Theory of Aster Models

Charles J. Geyer

April 6, 2024

Preface

This is a book about the theory of aster models. Its main intended readers are implementers of aster software. So far, that is just the author. The reason for the book is to go into all the gory details so the software can do the Right Thing.

Other readers may also find this book useful in having all of the theory of aster models in one place and presented with consistent notation. But they may have to skip a lot of material they don't want to read.

License

Contents

Chapter 1

Introduction

1.1 Background

Aster models (Geyer, Wagenius, and Shaw, 2007; Shaw, Geyer, Wagenius, Hangelbroek, and Etterson, 2008b; Geyer, Ridley, Latta, Etterson, and Shaw, 2013) are parametric statistical models specifically designed for life history analysis. They are exponential family models that generalize generalized linear models (GLM) that are also exponential family models (for example, logistic regression and Poisson regression with log link) in two ways

- in GLM components of the response vector are conditionally independent given covariate data but in aster models they need not be, and
- in GLM the conditional distributions of components of the response vector given covariate data all come from the same family but in aster models they need not.

As generalizations of GLM, aster models are also regression models. They model the conditional distribution of the response vector given covariate data. The marginal distribution of covariate data is not modeled.

In life history analysis, the data are about survival and reproduction of biological organisms. Thus aster models also generalize discrete time survival analysis (aster models model not only survival but also what happens conditional on survival). Aster models unify many disparate kinds of life history analysis that have appeared in the biological literature: comparison of Darwinian fitness between various groups (Geyer et al., 2007; Shaw et al., 2008b), estimation of fitness landscapes (Lande and Arnold, 1983; Mitchell-Olds and Shaw, 1987; Shaw et al., 2008b; Shaw and Geyer, 2010, Eck, Shaw,

Geyer, and Kingsolver, 2015), Leslie matrix analysis (Caswell, 2001), life table analysis in demography (Goodman, 1968), and estimation of population growth rates (Fisher, 1930; Lenski and Service, 1982; Shaw et al., 2008b; Eck et al., 2015). Aster models also generalize zero-inflated Poisson regression (Lambert, 1992), negative binomial regression (overdispersed Poisson regression), and zero-inflated negative binomial regression.

Aster models are a special case of graphical models (Lauritzen, 1996). In particular, they are statistical models for which the joint distribution of the response vector factorizes as a product of marginal and conditional distributions (equation (1.1) below). This makes aster models a special case of chain graph models (Lauritzen, 1996, Sections 2.1.1 and 3.2.3).

Aster models also have the predecessor-is-sample-size property (Section 1.12 below) that makes the joint distribution of the response vector an exponential family. This property can be seen to generalize unnamed properties of survival analysis, life-table analysis, Leslie matrix analysis, and population growth rate analysis (Section 1.13.3 below).

1.2 Software

Currently, all software for aster models is written in the R statistical computing language (R Core Team, 2023). There are two CRAN (`cran.r-project.org`) packages, `aster` (Geyer, 2021) and `aster2` (Geyer, 2017a). Both R and these packages can be installed in minutes on any computer, so any user can get started with aster models in almost no time.

R package `aster` is the most complete. It does everything except dependence groups and limiting conditional models.

R package `aster2` is the very incomplete. It does do dependence groups and limiting conditional models, but everything else is either missing or much harder to use than in R package `aster`. The only analyses known to us using R package `aster2` are Eck et al. (2015) and May, Shaw, Geyer, and Eck (2022).

So any aster model that can be done with R package `aster` should be done with that package.

1.3 Summary

Aster models combine three ideas,

- factorization of a joint distribution into a product of marginals and conditionals (Section 1.6 below),

- the predecessor is sample size property (Section 1.12 below), which says the conditioning variables in the conditional distributions in the factorization act like sample sizes, and
- the exponential family property (Section 1.16 below), which says the conditional distributions in the factorization are exponential families of distributions,

into one big idea. Together they make the joint distribution of the response vector also an exponential family. And this makes aster models as well-behaved as generalized linear models or log-linear models for categorical data analysis, even though they are far more more complicated.

1.4 Vectors and Subvectors

We adopt a notation from Lauritzen (1996) for subvectors, but fuss about it more. The idea that vectors can have arbitrary index sets, rather than indices that are consecutive numbers starting at one, appears earlier in Rockafellar (1988). The idea that vectors are functions occurs everywhere in functional analysis (Rudin, 1991, Appendix B).

As in set theory (Halmos, 1960, Section 8), if A and B are sets, then A^B denotes the set of all functions $B \to A$. In particular, if J is a finite set, then we let $\mathbb{R}^J$ denote the set of all functions $J \to \mathbb{R}$. This set can also be considered a finite-dimensional vector space. That functions $J \to V$ where J is any set and V is a field or a vector space can be considered vectors is the reason the study of infinite-dimensional topological vector spaces is called functional analysis.

Another way of looking at this distinction is that the usual view of finite-dimensional vector spaces is that they are $\mathbb{R}^d$ for some natural number d, which is tantamount to insisting that the index set for vectors in this space must be the set $\{1, \ldots, d\}$. Here we are saying the index set can be any finite set J.

Even though we consider vectors to be functions, we write evaluation of these functions y_j so it looks like usual notation. We even say that y_j is a component of the vector y rather than the value of the function y at the point j. But behind the scenes our vectors are also functions, and we could write $y(j)$ instead of y_j.

We need a notation for subvectors of a vector. If y is an element of $\mathbb{R}^J$ and $A \subset J$, then we let y_A denote the restriction of y to the set A. As such, it is an element of the vector space $\mathbb{R}^A$. Like all functions, it knows

its domain and codomain. It knows it is a function $A \to \mathbb{R}$. So it knows its components are y_j, $j \in A$. And these are also the components of y for $j \in A$. Since the components of y_A are a subset of the components of y, we say y_A is a *subvector* of y.

If we were to insist that all vectors, including subvectors, have index sets $\{1, \ldots, k\}$ for some natural number k, then we could not distinguish different subvectors of the same length, or at least could not without ugly and cumbersome extra decoration of the notation. (It is hard to explain how elegant this notation is with simple examples, but a perusal of Appendices B and C will show this notation is extremely powerful, and those appendices would be much longer and more confusing if we had to use conventional notation with indices going from 1 to k.)

Our notation does have the drawback that we have only the convention that lower case letters denote elements of sets and upper case letters denote sets to indicate that y_j is a component of a vector or subvector (the value of a function at the index j) and y_A is a subvector (the restriction of a function to the set A, so still a function, not the value of a function). We also consider any subscript notation that clearly denotes a set as indicating a subvector, for example, $y_{\{1,3,5\}}$ or $y_{\{j\}}$ or $y_{\{\, j \in J : j \prec i \,\}}$.

The conventional notation for restriction of a function in mathematics is $y|A$ or $y_{|A}$. If we used this, it would be more definite whether a subvector is intended. Perhaps our usage, which agrees with Lauritzen, is ill-advised. But we will continue with it.

It will not be important in what follows, but a special subvector $y_\varnothing$ has the empty index set. It is a function $\varnothing \to \mathbb{R}$. One might wonder what that is, but mathematicians define such. For any set A there is exactly one function $\varnothing \to A$, which is called the empty function. It has an empty graph (no argument-value pairs) because there are no possible argument values. That is, if $f : \varnothing \to A$ is an empty function, then it is valid to write $f(x)$ for all x in the domain of f, which is the empty set, so there are no such x. But there is the function. And there is exactly one such function (for each set A) because there is only one empty set and only one subset of $\varnothing \times A = \varnothing$ that is the graph of f. We also want to consider the vector space $\mathbb{R}^\varnothing$. Since this has only one element, the empty function $\varnothing \to \mathbb{R}$, this must be a vector space having only one element. Every vector space has a zero vector. So this single element must be the zero vector of the vector space $\mathbb{R}^\varnothing$. So the empty function $f : \varnothing \to \mathbb{R}$ must be the zero vector of this vector space. And this makes a crazy kind of mathematical sense: we do have $f(x) = 0$ for all x in the domain of f, just like for any other zero vector. The only tricky part is that there are no such x. But that is OK to mathematicians. Quantifiers

can range over any set, including the empty set, so the statement $f(x) = 0$ for all $x \in \varnothing$ is a true mathematical statement. We say $f(x) = 0$ holds *vacuously*, meaning there are no x that need to be considered. Whether one finds this paragraph interesting or just crazy, it is an inevitable part of the vectors-are-really-functions point of view. It is always there in the background, whether or not we pay any attention to it.

When we consider random vectors Y, then Y_A is also a random vector for any A that is a subset of the index set (domain) of Y. This includes $Y_\varnothing$. But since $Y_\varnothing$ has only one possible value, the empty function $y_\varnothing$, this must be a constant random vector (one that takes the same value with probability one).

Of course since $Y_\varnothing$ and $y_\varnothing$ are zero vectors of the vector space $\mathbb{R}^\varnothing$ in which they take values, we can write 0 instead of $Y_\varnothing$ or $y_\varnothing$.

