Maximum Likelihood Estimation in Gaussian Chain Graph Models under the Alternative Markov Property

Mathias Drton
Department of Statistics, The University of Chicago

Michael Eichler
Institute of Applied Mathematics, University of Heidelberg

Abstract

The AMP Markov property is a recently proposed alternative Markov property for chain graphs. In the case of continuous variables with a joint multivariate Gaussian distribution, it is the AMP rather than the earlier introduced LWF Markov property that is coherent with data-generation by natural block-recursive regressions. In this paper, we show that maximum likelihood estimates in Gaussian AMP chain graph models can be obtained by combining generalized least squares and iterative proportional fitting to an iterative algorithm. In an appendix, we give useful convergence results for iterative partial maximization algorithms that apply in particular to the described algorithm.

Key words: AMP chain graph, graphical model, iterative partial maximization, multivariate normal distribution, maximum likelihood estimation

1 Introduction

In graphical modelling, graphs are used to describe patterns of conditional independence. Undirected graphs encode the conditional independences underlying Markov random fields, and acyclic directed graphs encode the conditional independences underlying Bayesian networks. A generalization of both Markov random fields and Bayesian networks is provided by chain graphs that were introduced with the Markov/conditional independence interpretation described in Lauritzen & Wermuth (1989), Wermuth & Lauritzen (1990) and Frydenberg (1990); see also Lauritzen (1996, §5.4.1) and Edwards (2000,

§7.2). Graphical models jargon refers to the models induced by this Markov interpretation as LWF chain graph models. Recently, however, Andersson *et al.* (2001) have proposed an alternative Markov property (AMP) for chain graphs (see also Levitz *et al.*, 2001; Andersson & Perlman, 2004). In the case of continuous variables with a joint multivariate Gaussian = normal distribution, it is their AMP rather than the LWF Markov property that is coherent with data-generation by natural block-recursive regressions (Andersson *et al.*, 2001, §§1 and 5).

Statistical inference for LWF chain graph models is well developed, but this is not the case for AMP chain graph models. This paper considers maximum likelihood (ML) estimation in Gaussian AMP chain graph models. After reviewing these models in section 2, we derive, in section 3, the likelihood equations and the Fisher-information. Combining generalized least squares and iterative proportional fitting, we describe an iterative algorithm for solving the likelihood equations, which yields consistent and asymptotically efficient estimates. The convergence properties of this algorithm can be derived from convergence results for iterative partial maximization algorithms that are given in the Appendix. An application to university graduation data in section 4 illustrates AMP chain graph modelling. We conclude with the discussion in section 5.

2 Gaussian AMP chain graph models

Let $G = (V, E)$ be a mixed graph with finite vertex set V and an edge set E that may contain two types of edges, namely directed ($u \longrightarrow v$) and undirected ($u \multimap\!\!\!- v$) edges. The graph G is called a *chain graph* if it does not contain any semi-directed cycles, that is, it contains no path from v to v with at least one directed edge such that all directed edges have the same orientation. The vertex set of a chain graph can be partitioned into subsets $\tau \in \mathscr{T}$ such that all edges within each subset τ are undirected and edges between two different subsets $\tau \neq \tau'$ are directed. In the following, we assume that the partition $\tau \in \mathscr{T}$ is maximal, that is, any two vertices in a subset τ are connected by an undirected path. Then the subsets $\tau \in \mathscr{T}$ are unique and called the *chain components* of the graph G; compare figure 1 in section 4.

For a given chain graph G, we consider the class $\mathscr{P}(G)$ of normal distributions $\mathcal{N}(0, \Sigma)$ on $\mathbb{R}^V$ with positive definite covariance matrix Σ that satisfy the AMP Markov property (Andersson *et al.*, 2001, §4) with respect to G. Andersson *et al.* (2001, §5) described a parameterization of $\mathscr{P}(G)$ that associates one parameter with each vertex in V and each edge in E. More precisely, let $\Omega = (\Omega_{uv}) \in \mathbb{R}^{V \times V}$ be a positive definite matrix such that for any distinct vertices u and v

$$u \multimap\!\!\!- v \notin E \Rightarrow \Omega_{uv} = 0 \tag{1}$$

and let $B = (B_{uv})$ be an arbitrary matrix in $\mathbb{R}^{V\times V}$ such that for any vertices u and v

$$u \longrightarrow v \notin E \;\Rightarrow\; B_{vu} = 0. \tag{2}$$

For two such matrices Ω and B, we set

$$\Sigma(B,\Omega) = (I_V - B)^{-1}\,\Omega^{-1}\,(I_V - B')^{-1}, \tag{3}$$

where $I_V \in \mathbb{R}^{V\times V}$ denotes the identity matrix. A normal distribution $\mathcal{N}(0,\Sigma)$ with $\Sigma > 0$ satisfies the AMP Markov property if and only if there exist B and Ω such that (1) and (2) hold and $\Sigma = \Sigma(B,\Omega)$.

