JAN L.A. VAN RIJCKEVORSEL
JAN DE LEEUW

COMPONENT AND CORRESPONDENCE ANALYSIS

DIMENSION REDUCTION BY FUNCTIONAL APPROXIMATION

WILEY SERIES IN PROBABILITY AND MATHEMATICAL STATISTICS

WILEY

Component and Correspondence Analysis

Dimension Reduction by Functional Approximation

Edited by

JAN L.A. VAN RIJCKEVORSEL

Department of Statistics,
TNO Institute for Preventive Health Care
Leiden,
The Netherlands

JAN DE LEEUW

Department of Psychology and Mathematics
UCLA, Los Angeles, USA

JOHN WILEY & SONS

Chichester · New York · Brisbane · Toronto · Singapore

Library of Congress Cataloging in Publication Data:

Component and correspondence analysis: dimension reduction by
functional approximation/edited by Jan L. A. van Rijckevorsel, Jan
de Leeuw.

p. cm. — (Wiley series in probability and mathematical statistics)
Bibliography: p.
Includes index.
ISBN 0 471 91847 4
1. Principal components analysis. 2. Correlation (Statistics)
I. Rijckevorsel, Jan L. A. van II. Leeuw, Jan de. III. Series.
QA278.5.C657 1988
519.5′354—dc19

87–33981
CIP

British Library Cataloguing in Publication Data Available

Typeset by Cotswold Typesetting Ltd., Gloucester
Printed and bound in Great Britain by Anchor Brendon Ltd, Tiptree, Essex

Contents

List of Contributors

Dr Paul Bekker, *Dept of Econometrics, Nettelbosje* 2, *University of Groningen*, 9747 *AE Groningen, The Netherlands.*

Dr Philippe Besse, *Laboratoire de Statistique et Probabilités, Université Paul Sabatier*, 118 *Route de Narbonne*, 31062 *Toulouse Cedex, France*

Dr Jan de Leeuw, *Dept of Psychology and Mathematics, UCLA*, 405 *Hilgard Avenue, Los Angeles, CA* 90024, *USA.*

Dr J.-F. Martin, *Centre de Mathematiques, INSA, bâtiment* 403, 20 *Avenue Albert Einstein*, 69621 *Villeurbanne Cedex, France.*

Dr Jan L.A. van Rijckevorsel, *Dept of Statistics, TNO NIPG, Wassenaarseweg* 56, 2333 *AL Leiden, The Netherlands.*

Dr Suzanne Winsberg, 18 *rue Visconti, Paris* 75006, *France.*

References

Agarwal, G.G. and Studden, W.J. (1978). Asymptotic design and estimation using linear splines, *Commun. Statist.-Simula. Computa B7*, **4**, 309–319.

Andrews, D.F., Gnanadesikan, R. and Warner, J.L. (1971). Transformation of multivariate data, *Biometrika*, **27**, 825–40.

Aronszajn, N. (1950). Theory of reproducing kernel, *Transactions of the American Mathematical Society*, **68**, 337–404.

Bartlett, M.S. (1953). *Factor Analysis in Psychology as a Statistician Sees it*. Uppsala Symposium on Psychol. Factor Analysis. Uppsala: Almqvist & Wicksell.

Bekker, P. (1983). *Relationships Between Versions of Nonlinear Principal Component Analysis*. Leiden: Department of Data Theory.

Besse, P. (1979). *Etude descriptive des processus: Approximation et interpolation*. Thèse de 3ème Cycle. Toulouse: Université Paul Sabatier.

Besse, P. (1980). Deux exemples d'Analyse en Composantes Principales filtrantes, *Statistique et Analyse des Données*, **3**, 5–15.

Besse, P. (1986). Choice of a metric or choice of a model. The case of discrete events data. In F. de Antoni *et al.* (eds), *COMPSTAT. Proceedings in Computational Statistics*. Vienna: Physica-Verlag.

Besse, P. (1987). Choix de la métrique pour l'a.c.p. de séries d'évènements discrets, *Statistique et Analyse des données*, to appear.

Besse, P., Caussinus, H., Ferre, L. and Fine, J. (1987). Principal components analysis and optimization of graphical displays, submitted to *Statistics*.