1.5 Regression Notation

Strictly speaking, in regression theory, every probability and expectation is conditional on covariate data, at least on the part of the covariate data that is considered random rather than fixed by the design of the experiment. Thus to be hyperpedantic, we should always write

$$E(Y_A \mid \text{the part of covariate data that is random})$$
$$\Pr(Y_A \in B \mid \text{the part of covariate data that is random})$$

rather than $E(Y_A)$ or $\Pr(Y_A \in B)$. But, like most regression books, we will not do this. The dependence of probabilities and expectations on covariates is usually not made explicit in the notation.

This is especially important in aster models when components of the response vector depend on the values of other components, so we frequently write

$$E(Y_A \mid y_j)$$
$$\operatorname{pr}(y_A \mid y_j)$$

and the like (Section 1.6 below). And we do not want this dependence confused with dependence on covariate data.

When necessary for clarity, as in the discussion of fitness landscapes, which are regression functions, we can explicitly denote the dependence on covariate data in conditional probabilities and expectations.

1.6 Factorization

Let J be the index set of the response vector y of an aster model. We also need more variables that are not in the response vector. As we shall see they are treated as constants. Let N be the index set of all the variables, so $J \subset N$. Then there is a partition $\mathcal{G}$ of J and a function $q : \mathcal{G} \to N$, such that the joint distribution of y factorizes as

$$\operatorname{pr}(y) = \prod_{G \in \mathcal{G}} \operatorname{pr}(y_G \mid y_{q(G)}) \tag{1.1}$$

We emphasize that $q : \mathcal{G} \to N$ maps elements of $\mathcal{G}$, which are sets of indices, to elements of N, which are single indices, so each y_G in (1.1) is a subvector of the response vector y, and each $y_{q(G)}$ in (1.1) is a component of the vector of all the variables.

In this factorization, each component y_j of the response vector y appears exactly once "in front of the bar" in a conditional distribution on the right-hand side (because $\mathcal{G}$ is a partition of J so each $j \in J$ is in exactly one $G \in \mathcal{G}$). So every component of y is treated as random (the joint distribution of y is modeled). Random variables $y_{q(G)}$ that appear "behind the bar" in a conditional on the right-hand side may or may not be components of y. They are not if $q(G) \notin J$. The distribution of such random variables is not modeled by (1.1). So they are treated as constant random variables.

We say (1.1) is *valid* if what are denoted as conditional distributions on the right-hand side agree with the conditional distributions derived from the left-hand side (the joint distribution) by the usual operations of probability theory.

Theorem 1.1. *The factorization* (1.1) *is valid if and only if the partition* $\mathcal{G}$ *can be totally ordered by some total ordering* $<$ *such that* $q(G) \in H$ *implies* $G < H$.

A proof of this theorem is straightforward and given in Appendix A. It could also be derived from the discussion of chain graph models in Lauritzen (1996, equation 3.23).

In (1.1) we have been deliberately vague about what pr is supposed to mean, since there are many ways to specify probability distributions and any of them will do.

- If y is a discrete random vector, then pr could denote probability mass functions.

- If y is a continuous random vector, then pr could denote probability density functions.
- If y is a partly discrete and partly continuous continuous random vector (either some components discrete and some components continuous or some components a mixture of discrete and continuous) then pr could denote probability mass-density functions.
- No matter what, pr could denote cumulative distribution functions.
- No matter what, pr could denote probability measures and regular conditional probability measures (also called Markov kernels).

In any of these cases the multiplication indicated in (1.1) is actual multiplication of real-valued thingummies.

1.6.1 Topological Sort

The total order asserted to exist by the theorem need not be unique and usually is not unique. We can find such a total order using the algorithm called topological sort (Aho et al., 1983, Section 6.6).

This algorithm, given a partial order on a set, finds a total order that extends the partial order, or proves that none exists (and gives an error message). R function `tsort` in R package `pooh` (Geyer, 2017b) implements this algorithm.

The set we use here is $\mathcal{G}$ and the partial order on it is a list of pairs (G, H) of elements of $\mathcal{G}$ such that $q(G) \in H$. These are supplied to R function `tsort` as arguments `from` and `to`, which are vectors whose components are elements of $\mathcal{G}$ (coded somehow), that is, when `from[k]` is G, then `to[k]` is the H that contains $q(G)$. And every G such that $q(G) \in J$ occurs in `from`.

Vectors `from` and `to` could have length zero. This would happen if $q(G) \notin J$ for all $G \in \mathcal{G}$. This is only the most extreme case where `from` does not contain all elements of $\mathcal{G}$. When this happens, and we want the result to be a total order on $\mathcal{G}$, we can provide $\mathcal{G}$ (in any order) as the `domain` argument of R function `tsort`.

When invoked with these arguments, R function `tsort` returns a (not necessarily unique) total order on $\mathcal{G}$ compatible with the given partial order, that is, it agrees with Theorem 1.1. If the user has made a mistake and incorrectly specified the q function so there is no total order that satisfies Theorem 1.1, then `tsort` will give an error.

Current code in R packages `aster` and `aster2` does not actually use the topological sort algorithm but rather forces the user to input the data so

that the numerical order of the components of the response vector is the total order, that is, considering the index set of the response vector to be $\{1, \ldots, n\}$ for some integer n, current code requires $q(G) < j$ for any $j \in G$. It is up to the user to input the data in this way. The computer is no help.

But we could make the computer figure this out in future versions of the software.

1.6.2 Further Factorization

In Lauritzen (1996) chain graph factorizations like our (1.1) and his equation (3.23) can be further factorized, his equation (3.24). But in aster model theory, we shall never be interested in such further factorizations (even in cases where they are possible) and never use any notation that allows for them. So we will never have an analog of equation (3.24) in Lauritzen (1996). For us, factorization is our (1.1).

1.7 Graphs

Each factorization goes with a graph (Lauritzen, 1996, Section 3.2.3). The nodes of the graph are either the elements of N or the components of y corresponding to these elements (y_j for $j \in N$). There is a directed edge, also called an *arrow*, $q(G) \longrightarrow j$ (or if one prefers $y_{q(G)} \longrightarrow y_j$) for every $G \in \mathcal{G}$ and every $j \in G$. There is an undirected edge, also called a *line*, $j \text{ — } k$ (or if one prefers $y_j \text{ — } y_k$) for every $G \in \mathcal{G}$ and every $j, k \in G$ such that $j \neq k$.

As we have just seen, the function q determines the graph (the function q knows its domain $\mathcal{G}$ and codomain N). Conversely, the graph determines the function q.

- The set $J = \bigcup \mathcal{G}$ is the set of nodes of the graph that have incoming arrows (as we shall see, these nodes are called non-initial).

- The elements of $\mathcal{G}$ are the maximal connected components of the graph of lines having node set J. (The graph of lines is the graph obtained by keeping all the nodes and lines but removing all the arrows). If a node $j \in J$ has no incoming lines, then the singleton set $\{j\}$ is an element of $\mathcal{G}$. R function `weak` in R package `pooh` (Geyer, 2017b) can be used to find the dependence groups implied by a graph of lines.

- The graph of q is determined by the arrows: $(G, q(G))$ is an argument-value pair whenever there is an arrow $j \longrightarrow k$ with $j = q(G)$ and $k \in G$.

(If there is an arrow $j \longrightarrow k$ for any $k \in G$, then there must also be an arrow $j \longrightarrow k$ for every $k \in G$.)

Thus we can reason with with graphs or with q functions (which we will soon learn to call *predecessor functions*, Section 1.10 below). Graphs can be helpful, but we do not have to use them.

1.7.1 Exception

In theory, as stated above, there is a line between every pair of distinct elements of every dependence group and no other lines.

In practice, this leads to annoying and unnecessary clutter when we display graphs as figures. Because we never further factorize dependence groups (Section 1.6.2 above), we can find the dependence groups from the graph if we only include enough lines so that each dependence group is a connected subgraph of the graph of lines (again, this is the graph obtained by keeping all the nodes and lines but removing all the arrows).

This exception is illustrated in graph (1.16) below where only two lines rather than three are used to connect the nodes of each dependence group of size three.

1.8 Graphical Terminology

In aster theory, we say

- a node is *initial* if it has no incoming arrows or lines (when thinking about the graph) or if it is not an element of $J = \bigcup \mathcal{G}$ (when thinking about the function q),
- a node is *terminal* if it has no outgoing arrows (it may have outgoing lines and will have outgoing lines if it is an element of an element of $\mathcal{G}$ that is not a singleton set) (when thinking about the graph) or if it is not an element of $\{\, q(G) : G \in \mathcal{G} \,\}$ (when thinking about the function q),
- if there is an arrow $j \longrightarrow k$, then we say that j is the *predecessor* of k (or y_j is the predecessor of y_k),
- and, conversely, that k is a *successor* of j (or y_k is the successor of y_j).

