Metodološki zvezki, Vol. 1, No. 1, 2004, 119-130

Fitting Directed Graphical Gaussian Models with One Hidden Variable

Fulvia Pennoni[1]

Abstract

We discuss directed acyclic graph (DAG) models to represent the independence structure of linear Gaussian systems with continuous variables: such models can be interpreted as a set of recursive univariate regressions. Then we consider Gaussian models in which one of the variables is not observed and we show how the incomplete log-likelihood of the observed data can be maximized using the EM. As the EM algorithm does not provide the matrix of the second derivatives we show how to get an explicit formula for the observed information matrix. We illustrate the utility of the models with two examples.

1 Introduction

The analysis of multivariate data typically deals with complex association structures due to various direct and indirect relations among variables. The idea of graphical Markov models is to represent the independence structure of a multivariate random vector by a graph where the vertices correspond to variables and the absence of an edge between vertices stands for conditional or marginal independencies. In many applications some dependency structure between observed variables can be explained by supposing that their distribution arises after marginalizing over, and or conditioning on latent variables.

Such models are particularly of interest in the context in which one variable is not observed and some knowledge about the generating process of the data is available as for example for data collected in the social sciences. In such context appropriate estimation procedures have to be found to estimate the parameters of interest. We focus on maximum likelihood estimation of DAG models with one latent variable which can act as an intermediate, source or collision node. The estimation requires iterative solutions and thus appropriate algorithms.

The outline of the paper is as follows. In the first section we interpret a DAG for a Gaussian system as a set of recursive univariate regressions and we give some matrix notation. In Section 3 we show the observed data log-likelihood and briefly we discuss some identifiability problems. We also illustrate the steps of the EM algorithm for maximum likelihood estimation. Following Kiiveri (1987) we report

[1] University of Florence, Department of Statistics, "G. Parenti", Viale Morgagni, 59 - 50134 Florence, Italy; pennoni@ds.unifi.it

the explicit form for the second derivatives of observed log-likelihood and in the Appendix we show how to derive it. In Section 4 we give some examples of identifiable DAG's with one hidden variable using real data sets. Computations are carried out in `R` with the package `ggm` (Marchetti and Drton, 2003).

2 Gaussian directed acyclic graph models

Suppose $X = (X_1, X_2, ..., X_k)$ is a finite set of substantive variables of interest ordered in certain way, such that there exist a subset of indices $pa(i) \subseteq \{i+1, ..., k\}$, $i = 1, ..., k$, some independent random variables $\epsilon_1, \epsilon_2, ..., \epsilon_k$ and linear functions $f_1, f_2, ..., f_k$ such that

$$X_i = f_i(X_{pa(i)}, \epsilon_i), \qquad i = 1, ..., k \qquad [X_{pa(i)} \equiv \{X_j : j \in pa(i)\}].$$

The set of equations $X_i = f_i(X_{pa(i)}, \epsilon_i)$ prescribes a stepwise process for generating the distribution where a proper dependence of X_i is to be only on its potentially explanatory variables. The system is called *recursive* or a *univariate recursive regression system* or a *triangular system.*

This system can be represented by a directed acyclic graph (DAG) denoted by $G = (V, E)$ which consists of non empty finite set of vertices $V \equiv \{1, ..., k\}$ representing $X = (X_1, X_2, ..., X_k)$ and a set $E \subseteq V \times V$ of arrows $i \leftarrow j \in E$ iff $j \in pa(i)$ such that there are no direct path that start and end at the same variable. The multivariate distribution of X_v is called $G-Markovian$ if it fulfils the so called pairwise Markov property

$$X_a \perp\!\!\!\perp X_c | X_c \qquad \text{for all} \quad (a, b) \notin E, a \neq b.$$

For Gaussian distribution this is equivalent to the global Markov property (Lauritzen and Wermuth, 1989). An important property of a distribution satisfying the global directed Markov property associated with a DAG is that its joint density can be decomposed into conditional probabilities involving only variables and their parents according to the structure of the graph in the following way

$$p(x_1, ..., x_k) = \prod_{i=1}^{k} p(x_i | x_{pa(i)}).$$

Assuming that X is a vector of k mean centered random variables with Gaussian joint distribution with covariance matrix Σ, the recursive system can be written as

$$AX = \epsilon \qquad \text{cov}(\epsilon) = \Delta \tag{2.1}$$

where $A = \{-a_{rs}\}$ is upper triangular matrix with ones along the diagonals and with off-diagonal elements corresponding to partial regression coefficients between two variables given the parents, $-a_{rs} = \beta_{rs.pa(r)\setminus s}$ associated with a directed edge between $X_s \leftarrow X_r$; $\Delta = \text{cov}(\epsilon)$ is a nonsingular diagonal covariance matrix of the residuals with elements of partial variances $\delta_{rr} = \sigma_{rr.pa(r)}$ along the diagonal, representing the unexplained proportion of the variance of the dependent variable.