For a vertex $v \in V$, let $\mathrm{pa}(v) = \{u \in V \mid u \longrightarrow v \in E\}$ be the set of *parents* of v. Furthermore, we set $\mathrm{pa}(\tau) = \cup_{v\in\tau}\mathrm{pa}(v)$. Because of the nonexistence of semi-directed cycles, the joint distribution of X_V can be factorized as

$$f(x_V) = \prod_{\tau\in\mathscr{T}} f(x_\tau \mid x_{\mathrm{pa}(\tau)}), \quad x_V \in \mathbb{R}^V. \tag{4}$$

For $\tau \in \mathscr{T}$, the conditional distribution $f(x_\tau \mid x_{\mathrm{pa}(\tau)})$ is given by

$$X_\tau \mid X_{\mathrm{pa}(\tau)} \sim \mathcal{N}\big(B_\tau\, X_{\mathrm{pa}(\tau)}, \Omega_\tau^{-1}\big), \tag{5}$$

where $B_\tau = (B_{uv})_{u\in\tau, v\in\mathrm{pa}(\tau)}$ and $\Omega_\tau = (\Omega_{uv})_{u,v\in\tau}$ are submatrices of B and Ω, respectively. The conditional distribution corresponds to a *block-regression*, in which the block of variables X_τ is regressed on the parents $X_{\mathrm{pa}(\tau)}$.

The parameter (B_τ, Ω_τ) can be rewritten in vectorized form. Let $\beta_\tau = (B_{uv} \mid u \in \tau, v \in \mathrm{pa}(\tau))$ be the vector of unconstrained elements in B_τ. Subsequently, we write $B_\tau(\beta_\tau)$ for the matrix defined by β_τ and (2). Similarly, let ω_τ be the vector of elements of Ω_{uv} in Ω_τ such that either $u = v$, or $u < v$ and $u \multimap v \in E$. Furthermore, denote the dimension of β_τ and ω_τ by p_τ and q_τ, respectively. Then the parameter space for the parameter $(\beta_\tau, \omega_\tau)$ is

$$\Theta_\tau = \big\{(\beta_\tau,\omega_\tau) \in \mathbb{R}^{p_\tau+q_\tau} \mid \Omega_\tau(\omega_\tau) > 0\big\}, \tag{6}$$

where $\Omega_\tau(\omega_\tau) \in \mathbb{R}^{\tau\times\tau}$ is the matrix defined by ω_τ and (1). It follows from (4) and (5) that $\theta = (\beta_\tau,\omega_\tau)_{\tau\in\mathscr{T}}$ parameterizes $\mathscr{P}(G)$. Equation (15) below clarifies that θ is identifiable. The parameter space of $\mathscr{P}(G)$ is the Cartesian product $\Theta = \times_{\tau\in\mathscr{T}}\Theta_\tau$. This factorization of the parameter space together with the factorization of the joint density implies that the ML estimator (MLE) of the joint parameter θ can be obtained by computing, separately for every $\tau \in \mathscr{T}$, the MLE of (β_τ,ω_τ) in the block-regression (5). Furthermore, the Hessian of the likelihood function of the model $\mathscr{P}(G)$ is block-diagonal with one block for each one of the block-regressions indexed by $\tau \in \mathscr{T}$.

3 Maximum likelihood estimation

3.1 Likelihood equations

Let $X = (X_{v,i})_{v\in V, i\in N} \in \mathbb{R}^{V\times N}$ now be a data matrix whose column vectors, indexed by the set N, are independent and identically distributed according to some $P \in \mathscr{P}(G)$. Since, merely for notational convenience, the distributions in $\mathscr{P}(G)$ are assumed to be centered the sample covariance matrix is defined as

$$S = \frac{1}{n} X X',$$

where $n = |N|$ is the sample size. We assume that

$$n \geq \max_{\tau\in\mathscr{T}} \{|\tau| + |\mathrm{pa}(\tau)|\}$$

such that, with probability one, the submatrices $S_{\tau,\tau}$, $S_{\tau,\mathrm{pa}(\tau)}$, and the matrix $S(\beta_\tau)$ defined below are of full rank. This ensures that the MLE exists in each one of the block-regressions. Dividing by n and ignoring the additive constant $-(|V|/2)\log(2\pi)$, the log-likelihood function for the block-regression (5) is given by

$$\ell_n(\beta_\tau, \omega_\tau) = \frac{1}{2} \log|\Omega_\tau(\omega_\tau)| - \frac{1}{2} \operatorname{tr}\left[\Omega_\tau(\omega_\tau)\, S(\beta_\tau)\right], \tag{7}$$

where

$$\begin{aligned} S(\beta_\tau) &= \frac{1}{n}\left[X_\tau - B_\tau(\beta_\tau)\, X_{\mathrm{pa}(\tau)}\right]\left[X_\tau - B_\tau(\beta_\tau)\, X_{\mathrm{pa}(\tau)}\right]' \\ &= S_{\tau,\tau} - B_\tau(\beta_\tau) S_{\mathrm{pa}(\tau),\tau} - S_{\tau,\mathrm{pa}(\tau)} B_\tau(\beta_\tau)' + B_\tau(\beta_\tau) S_{\mathrm{pa}(\tau),\mathrm{pa}(\tau)} B_\tau(\beta_\tau)' \end{aligned}$$

is the sample covariance matrix of the residuals in the block-regression (5), and $X_A \in \mathbb{R}^{A\times N}$ denotes the submatrix of X that comprises all rows with index in A.

