International Encyclopedia of Statistical Science. Lovric, M. (ed.), New York: Springer, (2011), Part 13, 915-920

MULTIVARIATE STATISTICAL ANALYSIS

Nanny Wermuth [1]

Professor of Statistics, Department of Mathematical Sciences
Chalmers Technical University/University of Gothenburg, Sweden

Classical multivariate statistical methods concern models, distributions and inference based on the Gaussian distribution. These are the topics in the first textbook for mathematical statisticians by T.W. Anderson that was published in 1958 and that appeared as a slightly expanded 3rd edition in 2003. Matrix theory and notation is used there extensively to efficiently derive properties of the multivariate Gaussian or the Wishart distribution, of principal components, of canonical correlation and discriminant analysis and of the general multivariate linear model in which a Gaussian response vector variable Y_a has linear least-squares regression on all components of an explanatory vector variable Y_b.

In contrast, many methods for analysing sets of observed variables have been developed first within special substantive fields and some or all of the models in a given class were justified in terms of probabilistic and statistical theory much later. Among them are factor analysis, path analysis, structural equation models, and models for which partial-least squares estimation have been proposed. Other multivariate techniques such as cluster analysis and multidimensional scaling have been often used, but the result of such an analysis cannot be formulated as a hypothesis to be tested in a new study and satisfactory theoretical justifications are still lacking.

Factor analysis was proposed by psychologist C. Spearman (1904), (1926) and, at the time, thought of as a tool for measuring human intelligence. Such a model has one or several latent variables. These are hidden or unobserved and are to explain the observed correlations among a set of observed variables, called items in that context. The difficult task is to decide how many and which of a possibly large set of items to include into a model. But, given a set of latent variables, a classical factor analysis model specifies for a joint Gaussian distribution mutual independence of the observed variables given the latent variables. This can be recognized to be one special type of a graphical Markov model; see Cox and Wermuth (1996), Edwards (2000), Lauritzen (1996), Whittaker (1990).

[1]Dr Nanny Wermuth is Professor of Statistics, at the joint Department of Mathematical Sciences of Chalmers Technical University and the University of Gothenburg. She is a Past President, Institute of Mathematical Statistics (2008–2009) and Past President of the International Biometric Society (2000–2001). In 1992 she was awarded a Max Planck-Research Prize, jointly with Sir David Cox. She chaired the Life Science Committee of the International Statistical Institute (2001–2005) and was an Associate editor of the *Journal of Multivariate Analysis* (1998–2001) and *Bernoulli* (2007-2010). Professor Wermuth is an Elected member of the German Academy of Sciences and of the International Statistical Institute (1982), an elected Fellow of the American Statistical Association (1989), and of the Institute of Mathematical Statistics (2001). She is a co-author (with David R. Cox) of the text *Multivariate dependencies: models, analysis and interpretation* (Chapman and Hall, 1996).

Path analysis was developed by geneticist S. Wright (1923), (1934) for systems of linear dependence of variables with zero mean and unit variance. He used what we now call directed acyclic graphs to represent hypotheses of how the variables he was studying could have been generated. He compared correlations implied for missing edges in the graph with corresponding observed correlations to test the goodness of fit of such a hypothesis.

By now it is known, under which condition for these models in standardized Gaussian variables, maximum-likelihood estimates of correlations coincide with Wright's estimates via path coefficients. The condition on the graph is simple: there should be no three-node-two-edge subgraph of the following kind $\circ \longrightarrow \circ \longleftarrow \circ$. Then, the directed acyclic graph is said to be decomposable and captures the same independences as the concentration graph obtained by replacing each arrow by an undirected edge. In such Gaussian concentration graph models, estimated variances are matched to the observed variances so that estimation of correlations is equivalent to estimation of covariances.

Wright's method of computing implied path coefficients by 'tracing paths' has been generalized via a so-called separation criterion. This criterion, given by Geiger, Verma and Pearl (1990), permits to read off a directed acyclic graph all independence statements that are implied by the graph. The criterion takes into account that not only ignoring (marginalizing over) variables might destroy an independence, but also conditioning on common responses may render two formerly independent variables to be dependent. In addition, the separation criterion holds for any distribution generated over the graph.

