Statistics without probability: the French way Multivariate Data Analysis: Duality Diagrams

Susan Holmes

Last Updated October 31st, 2007

Outline

I. Historical Perspective: 1970's.
II. Maths but No Probability
III. Special Geometrical Tricks: visualization.
IV. Perturbation Methods
V. Summary and References

Part I

Introduction: History

Part I

Introduction: History

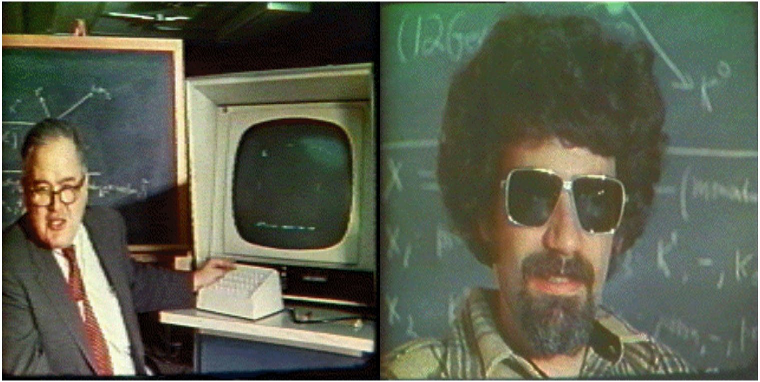
Thirty years ago....

1960's 1970's: Those were the days

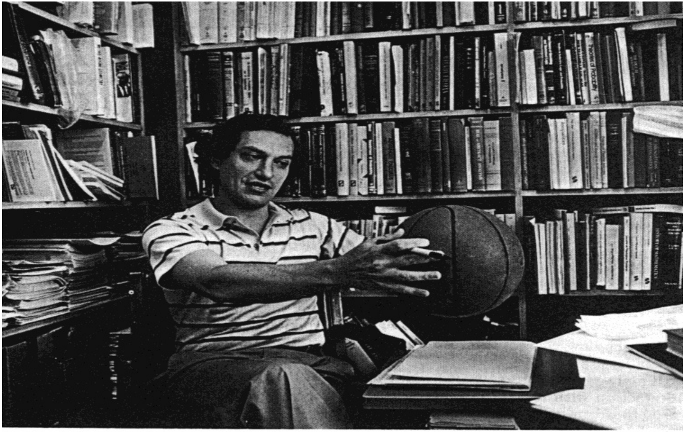
Here..

1960's 1970's: Those were the days

Here..



1960's 1970's: Those were the days



Persi Diaconis.

Here..

1960's 1970's: Those were the days

There

1960's 1970's: Those were the days

There

EDA=English for Analyse des Données

- Resist the mathematicians bid to take over statistics (Tukey).
- Take away the control of the statistics jobs by the probabilists in France.

EDA=English for Analyse des Données

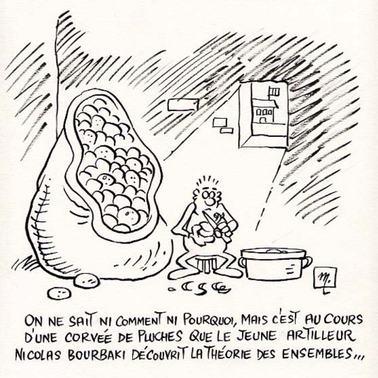
- Resist the mathematicians bid to take over statistics (Tukey).
- Take away the control of the statistics jobs by the probabilists in France. (They failed).

Abstraction Rules

- Bourbaki:

Abstraction Rules

- Bourbaki:



Le général Bourbaki, commandant de la garde impériale, puis de l'armée de l'Est. (Musée de l'armée, Paris).

- An overpowering triplet $(\Omega, \mathcal{A}, \mathcal{P})$

Abstraction Rules

- Bourbaki:

- Creation of a new triplet (X, Q, D) and an algebraic and geometric framework.

