ACTUALITÉ DE L'ANALYSE DES DONNÉES

Gilbert Saporta

Conservatoire National des Arts et Métiers, Paris

Au-delà de techniques particulières, l'analyse des données désigne un mouvement de pensée qui s'est développé de par le monde depuis les années 1960 en réaction aux abus du formalisme de la statistique mathématique, avec pour objectif de revenir aux fondamentaux de la statistique : décrire, comprendre, prévoir.

Parmi ses pères fondateurs, J.W.Tukey (1962) et J.P.Benzécri (1972) eurent une grande influence. Ils étaient très critiques avec ce qui était à l'époque la pensée dominante comme le montrent ces deux citations :

- « *He (Tukey) seems to identify statistics with the grotesque phenomenon generally known as mathematical statistics and find it necessary to replace statistics by data analysis* » (Anscombe, 1967).
- « *Statistique n'est pas probabilité. Sous le nom de statistique mathématique des auteurs (…) ont édifié une pompeuse discipline riche en hypothèses qui ne sont jamais satisfaites* » (Benzécri, 1972)

L'ambition de J.P.Benzécri allait d'ailleurs bien au-delà de la simple statistique quand il écrivait : « *L'analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature.* »

Des mouvements semblables sont nés au Japon avec C.Hayashi et ses élèves, aux Pays-Bas autour de J.de Leeuw. En Italie, l'analyse des données a été brillamment représentée par Alfredo Rizzi à qui cet article est dédié, et C.Lauro. Alfredo Rizzi s'est illustré tout particulièremenr par de nombreuses contributions à la classification. On se reportera à Rizzi (2008) pour une recension des travaux de l'école italienne.

L'analyse des données a rapidement montré son efficacité auprès d'utilisateurs de domaines très variés, confrontés au traitement de données que l'on qualifiait à l'époque de nombreuses, ce qui fait sourire maintenant.

Dans les pages qui suivent, je tenterai de montrer sans prétendre à l'exhaustivité, comment les méthodes de l'analyse des données ont évolué au cours de différentes époques qui se recouvrent, et continuent à être d'une puissance inégalée dans notre ère actuelle de données massives, toujours sans recourir à des hypothèses restrictives sur la distribution des observations.

1. Le temps des synthèses

Pour l'essentiel les méthodes d'analyse des données sont des méthodes de réduction de dimension : les méthodes de classification non supervisée ou *clustering* opèrent sur le nombre d'unités statistiques, tandis que les méthodes factorielles réduisent le nombre de variables en recherchant des combinaisons linéaires associées à de nouveaux axes de l'espace des individus.

On s'est rapidement aperçu que toutes les méthodes recherchant les valeurs et vecteurs propres de matrices liées à la dispersion d'un nuage (totale ou intra) ou de corrélation pouvaient s'exprimer comme cas particuliers de certaines techniques. Ainsi les analyses de correspondances (simple et multiple), l'analyse factorielle discriminante appelée en anglais *canonical discriminant analysis*, sont des analyses en composantes principales particulières. Il suffit pour cela d'étendre l'ACP classique en pondérant les unités et en introduisant des métriques. Le schéma de dualité introduit par Cailliez et Pagès (1976) est une manière abstraite de représenter les relations entre tableaux, matrices et espaces associés. Un article récent de De la Cruz, O. & Holmes, S.P. (2011) l'a remis en lumière.

Une autre synthèse est issue de la généralisation à plusieurs groupes de variables de l'analyse canonique proposée par J.D.Carroll (1968). Etant donnés p blocs de variables $\mathbf{X}_j$ on recherche des composantes $\mathbf{z}$ maximisant le critère suivant : $\sum_{j=1}^{p} R^2\left(\mathbf{z}, \mathbf{X}_j\right)$

On montre que toutes les méthodes factorielles, ainsi que la régression multiple sont des analyses canoniques particulières. Ce point de vue a été exploité par Bouroche & Saporta (1983).

L'approche PLS des modèles à équations structurelles reliant des blocs de variables fournit également un cadre global pour de nombreuses méthodes linéaires ainsi que l'a montré M.Tenenhaus (1999).

