Stat 8931 (Exponential Families) Lecture Notes

Aster Models

Charles J. Geyer

December 5, 2016

1 Introduction

1.1 Aster Models and Other Statistical Models

Aster models (Geyer, Wagenius, and Shaw, 2007; Shaw, Geyer, Wagenius, Hangelbroek, and Etterson, 2008) are statistical models designed especially for life history analysis, which is explained in Section 1.2 below. Aster models are implemented in the R statistical computing language (R Core Team, 2016) in the R packages `aster` and `aster2` (Geyer, 2015a,b) that are on CRAN (`http://cran.r-project.org/`) and hence easily installed by anyone who has R.

Like linear models (LM) and generalized linear models (GLM), aster models are regression models. They model the probability distribution of the response vector conditional on covariate data, which is thereby treated as nonrandom (the marginal distribution of covariates is not modeled).

One way to think about aster models is that they are a kind of generalized generalized linear models, that is, they are a generalization of (some types of) GLM (those that are exponential family models). In aster models components of the response vector are not necessarily conditionally independent given covariate data and their conditional distributions given covariate data do not necessarily come from the same family.

Another way to think about aster models is that they are a kind of graphical model. They are very simple graphical models that allow their joint distribution to be factored as a product of marginal and conditional distributions that are read off the graph. Lauritzen (1996) calls such graphical models "chain graph models," but aster models are even simpler than general chain graph models for reasons that will be revealed presently. Nothing of graphical model theory is required for understanding aster models.

Another way to think about aster models is that they are a generalization of (discrete time) survival analysis. Life history analysis is what you have when survival is an issue but not the main issue, which is what happens after survival.

1.2 Life History Analysis

The first published aster model (Geyer, et al., 2007) had this graph

$$\begin{tikzcd}
1 \arrow[r, "\text{Ber}"] & y_1 \arrow[r, "\text{Ber}"] \arrow[d, "\text{Ber}"] & y_2 \arrow[r, "\text{Ber}"] \arrow[d, "\text{Ber}"] & y_3 \arrow[d, "\text{Ber}"] \\
 & y_4 \arrow[d, "\text{0-Poi}"] & y_5 \arrow[d, "\text{0-Poi}"] & y_6 \arrow[d, "\text{0-Poi}"] \\
 & y_7 & y_8 & y_9
\end{tikzcd} \tag{1}$$

which is for one individual. There are 570 individuals in these data (R dataset `echinacea` in R package `aster`). So one can think of the full aster graph as 570 copies of this graph with the subscripts changed so each node (each y_j) has a different subscript.

The individuals are plants of the species *Echinacea angustifolia*, whose common name is narrow-leaved purple coneflower. These data were collected by the Echinacea Project (`http://echinaceaproject.org/`) a long-running project funded by the National Science Foundation (the PI's are the second and third authors of Geyer, et al. (2007)). The way (1) is laid out, variables in the first column (y_1, y_4, and y_7) are for 2002, those in the second column are for 2003, those in the third column are for 2004, those in the first row (y_1, y_2, and y_3) measure survival ($0 =$ dead, $1 =$ alive), those in the second row indicate flowering ($0 =$ no flowers, $1 =$ some flowers), those in the third row are flower head counts (actual number of flower heads).

Aster graphs can get a lot bigger than (1). The Echinacea Project now has data for years since 2004 (which extends the graph with many more "columns") and data for more life history stages (which extends the graph with more "rows"). Of course, the "rows" and "columns" are not part of the graphical structure. The only thing that matters is which nodes are connected by which arrows.

The node labels (the y_j) are random variables, components of the response vector. The arrows indicate conditional distributions. An arrow

$$y_i \longrightarrow y_j \tag{2}$$

indicates indicates the conditional distribution of y_j given y_i. An arrow

$$1 \longrightarrow y_i \tag{3}$$

indicates indicates the marginal distribution of y_i, because conditioning on a constant random variable is the same as not conditioning.

Labels on the arrows name the distribution. Ber is for Bernoulli (any zero-or-one-valued random variable), and 0-Poi is for zero-truncated Poisson (Poisson conditioned on being nonzero). This explanation of arrows and their distributions is incomplete and will be picked up again in Section 1.4.

Even from this one example one can see what “life history analysis” is about. There are life history stages. What happens at one stage depends on what happened before. Graphical models can capture some of this dependence.

Aster models are not only about life history of real biological organisms. Any statistical model that satisfies the assumptions (Sections 1.7 and 2.2 below) for aster models is an aster model.

1.3 Graphical Terminology

For an arrow (2) we say y_i is a *predecessor* of y_j, and, conversely, that y_j is a *successor* of y_i. For an arrow (3) we say the same, the only difference being that the predecessor is a constant random variable. A node having no predecessor is called an *initial* node. A node having no successor is called a *terminal* node.

This terminology is widely used in graph theory, but there is an alternative set of terms that is even more widely used, perhaps because it is more colorful. These are terms of biological origin: predecessor = parent, successor = child, initial = root, terminal = leaf.

Since the main applications of aster models are biological, graph theoretical terminology of biological origin is a source of real confusion. Does “child” refer to a graphical property or a biological property? Hence we have made a conscious decision to eschew all terms of biological origin in aster model theory.

Except the R package `aster` has one quirk. It uses “root” instead of “initial.” The *no terms of biological origin* policy had not yet been set when that package was written. The R package `aster2` does use “initial” instead of “root.”

Terminology of biological origin also has a term “ancestor” that means predecessor, predecessor of predecessor, predecessor of predecessor of predecessor, and so forth (any number of repetitions of “predecessor of”) and a term “descendant” that is its converse (y_j is a descendant of y_k if and only if y_k is an ancestor of y_j). We eschew these terms too. Occasions for using them when discussing aster models are are rare, and when they occur we use the long-winded expression in terms of predecessors and successors.

Terminology of biological origin also has a term “forest” that means a graph in which every node has at most one predecessor and a term "tree" that means a forest graph having a single initial node (so a "forest" is the disjoint union of one or more "trees"). We eschew these terms too, to avoid confusion when data is about real trees in real forests.

1.4 Predecessor is Sample Size

All aster models have the *predecessor is sample size* property. This is a very important property that separates them from all other graphical models.

For an arrow

$$y_i \xrightarrow{\text{What}} y_j$$

where the arrow label What is an abbreviation for the Whatever distribution,

- conditional on $y_i = 0$, the distribution of y_j is concentrated at zero,
- conditional on $y_i = 1$, the distribution of y_j is the Whatever distribution, and
- conditional on $y_i = n$ with $n > 1$, the distribution of y_j is the n-fold convolution of the Whatever distribution.

In short, the conditional distribution of y_j given y_i is the distribution of the sum of y_i independent and identically distributed (IID) random variables having the Whatever distribution. (By convention, a sum having zero terms is zero, and a sum having one term is that term.) Or, even shorter, the predecessor plays the role of sample size for the conditional distribution for an arrow. Or, shorter still, *predecessor is sample size.*

Note that arrow labels do not name the associated conditional distribution unless the predecessor is equal to one. For other predecessor values, we must work out what the distribution is. This can seem unnecessarily mysterious at first sight. For an arrow

$$y_i \xrightarrow{\text{Ber}} y_j$$

why not just say the conditional distribution of y_j given y_i is binomial with sample size y_i (because the sum of IID Bernoulli is binomial)? For one thing, it is not clear what sample size zero means without further explanation. For another thing, for an arrow

$$y_i \xrightarrow{\text{0-Poi}} y_j$$

the distribution of the sum of IID zero-truncated Poisson random variables is not a “brand name distribution.” And its probability mass function has no closed-form expression. So we could not label this arrow with the name of the conditional distribution of y_j given y_i because there is no such name.

A consequence of the predecessor is sample size property is that y_j that are predecessors (are at nonterminal nodes) must be nonnegative-integer-valued random variables.