The separation criterion for directed acyclic graphs has been translated into conditions for the presence of edge-inducing paths in the graph; see Marchetti and Wermuth (2009). Such an edge-inducing path is also association-inducing in the corresponding model, given some mild conditions on the graph and on the distributions generated over it; see Wermuth (2010). In the special case of only marginalising over linearly related variables, these induced dependences coincide with the path-tracing results given by Wright provided the directed acyclic graph model is decomposable and the variables are standardised to have zero means and unit variances. This applies not only to Gaussian distributions but also to special distributions of symmetric binary variables; see Wermuth, Marchetti and Cox (2009).

Typically however, directed acyclic graph models are defined for unstandardized random variables of any type. Then, most dependences are no longer appropriately represented by linear regression coefficients or correlations, but maximum-likelihood estimates of all measures of dependence can still be obtained by separately maximizing each univariate conditional distribution, provided only that its parameters are variation-independent from parameters of distributions in the past.

Structural equation models, developed in econometrics, can be viewed as another extension of Wright's path analyses. The result obtained by T. Haavelmo (1943) gave an important impetus. For his insight that separate linear least-squares estimation may be inappropriate for equations having strongly correlated residuals,

Haavelmo received a Nobel prize in 1989. It led to a class of models defined by linear equations with correlated residuals and to responses called endogenous. Other variables conditioned on and considered to be predetermined were named exogenous. Vigorous discussions of estimation methods for structural equations occurred during the first few Berkeley symposia on mathematical statistics and probability from 1945 to 1965.

Path analysis and structural equation models were introduced to sociological research via the work by O.D. Duncan (1966), (1975). Applications of structural equation models in psychological and psychometric research resulted from cooperations between A. Goldberger and K. Jöreskog; see Goldberger (1971), (1972) and Jöreskog (1973) (1981). The methods became widely used once a corresponding computer program for estimation and tests was made available; see also Kline (2010).

In 1962, A. Zellner published his results on seemingly unrelated regressions. He points out that two simple regression equations are not separate if the two responses are correlated and that two dependent endogenous variables need to be considered jointly and require simultaneous estimation methods. These models are now recognized as special cases of both linear structural equations and of multivariate regression chains, a subclass of graphical Markov models; see Cox and Wermuth (1993), Drton (2009), Marchetti and Lupparelli (2010).

But it was not until 40 years later, that a maximum-likelihood solution for the Gaussian distribution in four variables, split into a response vector Y_a and vector variable Y_b, was given and an example of a poorly fitting data set with very few observations for which the likelihood equations have two real roots; see Drton and Richardson (2004). For well-fitting data and reasonably large sample sizes, this is unlikely to happen; see Sundberg (2010). For such situations, a close approximation to the maximum-likelihood estimate has been given in closed form for the seemingly unrelated regression model, exploiting that it is a reduced model to the covering model that has closed-form maximum-likelihood estimates, the general linear model of Y_a given Y_b; see Wermuth, Cox and Marchetti (2006), Cox and Wermuth (1990).

For several discrete random variables of equal standing, i.e. without splits into response and explanatory variables, maximum-likelihood estimation was developed under different conditional independence constraints in a path-breaking paper by M. Birch (1963). This led to the formulation of general log-linear models, which were studied intensively among others by Haberman (1974), Bishop, Fienberg and Holland (1975), Sundberg (1975) and by L. Goodman, as summarized in a book of his main papers on this topic, published in 1978. His work was motivated mainly by research questions from the social and medical sciences.

For several Gaussian variables of equal standing, two different approaches to reducing the number of parameters in a model, were proposed at about the same time. T.W. Anderson put structure on the covariances, the moment parameters of a joint Gaussian distribution and called the resulting models, hypotheses linear in covariances; see Anderson (1973), while A.P. Dempster put structure on the canonical parameters with zero constraints on concentrations, the off-diagonal elements

of the inverse of a covariance matrix, and called the resulting models covariance selection models; see Dempster (1972).