The arguments we are dealing with apply also to the very much broader family of problems that are called *quasi linear* (Cox and Wermuth, 1996). It means that any dependence present has a linear component and like linear least square regression equations in a multivariate normal framework, any curvature and higher-order interactions present are such that a vanishing linear least-squares regression coefficient implies that no dependence of substantive importance is present.

A triangular decomposition of the covariance matrix Σ and of the concentration matrix Σ^{-1} is given by

$$\mathrm{cov}(X) = \Sigma = (A^{-1})\Delta(A^{-1})', \qquad \Sigma^{-1} = A'\Delta^{-1}A.$$

Here we consider the estimation of the unknown parameter Σ or equivalently (A, Δ) of a directed graphical Gaussian model based on an n independent and identically distributed observations $X^{(k)} = (X^1, ..., X^k)$ from X, with zero average constructed from the series of deviates from the mean.

Since our model assumes a zero mean, the empirical covariance matrix is definite to be

$$S = \sum_i X^i X^{i'}/n \qquad i = 1, ..., k.$$

We assume $n \geq p$ such that S is positive definite with probability one. Note that the case where the model also includes an unknown mean vector μ can be treated by estimating μ by the empirical mean vector $\bar{X}$.

The density function of X can be expressed as

$$f(x) = (2\pi)^{np/2}|\Sigma|^{-n/2}exp\{-\frac{n}{2}\mathrm{tr}(\Sigma^-1)S\},$$

see e.g. Edwards (2000). Considered as a function of the unknown parameters for fixed data x it gives the likelihood function. The log-likelihood of the model, apart from an additive constant

$$l_X(\Sigma) = \frac{n}{2}[\log|\Sigma^{-1}| - \mathrm{tr}(\Sigma^{-1}S)], \tag{2.2}$$

has to be maximized respect to Σ.

It can be shown that

$$\hat{a}_{rs} = -\hat{\beta}_{rs.pa(r)\backslash s} \qquad \hat{\delta} = \hat{\sigma}_{rr.pa(r)}$$

are the maximum likelihood estimates of A and Δ defined by linear regression estimates in the independent equations.

3 Maximum likelihood estimation with one unobserved variable

Supposing that we observe only a subset $Y^p = (Y^1, ..., Y^p)$ of the variables. The complete data can be seen as $X = (Y, Z)$ where Y denotes the observed components of X and Z denotes the unobserved component. When this is the case the corresponding DAG contains an hidden node.

The relevant log-likelihood function based on the observed components can be written as follows

$$l_Y(\Sigma) = \frac{n}{2}[\log|\Sigma_{yy}^{-1}| - \mathrm{tr}(\Sigma_{yy}^{-1} S_{yy})], \tag{3.1}$$

where Σ_{yy} denotes the submatrix referring to Y of the conformably partitioned covariance matrix of X, and S_{yy} is the observed covariance matrix.

The problem of what can be learned from the distribution of the observed variables about the joint distribution specified by the DAG involves identifiability conditions. If A and Δ can be uniquely reconstructed from the covariance matrix of the observed variables the system is said to be globally identified. Stanghellini and Wermuth (2003) give sufficient conditions for identifiability of DAG Gaussian models with one hidden node. They are formulated in terms of the joint distribution of the variables and based on properties of some conditional independence graphs induced by the model (see e.g Pennoni, 2004). If the sample covariance matrix is positive definite and the DAG considered satisfy one of the given conditions the likelihood surface is unimodal and when fitting the corresponding model a unique global maximum can be achieved.

Maximum likelihood analysis can be conceptualized as maximum likelihood estimation in a multivariate normal model with missing data (Dempster *et al.*, 1977). Following Kiiveri (1987) who first suggested the procedure in a discussion on Jöreskog paper (1981), we describe the maximum likelihood method for fitting such DAGs using the EM algorithm (Dempster *et al.*,1977). This is an iterative algorithm and each cycle, which consists of an E step followed by and M step, increases the likelihood of the parameters. The E step calculates the expected sufficient statistics given the observed data and the current estimate of the parameters and the M step determines the conditional expectations of the sufficient statistics as if they were the observed. For an application of the EM algorithm to estimate the factor analysis model see Rubin (1982).