L'école italienne développa très tôt des variantes non-symétriques de l'analyse des correspondances, en s'inspirant de l'analyse des redondances (Lauro & d'Ambra, 1984).

2. Le temps des méthodes typologiques ou *clusterwise*

La recherche de partitions en k classes d'un ensemble d'unités appartenant à un espace euclidien s'effectue le plus souvent à l'aide de l'algorithme des k-means : cette méthode converge très vite, même pour de vastes ensembles de données, mais pas forcément vers l'optimum global. Sous le nom de nuées dynamiques, E.Diday (1971) en a proposé de multiples extensions, où par exemple les représentants des classes peuvent être des groupes de points, des variétés etc.

La recherche simultanée de k classes et de modèles locaux en alternant k-means et modélisation est une manière géométrique et non probabiliste d'aborder les problèmes de mélange. La régression clusterwise en est le cas le plus connu : dans chaque classe on ajuste un modèle de régression et l'affectation aux classes se fait selon le meilleur modèle. Les méthodes clusterwise permettent de tenir compte d'une hétérogénéité non directement observable et sont particulièrement utiles pour les grands ensembles de données où la pertinence d'un modèle simple et global est contestable.

Dans les années 1970, E .Diday et ses collaborateurs développèrent des approches « typologiques » pour la plupart des techniques linéaires : ACP, régression (Charles, 1977) , discrimination.

Ces méthodes font à nouveau l'objet de nombreuses publications, en association avec des données fonctionnelles (Preda & Saporta, 2005) , des données symboliques (de Carvalho et al. , 2010), dans des cas multiblocs (De Roover et al., 2012).

3. Le temps des extensions à de nouveaux types de données

3.1 Les données fonctionnelles

Jean-Claude Deville (1974) montra que la décomposition de Karhunen-Loève n'était autre que l'ACP des trajectoires d'un processus, ouvrant ainsi la voie à l'analyse des données fonctionnelles (Ramsay & Silverman (1997). Le nombre de variables étant infini non dénombrable, on étend la notion de combinaison linéaire pour définir une composante principale, à une intégrale du type $\xi = \int_0^T f(t) X_t dt$, le facteur $f(t)$ est alors fonction propre de l'opérateur de covariance $\int_0^T C(t,s) f(s) ds = \lambda f(t)$.

Deville & Saporta (1980) étendirent ensuite cette méthode à l'analyse des correspondances. La réduction de dimension offerte par l'ACP permet alors de donner des solutions au problème de la régression sur des trajectoires, problème mal posé puisque le nombre d'observations est inférieur au nombre infini de variables. La régression PLS est cependant mieux adaptée dans ce dernier cas et permet de traiter des problèmes de classification supervisée (Costanzo & al. , 2006)

3.2 Les données symboliques

E.Diday est à l'origine de nombreux travaux qui ont permis d'étendre l'ensemble des méthodes d'analyse des données à de nouveaux types de données, dites symboliques. C'est le cas par exemple lorsque la cellule i,j d'un tableau de données n'est plus un nombre, mais un intervalle ou une distribution. Voici un exemple de tableau de données symboliques extrait de Billard & Diday (2006) :

w_u	Court Type	Player Weight	Player Height	Racket Tension
w_1	Hard	[65, 86]	[1.78, 1.93]	[14, 99]
w_2	Grass	[65, 83]	[1.80, 1.91]	[26, 99]
w_3	Indoor	[65, 87]	[1.75, 1.93]	[14, 99]
w_4	Clay	[68, 84]	[1.75, 1.93]	[24, 99]

On trouvera dans cet ouvrage une présentation des principaux résultats obtenus au cours des programmes européens SODAS et ASSO pendant une dizaine d'années. Il faut noter une importante participation des statisticiens italiens à ces travaux novateurs.