Part II

Diagrams and Duality

Geometry of Data Analysis: Rows and Columns

F. CAILLIEZ
Centre Technique Forestier Tropical

J.P. PAGES
Commissariat à l'Energie Atomique

INTRODUCTION

A

L'ANALYSE DES DONNEES

sous la direction de

G. MORLAT

Professeur à l'Institut de Statistique des Universités de Paris

avec des contributions de

J.C. AMIARD - J. ANDRES - M.F. BARA - J.M. BRAUN- J. BRENOT P. CAZES- J.DEHEDIN- B. DIOP- Y.ESCOUFIER- C. GUEGUEN- N. LACOURLY- J.P. MAILLES- B.MARCHADIER- M. PIETRI- E.ROY- G. SAPORTA- F. TESTU- R. THOMAS-

- 1976 -

SOCIETE de MATHEMATIQUES APPLIQUEES et de SCIENCES HUMAINES
9 rue Duban 75016 PARIS

A favorable review (Ramsay and de Leeuw)

Book Review (Psychometrika, 1983)
Quote: "*This remarkable book treats multivariate linear analysis in a manner that is with both distinctive and profoundly promising for future work in this field. With an approach that is strictly algebraic and geometric, it avoids almost all discussion of probabilistic notions, introduces a formalism that transcends matrix algebra, and offers a coherent treatment of topics not often found within the same volume. Finally, it achieves these things while remaining entirely accessible to nonmathematicians and including many excellent practical examples.*"

In summary Introduction à l'Analyse des Données offers a treatment of the multivariate linear model which is (a) metric and basis free, (b) offers a unified survey of both quantitative and certain qualitative procedures, (c) incorporates classical multidimensional scaling in a natural way, and (d) invites through its powerful formalism an extension in a number of valuable directions. We hope it will not be long before an English language counterpart appears.

Geometry of Data Analysis: Rows and Columns

I

PREFACE

"Pendant longtemps, j'ai cru que j'étais un statisticien qui s'intéressait aux inférences allant du particulier au général. Mais, attentif au développement de la statistique mathématique, j'ai trouvé des raisons d'étonnement et de doute".

Ainsi s'exprimait, il y a une dizaine d'années, John W. TUKEY, dans les premières phrases d'un article prophétique publié dans les : Annals of Mathematical Statistics, *sous un titre percutant :* The Future of Data Analysis.

Depuis cette époque, et tout particulièrement au cours des années les plus récentes, on a vu la gent statisticienne se scinder, grosso modo, en deux classes : la première catégorie est celle des statisticiens d'âge moyen qui ont appris et pratiqué la statistique mathématique classique, celle qui prétend formaliser l'induction, à la suite des statisticiens anglo-saxons, notamment, des années 1900 à 1950. La seconde classe est formée de gens en général plus jeunes, qui ont appris sous la même étiquette de "statistique" des techniques bien différentes, s'appuyant sur un outil mathématique purement algébrique, et

Who spent time at the labs in the mid 1960's?

Shepard
Tukey
Mallows
Kruskal

Who spent time at the labs in the mid 1960's?

Shepard
Tukey
Mallows
Kruskal
and Benzecri

A Geometrical Approach

i. The data are p variables measured on n observations.

ii. X with n rows (the observations) and p columns (the variables).

iii. D_n is an $n \times n$ matrix of weights on the "observations", which is most often diagonal.

iv Symmetric definite positive matrix Q,often

A Geometrical Approach

i. The data are p variables measured on n observations.

ii. X with n rows (the observations) and p columns (the variables).

iii. D_n is an $n \times n$ matrix of weights on the "observations", which is most often diagonal.

iv Symmetric definite positive matrix Q,often

$$Q = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & 0 & \dots \\ 0 & \frac{1}{\sigma_2^2} & 0 & 0 & \dots \\ 0 & 0 & \frac{1}{\sigma_3^2} & 0 & \dots \\ \dots & \dots & \dots & 0 & \frac{1}{\sigma_p^2} \end{pmatrix}.$$

Euclidean Spaces

These three matrices form the essential "triplet" $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ defining a multivariate data analysis.
Q and D define geometries or inner products in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively, through

$$
\begin{aligned}
x^t Q y &= < x, y >_Q \qquad && x, y \in \mathbb{R}^p \\
x^t D y &= < x, y >_D \qquad && x, y \in \mathbb{R}^n.
\end{aligned}
$$

