Computational Statistics & Data Analysis 19 (1995) 191–201
North-Holland

The EM algorithm for graphical association models with missing data

Steffen L. Lauritzen
Aalborg University, Aalborg, Denmark

Received March 1992
Revised July 1992

Abstract: It is shown how the computational scheme of Lauritzen and Spiegelhalter (1988) can be exploited to perform the E-step of the EM algorithm when applied to finding maximum likelihood estimates or penalized maximum likelihood estimates in hierarchical log-linear models and recursive models for contingency tables with missing data. The generalization to mixed association models introduced in Lauritzen and Wermuth (1989) and Edwards (1990) is indicated.

Keywords: Contingency tables; Belief networks; Decomposable models; Expert systems; Hierarchical models; Latent variables; Probability propagation; Recursive models

1. Introduction

The EM algorithm has been object of considerable interest since the paper of Dempster *et al.* (1977). It has proved a flexible tool for calculating maximum likelihood estimates in a variety of problems involving missing data or incomplete information in some sense.

In models with latent structure (Lazarsfeld and Henry, 1968) data are missing systematically. The algorithm was used in a special case of this by Dawid and Skene (1979).

In connection with log-linear models for contingency tables, the algorithm was studied by Fuchs (1982) extending work of Chen and Fienberg (1974, 1976) and Hocking and Oxspring (1974). Fuchs (1982) seems to consider the E-step of the algorithm as uninteresting, writing on p. 272 that "Similar formulas can be easily written for any log-linear model...". However, even though this is correct in principle, the computational effort involved in the E-step can be considerable, and efficient methods are needed, since this computation must be repeated many times. Some gain of efficiency is obtained by exploiting collapsibility

Correspondence to: S.L. Lauritzen, Institute for Electronic Systems, Aalborg University, Department of Mathematics and Computer Science, Fredrik Bajers vej 7, DK-9220, Aalborg Ø, Denmark.

SSDI 0167-9473(93)E0056-A

(Geng and Asano, 1988). In the present note it is shown how to exploit the computational scheme of Lauritzen and Spiegelhalter (1988) to perform the E-step of the algorithm in log-linear models. Later we show how this extends to recursive models for contingency tables. The methods extend and complement some of those developed by Lander and Green (1987) in a genetical context.

Section 2 reviews notation and basic estimation theory for log-linear models. Section 3 describes the EM algorithm as it appears in this problem. Section 4 gives a brief description of the computational procedure of Lauritzen and Spiegelhalter, and how it can be exploited. Section 5 discusses the algorithm applied to recursive models. Finally Section 6 indicates how the computations can be performed in the case of mixed graphical interaction models with both discrete and continuous variables.

2. Hierarchical log-linear models

We consider log-linear models for contingency tables in the case of multinomial sampling and use the notation from Darroch *et al.* (1980) or Lauritzen (1989).

Hence Δ is a finite set of *criteria* or *variables* with possible *level sets* I_δ, $\delta \in \Delta$. The set of *cells* of the table is the product $I = \times_{\delta \in \Delta} I_\delta$. In the case of complete observation, the basic data are the set $(n(i))_{i \in I}$ of counts which follow a multinomial distribution:

$$P\{N(i) = n(i),\ i \in I\} = \binom{N}{n(i),\ i \in I} \prod_{i \in I} p(i)^{n(i)}.$$

This is a consequence of the counts being obtained by adding up over N independent *cases* $i^1, \ldots, i^N$ such that

$$n(i) = \sum_{\nu=1}^{N} \chi^\nu(i), \qquad \text{where}$$

$$\chi^\nu(i) = \begin{cases} 1 & \text{if } i^\nu = i \\ 0 & \text{otherwise.} \end{cases}$$

A *hierarchical log-linear* model is specified by a *generating class* A of subsets of Δ, representing the restriction that

$$\log p \in \sum_{a \in A} F_a, \tag{1}$$

where F_a are the *factor subspaces* of functions that only depend on coordinates in a

$$x \in F_a \Leftrightarrow x(i_a) = x(j_a) \text{ whenever } i_a = j_a.$$

Here $i_a = (i_\delta)_{\delta \in a}$ is an element of the set of *a-marginal cells* $I_a = \times_{\delta \in a} I_\delta$. See for example Darroch and Speed (1983) for further details.