In mainstream graphical model theory, a different terminology is more widely used (Lauritzen, 1996) root = initial, leaf = terminal, parent = predecessor, child = successor. We do not use this terminology in aster model theory because it can cause serious confusion in biological applications.

As a general policy, we eschew all terminology based on biological analogies when there is an available alternative (even when that alternative is less popular).

In any aster graph every node has at most one predecessor and all nodes in the same $G \in \mathcal{G}$ must have the same predecessor (because q is a function that takes elements of $\mathcal{G}$ as arguments).

In mainstream graph theory, a chain graph with only arrows (no lines) having the at-most-one-predecessor property is called a *forest* and its maximal connected components are called *trees*, but we do not use this terminology either (avoiding serious confusion when the application involves data on real trees in real forests). It is enough to say that aster graphs have the at-most-one-predecessor property.

In mainstream graph theory, there is a term *ancestor* that means predecessor, or predecessor of predecessor, or predecessor of predecessor of predecessor, or predecessor of predecessor of predecessor of predecessor, or the same with arbitrarily many repetitions of "predecessor of." And there is a converse term *descendant*, that is, i is an ancestor of j if and only if j is a descendant of i.

In aster model theory we avoid these terms too (avoiding confusion when the application involves real biological organisms with real biological ancestors and real biological descendants). If we need the concepts, then we use the long-winded descriptions predecessor of predecessor of predecessor and so forth or successor of successor of successor and so forth. Fortunately, we rarely need these concepts. And when we do need these concepts we can avoid the cumbersome verbiage by using mathematical notation introduced in Section 1.11 below.

Finally, we need a term for $\mathcal{G}$ and its elements. The terminology we have been using in our writings about aster models is elements of $\mathcal{G}$ are *dependence groups*. The mainstream graphical models terminology (Lauritzen, 1996) is *chain components*. Both have two words and four syllables. Neither is very elegant. We don't like the "chain" terminology because we are not using general chain graph theory (aster models are very special chain graphs). Our term *dependence group* is not great, but we haven't thought of a better term.

1.8.1 Exception

R package `aster` uses "root" node for initial node. We hadn't completely thought through the terminology when that package was written, and we have kept this inconsistency for reasons of backward compatibility.

R package `aster2` does the Right Thing: uses "initial" node.

1.9 Two Kinds of Aster Graphs

The graphs for aster models are often very large with thousands or tens of thousands of nodes, but usually they are composed of isomorphic subgraphs. So drawing one of these isomorphic subgraphs is enough. If you've seen one, you've seen them all. (Graphs are isomorphic if a drawing of one can be laid on a drawing of the other with everything — nodes, lines, and arrows — matching up. Only the names of the nodes are different. And names of arrows, if any.)

An aster graph need not be composed of all isomorphic subgraphs, but the only published example of that is, as far as I know, Eck et al. (2015).

To distinguish these two kinds of graphs, we call the aster graph described in the preceding section the *full aster graph* (we consider the "full" redundant but the emphasis may help avoid confusion).

Certain subgraphs of the full aster graph, we then call graphs for "individuals" (in scare quotes for reasons to be explained presently). These are easier to recognize than describe.

Current aster software (Section 1.2 above) forces $q(G) \neq q(H)$ whenever $G \neq H$ and $q(G)$ and $q(H)$ are initial nodes. In this case, the graph for an "individual" (in scare quotes) consists of the subgraph consisting of one initial node and all of its successors or successors of successors or successors of successors of successors and so forth with arbitrarily many repetitions of "successors of" and all of the arrows and lines in the full graph connecting these nodes. (This is where the term "descendant" in its graph-theoretic sense would come in handy if we allowed ourselves to use it. The graph for an "individual" consists of one initial node, all of its descendant nodes, and all of the lines and arrows going between these nodes. But once we have the idea of the graph for an "individual" we no longer need the term "descendant.")

But aster theory as described so far does not force this convention. If $y_j = 1$, for all initial nodes j, which is the case with most (but not all) aster applications, then it would do no harm if all initial nodes were fused into one

initial node. That would invalidate nothing but the way we just described graphs for "individuals" (in scare quotes).

Thus we have to be a bit more careful. If G is a dependence group whose predecessor $q(G)$ is initial, then the graph for the "individual" (in scare quotes) containing G consists of $q(G)$, the nodes in G and their successors or successors of successors or successors of successors of successors and so forth with arbitrarily many repetitions of "successors of" and all of the arrows and lines in the full graph connecting these nodes. (And it would make this definition a little shorter if we allowed ourselves to use the word "descendant" in its graph-theoretic sense.)

There are two reasons why the scare quotes.

- In life history analysis, the graph for an "individual" ideally goes one or more times around the life cycle (exactly). Thus it may involve data not only for one biological individual but also for its offspring and perhaps offspring of offspring (if the experiment goes twice around the life cycle) or even perhaps more remote descendants (where here "descendants" means real biological descendants, not the graphical models idea of descendants).

- If the value of the constant y_j at the initial node of the graph for an "individual" is greater than one, then the data for this "individual" is actually cumulative data for y_j real biological individuals and perhaps their real biological descendants.

If one does not like our terminology of "'individual' in scare quotes," our advice is to just explain what data the graph is for. It may actually be for a biological individual, for a biological individual and its offspring, or n biological individuals and perhaps their offspring. Just say what it is.

Or we could use the characterization of Corollary B.3 in Appendix B, which says the subgraphs for "individuals" are stochastically independent subvectors of the response vector.

In general, the subgraphs for "individuals" are the minimal stochastically independent subvectors, in the sense that the data for an "individual" has no stochastically independent parts. But when limiting conditional models come into play, this is no longer the case. Thus independence of data for "individuals" (in scare quotes) is an important property of aster models, but it does not (in general) characterize subgraphs for "individuals."

1.10 The Other Predecessor Function

It is useful to have not only the set-to-index predecessor function q defined in Section 1.6 above but also the index-to-index predecessor function p defined as follows

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

Clearly, q determines p. The converse is not true because p knows nothing about dependence groups. But p and $\mathcal{G}$ together determine q.

1.11 Transitive Closure of Predecessor Relation

The *predecessor relation* on N is the index-to-index predecessor function p thought of as a relation, that is, thinking set-theoretically (Halmos, 1960, Section 7), as the set

$$\{ (j, p(j)) : j \in J \} \tag{1.2}$$

of pairs that satisfy the relation.

We need a notation from dynamical systems theory for repeated application of a function. If f is any function whose domain and codomain are the same, then it makes sense to compose f with itself. Then we let f^0 denote the identity function on the domain of f, let $f^1 = f$, $f^2 = f \circ f$, and, in general, $f^{n+1} = f^n \circ f$. So

$$\begin{aligned} f^0(x) &= x \\ f^1(x) &= f(x) \\ f^2(x) &= f(f(x)) \\ f^3(x) &= f(f(f(x))) \end{aligned}$$

and so forth.

We want to use this notation with the index-to-index predecessor function p, but it does not have the same domain and codomain. It is a function $J \to N$ and $J \subset N$, so its codomain is larger than its domain. (We could also consider p to be a partial function on N, but this would not help, so we don't bother.) This means

- $p^0(j) = j$ makes sense for all $j \in N$,
- $p^1(j) = p(j)$ makes sense for all $j \in J$, and

- $p^k(j) = p\big(p^{k-1}(j)\big)$ makes sense whenever $p^{k-1}(j) \in J$ but not otherwise; that is, it makes no sense when $p^{k-1}(j)$ is an initial node.

So $p^k(j)$ always makes sense for $k = 0$, and if it makes sense for some $k > 0$, then it also makes sense for all m such that $0 \le m \le k$. but it will never make sense for all nonnegative integers k, because for all j there exists a k such that $f^k(j)$ is an initial node (because all of our graphs have a finite number of nodes).

As (1.2) says, we are thinking of the predecessor relation as a relation on the set J. So in that context $p^k(j)$ only makes sense when $p^m(j) \in J$ for $m = 0, 1, \ldots, k - 1$.

The transitive closure of the predecessor relation is the smallest transitive relation R containing it. As with most relations, we prefer denoting this relation by infix notation: saying $j \succ k$ rather than $(j, k) \in R$, that is, $j \succ k$ means $k = p^n(j)$ for some positive integer n.

Theorem 1.2. *Under the conditions of Theorem 1.1, the transitive closure of the predecessor relation is a strict partial order.*

Proof. If $j \succ k$, then $j \in G$ for some $G \in \mathcal{G}$ and $k = p^n(q(G))$ for some natural number n ($n = 0$ is allowed).

If $k \in H$ for some $H \in \mathcal{G}$, then we have $G < H$ in the total ordering that Theorem 1.1 uses. Hence we cannot also have $k \succ j$ because that would imply $G < H$ and $H < G$ contradicting $<$ being a strict total order.

If $k \notin H$ for any $H \in \mathcal{G}$ then k has no predecessor (k is initial) and we cannot have $m \succ k$ for any node m.