4. L'analyse des données non-linéaire

4.1 l'ACP semi-linéaire

Dauxois et Pousse (1976) étendirent l'analyse en composantes principales et l'analyse canonique à des espaces hilbertiens. En simplifiant leur approche, au lieu de chercher comme en ACP des combinaisons linéaires de variance maximale $\arg\max V\left(\sum_{j=1}^{p} a_j x^j\right)$ sous contraintes $\|\mathbf{a}\| = 1$, on recherche des transformations non-linéaires Φ_j de carré intégrable de chacune des variables et le

critère devient $\arg\max V\left(\sum_{j=1}^{p}\Phi_j\left(x^j\right)\right)$ ou de manière équivalente, les Φ_j étant définies à une constante multiplicatives près, maximisant la somme des carrés de corrélations linéaires entre la composante principale *c* et les transformées des variables $\sum_{j=1}^{p}\rho^2\left(c,\Phi_j\left(x^j\right)\right)$.

Ce problème n'est bien posé qu'en dimension infinie, c'est-à-dire pour des variables *c* et x^j aléatoires. En dimension *n* finie, il faut restreindre les espaces de transformations Φ_j pour qu'ils soient de dimension finie. Un choix classique est alors celui de transformations splines, cf Besse (1988).

La recherche de transformations a fait l'objet de travaux de l'école néerlandaise synthétisé dans l'ouvrage publié sous le nom collectif de Gifi (1999).

4.2 La « kernelisation » de l'analyse des données

Dans la lignée des travaux de V. Vapnik, Schölkopf et al (1998) ont défini une ACP non-linéaire de la manière suivante où on transforme tout le vecteur $\mathbf{x}=(x^1,x^2,\ldots x^p)$. Chaque point de l'espace des individus E est envoyé dans un espace $\Phi(E)$ appelé espace étendu ou *feature space* muni d'un produit scalaire. La dimension de $\Phi(E)$ peut être très grande et la notion de variable se perd. On effectue alors une analyse de type *multidimensional scaling* métrique sur les points transformés selon la méthode de Torgerson qui est équivalente à l'ACP dans $\Phi(E)$. Tout repose sur le choix du produit scalaire dans $\Phi(E)$: si on prend un produit scalaire qui s'exprime aisément en fonction du produit scalaire de E , il n'est plus nécessaire de connaître la transformation Φ qui est alors implicite. Tous les calculs s'effectuent en dimension *n* . C'est le « *kernel trick* ».

Soit $k(\mathbf{x},\mathbf{y})$ un produit scalaire dans $\Phi(E)$ et $<\mathbf{x},\mathbf{y}>$ celui de E . Les choix suivants sont couramment utilisés (on vérifie les conditions de Mercer pour que la matrice symétrique de terme $k(\mathbf{x},\mathbf{y})$ aient ses valeurs propres non-négatives):

$k(\mathbf{x},\mathbf{y})=\left(<\mathbf{x},\mathbf{y}>+c\right)^d$ noyau polynomial

$k(\mathbf{x},\mathbf{y})=\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$ noyau gaussien ou RBF

$k(\mathbf{x},\mathbf{y})=\tanh\left(<\mathbf{x},\mathbf{y}>+c\right)$ noyau logistique ou sigmoïde

Il suffit alors de remplacer la matrice de Torgerson usuelle **W** par celle où chaque terme est $k(\mathbf{x},\mathbf{y})$, de la centrer en lignes et colonnes et d'en extraire les vecteurs propres pour obtenir les composantes principales dans $\Phi(E)$.

Une fois définie la kernel-PCA, de nombreux travaux suivirent « kernelisant » différentes méthodes, citons : l'analyse discriminante de Fisher par Baudat & Anouar (2000) trouvée indépendemment sous le nom de LS-SVM par Suykens & Vandewalle (1999) , la régression PLS de Rosipal & Trejo (2001), la classification non-supervisée avec les kernel k-means déjà proposée par Schölkopf et al., l'analyse canonique (Fyfe & Lai., 2001). Il est intéressant de noter que la plupart de ces développements proviennent non pas de statisticiens mais de chercheurs en Intelligence Artificielle ou Machine Learning.

5. Le temps des méthodes sparse

Lorsque le nombre de dimensions (ou variables) est très grand, l'ACP, L'ACM et les autres méthodes factorielles conduisent à des résultats difficilement interprétables : comment en effet donner un sens à une combinaison linéaire de plusieurs centaines de variables ? La recherche de combinaisons dites « sparse » limitées à un petit nombre de variables, c'est-à-dire avec un grand nombre de coefficients nuls fait l'objet de l'attention des chercheurs depuis une quinzaine d'années. Les premières tentatives imposant par exemple que les coefficients soient égaux à -1, 0 ou 1 conduisaient à des algorithmes non convexes difficiles d'emploi.

