1.41 + 1.81)^2 = 15.03$

- Distance computed with the first three axes:
  $d^2(\text{Arcens}, \text{Arvie}) \approx (-0.14 2.02)^2 + (1.41 + 1.81)^2 + (-0.04 - 2.02)^2 = 19.28$

- Distance computed with all eight axes:
  $d^2(\text{Arcens}, \text{Arvie}) = (-0.14 2.02)^2 + \cdots + (0.01 - 0.07)^2 = 21.93$
  Same as distance computed on the normed centered data.

Factorial axes and factors

Two views of a factorial axis α

– Factorial axis unit vector $\mathbf{u_\alpha}$.

– Factor $\psi_{\alpha i} = \sum_{j=1}^{p} u_{\alpha j}\, x_{ij}$

– $\mathbb{E}(\psi_\alpha) = 0$, $\mathrm{Var}(\psi_\alpha) = \lambda_\alpha$.

	u_1	u_2
Calcium	-0.04	-0.62
Magnésium	0.03	-0.58
Potassium	0.46	-0.14
Bicarbonates	0.51	0.00
Sulfates	0.08	-0.37
Fluorures	0.48	0.11
Sodium	0.51	0.14
Nitrates	-0.13	0.31
SumSqr	***1.00***	***1.00***

	Factor 1 Ψ_1	Factor 2 Ψ_2
Arcens	-0.14	1.41
Arvie	2.02	-1.81
Badoit	-0.84	-0.85
Beckerich	-1.29	0.18
Châteauneuf	1.26	-0.67
Eau de Perrier	-1.83	0.89
Faustine	-0.11	-0.60
La Salvetat	-1.36	-0.75
Perrier	-1.86	0.94
Puits St-Georges	-0.58	1.89
Pyrénées	-1.88	1.78
Quézac	0.60	-2.68
San Pellegrino	-1.27	-2.12
St-Diéry	0.24	-0.41
St-Jean	-0.37	0.68
St-Pierre	0.19	1.12
St-Yorre	6.32	1.13
Vernet	-0.79	1.27
Vernière	-0.34	-1.80
Vichy-Célestins	3.20	0.86
Wattwiller	-1.16	-0.46
Mean	***0.00***	***0.00***
Variance	***3.57***	***1.74***

Reconstruction

– On the factorial axis α

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} d_\alpha^2(i,i') = 2n\sum_{i=1}^{n} d_\alpha^2(i,G) = 2n\sum_{i=1}^{n}\psi_{\alpha i}^2 = 2n^2\mathsf{Var}(\psi_\alpha) = 2n^2\lambda_\alpha$$

– In the full space $\mathbb{R}^p$

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2(i,i') = \sum_{\alpha=1}^{p}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_\alpha^2(i,i') = 2n^2\sum_{\alpha=1}^{p}\lambda_\alpha = 2n^2 p$$

– Percent reconstruction on the first axis: $\frac{\lambda_1}{p}$.

– Percent reconstruction on the first two axes: $\frac{\lambda_1+\lambda_2}{p}$.

Computation of the PCA

Singular Value Decomposition of $X = V\,D\,U^\top$

- X: $n \times p$ data matrix of rank r.
- V: $n \times r$ orthogonal matrix.
- D: $r \times r$ diagonal matrix with elements $(\ldots \sqrt{\lambda_\alpha} \ldots)$.
- U: $p \times r$ orthogonal matrix.

Row PCA

- $\Sigma_{\text{row}} = X^\top X = U\,D\,V^\top V\,D\,U^\top = U\,D^2\,U^\top$.
- The unit vectors $\mathbf{u}_\alpha$ are the columns of U.
- The factors ψ_α are the columns of $X\,U = V\,D\,U^\top U = V\,D$.

Column PCA

- $\Sigma_{\text{col}} = X\,X^\top = V\,D\,U^\top U\,D\,V^\top = V\,D^2\,V^\top$.
- The unit vectors $\mathbf{v}_\alpha$ are the columns of V.
- The factors φ_α are the columns of $X^\top V = U\,D\,V^\top V = U\,D$.

Transition relations

Relation between row PCA and column PCA

- The following relations can be derived from the SVD equations.

$$\varphi\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X^\top \psi\alpha \qquad \psi\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X \varphi\alpha$$

$$\mathbf{u}\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \varphi\alpha \qquad \mathbf{v}\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \psi\alpha$$

Proof example

– Let $\mathbf{e}_\alpha$ be a basis vector of $\mathbb{R}^r$ and write

$$X\varphi_\alpha = (V\,D\,U^\top)(U\,D\,\mathbf{e}_\alpha) = V\,D\,D\,\mathbf{e}_\alpha = \sqrt{\lambda_\alpha}\,V\,D\,\mathbf{e}_\alpha = \sqrt{\lambda_\alpha}\,\psi_\alpha.$$

Summary

Normalization

– Correlation PCA: normalize mean and sdev (the default.)
– Covariance PCA: normalize mean (sometimes.)

PCA is a change of basis

– First factorial axis: direction with maximal dispersion.
– First factorial plane: plane with maximal dispersion.

Factor α

– Coordinates $\psi_{\alpha i}$ of all the observations on the axis α.

Dispersion

– Dispersion of the cloud on the first principal axis
  $=$ Variance of the first factor.
– Dispersion of the cloud on the first principal plane
  $=$ Sum of the variances of the first and second factors.

Contributions

Contribution of an observation to an axis

– How much an observation contributes to the definition of the axis?

$$CTR_\alpha(i) = \frac{\psi_{\alpha i}^2}{\sum_{i'=1}^{n} \psi_{\alpha i'}^2}$$

Contribution of a variable to an axis

– How much a variable contributes to the definition of the axis?

– Same thing using the PCA of the column points.

$$CTR_\alpha(j) = \frac{\varphi_{\alpha j}^2}{\sum_{j'=1}^{p} \varphi_{\alpha j'}^2}$$

Both types of contributions help interpreting the axes

Contributions of observations

			Coordinates				Contributions			
	Squared distance from origin	Fraction of dispersion	Axis 1	Axis 2	Axis 3	Axis 4	Axis 1	Axis 2	Axis 3	Axis 4
Arcens	4.4	2.6	-0.1	1.4	0.0	1.3	0.0	5.5	0.0	9.7
Arvie	13.1	7.8	2.0	-1.8	2.0	0.8	5.4	8.9	**17.4**	3.9
Badoit	7.1	4.2	-0.8	-0.9	1.5	-1.6	0.9	2.0	**10.2**	**13.3**
Beckerich	2.6	1.6	-1.3	0.2	-0.6	0.6	2.2	0.1	1.5	1.9
Châteauneuf	3.2	1.9	1.3	-0.7	-1.0	0.1	2.1	1.2	4.3	0.0
Eau de Perrier	5.8	3.5	-1.8	0.9	0.1	-1.0	4.5	2.2	0.0	5.4
Faustine	2.1	1.3	-0.1	-0.6	0.4	0.7	0.0	1.0	0.7	2.9
La Salvetat	7.4	4.4	-1.4	-0.7	-0.4	0.4	2.5	1.5	0.7	0.9
Perrier	6.3	3.8	-1.9	0.9	0.1	-1.1	4.6	2.4	0.0	7.1
Puits St-Georges	8.8	5.3	-0.6	1.9	1.3	-1.6	0.5	**9.8**	7.4	**14.5**
Pyrénées	7.1	4.2	-1.9	1.8	0.3	-0.4	4.7	8.7	0.3	0.7
Quézac	8.2	4.9	0.6	-2.7	0.5	0.0	0.5	**19.6**	1.1	0.0
San Pellegrino	14.6	8.7	-1.3	-2.1	-2.3	-1.2	2.1	**12.4**	**22.3**	7.9
St-Diéry	3.9	2.3	0.2	-0.4	1.6	0.5	0.1	0.5	**10.9**	1.6
St-Jean	1.1	0.7	-0.4	0.7	0.0	0.6	0.2	1.3	0.0	2.3
St-Pierre	3.0	1.8	0.2	1.1	-0.2	1.3	0.0	3.4	0.1	9.0
St-Yorre	**42.9**	**25.5**	6.3	1.1	-0.7	-1.0	**53.2**	3.5	2.2	5.3
Vernet	4.7	2.8	-0.8	1.3	-0.2	1.5	0.8	4.4	0.1	**12.5**
Vernière	3.7	2.2	-0.3	-1.8	0.4	0.0	0.2	8.9	0.5	0.0
Vichy-Célestins	12.0	7.1	3.2	0.9	-0.8	-0.3	**13.6**	2.0	2.7	0.5
Wattwiller	5.9	3.5	-1.2	-0.5	-2.0	0.3	1.8	0.6	**17.6**	0.5
	168	***100***					***100***	***100***	***100***	***100***

– *St Yorre* and *Vichy-Célestins* contribute most to the first axis.

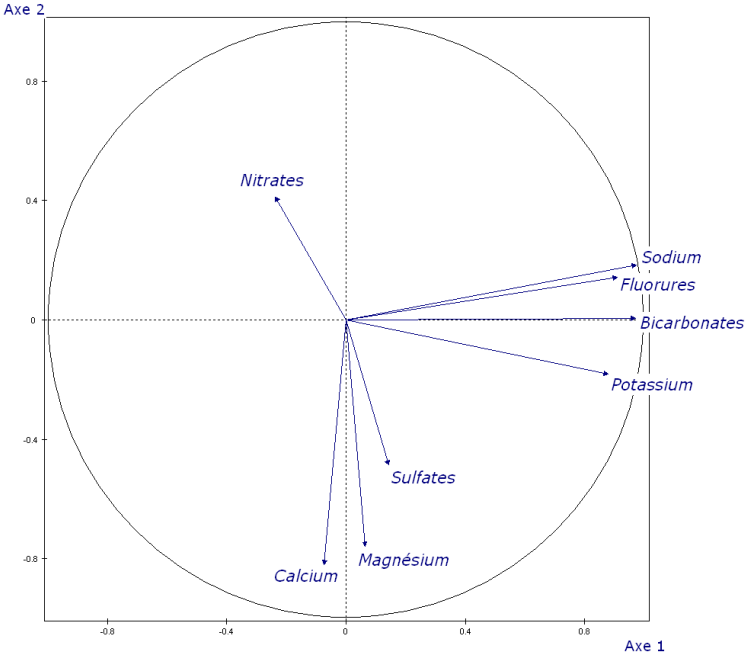
– *Quézac*, *San Pellegrino*, and *Puits St-Georges* contribute most to the second axis.

Contributions of variables

	Coordinates				Contributions			
	Axis 1	**Axis 2**	**Axis 3**	**Axis 4**	**Axis 1**	**Axis 2**	**Axis 3**	**Axis 4**
Calcium	-0.07	-0.82	-0.02	-0.29	0.16	**38.60**	0.04	9.73
Magnésium	0.06	-0.76	0.54	0.00	0.12	**33.06**	26.11	0.00
Potassium	0.88	-0.18	0.30	0.11	**21.60**	1.88	7.82	1.48
Bicarbonates	0.97	0.00	0.15	-0.11	**26.39**	0.00	1.90	1.52
Sulfates	0.14	-0.48	-0.73	-0.29	0.57	13.50	48.02	9.92
Fluorures	0.91	0.14	-0.28	-0.09	**23.14**	1.20	7.11	0.92
Sodium	0.97	0.18	0.00	-0.08	**26.46**	1.93	0.00	0.80
Nitrates	-0.24	0.41	0.32	-0.81	1.60	9.69	9.23	76.08
					100	***100***	***100***	***100***

– *Bicarbonates*, *Fluorures*, *Sodium*, and *Potassium* contribute most to the first axis.

– *Calcium* and *Magnesium* contribute most to the 2nd axis.



Simultaneously plotting rows and columns

Projecting the original axes

– Project points $\mathbf{e}_j = (\ldots, 0, 1, 0, 0, \ldots)$,

i.e. the tips of the corresponding unit vectors.

$$\psi_{\alpha,[j]} = \sum_{j'=1}^{p} \mathbb{I}(j = j')\, u_{\alpha j'} = u_{\alpha j}$$

	u_1	u_2
Calcium	-0.04	-0.62
Magnésium	0.03	-0.58
Potassium	0.46	-0.14
Bicarbonates	0.51	0.00
Sulfates	0.08	-0.37
Fluorures	0.48	0.11
Sodium	0.51	0.14
Nitrates	-0.13	0.31
SumSqr	***1.00***	***1.00***

– Relation with the column PCA:

$$\psi_{\alpha,[j]} = u_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}}\, \varphi_{\alpha j}$$

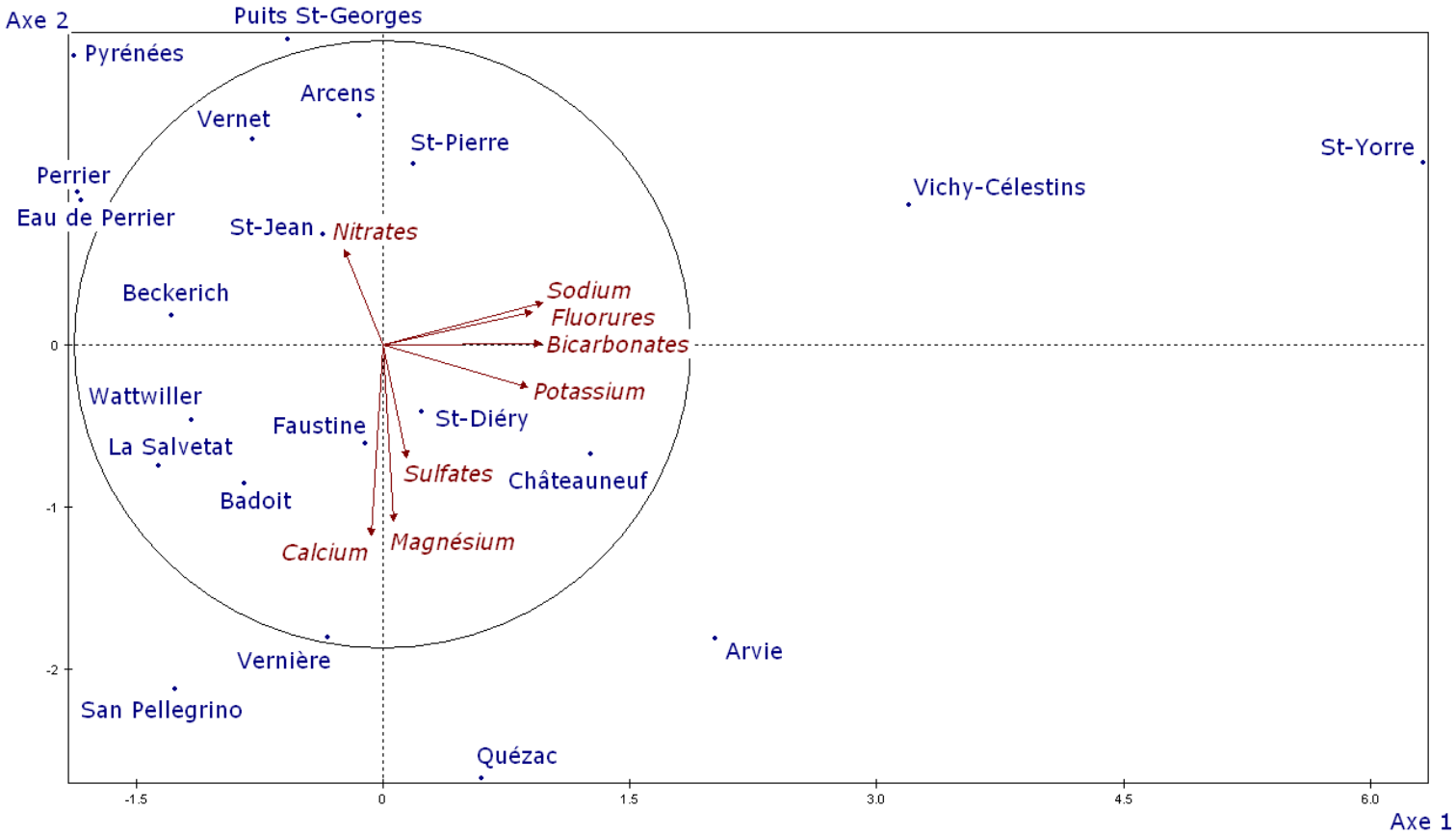
Correlation between variables and factorial axes

– The projected original axes reveal, up to a factor $\sqrt{\lambda_\alpha}$.

the correlation between the factorial axes and the original variable.

$$\frac{\mathrm{Correl}(x_{.j}, \psi_\alpha)}{\sqrt{\lambda_\alpha}} = \frac{\mathrm{Covar}(x_{.j}, \psi_\alpha)}{\lambda_\alpha} = \frac{\mathbf{e}_j^\top \Sigma\, \mathbf{u}_\alpha}{\lambda_\alpha} = \frac{\mathbf{e}_j^\top \lambda_\alpha \mathbf{u}_\alpha}{\lambda_\alpha} = u_{\alpha j} = \psi_{\alpha,[j]}$$

Simultaneously plotting rows and columns



– This particular plot uses a different scale for the points and the axes!