Let $P_\tau = \partial \operatorname{vec}(B_\tau)/\partial\beta_\tau'$ and $Q_\tau = \partial \operatorname{vec}(\Omega_\tau)/\partial\omega_\tau'$. Both P_τ and Q_τ have entries in $\{0,1\}$ and satisfy $\operatorname{vec}(B_\tau) = P_\tau\beta_\tau$ and $\operatorname{vec}(\Omega_\tau) = Q_\tau\omega_\tau$, respectively. Each column in P_τ has exactly one entry equal to one. A column in Q_τ has exactly one or exactly two entries equal to one depending on whether the associated element in ω_τ comes from the diagonal or the off-diagonal part of Ω_τ, respectively. With these two matrices the likelihood equations obtained by taking first derivatives with respect to β_τ and ω_τ can be written as

$$P_\tau' \left[\operatorname{vec}(\Omega_\tau\, S_{\tau,\mathrm{pa}(\tau)}) - (S_{\mathrm{pa}(\tau),\mathrm{pa}(\tau)} \otimes \Omega_\tau)\, P_\tau\, \beta_\tau\right] = 0 \tag{8}$$

and

$$Q_\tau' \operatorname{vec}\left[\Omega_\tau^{-1} - S(\beta_\tau)\right] = 0. \tag{9}$$

Equation (9) represents in a compact way the fact that the covariance associated with an undirected edge in the AMP chain graph is equal to its counterpart in $S(\beta_\tau)$, that is, it is equal to the empirical covariance of residuals computed for fixed β_τ. Thus, equation (9) parallels the well-known likelihood equations of undirected Gaussian graphical models.

3.2 Two-step estimation

If every vertex in $\mathrm{pa}(\tau)$ is adjacent to all vertices in τ, then no constraints on B_τ are imposed and P_τ becomes an identity matrix. In this case the first set of equations leads to the usual least squares estimator

$$\beta_\tau = (S^{-1}_{\mathrm{pa}(\tau),\mathrm{pa}(\tau)} \otimes I_\tau)\,\mathrm{vec}(S_{\tau,\mathrm{pa}(\tau)}) \;\Leftrightarrow\; B_\tau = S_{\tau,\mathrm{pa}(\tau)}\, S^{-1}_{\mathrm{pa}(\tau),\mathrm{pa}(\tau)}. \tag{10}$$

Thus the MLE of $(\beta_\tau, \omega_\tau)$ can be obtained by fitting an undirected graph model to the residuals computed using the regression coefficients estimates in (10). This can be done using iterative proportional fitting (Speed & Kiiveri, 1986; Whittaker, 1990, pp. 182–185), which generally will terminate in finitely many steps only if the subgraph G_τ induced by the chain component τ is decomposable.

In the case of general AMP chain graphs with constraints on B_τ, a similar two-step method can also be used for parameter estimation, as described in Edwards (2000, §7.5):

1. estimate β_τ by least squares by regressing each X_v, $v \in \tau$, on its parents $X_{\mathrm{pa}(v)}$,
2. estimate ω_τ by fitting an undirected graph model to the regression residuals.

For general AMP chain graphs with restrictions on B_τ, however, the two equations (8) and (9) for β_τ and ω_τ cannot be solved separately and the MLE differs from this two-step estimator.

3.3 Algorithm for maximum likelihood estimation

To compute the MLE, or rather a solution to the likelihood equations, in the general case, we consider an iterative method based on alternately maximizing the likelihood with respect to β_τ and ω_τ. Let $(\tilde{\beta}_\tau, \tilde{\omega}_\tau)$ be a consistent estimator. Then setting $\omega^{(1)} = \tilde{\omega}_\tau$ we define the sequence of estimators

$$\hat{\beta}^{(k+1)}_\tau = \underset{\beta_\tau \in \mathbb{R}^{p_\tau}}{\mathrm{argmax}}\, \ell_n(\beta_\tau, \hat{\omega}^{(k)}_\tau) \tag{11}$$

and

$$\hat{\omega}^{(k+1)}_\tau = \underset{\Omega_\tau(\omega_\tau)>0}{\mathrm{argmax}}\, \ell_n(\hat{\beta}^{(k+1)}_\tau, \omega_\tau) \tag{12}$$

for $k \geq 2$. Note that $\hat{\beta}_\tau^{(k+1)}$ can be computed in an explicit formula as the solution to (8) with Ω_τ substituted by $\Omega_\tau(\hat{\omega}_\tau^{(k)})$, which is

$$\hat{\beta}_\tau^{(k+1)} = \big\{ P_\tau' \left[S_{\mathrm{pa}(\tau),\mathrm{pa}(\tau)} \otimes \Omega_\tau(\hat{\omega}_\tau^{(k)}) \right] P_\tau \big\}^{-1} \big\{ P_\tau' \, \mathrm{vec}[\Omega_\tau(\hat{\omega}_\tau^{(k)}) \, S_{\tau,\mathrm{pa}(\tau)}] \big\}. \tag{13}$$

Similarly, $\hat{\omega}_\tau^{(k+1)}$ can be computed as the solution to (9) with β_τ substituted by $\hat{\beta}_\tau^{(k+1)}$. The equations in (9) then correspond to the likelihood equations of an undirected graph model for the undirected induced subgraph G_τ and the regression residuals as data. In other words the undirected graph model for G_τ has to be fitted to the sample covariance matrix $S(\hat{\beta}_\tau^{(k+1)})$, for which the iterative proportional fitting algorithm can be used.