Geyer, et al. (2007) mention an exception to this policy. If the conditional distribution for an arrow is infinitely divisible (for example, Poisson or normal with unknown mean and known variance), then the predecessor can be nonnegative-real-valued. But this feature has never been used, and the R package `aster` does not allow it.

1.5 The Name of the Game

The reason why random variables on the “bottom row” of the aster graph (1) are called “flower head” counts instead of “flower” counts is that *E. angustifolia* is in the family Asteraceae, the aster family (formerly called family Compositae, because members have composite flowers, also called inflorescences or flower heads). This very large family includes asters, chrysanthemums, dahlias, dandelions, daisies, sunflowers, and zinnias. The flowers gave aster models their name.

There was also a faint hint of analogy that the word “aster” means star in Greek and Latin, and some drawings of stars are graphs. But maybe no users get this connection.

1.6 Dependence Groups

General aster models have a feature called dependence groups, but they are not widely used. The R package `aster` does not implement them. The R package `aster2` does implement them but is unfinished, lacking many features of the other package. The basic plumbing is there, but not much else, so `aster2` is usable only by experts. All published analyses have used the `aster` package. except for (Eck, Shaw, Geyer, and Kingsolver, 2015).

The first paper about aster models (Geyer, et al., 2007) describes aster models with dependence groups, but the first submitted version, still to be found in one of the technical reports for that paper (Geyer, et al., 2005), did not. Dependence groups were added because the editor of the journal asked for more complication. No other paper about aster models used dependence groups until Eck et al. (2015). Your humble author taught a special topics course that ran for a whole semester without getting to dependence groups (`http://www.stat.umn.edu/geyer/8931aster/`).

We will do a first pass over the theory of aster models omitting dependence groups. That makes what is already a very complicated subject as simple as it can be.

1.7 The Aster Axioms (No Dependence Groups)

1.7.1 Graphical Axioms

An aster model without dependence groups is a graphical model whose graph has the following properties.

- It is directed. All edges have a direction going from one node to another. We denote the edges by arrows when drawing the graph and usually call them arrows too.
- It is acyclic. There is no path following arrows that returns to its starting point.
- Every node has at most one predecessor.

Let N denote the set of nodes of the graph, and let J denote the set of non-initial nodes of the graph. By the third axiom, every $j \in J$ has exactly one predecessor, which we denote $p(j)$. We may consider p a function $J \to N$. This function tells us everything there is to know about the graph. Its codomain is the node set, and its argument-value pairs specify the arrows. It is called the *predecessor function.*

1.7.2 Statistical Axioms

Each node of the graph is associated with a random variable. Let y_j, $j \in N$ denote the random variables.

The joint distribution of the random variables factors as a product of conditional distributions

$$\prod_{j \in J} \operatorname{pr}(y_j \mid y_{p(j)}) \tag{4}$$

(that this is a valid factorization follows from Lauritzen, 1996, Section 3.2.2). This is the *factorization axiom.*

Only y_j for $j \in J$ (for non-initial nodes) appear "in front of the bar" in one of these conditional distributions. Hence only these y_j are treated as random. Hence y_j for $j \in N \setminus J$ (for initial nodes) are treated as constant random variables.

Conditioning on a constant random variable is the same as not conditioning; such a conditional distribution is really a marginal distribution. Hence terms in (4) for which $p(j) \notin J$ are really marginal distributions.

Sometimes we use notation like $y_{p(j)}$ that makes these constants look like (constant) random variables. Sometimes we use notation like the 1 in the graph (1) that makes these constants look like constants. The constants are not considered part of the response vector of the aster model; when we write y for the response vector, its components are y_j, $j \in J$. Every component y_j of the response vector has a predecessor $y_{p(j)}$, which may or may not be another component of the response vector (if not, it is a constant at an initial node).

Each conditional distribution in (4) obeys the *predecessor is sample size* axiom (Section 1.4 above).

The final axiom is the *exponential family* axiom. Each conditional distribution in (4) is a one-parameter exponential family with y_j as the canonical statistic and $y_{p(j)}$ as the sample size. We let θ_j denote the canonical parameter and let c_j denote the cumulant function for sample size 1. This allows each arrow to be associated with a different exponential family and different canonical parameter from other arrows. Both c_j and θ_j depend on j.

Recall (Geyer, 2013, Section 8) that IID sampling from an exponential family produces another exponential family whose canonical parameter is the same as the original, whose canonical statistic is the sum of the canonical statistics for the IID sample, and whose cumulant function is the sample size times the original cumulant function. The log likelihood for an aster model without dependence groups is thus

$$l(\theta) = \sum_{j \in J} \left[y_j \theta_j - y_{p(j)} c_j(\theta_j) \right]. \tag{5}$$

In summary (so we have a nice list of axioms like in the preceding section), the statistical axioms are

- the factorization axiom,
- the predecessor is sample size axiom, and
- the exponential family axiom.

1.8 The Aster Transform (No Dependence Groups)

The log likelihood (5) is linear in the y's so the joint distribution of the response vector is also an exponential family, the y_j are its canonical statistics, whatever multiplies them must be the canonical parameters, and whatever is left over from the sum of canonical statistics times canonical parameters is the cumulant function of the joint distribution. Rewrite (5) as

$$\begin{aligned} l(\theta) &= \sum_{j \in J} \left[y_j \theta_j - y_{p(j)} c_j(\theta_j) \right] \\ &= \left(\sum_{j \in J} y_j \left[\theta_j - \sum_{\substack{k \in J \\ p(k) = j}} c_k(\theta_k) \right] \right) - \left(\sum_{\substack{j \in J \\ p(j) \notin J}} y_{p(j)} c_j(\theta_j) \right) \end{aligned}$$

to see that the terms in square brackets are the canonical parameters

$$\varphi_j = \theta_j - \sum_{\substack{k \in J \\ p(k) = j}} c_k(\theta_k) \tag{6}$$

and the cumulant function of the joint distribution is

$$c(\varphi) = \sum_{\substack{j \in J \\ p(j) \notin J}} y_{p(j)} c_j(\theta_j) \tag{7}$$

We call the change of parameter (6) the *aster transform.* It is an invertible change of parameter. The inverse is

$$\theta_j = \varphi_j + \sum_{\substack{k \in J \\ p(k) = j}} c_k(\theta_k). \tag{8}$$

How can that be an inverse with θ's on the right hand side? Traverse the graph in any order that visits successors before predecessors (there always is one because the graph is acyclic) using (8) to determine θ_j as a function of φ. Each time (8) is used, all of the k such that $p(k) = j$ are successors of j, so their θ_k have already been determined as a function of φ. So (8) works.

The fact that the aster transform is invertible clears up another mystery. It is all right for (7) to have φ as the free variable on the left hand side but only θ's on the right hand side because those θ's are a function of φ via the inverse aster transform. Also note that the $p(j)$ on the right hand side of

(7) are not in J so the $y_{p(j)}$ are constants rather than random variables, so (7) defines a deterministic (non-random) function, as a cumulant function for an unconditional distribution must be.

Written in terms of the φ's the aster log likelihood is very simple

$$l(\varphi) = \langle y, \varphi \rangle - c(\varphi) \tag{9}$$

where the term $\langle y, \varphi \rangle$ must have both vectors with the same index set J and the angle brackets mean duality pairing (Geyer, 2016, Section 5), that is,

$$\langle y, \varphi \rangle = \sum_{j \in J} y_j \varphi_j .$$

Geyer, et al. (2007) call θ the conditional canonical parameter vector and φ the unconditional canonical parameter vector. The analogy is not as close as the terminology seems to indicate. The components of the conditional canonical parameter vector are the canonical parameters for the one-parameter exponential families associated with the arrows of the graph. They are the canonical parameters for the conditional distributions in the factorization (4). The unconditional canonical parameter vector is the canonical parameter vector for the joint distribution of the aster model. Another way to say this is that the θ's are componentwise but not vectorwise canonical, and the φ's are vectorwise but not componentwise canonical. The names conditional and unconditional come from the θ's being the canonical parameters for the *conditional* distributions associated with the arrows and the φ's being the canonical parameters for the (*unconditional*) joint distribution.