An Algebraic Approach

- Q can be seen as a linear function from $\mathbb{R}^p$ to $\mathbb{R}^{p*} = \mathcal{L}(\mathbb{R}^p)$, the space of scalar linear functions on $\mathbb{R}^p$.
- D can be seen as a linear function from $\mathbb{R}^n$ to $\mathbb{R}^{n*} = \mathcal{L}(\mathbb{R}^n)$.
- $$\begin{tikzcd} \mathbb{R}^{p*} \arrow[r, "X"'] \arrow[d, shift left, "V"] & \mathbb{R}^n \arrow[d, shift right, "D"'] \\ \mathbb{R}^p \arrow[u, shift left, "Q"] & \mathbb{R}^{n*} \arrow[l, "X^t"] \arrow[u, shift right, "W"'] \end{tikzcd}$$

An Algebraic Approach

$$\begin{array}{ccc} \mathbb{R}^{p*} & \xrightarrow[X]{} & \mathbb{R}^{n} \\ Q \uparrow \;\; \downarrow V & & D \downarrow \;\; \uparrow W \\ \mathbb{R}^{p} & \xleftarrow[X^t]{} & \mathbb{R}^{n*} \end{array}$$

Duality diagram

i. Eigendecomposition of $X^t DXQ = VQ$

ii. Eigendecomposition of $XQX^t D = WD$

iii.

Notes

(1) Suppose we have data and inner products defined by Q and D :

$$(x, y) \in \mathbb{R}^p \times \mathbb{R}^p \longmapsto x^t Qy \;\; = < x, y >_Q \in \mathbb{R}$$

$$(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \longmapsto x^t Dy \;\; = < x, y >_D \in \mathbb{R}.$$

$$||x||^2_Q = < x, x >_Q = \sum_{j=1}^{p} q_j (x^{\cdot j})^2 \qquad ||x||^2_D = < x, x >_D = \sum_{j=1}^{p} p_i (x_{i\cdot})^2$$

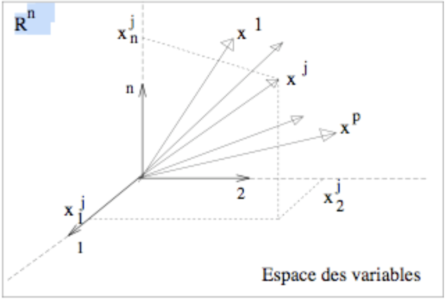
(2) We say an operator O is B-symmetric if $< x, Oy >_B = < Ox, y >_B$, or equivalently $BO = O^t B$.
The **duality diagram** is equivalent to $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ such that X is $n \times p$.
Escoufier (1977) defined as $XQX^t D = WD$ and $X^t DXQ = VQ$ as the characteristic operators of the diagram.

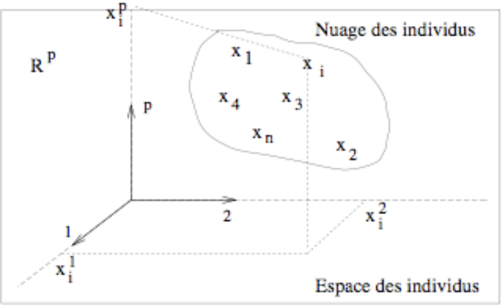
(3) $V = X^t DX$ will be the variance-covariance matrix, if X is centered with regards to D ($X'D\mathbf{1}_n = 0$).

Transposable Data

There is an important symmetry between the rows and columns of X in the diagram, and one can imagine situations where the role of observation or variable is not uniquely defined. For instance in microarray studies the genes can be considered either as variables or observations. This makes sense in many contemporary situations which evade the more classical notion of n observations seen as a random sample of a population. It is certainly not the case that the 30,000 probes are a sample of genes since these probes try to be an exhaustive set.

Two Dual Geometries





Properties of the Diagram

Rank of the diagram: X, X^t, VQ and WD all have the same rank. For Q and D symmetric matrices, VQ and WD are diagonalisable and have the same eigenvalues.