The convergence properties of the sequence $(\hat{\beta}_\tau^{(k)}, \hat{\omega}_\tau^{(k)})_{k\in\mathbb{N}}$ are discussed in the appendix. In particular, it follows that the sequence converges if there are only finitely many solutions to the likelihood equations (8) and (9). Note that the likelihood equations may indeed have multiple solutions; compare Drton & Richardson (2004) and Drton (2005) who consider seemingly unrelated regressions that are special cases of the block-regressions encountered here.

3.4 The Fisher-information

For $\tau \in \mathscr{T}$, the second derivatives of the log-likelihood function are

$$\begin{aligned}
\frac{\partial^2 \ell_n(\beta_\tau, \omega_\tau)}{\partial \beta_\tau \partial \beta_\tau'} &= -P_\tau' \big[S_{\mathrm{pa}(\tau),\mathrm{pa}(\tau)} \otimes \Omega_\tau \big] P_\tau, \\
\frac{\partial^2 \ell_n(\beta_\tau, \omega_\tau)}{\partial \omega_\tau \partial \omega_\tau'} &= -\frac{1}{2} \, Q_\tau' \big[\Omega_\tau^{-1} \otimes \Omega_\tau^{-1} \big] Q_\tau
\end{aligned}$$

and

$$\frac{\partial^2 \ell_n(\beta_\tau, \omega_\tau)}{\partial \beta_\tau \partial \omega_\tau'} = -P_\tau' \big[\big(S_{\mathrm{pa}(\tau),\tau} - S_{\mathrm{pa}(\tau),\mathrm{pa}(\tau)} B_\tau(\beta_\tau)' \big) \otimes I_\tau \big] Q_\tau.$$

Let $\theta = (\theta_\tau)_{\tau\in\mathscr{T}} = (\beta_\tau, \omega_\tau)_{\tau\in\mathscr{T}} \in \Theta$, and let Σ be the associated covariance matrix given by equation (3). Then the Fisher-information $\mathscr{I}(\theta)$ for the Gaussian AMP chain graph model $\mathscr{P}(G)$ is block-diagonal and the $\tau \times \tau$-block is equal to

$$\mathscr{I}(\theta)_{\tau,\tau} = \begin{pmatrix} P_\tau'(\Sigma_{\mathrm{pa}(\tau),\mathrm{pa}(\tau)} \otimes \Omega_\tau) P_\tau & 0 \\ 0 & \frac{1}{2} \, Q_\tau' \big(\Omega_\tau^{-1} \otimes \Omega_\tau^{-1} \big) Q_\tau \end{pmatrix}. \tag{14}$$

3.5 Consistency and asymptotic normality

In the following, let $\hat{\theta}_{\tau,n} = (\hat{\beta}_{\tau,n}, \hat{\omega}_{\tau,n})$ be the limit of the sequence $(\hat{\beta}_\tau^{(k)}, \hat{\omega}_\tau^{(k)})_{k\in\mathbb{N}}$ for sample size n and let $\hat{\theta}_n = (\hat{\theta}_{\tau,n})_{\tau\in\mathscr{T}}$. Should such a limit not exist then choose $\hat{\theta}_{\tau,n}$ as

an arbitrary accumulation point. In either situation, all $\hat{\theta}_{\tau,n}$ are roots to the likelihood equations (8) and (9). This, together with the fact that Gaussian AMP chain graph models form curved exponential families (theorem 1), leads to the asymptotic normality stated in theorem 2.

Theorem 1. *The Gaussian AMP chain graph model $\mathscr{P}(G)$ is a curved exponential family.*

Proof. The model $\mathscr{P}(G)$ is a subfamily of the regular exponential family of centered multivariate normal distributions with arbitrary positive definite covariance matrix. The parameter space $\Theta = \times_{\tau\in\mathscr{T}}\Theta_\tau$ of $\mathscr{P}(G)$ is an open set in a Euclidian space and in particular a smooth manifold. For $\theta = (\beta_\tau, \omega_\tau)_{\tau\in\mathscr{T}} \in \Theta$, let $B(\theta)$ be the matrix that is zero except for its $\tau \times \mathrm{pa}(\tau)$-submatrices, $\tau \in \mathscr{T}$, which are equal to $B_\tau(\beta_\tau)$, and let similarly $\Omega(\theta)$ be the block-diagonal matrix with blocks $\Omega_\tau(\omega_\tau)$, $\tau \in \mathscr{T}$. By equation (3), the mapping

$$\psi : \theta \mapsto \Sigma(\theta)^{-1} = [I_V - B(\theta)']\,\Omega(\theta)\,[I_V - B(\theta)]$$

maps the parameter $\theta = (\beta_\tau, \omega_\tau)_{\tau\in\mathscr{T}} \in \Theta$ in the parameter space of $\mathscr{P}(G)$ to $\Sigma(\theta)^{-1}$ with $\mathcal{N}(0, \Sigma(\theta)) \in \mathscr{P}(G)$. The inverse map of ψ is determined by the fact that

$$B(\theta)_{v,\mathrm{pa}(v)} = \Sigma(\theta)_{v,\mathrm{pa}(v)}[\Sigma(\theta)_{\mathrm{pa}(v),\mathrm{pa}(v)}]^{-1}, \qquad v \in V; \tag{15}$$

compare Richardson & Spirtes (2002, Theorem 8.7). It is now apparent that the mapping ψ is a diffeomorphism. Therefore, $\psi(\Theta)$ is a smooth manifold, which means that $\mathscr{P}(G)$ forms a curved exponential family (Kass & Vos, 1997, Definition 2.3.1, 4.2.1).

