

**Cours d'Analyse Multivariée Approfondie**  
**STA 201**  
**2009-2010**

## **Projet**

**Au choix** : un commentaire d'article ou une étude de logiciel.

1. Cas d'un article

Le travail personnel, de l'ordre de 15 pages dactylographiées, consiste à analyser un article scientifique : il ne s'agit pas de le traduire ni de le paraphraser mais d'en faire un exposé synthétique destiné à un statisticien qui ne l'a pas lu et veut savoir de quoi il retourne. Certains articles peuvent nécessiter une recherche de documents complémentaires. Une application sur des données personnelles des méthodes exposées dans l'article sera grandement appréciée. Lorsque les notations utilisées diffèrent trop de celles du cours, il convient de les reformuler.

2. Cas d'une étude de logiciel

Le travail pourra être fait à deux. Il consistera en une présentation de la méthode et du programme (ou de la procédure analysée) avec un « manuel » simple d'utilisation et une application à des données réelles (autres que les jeux d'essais fournis avec le logiciel).

## **Procédure de choix**

Chaque élève (ou binôme éventuellement pour les logiciels) régulièrement inscrit au cours propose par courrier électronique 3 sujets classés par ordre de préférence. Les premiers arrivés sont les premiers servis avec priorité pour les élèves qui assistent au cours (!). La liste est mise à jour au fur et à mesure : les sujets **surlignés en jaune** ne peuvent plus être choisis. Si les 3 choix sont déjà pris, on recommence.

## **Remise du travail :**

Au plus tard le 13 septembre 2010. Dans le cas d'une note de lecture d'un article, il est **impératif** d'y joindre une copie de l'article analysé. Les rapports seront envoyés par courrier électronique (sous forme d'un fichier **unique** de préférence en PDF) à [gilbert.saporta@cnam.fr](mailto:gilbert.saporta@cnam.fr), ou par courrier postal à G.Saporta : Chaire de Statistique Appliquée, case 441, 292 rue Saint Martin, 75141 Paris cedex 03

## Articles

La plupart des articles sont téléchargeables depuis les postes informatique du CNAM via <http://www.cnam.fr/bibliotheque/>

Pour la Revue de Statistique Appliquée, le site est <http://www.numdam.org/numdam-bin/feuilleter?j=RSA&sl=0>

1. Donald R. Jensen and Donald E. Ramirez: Anomalies in the Foundations of Ridge Regression, International Statistical Review, Volume 76, Issue 1, April 2008, pages 89–105

2. P. CAZES, A. CHOUAKRIA, E. DIDAY, Y. SCHEKTMAN - Extension de l'Analyse en Composantes Principales à des données de type intervalle, Revue de Statistique Appliquée. Vol. XLV, N° 3, 1997 attribué à Belaid Gater

3. F. FERRATY. - Estimation de transformations optimales en ACP curvilinéaire, Revue de Statistique Appliquée, Vol. XLV, N° 1, 1997

4. J. PAGÈS. Eléments de comparaison entre l'Analyse Factorielle Multiple et la méthode STATIS, Revue de Statistique Appliquée, Vol. XLIV, N° 4, 1996 attribué à Abdou Zaki

5. J.PAGÈS , Analyse factorielle de données mixtes, Revue de Statistique Appliquée, 52, n°4, 2004 attribué à HOANG Paul Anh Kiet

6. A.HYVARINEN, E.OJA – Independent component analysis: algorithms and applications, Neural networks, vol 13, 411-430, 2000, [http://www.cis.hut.fi/aapo/papers/IJCNN99\\_tutorialweb/](http://www.cis.hut.fi/aapo/papers/IJCNN99_tutorialweb/)

7. B.SCHOLKOPF, A.SMOLA, K.R. MULLER- Nonlinear Component Analysis as a Kernel Eigenvalue Problem, Neural Computation, 10:1299-1319, 1998. Technical Report No. 44, 1996, Max Planck Institut für biologische Kybernetik, Tübingen. <http://www.kernel-machines.org/papers/nlpca.ps.gz>

8. N. Louw and S.J. Steel: variable selection in kernel Fisher discriminant analysis by means of recursive feature elimination Computational Statistics & Data Analysis, Volume 51, Issue 3, December 2006, Pages 2043-2055