Besse, P. and Ramsay, J.O. (1986). Principal component analysis of sampled functions, *Psychometrika*, **51**, 2, 285–311.

Besse, P. and Vidal, C. (1982). Analyse des correspondances et codage par une Probabilité de transition, *Statistique et Analyse des Données*, **7**, 3.

Benzécri, J.P. *et al.* (1973). *Analyse des données* (2 vols). Paris: Dunod.

Benzécri, J.P. (1980). *Pratique de l'analyse des données* 1. Paris: Dunod.

Bezdek, J.C. (1987). Some nonstandard clustering algorithms. In P. Legendre and L. Legendre (eds), *Developments in Numerical Ecology*. New York: Springer.

Boneva, L., Kendall, D. and Stefanov, I. (1971). Spline transformations: three new diagnostic aids for the data analyst, *JRSS Ser B.*, **33**, 1–70.

Bordet, J.P. (1973). *Etudes de données geophysiques*. Thèse de 3ème Cycle. Paris: Université de Paris VI.

Bowman, A.W. (1980). A note on consistency of the kernel method for the analysis of categorical data. *Biometrika*, **67**, 3.

Box, G.E.P. and Cox, D.R. (1964). Analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B*, **26**, 211–52.

Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *JASA*, **80**, 580–598.

Burt, C. (1950). The factorial analysis of qualitative data, *British Journal of Statistical Psychology*, **3,** 166–85.
Carroll, J. D. (1968). A generalization of canonical correlation analysis to three or more sets of variables, *Proc. 76th Conv. APA*, 227–8.
Cattell, R. B. (1966). The scree test for the number of factors. *J. Multiv. Behav. Res.*, **1,** 245–76.
Chang, J.C. and Bargmann, R.E. (1974). *Internal Multidimensional Scaling of Categorical Variables*. Department of Statistics and Computer Science. Technical Report, University of Georgia.
Cliff, N. (1966). Orthogonal rotation to congruence, *Psychometrika*, **31,** 33–42.
Cogburn, R. and Davis, H.T. (1974). Periodical splines and spectral estimation, *Annals of Statistics*, **2,** 1108–26.
Conte, S. and De Boor, C. (1980). *Elementary Numerical Analysis*. 2nd edition. New York: McGraw-Hill.
Coolen, H., Van Rijckevorsel, J. and De Leeuw, J. (1982). An algorithm for nonlinear principal components analysis with B-splines by means of alternating least squares. In H. Caussinus *et al.* (eds), *COMPSTAT* 1982, part II. Vienna: Physika-Verlag.
Craddock, J.M. and Flood, C.R. (1969). Eigenvectors for representing the 500 mb. geopotential surface over the Northern Hemisphere. *Q.J.R. Met. Soc.*, **95,** 576–93.
Daudin, J.J., Duby, C. and Trecourt, P. (1987). Stability of principal components analysis studied by the bootstrap method, *Statistics*, to appear.
Dauxois, J. and Pousse, A. (1976). *Les analyses factorielles en calcul de probabilités et en statistique: Essai d'étude synthetique*. Thèse d'Etat. Toulouse: Université de Toulouse III.
De Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer.
De Leeuw, J. (1973). *Canonical Analysis of Categorical Data*. Leiden: Psychological Institute. Reissued by DSWO Press, Leiden, 1984.
De Leeuw, J. (1976). *HOMALS*. Paper presented at the Psychometric Society meeting, Murray Hill, NJ.
De Leeuw, J. (1982). Nonlinear principal component analysis. In H. Caussinus *et al.* (eds), *COMPSTAT* 82. *Proceedings in Computational Statistics*, part I. Vienna: Physika-Verlag.
De Leeuw, J. (1983). On the prehistory of correspondence analysis, *Statistica Neerlandica*, **37,** 161–164.
De Leeuw, J. (1984a). The Gifi-system of non-linear multivariate analysis. In E. Diday, M. Jambu, L. Lebart, J. Pages and R. Tomassone (eds), *Data Analysis and Informatics*, volume III. Amsterdam: North-Holland.
De Leeuw, J. (1984b). Statistical properties of multiple correspondence analysis. Submitted for publication.
De Leeuw, J. (1984c). Models of Data, *Kwantitatieve Methoden*, **5,** 17–31.
De Leeuw, J. and Tijssen, R. (1984). *Multivariate Analysis with Optimal Scaling*. Leiden: Department of Data Theory.
De Leeuw, J., Van Rijckevorsel, J. and Van der Wouden, H. (1981). Nonlinear principal component analysis with B-splines, *Methods of Operations Research*, **33,** 379–93.
De Leeuw, J. and Walter, J. (1977). *Optimal Scaling of Continuous Numerical Data*. Leiden: Department of Data Theory.
De Leeuw, J. and Van Rijckevorsel, J. (1980). Homals en Princals. In E. Diday *et al.* (eds), *Data Analysis and Informatics*. Amsterdam: North-Holland.
De Leeuw, J., Young, F.W. and Takane, Y. (1976). Additive structure in qualitative data: an alternating least squares method with optimal scaling features, *Psychometrika*, **41,** 471–503.