A hierarchical log-linear model is a regular exponential family in the sense of Barndorff-Nielsen (1978). Therefore the maximum likelihood estimates, in the

case of complete data, are obtained by equating the observed value of the sufficient statistic to its expectation, leading to the equations

$$Np(i_a) = n(i_a), \qquad a \in A,\ i_a \in I_a, \tag{2}$$

where $n(i_a)$ are the *marginal counts*, $n(i_a) = \Sigma_{j:\, j_a = i_a} n(j)$, see, for example, Andersen (1974) or Haberman (1974a).

The system of equations (2) does not always possess a solution within the class of probabilities specified, but if the model is extended to include weak limits of probabilities satisfying (1), it does (Lauritzen, 1989). In general, iterative algorithms such as iterative proportional scaling (Darroch and Ratcliff, 1972) have to be used, but in the case where A is *decomposable*, (2) has the explicit solution

$$\hat{p}(i) = \frac{\Pi_{a \in A} n(i_a)}{\Pi_{s \in S} n(i_s)^{\nu(s)}}. \tag{3}$$

where S is the set of *separators* $(s_j,\ j = 2, \ldots, n)$ obtained when organizing the elements of A in a sequence $(a_1, \ldots, a_n)$ having the *running intersection property*

$$s_j = a_j \cap (a_1 \cup \cdots \cup a_{j-1}) \subset a_i \text{ for some } i < j,$$

each separator s appearing $\nu(s)$ times in the sequence (Haberman, 1974a; Andersen, 1974; Darroch *et al.*, 1980).

3. The EM algorithm

The EM algorithm in the form that we need it in the present paper, is based on forming the conditional expectation of the log-likelihood function for complete data, given the observed data

$$Q(\theta' \mid \theta) = E_\theta\{\log f(X \mid \theta') \mid y\}, \tag{4}$$

where X is the random variable corresponding to the complete (unobserved) data having density f, whereas $y = g(x)$ is the observed data. When θ is fixed, the process of determining this expectation as a function of θ' is referred to as the E-step. The algorithm then alternates between the E-step and the M-step, which maximizes Q in θ'. The algorithm has generalizations called GEM algorithms which appear by not necessarily maximizing Q but only finding a value of θ' that makes $Q(\theta' \mid \theta)$ strictly increase over $Q(\theta \mid \theta)$. It is also possible to add a penalty to the log-likelihood function, calculating instead

$$Q^*(\theta' \mid \theta) = Q(\theta' \mid \theta) - J(\theta'),$$

at the E-step, where $J(\theta)$ is a penalty, for example obtained from a prior density proportional to $\exp\{-J(\theta)\}$ (Green, 1990). This leads to maximization of the penalized log-likelihood function $\log L(\theta) - J(\theta)$. In the case of a log-linear model, the log-likelihood function for the complete data is a linear function of the set of sufficient marginals

$$n(i_a), \qquad a \in A,\ i_a \in I_a.$$

Therefore the E-step is equivalent to calculating the expected marginal counts

$$n^*(i_a) = E_p\{N(i_a) \mid \text{observed data}\}.$$

Similarly the M-step can be identified with solving

$$np(i_a) = n^*(i_a), \qquad a \in A,\ i_a \in I_a, \tag{5}$$

for p, which maximizes the likelihood function, assuming the expected counts were the true counts. The equation (5) was found in the exponential family case by Martin-Löf (1966), and established by Sundberg (1972, 1974). For the log-linear model it can be found in Haberman (1974b) just in a somewhat different notation. Sundberg (1976) also exploited the results to construct the EM algorithm for exponential families and study its properties.

The M-step is computationally equivalent to solving the likelihood equations (2). Here we show how to perform the E-step.