In either of the preceding cases we never have $k \succ j$ and $j \succ k$. Since $\succ$ is a transitive relation by definition, it is a strict partial order (Halmos, 1960, Section 14). □

Corollary 1.3. *The transitive closure of the predecessor relation is compatible with the total order on the family of dependence groups defined in Theorem 1.1 in the sense that $j \in G \in \mathcal{G}$ and $k \in H \in \mathcal{G}$ and $j \succ k$ implies $G < H$.*

The non-strict counterpart of this relation is the reflexive transitive closure of the predecessor relation, which is denoted $\succeq$. We have $j \succeq k$ if and only if $j \succ k$ or $j = k$.

The inverse of a relation R considered as a set of argument-value pairs reverses the order in the pairs, that is $(k, j) \in R^{-1}$ if and only if $(j, k) \in R$. As usual, we denote the inverse of a relation by turning its infix notation around: $\prec$ is the inverse of $\succ$ and $\preceq$ is the inverse of $\succeq$.

The inverse of the predecessor relation is the successor relation, so $\prec$ is the transitive closure of the successor relation and $\preceq$ is the reflexive transitive closure of the successor relation.

The choice of whether the transitive closure of the predecessor relation is denoted $\succ$ or $\prec$ is arbitrary. Either choice works so long as one keeps straight which is which. Our choice is influenced by an arbitrary choice in the source code for R package `aster`. When the predecessor function is encoded (as the argument `pred` to the R function `aster`) it is required that predecessors have lower indices than successors (come before them in the `pred` vector). Thus we want to think of predecessors as "less than" successors in some sense. Hence our decision to make $p(j) \prec j$.

In graphical model theory, $\succ$ is called the ancestor relation, $\prec$ the descendant relation, $\succeq$ the ancestor-or-self relation, and $\preceq$ the descendant-or-self relation. But, as stated in Section 1.8 above, our policy is to avoid these terms to avoid confusion in biological applications. If we need words rather than symbols, we have to use the long winded ones: "reflexive transitive closure of the predecessor relation" and so forth.

1.12 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. There is a long history of models that have this property. Life table analysis and discrete time survival analysis have it. So does Leslie matrix analysis (Caswell, 2001) and other methods of estimation of population growth rate (Fisher, 1930; Goodman, 1968; Lenski and Service, 1982). But none of those models were regression models, nor did they have the generality of aster models in their graphical structure. They do have the basic relationship of conditional and unconditional means implied by this property (Section 1.13.3 below) but nothing else of aster model theory.

For one conditional distribution in the factorization (1.1), say for the conditional distribution of y_G given $y_{q(G)}$,

- conditional on $y_{q(G)} = 0$, the distribution of y_G is concentrated at zero (the zero vector having all components equal to zero),
- conditional on $y_{q(G)} = 1$, the distribution of y_G is whatever this distribution is designated to be, and
- conditional on $y_{q(G)} = n$ with the $n > 1$, the distribution of y_G is the n-fold convolution of the distribution for sample size one.

In short, the conditional distribution of y_G given $y_q(G)$ is the distribution of the sum of $y_{q(G)}$ independent and identically distributed (IID) random vectors having whatever the distribution is for sample size one. (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 this conditional distribution. Or, shorter still, *predecessor is sample size*.

Note that we name families for dependence groups by the conditional distribution for sample size one. This is an unusual practice. It is not the way families are named for generalized linear models. And it can seem unnecessarily mysterious at first sight.

All of our example graphs in Section 1.14 below have Bernoulli arrows. For such 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 (although R functions for the binomial distribution do understand sample size zero to mean the same thing we do). 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.

One 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. There is an exception to this requirement that will be discussed in Section 1.17 below, but that exception has never been used.

Another consequence of the predecessor-is-sample-size property is the following section.

1.13 Conditional and Unconditional Mean Values

1.13.1 Unconditional

Let y be the response vector of an aster model. We define a parameter vector $\mu = E(Y)$. This is the vector having components

$$\mu_j = E(Y_j), \qquad j \in J, \tag{1.3}$$

where, as usual, J is the index set of the response vector (the set of non-initial nodes of the full aster graph).

This is called the *unconditional mean value parameter vector*. This name is getting us a little bit ahead of ourselves. At this point, we don't even know these means exist. (We will eventually find out they do exist.) And, at this point, we don't know that means parameterize aster models, since we haven't yet even completely specified what the distribution of an aster model is. We know the fundamental factorization (1.1), and we know each of those factors obeys the predecessor-is-sample-size property, but we don't yet know anything more. (We will eventually find out means do parameterize aster models.)

For now we will just assume these means exist.

1.13.2 Conditional

We define another parameter vector ξ having components

$$\xi_j = E(Y_j \mid Y_{p(j)} = 1), \qquad j \in J, \tag{1.4}$$

if this expression makes sense. It will not make sense when the conditioning event has probability zero (so the conditional expectation can be defined arbitrarily). In that case we have to use a different definition that does not come with an equation. The predecessor-is-sample-size property says that y_j is the sum of $y_{p(j)}$ IID random variables, and we say ξ_j is the mean of those random variables.

This is called the *conditional mean value parameter vector*. As in the preceding section, this name is getting us a little bit ahead of ourselves. At this point, we don't even know these means exist. (We will eventually find out they do exist.) And, at this point, we don't know that means (conditional or unconditional) parameterize aster models. But the next section will show the unconditional means determine conditional means and vice versa. So if μ parameterizes, then so does ξ, and vice versa.

1.13.3 The Combination of the Two

It follows from the predecessor-is-sample-size property and linearity of expectation that

$$E(Y_j \mid y_{p(j)}) = \xi_j y_{p(j)}, \qquad j \in J. \tag{1.5}$$

Then it follows from the iterated expectation axiom of conditional probability

$$E(Y_j) = E\{E(Y_j \mid Y_{p(j)})\} \tag{1.6}$$

that

$$\mu_j = \xi_j \mu_{p(j)}, \qquad j \in J. \tag{1.7}$$

This is the fundamental recursive relation that shows (as we examine in more detail presently) how μ is determined by ξ and vice versa.

To map from ξ to μ we use (1.7) 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 (1.7) to determine μ_j as a function of ξ (having already determined $\mu_{p(j)}$ when its node was visited previously). To get the recursion started, we need the mean values at initial nodes, which are given by

$$\mu_j = y_j, \qquad j \in N \setminus J, \tag{1.8}$$

because the mean value of a constant random variable is its constant value.

Using the reflexive transitive closure of the successor relation $\preceq$ we can rewrite the above as follows

$$\mu_j = \left(\prod_{\substack{i \in J \\ i \preceq j}} \xi_i \right) \left(\prod_{\substack{i \in N \setminus J \\ i \preceq j}} \mu_i \right), \qquad j \in J, \tag{1.9}$$

where we note that the second product always has exactly one term: there is always exactly one initial node i such that $i \preceq j$. We could also rewrite (1.9) as

$$\mu_j = \left(\prod_{\substack{i \in J \\ i \preceq j}} \xi_i \right) \left(\prod_{\substack{i \in N \setminus J \\ i \preceq j}} y_i \right), \qquad j \in J, \tag{1.10}$$

by (1.8).

To map from μ to ξ, rewrite (1.7) as

$$\xi_j = \frac{\mu_j}{\mu_{p(j)}} \tag{1.11}$$

but for this to make sense, we must know that $\mu_{p(j)}$ is never zero.

We will eventually find out that $\mu_{p(j)}$ is never zero except in limiting conditional models. So we do not always have this property. Thus (1.11) makes sense when $\mu_{p(j)}$ is never zero, but otherwise some components of ξ are not determined by μ.

Conversely, multiplication by zero is not a problem (unlike division by zero), so (1.9) and (1.10) always determine μ as a function of ξ.

Everything in this section up to this point is an elementary consequence of the laws of conditional and unconditional expectation and the predecessor-is-sample-size property. Consequently, everything in this section up to this point is also true of all previous models in survival analysis and demography that have also had this property cited in Sections 1.1 and 1.12 above.

1.13.4 Confession

Geyer et al. (2007) did not define ξ the way we do here. Instead they used that Greek letter to denote (1.5). A referee said this definition is dumb. It makes ξ a function of both random variables and parameters, so it is not a parameter, and one shouldn't use Greek letters for things that aren't parameters. We didn't listen then and managed to get the paper published overriding this objection. But now we agree with the referee.

The vector ξ as defined here is an important parameterization of aster models (this has been realized since Geyer, 2010).

R package `aster` used the same dumb definition until version 1.0-2 of the package, when a new optional argument `is.always.parameter` was added to the method of R generic function `predict` that handles aster model objects. And, for reasons of backward compatibility, the dumb definition is still the default. One must use the optional argument `is.always.parameter = TRUE` to estimate ξ as defined in this section.

R package `aster2` and recent papers and technical reports use the definition presented here (the conditional mean value parameter vector is defined as we do here if they mention conditional mean value parameters at all).