De Vore, R. (1977). Monotone approximation by splines, *SIAM J. Math. Anal.*, **8,** 891–905.

Deville, C. (1977). Un exemple catastrophique d'analyse factorielle et son explication. In E. Diday (ed), *Premières Journées Internationales, Analyse des Données et Informatique.* Versailles: Inria, pp. 557–65.

Duc-Jacquet, M. (1973). *Approximation des fonctionnelles linéaires sur des espaces hilbertiens auto-reproduisants*, unpublished doctoral dissertation, Grenoble.

Eagle, A. (1928). On the relation between the fourier constants of a periodic function and the coefficients determined by harmonic analysis. *Phil. Mag.*, **5,** 113–32.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, **1,** 211–18.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests, Grundlagen und Anwendungen.* Berne: Verlag Hans Huber.

Gallego, F.J. (1980). *Un codage flou pour l'analyse des correspondances.* Thèse de 3ème Cycle. Paris: Université de Paris VI.

Gautier, J.M. and Saporta, G. (1982). About fuzzy discrimination. In H. Caussinus *et al.* (eds), *Compstat* 1982. *Proceedings in Computational Statistics.* Vienna: Physica-Verlag.

Ghermani, B.M., Roux, C. and Roux, M. (1977). Sur le codage logique des données héterogènes, *Les Cahiers de l'Analyse des Données*, **1,** 115–18.

Gifi, A. (1981a). *Non-linear Multivariate Analysis.* Leiden: Department of Data Theory.

Gifi, A. (1981b). *HOMALS User's Guide.* Leiden: Department of Data Theory.

Gifi, A. (1982). *PRINCALS User's Guide.* Leiden: Department of Data Theory.

Gifi, A. (1988). *Non-linear Multivariate Analysis.* Leiden: DSWO Press.

Good, I.J. and Gaskins, R.A. (1971). Non-parametric roughness penalties for probability densities, *Biometrika*, **58,** 255–77.

Greenacre, M.J. (1981). Practical correspondence analysis. In V. Barnett (ed), *Interpreting Multivariate Data.* New York: Wiley.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis.* New York: Academic Press.

Guitonneau, G.G. and Roux, M. (1977). Sur la taxinomie du genre erodium, *Les Cahiers de l'Analyse des Données*, **1,** 97–113.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst *et al.* (eds), *The Prediction of Personal Adjustment.* New York: Social Science Research Council.

Harman, H.H. (1976). *Modern Factor Analysis.* Chicago: University of Chicago Press (revised edn).

Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Ann. Inst. Statist. Math.*, **5,** 121–143.

Holmquist, N.D., McMahan, C.A. and Williams, O.D. (1967). Variability in classification of carcinoma in situ of the uterine cervix, *Arch. Path.*, **84,** 334–45.

Horst, P. (1965). *Factor Analysis of Data Matrices.* New York: Holt, Rinehart & Winston.

Hudson, D.J. (1966). Fitting segmented curves whose join points have to be estimated, *JASA*, **61,** 1097–1129.

Jambu, M. and Lebeaux, M.O. (1983). *Cluster Analysis and Data Analysis.* Amsterdam: North-Holland.

Jolliffe, I. T. (1986). *Principal Component Analysis.* New York: Springer-Verlag.

Kettenring, J.R. (1971). Canonical analysis of several sets of variables, *Biometrika*, **58,** 433–60.