Assume that we have N independent observations $i^1_{b^1}, i^2_{b^2}, \ldots, i^N_{b^N}$, such that for case υ we have only observed the value of variables in the set b^υ. Then we find

$$\begin{aligned} n^*(i_a) &= \sum_{\upsilon=1}^{N} E_p\{\chi^\upsilon(i_a) \mid i^1_{b^1}, \ldots, i^N_{b^N}\} \\ &= \sum_{\upsilon=1}^{N} E_p\{\chi^\upsilon(i_a) \mid i^\upsilon_{b^\upsilon}\} \\ &= \sum_{\upsilon=1}^{N} p(i_a \mid i^\upsilon_{b^\upsilon}), \end{aligned} \tag{6}$$

where we have let

$$\chi^\upsilon(i_a) = \begin{cases} 1 & \text{if } i^\upsilon_a = i_a \\ 0 & \text{otherwise.} \end{cases}$$

What remains to be observed is that the Lauritzen-Spiegelhalter procedure for probability propagation is an efficient method of calculating the individual terms in (6). We describe this in more detail in the next section. For computational efficiency one would of course collect cases with identical observations in groups and only calculate $p(i_a \mid i_b)$ once for each of these identical terms.

4. Probability propagation

This section describes briefly the procedure of Lauritzen and Spiegelhalter (1988), essentially in the form given by Jensen *et al.* (1990) and implemented in the program HUGIN (Andersen *et al.*, 1989). The reader is referred to these references, Dawid (1992), or Spiegelhalter *et al.* (1993) for further details. Only the initialization differs since we begin with a log-linear representation instead of conditional probability tables.

First the *2-section graph* of the generating class A is formed, being the undirected graph with the elements of Δ as vertices and edges between pairs $\{\delta_1, \delta_2\} \subseteq \Delta$ having $\{\delta_1, \delta_2\} \subseteq a$ for some $a \in A$.

Then this graph has to be triangulated, *i.e.* edges are added until all cycles of length four or more possess chords. Kjærulff (1990, 1992) investigates good heuristic algorithms for doing this.

The set C of cliques (maximal complete subsets) of the triangulated graph is arranged in a *junction tree* of belief universes having the property that the intersection $c \cap d$ of any two cliques (universes) are subsets of all cliques on the path between c and d in the tree. This is essentially equivalent to arranging the cliques in a sequence with running intersection property.

From the log-linear representation of p, a factorization of the type

$$p(i) = \frac{\prod_{c \in C} \psi_c(i_c)}{\prod_{s \in S} \psi_s(i_s)}, \tag{7}$$

is obtained, where S is the list of *separators*, being identical to the intersections of pairs of neighbouring universes in the junction tree. The 'same' separator set may appear several times in S, but the functions ψ_s, may be different, compare with (3).

The functions ψ_u, $u \in C \cup S$, are stored as *potential tables* associated with each universe/separator.

Evidence is entered to the belief universes in the way that if i_δ has been observed for a given case, then for some $c \in C$ containing δ, the function ψ_c is multiplied with the corresponding indicator function to obtain $\tilde{\psi}_c$ where then

$$\tilde{\psi}_c(i_c) = \begin{cases} \psi_c(i_c) & \text{if } i_\delta = i_\delta^v \\ 0 & \text{otherwise.} \end{cases}$$

This is done for all variables $\delta \in b^v$, where b^v is the set of variables observed at case v.

The calculations are performed via a message passing scheme. The basic operation is that a universe a *absorbs* information from a neighbour b as follows

$$\tilde{\psi}_a(i) \leftarrow \tilde{\psi}_a(i) \frac{\sum_{j_b:\, j_{b\cap a} = i_{b\cap a}} \tilde{\psi}_b(j_b)}{\tilde{\psi}_{a\cap b}(i_{a\cap b})}$$

$$\tilde{\psi}_{a\cap b}(i_{a\cap b}) \leftarrow \sum_{j_b:\, j_{b\cap a} = i_{b\cap a}} \tilde{\psi}_b(j_b).$$

Thus the potential functions associated with the absorbing universe and the separator change. The expression (7) clearly remains invariant under the operation.

The message passing scheme involves two passes. First a universe r is chosen as the root of the tree and r *collects evidence* by asking each of its neighbours for

a message. Before these neighbours send, they in turn ask their neighbours *etc.* until the requests reach the leaves of the tree. Messages are then sent towards the root r.