9. Jean-François DURAND :La régression Partial Least-Squares boostée. La revue MODULAD, numéro 38, 2008. <http://www-rocq.inria.fr/axis/modulad/archives/numero-38/Durand-38/durand.pdf> attribué à Saïdé Yekken

10. P.BASTIEN, M.TENENHAUS, V.ESPOSITO-VINZI : PLS Generalized linear regression , Computational Statistics and Data Analysis, 48, 17-46, (2005) attribué à Dalice FERET ANVELET

11. Abba M. KRIEGER, Paul E. GREEN - A Generalized Rand-Index Method for Consensus Clustering of Separate Partitions of the Same Data Base, J. of Classification 16:63-89 (1999).

12. A.D. GORDON, M. VICHI - Partitions of Partitions, J. of Classification 15:265-285 (1998)

13. G. DIANA, C. TOMMASI – Cross validation methods in principal component analysis: a comparison., *Statistical Methods and Applications*, vol 11, n°1, 71-82 (2002) attribué à Aurélia Petit

14. A. BEN-HUR, D. HORN, H. T. SIEGELMANN, V. VAPNIK, Support Vector Clustering, *Journal of Machine Learning Research*, 2:125-137, 2001.

<http://www.ai.mit.edu/projects/jmlr/papers/volume2/horn01a/rev1/abstract.html>

15. Sugnet Gardner and Niël J. le Roux, Extensions of Biplot Methodology to Discriminant Analysis, *Journal of Classification*, 22, n°1, pp 59 – 86, (2005)

16. Jonathan N. Crook, David B. Edelman, Lyn C. Thomas: Recent developments in consumer credit risk assessment, *European Journal of Operational Research* 183 (2007) 1447–1465 attribué à Didier FRA

17. Van Gesteel, and al. A support vector machine approach to credit scoring, <http://www.geocities.com/joaogarcia18/BANKFINVer4.pdf> attribué à Pascaline Wolasse

18. Thaddeus Tarpey, Kimberly K. J. Kinader : Clustering Functional Data , *Journal of Classification*, 20, n°1 pp. 93 - 114 (2003)

19. Eric J. Beh and Luigi D'Ambra : Some Interpretative Tools for Non-Symmetrical Correspondence Analysis, *Journal of Classification*, Volume 26, Number 1 / avril 2009, 55-76

20. Joost van Rosmalen, Patrick J. F. Groenen, Javier Trejos and William Castillo: Optimization Strategies for Two-Mode Partitioning, *Journal of Classification*, Volume 26, Number 2 / août 2009, 155-181

21. Roman Rosipal, Leonard J. Trejo: Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space , *JMLR Special Issue on Kernel Methods*, 2001 <http://www.jmlr.org/papers/v2/rosipal01a.html> attribué à Serge AKAI

22. Donatella Vicari and Maurizio Vichi : Structural Classification Analysis of Three-Way Dissimilarity Data, *Journal of Classification*, Volume 26, Number 2 / août 2009, 121-154

23. A. Buja, D. Swayne: Visualization Methodology for Multidimensional Scaling, *Journal of Classification*, Volume 19, Number 1, 2002, Pages: 7 – 43

24. S. Van Huffel, Total least squares <ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/markovsky/reports/05-34.ps.gz>

25. Michel van de Velden, Optimal Scaling of Paired Comparison Data , *Journal of Classification*, 21, n°1 pp89 – 109, (2004)

26. Jieping Ye, Characterization of a Family of Algorithms for Generalized Discriminant Analysis on Undersampled Problems , *Journal of Machine Learning Research*, 6:483--502, 2005. , <http://jmlr.csail.mit.edu/papers/volume6/ye05a/ye05a.pdf>

27. Ana M. Aguilera, Manuel Escabias and Mariano J. Valderrama ,

Using principal components for estimating logistic regression with high-dimensional multicollinear data, Computational Statistics and Data Analysis, Volume 50, Issue 8, Pages 1905-1924, (2006) attribué à Vincent Touzé

28. S. Bougeard, M. Hanafi and E.M. Qannari: Continuum redundancy-PLS regression: A simple continuum approach, Computational Statistics and Data Analysis, volume 52, n°7, (2008) pages 3686-3696