1.9 Zero-Inflated Poisson

Readers may have wondered why the graph (1) has its “middle layer”. The variables y_4, y_5, and y_6 are a function of the variables y_7, y_8, and y_9, respectively, ($y_j = 1$ if and only if $y_{j+3} > 0$). So why were these variables inserted in the graph?

Consider just the subgraph

$$y_1 \xrightarrow{\text{Ber}} y_4 \xrightarrow{\text{0-Poi}} y_7 \tag{10}$$

The conditional distribution of y_7 given y_1 (both arrows combined) is zero-inflated Poisson (Lambert, 1992). Since $y_7 = 0$ if and only if $y_4 = 0$, and the probability of this event can be anything (because the Bernoulli arrow does not have any restrictions on its parameter), it is, strictly speaking, zero-inflated-or-deflated Poisson, but we will not be this fussy about terminology.

So having arrows arranged like this is just the aster way of getting zero-inflated Poisson random variables into the model.

Although we have zero-inflated Poisson distributions in aster model, we do not give them their usual parameterization (Lambert, 1992). Neither θ nor φ parameterizes the distribution associated with the graph (10) in the usual parameterization for zero-inflated Poisson.

1.10 Exponential Families and Canonical Affine Submodels

Suppose (9) is the log likelihood of an exponential family that is not necessarily an aster model. As in LM and GLM we consider linear submodels in which the canonical parameters are expressed as a linear function of other parameters, called *regression coefficients*. Actually, LM and GLM as implemented by the R functions `lm` and `glm` allow submodels to express canonical parameters as affine functions of regression coefficients through the `offset` argument. Thus, strictly speaking, they should be called "affine models" and "generalized affine models." But offsets are rarely used, so the terminology LM and GLM with L for linear persists. In aster model theory we do call affine submodels "affine" rather than "linear."

A *canonical affine submodel* has parameterization

$$\varphi = a + M\beta, \tag{11}$$

where a is a known vector and M is a known matrix; a is called the *offset* vector and M is called the *model matrix* by the R functions `lm` and `glm`. M is called the *design matrix* by others. We use the terminology favored by R.

The log likelihood for the canonical affine submodel is

$$\begin{aligned} l(\beta) &= \langle y, a + M\beta \rangle - c(a + M\beta) \\ &= \langle y, a \rangle + \langle M^T y, \beta \rangle - c(a + M\beta), \end{aligned}$$

and the term $\langle y, a \rangle$ that does not contain the parameter can be dropped giving the log likelihood

$$l(\beta) = \langle M^T y, \beta \rangle - c(a + M\beta), \tag{12}$$

and this shows the canonical affine submodel is again an exponential family with canonical statistic vector $M^T y$, canonical parameter vector β, and cumulant function

$$\beta \mapsto c(a + M\beta).$$

In short, canonical affine submodels take us from full exponential families to other full exponential families.

1.11 Aster Models and Canonical Affine Submodels

The R package `aster` has a terminological quirk. It calls the offset (the a in (11)) the *origin*. Your humble author, who is also the author of the `aster` package, was unclear at that time about the relationship to offsets in LM and GLM.

Up to now we have only considered aster models having one parameter per component of the response vector. We call them *saturated models*. Each is the same model whether we parameterize it with θ or with φ. Now we consider canonical affine submodels.

Since aster models have two canonical parameterizations, they have two kinds of canonical affine submodels.

The unconditional canonical parameter vector φ is the canonical parameter vector when the joint distribution of the aster model is considered as an exponential family. Hence the *unconditional canonical affine submodel*, given by (11), is the one that takes exponential families to exponential families, resulting in a submodel that has all exponential family properties. For short, we will call such models *unconditional aster models* (UAM).

But there is also the temptation to use the conditional canonical parameter vector θ in an affine submodel with parameterization

$$\theta = a + M\beta.$$

For short, we will call such models *conditional aster models* (CAM). Of course, even if the a and M are the same as in (11), this produces a radically different submodel. This kind of submodel does not take full exponential families to full exponential families. Of course, being a smooth submodel of the exponential family saturated model, it is a so-called curved exponential family. But those do not have all exponential family properties. For example, there is, for general curved exponential families, no guarantee that the log likelihood is concave.

For conditional aster aster models, however, there is such a guarantee. The *associated independent model* (AIM) of the CAM is the model that treats all predecessors as known constants. This makes no sense probabilistically because the very same variables occur as both successors and predecessors. But it does make sense algebraically. If the log likelihood of the CAM is given by (5), then the log likelihood of the AIM is given by

$$l(\theta) = \sum_{j \in J} \left[y_j \theta_j - n_j c_j(\theta_j) \right], \tag{13}$$

which is just (5) with every $y_{p(j)}$ replaced by n_j. Now (13) has the form of a full exponential family with independent components of the response vector y_j. So canonical affine submodels of it are also full exponential families and have all exponential family properties. In particular, the log likelihood (5) or (13) is strictly concave and the MLE is unique if it exists.

We stess that the AIM makes no sense probabilistically, so exponential family properties that are probabilistic (the sufficient dimension reduction property and the maximum entropy property) are not properties of the AIM or CAM. But the AIM does make sense algebraically, so exponential family properties that are algebraic (concavity of the log likelihood and multivariate monotonicity of the map from canonical to mean-value parameters) are properties of the AIM or CAM.

1.12 Conditioning and Notation

As mentioned in Section 1.1, aster models are regression models, but we have not mentioned covariate data since. That is because saturated aster models do not involve covariate data. Now that we have canonical affine submodels, we incorporate covariate data in the same way that LM and GLM do: the model matrix depends on covariate data. The offset vector could also depend on covariate data, but usually does not.

How do we want to indicate this dependence? Do we want to rewrite the factorization (4) as

$$\prod_{j \in J} \operatorname{pr}(y_j \mid y_{p(j)} \text{ and covariate data})$$

and, similarly, put conditioning on covariate data into every probability and expectation? Our answer is no. In this we follow many treatments of LM and GLM. The explicit dependence of models on model matrices is enough. And anyway, covariate data need not be random. Some of it is designed (values chosen by experimenters). And avoidance of explicit mention of conditioning on the part of the covariate data that is random avoids this issue.

In aster models, there is an even more important reason for avoiding explicit indication of conditioning on covariate data. That is the more important conditioning of some components of the response vector on other components (of successors on predecessors), which we denote explicitly, as in (4). We want a clear distinction between the unconditional expectation of y_j and the conditional expectation of y_j given $y_{p(j)}$, and we do not want to confuse this distinction by calling every probability and expectation conditional (on part of the covariate data and perhaps on some components of the response vector).

In summary, model matrices depend on covariate data, submodels depend on their model matrices, hence what we explicitly write as the unconditional distribution of the response vector (4) is implicitly the conditional distribution of the response vector given (the random part of) the covariate data. But this is only implicitly indicated. In this we follow many treatments of LM and GLM.

This is an arbitrary place to end the introduction. We have met aster models, the aster transform, conditional and unconditional canonical parameters, and conditional and unconditional canonical affine submodels.

REVISED DOWN TO HERE

2 Theory

Now we are going to redo part of the introduction, adding dependence groups. The reason is that theory for aster models with and without dependence groups is very similar once you get past the additional complications that dependence groups entail. We have to redo what we have already done, but we do not want to do everything twice. So on to dependence groups.

2.1 Why Dependence Groups?

In an aster model without dependence groups, nodes are conditionally independent given their predecessors; this is inherent in the factorization (4). There are at least two good reasons to relax this assumption.

2.1.1 Normal?

At terminal nodes of the graph, we may want to have normal random variables or other continuous random variables (that have exponential family distributions). (Recall from Section 1.4 that non-terminal nodes must be nonnegative-integer-valued.)