  The circle indicates the location of the "unit" sphere.

– Compare with the column PCA (slide 13.)

Continuous supplementary variables

Mineral water	Price
Arcens	0.34
Arvie	0.44
Badoit	0.64
Beckerich	0.17
Châteauneuf	0.58
Eau de Perrier	0.72
Faustine	0.2
La Salvetat	0.38
Perrier	0.94
Puits St-Georges	0.35
Pyrénées	0.3
Quézac	0.52
San Pellegrino	0.65
St-Diéry	0.32
St-Jean	0.4
St-Pierre	0.32
St-Yorre	0.53
Vernet	0.36
Vernière	0.39
Vichy-Célestins	0.59
Wattwiller	0.77

Price axis

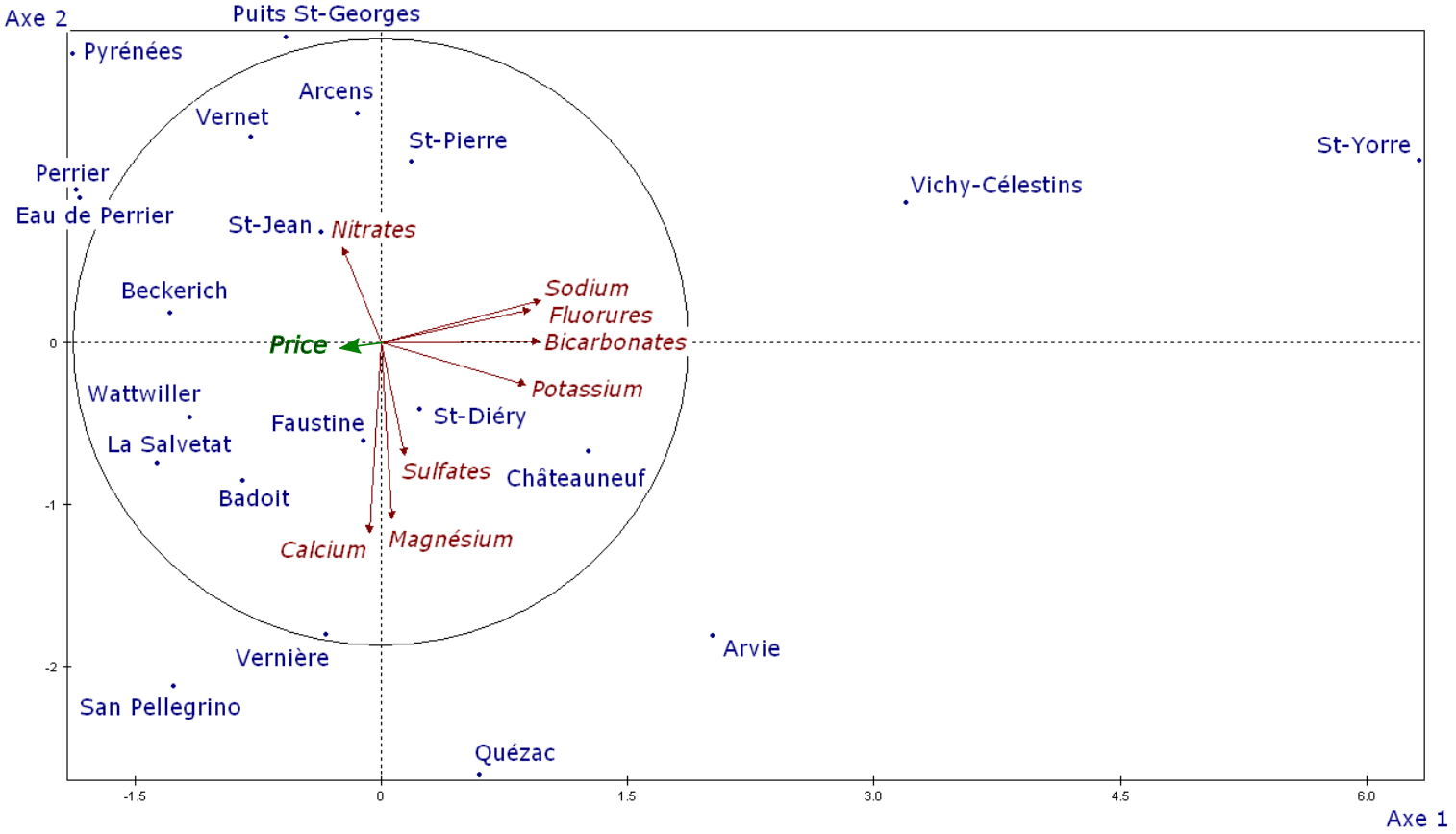
– Use the correlation formula

$$\psi_{\alpha,[\text{price}]} = \frac{\text{Correl}(\text{price}, \psi_\alpha)}{\sqrt{\lambda_\alpha}}$$

Relation with the unit sphere

$$\sum_{\alpha=1}^{p} \psi_{\alpha,[j]}^2 = 1 \qquad \sum_{\alpha=1}^{p} \psi_{\alpha,[\text{price}]}^2 < 1$$

Continuous supplementary variables



– The price of a sparkling spring water bears little relation to its mineral content.

Categorical supplementary variables

Mineral water	Region
Arcens	Rhône-Alpes
Arvie	Auvergne
Badoit	Rhône-Alpes
Beckerich	Other
Châteauneuf	Auvergne
Eau de Perrier	Languedoc-Roussillon
Faustine	Rhône-Alpes
La Salvetat	Other
Perrier	Languedoc-Roussillon
Puits St-Georges	Rhône-Alpes
Pyrénées	Other
Quézac	Languedoc-Roussillon
San Pellegrino	Other
St-Diéry	Auvergne
St-Jean	Rhône-Alpes
St-Pierre	Rhône-Alpes
St-Yorre	Auvergne
Vernet	Other
Vernière	Languedoc-Roussillon
Vichy-Célestins	Auvergne
Wattwiller	Other

Partition observation in groups

– e.g. one group per region β.

Per category barycenter

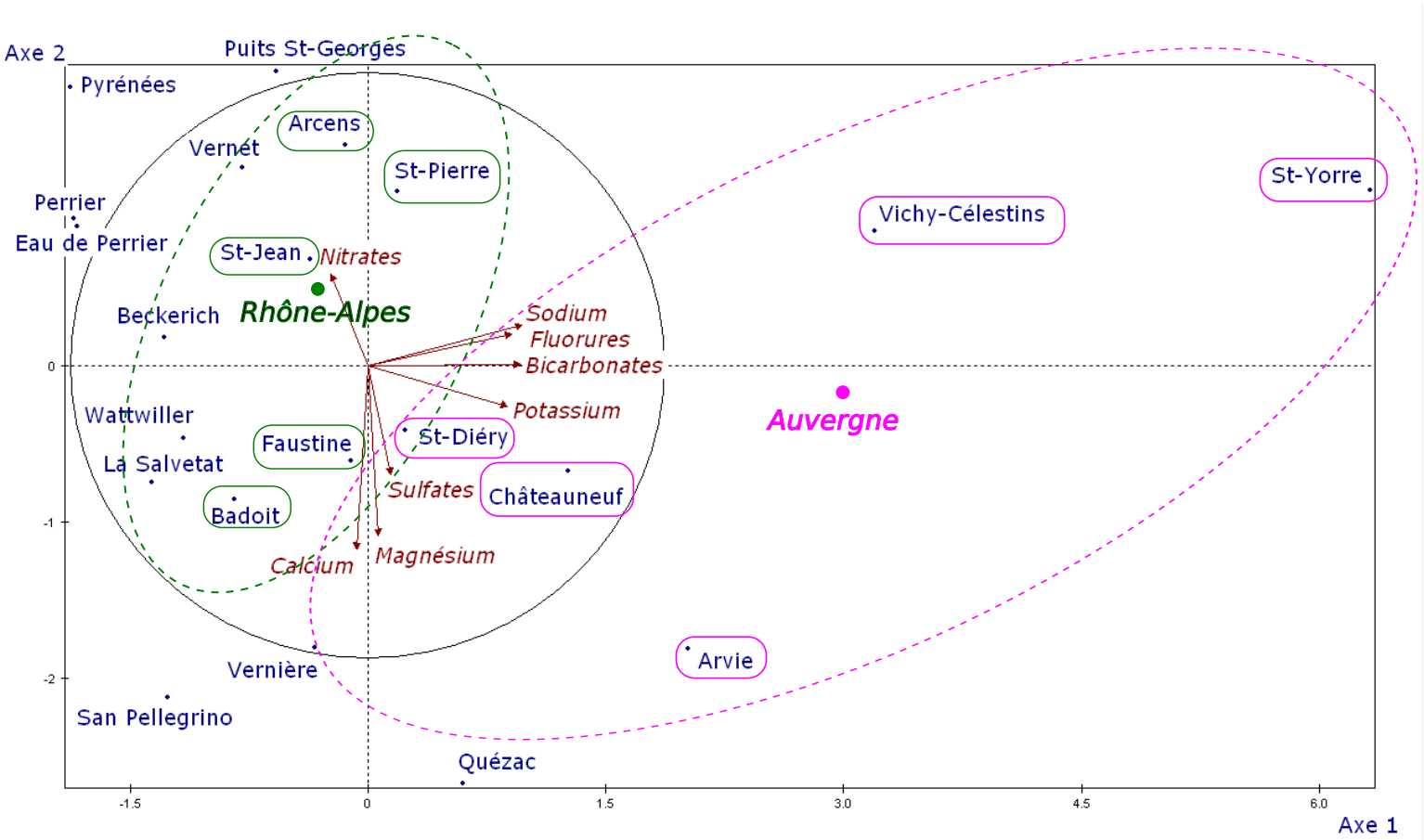
$$\bar{\psi}_{\alpha,[\beta]} = \frac{1}{n_\beta} \sum_{i \in \beta} \psi_{\alpha i}$$

Per category variance ellipsoid

$$\Sigma_{[\beta]} = \frac{1}{n_\beta} \sum_{i \in \beta} (\psi_{\alpha i} - \bar{\psi}_{\alpha,[\beta]})(\psi_{\alpha i} - \bar{\psi}_{\alpha,[\beta]})^\top$$

Use the central limit theorem to ascertain whether a group effect is significative !

Groups are often more interesting



– The barycenter of the six "Rhônes-Alpes" springs is close to the origin.

– The barycenter of the five "Auvergne" springs is high on the first axis.

Supplementary observations

What about Champagne?

	Concentration	Source
Calcium	86 mg/l	(Jos et. al., Talenta 63. 2004)
Magnesium	83 mg/l	(Jos et. al., Talenta 63. 2004)
Potassium	339 mg/l	(Jos et. al., Talenta 63. 2004)
Bicarbonate	1229 mg/l	(estimated: mean bicarbonates in water)
Sulfates	80 mg/l	(estimated: typical sulfites in wine)
Fluorures	1.5 mg/l	(estimated: mean fluorures in water)
Sodium	10 mg/l	(Jos et. al., Talenta 63. 2004)
Nitrates	2 mg/l	(estimated: mean nitrates in spring water)

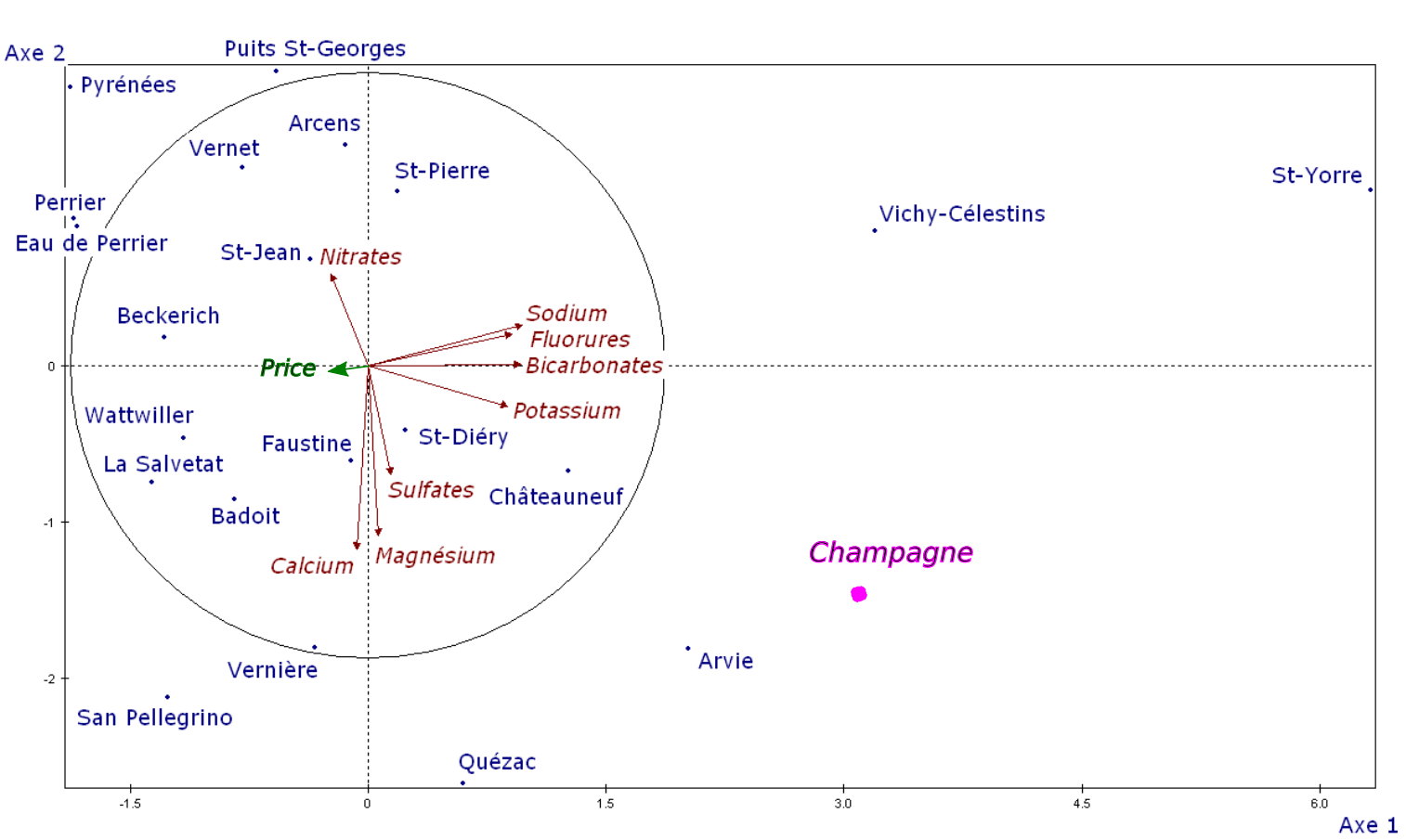
Changes of basis

Raw data	Calcium	Magnésium	Potassium	Bicarbonates	Sulfates	Fluorures	Sodium	Nitrates
Champagne	86	83	339	1229	80	1.5	10	1.98

Centered normed	Calcium	Magnésium	Potassium	Bicarbonates	Sulfates	Fluorures	Sodium	Nitrates
Champagne	-0.55	1.54	8.1	0	-0.13	-0.02	-0.79	0

Factors	Ψ_1	Ψ_2
Champagne	3.4163	-1.7203

Supplementary observations



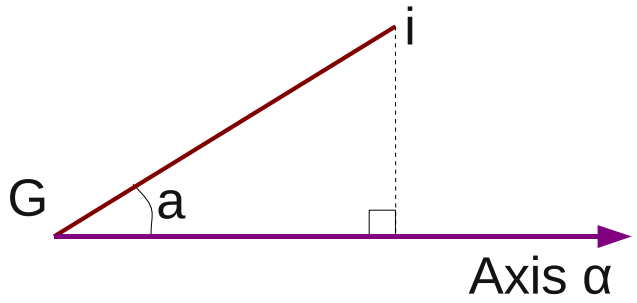
– The price of Champagne is not related to its mineral content (surprise!)