1.14 Some Aster Graphs

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

$$\begin{array}{ccccccc} 1 & \xrightarrow{\text{Ber}} & y_1 & \xrightarrow{\text{Ber}} & y_2 & \xrightarrow{\text{Ber}} & y_3 \\ & & \downarrow \text{Ber} & & \downarrow \text{Ber} & & \downarrow \text{Ber} \\ & & y_4 & & y_5 & & y_6 \\ & & \downarrow \text{0-Poi} & & \downarrow \text{0-Poi} & & \downarrow \text{0-Poi} \\ & & y_7 & & y_8 & & y_9 \end{array} \tag{1.12}$$

which is for one individual. There are 570 individuals in the data set, which is included in the R package `aster`. So one can think of the full aster graph as 570 copies of this graph with the subscripts changed so the nodes (the y_j) are all different.

Because this graph has only arrows, no lines, each node is a dependence group all by itself.

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 co-PI's are the second and third authors of Geyer et al. (2007)). The way (1.12) 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).

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.

Aster graphs for "individuals" can get a lot bigger than (1.12). 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").

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

$$y_i \longrightarrow y_j \tag{1.13}$$

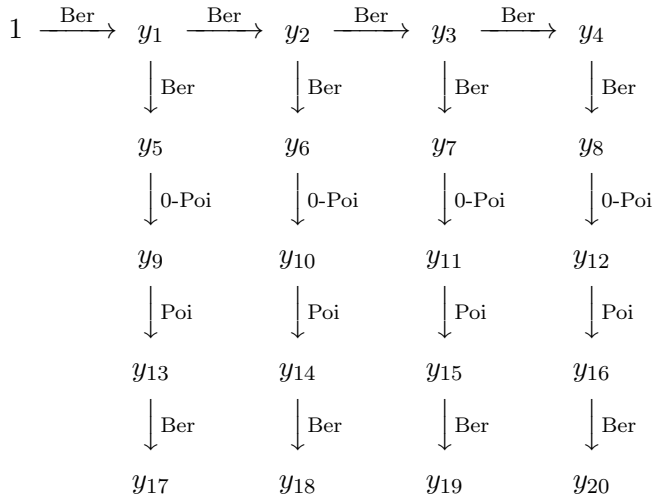
indicates the conditional distribution of y_j given y_i. An arrow

$$1 \longrightarrow y_i \tag{1.14}$$

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 1.12.

Here is a more complicated aster graph from Shaw and Geyer (2010)



(1.15)

Again, because this graph has only arrows, no lines, each node is a dependence group all by itself. The label Poi on arrows indicates the Poisson distribution.

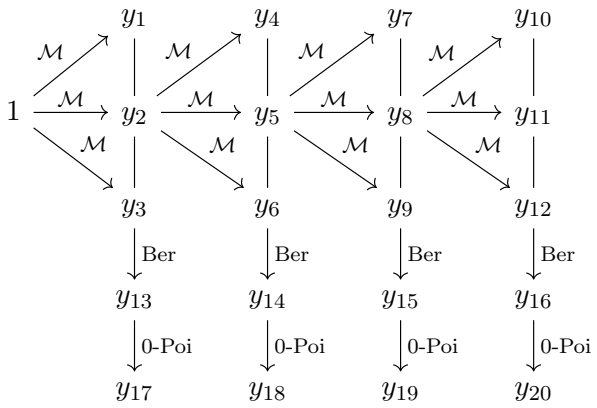
This graph is for simulated data, which Shaw and Geyer (2010) used because at the time no data for aster models as complicated as (1.15) had been collected by biologists, and it was important to give such an illustration of the possibilities of aster models. Like in (1.12) the "columns" in (1.15) are for data in successive years. The first three "rows" of (1.15) can be taken to be the same as those of (1.12): survival, flowering indicator variables, and flower counts. The fourth row of (1.15) is seed counts, and the fifth row is number of seeds that germinate (produce new plants). Of course, since the data are simulated, the story about these variables is just a story. It could be told differently, and Shaw and Geyer (2010) do have a story where the same graph could be for data about an animal rather than a plant.

This graph did serve as a good example of what was possible. Stanton-Geddes, Tiffin, and Shaw (2012b, in the on-line appendix) discuss an aster model with seven life history stages (one of which is artificial, modeling random sampling in the data collection process — for more on this see

Chapter 3 below — hence only six are about life history of the organisms) and thus would be like the graph (1.15) except with seven "rows." Because the organism in question (*Chamaecrista fasciculata*, common name partridge pea) was an annual plant, there is only one "column." (As mentioned above, "rows" and "columns" are not part of the graphical structure — the only thing that matters is which nodes are connected by which arrows and lines (if any) — and Stanton-Geddes et al. actually laid out their graph in a row.)

Statistically, there are some important differences between these graphs. Graph (1.15) has both Poisson (Poi) and zero-truncated Poisson (0-Poi) arrows and hence illustrates when to use which (see also Section 1.14.1 below). In graph (1.12) every predecessor node is Bernoulli, but in graph (1.15) y_{13} through y_{16} are non-Bernoulli predecessor nodes. So (1.15) shows that predecessor values can be any nonnegative integer.

The graph (1.16) 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 (a moth) that does have these life history stages. Those 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 clean an example as the graph (1.16).



(1.16)

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 nodes 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 $j \geq 13$.

As stated in Section 1.7.1 above, this graph does not follow the theory, which requires a line connecting each pair of distinct elements of each dependence group and which would require us to add lines y_1 — y_3 and y_4 — y_6 and y_7 — y_9 and y_{10} — y_{12} to the graph. But, also as stated in Section 1.7.1 above, we don't need these lines to infer the dependence groups from the graph. The lines y_1 — y_2 and y_2 — y_3 that are in the graph say y_1 and y_3 are in the same dependence group as y_2, and since there are no other lines to these nodes, they must constitute a dependence group. Since there doesn't seem to be any room in the picture (1.16) for these additional lines required by theory, we omit them.

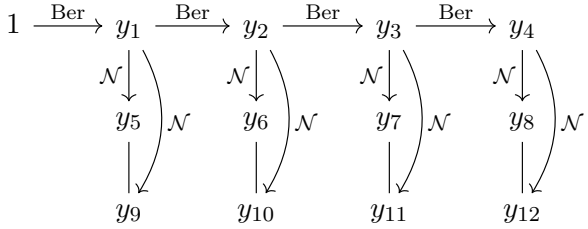
Each of the multi-node dependence groups has a conditional multinomial distribution with, as usual, predecessor as sample size. Since each predecessor is zero-or-one-valued, if a predecessor (say y_2) is equal to one, then exactly one of its three successor nodes (y_4, y_5, and y_6) is equal to one, and, 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_3, y_4, y_6, y_7, y_9, y_{10}, y_{11}, and y_{12}) 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 (as in Eck et al., 2015) is life history data for an insect. As in our graphs without dependence groups, "columns" of the graph are for different times (days, perhaps). 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 before. It is not necessary for every individual to have the same graph (here females and males have different graphs). And we have non-singleton dependence

groups, multinomial "switches" between different life history stages.

Here is yet another graph illustrating normal dependence groups.



(1.17)

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. Why the normal distribution has two nodes in the graph is completely explained in Section D.10 in Appendix D. If we had a one-parameter subfamily of the normal family, then that would have only one node; see Section D.5 for that.

An aster model with two-parameter normal dependence groups, like graph (1.17), is seen in May et al. (2022). An aster model with one-parameter normal arrows (normal location family) is seen in Warwell and Shaw (2017).

1.14.1 Zero-Inflated Poisson

Readers may have wondered why the graph (1.12) has its middle "row." 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 $y_{j+3} > 0$, $j = 4, 5, 6$). So why were these variables inserted in the graph?

Consider just the subgraph

$$y_1 \xrightarrow{\text{Ber}} y_4 \xrightarrow{\text{0-Poi}} y_7 \tag{1.18}$$

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 this bit of terminology.

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

Why can we not just have zero-inflated Poisson arrows? Because zero-inflated Poisson is not an exponential family. It can be factored into a product of two exponential families, like (1.18) does. But it is not itself exponential family. Hence we cannot have arrows like that in an aster model because of the exponential family assumption (Section 1.16 below). If we want zero-inflated Poisson or zero-inflated negative binomial in aster models, then this is the way we have to do it.

Although we have zero-inflated Poisson distributions in aster models, we do not give them their usual parameterization (Lambert, 1992). No aster model parameterization (summarized in Section 1.21 below) has this usual parameterization.

1.14.2 On Not Overusing Zero-Truncated Poisson

The zero-truncated Poisson distribution is for when *by definition of the model* a count of zero cannot occur unless the predecessor is zero. It is not for the case where expected count is so high that zeros occur infrequently. It is for the case where zeros are impossible (except when predecessors are zero).

Some users have been confused about this, using zero-truncated distributions where they are not appropriate.

