THE MULTIVARIATE ANALYSIS RESEARCH GROUP

Carles M Cuadras
Departament d'Estadística
Facultat de Biologia
Universitat de Barcelona

The set of statistical methods known as Multivariate Analysis covers a wide group of theoretical and applied methods, including factor analysis, classification, multivariate analysis of variance and multidimensional scaling. Since 1980, and especially since 1989, a group of researchers in Barcelona has been studying geometric aspects of statistics, with applications in statistical inference, regression and multivariate analysis.

Distance-based regression: In this approach a response variable is predicted using several explanatory variables on quantitative and qualitative measurement scales. The key idea is to define a distance between observations and to project the response variable on the principal dimensions obtained via multidimensional scaling. This distance-based (DB) model, which may be named «principal coordinate regression», generalizes and improves the multiple regression model, as well as the non-linear regression model by using suitable distance functions (Cuadras, 1989; Cuadras and Arenas, 1990; Cuadras, Arenas and Fortiana, 1996). Related results can be found in Cuadras (1993) and Fortiana and Cuadras (1997).

Distance-based discriminant analysis: The classic problem of allocating an observation to two or more known groups, can also be solved by using the DB approach. Using only distance matrices, one for each group, we can define a proximity function which reduces to the classic discriminant function for particular distances, such as Euclidean or Mahalanobis. This method allows us to handle nominal variables and missing data, and approach the problem of typicality in classification (Cuadras, 1989; Cuadras, 1992; Cuadras, Fortiana and Oliva, 1997; Cuadras, Atkinson and Fortiana, 1997; Cuadras and Fortiana, 2000; Villarroya, Rios and Oller, 1995).

Related metric scaling: A natural extension of this DB approach is to define a joint distance as a function of two given distances on the same set, which satisfies some specific rules (e.g., additivity in the case of independence) and preserves the redundancy between both distances. This approach allows us to relate distances and to represent multivariate data under two different kinds of information (Cuadras and Fortiana, 1995, 1996; Cuadras, 1998; Arenas *et al.*, 2000).

Principal components of a random variable: Just as we can obtain the principal components of a finite set of variables, we can define and obtain the principal directions of a Bernoulli process associated with a continuous random variable. This construction is useful in goodness-of-fit assessment, in expanding a random variable in terms of its principal components, in constructing bivariate distributions with given marginals and in studying the asymptotic distribution of some statistics related to Rao's quadratic entropy (Cuadras and Fortiana, 1995; Cuadras and Lahlou, 2000).

Distributions with given marginals: The construction of joint distributions given the marginals and some dependence parameters, is of great interest in probability and statistics. Families of distributions have been obtained when an interdependence matrix between marginals is given and when the regression curve is given. A continuous extension of correspondence analysis has also been derived (Cuadras and Auge, 1981; Ruiz-Rivas and Cuadras, 1988; Cuadras, 1992; Cuadras, 1996; Cuadras and Fortiana, 1997; Cuadras, Fortiana and Greenacre, 2000; Cuadras, 2002).

Differential geometry in statistics: If we interpret a statistical model as a Riemannian manifold, and define a distance between parameters through geodesics by using the Fisher information matrix as the metric tensor, we obtain the Rao distance, a natural extension of the Mahalanobis distance. This allows us to study the geometry of any regular statistical model and approach some problems in statistical inference, such as intrinsic estimation, invariance, testing of hypotheses and representing parametric models. (Oller and Cuadras, 1985; Oller, 1989; Calvo and Oller, 1990; Oller and Corcuera, 1995; Rios, Villarroya and Oller, 1992; Villarroya and Oller, 1993; Villarroya, Rios and Oller, 1995; Garcia and Oller, 2001).

References

1. Arenas, C. and Cuadras, C. M. (2002). «Recent statistical methods based on distances». *Contributions to Science*, in press.
2. Arenas, C., Escudero, T., Mestres, F., Coll, M. D. and Cuadras, C. M. (2000). «Cacromos: a fortran program to reconstruct the position of human chromosomes». *Hereditas*, 132, 157-159.
3. Barata, C., Baird, D., Miñarro, A. and Soares, A. (2000). «Do genotype responses always converge from lethal to nonlethal toxicant exposure levels? Hypothesis tested using clones of Daphnia magna straus». *Environmental Toxicology and Chemistry*, 19, 2314-2322.
4. Burbea, J., Oller, J. M. and Reverter, F. (2002). «Some remarks on the information geometry of the gamma distribution». *Communications in Statistic: Theory and Methods*, 31, 1959-1975.
5. Calvo, M. and Oller, J. M. (1990). «A distance between multivariate normal distribution based in an embedding into the Siegel group». *Journal of Multivariate Analysis*, 35, 223-242.