– How do we know if the Champage point is meaningful ?

Squared cosines

Quality of the projection on a factorial axis

$$QLT(i) = \cos^2(a) = \frac{\psi_{\alpha i}^2}{d^2(i,G)} = \frac{\psi_{\alpha i}^2}{\sum_{\alpha'=1}^{p} \psi_{\alpha' i}^2}$$

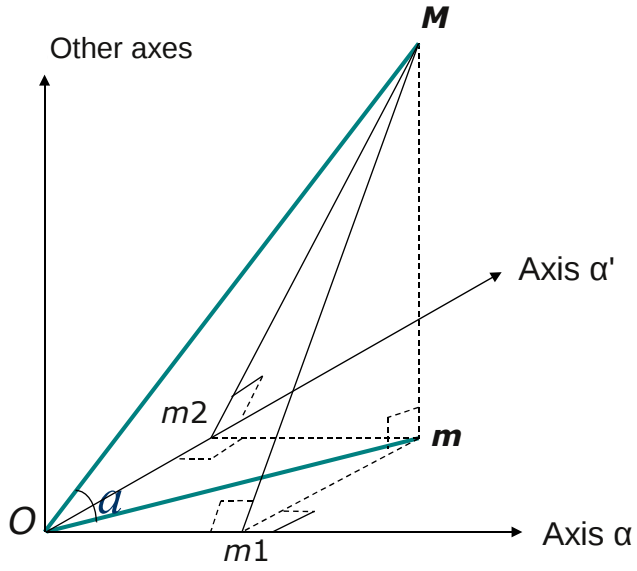


Quality of the projection on a factorial plane

– Squared cosines add up.

$$QLT_2(i) = \cos^2(a) = \frac{\psi_{\alpha i}^2 + \psi_{\alpha' i}^2}{d^2(i,G)}$$

$$= QLT_{(\text{for axis } \alpha)}(i) + QLT_{(\text{for axis } \alpha')}(i)$$



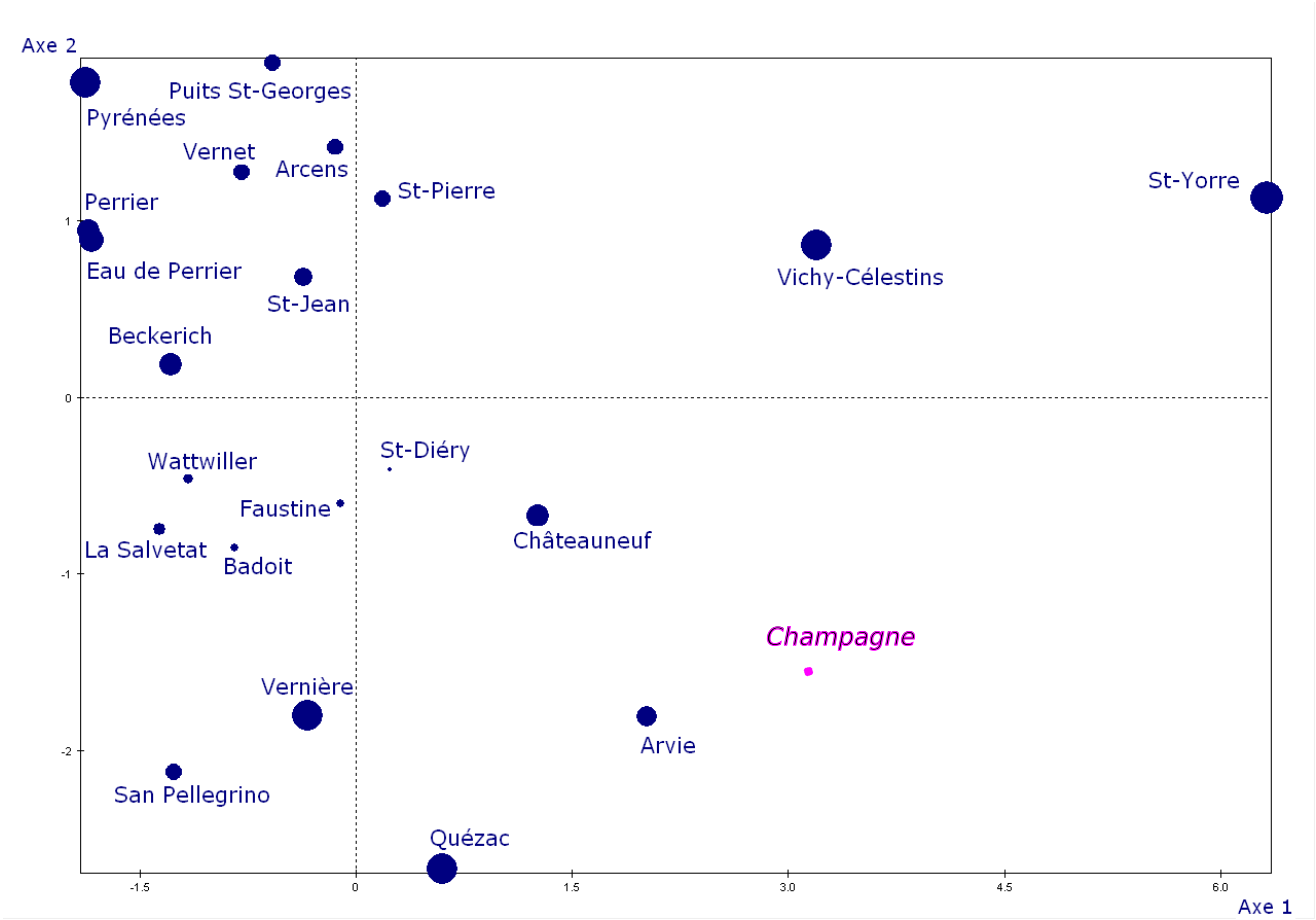
Squared cosines

		Coordinates		Squared cosines		
Name	**Squared distance to origine**	**Axis 1**	**Axis 2**	**Axis 1**	**Axis 2**	***QLT2***
Arcens	4.4	-0.14	1.41	0.00	0.45	*0.45*
Arvie	13.1	2.02	-1.81	0.31	0.25	*0.56*
Badoit	7.1	-0.84	-0.85	0.10	0.10	*0.20*
Beckerich	2.6	-1.29	0.18	0.62	0.01	*0.64*
Châteauneuf	3.2	1.26	-0.67	0.50	0.14	*0.64*
Eau de Perrier	5.8	-1.83	0.89	0.58	0.14	*0.71*
Faustine	2.1	-0.11	-0.60	0.01	0.17	*0.18*
La Salvetat	7.4	-1.36	-0.75	0.25	0.08	*0.33*
Perrier	6.3	-1.86	0.94	0.54	0.14	*0.68*
Puits St-Georges	8.8	-0.58	1.89	0.04	0.40	*0.44*
Pyrénées	7.1	-1.88	1.78	0.50	0.45	***0.95***
Quézac	8.2	0.60	-2.68	0.04	0.88	***0.92***
San Pellegrino	14.6	-1.27	-2.12	0.11	0.31	*0.42*
St-Diéry	3.9	0.24	-0.41	0.01	0.04	*0.06*
St-Jean	1.1	-0.37	0.68	0.12	0.42	*0.54*
St-Pierre	3.0	0.19	1.12	0.01	0.41	*0.42*
St-Yorre	**42.9**	6.32	1.13	0.93	0.03	***0.96***
Vernet	4.7	-0.79	1.27	0.13	0.34	*0.48*
Vernière	3.7	-0.34	-1.80	0.03	0.88	***0.91***
Vichy-Célestins	12.0	3.20	0.86	0.85	0.06	***0.91***
Wattwiller	5.9	-1.16	-0.46	0.23	0.04	*0.26*
Champagne	69.0	3.42	-1.72	0.17	0.04	*0.21*

– *Pyrénées*, *Quézac*, *St-Yorre*, *Vernière*, and *Vichy-Célestins* are well represented in the first factorial plane.

– *Champagne* is far behind but not the worst.

Squared cosines



Name	*QLT2*
Arcens	*0.45*
Arvie	*0.56*
Badoit	*0.20*
Beckerich	*0.64*
Châteauneuf	*0.64*
Eau de Perrier	*0.71*
Faustine	*0.18*
La Salvetat	*0.33*
Perrier	*0.68*
Puits St-Georges	*0.44*
Pyrénées	***0.95***
Quézac	***0.92***
San Pellegrino	*0.42*
St-Diéry	*0.06*
St-Jean	*0.54*
St-Pierre	*0.42*
St-Yorre	***0.96***
Vernet	*0.48*
Vernière	***0.91***
Vichy-Célestins	***0.91***
Wattwiller	*0.26*

Champagne	*0.21*

How many axes?

For visualisation

– Always investigate at least the first three factorial axes.

– Plot multiple factorial planes (12,23,13,. . .)

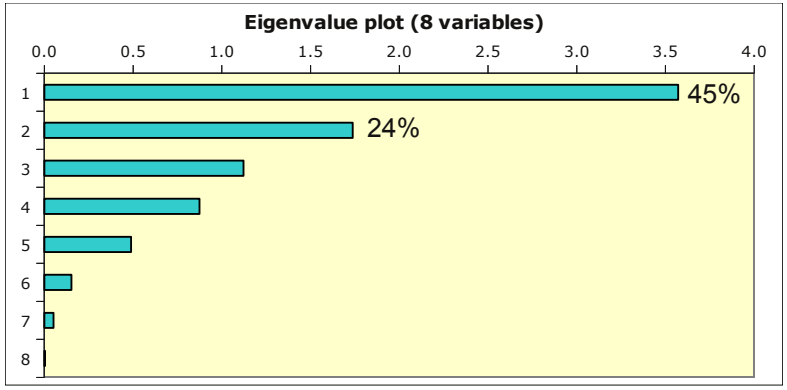
Heuristics

– Search the “elbow” in the plot of decreasing eigenvalues.

– Discard axes with eigenvalue smaller than 1.

Stability

– Evaluated with bootstrap procedures.

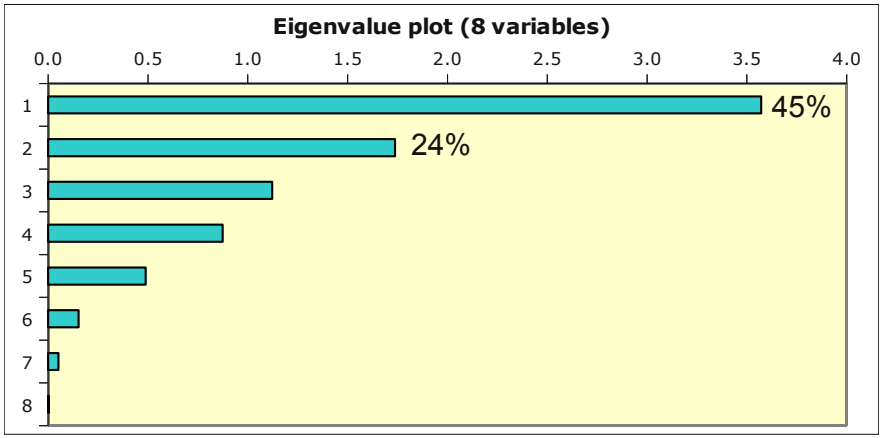


How many axes?

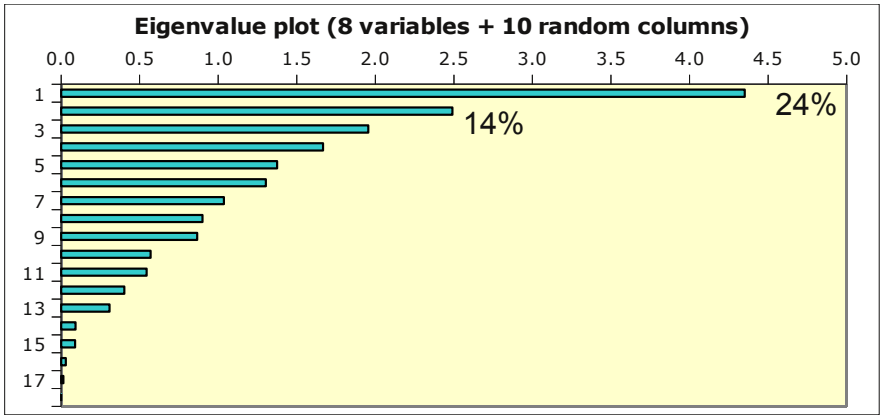
Percentages of variance can be misleading

– Two examples with the same information but different noise levels.

Original data.



Same data with 10 additional random columns.



II. Correspondence Analysis

Hair and eye colors

We know for 592 english women

– the color of their eyes

– the color of their hair.

n

		Hair color				
		Dark	Auburn	Red	Blond	*Totals*
Eyes color	Brown	68	119	26	7	***220***
	Hazel	15	54	14	10	***93***
	Green	5	29	14	16	***64***
	Blue	20	84	17	94	***215***
	Totals	***108***	***286***	***71***	***127***	***592***

p

Contingency table $[x_{ij}]$

$$x_{i\bullet} = \sum_{j=1}^{p} x_{ij} \qquad x_{\bullet j} = \sum_{i=1}^{n} x_{ij} \qquad x_{\bullet\bullet} = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}$$

Row and column profiles

Rows and columns are best described by their histograms.

Row profiles

		Hair color					
		Dark	Auburn	Red	Blond	*Totals*	*Mass*
Eyes color	Brown	31%	54%	12%	3%	*100%*	*37.16%*
	Hazel	16%	58%	15%	11%	*100%*	*15.71%*
	Green	8%	45%	22%	25%	*100%*	*10.81%*
	Blue	9%	39%	8%	44%	*100%*	*36.32%*
Average row profile		*18.24%*	*48.31%*	*11.99%*	*21.45%*		

$$r_{ij} = \frac{x_{ij}}{x_{i\bullet}} \qquad m_i = \frac{x_{i\bullet}}{x_{\bullet\bullet}}$$

$$c_j = \frac{x_{\bullet j}}{x_{\bullet\bullet}}$$

Column profiles

		Hair color				Average column profile
		Dark	Auburn	Red	Blond	
Eyes color	Brown	63%	42%	37%	6%	*37%*
	Hazel	14%	19%	20%	8%	*16%*
	Green	5%	10%	20%	13%	*11%*
	Blue	19%	29%	24%	74%	*36%*
	Totals	*100%*	*100%*	*100%*	*100%*	
	Mass	*18.24%*	*48.31%*	*11.99%*	*21.45%*	

$$o_{ij} = \frac{x_{ij}}{x_{\bullet j}} \qquad m_i = \frac{x_{i\bullet}}{x_{\bullet\bullet}}$$

$$c_j = \frac{x_{\bullet j}}{x_{\bullet\bullet}}$$

Row and column profiles

Row profiles

		Hair color					
		Dark	Auburn	Red	Blond	*Totals*	*Mass*
Eyes color	Brown	31%	54%	12%	3%	*100%*	*37.16%*
	Hazel	16%	58%	15%	11%	*100%*	*15.71%*
	Green	8%	45%	22%	25%	*100%*	*10.81%*
	Blue	9%	39%	8%	44%	*100%*	*36.32%*
Average row profile		*18.24%*	*48.31%*	*11.99%*	*21.45%*		

$$r_{ij} = \frac{x_{ij}}{x_{i\bullet}} \qquad m_i = \frac{x_{i\bullet}}{x_{\bullet\bullet}}$$

$$c_j = \frac{x_{\bullet j}}{x_{\bullet\bullet}}$$

Remarks

– The "mass" indicates the relative importance of each row.

– The "average row profile" is not the mean of the row profiles.

– The "average row profile" is the weighted mean of the row profiles.

$$\frac{1}{n}\sum_{i=1}^{n} r_{ij} \neq c_j \qquad \sum_{i=1}^{n} m_i\, r_{ij} = \sum_{i=1}^{n} \frac{x_{i\bullet}}{x_{\bullet\bullet}}\,\frac{x_{ij}}{x_{i\bullet}} = c_j$$

– PCA on the row profiles?