1.14.3 Not Really Missing Data

The property that predecessor zero implies successor zero (because a sum having zero terms is zero by convention, Section 1.12 above) takes care of what people formerly conceived of as a missing data problem: when $y_{p(j)} = 0$ (for concreteness, say this means the "individual" is dead), you cannot "observe" y_j.

Nevertheless we can infer $y_j = 0$ (if y_j is flower count, then we are inferring that dead plants have no flowers; if y_j is survival for the following year, then we are inferring that dead plants stay dead).

So that is not a true missing data problem from the aster model perspective. Researchers do need to be aware of the need to code their data this way (dead individuals have 0 flowers not `NA` flowers, `NA` being the R value for missing data).

Many researchers have been confused about this when first introduced to aster models. If different individuals live different numbers of years, don't we need different graphs for individuals to reflect this? No. We just have a long enough graph to accommodate all life spans. After the individual is dead the data are just zero (not `NA`).

As we shall see (Section 1.16 below) the property that predecessor zero implies successor zero makes likelihood inference automatically do the Right Thing. We do not have to go into contortions to get it to do the Right Thing. It just does the Right Thing automatically.

1.14.4 Really Missing Data

If we did have truly missing data (not observable and not inferable), then we would have a problem that aster models are not equipped to solve.

R function `aster` in R package `aster` does not allow `NA` values in the data it analyzes. Its help page says

> It was almost always wrong for aster model data to have `NA` values. Although theoretically possible for the R formula mini-language to do the right thing for an aster model with `NA` values in the data, usually it does some wrong thing. Thus, since version 0.8-20, this function and the `reaster` function give errors when used with data having `NA` values. Users must remove all `NA` values (or replace them with what they should be, perhaps zero values) "by hand".

R function `asterdata` in R package `aster2` also does not allow `NA` values in aster data for analyses done by that package.

1.14.5 Bernoulli versus Multinomial

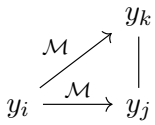
Bernoulli arrows and multinomial dependence groups are closely related but work differently. Bernoulli is related to multinomial but different.

In (1.16) every arrow labeled with an $\mathcal{M}$ is Bernoulli marginally, but the whole point of dependence groups is that the components of the response vector in a dependence group are not conditionally independent given their predecessors, unlike the Bernoulli arrows labeled Ber in (1.16) or in any of the graphs in this section (by Lemma B.1).

Conversely, if

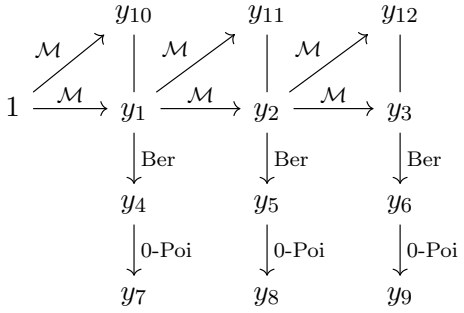
$$y_i \xrightarrow{\text{Ber}} y_j$$

is a Bernoulli arrow, then we could replace this with a multinomial dependence group that does the same thing



where k is some index that hasn't been used in the rest of the graph.

For example, we could change graph (1.12) to



(1.19)

and all of the components of the response vector that have the same indices would have the same interpretation and the same values in the same data in both graphs (1.12) and (1.19). Then the additional nodes y_{10}, y_{11}, and y_{12} are determined by the properties of the multinomial distribution and the predecessor-is-sample size property (Section 1.12 above)

$$\begin{aligned} y_{10} &= 1 - y_1 \\ y_{11} &= y_1 - y_2 \\ y_{12} &= y_2 - y_3 \end{aligned} \tag{1.20}$$

The aster models with graphs (1.12) and (1.19) can be made equivalent if parameterized to make that so (I think, no proof here), but they don't have to be equivalent (I think). Hence scientists have to decide which to use. The argument of simplicity argues for (1.12) (no dependence groups). But arguments can be made both ways.

Not only can we replace Bernoulli arrows with multinomial dependence groups, we can, at least partially replace multinomial dependence groups with Bernoulli arrows. We could replace the graph (1.16) with the graph shown in Figure 1.1 (p. 28). In graph (1.16) and the graph in Figure 1.1 all of the components of the response vector that have the same indices have the same interpretation and the same values in the same data *except* for y_1,

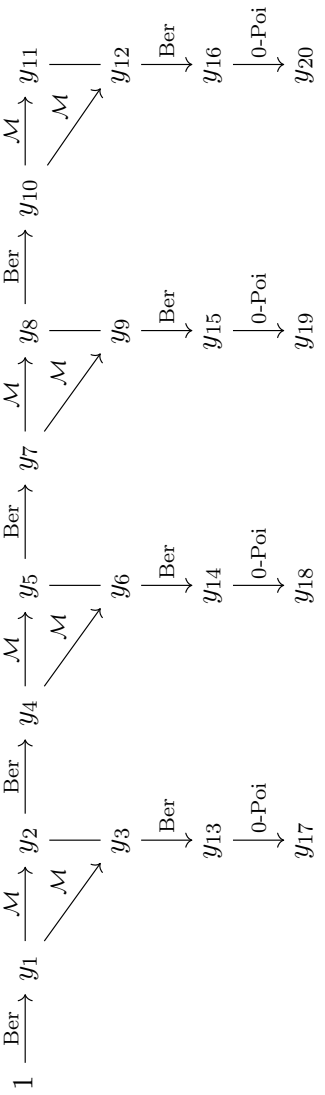


Figure 1.1: Alternative to Graph (1.16).

y_4, y_7, and y_{10} which have the *opposite* interpretation and opposite values in the two graphs. In graph (1.16) the value one for y_1, y_4, y_7, and y_{10} means *dead*. In the graph in Figure 1.1 the value one for these variables means *alive*.

The aster models with the graph (1.16) and the graph in Figure 1.1 cannot (I think) be made equivalent by some choice of parameterization. Hence scientists have to decide which to use. Now the argument of simplicity seems to argue for (1.16). In either case we have dependence groups (we must have to track life history stages). So as long as we have to have dependence groups, we might as well make them do as much work as possible, tracking mortality as well as progress through life history stages. But arguments can be made both ways.

1.14.6 Bernoulli versus Bernoulli

Any composition of Bernoulli arrows is another Bernoulli. In the following graph

$$y_1 \xrightarrow{\text{Ber}} y_2 \xrightarrow{\text{Ber}} y_3 \xrightarrow{\text{Ber}} y_4 \tag{1.21}$$

the conditional distribution of y_4 given $y_1 = 1$ is Bernoulli. If $y_1 = 1$, then y_2 is either zero or one, hence so is y_3, hence so is y_4, and any zero-or-one-valued random variable is Bernoulli.

Hence, we can replace the subgraph (1.21) by the single arrow

$$y_1 \xrightarrow{\text{Ber}} y_4 \tag{1.22}$$

(deleting the components y_2 and y_3 from the response vector).

And this change-of-data gives a model in which the conditional distribution of y_4 given $y_1 = 1$ is Bernoulli in either case.

But when these are subgraphs of a larger graph, the corresponding aster models cannot (I think) be made equivalent by some choice of parameterization. Hence scientists have to decide which to use.

Of course, had (1.21) been different

$$\begin{array}{ccccccc} y_1 & \xrightarrow{\text{Ber}} & y_2 & \xrightarrow{\text{Ber}} & y_3 & \xrightarrow{\text{Ber}} & y_4 \\ & & \downarrow & & & & \\ & & \vdots & & & & \end{array}$$

so y_2 had a successor other than y_3 (or similarly for y_3) then y_2 and y_3 could not have been removed from the data.

The observation of this section is only for when we have a straight line of Bernoullis like (1.21) with no branches. But this case is seen surprisingly often in published aster analyses. Again, scientists have to decide which to use.

1.14.7 Thinned Poisson

A Poisson followed by a Bernoulli is Poisson (this is well known in spatial statistics). In the graph

$$y_1 \xrightarrow{\text{Poi}} y_2 \xrightarrow{\text{Ber}} y_3 \tag{1.23}$$

the conditional distribution of y_3 given $y_1 = 1$ is Poisson. Thus this graph can be replaced by the single arrow

$$y_1 \xrightarrow{\text{Poi}} y_3 \tag{1.24}$$

(deleting the component y_2 from the response vector),

And this change-of-data gives a model in which the conditional distribution of y_4 given $y_1 = 1$ is Poisson in either case. Again, scientists have to decide which to use.

There are, as far as I can see no general principles for which aster graph is the one and only Right Thing for any particular data. Multiple different aster models may have some rationale supporting them.

1.15 Exponential Families of Distributions

This is a brief overview of the theory of exponential families of distributions, just enough to allow us to finish the basic theory of aster models. We will do some more exponential family theory later as the need arises.

1.15.1 Definition

The usual definition of exponential families of distributions (Barndorff-Nielsen, 1978, Chapter 8; Brown, 1986, Chapter 1; Geyer, 1990, Chapter 1) involves probability mass-density functions (or measure-theoretic probability density functions with respect to an arbitrary positive measure). Here we give a simpler definition from Geyer (2009).