Nowadays, log-linear models and covariance selection models are viewed as special cases of concentration graph models and zero constraints on the covariance matrix of a Gaussian distribution as special cases of covariance graph models. Covariance and concentration graph models are graphical Markov models with undirected graphs capturing independences. A missing edge means marginal independence in the former and conditional independence given all remaining variables in the latter; see also Wermuth and Lauritzen (1990), Wermuth and Cox (1998), (2004), Wermuth (2010).

The largest known class of Gaussian models that is in common to structural equation models and to graphical Markov models are the recursive linear equations with correlated residuals. These include linear summary graph models of Wermuth (2010), linear maximal ancestral graph of Richardson and Spirtes (2002), linear multivariate regression chains, and linear directed acyclic graph models. Deficiencies of some formulations start to be discovered by using algebraic methods. Identification is still an issue to be considered for recursive linear equations with correlated residuals, since so far only necessary or sufficient conditions are known but not both. Similarly, maximum-likelihood estimation still needs further exploration; see Drton, Eichler and Richardson (2009).

For several economic time series, it became possible to judge whether such fluctuating series develop nevertheless in parallel, that is whether they represent cointegrating variables because they have a common stochastic trend. Maximum-likelihood analysis for cointegrating variables, formulated by Johansen (1988, 2009), has led to many important applications and insights; see also Hendry and Nielsen (2007).

Algorithms and corresponding programs are essential for any widespread use of multivariate statistical methods and for successful analyses. In particular, iterative proportional fitting, formulated by Bishop (1964) for log-linear models, and studied further by Darroch and Ratcliff (1972), was adapted to concentration graph models for CG(conditional Gaussian)-distributions (Lauritzen and Wermuth, 1989) of mixed discrete and continuous variables by Frydenberg and Edwards (1989).

The EM(expectation-maximization)-algorithm of Dempster, Laird and Rubin (1977) was adapted to Gaussian directed acyclic graph models with latent variables by Kiiveri (1987) and to discrete concentration graph models with missing observation by Lauritzen (1995).

With the TM-algorithm of Edwards and Lauritzen (2001), studied further by Sundberg (2002), maximum-likelihood estimation became feasible for all chain graph models called blocked concentration chains in the case these are made up of CG(conditional Gaussian)-regressions (Lauritzen and Wermuth, 1989).

For multivariate regression chains of discrete random variables, maximum-likelihood estimation has now been related to the multivariate logistic link function by Marchetti and Lupparelli (2010), where these link functions provide a common framework and corresponding algorithm for generalized linear models, which

include among others linear, logistic and probit regressions as special cases; see McCullagh and Nelder (1989), Glonek and McCullagh (1995).

Even in linear models, estimation may become difficult when some of the explanatory variables are almost linear functions of others, that is if there is a problem of multicollinearity. This appears to be often the case in applications in chemistry and in the environmental sciences. Thus, in connection with consulting work for chemists, Hoerl and Kennard (1970) proposed the use of ridge-regression instead of linear least-squares regression. This means for regressions of vector variable Y on X, to add to $X^{\mathrm{T}}X$ some positive constant k along the diagonal before matrix inversion to give as estimator $\tilde{\beta} = (kI + X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$.

Both ridge-regression and partial-least-squares, proposed as an estimation method in the presence of latent variables by Wold (1980), have been recognized by Björkström and Sundberg (1999) to be shrinkage estimators and as such special cases of Tykhonov (1963) regularization.

More recently, a number of methods have been suggested which combine adaptive skrinkage methods with variable selection. A unifying approach which includes the least-squares estimator, shrinkage estimators and various combinations of variable selection and shrinkage has recently been given via a least squares approximation by Wang and Leng (2007). Estimation results depend necessarily on the chosen formulations and the criteria for shrinking dependences and for selecting variables.

Many more specialised algorithms and programs have been made available within the open access programming environment R, also those aiming to analyse large numbers of variables for only few observed individuals. It remains to be seen, whether important scientific insights will be gained by their use.

References

[1] Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* New York: Wiley. (2003) 3rd ed., New York: Wiley.