Centering the columns

Subtracting the average row profile

		Hair color				
		Dark	Auburn	Red	Blond	***Mass***
Eyes color	Brown	0.127	0.058	-0.002	-0.183	***37.16%***
	Hazel	-0.021	0.098	0.031	-0.107	***15.71%***
	Green	-0.104	-0.030	0.099	0.035	***10.81%***
	Blue	-0.089	-0.092	-0.041	0.223	***36.32%***
Average row profile		***18.24%***	***48.31%***	***11.99%***	***21.45%***	

$$r_{ij} - c_j = \frac{x_{ij}}{x_{i\bullet}} - \frac{x_{\bullet j}}{x_{\bullet\bullet}}$$

– This is smarter than subtracting the column average.
Aggregating two identical rows does not change the result.

– This table is centered only if we take the masses into account.
From now on we must always take the masses into account.
For instance, when computing covariances.

Rescaling the columns (1)

Standard deviation of the columns?

– The standard deviation of a column is a bad measure.
  The difference between 43% and 44% is a small difference.
  The difference between 1% and 2% is a big difference.

The binomial argument

– The x_{ij} are counting events whose probability is roughly c_j.
– The standard deviation of the r_{ij} would then be $\sqrt{c_j(1-c_j)}$.
– The c_j are usually well below one because $\sum_j c_j = 1$.

– Conclusion: divide the columns by $\sqrt{c_j}$.

– This will all make more sense later. . .

Rescaling the columns (2)

Normalized row profiles

		Hair color				
		Dark	Auburn	Red	Blond	***Mass***
Eyes color	Brown	0.30	0.08	-0.01	-0.39	***37.16%***
	Hazel	-0.05	0.14	0.09	-0.23	***15.71%***
	Green	-0.24	-0.04	0.29	0.08	***10.81%***
	Blue	-0.21	-0.13	-0.12	0.48	***36.32%***
Average row profile		***18.24%***	***48.31%***	***11.99%***	***21.45%***	

$$y_{ij} = \frac{r_{ij} - c_j}{\sqrt{c_j}} = \sqrt{\frac{1}{c_j}} \left(\frac{x_{ij}}{x_{i\bullet}} - \frac{x_{\bullet j}}{x_{\bullet\bullet}} \right)$$

Euclidian distance between two normalized rows

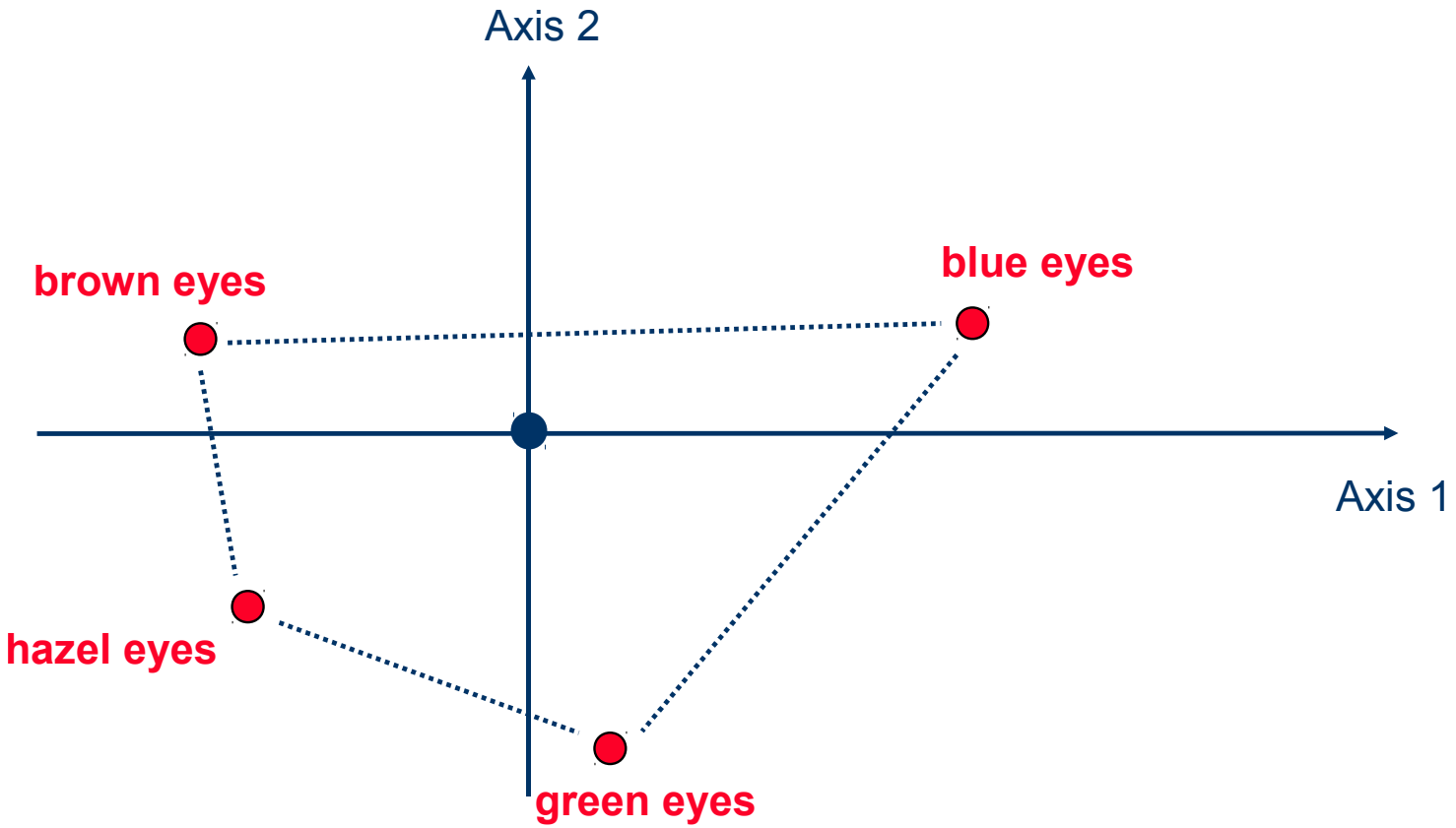
$$d(i,i') = \sum_{j=1}^{p} \frac{(r_{ij} - r_{i'j})^2}{c_j} = \sum_{j=1}^{p} \frac{x_{\bullet\bullet}}{x_{\bullet j}} \left(\frac{x_{ij}}{x_{i\bullet}} - \frac{x_{i'j}}{x_{i'\bullet}} \right)^2$$

– This is called the χ^2 distance.

– This will all make more sense later...

Principal component analysis

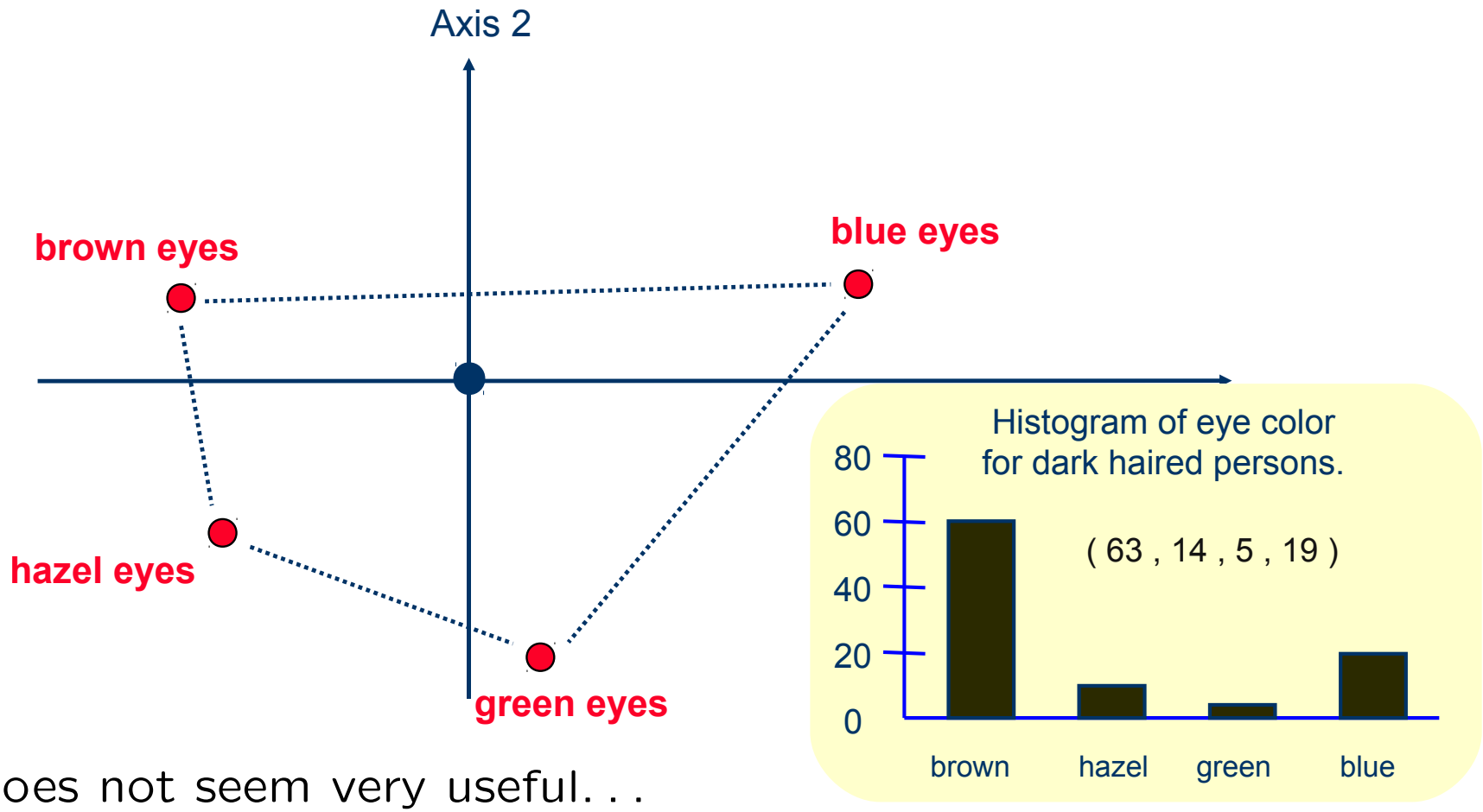
– Compute the covariance matrix (weighted by the masses)
– Diagonalize and project on the two first axes.



– This does not seem very useful. . .

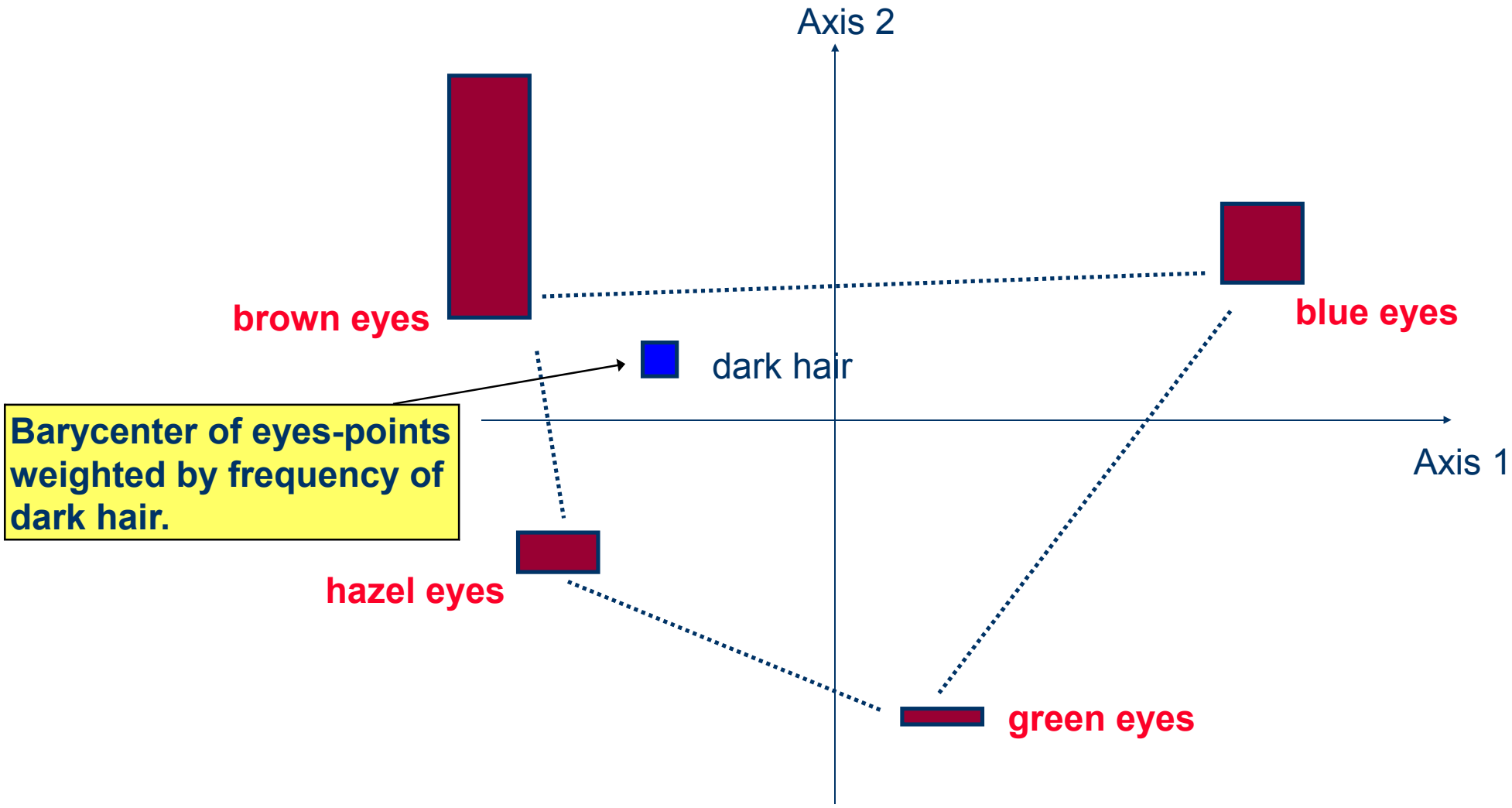
Principal component analysis

– Compute the covariance matrix (weighted by the masses)
– Diagonalize and project on the two first axes.

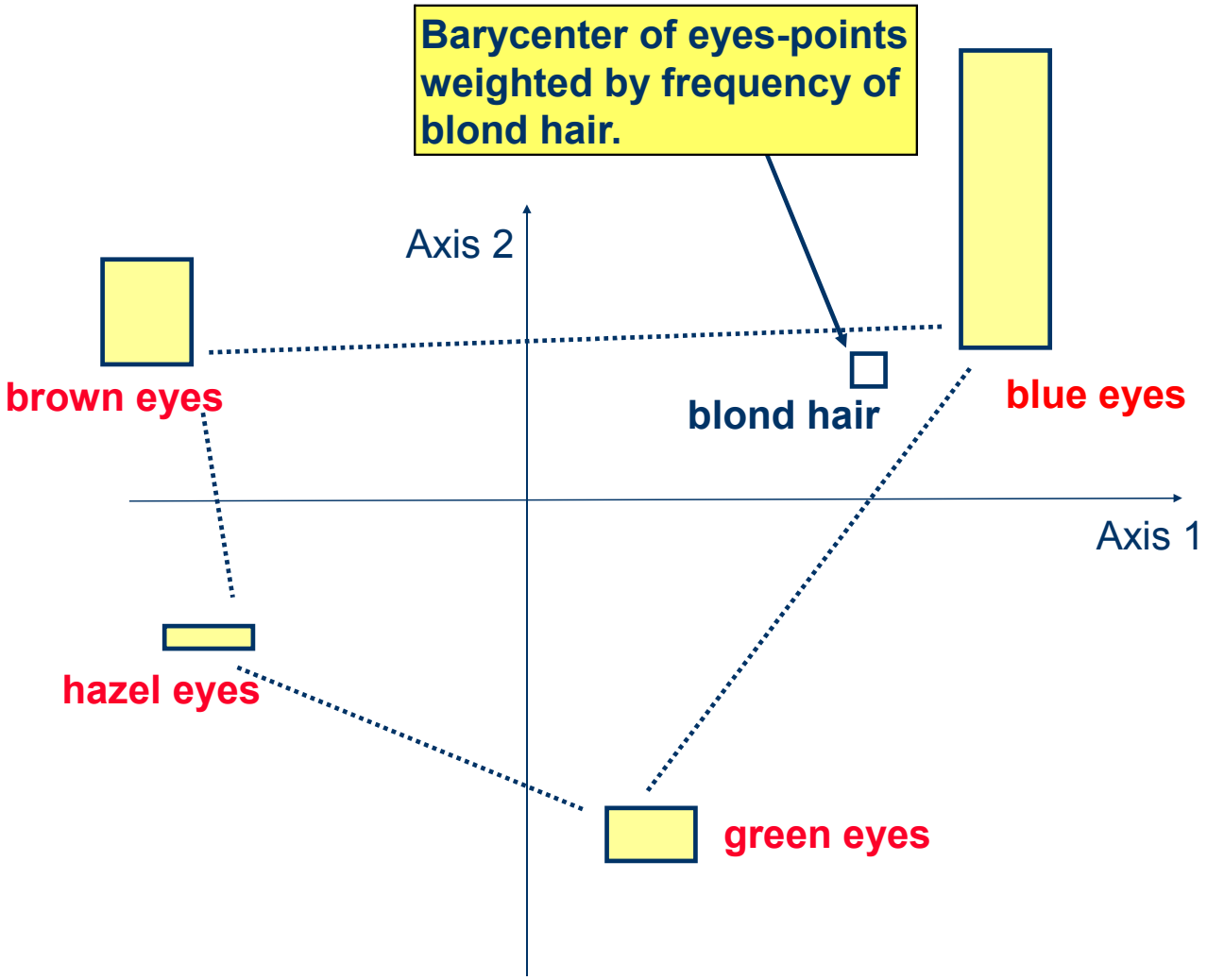


– This does not seem very useful. . .