But aster models without dependence groups require one-parameter exponential families. So we could put in normal location families, and the R package aster does implement this. That assumes the variance for these arrows is known, which is annoying (because it never really is).

If we treat the univariate normal distribution as a two-parameter exponential family, then it has a two-dimensional canonical statistic, and its components, for sample size 1, are x and x^2, where x is what is usually considered the univariate normal random variable. And these are dependent random variables. So we either have to use some special magic for normal random variables, or we have to figure out how to put multiparameter exponential families into aster models.

One might think that the "special magic" of introducing scale parameters the way GLM do is the right way to deal with two-parameter normal components of the response vector, but scale parameters take us outside of exponential families and the aster transform would no longer work. Hence we go in another direction.

2.1.2 Multinomial?

Every univariate marginal of a multinomial random vector is a binomial random variable. But even a two-dimensional multinomial distribution is two-dimensional, so the binomial distribution is not a special case of the multinomial distribution.

A multinomial distribution is degenerate. It is the distribution of the random vector of category counts for IID individuals classified into categories with the categories being mutually exclusive and exhaustive, so every individual goes in exactly one category, and the category counts sum to the sample size. If x is one category count and there are two categories, then the other category count is $n - x$, where n is the sample size. Both x and $n - x$ are binomial, the multinomial random vector is $(x, n - x)$.

If the sample size is $n = 1$ and there are k categories, then a multinomial random vector serves as a k-way switch. Exactly one component is equal to one, and the rest are equal to zero. It indicates the category in which the individual ($n = 1$) is classified. This k-way switch can be useful in life history with, for example, animals with life history stages, such as (in insects) larva, pupa, and adult.

In most of statistics we do not want degenerate random vectors. We deal with this by dropping one of the components of a multinomial random vector to get a non-degenerate random vector. With aster models we cannot do that. Random variables correspond to nodes of the graph. Dropping random variables changes the graph. If we were to drop one category a k-way switch would become $(k-1)$-way. Not what was needed. Hence we have to deal with exponential families that have degenerate distributions of their canonical statistics and hence nonidentifiable canonical parameterizations.

We continue this thread in Section 2.6.8 below.

2.2 The Aster Axioms (With Dependence Groups)

2.2.1 Dependence Groups

Generalizing the two examples just presented, we will take a *dependence group* to be a subset of the response vector that has a joint exponential family distribution conditional on some other component (the predecessor), and we will still assume predecessor is sample size (Sections 1.4 and **??** above).

We need our dependence groups to be disjoint. And there is no reason to exclude dependence groups of size one, so every variable goes in some dependence group. Then the dependence groups form a partition $\mathcal{G}$ of the set J of non-initial nodes of the graph.

In order for predecessor is sample size to work in this new context, all of the nodes in a dependence group must have the same predecessor. For $G \in \mathcal{G}$ we write this predecessor as $q(G)$. Then we may consider q a function $\mathcal{G} \to N$. It is also useful to consider the function p we had before, which is now defined in terms of q by

$$p(j) = k \quad \text{if and only if} \quad j \in G \text{ and } q(G) = k$$

So now we have two predecessor functions: the set-to-node predecessor function q and the node-to-node predecessor function p. Since q determines p, the function q tells us everything there is to know about the dependence structure of the model. (We want to say, as we did before, that the predecessor function tells us everything there is to know about the graph, but we haven't said what the graph is yet. Still, as before, the predecessor function is all we need to know to specify the statistical model, so we will do statistical axioms first, graphical axioms later.)

The node-to-node predecessor function p no longer tells us everything there is to know about the dependence when there are dependence groups. But it is still useful in some situations.

2.2.2 Statistical Axioms

Let y_G denote the subvector of the response vector consisting of the components of the response vector in dependence group G. And we use similar notation for parameter vectors.

Note that we cannot consider these subvectors as having subscripts 1, 2, ... If G_1 and G_2 are the index sets for two different dependence groups, then we can never have $y_{G_1} = y_{G_2}$ even if the values of these subvectors are the same. The vectors have to know what their index sets are. The formal mathematical way to think of this is that y_G is a function $G \to \mathbb{R}$ rather than a tuple. And then we might as well think of y as a function $J \to \mathbb{R}$.

We now have two kinds of subscripts: y_j denotes a component of the response vector, and y_G denotes a subvector of the response vector, and we only have the convention dictating lower case letters for elements of sets and upper case letters for sets to help us keep straight which is which.

The joint distribution of the random variables factors as a product of conditional distributions

$$\prod_{G \in \mathcal{G}} \operatorname{pr}(y_G \mid y_{q(G)}) \tag{14}$$

This is still called the *factorization axiom.*

We now want to rely on Section 3.2.3 in Lauritzen (1996) to see that this is a valid factorization (that what are denoted as conditional distributions actually are conditional distributions). But again this is getting a bit ahead of ourselves, as we haven't said what the graph is yet.

The factorization with dependence groups (14) is very similar to the factorization without dependence groups (4), the only differences are that now we have subscripts that are subsets, and we use the set-to-node predecessor function. It takes a bit of getting used to that our subscripts in (14) denote sets G and run over a family of sets $\mathcal{G}$. But having swallowed that, the rest is very similar to the case where there are no dependence groups.

As before, only y_j for $j \in \bigcup \mathcal{G} = J$ appear "in front of the bar" in one of these conditional distributions. Hence only these y_j are treated as random. Hence y_j for $j \in N \setminus J$ are treated as constant random variables.

As before, conditioning on a constant random variable is the same as not conditioning, so such a conditional distribution is really a marginal distribution. Hence terms in (14) for which $q(G) \notin J$ are really marginal distributions.

Each conditional distribution in (14) obeys the *predecessor is sample size* axiom (Section 1.4 above). The only difference is now that this is the conditional distribution of a random vector y_G given a random scalar $y_{q(G)}$.

The final axiom is the *exponential family* axiom. Each conditional distribution in (14) is an exponential family with y_G as the canonical statistic vector and $y_{q(G)}$ as the sample size. We let θ_G denote the canonical parameter vector and let c_G denote the cumulant function for sample size 1, so

$$\theta_G \mapsto y_{q(G)} c_G(\theta_G)$$

is the cumulant function for sample size $y_{q(G)}$.

The log likelihood for an aster model with dependence groups is thus

$$l(\theta) = \sum_{G \in \mathcal{G}} \left[y_G \theta_G - y_{q(G)} c_G(\theta_G) \right]. \tag{15}$$

The log likelihood with dependence groups (15) is very similar to the log likelihood without dependence groups (5), the only differences are that now we have subscripts that are subsets, and we use the set-to-node predecessor function. It takes a bit of getting used to that our subscripts in (14) denote sets G and run over a family of sets $\mathcal{G}$. But having swallowed that, the rest is very similar to the case where there are no dependence groups.

2.2.3 Graphical Axioms

Lauritzen (1996), his equation (3.23), gives the most general factorization of a joint distribution. It is associated with a chain graph, which is a kind of general graph.

A general graph has both *directed edges*, also called *arrows*, which are drawn as arrows, and *undirected edges*, also called *lines*, which are drawn as line segments (no arrowheads). As before, we say arrows go from predecessors to successors. Lauritzen uses the alternate terminology, saying they go from parents to children. Now we also have lines, which Lauritzen says go between *neighbors*. And we will use the same terminology, at least temporarily.

A general graph that corresponds to a factorization is a *chain graph* (Lauritzen, 1996, Section 2.1.1) the node set has a partition $\mathcal{C}$, whose elements are called *chain components*, which is totally ordered by a relation $\prec$ such that there are arrows only between nodes in different chain components, there are lines only between nodes in the same chain component, and arrows agree with the total ordering ($j \longrightarrow k$ only if $j \in C_1$ and $k \in C_2$ and $C_1 \prec C_2$). From the graph, one determines the chain components by finding the connectivity components of the graph obtained from the original graph by keeping the lines and deleting the arrows. There may be many total orders compatible with the arrows. The algorithm called *topological sort* (Aho, et al., 1983, Section 6.6) finds one or proves that none exists (in which case we do not have a chain graph).