Theorem 2. *Let $\theta = (\theta_\tau)_{\tau\in\mathscr{T}}$, $\theta_\tau = (\beta_\tau, \omega_\tau)$, be the true parameter. Then $\hat{\theta}_n \to \theta$ in probability, the estimates $\hat{\theta}_{\tau,n}$, $\tau \in \mathscr{T}$, are asymptotically independent, and for each $\tau \in \mathscr{T}$,*

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_{\tau,n} - \beta_\tau \\ \hat{\omega}_{\tau,n} - \omega_\tau\end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, [\mathscr{I}(\theta)_{\tau,\tau}]^{-1}\big)$$

with $\mathscr{I}(\theta)_{\tau,\tau}$ given in (14).

Proof. The estimators $\hat{\theta}_n$ are roots to the likelihood equations, computed in iterations starting at consistent estimates. Theorems 2.4.1, 2.6.1, 2.6.7, and 2.6.12 (see also Corollaries 2.4.2 and 2.6.2) in Kass & Vos (1997) imply that in one-parameter curved exponential families such roots to the likelihood equations are consistent and asymptotically normal with asymptotic variance equal to the inverse of the Fisher-information. As indicated before the statement of Theorem 4.2.4 in Kass & Vos (1997), these results extend to multi-parameter families, which yields our claim.

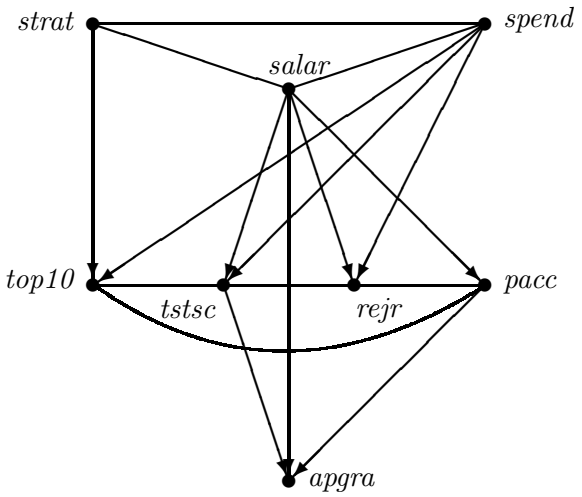

Figure 1: Chain graph with the three chain components {*spend, strat, salar*}, {*top10, tstsc, rejr, pacc*}, and {*agpra*}.

4 Example: University graduation rates

We illustrate our maximum likelihood procedure using the data in Druzdzel & Glymour (1999), which stem from a study for college ranking carried out in 1993. Based on $n = 159$ universities, Druzdzel & Glymour (1999, Table 3) state a correlation matrix for eight variables that are

spend	average spending per student,
strat	student-teacher ratio,
salar	faculty salary,
rejr	rejection rate,
pacc	percentage of admitted students who accept university's offer,
tstsc	average test scores of incoming students,
top10	class standing of incoming freshmen, and
apgra	average percentage of graduation.

Figure 1 shows a chain graph for these variables. This graph has the chain components $\tau_1 = \{spend, strat, salar\}$, $\tau_2 = \{top10, tstsc, rejr, pacc\}$, and $\tau_3 = \{apgra\}$.

It was selected via the SIN model selection procedure described in Drton & Perlman (2004a,b). More precisely, we used SIN model selection with simultaneous significance level 0.15 fixing the chain components τ_1, τ_2, and τ_3 *a priori* in the temporal order $\tau_1 < \tau_2 < \tau_3$. In the resulting AMP chain graph we deleted the undirected edge between *top10* and *rejr*, and introduced the undirected edge between *top10* and *pacc*, creating a non-decomposable chain component τ_2. Furthermore, we deleted the edge between *salar* and *top10* to create the edge constellation

$$salar \longrightarrow top10 \multimap tstsc \longleftarrow strat.$$

The induced subgraph over the four vertices *salar*, *strat*, *top10* and *tstsc*, which also contains the edge *salar* — *strat*, forms what is called a 2-biflag by Andersson *et al.* (2001); compare their figure 5(d). Therefore, by theorem 4 in Andersson *et al.* (2001), the AMP and LWF Markov properties differ for the graph in figure 1.

The block-regression for τ_1 is trivial as $\mathrm{pa}(\tau_1) = \emptyset$ and the undirected induced subgraph G_{τ_1} is complete, and thus the MLE of Ω_{τ_1} is simply the inverse of S_{τ_1,τ_1}. The block-regression for τ_3 is also simple as τ_3 contains only a single vertex. In this case, the MLE of β_{τ_3} and ω_{τ_3} can be computed by regressing the single variable in τ_3, here the variable *apgra*, on all its parents, here the variables *pacc*, *salar*, and *tstsc*. The vector of least squares estimates of the regression coefficients is the MLE of β_{τ_3} and the inverse of the estimated conditional variance is the MLE of ω_{τ_3}.