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r \geq 0 \geq \cdots \geq 0.$$

Eigendecomposition of the diagram: VQ is Q symmetric, thus we can find Z such that

$$VQZ = Z\Lambda, Z^t QZ = \mathcal{I}_p, \text{ where } \Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p). \quad (1)$$

Practical Computations

Cholesky decompositions of Q and D, (symmetric and positive definite) $H^tH = Q$ and $K^tK = D$.
Use the singular value decomposition of KXH:

$$KXH = UST^t, \qquad \text{with } T^tT = \mathcal{I}_p, U^tU = \mathcal{I}_n, S \text{ diagonal.}$$

Then $Z = (H^{-1})^tT$ satisfies

$$VQZ = Z\Lambda, Z^tQZ = \mathcal{I}_p$$

with $\Lambda = S^2$.
The renormalized columns of Z, $A = SZ$ are called the principal axes and satisfy:

$$A^tQA = \Lambda.$$

Practical Computations

Similarly, we can define $L = K^{-1}U$ that satisfies

$$WDL = L\Lambda, L^t DL = \mathcal{I}_n, \text{ where } \Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_r, 0, \ldots, 0). \tag{2}$$

$C = LS$ is usually called the matrix of principal components. It is normed so that

$$C^t DC = \Lambda.$$

Transition Formulæ:

Of the four matrices Z, A, L and C we only have to compute one, all others are obtained by the transition formulæ provided by the duality property of the diagram:

$$XQZ = LS = C \qquad X^t DL = ZS = A$$

French Features

Inertia: $Trace(VQ) = Trace(WD)$
(inertia in the sense of Huyghens inertia formula for instance).
Huygens, C. (1657),
De ratiociniis in ludo alea, printed in Exercitationium mathematicaram by F. van Schooten. Elsevirii, Leiden.

$$\sum_{i=1}^{n} p_i d^2(x_i, a)$$

Inertia with regards to a point a of a cloud of p_i-weighted points.
PCA with $Q = \mathcal{I}_p$, $D = \frac{1}{n}\mathcal{I}_n$, and the variables are centered, the inertia is the sum of the variances of all the variables.
If the variables are standardized (Q is the diagonal matrix of inverse variances), then the inertia is the number of variables p.
For correspondence analysis the inertia is the Chi-squared statistic.

Dimension Reduction: Eckart-Young

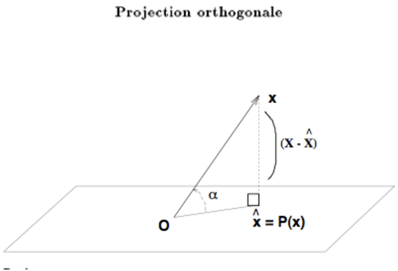
$$X^{[k]} = US^{[k]}V'$$

is the best k rank approximation to X.

Geometric Interpretations of Statistical Quantities

- $$\bar{x} = x' D \mathbf{1}_n \qquad \text{call } \tilde{X} = (\mathbb{I} - 1_n D 1_n') X$$
- $$\hat{\sigma_x}^2 = \sum_i p_i (x_i - \bar{x}) = ||\tilde{x}||_D^2$$
- $$\text{covariance } \operatorname{cov}(x, y) = < \tilde{x}, \tilde{y} >_D$$
- $$\text{correlation } r_{xy} = \frac{< \tilde{x}, \tilde{y} >_D}{||\tilde{x}||_D ||\tilde{y}||_D} = cos(\tilde{x}, \tilde{y})$$

Quality of Representations



- The cosine again, $cos(x, y) = \frac{<x,y>}{||x||||y||}$
- $$cos^2 \alpha = \frac{||\hat{x}||^2}{||x||^2}$$

  tells us how well x is represented by its projection.

Inertia and Contributions

- $$In(X) = ||X||^2 = \sum_i p_i ||x_{i.}||^2 = \sum_j q_j ||x^{.j}||^2 = \sum_{\ell=1}^{p} \lambda_\ell$$
- Contribution of an observation to the total inertia: $\frac{p_i ||x_{i.}||^2}{||X||^2}$
- Contribution of a variable to the total inertia: $\frac{q_j ||x^{.j}||^2}{||X||^2}$

Inertia and Contributions

- Contribution of the kth axis to variable j: $\frac{\lambda_k v_{kj}^2}{||x^j||_D^2}$
- Contribution of variable j to the kth axis $q_j v_{kj}^2$.
- Contribution of the kth axis to observation i: $\frac{\lambda_k u_{ik}^2}{||x_i||_Q^2}$
- Contribution of observation i to the kth axis $p_i u_{ik}^2$.

Comparing Two Diagrams: the RV coefficient

Many problems can be rephrased in terms of comparison of two "duality diagrams" or put more simply, two characterizing operators, built from two "triplets", usually with one of the triplets being a response or having constraints imposed on it. Most often what is done is to compare two such diagrams, and try to get one to match the other in some optimal way.