La transposition à l'ACP de la régression LASSO de Tibshirani (1996) a permis de donner des solutions exactes et élégantes. Rappelons que le LASSO consiste à effectuer une régression avec une pénalisation L^1 sur les coefficients, ce qui permet de gérer facilement la multicolinéarité et la grande dimension.

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

Il suffit alors (Zou et al. (2006)) de modifier un critère de l'ACP d'un tableau $\mathbf{X}$ et de composantes principales $\mathbf{z}$:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$$

La première contrainte en norme L^2 ne fait qu'indiquer que le vecteur « propre » doit être de norme 1 ; c'est la deuxième contrainte en norme L^1 qui provoque la sparsité quand le multiplicateur de Lagrange λ_1 varie. Numériquement la solution s'obtient en alternant SVD ($\boldsymbol{\beta}$ fixé, pour trouver $\mathbf{z}$) et elastic-net pour trouver $\boldsymbol{\beta}$ à $\mathbf{z}$ fixé, jusqu'à convergence.

Les positions des coefficients nuls ne sont pas les mêmes pour les différentes composantes. La sélection des variables se fait donc dimension par dimension. Si l'interprétabilité augmente, la contrepartie est la perte de propriétés caractéristiques de l'ACP comme l'orthogonalité des composantes principales et (ou) des facteurs principaux.

Depuis lors, des variantes sparse de nombreuses méthodes ont été développées, comme la sparse-PLS de Chun & Keleş (2009) , la sparse discriminant analysis de Clemmensen et al. (2008), la sparse canonical analysis par Hardoon & Shawe-Taylor (2009) et simultanément par Witten et al. (2009) et l'analyse des correspondances multiples sparse (Bernard et al, 2012).

On peut voir les méthodes sparses comme des cas particuliers de régularisation avec des contraintes L^1. On trouvera dans Tenenhaus & Tenenhaus (2011) un cadre général pour la régularisation de différentes méthodes qui aurait pu figurer également dans le § 1 « Le temps des synthèses ».

Conclusion

L'analyse des données reste un domaine très vivant de la statistique qui ne se limite plus à la trilogie analyse en composantes principales, analyse des correspondances et classification. On aurait pu également évoquer les méthodes multiway, multiblocs, les analyses procustéennes etc. Les

chercheurs développent des algorithmes qui se parallélisent pour traiter les données massives. Remarquons à ce propos que la décomposition en valeurs singulières et la méthode des k-means bien programmées s'adaptent au contexte Big Data. La mise à disposition systématique de packages R assure une large diffusion aux nouveautés théoriques. Au-delà des techniques particulières, l'esprit des pionniers de l'analyse des données est toujours présent : « les données doivent précéder les modèles et non l'inverse » écrivait J.P.Benzécri ce que l'on retrouve avec l'expression « data driven ». C'est d'autant plus nécessaire que les données massives demandent d'être analysées sans *a priori*.

Références

Anscombe F. J. (1967): Topics in the investigation of linear relations. *J. Roy. Stat. Soc*., vol. B 29, 1 – 52

Baudat, G.; Anouar, F. (2000): Generalized discriminant analysis using a kernel approach. *Neural Computation* 12 (10): 2385–2404

Benzécri, J.P. (1972) : *l'Analyse des Données*, tome 2, Dunod, Paris

Bernard, A. , Guinot, C. , Saporta, G. (2012) Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis, *Proceedings of Compstat 2012*, 99-106

Besse, Ph. (1988): Spline functions and optimal metric in linear principal components analysis. *In Components and Correspondence Analysis* (Van Rijkevorsel et al, Eds.). Wiley, London.

Billard, L., Diday, E. (2012): *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons.