[2] Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1**, 135–141.

[3] Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc.* B **25**, 220–233.

[4] Bishop, Y.M.M. (1967). Multidimensional contingency tables: cell estimates. Ph.D. dissertation. Department of Statistics. Harvard University.

[5] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete multivariate analysis: theory and practice.* Cambridge, Mass.: MIT Press.

[6] Björkstrm̈, A. and Sundberg, R. (1999). A generalized view on continuum regression. *Scand. J. Statist.* **26**, 17–30.

[7] Cox, D.R. and Wermuth, N. (1990). An approximation to maximum-likelihood estimates in reduced models. *Biometrika* **77**, 747–761.

[8] Cox, D.R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Science* **8**, 204–218; 247–277.

[9] Cox, D.R. and Wermuth, N. (1996). *Multivariate dependencies: models, analysis, and interpretation.* London: Chapman and Hall.

[10] Darroch, J.N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Ann. Math. Statist.* **43**, 1470–1480.

[11] Dempster, A.P. (1972). Covariance selection. *Biometrics* **28**, 157–175.

[12] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* **39**, 1–38.

[13] Drton, M. (2009). Discrete chain graph models. *Bernoulli* **15**, 736–753.

[14] Drton, M. and Richardson, T.S. (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regression model. *Biometrika* **91**, 383–392.

[15] Drton, M., Eichler, M. and Richardson, T.S. (2009). Computing maximum likelihood estimates in recursive linear models. *J. Mach. Learn. Res.* **10**, 2329–2348.

[16] Duncan, O.D. (1966). Path analysis: sociological examples. *Amer. J. Sociol.* **72** 1–12.

[17] Duncan, O.D. (1975). *Introduction to structural equation models.* New York: Academic Press.

[18] Edwards, D. (2000). *Introduction to graphical modelling.* (2nd edition), New York: Springer.

[19] Edwards, D. and Lauritzen, S.L. (2001). The TM algorithm for maximising a conditional likelihood function. *Biometrika* **88**, 961–972.

[20] Frydenberg, M. and Edwards, D. (1989) A modified iterative proportional scaling algorithm for estimation in regular exponential families *Comput. Statist. Data Analy.* **8**, I143–153

[21] Frydenberg, M. and Lauritzen, S.L. (1989). Decomposition of maximum likelihood in mixed interaction models. *Biometrika* **76**, 539–555.

[22] Geiger, D., Verma, T.S. and Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks* **20**, 507–534.

[23] Glonek, G.F.V. and McCullagh, P. (1995). Multivariate logistic models. *J. Roy. Statist. Soc.* B, **57**, 533–546.

[24] Goldberger, A.S. (1971). Econometrics and psychometrics: a survey of communalities. *Psychometrika* **36**, 83–107.

[25] Goldberger, A.S. (1972). Structural equation methods in the social sciences. *Econometrica* **40**, 979–1002.

[26] Goodman, L.A. (1978). *Analyzing qualitative/categorical data.* Cambridge Mass.: Abt Books.

[27] Haberman S.J. (1974). *The analysis of frequency data.* Chicago: U Chicago Press.

[28] Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1–12. Reprinted in: Hendry, D.F. and Morgan, M.S. (eds.) (1995). *The foundations of econometric analysis.* Cambridge: Cambridge U Press. 477?490.

[29] Hendry, D.F. and Nielsen, B. (2007). *Econometric modeling: a likelihood approach.* Princeton, Princeton Univ. Press.

[30] Hoerl, A.E. and Kennard, R.N. (1970). Ridge regression. Biased estimation for non- orthogonal problems. *Technometrics* **12**, 55–67.

[31] Johansen, S. (1988). Statistical analysis of cointegration vectors. *J. Econ. Dyn. Contr.* **12**, 231–254. Reprinted in: Engle, R.F. and Granger, C.W.J. (Eds.) (1991). *Long-run economic relationships, readings in cointegration.* Oxford: Oxford Univ. Press (1991). 131–152.