In the following we explicitly define the E and the M step of the algorithm and present a simple matrix expression for carrying out the computations.

The computations required are particularly straightforward: in the E-step we must compute $Q(\Sigma|\Sigma_r)$ the conditional expected value of the complete data log-likelihood to the observed data Y and a guessed initial value of complete data covariance matrix Σ_r

$$Q(\Sigma|\Sigma_r) = E[l(\Sigma, |Y_1, ..., Y_p, \Sigma_r)].$$

It can be shown that

$$Q(\Sigma|\Sigma_r) = \frac{n}{2}[\ln|\Sigma^{-1}| - \mathrm{tr}[(\Sigma^{-1})E(S|Y, \Sigma_r)]] \tag{3.2}$$

where the expected complete data covariance matrix given Y can be written as

$$E(S|Y, \Sigma_r) = \begin{pmatrix} S_{yy} & S_{yy}B' \\ . & BS_{yy}B' + (\sigma^{zz})^{-1} \end{pmatrix} = C(S_{yy}|\Sigma_r) \tag{3.3}$$

where the element of the concentration matrix Σ^{-1} corresponding to the missing data Z are noted σ^{zz} and where $B = -(\Sigma^{zz})^{-1}\Sigma^{zy} = \Sigma_{zy}\Sigma_{yy}^{-1}$ are the regression coefficients between Z and Y.

Therefore in the M-step we maximize $Q(\Sigma|\Sigma_r)$ as a function of Σ to produce an improved estimate Σ_{r+1}. This maximization is carried out by fitting the linear recursive regression equations specified by the DAG.

The generalized likelihood ratio test for directed graphical Gaussian models against the saturated model, the deviance at convergence is

$$D = n[\mathrm{tr}(S_{yy}\hat{\Sigma}^{-1}) - \ln|S_{yy}\hat{\Sigma}^{-1}| - m],$$

which has an asymptotic χ^2 distribution with $df = [m(m+1)/2 - m - k]$ degrees of freedom, where m is the number observed variables and k is the number of edges in the DAG.

The EM has the advantage of numerical stability which leads to a steady increase in the likelihood of the data. A negative feature is that it may require many iterations to converge, it is characterized by a slow convergence rate in a neighborhood of the optimal point. It is also sensible to the starting values and it is convenient to choose multiple random starting values. One major shortcoming is that the observed information matrix is not obtained as a by-product of the algorithm, which is useful to get an estimate of the precision of the estimated parameters to construct confidence intervals and to construct various tests of significance.

As illustrated above the EM finds the value of θ, where $\theta = (\theta_1, ..., \theta_h)$ is the vector of the unknown parameters, $\hat{\theta}$ that maximizes $l_Y(\theta)$, that is the maximum likelihood for θ based on the observed data Y. Following Dempster *et al.* (1977) the observed log-likelihood $l_Y(\theta)$ can be decomposed as

$$l_Y(\theta) = Q(\theta|\theta') - H(\theta|\theta')$$

which leads to a simple expression for the second derivative matrix of the observed log-likelihood derived in terms of the criterion function invoked by the EM algorithm. Minus the second derivative of the log-likelihood is made of two parts

$$-\frac{\partial^2 l_Y}{\partial\theta\partial\theta} = -\frac{\partial^2 Q(\theta|\theta')}{\partial\theta\partial\theta} - \Big(-\frac{\partial^2 H(\theta|\theta')}{\partial\theta\partial\theta}\Big)$$

where Q is as in (3.2) and H is the expected value of the conditional density of the complete data X given the observed data Y (Tanner, 1996). Referring to $-Q$ as the *complete information* and to $-H$ as the *missing information* it has the following appealing interpretation: the *observed information* is equal to the complete information minus the missing information due to the unobserved components, which has been called the "missing information principle" by Orchard and Woodbury (1972). A basic result due to Louis (1982) is that if the distribution of the complete data is in a regular exponential family $-\partial^2 H/\partial\theta\partial\theta = Var_{X|Y}(\partial l_X/\partial\theta)$ the second derivative of the log-likelihood of the observed data can be expressed entirely in terms of the complete data log-likelihood

$$-\frac{\partial^2 l_Y}{\partial\theta\partial\theta} \quad = \quad -E_{X|Y}\Big[\frac{\partial^2 l_X}{\partial\theta\partial\theta}\Big] - Var_{X|Y}\Big(\frac{\partial l_X}{\partial\theta}\Big) \tag{3.4}$$

the amount of information lost by observing only Y is determined by the conditional variance of the complete data log-likelihood given Y.