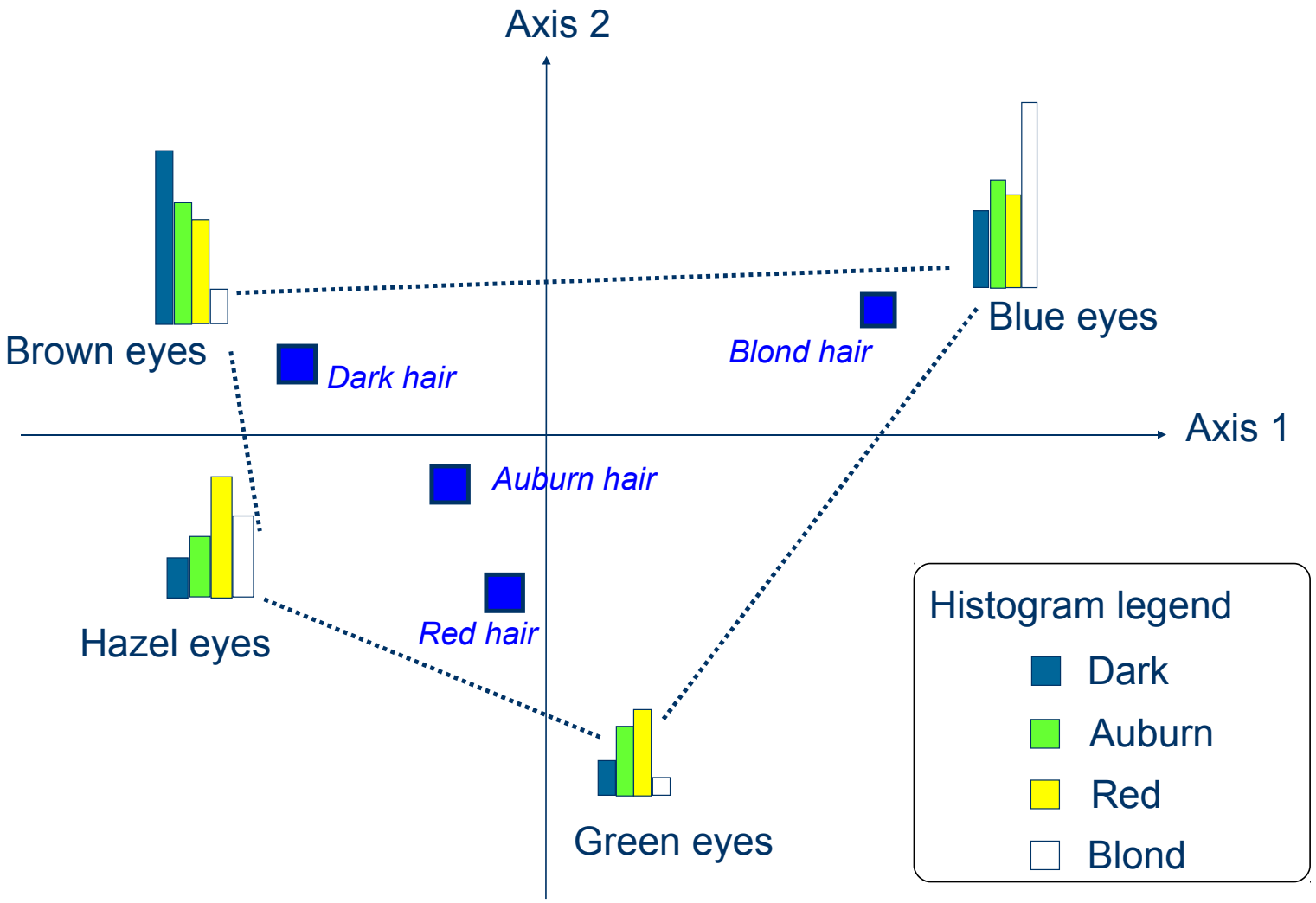
Placing the columns in the row space



Placing the columns in the row space

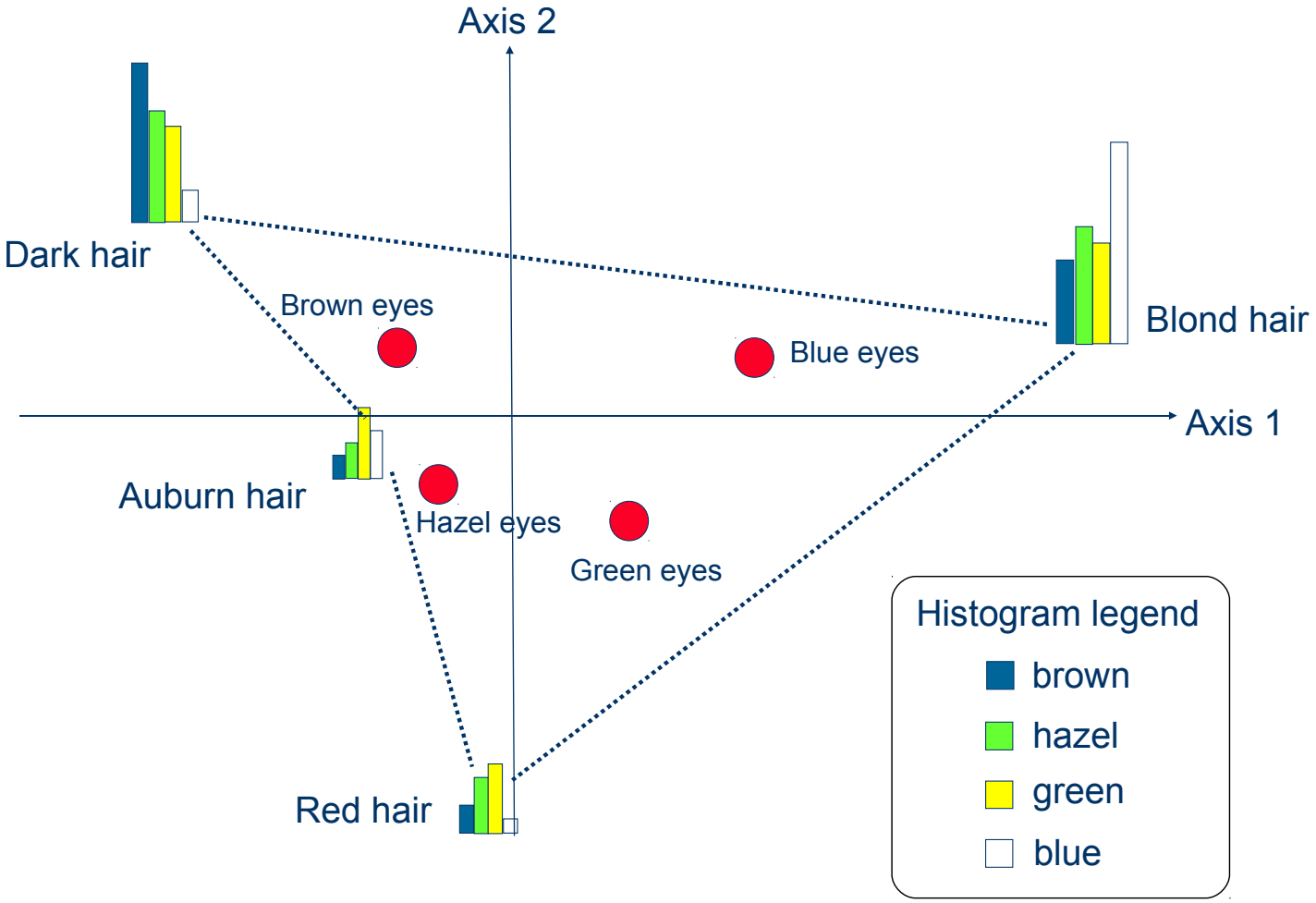


Placing the columns in the row space



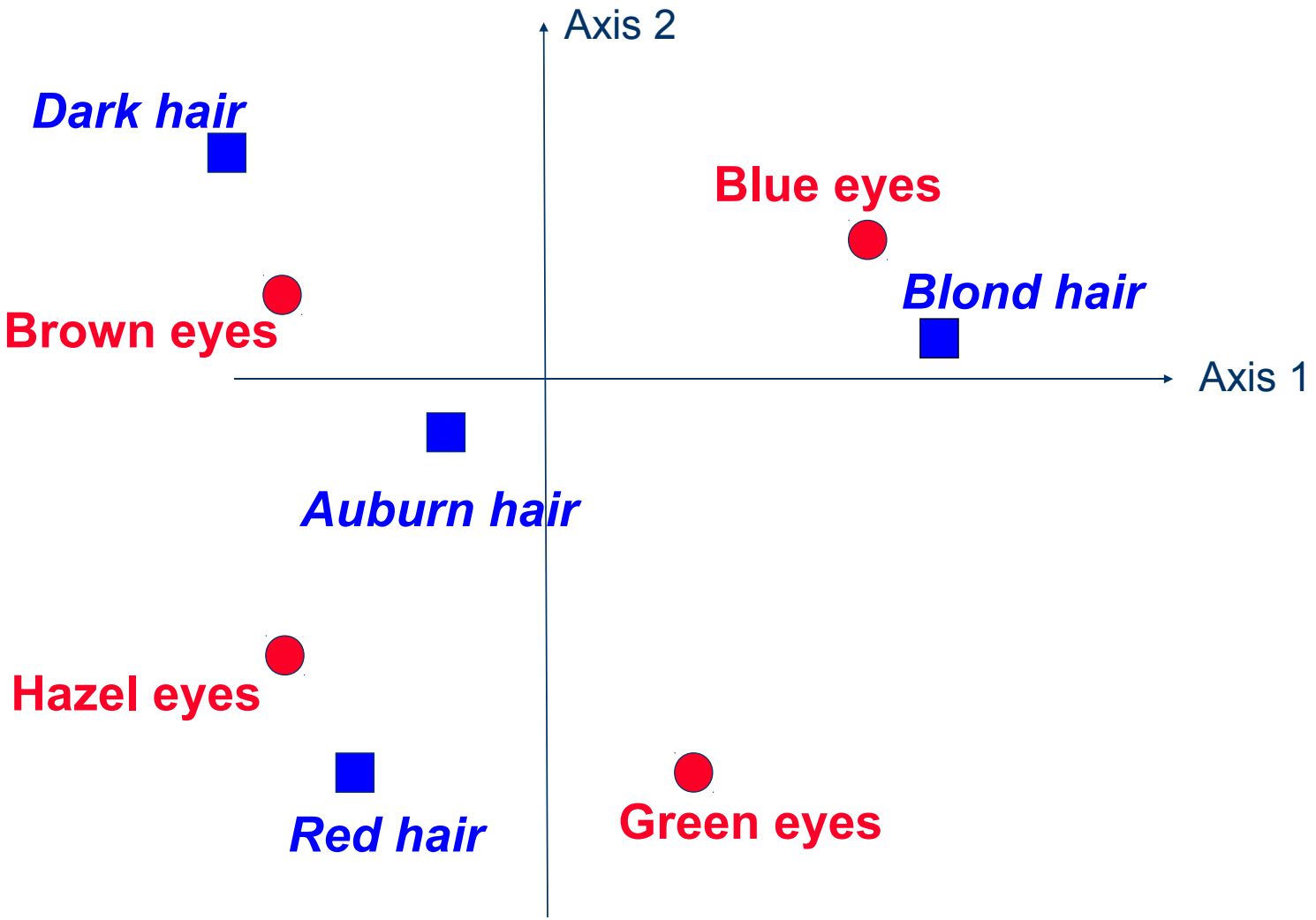
Placing the rows in the column space

– Same thing with the columns, including centering and rescaling.

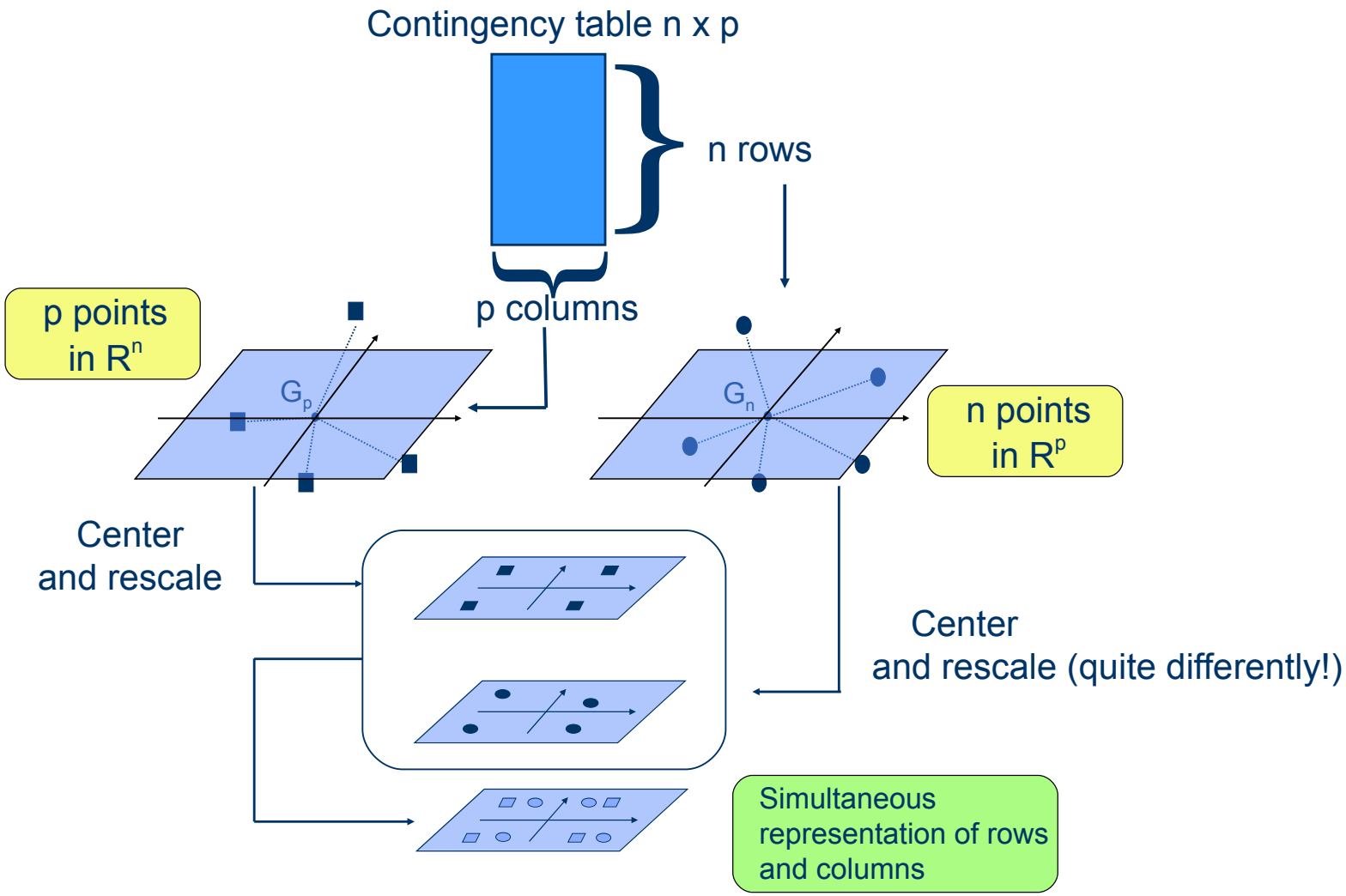


– Same shapes, different scale. . .

