Correlated Child Nodes in Aster Models

By
Charles J. Geyer
Technical Report No. 653
School of Statistics
University of Minnesota
April 25, 2006

Abstract

This technical report is an addendum to Technical Report (TR) No. 644 *Aster Models for Life History Analysis* by Charles J. Geyer, Stuart Wagenius, and Ruth G. Shaw. It is not intended to stand on its own. It simplifies and generalizes the models of TR No. 644, simplifying by getting rid of the notion of individual and the corresponding index in all notation and generalizing by allowing child nodes to be conditionally dependent given the parent. The simplification was inserted into the draft journal article in TR No. 644 during revision. The generalization was not because of space limitations.

1 Introduction

What would happen to the theory in the aster model draft journal article in Technical Report (TR) No. 644 if we allowed correlated child nodes? By this we mean child nodes conditionally correlated given the parent node. For specificity, say we allow a set of nodes that are children of one parent to be multinomial with sample size the parent variable.

1.1 Old Graphical Model

We describe conditional dependence structure graphically. Aster models described in TR No. 644 have directed acyclic graphs in which each node has at most one parent. Node j in the graph is associated with a variable X_j.

If an arrow in the graph points from node m to node j, then we say m is a *parent* of j and j is a *child* of m. Nodes without parents are called *root* nodes. Nodes without children are called *terminal* nodes.

Each arrow in the graph represents a conditional distribution of the variable at the child node given the variable at the parent node. Thus the joint distribution of all the variables is the product of conditionals, one conditional for each arrow in the graph.

Divide the nodes of the graph into disjoint sets, the root nodes F and non-root nodes J, and introduce a function $p : J \to J \cup F$ that maps nodes j to their parents $p(j)$. The diagram of p is the graph of the graphical model with all the arrows are reversed.

With this notation, we can write the joint distribution as

$$\prod_{j \in J} \Pr\{X_j | X_{p(j)}\}. \tag{1}$$

The fact that the variables X_j for $j \in F$ do not appear in front of the bar in (1) means these variables are treated as nonrandom in that we are conditioning on them and not modelling their distributions.

1.2 Individuals and Connectivity Components

A general graph has both directed edges (*arrows*) and undirected edges (*lines*). The *undirected version* of a directed graph is obtained by converting arrows to lines.

We say nodes of the graph are *connected* if there is a path following arrows or lines from one to the other. A *connectivity component* of the graph is a maximal connected set.

In aster models described in TR No. 644 as revised for publication, individuals correspond to connectivity components of the undirected version of the graph of the model. Data on different components are independent.

Thus these components can be thought of as individuals. In TR No. 644 there was an unnecessary assumption that all individuals (components) were the same, and this assumption was also embedded in the R package `aster`. In the revision of the draft journal article we dropped this assumption because (1) it confused the referee (2) it was unnecessary, (3) it made the notation uglier and harder to explain.

As yet, we have done nothing to revise the R package, although nothing in the package *requires* a connected graph. One could do a general aster model, as described in the revised draft journal article, by declaring that there was only one individual. For the *Echinacea* data, the original description had 570 individuals and a graph with 9 non-root nodes and one-root node (Figure 1 in the revised draft journal article and TR No. 644). But we could have just as well have said there was only one individual and the graph consisted of 570 identical copies of the graph in that Figure 1. Both specify exactly the same model. The `aster` package should probably be revised to reflect this change.

The result of this, is that whereas TR No. 644 had data X_{ij} with i indexing individuals and j indexing nodes of the graph, this TR and the revised draft journal article have just X_j with j indexing nodes of the graph (and referring to different "individuals" when in different components).

1.3 Old Log Likelihood

Each of the conditional distributions in (1) was a one-parameter exponential family (perhaps a different such family for each j) with X_j the canonical statistic and the dependence on $X_{p(j)}$ being that X_j is the sum of $X_{p(j)}$ independent and identically distributed random variables. In order that this make sense the variables X_j for $j \in p(J)$, where $p(J)$ denotes the set of non-terminal nodes, must be nonnegative-integer-valued.

Then the log likelihood for the whole family has the form

$$\sum_{j \in J} \bigl[X_j \theta_j - X_{p(j)} \psi_j(\theta_j) \bigr] \tag{2}$$

where θ_j is the canonical parameter for the j-th conditional family and ψ_j is the *cumulant function* for that family.

1.4 New Graphical Model

In order to have the graph reflect dependence among children, we must allow undirected edges (lines) between such children. So we now have a general graph with both arrows and lines.

However, it is a very special general graph. First it is a *chain graph* (Lauritzen, 1996, p. 7). The graph can be partitioned into a family of subsets

$\mathcal{G}$ that has a strict total ordering $\prec$ compatible with the graph in the following sense.