It is important to emphasize that the variance-covariance obtained is based on the first and second derivatives of the observed data log-likelihood and thus is guarantee to be inferentially valid only asymptotically.

Kiiveri (1982) calculated an explicit form for the above expression

$$\frac{\partial^2 l_Y}{\partial\theta_i\partial\theta_j} = \frac{1}{2}tr\Big(\Sigma^{ij}(\Sigma - C) - \Sigma^i\Sigma\Sigma^j\Sigma\Big) + \frac{1}{2}tr(\Sigma^i C\Sigma^j C - \Sigma^i\tilde{C}\Sigma^j\tilde{C})\Big) \tag{3.5}$$

where $C = C(S_{yy}|\Sigma)$; $\tilde{C} = [C(S_{yy}|\Sigma) - H]$, where $H = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{\sigma^{zz}} \end{pmatrix}$; and $\Sigma^{ij} = \partial\Sigma^{-1}/\partial\theta_i\partial\theta_j$ and $\Sigma^i = \partial\Sigma^{-1}/\partial\theta_i$.

In Appendix it is shown how to get such a result and also the explicit formulas for the second derivatives of the observed data log-likelihood for the adopted decomposition to Σ^{-1}.

4 Examples

We illustrate the fitting of the models described above on same examples. The computations were carried out using the `R` package `ggm` (Marchetti and Drton, 2003). We implemented the `R` code to add to the existing routines of such package to compute the standard errors for the estimated parameters (cf. Pennoni, 2004).

Example 1: Criminological research. Let us consider an example from criminological research described by Smith and Patterson (1984). Random samples of persons in sixty residential neighborhoods were interviewed regarding victimization experiences, neighborhood safety and evaluation of police performance. The sample was 1500 people living alone. The seven variables observed were as follows:
- Y_4 number of self reported prior victimizations in the last twelve months,
- Y_5 respondent's age,
- Y_6 respondent sex,
- Y_7 the rate of personal and property victimization per 100 households in the respondent's neighborhood.

The following variables were responses to three questions asking respondents how likely they thought it was that they would be victims of
- Y_3 vandalism during the next year
- Y_2 burglary
- Y_1 robbery.

Response categories on these items ranged from "not at all likely' to "very likely'. The proposed model was that variables Y_1, Y_2, Y_3 acted as indicators of a latent variable named *perceived risk of victimization*.

The estimated residual variances for Y_1, Y_2, Y_3 and Z are shown in Table 1 with their standard errors and z-values.

Considering a system of linear equations represented with the DAG in Figure 1, where Z is the latent variable we want to estimate the relevant regression coefficients

Table 1: Estimated partial residual variances for the model in example 1.

	δ	s.e	z
$\delta_{1.Z}$	0.4691	(0.0303)	15.4818
$\delta_{2.Z}$	0.3611	(0.0132)	27.3560
$\delta_{3.Z}$	0.4390	(0.0161)	27.2671

Table 2: Estimated regression coefficients for the model in the example 1.

Arrows	Estimates	s.e	z
$Y_1 \leftarrow Z$	0.6805	0.0486	14.0021
$Y_2 \leftarrow Z$	0.7350	0.0503	14.6123
$Y_3 \leftarrow Z$	0.6887	0.0480	14.3479
$Z \leftarrow Y_4$	0.2541	0.0303	8.3861
$Z \leftarrow Y_5$	-0.1343	0.0305	-4.4033
$Z \leftarrow Y_6$	-0.0168	0.0305	-0.5508
$Z \leftarrow Y_7$	0.2474	0.0299	8.2742
$Y_1 \leftarrow Y_5$	0.1165	0.0200	5.8250
Deviance	13.76	*df* 7	$p < 0.06$