When the root has received its messages it *distributes evidence* by sending to all its neighbours who again send to their neighbours and so on. When the distributed evidence reaches the leaves of the tree, the process terminates and all universes hold modified potentials $\tilde{\psi}_u$ that satisfy

$$\tilde{\psi}_u(i_u) \propto p(i_u \mid i_{b^v}^v) \quad \text{for all} \quad u \in C \cup S,$$

i.e. these are proportional to the conditional probabilities given the evidence. The normalizing constant is equal to the probability of the evidence

$$z^v = \sum_{i_u \in I_u} \tilde{\psi}_u(i_u) = p(i_{b^v}^v), \tag{8}$$

independently of u. For a proof of these assertions see Jensen *et al.* (1990).

Since all $a \in A$ are subsets of some $c \in C$, the terms in (6) can now be simply calculated as

$$p(i_a \mid i_{b^v}^v) = \sum_{j_c:\, j_a = i_a} \tilde{\psi}_c(j_c)/z^v,$$

for a suitably chosen c.

From (8) it also follows that the log-likelihood function can be obtained basically without computation through

$$\log L(p) = \sum_{v=1}^{N} \log p(i_{b^v}^v) = \sum_{v=1}^{N} \log z^v. \tag{9}$$

This is useful for monitoring the behaviour of the EM algorithm.

In the case of a decomposable hierarchical model we have $A = C$ and the M-step is trivial since it follows from (3) that we as potentials in the next iteration can use

$$\psi_u(i_u) = n^*(i_u)/N \quad \text{for} \quad u \in C \cup S.$$

In the general case, the next values of $\psi_u(i_u)$ must be iteratively calculated from $n^*(i_u)$, $u \in A$, using for example iterative proportional scaling as mentioned in the previous section. Note that it is not necessary to do a full iterative proportional scaling. By only using one cycle of proportional scaling, the expected log-likelihood increases strictly, leading to a GEM algorithm.

5. Recursive models

It is tempting to use the technique to estimate conditional probabilities in the recursive graphical models of Wermuth and Lauritzen (1983), in particular since these are used for constructing probabilistic expert systems (Pearl 1988; Andreassen *et al.* 1989). In the expert system literature such Markov probabilities

are known as *influence diagrams*, *Bayesian belief networks*, *causal probabilistic networks* or similar terms (Oliver and Smith, 1990; Shafer and Pearl, 1990; Spiegelhalter *et al.*, 1993).

A *recursive graphical model* is specified by the unknown probability distribution p belonging to the set of distributions that obey the Markov property with respect to a directed acyclic graph (Kiiveri *et al.*, 1984; Lauritzen *et al.*, 1990). This is equivalent to assuming the existence of a factorization of p into conditional probabilities as

$$p(i) = \prod_{\delta \in \Delta} p\left(i_\delta \mid i_{\mathrm{pa}(\delta)}\right), \tag{10}$$

where $\mathrm{pa}(\delta)$ denotes the set of parents of δ.

If the recursive model has links between all parents that are common to a node, the model is equivalent to a decomposable log-linear model (Wermuth and Lauritzen, 1983), and the procedure previously described applies directly. A special case of this is when the directed graph is a *causal tree*. *i.e.* it contains no loops, in which case the calculations are particularly simple, as exploited in this context by Golmard and Mallet (1991).

To identify the E-step, we obtain from (10) a factorization of the likelihood function for complete data as

$$\begin{aligned} L(p) &\propto \prod_{i \in I} \left\{ \prod_{\delta \in \Delta} p\left(i_\delta \mid i_{\mathrm{pa}(\delta)}\right) \right\}^{n(i)} \\ &= \prod_{\delta \in \Delta} \prod_{i_{\mathrm{cl}(\delta)} \in I_{\mathrm{cl}(\delta)}} p\left(i_\delta \mid i_{\mathrm{pa}(\delta)}\right)^{n i_{\mathrm{cl}(\delta)}} = \prod_{\delta \in \delta} L_\delta(p), \end{aligned} \tag{11}$$

where $\mathrm{cl}(\delta) = \delta \cup \mathrm{pa}(\delta)$. We then take conditional expectation of the log-likelihood function and get, using (4)

$$Q(p' \mid p) = \sum_{\delta \in \delta} \sum_{i_{\mathrm{cl}(\delta)} \in I_{\mathrm{cl}(\delta)}} n^*\left(i_{\mathrm{cl}(\delta)}\right) \log\, p'\left(i_\delta \mid i_{\mathrm{pa}(\delta)}\right).$$

Thus the E-step can be performed exactly as before with the only difference that the junction tree is initialized with potential functions from the conditional probability tables as described by Jensen *et al.* (1990).