- All edges between nodes in the same subset are undirected.
- All edges between nodes in different subsets are directed.
- If G and G' are elements of $\mathcal{G}$ such that $G \prec G'$, then every arrow between $g \in G$ and $g' \in G'$ goes from g' to g.

The elements of $\mathcal{G}$ are called *chain components*.

We also impose restrictions specific to aster models. We need two additional bits of terminology. Nodes connected by a line are *neighbors* (note, this does not include nodes connected by arrows — parents and children). An undirected graph is *complete* if every pair of nodes are neighbors.

- Each node has at most one parent.
- Nodes having no parent (*root* nodes) have no neighbors.
- Each chain component is a complete undirected graph.
- Every node in a chain component has the same parent.

The last item allows us to redefine the parent function, writing $p(G)$ for the parent of the nodes in the chain component G. Since root nodes are singleton chain components and have no parents p is formally a function from the family of non-root chain components

$$\mathcal{J} = \{ G \in \mathcal{G} : G \subset J \}$$

to the graph $J \cup F$.

1.5 New Log Likelihood

Section 3.2.3 in Lauritzen (1996) says that the joint probability distribution for a chain graph model (including our new aster model) can be decomposed as the product

$$\prod_{G \in \mathcal{J}} \Pr\{\mathbf{X}_G | X_{p(G)}\}. \tag{3}$$

where $\mathbf{X}_G$ denotes the vector with components X_j, $j \in G$. This is (3.23) in Lauritzen (1996). This decomposition is subject to a condition that allows further factorization of each term in (3), which is (3.24) in Lauritzen (1996), but our assumption that each $G \in \mathcal{J}$ is complete makes that further decomposition trivial (a 'product' with only one term). Thus our (3) above is the analog of our old factorization (1).

We now assume that each of the conditional distributions in (3) is an exponential family (perhaps a different such family for each G) with $\mathbf{X}_G$ the canonical statistic and the dependence on $X_{p(G)}$ being that $\mathbf{X}_G$ is the sum of $X_{p(G)}$ independent and identically distributed random vectors. In order that this make sense the variables at non-terminal nodes, must be, as before, nonnegative-integer-valued.

Then the log likelihood for the whole family has the form

$$\sum_{G \in \mathcal{J}} \left[\sum_{j \in G} X_j \theta_j - X_{p(G)} \psi_G(\boldsymbol{\theta}_G) \right] = \sum_{j \in J} X_j \theta_j - \sum_{G \in \mathcal{J}} X_{p(G)} \psi_G(\boldsymbol{\theta}_G) \qquad (4)$$

where $\boldsymbol{\theta}_G$ is the canonical parameter vector for the G-th conditional family, having components θ_j, $j \in G$, and ψ_G is the *cumulant function* for that family.

1.6 Multinomial

We use a nonidentifiable but symmetric parameterization for the multinomial. if $\mathbf{X} = (X_1, \dots, X_d)$ is multinomial with sample size n and probability vector $\mathbf{p} = (p_1, \dots, p_d)$, then we take the map from canonical to mean value parameters $\boldsymbol{\tau} = n\mathbf{p}$ to be given by

$$p_j = \frac{\exp(\theta_j)}{\sum_{k=1}^{d} \exp(\theta_k)}.$$

We must at some point impose one linear constraint (or non-linear?) on the canonical parameters to force identifiability (or do without identifiability?)

The ψ_G for a multinomial group is given by

$$\psi_G(\boldsymbol{\theta}_G) = -\log \left(\sum_{j \in G} \exp(\theta_j) \right).$$

2 Theory

2.1 Reparameterization

Now we rewrite the unnumbered displayed equation above (1.5) of TR No. 644 and (5) of the revised draft journal article

$$\sum_{j \in J} X_j \left[\theta_j - \sum_{G \in p^{-1}(\{j\})} \psi_G(\boldsymbol{\theta}_G) \right] - \sum_{G \in p^{-1}(F)} X_{p(G)} \psi_G(\boldsymbol{\theta}_G)$$

and this makes the reparameterization (1.5) of TR No. 644 and (5) of the revised draft journal article become

$$\varphi_j = \theta_j - \sum_{G \in p^{-1}(\{j\})} \psi_G(\boldsymbol{\theta}_G), \qquad j \in J. \tag{5}$$

Changes are essentially only notational so far.