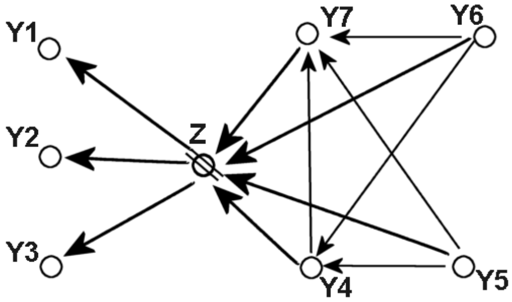

Figure 1: DAG for Criminological example.

involving Z; the arrows between Y_4, Y_5, Y_6 and Y_7 are considered as dependencies of some interest and they do not imply causal relationship. The dag satisfies the conditions for global identifiability which can be checked in ggm, then the model is globally identified and the parameters can be uniquely estimated up to the sign of the coefficients involving the latent variable. Fitting this model with the residual variance of the latent variable constrained to be one we get a deviance of 43.63 on 8 degrees of freedom. It can be seen from Table 2 that a significant fit can be achieved by adding a direct edge from age (Y_5) to the perceived risk of robbery (Y_1). The results from the new model are similar to those of the previous model with the addition of a positive effect of the respondent's age on the robbery indicator of perceived risk as displayed in Table 2. It can be seen the non significant z-statistic for the regression coefficient of sex at the 5% level; it appears that prior victimization (Y_4) and victimization rate (Y_7) have the greatest effect on the latent variable.

Table 3 gives the observed correlation matrix (upper triangular) and the estimated correlation matrix (lower triangular) for the last models.

Table 3: Observed (upper diagonal) and Estimated (lower diagonal) covariance matrix.

Y_1	–	0.575	0.540	0.169	-0.014	-0.023	0.224	
Y_2	0.575	–	0.598	0.240	−0.144	−0.088	0.215	
Y_3	0.539	0.599	–	0.246	−0.128	−0.092	0.182	
Y_4	0.198	0.237	0.222	–	−0.184	−0.148	0.168	
Y_5	−0.014	−0.141	−0.132	−0.184	–	0.236	−0.027	
Y_6	−0.048	−0.082	−0.077	−0.148	0.236	–	−0.102	
Y_7	0.198	0.217	0.203	0.168	−0.027	−0.102	–	
Z	0.783	0.869	0.814	0.323	−0.191	−0.111	0.295	1.182
	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Z

Example 2: Measurement problems. As a further illustration of fitting a DAG Gaussian model an example from Sewel, Haller & Ohlendorf (1970) and Wiley (1973) is be used. Most measurements used in behavioral and social sciences contains sizeable measurements errors which if not taken into account can cause several bias in results. The following example illustrates the problems with measurement errors in observed variables. The purpose is to describe how well the observed indicators serve as measurement instruments for the latent variable. A sample of 3500 was recorded on the following items:
-MA Mental ability,
-SES Socio-economic status,
-AP Academic performance,
-SO Significant others'influence,
-EA Educational aspiration.
The postulated model includes a measurement error in SO, the graphical model in Figure 2.

The correlation matrix between the observed variables is shown in Table 4.

The model is globally identified because it satisfy some of the sufficient conditions given in Stanghellini and Wermuth (2003). Fitting the model with Z scaled to have unit variance we get results shown in Table 5, where the standard errors and the Wald test are also reported.

Table 4: Correlation matrix for variables in Educational example.

MA	1.00				
SES	.288	1.00			
AP	.589	.194	1.00		
SO	.438	.359	.473	1.00	
EA	.418	.380	.459	.611	1.00

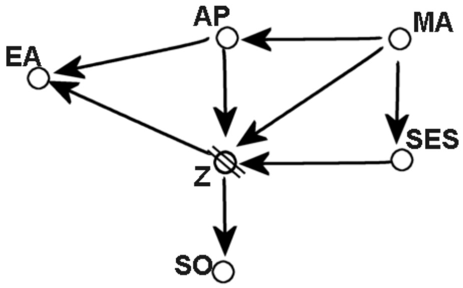

Figure 2: Graphical model for Educational example.

Table 5: Estimates for Educational example.

	Estimates	s.e	Z
$EA \leftarrow Z$	0.556	0.009	61.682
$EA \leftarrow AP$	0.035	0.013	2.660
$Z \leftarrow AP$	0.617	0.021	29.468
$Z \leftarrow MA$	0.303	0.021	14.155
$Z \leftarrow SES$	0.479	0.018	27.100
$SO \leftarrow Z$	0.532	0.007	72.602
$SES \leftarrow MA$	-0.288	0.016	-17.192
$AP \leftarrow MA$	-0.589	0.014	-43.100
Deviance	7.1459	*df* 2	

Table 6: Estimated partial residual variance in Educational model.