Koyak, R. (1985). *Optimal Transformations for Multivariate Linear Reduction Analysis.* Unpublished PhD thesis. University of California, Berkeley, California: Dept. of Statistics.

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data, *Journal of the Royal Statistical Society*, series B, **27**, 251–63.
Kruskal, J.B. and Shepard, R. N. (1974). A nonmetric variety of linear factor analysis, *Psychometrika*, **39**, 123–57.
Lafaye de Michaux, D. (1978). *Approximations d'analyses canoniques nonlineaires de variables aléatoires et analyses factorielles privelégiantes*. Thèse de Docteur-Ingenieur. Nice: Université de Nice.
Lancaster, H.O. (1969). *The Chi-squared Distribution*. New York: Wiley.
Lebart, L., Morineau, A. and Warwick, K.M. (1984). *Multivariate Descriptive Statistical Analysis*. New York: Wiley.
Le Foll, Y. (1979). *Sur les propriétés de l'analyse des correspondances pour diverses formes complètes de données*. Thèse de 3ème Cycle. Paris: Université de Paris VI.
Mallet, J.L. (1982). Propositions for fuzzy characteristics functions in data analysis. In H. Caussinus *et al.* (eds), *Compstat* 82. *Proceedings in Computational Statistics*. Vienna: Physica-Verlag.
Martin, J.F. (1980) *Le codage flou et ses applications en statistique*. Thèse de 3ème Cycle. Pau: Université de Pau et des Pays de l'Adour.
Martin, J.F. (1981). Formalisation et étude de codages flous, *C.R. Acad. Sci. Paris*, **292**, 299–301.
Maung, K. (1941). Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colors of Scottish school children. *Ann. Eugen.*, **11**, 189–223.
McDonald, R.P. (1968). A unified treatment of the weighting problem, *Psychometrika*, **33**, 351–81.
Munoz-Perez, F. (1982). L'évolution de la fécondité dans les pays industrialisés dépuis 1971, *Population*, 37iéme année, **3**, 483–512.
Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and its Applications*. Toronto: University of Toronto Press.
Park, S.H. (1978). Experimental designs for fitting segmented polynomial regression models, *Technometrics*, **20**, 2, 151–4.
Poirier, D.J. (1973). Piecewise regression using cubic splines. *JASA*, **68**, 515–24.
Poirier, D.J. (1976). *The Econometrics of Structural Change*. Amsterdam: North-Holland.
Quade, W. and Collatz, L. (1938). Zur Interpolationstheorie der Reelen Periodischen Funktionen, *Aka, Wiss.* (*Math-Phys. Kl*), **30**, 383–429.
Rao, C.R. (1964). The use and interpretation of Principal Component Analysis in applied research, *Sankhyā*, **26(A)**, 329–58.
Ramsay, J.O. (1977). Monotonic weighted power transformations to additivity, *Psychometrika*, **42**, 83–109(a).
Ramsay, J.O. (1982). When data are functions, *Psychometrika*, **47**, 379–96.
Runge, C. (1901). Uber Die Darstellung Willkürlichen Funktionen und die Interpolation zwischen Aquidistanten Ordinaten, *Z. Angew, Math. Phys.*, **46**, 224–43.
Ruspini, E. (1969). A new approach to clustering, *Inf. and Control*, **15**, 22–32.
Saporta, G. (1985). Data analysis for numerical and categorical individual time-series. *Applied Stochastic Models and Data Analysis*, **1**, 109–119.
Schoenberg, I.J. (1946a). Contributions to the problem of approximation of equidistant data by analytic functions, Part A, *Quart. Appl. Math.*, **4**.
Schoenberg, I.J. (1946b). Contributions to the problem of approximation of equidistant data by analytic functions, Part B, *Quart. Appl. Math.*, **4**.
Schriever, B.F. (1986). *Order Dependence*, CWI-Tract 20. Amsterdam: Centre for Mathematics and Computer Science.
Schumaker, L. (1981). *Spline Functions: Basic Theory*. New York: Wiley.

Shapiro, H.S. (1971). *Topics in Approximation Theory*. New York: Springer-Verlag.

Smith, P. (1979). Splines as a useful statistical tool, *The American Statistician*, **33**, 57–62.

Styan, G.P. (1973). Hadamard products and multivariate statistical analysis, *Linear Algebra and its Applications*, **6**, 217–240.