Now what about that lack of identifiability? It arises (or is dual to, whatever) from a constraint on canonical statistics

$$\sum_{j \in G} X_j = X_{p(G)}$$

for G that index a multinomial group. The natural thing to do is to impose a linear constraint

$$\sum_{j \in G} \theta_j = 0 \tag{6a}$$

for such a G if we are using a conditional model and impose instead

$$\sum_{j \in G} \varphi_j = 0 \tag{6b}$$

for such a G if we are using an unconditional model. Note that (6a) and (6b) cannot be imposed simultaneously. They conflict. We impose one or the other. We impose the constraint on the model we are fitting. Note that this constrains the model matrix for a canonical linear model.

2.2 Mean Value Parameters

Now consider the formula (1.12) in TR No. 644 and (12) in the revised draft journal article for unconditional mean value parameters. It seems we will no longer be able to write a simple formula, but must give a recursive formula, as we did for variances and covariances

$$E_{\boldsymbol{\varphi}}\{X_j\} = E_{\boldsymbol{\varphi}}\{X_{p(j)}\} \frac{\partial \psi_G(\boldsymbol{\theta}_G)}{\partial \theta_j}, \qquad j \in G \in \mathcal{J} \tag{7}$$

where (as before) $\boldsymbol{\theta}$ is determined from $\boldsymbol{\varphi}$ by solving (5).

2.3 Fisher Information

2.3.1 Variances

The recursive formula for variances in (1.16a) in TR No. 644 and (17) in the draft aster journal article becomes

$$\operatorname{var}_{\boldsymbol{\varphi}}\{X_j\} = \frac{\partial^2 \psi_G(\boldsymbol{\theta}_G)}{\partial \theta_j^2} E_{\boldsymbol{\varphi}}\{X_{p(G)}\} + \left(\frac{\partial \psi_G(\boldsymbol{\theta}_G)}{\partial \theta_j} \right)^2 \operatorname{var}_{\boldsymbol{\varphi}}\{X_{p(G)}\}, \tag{8}$$

where (as before) $j \in G \in \mathcal{J}$ and $\boldsymbol{\theta}$ is determined from $\boldsymbol{\varphi}$ by solving (5).

2.3.2 Covariances

The recursive formula for covariances (1.16b) in TR No. 644 and (18) in the draft aster journal article becomes even more complicated because of the chain components. As before, we fix $j \in G \in \mathcal{J}$ and $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ related by (5).

Let i be a node distinct from i. Now we get obnoxious case splitting. We consider the following cases.

(i) i is not a member of G or a descendant of a member of G.

(ii) $i \in G$ but $i \neq j$.

Case (i) is just what is considered in the aster draft journal article, but case (ii) is new, arising only with correlated neighbor nodes. In case (i) we rewrite (1.16b) in TR No. 644 and (18) in the draft aster journal article as

$$\operatorname{cov}_{\boldsymbol{\varphi}}\{X_i, X_j\} = \frac{\partial \psi_G(\boldsymbol{\theta}_G)}{\partial \theta_j} \operatorname{cov}_{\boldsymbol{\varphi}}\bigl\{X_i, X_{p(G)}\bigr\}. \tag{9}$$

In case (ii) correlation of these variables becomes

$$\operatorname{cov}_{\boldsymbol{\varphi}}\{X_i, X_j\} = \frac{\partial^2 \psi_G(\boldsymbol{\theta}_G)}{\partial \theta_i \partial \theta_j} E_{\boldsymbol{\varphi}}\bigl\{X_{p(G)}\bigr\} + \frac{\partial \psi_G(\boldsymbol{\theta}_G)}{\partial \theta_i} \frac{\partial \psi_G(\boldsymbol{\theta}_G)}{\partial \theta_j} \operatorname{var}_{\boldsymbol{\varphi}}\bigl\{X_{p(G)}\bigr\}. \tag{10}$$

2.4 Summary

Except for equation (10), which is entirely new, arising only from the conditional correlation of neighbor nodes given parents, everything else in the original TR and draft journal article goes through with merely notational changes.

3 Discussion

The motivation for this was a challenge to generalize aster models posed by a journal editor. We have no example applications.

The simplest aster model of this new type that would not be also be the old type would have a 'switch' node. The multinomial random vector would have two nodes and simple size which was Bernoulli. If X_1 and X_2 were the two child nodes, then we would have $X_1 + X_2 = X_{p(1)}$ so when $X_{p(1)} = X_{p(2)} = 1$ we would have either $X_1 = 1$ and $X_2 = 0$ or $X_1 = 0$ and $X_2 = 1$. The descendants of node 1 or node 2 would all be zero, and the descendants of the other could be nonzero.

Ruth Shaw said that when she gave a talk at University of California, Davis Beverly Ajie asked about something like this and whether it could be used for aster models for animal behavior data.

References

LAURITZEN, S. L. (1996). *Graphical Models.* New York: Oxford University Press.