The remaining block-regression for τ_2 is non-trivial. We apply the ML estimation algorithm described in section 3.3, starting with the identity matrix as initial estimate of Ω_{τ_2} and iterating until convergence to find the estimates stated in the columns labelled "MLE" in table 1. Note that we cannot guarantee that these estimates constitute the global maximum of the likelihood function. However, using these estimates to evaluate the deviance of the AMP chain graph model yields a value of 16.89, which compared to 11 degrees of freedom indicates a reasonable fit.

Table 1 also states the two-step estimates obtained as described in section 3.2. These estimates coincide with the estimates after two steps of the ML estimation algorithm, provided the algorithm is started at a diagonal matrix. The two steps of the ML estimation algorithm consist of one step estimating β_τ assuming a diagonal matrix Ω_τ, i.e. assuming independence of all variables in the chain component τ, and one step estimating ω_τ using the newly found estimate of β_τ. The two-step estimates are fairly close to the MLEs, all differences being clearly smaller than two standard errors. The deviance based on the two-step estimates would be 19.18. Interestingly, the two-step estimates and the MLEs for the variance parameters ω_τ are identical in two digits of precision with the exception of the conditional variances ω_{top10}, ω_{tstsc} and the inverse

Table 1: *MLEs, their standard errors computed from the Fisher information matrix, and the two-step estimates for the block-regression for chain-component* τ_2.

Parameter	MLE	SE	2-step	Parameter	MLE	SE	2-step
$\beta_{pacc \leftarrow salar}$	-0.53	0.07	-0.52	ω_{pacc}	1.46	0.16	1.46
$\beta_{rejr \leftarrow salar}$	0.26	0.09	0.30	ω_{rejr}	1.64	0.18	1.64
$\beta_{rejr \leftarrow spend}$	0.30	0.09	0.27	ω_{top10}	2.99	0.33	2.92
$\beta_{top10 \leftarrow spend}$	0.98	0.08	0.99	ω_{tstsc}	3.39	0.37	3.34
$\beta_{top10 \leftarrow strat}$	0.44	0.07	0.45	$\omega_{pacc,rejr}$	-0.33	0.12	-0.33
$\beta_{tstsc \leftarrow salar}$	0.26	0.06	0.36	$\omega_{pacc,top10}$	-0.16	0.14	-0.16
$\beta_{tstsc \leftarrow spend}$	0.49	0.07	0.43	$\omega_{rejr,tstsc}$	-0.65	0.16	-0.65
				$\omega_{top10,tstsc}$	-1.76	0.28	-1.69

covariance $\omega_{top10,tstsc}$ that all involve the variables *top10* and *tstsc* that are part of the biflag.

5 Discussion

The likelihood function of a Gaussian AMP chain graph model can be factored into the product of conditional likelihood functions. Each chain component of the graph gives rise to one factor in this factorization. The iterative algorithm we proposed for ML estimation in Gaussian AMP chain graph models takes advantage of this fact and treats each chain component separately. For a given chain component, the algorithm alternates between estimating regression coefficients while fixing a covariance matrix and estimating the (restricted) covariance matrix while fixing regression coefficients. To perform the former task of estimating regression coefficients we use a generalized least squares formula, whereas the iterative proportional fitting algorithm is used to perform the latter task of estimating a covariance matrix.

The algorithm calls upon repeated runs of iterative proportional fitting in order to fit the block-regression model associated with a given chain component. This is in contrast to the case of LWF chain graph models, for which the ML estimates of the parameters associated with a chain component can be computed by running iterative proportional fitting only once (Lauritzen, 1996, §5.4.1, Proposition 6.33). However, the undirected graph on which iterative proportional fitting is run must be derived from the original LWF chain graph in a process called moralization. In general, this derived undirected graph contains also vertices outside the considered chain component and may feature

larger cliques than the undirected subgraph induced by the chain component, on which iterative proportional fitting is run when fitting AMP chain graph models.

The developed methodology for ML fitting of AMP chain graph models permits in particular to compare two models based on different chain graphs via likelihood ratio tests and information criteria. However, one may also be interested in testing parameter equality in a given model. If parameters are set equal in a curved exponential family, then the resulting submodel is again a curved exponential family. Therefore, the ML estimates in the submodel are asymptotically normal, and the problem of testing parameter equality can be addressed by a likelihood ratio test. For the computation of ML estimates in such submodels, the algorithm we proposed for fitting AMP chain graph models needs to be extended to incorporate equality constraints amongst subsets of the parameters. If parameter equality occurs between regression coefficients that appear in the same matrix B_τ, then the generalized least squares step of the algorithm can easily be adapted to deal with this new situation. The required changes consist solely of removing all but one of the identical entries of the vector β_τ and altering the matrix P_τ accordingly. With these changes, formula (11) still applies. If parameter equality occurs between entries of the matrix Ω_τ then the iterative proportional fitting step of the algorithm has to be adapted. This can be done as described in Højsgaard & Lauritzen (2005) who treat parameter equality in undirected graphical models. Finally if parameter equality occurs between parameters appearing in different matrices B_τ and $B_{\bar{\tau}}$, or Ω_τ and $\Omega_{\bar{\tau}}$, then the block-regressions can no longer be treated separately. In this case the extension of the presented algorithm requires additional work.