To compare two symmetric operators, there is either a vector covariance as inner product

$covV(O_1, O_2) = Tr(O_1 O_2) =< O_1, O_2 >$ or a vector correlation [Escoufier, 1977]

$$RV(O_1, O_2) = \frac{Tr(O_1 O_2)}{\sqrt{Tr(O_1^t O_1) tr(O_2^t O_2)}}.$$

If we were to compare the two triplets $\left(X_{n \times 1}, 1, \frac{1}{n} I_n\right)$ and $\left(Y_{n \times 1}, 1, \frac{1}{n} I_n\right)$ we would have $RV = \rho^2$.

PCA: Special case

PCA can be seen as finding the matrix Y which maximizes the *RV* coefficient between characterizing operators, that is, between $(X_{n\times p}, Q, D)$ and $(Y_{n\times q}, I, D)$, under the constraint that Y be of rank $q < p$.

$$RV\left(XQX^tD, YY^tD\right) = \frac{Tr\left(XQX^tDYY^tD\right)}{\sqrt{Tr\left(XQX^tD\right)^2 \, Tr\left(YY^tD\right)^2}}.$$

This maximum is attained where Y is chosen as the first q eigenvectors of XQX^tD normed so that $Y^tDY = \Lambda_q$. The maximum RV is

$$RVmax = \frac{\sum_{i=1}^{q} \lambda_i^2}{\sum_{i=1}^{p} \lambda_i^2}.$$

Of course, classical PCA has $D = \frac{1}{n}\mathcal{I}$, $Q = \mathcal{I}$, but the extra flexibility is often useful. We define the distance between triplets (X, Q, D) and (Z, Q, M) where Z is also $n \times p$, as the distance deduced from the RV inner product between operators XQX^tD and ZMZ^tD.

One Diagram to replace Two Diagrams

Canonical correlation analysis was introduced by Hotelling[Hotelling, 1936] to find the common structure in two sets of variables X_1 and X_2 measured on the same observations. This is equivalent to merging the two matrices columnwise to form a large matrix with n rows and $p_1 + p_2$ columns and taking as the weighting of the variables the matrix defined by the two diagonal blocks $(X_1^t D X_1)^{-1}$ and $(X_2^t D X_2)^{-1}$

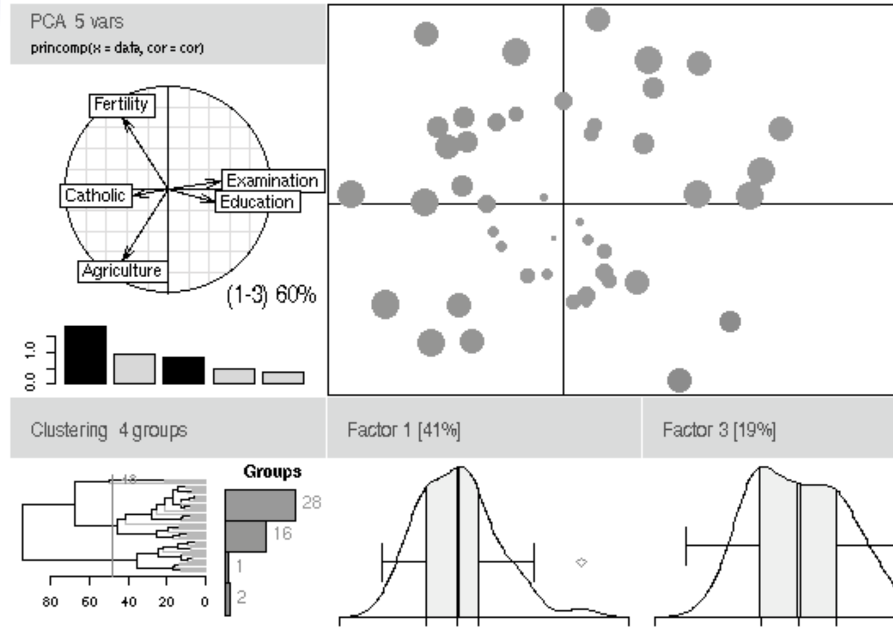
$$Q = \left(\begin{array}{c|c} (X_1^t D X_1)^{-1} & 0 \\ \hline 0 & (X_2^t D X_2)^{-1} \end{array} \right)$$