29. Raffaella Piccarreta : Classification trees for ordinal variables, Computational Statistics, Volume 23, Number 3 (2008) pages 407-427

30. Joaquín Pacheco, Silvia Casado, Laura Núñez and Olga Gómez  
Analysis of new variable selection methods for discriminant analysis • Computational Statistics and Data Analysis, Volume 51, n°3, Pages 1463-1478

31. António Pedro Duarte Silva and Paula Brito , Linear discriminant analysis for interval data, Computational Statistics, Volume 21, Number 2 / June, 2006 , 289-308

32. Govaert, G., Classification binaire et modèles. Revue de Statistique Appliquée, 38 no. 1 (1990), p. 67-81

33. Marcotorchino, F.; Michaud, P. Agrégation de similarités en classification automatique. Revue de Statistique Appliquée, 30 no. 2 (1982), p. 21-44

34. Celeux, G. Classification et modèles. Revue de Statistique Appliquée, 36 no. 4 (1988), p. 43-57

35. M. Escabias, A.M. Aguilera, M.J. Valderrama :Functional PLS logit regression model. Computational Statistics & Data Analysis, Volume 51, Issue 10, 15 June 2007, Pages 4891-4902

36. Michael J. Brusco, Renu Singh and Douglas Steinley : Variable Neighborhood Search Heuristics for Selecting a Subset of Variables in Principal Component Analysis, Psychometrika, Volume 74, Number 4 / décembre 2009, 705-726

37. Heungsun Hwang:Regularized Generalized Structured Component Analysis, Psychometrika, Volume 74, Number 3, 517-530

38. Ke-Hai Yuan: Identifying Variables Responsible for Data not Missing at Random; Psychometrika, Volume 74, Number 2, 233-256

39. Douglas Steinley and Michael J. Brusco: Selection of Variables in Cluster Analysis: An Empirical Comparison of Eight Procedures, Psychometrika, Volume 73, Number 1 / mars 2008, 125-144 attribué à Violaine Guérault

## Logiciels

1. Procédure MODECLUS de SAS attribué à Vivien El Jalti
2. Procédure ACECLUS de SAS attribué à Vincent Gigliobianco et Serge Nakache
3. Procédure MDS de SAS attribué à MELLAK KARIMA et STUHLIK MARION
4. Procédure VARCLUS de SAS attribué à Eric LAZARO et YIN Xuehua
5. Procédure PLS de SAS attribué à Frederik Cassor
6. Procédure GAM de SAS attribué à Olivier Moreau
7. Procédure FACTOR de SAS attribué à Thierno Ousmane BA et Aurelien MOUENET
8. Procédure LOESS de SAS attribué à Yannis Manoël GHOZZI et Brahim Mizmami
9. Procédure KDE de SAS attribué à Aïmad Brahim
10. Procédures PHREG et TPHREG de SAS attribué à Cécile Welter-Nicol et Elodie Lalande
11. Procédures MI et MIANALYZE de SAS attribué à Emmanuelle BEGUINET Karine BLACHON
12. Procédure ROBUSTREG de SAS attribué à Guillaume Groleau
13. Procédure QUANTREG de SAS attribué à LEBAMBA-MOUSSA OUSMANE ET OBOYO EMIL
14. Procédure PRINQUAL de SAS attribué à Nafessa MFOIHAYA
15. Procédure TRANSREG de SAS attribué à De Sousa Cardoso Cyril et Josnin Rémi

*La documentation de SAS V9 est accessible à l'adresse:*

[http://support.sas.com/documentation/cdl\\_main/index.html](http://support.sas.com/documentation/cdl_main/index.html)

16. Arbres et graphes de décision: SIPINA et TANAGRA <http://eric.univ-lyon2.fr/~ricco/sipina.html> <http://eric.univ-lyon2.fr/~ricco/tanagra> attribué à Klara Vinceneux et Elodie Kranklader
17. Classes latentes : logiciel LatentGold version de démo gratuite téléchargeable à l'adresse : <http://www.statisticalinnovations.com/> attribué à Mehdi taleb et Emilie Vicente
18. Classes latentes : PROC LCA développée à Penn State University : <http://methodology.psu.edu/> attribué à Youssef Benzakour et Davy Mendoume