Lauritzen's equation (3.23) matches up with our (14) if our dependence groups are his chain groups and our $q(G)$ is his $\mathrm{pa}(G)$, except that our dependence groups partition the set of non-initial nodes and his chain groups partition the set of all nodes. This is not a problem because if we consider each initial node j to be a chain group by itself, then the additional terms Lauritzen has in his factorization are of the form $\mathrm{pr}(y_j \mid \varnothing)$ because j has no predecessor. And conditioning on nothing is the same as an unconditional distribution and the unconditional distribution of a constant random variable gives probability one to the (constant) value, and multiplication by one disappears.

Lauritzen (1996, Section 3.2.3) imposes a further condition on chain graph that governs whether the basic chain graph factorization, his equation (3.23), factorizes further, his equation (3.24). We do not want any further factorization. To guarantee this, we need a line between every two nodes in the same dependence group and an arrow $p(j) \longrightarrow j$ for every non-initial node j.

So we finally have our graph axioms.

- The dependence groups partition the set of non-initial nodes.
- Each dependence group has exactly one predecessor.
- There is an arrow to each node in a dependence group from the predecessor of the dependence group.
- There is a line between every two nodes in the same dependence group.
- The dependence groups can be totally ordered consistently with the predecessor relation. (Equivalently, the directed graph obtained by keeping the arrows and deleting the lines is acyclic.)

These axioms are more complicated than those without dependence groups (Section 1.7.1), so graphs with dependence groups are less intuitive than those without. Perhaps there is a simpler and more intuitive way to present the graph, but we stick with accepted graphical model theory with one exception (which follows).

When drawing the graphs, the fourth requirement here makes for a graph that is a little too cluttered. We only draw enough lines so that the whole dependence group is connected. As mentioned above, not having a line between each node in a dependence group would, in mainstream graphical models theory, indicate a "further factorization" (Lauritzen, 1996, equation (3.24)) where some of the terms in our (14) would themselves factorize.

Since in aster model theory we never want such "further factorization," we simply ignore this distinction. For us any nodes connected by a line are in the same dependence group, so the dependence groups are the maximal connected sets in the graph of lines (arrows ignored).

2.3 Life History Analysis with Dependence Groups

The graph (16) comes from a still unpublished manuscript for a book about aster models. It was the first graph for a model for an animal having life history stages like an insect's larva, pupa, and adult. We present this graph for hypothetical data even though a similar model has been fit to real data by Eck et al. (2015). Those data are for the tobacco hornworm *Manducca sexta*, which is an insect (moth) that does have these life history stages. These data were not collected with the intention of using an aster model (which were very new when the experiment was done) and so were not ideal for aster analysis. Although an aster analysis was done by Eck et al. (2015), it does not serve as quite as good an example as the graph (16).

$$\begin{tikzcd}
 & y_1 \arrow[d, dash] & y_4 \arrow[d, dash] & y_7 \arrow[d, dash] & y_{10} \arrow[d, dash] \\
1 \arrow[ur, "\mathcal{M}"] \arrow[r, "\mathcal{M}"] \arrow[dr, "\mathcal{M}"] & y_2 \arrow[ur, "\mathcal{M}"] \arrow[r, "\mathcal{M}"] \arrow[dr, "\mathcal{M}"] \arrow[d, dash] & y_5 \arrow[ur, "\mathcal{M}"] \arrow[r, "\mathcal{M}"] \arrow[dr, "\mathcal{M}"] \arrow[d, dash] & y_8 \arrow[ur, "\mathcal{M}"] \arrow[r, "\mathcal{M}"] \arrow[dr, "\mathcal{M}"] \arrow[d, dash] & y_{11} \arrow[d, dash] \\
 & y_3 \arrow[d, "\text{Ber}"] & y_6 \arrow[d, "\text{Ber}"] & y_9 \arrow[d, "\text{Ber}"] & y_{12} \arrow[d, "\text{Ber}"] \\
 & y_{13} \arrow[d, "\text{0-Poi}"] & y_{14} \arrow[d, "\text{0-Poi}"] & y_{15} \arrow[d, "\text{0-Poi}"] & y_{16} \arrow[d, "\text{0-Poi}"] \\
 & y_{17} & y_{18} & y_{19} & y_{20}
\end{tikzcd} \qquad (16)$$

As always, the constant 1 at the initial node of the graph indicates that the graph is for one individual. In addition to the notations Ber = Bernoulli and 0-Poi = zero-truncated Poisson, which we have already met, we now also have $\mathcal{M}$ = multinomial. Lines without arrowheads are "lines" connecting "neighbors" in the same dependence group. Hence the dependence groups containing more than one node are $\{1, 2, 3\}$, $\{4, 5, 6\}$, $\{7, 8, 9\}$, and $\{10, 11, 12\}$. Other nodes are dependence groups all by themselves; $\{j\}$ is a dependence group for $j \geq 13$.

Each of the multinode dependence groups has a conditional multinomial distribution with, as usual, predecessor as sample size. Since each predecessor is zero-or-one-valued, if a predecessor (say y_2) is equal to one, then exactly one of its three successor nodes (y_4, y_5, and y_6) is equal to one, and the probabilities are determined by the corresponding parameters (θ_4, θ_5, and θ_6), and, if this predecessor is equal to zero, then all of its three successor nodes are also equal to zero. In effect, exactly one of the "exterior nodes" of this group of switches (y_1, y_4, y_7, y_{10}, y_{11}, y_{12}, y_9, y_6, and y_3) is equal to one. There is one path taken by any particular individual, from the initial node (marked 1) through these four multinomial dependence groups.

The intended application for this graph (Eck et al., 2015) is life history data for an insect. As in our graphs without dependence groups (Section 1.2 above), "columns" of the graph are for different times (here days, there years). Nodes in the top "row" of this graph (y_1, y_4, y_7, and y_{10}) indicate death. Nodes in the second "row" of this graph (y_2, y_5, y_8, and y_{11}) indicate the individual is a larva (caterpillar). Nodes in the third "row" of this graph (y_3, y_4, y_9, and y_{12}) indicate the individual is an adult (moth, with wings, flying around trying to mate). Nodes in the fourth "row" of this graph (y_{13} through y_{16}) indicate mating success. Nodes in the bottom "row" of this graph (y_{17} through y_{20}) count number of eggs laid. So this graph is for female individuals. In Eck et al. (2015) the same graph with only the multinomial dependence groups (nodes 1 through y_{12}) is used for male individuals because the sex of individuals was not determined before they reached adulthood.

So this graph illustrates two important points not seen in Section 1.2. It is not necessary for every individual to have the same graph (here females and males have different graphs). And we have dependence groups, multinomial "switches" between different life history stages.

Here is yet another graph illustrating normal dependence groups.

$$\begin{tikzcd}
1 \arrow[r, "\mathrm{Ber}"] & y_1 \arrow[r, "\mathrm{Ber}"] \arrow[d, "\mathcal{N}"'] \arrow[dd, bend left, "\mathcal{N}"] & y_2 \arrow[r, "\mathrm{Ber}"] \arrow[d, "\mathcal{N}"'] \arrow[dd, bend left, "\mathcal{N}"] & y_3 \arrow[r, "\mathrm{Ber}"] \arrow[d, "\mathcal{N}"'] \arrow[dd, bend left, "\mathcal{N}"] & y_4 \arrow[d, "\mathcal{N}"'] \arrow[dd, bend left, "\mathcal{N}"] \\
& y_5 \arrow[d, dash] & y_6 \arrow[d, dash] & y_7 \arrow[d, dash] & y_8 \arrow[d, dash] \\
& y_9 & y_{10} & y_{11} & y_{12}
\end{tikzcd} \tag{17}$$

Here the top "row" indicates survival. And the next two rows are for normally distributed something or other given survival. Here we model the normal as two-node dependence groups $\{5, 9\}$, $\{6, 10\}$. $\{7, 11\}$, and $\{8, 12\}$ because we do not want to assume variance is known. As we shall see, this permits but does not require, modeling variance as a function of covariates.