$$
\begin{tikzcd}
\mathbb{R}^{p_1*} \arrow[r, "X_1"'] \arrow[d, shift left=3, "V_1"] & \mathbb{R}^{n} \arrow[d, shift right=3, "D"'] \\
\mathbb{R}^{p_1} \arrow[u, shift left=3, "\mathcal{I}_{p_1}"] & \mathbb{R}^{n*} \arrow[l, "X_1^t"] \arrow[u, shift right=3, "W_1"']
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathbb{R}^{p_2*} \arrow[r, "X_2"'] \arrow[d, shift left=3, "V_2"] & \mathbb{R}^{n} \arrow[d, shift right=3, "D"'] \\
\mathbb{R}^{p_2} \arrow[u, shift left=3, "\mathcal{I}_{p_2}"] & \mathbb{R}^{n*} \arrow[l, "X_2^t"] \arrow[u, shift right=3, "W_2"']
\end{tikzcd}
$$

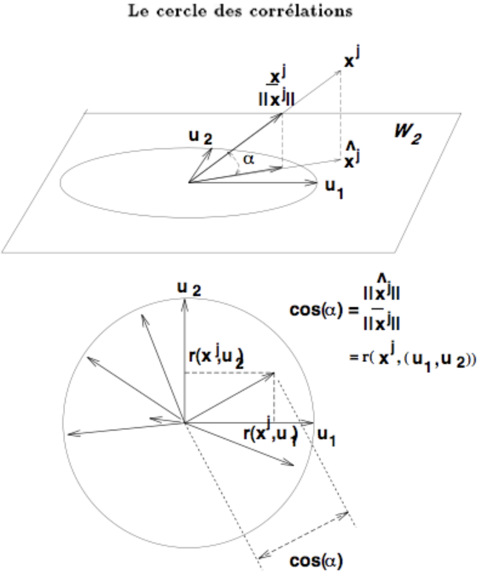
$$
\begin{tikzcd}
\mathbb{R}^{p_1+p_2*} \arrow[r, "{[X_1;X_2]}"'] \arrow[d, shift left=3, "V"] & \mathbb{R}^{n} \arrow[d, shift right=3, "D"'] \\
\mathbb{R}^{p_1+p_2} \arrow[u, shift left=3, "Q"] & \mathbb{R}^{n*} \arrow[l, "{[X_1;X_2]^t}"] \arrow[u, shift right=3, "W"']
\end{tikzcd}
$$

This analysis gives the same eigenvectors as the analysis of the triple
$(X_2^t D X_1, (X_1^t D X_1)^{-1}, (X_2^t D X_2)^{-1})$, also known as the canonical

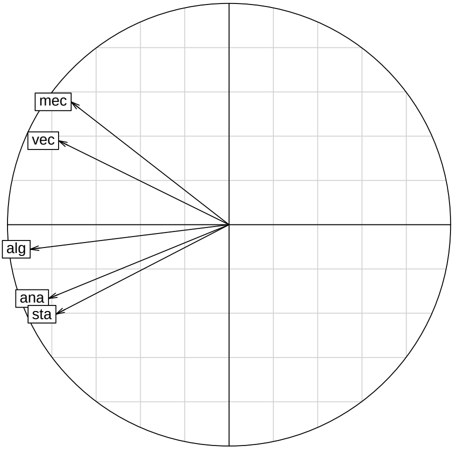
correlation analysis of X_1 and X_2.



Circle of Correlations



Circle of Correlations for Score data



Part IV

Perturbation for Validation and Discovery

Internal and External Methods

- Cross Validation: Leave one variable out, or leave one observation out, or both.
- Bootstrap, bootstrap rows or columns, partial bootstrap, where can we compare them?
- Convex Hulls.
- Procrustes solutions for data cubes (STATIS).

Successful Perturbative Method for Non-hierarchical Clustering

Dynamical Clusters: Edwin Diday, 1970, 1972. [Diday, 1973]

- Repeated k-means with fixed class sizes.
  - Choose a set of k nuclei (usually from the data).
  - Partition the data as the nearest neighbors to each of the k points.
  - For each partition define its centroid.
  - Iterate the above 2 steps until convergence.
- This process gives a set of clusters.
- Organize these clusters according to sets of 'strong forms" the ones that were always together (or mostly) together.