[32] Johansen, S. (2009). Cointegration. Overview and development. In: Andersen, T.G., Davis, R., Kreiss, J-P. and Mikosch T. (Eds.) *Handbook of financial time series.* New York: Springer. 671–693.

[33] Jöreskog, K.G. (1973). A general method for estimating alinear structural equation system. In: Goldberger, A.S. and Duncan, O.D. *Structural equationmodels in the social sciences.* New York: Seminar Press. 85–112.

[34] Jöreskog, K.G. (1981). Analysis of covariance structures. *Scan. J. Statist.***8**, 65–92.

[35] Kiiveri, H.T. (1987). An incomplete data approach to the analysis of covariance structures. *Psychometrika* **52**, 539–554.

[36] Kline, R.B. (2010). *Principles and practice of structural equation modeling,* (3rd edition) New York: Guilford Press.

[37] Lauritzen, S. L. (1995). The EM-algorithm for graphical association models with missing data. *Comp. Statist. Data Anal.* **1**, 191–201.

[38] Lauritzen, S. L. (1996). *Graphical models.* Oxford: Oxford Univ. Press.

[39] Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**, 31–57.

[40] Marchetti, G.M. and Lupparelli, M. (2010). Chain graph models of multivariate regression type for categorical data. Submitted and available on ArXiv, http://arxiv.org/abs/0906.2098v2.

[41] Marchetti, G.M. and Wermuth, N. (2009). Matrix representations and independencies in directed acyclic graphs. *Ann. Statist.* **47**, 961–978.

[42] McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models,* 2nd edition, Boca Raton: Chapman & Hall/CRC.

[43] Richardson, T.S. and Spirtes, P. (2002). Ancestral Markov graphical models. *Ann. Statist.* **30**, 962–1030.

[44] Spearman, C. (1904). General intelligence, objectively determined and measured. *Amer. J. Psych.* **15**, 201–293.

[45] Spearman, C. (1926). *The Abilities of Man.* New York: Macmillan.

[46] Sundberg, R. (1975). Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests. *Scand. J. Statist.* **2**, 71–79.

[47] Sundberg, R. (2002). The convergence rate of the TM algorithm of Edwards and Lauritzen. *Biometrika* **89**, 478–483.

[48] Sundberg, R. (2010). Flat and multimodal likelihoods and model lack of fit in curved exponential families. *Scand. J. Statistics*, to appear.

[49] Tikhonov, A.N. (1963). Solution of ill-posed problems and the regularization method. (Russian) *Dokl. Akad. Nauk SSSR* **153**, 49–52.

[50] Wang, H. and Leng, C. (2007). Unified lasso estimation via least square approximation. *J. Amer. Statist. Assoc.* **102**, 1039–1048.

[51] Wermuth, N. (2010). Probability distributions with summary graph structure. Submitted and available on ArXiv, http://arxiv.org/abs/1003.3259.

[52] Wermuth, N. and Cox, D.R. (1998). On association models defined over independence graphs. *Bernoulli* **4**, 477–495.

[53] Wermuth, N. and Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems. *J. Roy. Statist. Soc. Ser. B* **66**, 687–717.

[54] Wermuth, N. and Lauritzen, S.L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discusssion). *J.Roy. Statist. Soc.* **B 52**, 21–75.

[55] Wermuth, N., Marchetti, G.M. and Cox, D.R. (2009). Triangular systems for symmetric binary variables. *Electr. J. Statist.* **3**, 932–955.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics.* Chichester: Wiley.

[56] Wold, H.O.A. (1954). Causality and econometrics. *Econometrica* **22**, 162–177.

[57] Wold, H.O.A. (1980). Model construction and evaluation when theoretical knowledge is scarce: theory and application of partial least squares. In: Kmenta, J. and Ramsey, J. (eds.) *Evaluation of econometric models.* New York: Academic Press, 47–74.

[58] Wright, S. (1923). The theory of path coefficients: a reply to Niles' criticism. *Genetics* **8**, 239–255.

[59] Wright, S. (1934). The method of pathcoefficients. *Ann. Math. Statist.* **5**, 161–215.

[60] Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57**, 348–368.