Simultaneous representation



Summary


Contingency table n x p
n rows
p columns
p points in R^n
G_p
G_n
n points in R^p
Center and rescale
Center and rescale (quite differently!)
Simultaneous representation of rows and columns

Duality of the row and column analysis

Standard PCA

– Center and rescale the columns (mean and sdev).
– Diagonalize the row covariance of the normalized table.
– Diagonalize the column covariance of the same normalized table.
– Dual representation arise from the properties of diagonalization.

Correspondence Analysis

– Center and rescale the columns, diagonalize the row covariance.
– Center and rescale the rows, diagonalize the column covariance.
– Why do we get a dual representation?

Duality of the row and column analysis

Weighted covariance

– We diagonalize the weighted covariance matrix $\Sigma = Y^\top D_m Y$ where Y is the normalized row profile matrix and $D_c = \mathrm{diag}(m_1 \ldots m_p)$.

– We can write $\Sigma = Z^\top Z$ with $Z = D_m^{\frac{1}{2}} Y$.

Divergence matrix Z for the row analysis

$$z_{ij} = \sqrt{\frac{m_i}{c_j}} \left(\frac{x_{ij}}{x_{i\bullet}} - \frac{x_{\bullet j}}{x_{\bullet\bullet}} \right) = \sqrt{\frac{x_{i\bullet}}{x_{\bullet j}}} \left(\frac{x_{ij}}{x_{i\bullet}} - \frac{x_{\bullet j}}{x_{\bullet\bullet}} \right) = \frac{x_{ij} - \frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}}{\sqrt{x_{i\bullet} x_{\bullet j}}} = \frac{\frac{x_{ij}}{x_{\bullet\bullet}} - m_i c_j}{\sqrt{m_i c_j}}$$

Divergence matrix Z for the column analysis

$$z_{ij} = \sqrt{\frac{c_j}{m_i}} \left(\frac{x_{ij}}{x_{\bullet j}} - \frac{x_{i\bullet}}{x_{\bullet\bullet}} \right) = \sqrt{\frac{x_{\bullet j}}{x_{i\bullet}}} \left(\frac{x_{ij}}{x_{\bullet j}} - \frac{x_{i\bullet}}{x_{\bullet\bullet}} \right) = \frac{x_{ij} - \frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}}{\sqrt{x_{i\bullet} x_{\bullet j}}} = \frac{\frac{x_{ij}}{x_{\bullet\bullet}} - m_i c_j}{\sqrt{m_i c_j}}$$

The dual representation exists because this is the same matrix.

The χ^2 test of independence

– The real data table:

$[x_{ij}]$

Eyes color	Hair color: Dark	Auburn	Red	Blond	Totals
Brown	68	119	26	7	***220***
Hazel	15	54	14	10	***93***
Green	5	29	14	16	***64***
Blue	20	84	17	94	***215***
Totals	***108***	***286***	***71***	***127***	***592***

– The theoretical data table assuming independence:

$[x_{\bullet\bullet} \times m_i\, c_j]$

Eyes color	Hair color: Dark	Auburn	Red	Blond	Totals
Brown	40.1	106.3	26.4	47.2	***220***
Hazel	17.0	44.9	11.2	20.0	***93***
Green	11.7	30.9	7.7	13.7	***64***
Blue	39.2	103.9	25.8	46.1	***215***
Totals	***108***	***286***	***71***	***127***	***592***

– The *inertia* $\mathcal{I} = \sum_{ij} z_{ij}^2 = \sum_{ij} \frac{1}{m_i\, c_j} \left(\frac{x_{ij}}{x_{\bullet\bullet}} - m_i\, c_j \right)^2$

measures how dependent are the rows and columns.

– Correspondence analysis finds the axes that best display this dependence !

The numerical recipe

Compute the divergence matrix

$$Z = \left[\frac{\frac{x_{ij}}{x_{\bullet\bullet}} - m_i c_j}{\sqrt{m_i c_j}} \right]$$

Singular Value decomposition

$$Z = V\ D\ U^\top \qquad \text{with} \quad D = \text{diag}(\sqrt{\lambda_\alpha})$$

Compute the factors

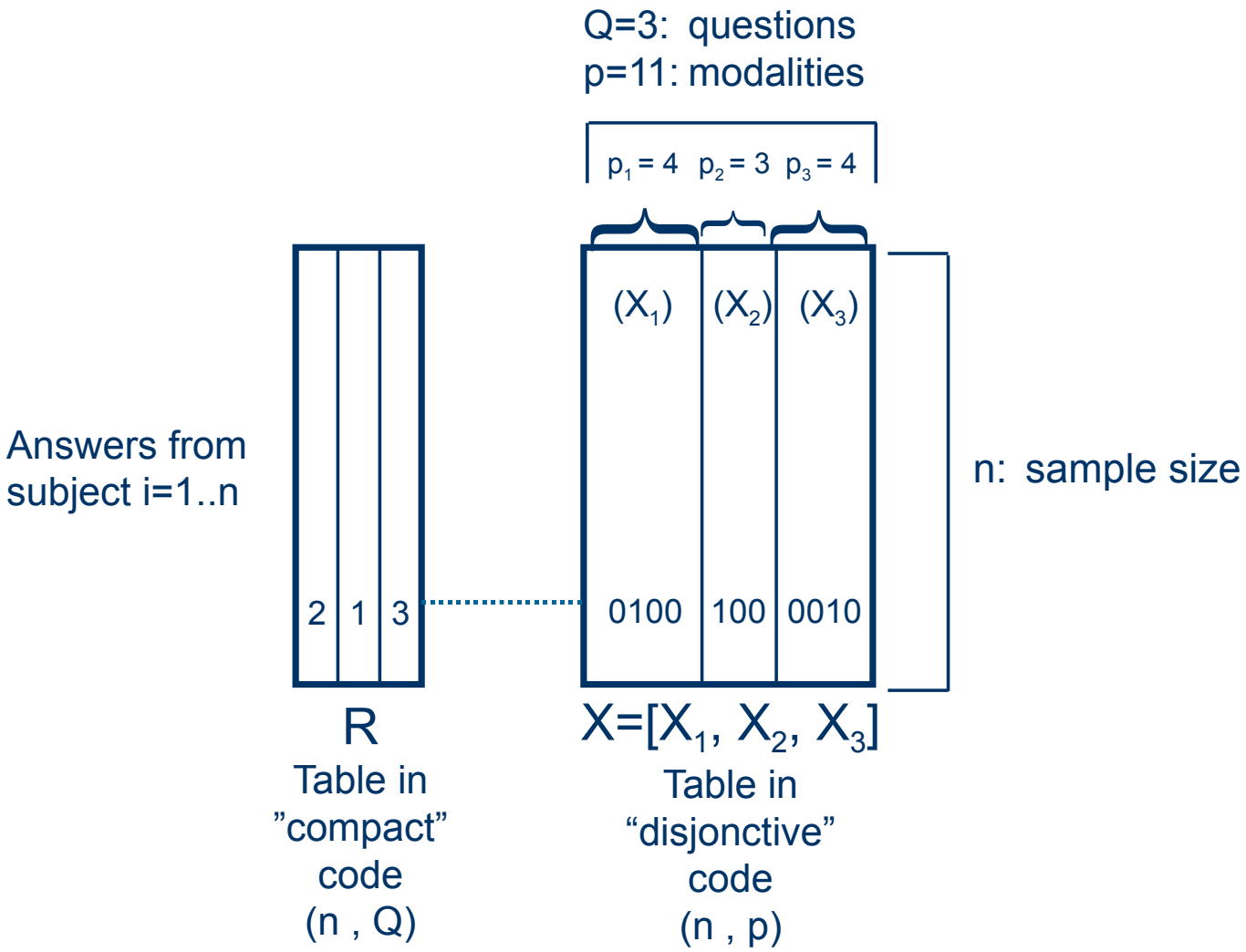
$$\psi = D_m^{-\frac{1}{2}}\ V\ D \qquad \varphi = D_c^{-\frac{1}{2}}\ U\ D$$

Transition relations

$$\psi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \frac{x_{ij}}{x_{i\bullet}}\, \varphi_{\alpha j} \qquad \varphi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{x_{ij}}{x_{\bullet j}}\, \psi_{\alpha i}$$