To identify the M-step, note that the expression for Q is identical to the log-likelihood just with observed data replaced by estimated. The factorization (11) displays the likelihood function as a product of likelihood functions L_δ.

Since the model sets no further restrictions on the conditional probabilities, the joint likelihood function can be maximized by maximizing each of the factors. We therefore obtain that the M-step lets

$$p\left(i_v \mid i_{\mathrm{pa}(v)}\right) = n^*\left(i_{\mathrm{cl}(v)}\right) / n^*\left(i_{\mathrm{pa}(v)}\right). \tag{12}$$

In this formula $n(i_0) = n$. This will appear in the denominator whenever a variable δ has no parents. As in the case of log-linear models, the likelihood function is monitored using (9).

Assuming the conditional probabilities $p(\cdot \mid i_{\mathrm{pa}(v)})$ to be independent and Dirichlet distributed with parameters $\alpha(i_{\mathrm{cl}(v)})$ leads to the penalty

$$-J(p) = \sum_{\delta \in \Delta} \sum_{i_{\mathrm{cl}(\delta)} \in I_{\mathrm{cl}(\delta)}} \alpha(i_{\mathrm{cl}(\delta)}) \log p(i_\delta \mid i_{\mathrm{pa}(\delta)}).$$

The penalized likelihood has maximum in the posterior mode which can be found iteratively by replacing (12) with

$$p(i_v \mid i_{\mathrm{pa}(v)}) = \frac{n^*(i_{\mathrm{cl}(v)}) + \alpha(i_{\mathrm{cl}(v)}) - 1}{n^*(i_{\mathrm{pa}(v)}) + \alpha(i_{\mathrm{pa}(v)}) - |I_v|},$$

provided this remains positive. Here $\alpha(i_{\mathrm{cl}(v)})$ can be interpreted as prior counts. If these counts are less than 1, the posterior distribution may not have a mode in the interior and the above expression can turn negative. Hence it seems more suitable to penalize the likelihood by interpreting the α values as counts, leading to the iteration

$$p(i_v \mid i_{\mathrm{pa}(v)}) = \frac{n^*(i_{\mathrm{cl}(v)}) + \alpha(i_{\mathrm{cl}(v)})}{n^*(i_{\mathrm{pa}(v)}) + \alpha(i_{\mathrm{pa}(v)})}.$$

The prior distribution described above has also been used in Spiegelhalter and Lauritzen (1990) for an approximate Bayesian approach to sequential estimation in recursive models. A brief empirical comparison of this procedure to maximum penalized likelihood using the EM algorithm is given in Spiegelhalter *et al.* (1993).

6. Miscellaneous

It has been shown how the procedure of Lauritzen and Spiegelhalter (1988) can be used to calculate the terms in the E-step of the EM algorithm for hierarchical log-linear models and recursive models. The present note does not throw any additional light on the convergence properties of the EM algorithm which are discussed in Wu (1983). Asymptotic properties of estimates obtained in this way is treated in Sundberg (1972, 1974).

The algorithm has been implemented (Thiesson, 1991) using HUGIN for the propagation calculations and tested in several examples. Experience indicates that with data missing massively and systematically, the likelihood function has a number of local maxima and straight maximum likelihood gives results with unsuitably extreme probabilities. The penalized likelihood seems to perform much better.

The EM algorithm is known to converge relatively slowly when it is getting close. It turns out that the gradient of the likelihood function can be calculated with essentially the same amount of work as is involved in the E-step of the EM algorithm. It is therefore conceivable that algorithms exploiting the gradient could be preferable.

As an alternative, the related computational scheme described in Shenoy and Shafer (1990) could be used for the probability propagation in the E-step.

Finally it deserves mention that the procedure can be generalized to work for the mixed graphical association models of Lauritzen and Wermuth (1989) and the mixed hierarchical models of Edwards (1990) in cases where the joint distribution of the random variables involved is a CG-distribution. For these models the propagation scheme of Lauritzen (1992) can be used to calculate the conditional moments in the E-step. The explicit formulae in Frydenberg and Lauritzen (1989) may then be used for the M-step in the decomposable case, whereas iteration using the algorithm of Frydenberg and Edwards (1989) – implemented in the program MIM (Edwards 1989) – must be used in general. We abstain from giving a further description of the mixed case here. The case is quite analogous, although with somewhat more complex detail. In the special case of a saturated and homogeneous model, the algorithm has been discussed by Little and Schluchter (1985).