Bouroche, J.M., Saporta G. (1983) : *L'analisi dei dati,* CLU, Napoli

Cailliez, F., Pagès, J.P. (1976) : *Introduction à l'analyse des données*, Smash, Paris

Carroll, J.D. (1968): Generalization of canonical analysis to three or more sets of variables, *Proc. Am.Psych.Assoc.*, 227-228

Charles, C., (1977) : Régression typologique et reconnaissance des formes. Ph.D., Université Paris IX

Chiandotto, B., Rizzi, A. (2008) : Recent contributions of Italian statisticians to data analysis, *Metron,* 53-

Chun, H. , Keleş, S. (2010) : Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B ,* 72, 3-25

Clemmensen, L., Hastie, T. and Ersboell, K. (2008) Sparse discriminant analysis, Technical report, IMM, Technical University of Denmark

Costanzo, D., Preda, C., Saporta, G. (2006): Anticipated prediction in discriminant analysis on functional data for binary response, *in COMPSTAT'06, A.Rizzi ed.* , Physica-Verlag, 821-828

Dauxois, J. & Pousse, A. (1976) : *Les Analyses factorielles en calcul des Probabilités et en Statistique : Essai d'étude synthétique.* Thèse de Doctorat d'État, Université Paul Sabatier,Toulouse.

de Carvalho, F. A. T. , Saporta, G. , Queiroz, D. N. (2010): A Clusterwise Center and Range Regression Model for Interval-Valued Data, *Proceedings COMPSTAT'2010*, Springer, 461-468

De la Cruz, O., Holmes, S.P. (2011): The Duality Diagram in Data Analysis: Examples of Modern Applications, *Annals of Applied Statistics* ;5(4): 2266-2277

De Roover, K. , Ceulemans, E., Timmerman, M.E., Vansteelandt, K., Stouten, J., Onghena, P. (2012) : Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. *Psychol Methods*. 17(1):100-119.

Deville J.C., (1974) : Méthodes statistiques et numériques de l'analyse harmonique, *Annales de l'INSEE,* 15, 3-101

Deville J.C., Saporta, G. (1980): Analyse harmonique qualitative, in *Data Analysis and Informatics, E.Diday*, North-Holland, 375-389

Diday, E. (1971) : Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Revue de Statistique Appliquée*, 19 ,2, 19-33

Diday, E. (1974) : Introduction à l'analyse factorielle typologique. *Revue de Statistique Appliquée*, 22 4, 29-38

Fyfe, C., & Lai, P. L. (2001). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10, 365–374

Gifi, A. (1990): *Non-linear multivariate analysis*, Wiley, New York

Hardoon, D.R., Shawe-Taylor, J. (2009) : Sparse Canonical Correlation Analysis, arXiv:0908.2724

Lauro, N., D'ambra, L. (1984) : L'analyse non symétrique des correspondances. *In: Diday, E. eds. Data Analysis and Informatics.* Elsevier, North Holland, 433-446

Preda, C. , Saporta, G. (2005): Clusterwise PLS regression on a stochastic process, *Computational Statistics and Data Analysis*, 49, 99-108

Ramsay, J.O. , Silverman, B. (1997): *Functional data analysis*, Springer

Rizzi, A. (2008) : Italian contributions to Data Analysis, *Electronic Journal for History of Probability and Statistics*, **4**,2

Rosipal, A., Trejo, L. (2001): Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space, *Journal of Machine Learning Research,* 2 , 97-123

Schölkopf, B., Smola,A., Müller, K.L. (1998): Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, 10, 5 , 1299-1319

Suykens, J.A.K.; Vandewalle, J. (1999) : Least squares support vector machine classifiers", *Neural Processing Letters*, 9 (3), 293-300

Tenenhaus, A., Tenenhaus, M. (2011): Regularized Generalized Canonical Correlation Analysis, *Psychometrika*, 76, 2, 257-284

Tenenhaus, M. (1999) : L'approche PLS, *Revue de Statistique Appliquée*, vol. XLVII, n°2, pp.5-40

Tibshirani, R. (1996) : Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Series B, 58, 267-288

Tukey, J.W. (1962): The Future of Data Analysis, *Ann. Math. Statist*. **33**, 1, 1-67

Witten,D., Tibshirani, R., and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10(3):515-534.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15, 265-286