We see that the aster formalism suggests new possibilities. In order to have a two-parameter normal distribution, we need two-node dependence groups. But this makes us treat variance parameters as just like other parameters, so we can model them too.

Actually, we shouldn't say "variance" parameters. As the reader should know, the canonical parameters for the two-parameter normal are not the usual parameters. The log likelihood for sample size one is

$$-\frac{(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(\sigma)$$

and the terms containing both parameters and data are

$$-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2}$$

so, if we take the canonical statistic vector to be (x, x^2), then the canonical parameter vector is $(\mu/\sigma^2, -1/2\sigma^2)$, and these are what we model as a function of covariates.

2.4 The Aster Transform (With Dependence Groups)

Now that we have gotten past what dependence groups are and why we would want them and the peculiarities of their notation (with subscripts that are sets ranging over families of sets), everything moves rapidly. The aster transform looks almost the same with and without dependence groups. The aster transform without dependence groups (6) becomes with dependence groups

$$\varphi_j = \theta_j - \sum_{\substack{G \in \mathcal{G} \\ q(G) = j}} c_G(\theta_G), \tag{18}$$

and the cumulant function of the joint distribution without dependence groups (7) becomes with dependence groups

$$c(\varphi) = \sum_{\substack{G \in \mathcal{G} \\ q(G) \notin J}} y_{q(G)} c_G(\theta_G). \tag{19}$$

As before, (18) is an invertible change of parameter, and its inverse (the inverse aster transform) is given by

$$\theta_j = \varphi_j + \sum_{\substack{G \in \mathcal{G} \\ q(G) = j}} c_G(\theta_G), \tag{20}$$

which works when nodes are visited in any order that has successors before predecessors.

We still call θ the conditional canonical parameter vector and φ the unconditional canonical parameter vector and the log likelihood for φ still has the simple expression (9).

2.5 Canonical Affine Submodels (With Dependence Groups)

Nothing in Section 1.11 needs to be rewritten, since it only used vector and matrix notation. The aster transform is there working behind the scenes, but we don't see it in the notation.

2.6 More Theory of Exponential Families

2.6.1 Densities

We mostly do not need the densities for an exponential family. As we saw so far, we mostly can reason with log likelihoods.

But it will help with some things if we have densities. For that we need a dominating measure of the family, and we can always pick that to be a probability measure in the family (because all distributions in the family are absolutely continuous with respect to each other). Suppose the dominating measure is taken to be the distribution in the family with canonical parameter vector φ^*. Then the densities are

$$f_\varphi(\omega) = e^{l(\varphi) - l(\varphi^*)} = e^{\langle y(\omega), \varphi - \varphi^* \rangle - c(\varphi) + c(\varphi^*)} \tag{21}$$

(recall that y is just a vector statistic, a function of the whole data ω not involving parameters).

2.6.2 Cumulant Functions

In order for (21) to integrate to one, we must have

$$c(\varphi) = c(\varphi^*) + \log E_{\varphi^*}\left(e^{\langle y, \varphi - \varphi^* \rangle}\right) \tag{22}$$

This, thought of as a function of φ for fixed φ^*, determines the cumulant function up to an unknown constant $c(\varphi^*)$, which is arbitrary. (An arbitrary constant can be added to the cumulant function without changing anything else about the model).

2.6.3 Full Families

Following Barndorff-Nielsen (1978) we think of (22) as defining the cumulant function up to an arbitrary constant $c(\varphi^*)$ for all φ in the vector space where the canonical parameter takes values. For φ such that the expectation in (22) does not exist, we define $c(\varphi) = +\infty$. The *full* exponential family having this cumulant function has canonical parameter space

$$\Phi = \{\, \varphi : c(\varphi) < +\infty \,\}. \tag{23}$$

Then (21) defines a distribution for every $\varphi \in \Phi$.

2.6.4 Moment Generating Functions

The moment generating function (MGF) for the canonical statistic vector is defined by

$$\begin{aligned} M_\varphi(t) &= E_\varphi\left(e^{\langle y, t \rangle}\right) \\ &= E_{\varphi^*}\left(e^{\langle y, t + \varphi - \varphi^* \rangle - c(\varphi) + c(\varphi^*)}\right) \\ &= e^{c(\varphi + t) - c(\varphi)} \end{aligned}$$

provided this defines an MGF, that is, provided M_φ is finite on a neighborhood of zero, which happens when the cumulant function c is finite on a neighborhood of φ, which happens when φ is an interior point of the full canonical parameter space (23).

The MGF is so called because its partial derivatives evaluated at zero are ordinary moments. For an exponential family, partial derivatives of M_φ evaluated at zero are ordinary moments of the canonical statistic vector.

2.6.5 Cumulant Generating Functions

The logarithm of the MGF is called the *cumulant generating function* (CGF) because its partial derivatives evaluated at zero are (by definition) the *cumulants* of the distribution. Cumulants of order m are polynomial functions of the ordinary moments up to order m and vice versa (Cramér, 1951, Section 15.10). For an exponential family, the CGF is

$$K_\varphi(t) = c(\varphi + t) - c(\varphi)$$

Note that partial derivatives of K_φ evaluated at zero are partial derivatives of c evaluated at φ. Hence the name cumulant function.

2.6.6 Regular Full Families

A full exponential family is *regular* if its canonical parameter space (23) is an open set. Every distribution in a full exponential family has an MGF and a CGF if and only if the family is regular. All aster models that have been discussed in print or that are implemented in software are full regular exponential families.

An aster model is a full regular exponential family if each conditional exponential family for each dependence group is a full regular exponential family. A proof of this (which is long but merely establishes that what seems obvious actually is obvious) can be found in Appendix E of Geyer, et al. (2005). (This proof is only for aster models without dependence groups, because dependence groups had not yet been added when that technical report was written, but is easily modified in an obvious way for dependence groups.)

2.6.7 Cumulants

We are only interested in the first two cumulants, which happen to be the mean and variance

$$E_\varphi(y) = \nabla c(\varphi) \tag{24a}$$

$$\operatorname{var}_\varphi(y) = \nabla^2 c(\varphi) \tag{24b}$$

These can also be derived via the Bartlett identities (identities for log likelihood derivatives derived by differentiating under the integral sign the fact that probability densities integrate to one), but that involves an argument that differentiation under the integral sign is always valid in regular exponential families and does not explain the term "cumulant function" so we prefer this route.

Of course, (24a) and (24b) only are valid for φ in the interior of the full canonical parameter space (23). Hence they are only valid for all φ if the family is regular.

REVISED DOWN TO HERE

2.6.8 Identifiability

In a full exponential family, if φ_1 and φ_2 are values of the canonical parameter that correspond to the same probability distribution, then $\varphi_2 - \varphi_1$ is called a *direction of constancy* of the family. This is only one of eight equivalent characterizations of this concept tied together by Theorem 1 in Geyer (2009). Another characterization is that δ is a direction of constancy if and only if φ and $\varphi + s\delta$ correspond to the same distribution for all φ in the full canonical parameter space and all real s. Another characterization is that δ is a direction of constancy if and only if $\langle y, \delta \rangle$ is a constant random variable, where y is the canonical statistic vector.

It follows from the characterization we adopted as a definition here that directions of constancy completely characterize identifiable parameterizations in an exponential family. The only kind of nonidentifiability an exponential family can have is that if φ_1 and φ_2 are not identifiable, then $\varphi_2 - \varphi_1$ is a direction of constancy.