Acknowledgements

Comments and encouragement from Philip Dawid and David Spiegelhalter are greatly appreciated. I am also grateful to Rolf Sundberg for commenting on an earlier version of the manuscript.

The research has received partial support from the SCIENCE programme of the EEC as well as the Danish Natural Science Research Council.

References

Andersen, A.H. (1974), Multidimensional contingency tables, *Scandinavian Journal of Statistics*, **1**, 115–27.

Andersen, S.K., Olesen, K.G., Jensen, F.V., and Jensen, F. (1989), HUGIN – A shell for building Bayesian belief universes for expert systems, In *Proceedings of the 11th international joint conference on artificial intelligence*, pp. 1080–5. Also reprinted in Shafer and Pearl (1990).

Andreassen, S., Jensen, F., Andersen, S., Falck, B., Kjærulff, U., Woldbye, M., Sørensen, A., Rosenfalck, A., and Jensen, F. (1989), MUNIN – An EMG assistant, In *Computer-aided electromyography and expert systems in diagnosis*, (ed. J.E. Desmedt). Elsevier Science Publishers B.V. (North-Holland), Amsterdam.

Barndorff-Nielsen, O.E. (1978), *Information and exponential families in statistical theory*, John Wiley and Sons, New York.

Chen, T. and Fienberg, S.E. (1974), Two-dimensional contingency tables with both completely and partially cross-classified data, *Biometrics*, **30**, 629–42.

Chen, T. and Fienberg, S.E. (1976), The analysis of contingency tables with incompletely classified data, *Biometrics*, **32**, 133–44.

Darroch, J.N., Lauritzen, S.L., and Speed, T.P. (1980), Markov fields and log linear models for contingency tables, *Annals of Statistics*, **8**, 522–39.

Darroch, J.N. and Ratcliff, D. (1972), Generalized iterative scaling for log-linear models, *Annals of Mathematical Statistics*, **43**, 1470–80.

Darroch, J.N. and Speed, T.P. (1983), Additive and multiplicative models and interactions, *Annals of Statistics*, **11**, 724–38.

Dawid, A.P. (1992), Applications of a general propagation algorithm for probabilistic expert systems, *Statistics and Computing*, **2**, 25–36.

Dawid, A.P. and Skene, A.M. (1979), Maximum likelihood estimation of observer error-rates using the EM algorithm, *Applied Statistics*, **28**, 20–8.

Dempster, A.P., Laird, N., and Rubin, D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society*, Series B, **39**, 1–38.

Edwards, D. (1989), *A guide to* MIM, Manual.

Edwards, D. (1990), Hierarchical interaction models (with discussion), *Journal of the Royal Statistical Society, Series B*, **52**, 3–20 and 51–72.

Frydenberg, M. and Edwards, D. (1989), A modified iterative proportional scaling algorithm for estimation in regular exponential families, *Computational Statistics and Data Analysis*, **8**, 143–53.

Frydenberg, M. and Lauritzen, S.L. (1989), Decomposition of maximum likelihood in mixed interaction models, *Biometrika*, **76**, 539–55.

Fuchs, C. (1982), Maximum likelihood estimation and model selection in contingency tables with missing data, *Journal of the American Statistical Association*, **77**, 270–8.

Geng, Z. and Asano, C. (1988), Recursive procedures for hierarchical loglinear models on high-dimensional contingency tables, *Journal of the Japanese Society for Computational Statistics*, **1**, 17–26.

Golmard, J.-L. and Mallet, A. (1991), Learning probabilities in causal trees from incomplete databases, *Revue d'Intelligence Artificielle*, **5**, 93–106.

Green, P.J. (1990), On use of the EM algorithm for penalized likelihood estimation, *Journal of the Royal Statistical Society*, Series B, **52**, 443–52.

Haberman, S.J. (1974a), *The analysis of frequency data*, University of Chicago Press, Chicago.

Haberman, S.J. (1974b), Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations, *Annals of Statistics*, **2**, 911–24.