Part V

Conclusions: Data Analysis and Data Integration

Computer Intensive Data Analysis Today

i. Interactive.

ii. Iteration.

iii. Nonparametric.

iv Heterogeneous Data.

v Kernel Methods.

Computer Intensive Data Analysis Today

i. Interactive.
ii. Iteration.
iii. Nonparametric.
iv Heterogeneous Data.
v Kernel Methods.

No more in statistics departments at Universities in France
All in institutes, INRA, INRIA, INSERM,ENSET,ENSAEE, ...à suivre...

Part VI

V. References

Reading

Few references in English explaining the duality/operator point of view.

H. 2006, Multivariate Data Analysis: the French way.[Holmes, 2006]

French: Escoufier [Escoufier, 1987, Escoufier, 1977]. Fréderique Glaçon's PhD thesis [Glaçon, 1981] (in French) on data cubes.

Masters level textbooks on the subject for many details and examples:

- Brigitte Escofier and Jérôme Pagès [Escofier and Pagès, 1998] do not delve into the Duality Diagram
- [Lebart et al., 2000] is one of the broader books on multivariate analyses
- Cailliez and Pagès [Cailliez and Pages., 1976] is hard to find, but was the first textbook completely based on the diagram approach, as was the case in the earlier literature they use transposed matrices.
- Stability studies: [Holmes, 1985],[Lebart et al., 2000].

Software

The methods described in this article are all available in the form of R packages which I recommend.

- `ade4` [Chessel et al., 2004] However, a complete understanding of the duality diagram terminology and philosophy is necessary.
  One of the most important features in all the '`dudi.*`' functions is that when the argument `scannf` is at its default value `TRUE`, the first step imposed on the user is the perusal of the scree plot of eigenvalues.
- `vegan` ecological community

Functions

- Principal Components Analysis (PCA) is available in `prcomp` and `princomp` in the standard package `stats` as `pca` in `vegan` and as `dudi.pca` in `ade4`.
- Two versions of PCAIV are available, one is called Redundancy Analysis (RDA) and is available as `rda` in `vegan` and `pcaiv` in `ade4`.
- Correspondence Analysis (CA) is available in `cca` in `vegan` and as `dudi.coa` in `ade4`.
- Discriminant analysis is available as `lda` in `stats`, as `discrimin` in `ade4`
- Canonical Correlation Analysis is available in `cancor` in `stats` (Beware `cca` in `ade4` is Canonical Correspondence Analysis).
- STATIS (Conjoint analysis of several tables) is available in `ade4`.

Cailliez, F. and Pages., J. P. (1976).
Introduction à l'analyse des donnés.
SMASH, Paris.

Chessel, D., Dufour, A. B., and Thioulouse., J. (2004).
The ade4 package - i: One-table methods.
R News, 4(1):5–10.

Diday, E. (1973).
The dynamic clusters method in nonhierarchical clustering.
International Journal of Computer and Information Sciences, 2(1):62—-88.

Escofier, B. and Pagès, J. (1998).
Analyse factorielles simples et multiples : Objectifs, méthodes et interprétation.
Dunod.

Escoufier, Y. (1977).
Operators related to a data matrix.

In Barra, J. e. a., editor, *Recent developments in Statistics.*, pages 125–131. North Holland,.

Escoufier, Y. (1987).
The duality diagram: a means of better practical applications.
In In Legendre, P. and Legendre, L., editors, *Development. in numerical ecology.*, pages 139–156.

Glaçon, F. (1981).
Analyse conjointe de plusieurs matrices de données. Comparaison de différentes méthodes.
PhD thesis, Grenoble.

Holmes, S. (1985).
Outils Informatiques pour l'Evaluation de la Pertinence d'un Resultat en Analyse des Données.
PhD thesis, Montpellier, USTL.

Holmes, S. (2006).
Multivariate analysis: The french way.
Feschrift for David Freedman, IMS.

Hotelling, H. (1936).
Relations between two sets of variables.
Biometrika, 28:321–327.

Lebart, L., Piron, M., and Morineau, A. (2000).
Statistique exploratoire multidimensionnelle.
Dunod, Paris, France.

Acknowledgements

Chessel, Thioulouse, ADE4 team.
Persi Diaconis.
Elisabeth Purdom.
Yves Escoufier.
NSF-DMS award 02-41246 to SH.