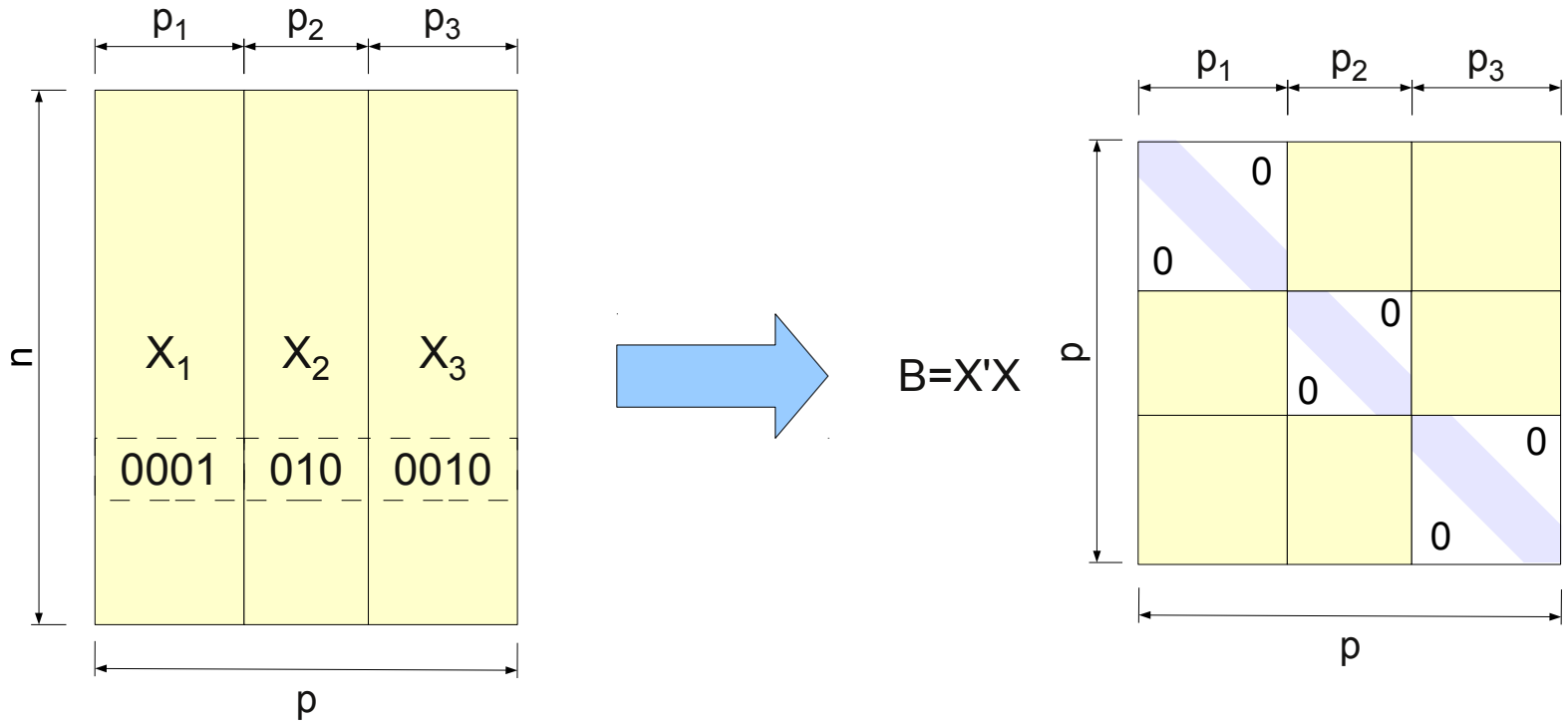
III. Multiple Correspondence Analysis

Polling n subjects with a questionnaire

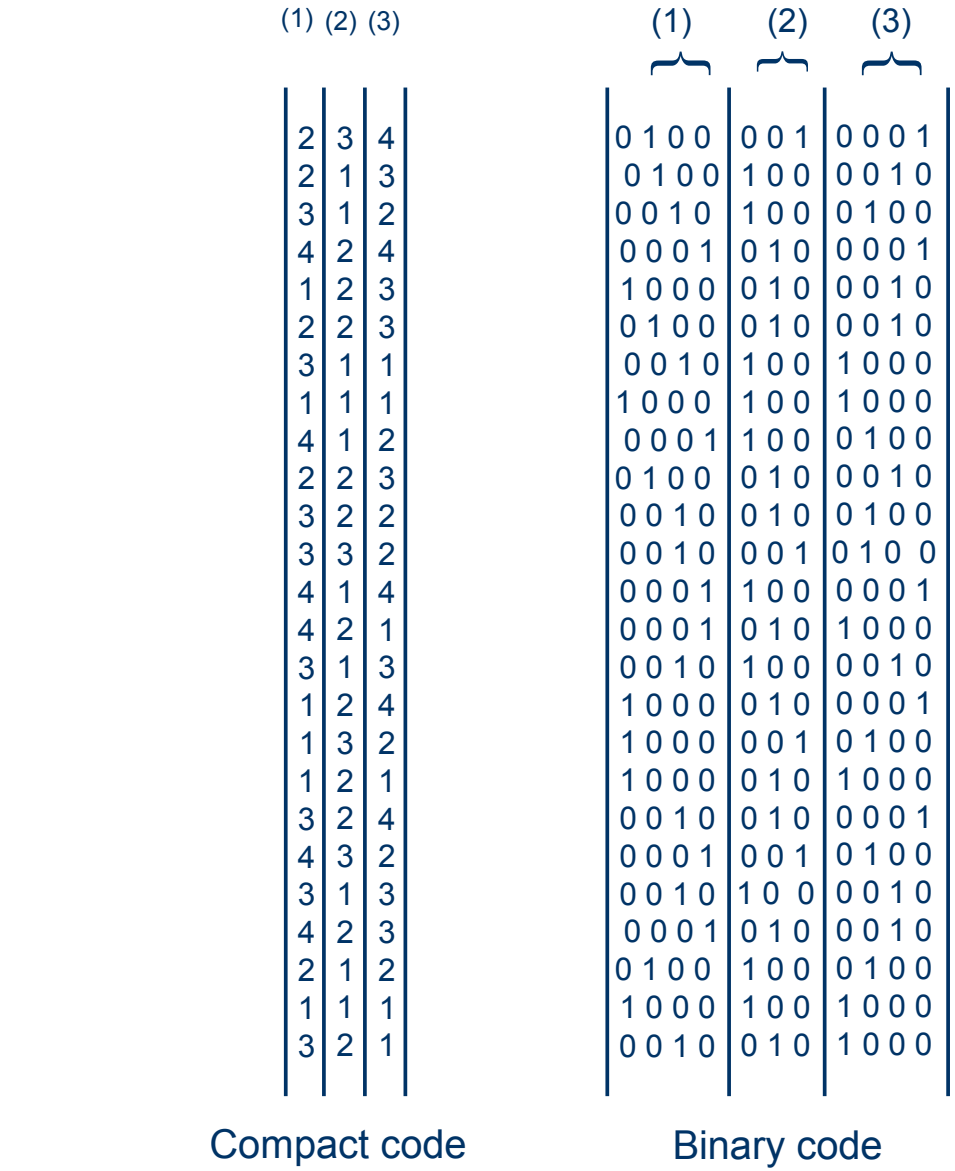


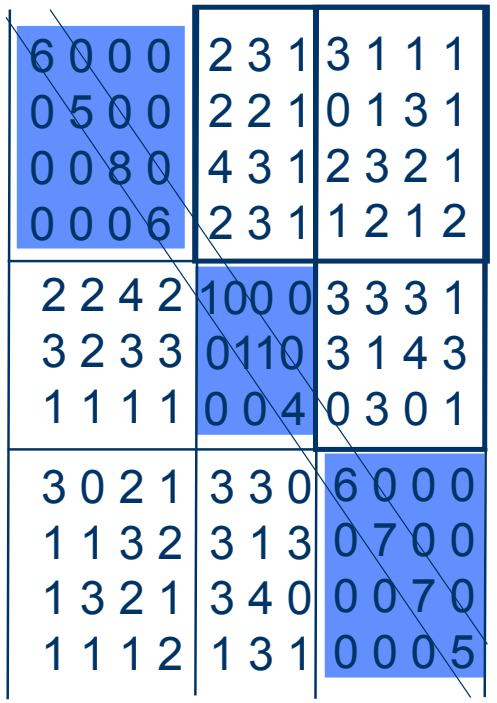
Burt table

Multiple contingency table



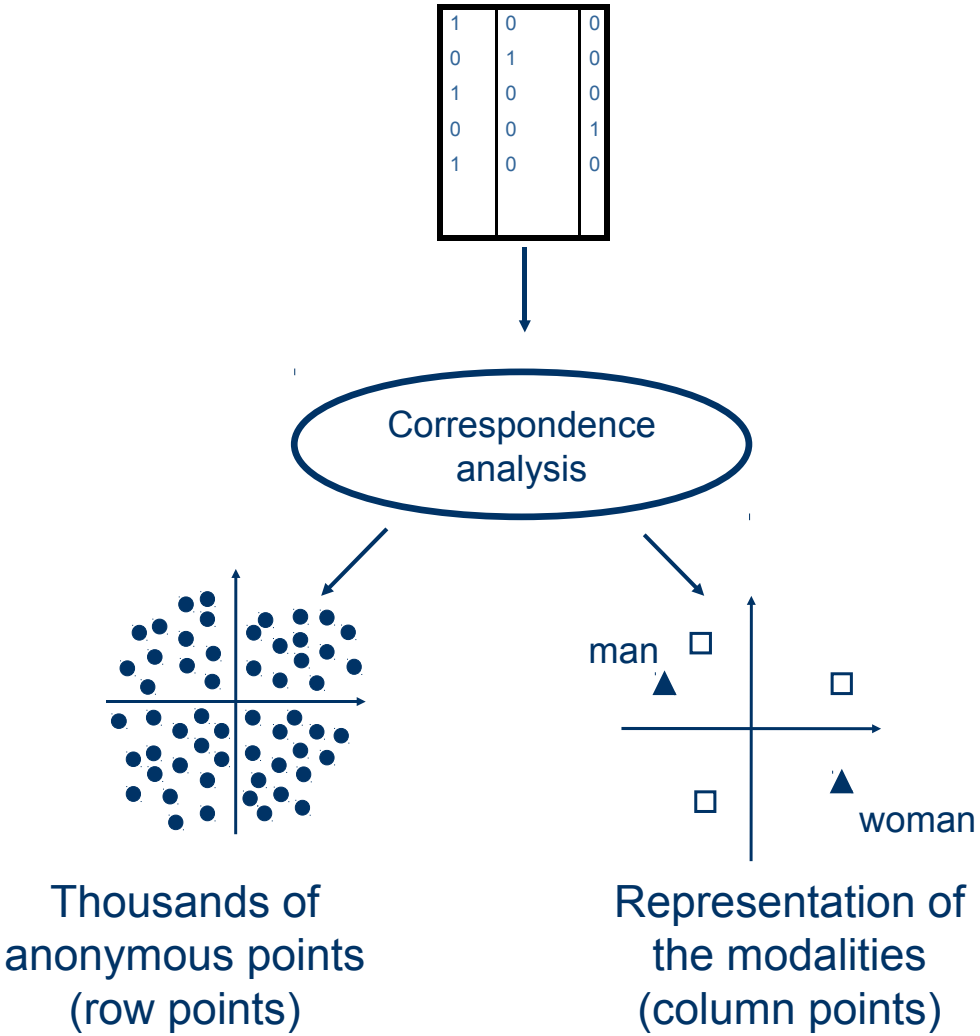
Encoding the Burt table





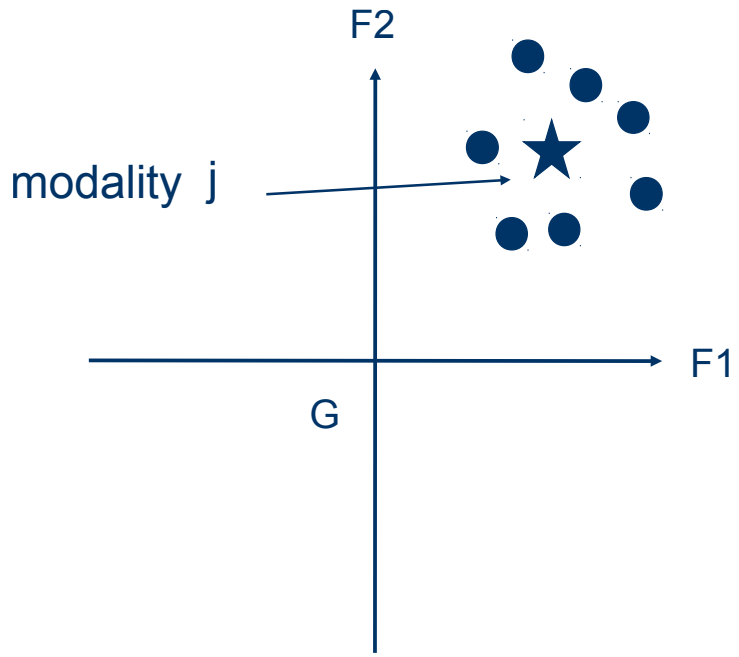
Multiple contingency table
Burt table

Multiple correspondence analysis



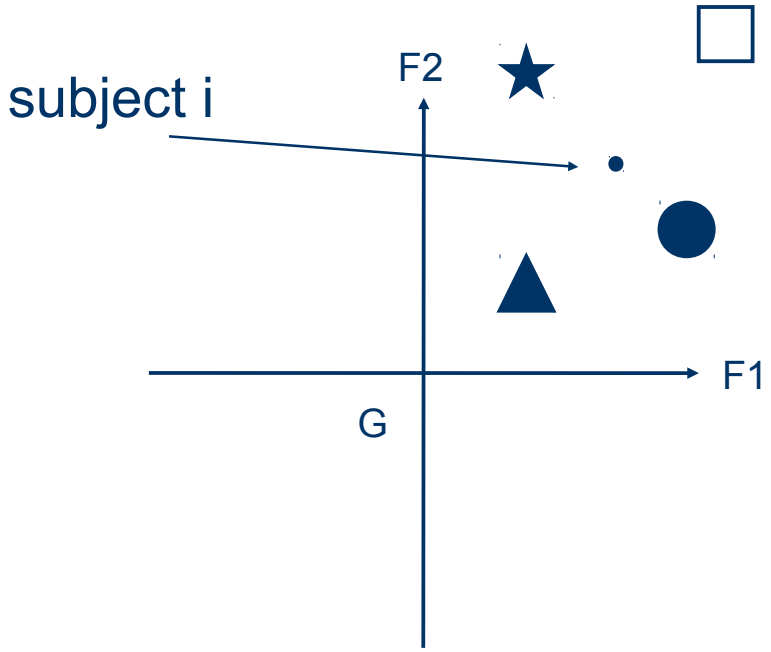
Transition relations for MCA

$$\varphi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \left(\begin{array}{c} \text{Mean of coordinates } \psi_{\alpha i} \text{ for the} \\ \text{subjects who chose modality } j \end{array} \right)$$



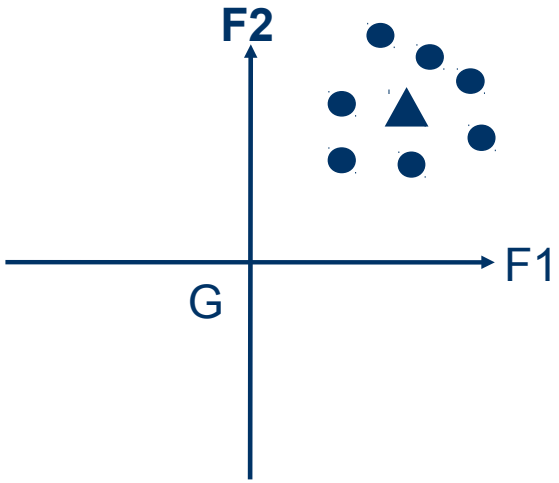
Transition relations for MCA

$$\psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \left(\begin{array}{c} \text{Mean of coordinates } \varphi_{\alpha j} \text{ for the} \\ \text{modalities selected by subject } i \end{array} \right)$$

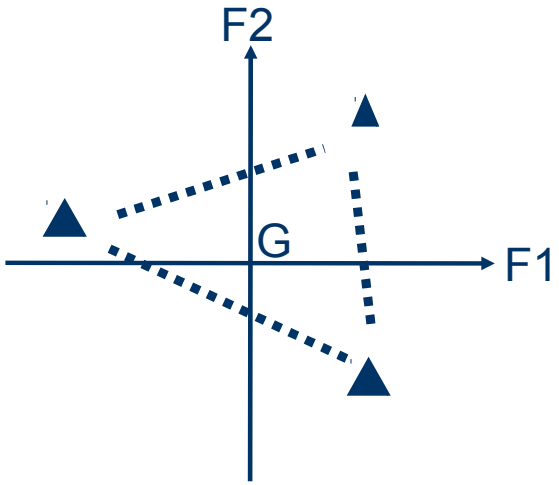


Essential properties

1. A modality is the mean of the subjects that selected it, up to the $\sqrt{\lambda_\alpha}$ coefficient.



2. The weighted barycenter of the modalities of a question is the origin.



Inertia

Matrix X contains only zeroes and ones

$x_{i\bullet} = Q$ number of questions, $x_{\bullet\bullet} = nQ$, $m_i = \frac{1}{n}$, $c_j = \frac{x_{\bullet j}}{nQ}$

Inertia for modality j : $\mathcal{I}_j = \sum_i \frac{c_j}{m_i} \left(\frac{x_{ij}}{x_{\bullet j}} - m_i \right)^2 = \frac{1}{Q} \left(1 - \frac{x_{\bullet j}}{n} \right)$.

The inertia increases when few subjects pick the modality.

Inertia for variable q : $\mathcal{I}_q = \sum_{j \in q} \mathcal{I}_j = \frac{1}{Q} \sum_{j \in q} \left(1 - \frac{x_{\bullet j}}{n} \right) = \frac{p_q - 1}{Q}$,

where p_q is the number of modalities for the question.

The inertia increases when the variable has many modalities.

Total inertia : $\mathcal{I} = \sum_q \mathcal{I}_q = \frac{p - 1}{Q}$.

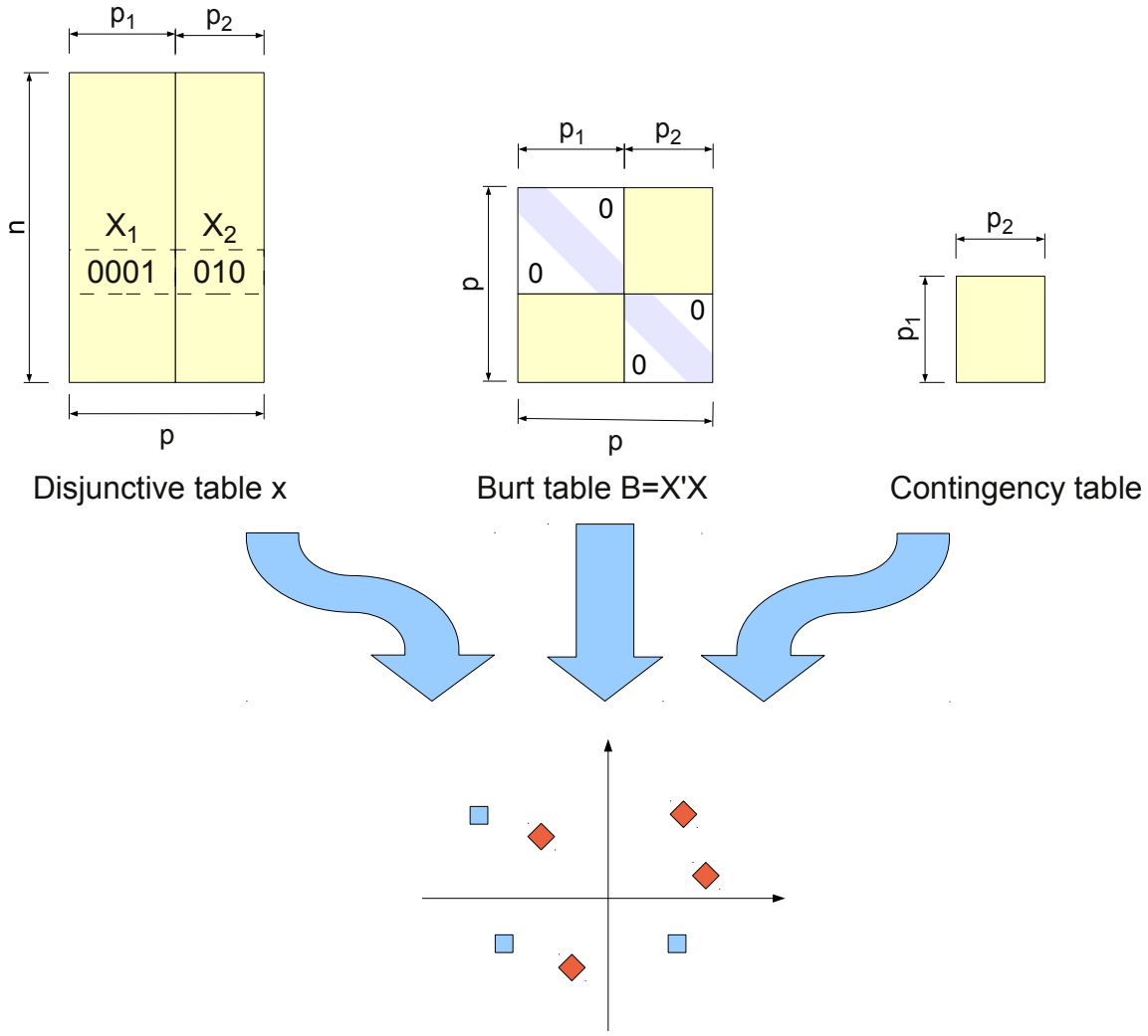
Practical consequences

Tricks that improve the MCA results

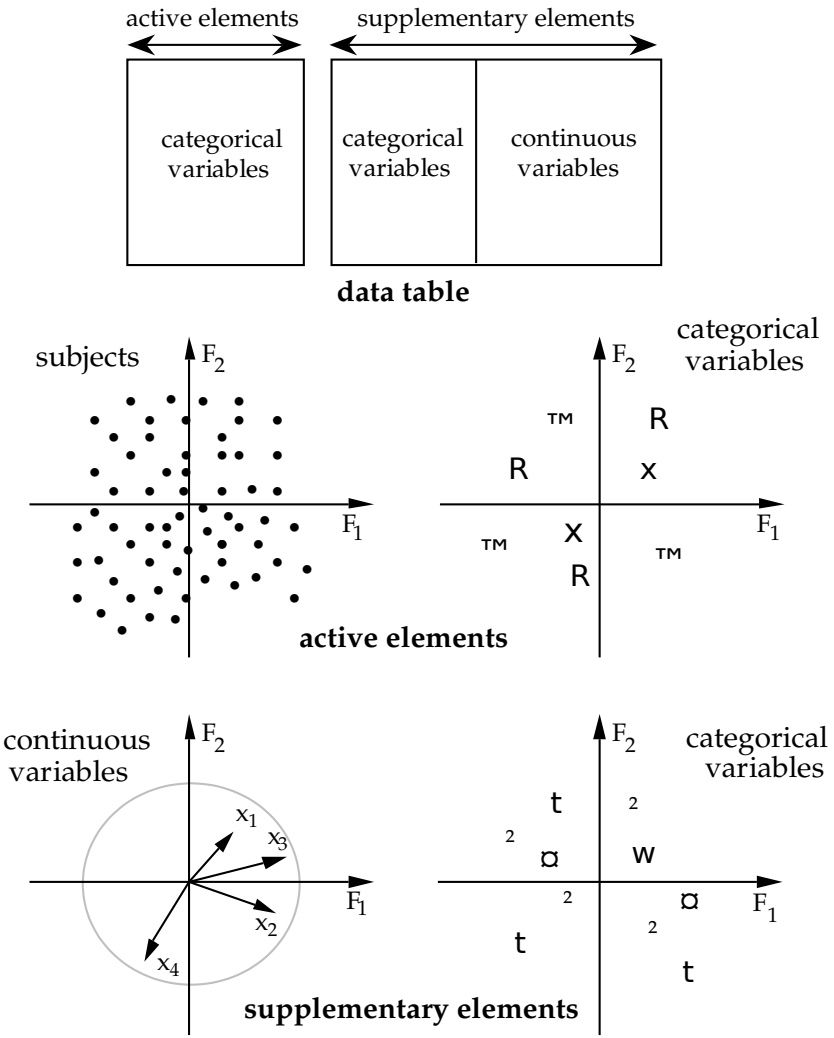
- The inertia of a modality increases when few subjects pick it.
  $\Longrightarrow$ *aggregate rare modalities, or make them supplementaries* .
  $\Longrightarrow$ *bin continuous variables using quantiles*.