$\delta_{EA.AP,Z}$	$\delta_{SO.Z}$	$\delta_{Z.AP,MA,SES}$	$\delta_{AP.MA}$	$\delta_{SES.MA}$	δ_{MA}
0.378	0.399	0.474	0.653	0.917	1.00

The estimated residual variances are in Table 6. It can be seen that the residual variance for SO is 0.399 it means that the reliability of SO is only 0.601.

Acknowledgements

I am grateful to Giovanni M. Marchetti for many fruitful comments and suggestions.
I also thank Harri Kiiveri to provide his Ph.D thesis and many advices.
This work was supported by MIUR Italy, grant scheme PRIN 2002.

References

[1] Cox, D.R. and Wermuth, N. (1996): *Multivariate Dependencies - Models, Analysis and interpretation.* London: Chapman & Hall.

[2] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1-38.

[3] Edwards, D. (2000): *Introduction to graphical modelling.* Second-Edition. New York: Springer-Verlag.

[4] Jöreskog, K.G. (1981): Analysis of covariance structures. With discussion. *Scandinavian Journal of Statistics*, **8**, 65-92.

[5] Kiiveri, H.T. (1982): A unified approach to causal models. Ph.D thesis. University of Western Australia.

[6] Kiiveri, H.T. (1987): An incomplete data approach to the analysis of covariance structure. *Psychometrika*, **52**, 539-554.

[7] Lauritzen, S.L. (1996): *Graphical Models.* Oxford: Clarendon Press.

[8] Lauritzen, S.L. and Wermuth, N. (1989): Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, **17**, 31-57.

[9] Little, R.J.A. and Rubin, D.B. (2002): *Statistical Analysis with missing data.* 2nd Edition, Wiley.

[10] Louis, T.A. (1982): Finding the observed information matrix when using the EM-algorithm. *Journal of the Royal Statistical Society*, **44**, 226-233.

[11] Marchetti, G.M. and Drton, M. (2003): `ggm`: an `R` pachakage for Gaussian graphical models, URL: `http://cran.r-project.org/`.

[12] Orchard, T. and Woodbury, M.A. (1972): A missing information principle: Theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 697-715.

[13] Pearl, J. (1988): *Probabilistic reasoning in Intelligent Systems.* S. Mateo, CA: Morgan Kaufmann.

[14] Pennoni, F. (2004): Issues on the estimation of Latent Variable and Latent Class models with Social Science Applications. *PhD Thesis*, Department of Statistics, University of Florence, URL: `http://www.ds.unifi.it/pennoni`.

[15] Rubin, D.B. and Thayer, D.T. (1982): EM Algorithms for Maximum Likelihood Factor Analysis. *Psychometrika*, **47**, 69-76.

[16] Smith, D.A. and Patterson, E.B. (1984): Application and generalization of mimic models to criminological research. *Journal of Research in Crime and Deliquency*, **21**, 333-352.

[17] Sewell, W.H., Haller, A.O., and Ohlendorf, G.W. (1970): The educational and early occupational status attainment process: revisions and replications. *American Sociological Review*, **35**, 1014-1027.

[18] Stanghellini, E. and Wermuth, N. (2003): On the identification of path-analysis models with one hidden variable. *Accepted for publication in Biometrika.*

[19] Tanner, M.A. (1996): *Tools for Statistical Inference.* New York: Springer.

[20] Wiley, D.E. (1973): The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger and O. D. Duncan (Eds.): *Structural Equation Models in the Social Sciences.* New York: Academic Press, 69-83.

Appendix

To simplify the notation we write l_X for $l_X(\Sigma)/n$ and l_Y for $l_Y(\Sigma)/n$. The first and the second derivatives of (3.2) when Σ is a function of a vector of parameters θ are

$$\frac{\partial l_X}{\partial \theta} = \frac{1}{2}\mathrm{tr}(\Sigma^i(\Sigma - S))$$

$$\frac{\partial^2 l_X}{\partial \theta_i \partial \theta_j} = -\frac{1}{2}\mathrm{tr}(\Sigma^{ij}(\Sigma - S) - \Sigma^i \Sigma \Sigma^j \Sigma)$$