Appendix: Iterative partial maximization

The algorithm for ML estimation proposed in this paper is an iterative partial maximization algorithm in the sense of Lauritzen (1996, Appendix A.4). Partial maximization refers to a maximization of the likelihood function over a section in the parameter space. In an iterative partial maximization algorithm, one repeatedly performs a sequence of partial maximizations. In this appendix, we generalize the convergence results in Lauritzen (1996, Appendix A.4) by not assuming the existence of a unique local (and global) maximum of the likelihood function.

Let $L : \theta \to \mathbb{R}$ be a differentiable real-valued function on an open set $\Theta \subseteq \mathbb{R}^d$. In the context of ML estimation, L constitutes the (log-)likelihood function and Θ is the parameter space of a statistical model. Assume that there exists θ_0 such that $\Theta_0 = \{\theta \in \Theta \mid L(\theta) \geq L(\theta_0)\}$ is compact. Then L has a (not necessarily unique) global maximum in Θ_0. For functions $g_i : \Theta \to \mathbb{R}^{d_i}$, $i = 1, \ldots, k$ and $\theta^* \in \Theta$, we define sections $\Theta_i(\theta^*)$ in

Θ by

$$\Theta_i(\theta^*) = \{\theta \in \Theta \mid g_i(\theta) = g_i(\theta^*)\}.$$

We assume that the maximum of L over the section $\Theta_i(\theta^*)$ is uniquely attained for all $\theta^* \in \Theta$ and $i = 1, \ldots, k$ and that the associated mapping

$$T_i(\theta^*) = \operatorname*{argmax}_{\theta \in \Theta_i(\theta^*)} L(\theta)$$

from Θ into itself is continuous for all $i = 1, \ldots, k$. Moreover, we assume that if θ^* maximizes L over all sections $\Theta_i(\theta^*)$, and consequently satisfies $\theta^* = T_i(\theta^*)$ for $i = 1, \ldots, k$, then θ^* solves the likelihood equations

$$\left.\frac{\partial L(\theta)}{\partial \theta}\right|_{\theta=\theta^*} = 0. \tag{16}$$

Let $\theta_0 \in \Theta$ be a starting value such that Θ_0 is compact and define

$$\theta_{n+1} = S(\theta_n) = T_k \cdots T_1(\theta_n), \qquad n \geq 0.$$

By definition of Θ_0, we have $\theta_n \in \Theta_0$ for all $n \geq 0$. Let $\mathcal{A}_\infty$ be the set of accumulation points of the sequence $(\theta_n)_{n\in\mathbb{N}}$. Since Θ_0 is compact, we have $\mathcal{A}_\infty \subseteq \Theta_0$. The following results discuss the properties of $\mathcal{A}_\infty$. It is a special case of the convergence theorem in Zangwill (1969, Chapter 4).

Proposition 1. *The sequence* $\big(L(\theta_n)\big)_{n\in\mathbb{N}}$ *of values of the likelihood function converges to a limit* $\ell_\infty \in \mathbb{R}$. *Furthermore, if* $\alpha \in \mathcal{A}_\infty$ *then* $L(\alpha) = \ell_\infty$ *and* α *satisfies* (16).

Proof. Since the sequence $(L(\theta_n))_{n\in\mathbb{N}}$ is monotonously increasing and bounded, it converges to a limit ℓ_∞. By continuity of L, this also implies $L(\alpha) = \ell_\infty$ for all $\alpha \in \mathcal{A}_\infty$.

Next, since $S = T_k \cdots T_1$ is continuous, $S(\mathcal{A}_\infty)$ is the set of accumulation points of $(S(\theta_n)) = (\theta_{n+1})$. Consequently $S(\mathcal{A}_\infty) = \mathcal{A}_\infty$ and $L(S(\alpha)) = \ell_\infty$ for all $\alpha \in \mathcal{A}_\infty$. By definition of T_i, we now obtain for arbitrary $\alpha \in \mathcal{A}_\infty$,

$$\ell_\infty = L\big(T_k \cdots T_1(\alpha)\big) \geq L\big(T_{k-1} \cdots T_1(\alpha)\big) \geq L\big(T_1(\alpha)\big) \geq L\left(\alpha\right) = \ell_\infty,$$

which implies $T_i(\alpha) = \alpha$ for all $i = 1, \ldots, k$ because of uniqueness of the maximum over $\Theta_i(\alpha)$. Thus α maximizes L over all sections and hence satisfies equations (16) by assumption.

For the next theorem, recall that a compact set is said to be connected if it cannot be partitioned into two nonempty compact sets (see also Ostrowski, 1966, Theorem 28.1).

Theorem 3. *$\mathcal{A}_\infty$ is a compact and connected subset of Θ_0.*

Proof. Since $\mathcal{A}_\infty$ is a subset of a compact set, it suffices to show that $\mathcal{A}_\infty$ is closed. Let $\alpha^* \in \overline{\mathcal{A}_\infty}$. Then for any $\varepsilon > 0$ there exists $\alpha \in \mathcal{A}_\infty$ such that $\alpha \in B_\varepsilon(\alpha^*)$. Similarly, since α is an accumulation point of (θ_n), there exists for every $\delta > 0$ some $n_\delta \in \mathbb{N}$ such that $\theta_{n_\delta} \in B_\delta(\alpha)$. Since $B_\varepsilon(\alpha^*)$ is open, we can choose δ small enough such that $B_\delta(\alpha) \subseteq B_\varepsilon(\alpha^*)$, which implies $\theta_{n_\delta} \in B_\varepsilon(\alpha^*)$. Since ε was arbitrary, $\alpha^* \in \mathcal{A}_\infty$, which establishes the closedness of $\mathcal{A}_\infty$.