Hocking, R.R. and Oxspring, H.H. (1974), The analysis of partially categorized contingency data, *Biometrics*, **30**, 469–83.

Jensen, F.V., Lauritzen, S.L., and Olesen, K.G. (1990), Bayesian updating in causal probabilistic networks by local computation, *Computational Statistics Quarterly*, **4**, 269–82.

Kiiveri, H., Speed, T.P., and Carlin, J.B. (1984), Recursive causal models, *Journal of the Australian Mathematical Society* Series, A, **36**, 30–52.

Kjærulff, U. (1990), Graph triangulation – algorithms giving small total state space, Technical Report R 90-09, University of Aalborg, Denmark.

Kjærulff, U. (1992), Optimal decomposition of probabilistic networks by simulated annealing, *Statistics and Computing*, **2**, 19–24.

Lander, E.S. and Green, P. (1987), Construction of multilocus genetic linkage maps in humans, *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 2363–7.

Lauritzen, S.L. (1989), Lectures on contingency tables. 3rd edition, Technical Report R-89-29, Institute for Electronic Systems, Aalborg University. 1st. edition 1979, 2nd. edition 1982.

Lauritzen, S.L. (1992), Propagation of probabilities, means and variances in mixed graphical association models, *Journal of the American Statistical Association*, **87**, 1098–1108.

Lauritzen, S.L., Dawid, A.P., Larsen, B.N., and Leimer, H.-G. (1990), Independence properties of directed Markov fields, *Networks*, **20**, 491–505.

Lauritzen, S.L. and Spiegelhalter, D.J. (1988), Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *Journal of the Royal Statistical Society*, Series B, **50**, 157–224.

Lauritzen, S.L. and Wermuth, N. (1989), Graphical models for associations between variables, some of which are qualitative and some quantitative, *Annals of Statistics*, **17**, 31–57.

Lazarsfeld, P.F. and Henry, N.W. (1968), *Latent structure analysis*, Houghton Mifflin, Boston, Mass.

Little, R.J.A. and Schluchter, M.D. (1985), Maximum likelihood estimation for mixed continuous and categorical data with missing values, *Biometrika*, **72**, 497–512.

Martin-Löf, P. (1966), Statistics from the point of view of statistical mechanics, Notes by Ole Jørsboe from lectures at Aarhus University.

Oliver, R.M. and Smith, J.Q. (1990), *Influence diagrams, belief nets and decision analysis*, John Wiley and Sons, Chichester.

Pearl, J. (1988), *Probabilistic inference in intelligent systems*, Morgan Kaufmann, San Mateo.

Shafer, G.R. and Pearl, J. (ed.) (1990), *Readings in uncertain reasoning*, Morgan Kaufmann, San Mateo, California.

Shenoy, P.P. and Shafer, G.R. (1990), Axioms for probability and belief-function propagation, In *Uncertainty in artificial intelligence IV*, (ed. R.D. Shachter, T.S. Levitt, L.N. Kanal, and J.F. Lemmer), pp. 169–98. North-Holland, Amsterdam.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., and Cowell, R.G. (1993), Bayesian analysis in expert systems, *Statistical Science* 8, 219–247.

Spiegelhalter, D.J. and Lauritzen, S.L. (1990), Sequential updating of conditional probabilities on directed graphical structures, *Networks*, **20**, 579–605.

Sundberg, R. (1972), *Maximum likelihood theory and applications for distributions generated when observing a function of an exponential family variable*, PhD thesis, University of Stockholm.

Sundberg, R. (1974), Maximum likelihood theory for incomplete data from an exponential family, *Scandinavian Journal of Statistics*, **1**, 49–58.

Sundberg, R. (1976), An iterative method for solution of the likelihood equations for incomplete data from exponential families, *Communications in Statistics – Simulation and Computation*, **B5**, 55–64.

Thiesson, B. (1991), (G)EM algorithms for maximum likelihood in recursive graphical association models, Master's thesis, Department of Mathematics and Computer Science, Aalborg University.

Wermuth, N. and Lauritzen, S.L. (1983), Graphical and recursive models for contingency tables, *Biometrika*, **70**, 537–52.

Wu, C.F.J. (1983), On the convergence of the EM algorithm, *Annals of Statistics*, **11**, 95–103.