A statistical model is a family of probability distributions. A statistical model is an *exponential family of distributions* if it has a log likelihood of the form

$$l(\theta) = \langle y, \theta \rangle - c(\theta) \tag{1.25}$$

where

- y is a vector-valued statistic, which is called the *canonical statistic*,
- θ is a vector-valued parameter, which is called the *canonical parameter*,
- c is a real-valued function, which is called the *cumulant function*, and
- $\langle \cdot, \cdot \rangle$ is a bilinear form that places the vector space where y takes values and the vector space where θ takes values in duality.

In equation (1.25) we have used the rule that additive terms in the log likelihood that do not contain the parameter may be dropped. Such terms have been dropped in (1.25).

In aster model theory, we always have

$$\langle y, \theta \rangle = \sum_{j \in J} y_j \theta_j$$

where J is the common index set for y and θ so both are vectors in $\mathbb{R}^J$. But the angle brackets notation, which comes from functional analysis (Rudin, 1991), is used to indicate that we don't think of the the vector space where y takes values and the vector space where θ takes values as being the same. They are dual vector spaces, not the same.

This means that $\langle y, y \rangle$ or $\langle \theta, \theta \rangle$ are glaring errors, unlike what would be the case if we said that $\langle \cdot, \cdot \rangle$ was an inner product on $\mathbb{R}^J$ or some such. In any event, your humble author has been using this notation since his thesis (Geyer, 1990) and is not going to stop now. Moreover, most aster papers also use this notation (if they discuss aster theory at all).

Although we usually say "the" canonical statistic, "the" canonical parameter, and "the" cumulant function, these are not uniquely defined:

- any one-to-one affine function of a canonical statistic vector is another canonical statistic vector,
- any one-to-one affine function of a canonical parameter vector is another canonical parameter vector, and
- any real-valued affine function plus a cumulant function is another cumulant function.

(Affine functions are defined in Section 1.15.3 below.) These possible changes of statistic, parameter, or cumulant function are not algebraically independent. Changes to one may require changes to the others to keep a log likelihood of the form (1.25) above.

Usually no fuss is made about this nonuniqueness. One fixes a choice of canonical statistic, canonical parameter, and cumulant function and leaves it at that.

The cumulant function may not be defined by (1.25) on the whole vector space where θ takes values. In that case it can be extended to this whole vector space by

$$c(\theta) = c(\psi) + \log \left\{ E_\psi \left(e^{\langle Y, \theta - \psi \rangle} \right) \right\} \tag{1.26}$$

where θ varies while ψ is fixed at a possible value of the canonical parameter vector, and the expectation and hence $c(\theta)$ are assigned the value ∞ for θ such that the expectation does not exist (Geyer, 2009, equation (5)).

The family is *full* if its canonical parameter space is

$$\Theta = \{ \, \theta : c(\theta) < \infty \, \} \tag{1.27}$$

and a full family is *regular* if its canonical parameter space is an open subset of the vector space where θ takes values.

Almost all exponential families used in real applications are full and regular. So-called *curved exponential families* (smooth non-affine submodels of full exponential families) are not full. Constrained exponential families (Geyer, 1991) are not full. A few exponential families used in spatial statistics are full but not regular (Geyer and Møller, 1994).

Many people use "natural" everywhere this book uses "canonical." In this we are following Barndorff-Nielsen (1978). It also goes with our policy of avoiding terminology used in biology if alternatives are available.

Many people also use an older terminology that says a statistical model is *in the* exponential family, where we say a statistical model is *an* exponential family. Thus the older terminology says *the* exponential family is the collection of all of what the newer terminology calls exponential families. The older terminology names a useless mathematical object, a heterogeneous collection of statistical models not used in any application. The newer terminology names an important property of statistical models. If a statistical model is a regular full exponential family, then it has all of the properties discussed here. If a statistical model is an exponential family (not necessarily full or regular), then it has many of the properties discussed here. Presumably, that is the reason for the newer terminology. In this we are again following Barndorff-Nielsen (1978).

1.15.2 Independent and Identically Distributed

Suppose we have an exponential family with vector canonical statistic z, vector canonical parameter θ, and log likelihood

$$l(\theta) = \langle z, \theta \rangle - c(\theta)$$

And suppose we have an IID sample from this family with log likelihood

$$\begin{aligned} l_n(\theta) &= \sum_{i=1}^{n} \left[\langle z_i, \theta \rangle - c(\theta) \right] \\ &= \left\langle \sum_{i=1}^{n} z_i, \theta \right\rangle - nc(\theta) \end{aligned} \tag{1.28}$$

where subscripts indicate a sequence of vectors rather than components of one vector: z_1, z_2, ... are IID random vectors. This tells us that IID sampling from an exponential family gives another exponential family with

- canonical statistic vector that is the sum of the canonical statistic vectors for the elements of the sample,
- the same canonical parameter vector, and
- cumulant function that is n times the cumulant function for the elements of the sample.

A lot of "addition rules" for "brand name distributions" are special cases. For example, sum of IID Bernoulli is binomial, sum of IID Poisson is Poisson, sum of IID exponential is gamma.

The simple math in this section is important for aster models. The predecessor-is-sample-size property (Section 1.12 above) requires us to know n-fold convolutions of the distributions associated with arrows at least well enough to write down the log likelihood. This section tells how that is done.

1.15.3 Canonical Affine Submodels

In this section we consider *canonical affine submodels* of exponential families. If θ is the canonical parameter vector, then these submodel parameterizations have the form

$$\theta = a + M\beta, \tag{1.29}$$

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* in the terminology of R functions `lm` and `glm`. M is called the *design matrix* by others. We use the terminology favored by R. The submodel parameter vector β is called the *coefficients* vector by R. We will find another name for it in this section.

The term "canonical affine submodel" was introduced by Geyer et al. (2007), but before that "affine hypothesis" was used by Barndorff-Nielsen (1978, Section 8.2). Usually these are called linear models or generalized linear models or log-linear models in various settings.

There are two notions of linear used in mathematics. There is a sharp dividing line at the beginning of linear algebra. In calculus and lower level mathematics (including pre-college) a linear function is one with a flat graph. In linear algebra and all higher level mathematics a linear function is one that preserves the vector space operations. In particular, if f is a linear function in this higher-level sense, then $f(0) = 0$. If one needs the lower-level sense in higher-level mathematics, it is called an affine function. An affine function between vector spaces is a linear function plus a constant function. If $a \neq 0$ in (1.29), then this is a linear change-of-parameter in the lower-level sense but an affine change-of-parameter in the higher-level sense. It is unclear (to me) whether the "linear" in linear models, generalized linear models, and log-linear models is the lower-level sense allowing $a \neq 0$ in (1.29) or the higher-level sense assuming $a = 0$ in (1.29), because $a = 0$ in most applications. We follow Geyer et al. (2007) and Barndorff-Nielsen (1978) in calling these models affine.

The log likelihood for the canonical affine submodel is

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

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

$$l(\beta) = \langle y, M\beta \rangle - c(a + M\beta)$$

and, because

$$\langle y, M\beta \rangle = y^T(M\beta) = y^T M\beta = (M^T y)^T \beta = \langle M^T y, \beta \rangle$$

we can write the submodel log likelihood in exponential family form

$$l(\beta) = \langle M^T y, \beta \rangle - c_{\text{sub}}(\beta) \tag{1.30}$$

where

$$c_{\text{sub}}(\beta) = c(a + M\beta), \qquad \text{for all } \beta. \tag{1.31}$$

This shows the canonical affine submodel is itself an exponential family with

- canonical statistic vector $M^T y$,
- canonical parameter vector β, and
- cumulant function c_{sub}.

It is helpful to have a term saying what model a canonical affine submodel is a submodel of. In this context, we call that model the *saturated model*. It is the special case of (1.29) with $a = 0$ and M the identity matrix, so $\theta = \beta$. The saturated model is the largest possible canonical affine submodel (of itself).

A canonical affine submodel is a full exponential family if its parameter space is

$$\{\, \beta \in \mathbb{R}^K : c(a + M\beta) < \infty \,\}$$

where K is the index set of β and c is the cumulant function of the saturated model. Because affine functions are continuous and the preimage of an open set through a continuous function is open, the full canonical affine submodel is regular if the saturated model is a regular full exponential family.

1.15.4 Moment and Cumulant Generating Functions

The *moment generating function* (MGF) of a random vector Y is given by

$$M(t) = E\left(e^{\langle Y, t\rangle}\right)$$

provided that the function M so defined is finite on a neighborhood of zero; otherwise we say the random vector Y does not have an MGF. Clearly, the MGF only depends on the distribution of Y, so we also say this is the MGF of this distribution.