By Corollary 2 in Geyer (2009) if $\hat{\varphi}_1$ and $\hat{\varphi}_2$ are canonical parameter values, and each is a maximum likelihood estimate (MLE), then $\hat{\varphi}_2 - \hat{\varphi}_1$ is a direction of constancy of the family. Thus we have identifiability of a sort where it really counts: the probability distribution associated with the MLE is always unique. If $\hat{\varphi}$ is an MLE and δ is a direction of constancy, then $\hat{\varphi} + s\delta$ is also an MLE for all real s, but all of these parameter values correspond to the same distribution.

Thus (when dealing with exponential families) we can learn to live without identifiability. Nonidentifiability is, as Geyer (2009) says, "merely a computational nuisance." If we are going to impose identifiability at some point (merely to simplify computation), then we should do it as late as possible, not allowing it to distort our theoretical discussion.

REVISED DOWN TO HERE

2.7 Identifiability of Aster Models

So what does this have to do with aster models? It depends on which R package we are using.

2.7.1 Package 1

Aster models allowed by the R package `aster` are identifiable unless the model matrix is not full rank, in which case they have the same sort of non-identifiability that can arise in GLM and LM, and this non-identifiability is handled in the same way. Some columns of the model matrix are dropped to obtain a full rank model matrix having the same column space as the original model matrix. Then the model is identifiable.

The only difference in the behavior between the aster package and R core functions is that the R function summary has lines with all values NA for the dropped columns when summarizing an R object of class "lm" or "glm" but has no such lines when the object is of class "aster". (The method summary.aster has the behavior that the methods summary.lm and summary.glm used to have years ago.) One can find the names of the dropped columns in the dropped component of an R object of class "aster".

2.7.2 Package 2

Aster models allowed by the R package aster2 are also non-identifiable when the model matrix is not full rank, but they also allow two other kinds of non-identifiability: those resulting from having multinomial dependence groups and those of limiting conditional models.

Multinomial Dependence Groups Multinomial dependence groups are described in Section 2.1.2 above. If y_G is the canonical statistic for a multinomial dependence group, then

$$\sum_{j \in G} y_j = y_{q(G)} \tag{25}$$

with probability one. The left-hand side is $\langle y_G, \delta_G \rangle$, where δ_G is the vector having all components equal to one. Thus δ_G is a direction of constancy of the conditional distribution of y_G given $y_{q(G)}$. There being no other linear function of the canonical statistic vector that is almost surely constant, this is the only direction of constancy. (This statement is wrong for limiting conditional models. They can have additional constraints $y_j = 0$ almost surely for some $j \in G$.)

That describes directions of constancy for the conditional distributions for dependence groups. It also describes the directions of constancy for the conditional canonical parameter vector. For each multinomial dependence group G define a vector ζ^G having components

$$\zeta^G_j = \begin{cases} 1, & j \in G \\ 0, & j \in J \setminus G \end{cases} \tag{26}$$

And θ and $\theta + \delta$ induce the same conditional distributions for all dependence groups if and only if and only if δ is a linear combination of these ζ^G vectors.

The same degeneracy (25) is the key to directions of constancy for the unconditional canonical parameter vector. The only difference is that when we are thinking unconditionally $y_{q(G)}$ may or may not be random depending on whether or not $q(G)$ is a non-initial node. For each multinomial dependence group G define a vector ζ^G having components

$$\zeta_j^G = \begin{cases} 1, & j \in G \\ -1, & j = q(G) \\ 0, & j \in J \setminus (G \cup \{q(G)\}) \end{cases} \tag{27}$$

Notice that ζ^G has a component equal to -1 if $q(G)$ is not an initial node and does not have such a component if $q(G)$ is an initial node (then the middle case does not hold for any $j \in J$). And φ and $\varphi + \delta$ induce the same joint distribution of the response vector for all dependence groups if and only if and only if δ is a linear combination of these ζ^G vectors.

None of this complexity can be avoided. As hinted at in Section 2.1.2 above, we cannot avoid directions of constancy and keep all of the components of the multinomial in the response vector, which we have to do to have the multinomial play its role as multiway switch.

If we have a nonidentifiable conditional canonical parameterization θ, then the aster transform (18) takes us to a nonidentifiable unconditional canonical parameterization φ. Moving to an unconditional affine submodel may take us to an identifiable submodel parameterization β or may introduce additional directions of constancy if the model matrix is not full rank. In any event, the theory of directions of constancy for exponential families keeps nonidentifiability under control.

A canonical affine submodel (conditional or unconditional) will be identifiable if and only if the column space of the model matrix contains no direction of constancy of the saturated model, that is, it contains none of the ζ^G defined above (in different ways for conditional and unconditional submodels).

Limiting Conditional Models

2.8 Mean Value Parameters

2.8.1 Regular Full Exponential Families

By (24a), the map h defined by

$$h(\varphi) = \nabla c(\varphi) = E_\varphi(y) \tag{28}$$

takes the canonical parameter φ of a regular full exponential family to the mean of the canonical statistic y for the distribution having this parameter value. The Jacobian matrix of this change of parameter is defined by

$$\nabla h(\varphi) = \nabla^2 c(\varphi) = \operatorname{var}_\varphi(y)$$

by (24b). If δ is a direction of constancy, we have

$$h(\varphi + s\delta) = h(\varphi)$$

for all real s because φ and $\varphi + s\delta$ correspond to the same distribution, hence to the same means. Conversely, if δ is not a direction of constancy, we have

$$\left. \frac{d\delta^T h(\varphi + s\delta)}{ds} \right|_{s=0} = \delta^T \bigl[\nabla h(\varphi)\bigr] \delta = \delta^T \bigl[\operatorname{var}_\varphi(y)\bigr] \delta = \operatorname{var}_\varphi(\delta^T y) > 0$$

because $\delta^T y$ is a constant random variable if and only if δ is a direction of constancy (Section 2.6.8 above). Hence by the fundamental theorem of calculus

$$\delta^T h(\varphi + s\delta) > \delta^T h(\varphi)$$

whenever $\delta \neq 0$ and $s > 0$. This proves that $h(\varphi_1) = h(\varphi_2)$ if and only if $\varphi_2 - \varphi_1$ is a direction of constancy.

The parameter $\mu = h(\varphi)$ is thus always identifiable, and the map h is one-to-one if and only if the canonical parameterization is identifiable. This parameter is called the *mean value* parameter of the exponential family.

The proof just given (that the mean value parameterization is always identifiable) can be extended to nonregular full exponential families with the proviso that a distribution in the family having canonical parameter value φ that is on the boundary of the canonical parameter space (23) need not have a mean value (the expectation of the canonical parameter vector need not exist). For the mean values that do exist, each different mean corresponds to a different distribution.

2.8.2 Aster Models: Unconditional Mean Values

Assuming our aster model is a regular full exponential family (which all currently implemented are), Section 2.8.1 above applies, and

$$\mu = \nabla c(\varphi) = E_\varphi(y) \tag{29}$$

is an identifiable parameterization of the aster model. We call μ the *unconditional mean value parameter.*

2.8.3 Aster Models: Conditional Mean Values

Again assuming our aster model is a regular full exponential family, Section 2.8.1 above also applies to each dependence group. Define

$$\xi_G = \nabla c_G(\theta_G) = E_\varphi(y_G \mid y_{q(G)} = 1). \tag{30}$$

Then the vector ξ having subvectors ξ_G defined this way is called the *conditional mean value parameter.*

There is a technical quibble that goes with (30). It could be the case that the distribution of $y_{q(G)}$ gives probability zero to the value 1, in which case $E_\varphi(y_G \mid y_{q(G)} = 1)$ makes no sense (is undefined), but $\nabla c_G(\theta_G)$ is still well defined and defines ξ_G. In this case, we can still interpret ξ_G but must describe it more longwindedly: y_G is, by the predecessor is sample size property, the sum of $y_{q(G)}$ IID random vectors whose distribution has cumulant function c_G and ξ_G is the mean vector of each of those IID random vectors. Whether or not the event $y_{q(G)} = 1$ can occur, we have, from this latter description

$$E_\varphi(y_G \mid y_{q(G)}) = y_{q(G)} \xi_G$$

and we can rewrite this using the node-to-node predecessor function

$$E_\varphi(y_j \mid y_{p(j)}) = y_{p(j)} \xi_j. \tag{31}$$

Now we come to a confession. Geyer, et al. (2007) did not define the conditional mean value parameters the way we do here. Instead they called (31) the conditional mean value parameter. A referee said this definition is dumb. It is a function of both random variables $y_{p(j)}$ and parameters ξ_j and so shouldn't be called a parameter. We didn't listen then and managed to get the paper published overriding this objection. But now we agree with the referee. The R package `aster` uses the same dumb definition. The R package `aster2` and recent papers and technical reports use the definition presented here (the conditional mean value parameter vector is ξ) if they mention conditional mean value parameters at all.