- The inertia of a question increases when it has many possible answers.
  $\Longrightarrow$ *balance the number of modalities for all variables*.

Equivalences for the case of two variables



Supplementary elements



IV. Application example: Semiometrie

Semiometry

Introduced by J. F. Steiner in the 70s.

Semiometrie is the use of words to describe lifestyles and values.

Ludovic Lebart

Semiometry

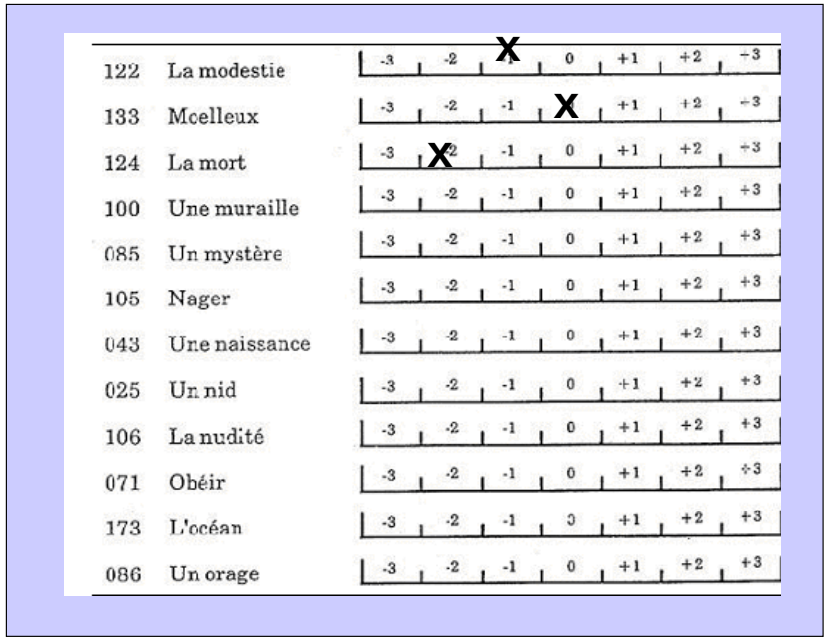
The basic idea is to insert in a marketing questionnaire a series of questions consisting uniquely of words.

Questionnaires in 5 languages

FRENCH	ENGLISH	GERMAN	SPANISH	ITALIAN
l'absolu	absolute	absolut	el absoluto	l'assoluto
l'acharnement	persistence	hartnaeckig	el empeno	l'accanimento
acheter	to buy	kaufen	comprar	comprare
admirer	to admire	bewundern	admirar	ammirare
adorer	to love	anbeten	adorar	adorare
l'ambition	ambition	der ehrgeiz	la ambicion	l'ambizione
l'âme	soul	die seele	el alma	l'anima
l'amitié	friendship	die freundschaft	la amistad	l'amicizia
l'angoisse	anguish	die angst	la angustia	l'angoscia
un animal	animal	ein tier	un animal	un animale
un arbre	tree	ein baum	un arbol	un albero
l'argent	silver	das geld	el dinero	il denaro
une armure	armour	die ruestung	una armadura	un'armatura
l'art	art	die kunst	el arte	l'arte

Semiometry

The subjects must rate these words on a seven levels scale.
– from *most disagreeable or unpleasant*
– to *most agreeable or pleasant*



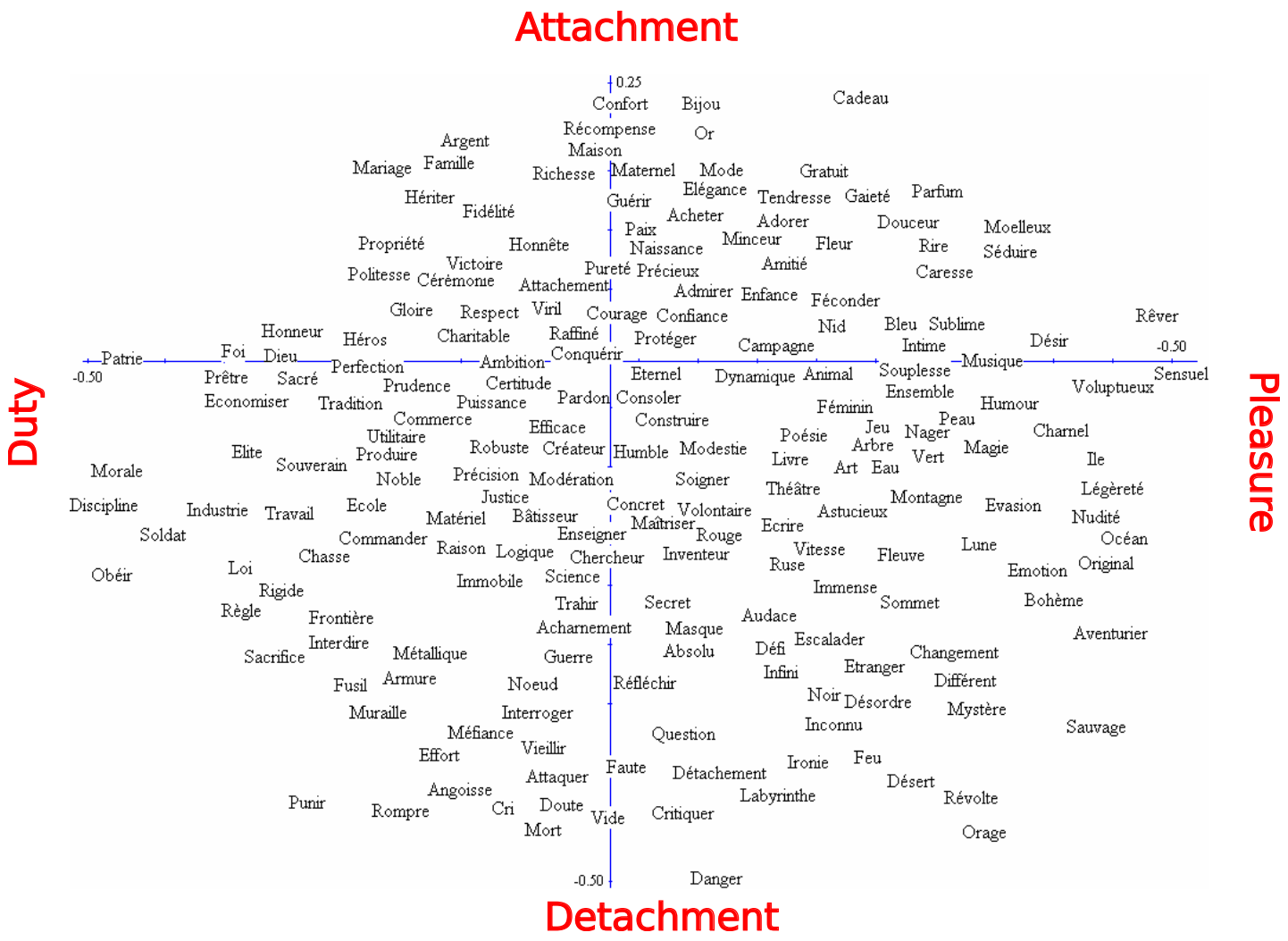
Facsimile of a questionnaire

Principal component analysis

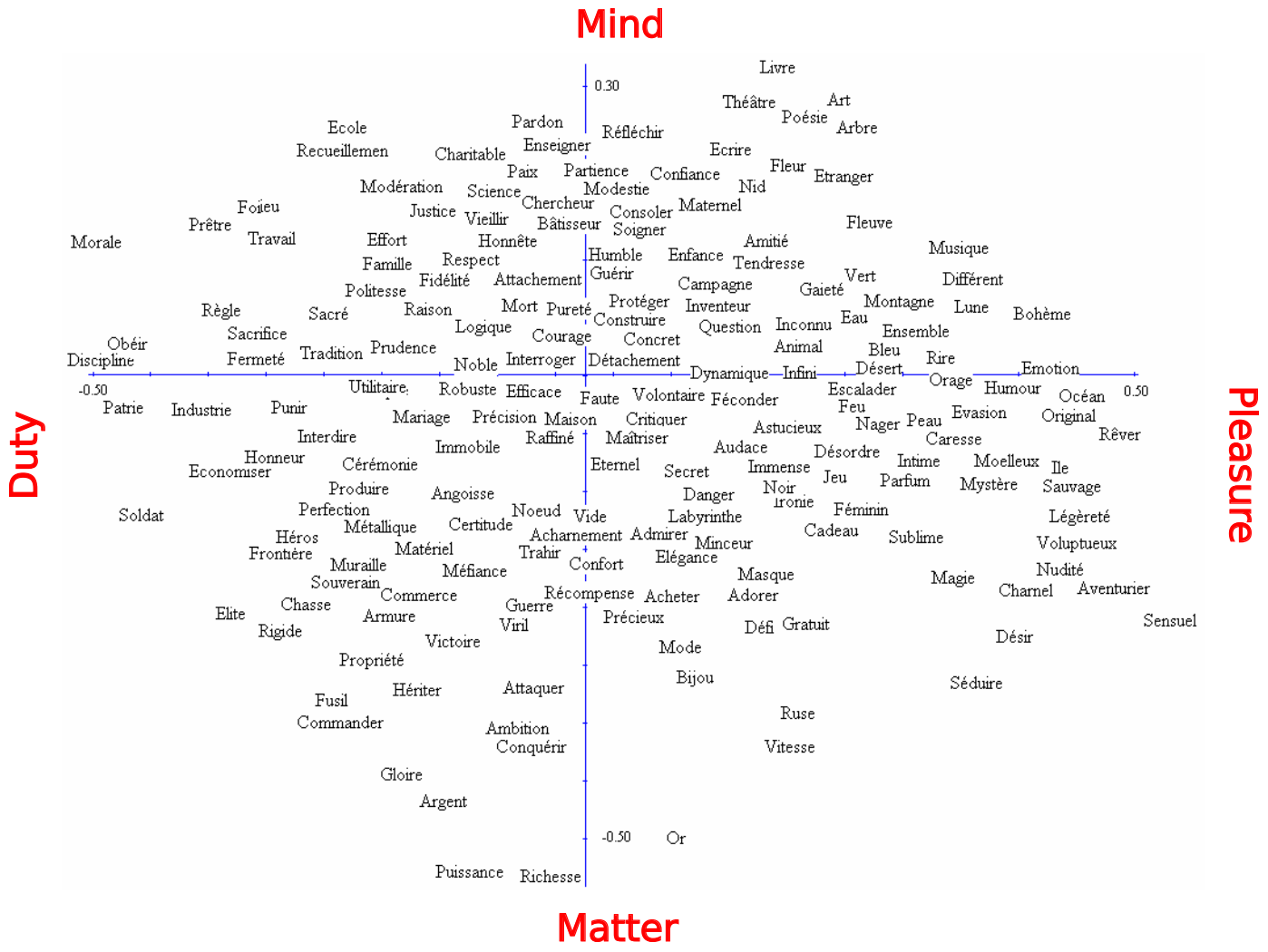
– The first axis is not interesting.
  It just orders words from “bad” to “good”.
  We already know which words are “bad” to “good”.

– The next five axes are highly meaningful.
  They are robust across studies.
  They are robust across languages.
  They are robust across countries.

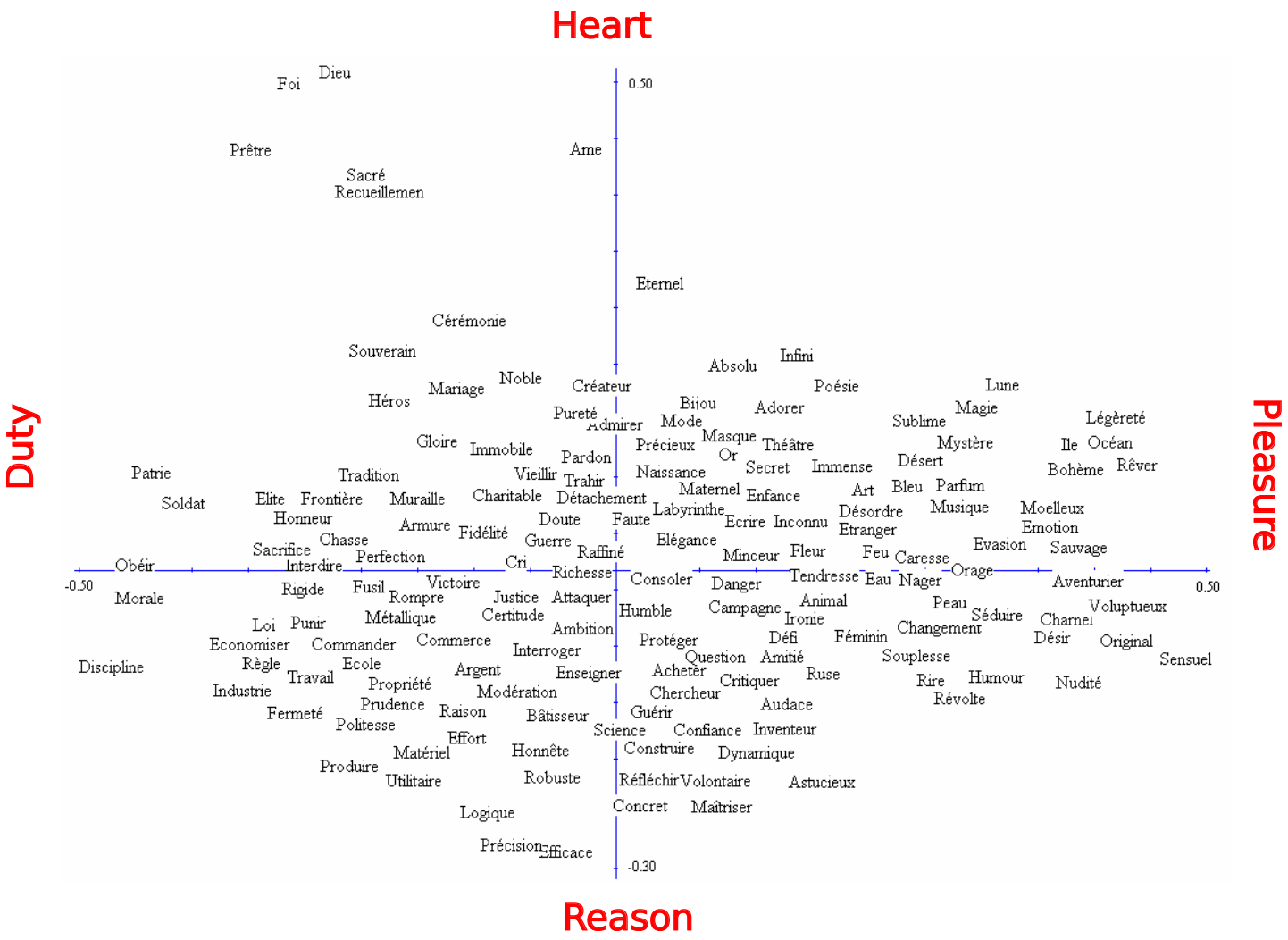
Semiometric plane (2,3)



Semiometric plane (2,4)



Semiometric plane (2,5)



Semiometric plane (2,6)

How is this useful?

Politics

Plot groups of voters as supplementary variables
– those who say they vote for you.
– those who say they are undecided.
– those who say they'll never vote for you.
– those who say they vote for someone else. . .

Determine a target population of voters to convert.
Read which keywords make them tick. . .

Your competitors are also using the same methods.
Your competitors are tracking your moves.
The semiotic space is a political chess board.

How is this useful?

Marketing

Plot groups of customers as supplementary variables
– those who buy your product again and again.
– those who buy your product sometimes.
– those who buy your competitors' products.
– etc. . . .

Determine a target population of customers to convince.
Read how to advertise most effectively. . .

Your competitors are also using the same methods.
Your competitors are tracking your moves.
The semiotic space is a marketing chess board.