Next, let $B_\varepsilon(\mathcal{A}_\infty) = \{\theta \in \Theta \mid d(\theta, \mathcal{A}_\infty) < \varepsilon\}$ where $d(A, B)$ is the distance between two subsets A and B in $\mathbb{R}^d$. Then for every $\varepsilon > 0$ there exists $n_\varepsilon \in \mathbb{N}$ such that $\theta_n \in B_\varepsilon(\mathcal{A}_\infty)$ for all $n \geq n_\varepsilon$.

Now suppose that $\mathcal{A}_\infty$ can be partitioned into two compact sets A and B. Then $d(A, B) > 0$ and we set $\delta = d(A, B)/2$. Furthermore, because of uniform continuity of S on Θ_0, for all $\delta > 0$ there exists $\varepsilon' > 0$ such that for all $\alpha \in \mathcal{A}_\infty$, $\theta_n \in B_{\varepsilon'}(\alpha)$ implies $\theta_{n+1} = S(\theta_n) \in B_\delta(\alpha)$.

Then if $n > n_\varepsilon$ and $\theta_n \in B_\varepsilon(A)$, we have

$$\theta_{n+1} \in B_\delta(A) \cap B_\varepsilon(\mathcal{A}_\infty) = B_\varepsilon(A),$$

since $d(A, B) > \delta$. Thus $\theta_n \in B_\varepsilon(A)$ for all $n > n_\varepsilon$ and hence $B = \varnothing$ which concludes the proof.

Corollary 1. *If $\mathcal{A}_\infty$ is finite, then $\mathcal{A}_\infty = \{\theta^*\}$ for some $\theta^* \in \Theta_0$ and the sequence $(\theta_n)_{\theta \in \mathbb{N}}$ converges to θ^*.*

Proof. Any connected finite set must be a singleton.

Corollary 2. *If the likelihood equations* (16) *have only finitely many solutions that lie on the same contour of the likelihood function L, then the sequence $(\theta_n)_{n \in \mathbb{N}}$ converges to one solution θ^*.*

Proof. This follows from Proposition 1 and Corollary 1.

References

Andersson, S. A., Madigan, D. & Perlman, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Statist.* **28**, 33–85.

Andersson, S. A. & Perlman, M. D. (2004). Characterizing Markov equivalence classes for AMP chain graph models. Technical Report 453, Department of Statistics, University of Washington. Available at http://www.stat.washington.edu/www/research/reports/.

Drton, M. (2005). Computing all roots of the likelihood equations of seemingly unrelated regressions. *J. Symbolic Comput.*, in press.

Drton, M. & Perlman, M. D. (2004a). Model selection for Gaussian concentration graphs. *Biometrika* **91**, 591–602.

Drton, M. & Perlman, M. D. (2004b). A SINful approach to Gaussian graphical model selection. Technical Report 457, University of Washington.
Available at http://www.stat.washington.edu/www/research/reports/.

Drton, M. & Richardson, T. S. (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika* **91**, 383–392.

Druzdzel, M. J. & Glymour, C. (1999). Causal inferences from databases: Why universities lose students. In C. Glymour & G. F. Cooper (Eds.), *Computation, Causation, and Discovery*, chapter 19, pp. 521–539. AAAI Press, Menlo Park, CA.

Edwards, D. M. (2000). *Introduction to graphical modelling* (Second ed.). Springer-Verlag, New York.

Frydenberg, M. (1990). The chain graph Markov property. *Scand. J. Statist.* **17**, 333–353.

Højsgaard, S. & Lauritzen, S. (2005). Restricted concentration models – Gaussian models with concentration parameters restricted to being equal. In R. G. Cowell & Z. Ghahramani (Eds.), *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 152–157. Society for Artificial Intelligence and Statistics.
Available at http://www.gatsby.ucl.ac.uk/aistats/.

Kass, R. E. & Vos, P. W. (1997). *Geometrical foundations of asymptotic inference*. Wiley, New York.

Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press, Oxford, UK.

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

Levitz, M., Perlman, M. D. & Madigan, D. (2001). Separation and completeness properties for AMP chain graph Markov models. *Ann. Statist.* **29**, 1751–1784.

Ostrowski, A. M. (1966). *Solution of equations and systems of equations*. Academic Press, New York.

Richardson, T. S. & Spirtes, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30**, 962–1030.

Speed, T. P. & Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14**, 138–150.

Wermuth, N. & Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. Roy. Statist. Soc. Ser. B* **52**, 21–50, 51–72.

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

Zangwill, W. I. (1969). *Nonlinear programming: A unified approach*. Prentice-Hall Inc., Englewood Cliffs, NJ.

Mathias Drton, Department of Statistics, The University of Chicago, 5734 S. University Avenue, Chicago IL, 60637, U.S.A.
E-mail: drton@galton.uchicago.edu