where $\Sigma^i = \partial \Sigma^{-1}/\partial \theta_i$ and $\Sigma^{ij} = \partial \Sigma^{-1}/\partial \theta_i \partial \theta_j$. For the parametrization considered $\Sigma^{-1} = A'\Delta^{-1}A$ the explicit derivatives have the form

$$\frac{\partial l_X}{\partial A_{ji}} = [(\Sigma - S)A\Delta^{-1}]_{ji}; \qquad \frac{\partial l_X}{\partial \Delta_{ii}} = \frac{1}{2}[A(\Sigma - S)A']_{ii};$$

$$\frac{\partial^2 l_X}{\partial A_{ji}\partial A_{ml}} = -(A^{lj}A^{im} + S_{li}\Delta^{-1}_{mj}); \qquad \frac{\partial^2 l_X}{\partial \Delta_{ii}\partial \Delta_{ll}} = -\frac{1}{4}(\Delta^{il}\Delta^{il} + \Delta^{il}\Delta^{il})$$

where A^{ji} denotes the $(j,i)th$ element of A^{-1} and Δ^{il} is the $(i,l)th$ element of Δ^{-1}.

To get the derivatives of the incomplete log-likelihood l_Y as in (3.1) Dempster *et al.* (1977) showed that

$$\frac{\partial l_Y}{\partial \theta} = E_{X|Y}\Big(\frac{\partial l_X}{\partial \theta}\Big)$$

the observed score is equal to the expected score of the complete data log-likelihood conditioned on the observed data. This expression becomes

$$\frac{\partial l_Y}{\partial \theta_i} = \frac{1}{2}\mathrm{tr}\Big(\Sigma^i(\Sigma - C)\Big)$$

where $C = C(S_{yy}|\Sigma)$ is defined in (3.3). The first part of the right hand side of (3.4) is minus the conditional expected value of second derivative of (3.2)

$$-E_{X|Y}\Big[\frac{\partial^2 l_X}{\partial\theta_i\partial\theta_j}\Big] = -\frac{1}{2}\mathrm{tr}(\Sigma^{ij}(\Sigma - C) - \Sigma^i\Sigma\Sigma^j\Sigma).$$

The second part of the right hand side of (3.4) can be written

$$-Var_{X|Y}\Big(\frac{\partial l_X}{\partial\theta}, \frac{\partial l_X}{\partial\theta'}\Big) = -E_{X|Y}\Big\{\Big[\frac{\partial l_X}{\partial\theta} - E_{X|Y}\Big(\frac{\partial l_X}{\partial\theta}\Big)\Big]\Big[\frac{\partial l_X}{\partial\theta} - E_{X|Y}\Big(\frac{\partial l_X}{\partial\theta}\Big)\Big]'\Big\} =$$

$$= -E_{X|Y}\Big\{\frac{1}{2}\mathrm{tr}\Big[(\Sigma^i(C - S))(\Sigma^i(C - S))'\Big]\Big\} = -\frac{1}{2}\mathrm{tr}\Big(\Sigma^i C\Sigma^j C - \Sigma^i\tilde{C}\Sigma^j\tilde{C}\Big).$$

So (3.4) is established. In the parameterizations (A, Δ) we get the second derivatives

$$\frac{\partial^2 l_Y}{\partial A_{ji}\partial A_{ml}} = -(A^{lj}A^{im} + C_{li}\Delta^{-1}_{mj}) + C_{li}[\Delta^{-1}A'CA\Delta^{-1}]_{mj} +$$

$$+[CA\Delta^{-1}]_{mi}[\Delta^{-1}A'C]_{lj} - \tilde{C}_{li}[\Delta^{-1}A'\tilde{C}A\Delta^{-1}]_{mj} - [\tilde{C}A\Delta^{-1}]_{mi}[\Delta^{-1}A'\tilde{C}]_{lj}$$

And the derivatives respect to Δ^{-1}

$$\frac{\partial^2 l_Y}{\partial\Delta_{ii}\partial\Delta_{ll}} = -\frac{1}{4}\Big(\Delta^{il}\Delta^{il} + \Delta^{il}\Delta^{il}\Big) + \frac{1}{4}\Big\{[A'CA]_{il}[A'CA]_{il}$$

$$+[A'CA]_{il}[A'CA]_{il} - [A'\tilde{C}A]_{il}[A'\tilde{C}A]_{il} - [A'\tilde{C}A]_{il}[A'\tilde{C}A]_{il}\Big\}.$$