6. Calvo, M. and Oller, J. M. (2002). «A distance between elliptical distributions based in an embedding into the Siegel group». *Journal of Computational and Applied Mathematics*, 145, 319-334.
7. Calvo, M., Villarroya, A. and Oller, J. M. (2002). «A biplot method for multivariate normal populations with unequal covariance matrix». *TEST*, 11, 143-165.
8. De la Cruz, X. and Calvo, M. (2001). «Use of surface area computations to describe atom-atom interactions». *Computer-Aided Molecular Design*, 15, 521-532.
9. Cuadras, C. M. (1989). «Distance analysis in discrimination and classification using both continuous and categorical variables», in Y. Dodge (ed.), *Statistical Data Analysis and Inference*, 459-473. Elsevier Science Publishers, Amsterdam.
10. Cuadras, C. M. (1992). «Probability distributions with given multivariate marginals and given dependence structure». *Journal of Multivariate Analysis*, 42, 51-66.
11. Cuadras, C. M. (1992). «Some examples of distance based discrimination». *Biometrical Letters*, 29, 1-18.
12. Cuadras, C. M. (1993). «Interpreting an inequality in multiple regression». *The American Statistician*, 47, 256-258.
13. Cuadras, C. M. (1996). «A distribution with given marginals and given regression curve», in: L. Rüschendorf, B. Schweizer and D. Taylor (eds.), *Distributions with Fixed Marginals and Related Topics*, 76-83. IMS-Lecture Notes-Monograph Series, Hayward, California.
14. Cuadras, C. M. (1998). «Multidimensional dependencies in classification and ordination», in: K. Fernández-Aguirre and A. Morineau (eds.), *Analyse Multidimensionnelles des Données*, 15-25. CISIA, Saint Mandé, France.
15. Cuadras, C. M. (1998). «Some cautionary notes on the use of principal components regression (Revisited)». *The American Statistician*, 52, 371.
16. Cuadras, C. M. (2000). «Distributional relationships arising from simple trigonometric formulas (Revisited)». *The American Statistician*, 54, 87.
17. Cuadras, C. M. (2000). «Discussion on "Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests" by E. del Barrio, J. A. Cuesta-Albertos and C. Matrán». *TEST*, 9, 1-96.
18. Cuadras, C. M. (2002). «On the covariance between functions». *Journal of Multivariate Analysis*, 81, 19-27.
19. Cuadras, C. M. (2002). «Correspondence analysis and diagonal expansions in terms of distribution functions». *Journal of Statistical Planning and Inference*, 103, 137-150.
20. Cuadras, C. M. (2002). «Diagonal distributions via diagonal expansions and tests of independence». In *Distributions with given marginals and statistical modelling*, (C. M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena, Eds.), Kluwer Academic Publishers Dordrecht, 35-42.
21. Cuadras, C. M. (2002). «Discussion on "Skewed multivariate models related to hidden truncation and/or selective reporting" by B. C. Arnold and R. J. Beaver». *TEST*, 11, 7-54.

22. Cuadras, C. M. (2002). «Geometrical understanding of the Cauchy distribution». *Qüestiió*, 26, 283-287.
23. Cuadras, C. M., Fortiana, J. and J. A. Rodriguez-Lallena (eds.) (2002). *Distributions with given marginals and statistical modelling*. Kluwer Academic Publishers, Boston/Dordrecht/London, 272 pp.
24. Cuadras, C. M. and Lahlou, Y. (2000). «Some orthogonal expansions for the logistic distribution». *Communications in Statistics: Theory and Methods*, 29, 2643-2663.
25. Cuadras, C. M. and Lahlou, Y. (2002). «Principal components of the Pareto distribution». In *Distributions with given marginals and statistical modelling* (C. M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena, Eds.), Kluwer Academic Publishers, Dordrecht, 43-50.
26. Cuadras, C. M. and Cuadras, D. (2001). «Principal directions for the normal random variable». *Statistical Review*, 2, 101-102.
27. Cuadras, C. M. and Cuadras, D. (2002). «Orthogonal expansions and distinction between logistic and normal». In *Goodness-of-fit Tests and Validity Models*. C. Huber-Carol, N. Balakrishnan, M. S. Nikulin, M. Mesbah, eds., 325-338, Birkhauser, Boston.
28. Cuadras, C. M. and Arenas, C. (1990). «A distance based regression model for prediction with mixed data». *Communications in Statistics: Theory and Methods*, 19, 2261-2279.
29. Cuadras, C. M., Arenas, C. and Fortiana, J. (1996). «Some computational aspects of a distance-based model for prediction». *Communications in Statistics: Simulation and Computation*, 25, 593-609.
30. Cuadras, C. M., Atkinson, R. A. and Fortiana, J. (1997). «Probability densities from distances and discriminant analysis». *Statistics and Probability Letters*, 33, 405-411.
31. Cuadras, C. M. and Augé, J. (1981). «A continuous general multivariate distribution and its properties». *Communications in Statistics: Theory and Methods*, A10, 339-353.
32. Cuadras, C. M. and Fortiana, J. (1995). «A continuous metric scaling solution for a random variable». *Journal of Multivariate Analysis*, 52, 1-14.
33. Cuadras, C. M. and Fortiana, J. (1996). «Weighted continuous metric scaling», in A. K. Gupta and V. L. Girko (eds.), *Multidimensional Statistical Analysis and Theory of Random Matrices*, 27-40. VSP, Zeist, The Netherlands.
34. Cuadras, C. M. and Fortiana, J. (1997). «Continuous scaling on a bivariate copula», in V. Benes and J. Stepan (eds.), *Distributions with Given Marginals and Moment Problems*, 137-142. Kluwer Academic Publishers, The Netherlands.
35. Cuadras, C. M. and Fortiana, J. (2000). «The importance of geometry in multivariate analysis and some applications», in C. R. Rao and G. Szekely, (eds.), *Statistics for the 21st Century*, 93-108. Marcel Dekker, N. York.