2.8.4 The Combination of the Two

Taking expectations in (31) gives

$$\mu_j = \mu_{p(j)} \xi_j. \tag{32}$$

This holds for all $j \in J$ if we define μ_j when j is an initial node to be the mean of the constant random variable y_j at the initial node, and, of course, the expectation of a constant is a constant, $\mu_j = y_j$ when j is initial.

Equation (32) implicitly characterizes the mapping between μ and ξ, which is invertible.

To map from ξ to μ we use (32) recursively

$$\begin{aligned} \mu_j &= \xi_j \mu_{p(j)} \\ &= \xi_j \xi_{p(j)} \mu_{p(p(j))} \\ &= \xi_j \xi_{p(j)} \xi_{p(p(j))} \mu_{p(p(p(j)))} \end{aligned}$$

and so forth. With as many recursive applications as necessary. In practice, the computer traverses the graph in any order that visits predecessors before successors using (32) to determine μ_j as a function of ξ ($\mu_{p(j)}$ having already been determined when its node was visited previously).

To map from μ to ξ, rewrite (32) as

$$\xi_j = \frac{mu_j}{\mu_{p(j)}} \tag{33}$$

but for this to make sense, we must know that $\mu_{p(j)}$ is never zero. We do know that every predecessor $y_{p(j)}$ is nonnegative-valued, hence $\mu_{p(j)} = 0$ if and only if $y_{p(j)} = 0$ almost surely. But that happens only in one of three cases, which must be ruled out separately. Case I: $p(j)$ is an initial node, so $y_{p(j)}$ is constant, and that constant is zero. This case is not allowed by aster software, either R package `aster` or R package `aster2`. Case II: the conditional distribution of $y_{p(j)}$ given $y_{p(p(j))}$ is degenerate, concentrated at zero. This case is not allowed by aster software, either R package `aster` or R package `aster2`, except that `aster2` does allow it for limiting conditional models, in which case (33) gives zero over zero and ξ_j is undefined. But we do not have to worry about this case until limiting conditional models are discussed. Case III: $\mu_{p(p(j))} = 0$. This case we rule out by induction. If no initial node j has $\mu_j = 0$ and if no conditional distribution of $y_{p(j)}$ given $y_{p(p(j))}$ is concentrated at zero, then we cannot have $\mu_{p(j)} = 0$ by induction over the graph (moving from predecessors to successors).

In summary (33) is valid for all aster models so long as no initial node is allowed to be zero (which aster software enforces) and so long as no conditional distribution for a predecessor is allowed to be degenerate, concentrated at zero (which aster software enforces). But (33) need not be valid for limiting conditional distributions of aster models (which we have not discussed yet).

2.8.5 Mapping Mean Value Parameters to Canonical Parameters

We have seen that the mapping h defined by (28) is invertible if and only if the canonical parameterization is identifiable. If invertible, how do we calculate the inverse? Essentially, this is the same procedure as maximum likelihood, except we use mean values rather than data. Let μ be a possible value of the unconditional mean value parameter vector, and define a function l by

$$l(\varphi) = \mu^T \varphi - c(\varphi)$$

where c is the cumulant function of the family. This is like the log likelihood of a general exponential family (**??**) except we have replaced data y by its expected value μ and we are denoting the canonical parameter by φ instead of θ. The cumulant function c is always convex because its second derivative matrix is always positive semidefinite (24b). Hence l is always concave. It follows that any zero of ∇l maximizes l. But

$$\nabla l(\varphi) = \mu - \nabla c(\varphi) = \mu - h(\varphi)$$

where h is the mapping h defined by (28) from canonical to mean value parameter, and to say that μ is a possible mean value is to say that $\mu = h(\varphi)$ for some φ. Hence a zero of ∇l always exists. It will be unique if the canonical parameterization is identifiable, and otherwise any φ such that $\mu = h(\varphi)$ is a solution (and all solutions differ by a direction of constancy).

The same analysis holds for conditional distributions for dependence groups, only the notation is more cluttered. Let G be a dependence group and ξ_G its conditional mean value parameter vector, and define

$$l_G(\theta_G) = \xi_G^T \theta_G - c_G(\theta_G)$$

Then maximizers always exists and are any θ_G satisfying $\xi_G = \nabla c_G(\theta_G)$.

References

Aho, A. V., Hopcroft, J. E., and Ullman, J. D. (1983). *Data Structures and Algorithms.* Addison-Wesley, Reading, MA.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families.* Wiley, Chichester.

Cramér, H. (1951). *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

Eck, D., Shaw, R. G., Geyer, C. J., and Kingsolver, J. (2015). An integrated analysis of phenotypic selection on insect body size and development time. *Evolution*, **69**, 2525–2532.

Geyer, C. J. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, **3**, 259–289.

Geyer, C. J. (2013). Stat 8112 Lecture Notes: Asymptotics of Exponential Families. `http://www.stat.umn.edu/geyer/8112/notes/expfam.pdf`.

Geyer, C. J. (2015a). R package `aster` (Aster Models), version 0.8-31. `http://cran.r-project.org/package=aster`.

Geyer, C. J. (2015b). R package `aster2` (Aster Models), version 0.2-1. `http://cran.r-project.org/package=aster2`.

Geyer, C. J. (2016). Stat 8931 (Exponential Families) Lecture Notes: The Eggplant that Ate Chicago (the Notes that Got out of Hand). `http://www.stat.umn.edu/geyer/8931expfam/convex.pdf`.

Geyer, C. J., Ridley, C. E., Latta, R. G., Etterson, J. R., and Shaw, R. G. (2013). Local adaptation and genetic effects on fitness: Calculations for exponential family models with random effects. *Annals of Applied Statistics*, **7**, 1778–1795.

Geyer, C. J., Wagenius, S., and Shaw, R. G. (2005). Aster models for life history analysis. Technical Report No. 644. School of Statistics, University of Minnesota. `http://www.stat.umn.edu/geyer/aster/`.

Geyer, C. J., Wagenius, S., and Shaw, R. G. (2007). Aster models for life history analysis. *Biometrika*, **94**, 415–426.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

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

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`.

Shaw, R. G., and C. J. Geyer (2010). Inferring fitness landscapes. *Evolution*, **64**, 2510–2520.

Shaw, R. G., Geyer, C. J., Wagenius, S., Hangelbroek, H. H., and Etterson, J. R. (2008). Unifying life history analysis for inference of fitness and population growth. *American Naturalist*, **172**, E35–E47.

Shaw, R. G., Wagenius, S., and Geyer, C. J. (in press). *Echinacea angustifolia* and its specialist aphid: the roles of plant phenotype and genotype. To appear in *Journal of Ecology*.

Stanton-Geddes, J., Tiffin, P., and Shaw, R. G. (2012). Role of climate and competitors in limiting fitness across range edges of an annual plant. *Ecology*, **93**, 1604–1613. Supplemental material, *Ecological Archives*, E093-142-A1. `http://esapubs.org/archive/ecol/E093/142/appendix-A.htm`.