Takane, Y., Young, F.W. and De Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features, *Psychometrika*, **43**, 278–81.

Tenenhaus, M. (1977). Analyse en composantes principales d'un ensemble de variables nominales et numériques, *Revue de Statistique Appliqué*, **25**, 39–56.

Tenenhaus, M., and Young, F.W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data, *Psychometrika*, **50**, 91–119.

Thurstone, L.L. (1947). *Multiple Factor Analysis*. Chicago: University of Chicago.

Tijssen, R. (1984). *A New Approach to Nonlinear Canonical Correlation Analysis*. Leiden: Department of Data Theory.

Van de Geer, J.P. (1986). *Introduction to Linear Multivariate Analysis*. (2 Vols). Leiden: DSWO Press.

Van der Heijden, P.G.M. (1987). *Correspondence Analysis of Longitudinal Categorical Data*. Leiden: DSWO Press.

Van Rooij, P.L.J. and Schurer, F. (1974). A bibliography on spline functions. In K. Böhmer *et al.* (eds), *Tagung über Spline Funktionen*. Mannheim: Bigliographisches Institut, pp. 315–415.

van Rijckevorsel, J.L.A. (1982). Canonical analysis with B-splines. In H. Caussinus *et al.* (eds), *Compstat* 1982. *Proceedings in Computational statistics*. Vienna: Physica-Verlag, pp. 393–8.

van Rijckevorsel, J.L.A. (1987). *The Application of Horseshoes and Fuzzy Coding in Multiple Correspondence Analysis*. Leiden: DSWO Press.

van Rijckevorsel, J.L.A. and van Kooten, G. (1985). Smooth PCA of economic data, *Computational Statistics Quarterly*, **2**, 143–72.

Wahba, G. (1971). A polynomial algorithm for density estimation, *Annals of Mathematical Statistics*, **42**, 1870–86.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression, *JRSC*, Ser. B, **40**, 364–72.

Wegman, E.J. and Wright, I. (1983). Splines in statistics, *JASA*, **78**, 351–65.

Winsberg, S. and Kruskal, J.O. (1986). Easy to generate metrics for use with sampled functions. In F. De Antoni *et al.* (eds), *COMPSTAT* 86. Proceedings in Computational Statistics. Vienna: Physica-Verlag.

Winsberg, S. and Ramsay, J.O. (1980). Monotonic transformations to additivity using splines, *Biometrika*, **67**, 669–74.

Winsberg, S. and Ramsay, J.O. (1981). Analysis of pairwise preferences data using integrated B-splines, *Psychometrika*, **46**, 171–86.

Winsberg, S. and Ramsay, J.O. (1982). Monotone splines: a family of transformations useful for data analysis. In H. Caussinus *et al.* (eds), *Compstat* 82. *Proceedings in Computational Statistics*. Vienna: Physica Verlag, pp. 451–6.

Winsberg, S. and Ramsay, J.O. (1983). Monotone spline transformations for dimension reduction, *Psychometrika*, **48**, 575–95.

Wold, S. (1974). Spline functions in data analysis, *Technometrics*, **16**, 1–11.

Wold, H. and Lyttkens, E. (1969). Nonlinear iterative partial least squares (Nipals) estimation procedures, *Bulletin of the International Statistical Institute*, **43**, 29–51.

Wright, I.W. and Wegman, E.J. (1980). Isotonic, convex and related splines, *Annals of Statistics*, **8**, 1023–35.

Young, F.W., De Leeuw, J. and Takane, Y. (1976). Regression with qualitative variables: an alternating least squares method with optimal scaling features, *Psychometrika*, **41**, 505–529.

Young, F.W., Takane, Y. and De Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: an alternating least squares methods with optimal scaling features, *Psychometrika*, **43**, 279–81.

Young, F.W. (1981). Quantitative analysis of qualitative data, *Psychometrika*, **46**, 357–88.

Young, F.W., De Leeuw, J. and Takane, Y. (1980). Quantifying qualitative data. In E. D. Lantermann and H. Feger (eds), *Similarity and Choice. Papers in Honour of Clyde Coombs*. Berne: Hans Huber.

Zadeh, L.A. (1965). Fuzzy sets, *Inf. and Control*, **8**, 338–53.

Index