36. Cuadras, C. M., Fortiana, J. and Greenacre, M. J. (1999). «Continuous extensions of matrix formulations in correspondence analysis, with applications to the FGM family of distributions», in R. D. H. Heijmans, D. S. G. Pollock and A. Satorra, (eds.), *Innovations in Multivariate Statistical Analysis*, 101-116. Kluwer Academic Publishers, The Netherlands.
37. Cuadras, C. M., Fortiana, J. and Oliva, F. (1997). «The proximity of an individual to a population with applications in discriminant analysis». *Journal of Classification*, 14, 117-136.
38. Cuadras, C. M. and Fortiana, J. (1993). «Continuous metric scaling and prediction», in C. M. Cuadras and C. R. Rao (eds.), *Multivariate Analysis: Future Directions*, 2, Elsevier, Amsterdam, 47-66.
39. Cuadras, C. M. and Fortiana, J. (1994). «As certeining the underlying distribution of a data set», in *Selected Topics on Stochastic Modelling*, R. Gutierrez and M. J. Valderrama (eds.), World Scientific, Singapore, 223-230.
40. Cuadras, C. M. and Fortiana, J. (1998). «Visualizing categorical data with related metric scaling». In *Visualization of Categorical Data*, ch. 25, J. Blassius and M. Greenacre (eds.), Ac. Press, N. York, 365-376.
41. Cuadras, C. M. and Rao, C. R. (eds.) (1993). *Multivariate Analysis: Future Directions* 2, Elsevier, Amsterdam, 488 pp.
42. Cubedo, M. and Oller, J. M. (2002). «Hypothesis testing: a model selection approach». *Journal of Statistical Planning and Inference*, 108, 3-21.
43. Fortiana, J. and Grané, A. (2002). «A scale-free goodness-of-fit statistic for the exponential distribution based on maximum correlations». *Journal of Statistical Planning and Inference*, 108, 85-97.
44. Fortiana, J. and Grané, A. (2002). «Goodness-of-fit tests based on maximum correlations and their orthogonal decompositions». *Journal of Royal Statistical Society B*, in press.
45. García, G. and Oller, J. M. (2001). «Minimum Riemannian risk equivariant estimator for the univariate normal model». *Statistics & Probability Letters*, 52 109-113.
46. Grané, A. and Fortiana, J. (2002). «Maximum correlations and tests of goodness-of-fit». In *Distributions with given marginals and statistical modelling* (C. M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena, eds.), Kluwer Academic Publishers, Dordrecht, 113-122.
47. Ruiz-Rivas, C. and Cuadras, C. M. (1988). «Inference properties of a one-parameter curved exponential family of distributions with given marginals». *Journal of Multivariate Analysis*, 27, 447-456.
48. Fortiana, J. and Cuadras, C. M. (1997). «A family of matrices, the discretized Brownian Bridge and distance-based regression». *Linear Algebra and its Applications*, 263, 173-188.
49. Oller, J. M. (1989). «Some geometrical aspects of data analysis and statistics», in: Y. Dodge (ed.), *Statistical Data Analysis and Inference*, 41-58. Elsevier Science Publishers, Amsterdam.

50. Oller, J. M. and Corcuera, J. M. (1995). «Intrinsic analysis of statistical estimation». *Annals of Statistics*, 23, 1562-1581.
51. Ríos, M., Gracia, J. M., Sánchez, J. A. and Pérez, D. (2000). «A statistical analysis of the seasonality in Pulmonary Tuberculosis». *European Journal of Epidemiology*, 16, 483-488.
52. Oller, J. M. and Cuadras, C. M. (1985). «Rao's distance for negative multinomial distributions». *Sankhyã*, A 47, 75-83.
53. Ríos, M., Villarroya, A. and Oller, J. M. (1992). «Rao distance between multivariate linear models and their applications to the classification of response curves». *Computational Statistics and Data Analysis*, 13, 431-441.
54. Villarroya, A. and Oller, J. M. (1993). «Statistical tests for the inverse Gaussian distribution based on Rao distance». *Sankhyã*, A55, 80-103.
55. Villarroya, A., Ríos, M. and Oller, J. M. (1995). «Discriminant analysis algorithm based on a distance funtion and a Bayesian decision». *Biometrics*, 51, 908-919.