The MGF of the distribution of the canonical statistic of an exponential family corresponding to canonical parameter θ is given by

$$M_\theta(t) = E_\theta\left(e^{\langle Y, t\rangle}\right)$$

We rearrange (1.26) obtaining

$$e^{c(\theta) - c(\psi)} = E_\psi\big(e^{\langle Y, \theta - \psi\rangle}\big)$$

and then change θ to $\theta + t$ and ψ to θ in that order obtaining

$$e^{c(\theta+t)-c(\theta)} = E_\theta\big(e^{\langle Y,t\rangle}\big)$$

so the MGF of the distribution of Y for parameter vector θ is

$$M_\theta(t) = e^{c(\theta+t)-c(\theta)}$$

provided M_θ is finite on a neighborhood of zero, which is when c is finite on a neighborhood of θ, which is when θ is in the interior of the full canonical parameter space (1.27). Hence every distribution in a *regular* full exponential family has an MGF.

The *cumulant generating function* (CGF) of a random vector Y is the log of the MGF provided the MGF exists. If a distribution has no MGF, then it has no CGF either. Hence the CGF of the distribution of Y for parameter vector θ is

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

provided K_θ is finite on a neighborhood of zero, which is when θ is in the interior of the full canonical parameter space (1.27). Hence every distribution in a *regular* full exponential family has a CGF.

The MGF is so called because its derivatives evaluated at zero give moments. The CGF is so called because its derivatives evaluated at zero give cumulants. Cumulants are polynomial functions of moments and vice versa. Thus a distribution that has an MGF or a CGF has moments and cumulants of all orders. But we will not use moments and cumulants higher than second order (next section) in our study of aster models.

Observe that derivatives of K_θ evaluated at zero are derivatives of c evaluated at θ. Hence the cumulant function is so called because its derivatives evaluated at θ give cumulants. (Of course, this is only valid when θ is in the interior of the full canonical parameter space (1.27)).

This also shows that cumulant functions are infinitely differentiable at every point in the interior of the full canonical parameter space (1.27)), and cumulant functions of a regular full exponential family are infinitely differentiable at every point of the full canonical parameter space (1.27)).

1.15.5 Mean and Variance

First and second order cumulants are mean and variance

$$E_\theta(Y) = c'(\theta) \tag{1.32a}$$

$$\operatorname{var}_\theta(Y) = c''(\theta) \tag{1.32b}$$

(Barndorff-Nielsen, 1978, Theorem 8.1). Of course, these formulas are valid only where c is differentiable, which is when θ is in the interior of the full canonical parameter space (1.27). On the boundary of the full canonical parameter space, the derivatives on the right-hand sides of these formulas do not exist and the cumulants on the left-hand sides might or might not exist but in any case are not given by these formulas. In a *regular* full exponential family these formulas are good for all values of θ in the full canonical parameter space (1.27).

In (1.32a) Y is a vector and its expectation is a vector having the same index set: $\mu = E_\theta(Y)$ is the vector having components $\mu_j = E_\theta(Y_j)$. The right-hand side of (1.32a) must also be a vector having the same index set: its components are $\partial c(\theta)/\partial\theta_j$.

In (1.32b) Y is a vector and its variance is a square symmetric matrix having components $\operatorname{cov}_\theta(Y_j, Y_k)$. The right-hand side of (1.32b) must also be a matrix having the same index set: its components are $\partial^2 c(\theta)/\partial\theta_j\partial\theta_k$.

Many people do not like our terminology calling the left-hand side of (1.32b) the *variance* matrix of Y. Names in common use are *covariance* matrix (but this is horrible terminology because it uses up a name that should be used for the covariance of two random vectors), *variance-covariance* matrix, and *dispersion* matrix. We use our terminology, because it denotes the multivariate analog of the variance of a random variable. This can be seen everywhere in probability theory. If you can accept our notation $\operatorname{var}_\theta(Y)$ for this mathematical object, then the same formulas work for univariate and multivariate Y.

1.16 Aster Models and Exponential Families

We finally get to the final axiom of aster models. Each conditional distribution in the factorization (1.1) actually stands for parametric family of distributions, and each of these is an exponential family of distributions. As explained in Section 1.12 above, the distribution for these conditional families is determined by the distribution for sample size one, and the distributions for dependence group G for sample size one form an exponential family having

- canonical statistic y_G,
- canonical parameter θ_G, and
- cumulant function c_G.

This notation means that θ is a vector with the same index set as the response vector y and that θ_G are subvectors of θ in the same way as y_G are subvectors of y. The fact that the cumulant functions are subscripted means every dependence group may correspond to a different exponential family, and we saw this in the examples (Section 1.14 above).

Now Section 1.15.2 above tells us the cumulant function for sample size n is n times the cumulant function for sample size one. Note that this is valid for $n = 1$ and $n = 0$ too. For the latter we calculate using (1.26) in the case that y is the random vector concentrated at zero

$$c(\theta) = c(\psi) + \log\{E_\psi(1)\} = c(\psi) + \log(1) = c(\psi)$$

so the cumulant function of the random vector concentrated at zero is a constant function ($c(\psi)$ is arbitrary but ψ is a constant in this equation), and to obtain agreement with the assertion that the cumulant function for sample size n is n times the cumulant function for sample size one, we take this arbitrary constant to be zero, so the cumulant function for sample size zero is always the zero function, which always has the value zero.

Now the predecessor-is-sample-size property says that the sample size for dependence group G is $y_{q(G)}$, so the cumulant function for this sample size is $y_{q(G)}$ times the cumulant function for sample size one.

Because the log of a product is the sum of the logs, the log likelihood for an aster model is the sum of terms, one term for each term in the factorization (1.1), and each term looks like (1.25) except with subscripts for the dependence groups, that is,

$$l(\theta) = \sum_{G \in \mathcal{G}} \left[\langle y_G, \theta_G \rangle - y_{q(G)} c_G(\theta_G) \right] \tag{1.33}$$

(by the predecessor-is-sample-size property, the conditional distribution of y_G given $y_{q(G)}$ is the sum of IID random vectors having cumulant function c_G, hence the cumulant function in each term is $y_{q(G)}$ times the cumulant function for sample size one).

As was mentioned in Section 1.14.3 above, probability theory just does the Right Thing in case $y_{q(G)} = 0$. By the predecessor-is-sample-size property, this implies $y_G = 0$ too, so the whole contribution to the log likelihood of the term for dependence group G is zero (when, as we are discussing here, the predecessor is zero). And zero is the correct contribution because the conditional distribution of y_G given $y_{q(G)} = 0$ is concentrated at zero, that is, it is zero with probability one, so the contribution to the log likelihood should be $\log(1) = 0$.

So there is no need to do anything special about the case where predecessor equals zero. All of our formulas are correct in this case too.

1.17 Infinitely Divisible Aster Families

As stated in Section 1.12 above, one consequence of the predecessor-is-sample-size property is that components of the response vector that are predecessors must be nonnegative-integer-valued random variables.

Except Geyer et al. (2007) note an exception to this rule. If the distribution in question is infinitely divisible, then r-fold convolution makes sense for any $r \geq 0$. A distribution having moment generating function m is infinitely divisible if and only if $m(\,\cdot\,)^r$ is a moment generating function for all real $r \geq 0$ (Cuppens, 1975, Corollary 4.2.2). Then the distribution having moment generating function $m(\,\cdot\,)^r$ is defined to be the r-fold convolution of the distribution having moment generating function m. Hence a regular full exponential family having cumulant function c is infinitely divisible if and only if $rc(\,\cdot\,)$ is a cumulant function for all real $r \geq 0$. R package `aster` (Geyer, 2021) has several infinitely divisible families: Poisson, negative binomial, and normal-location (normal with unknown mean and known variance). Hence, if the conditional distribution of y_G given $y_{q(G)} = 1$ is infinitely divisible, then the distribution of $y_{q(G)}$ can be nonnegative-real-valued.

This observation, however, has played no role in applications of aster models because R packages `aster` and `aster2` do not implement any families of nonnegative-valued random variables that are not also integer-valued. Such families exist (the gamma distribution, for example), so if they were implemented, there could be a role for this observation.

1.18 The Aster Transform

The aster log likelihood (1.33) does not have exponential family form (1.25). But since (1.33) is affine in y, it must be an exponential family having canonical statistic vector y. In order to figure out the canonical parameter and cumulant function for this family we rewrite (1.33) as follows

$$l(\theta) = \left(\sum_{j \in J} y_j \left[\theta_j - \sum_{\substack{G \in \mathcal{G} \\ q(G) = j}} c_G(\theta_G) \right] \right) - \left(\sum_{\substack{G \in \mathcal{G} \\ q(G) \notin J}} y_{q(G)} c_G(\theta_G) \right) \tag{1.34}$$

and now we see we do have exponential family form with the canonical parameters (of the exponential family which is the joint distribution of y) being the terms in square brackets (the multipliers of the components of y)

$$\varphi_j = \theta_j - \sum_{\substack{G \in \mathcal{G} \\ q(G) = j}} c_G(\theta_G), \qquad j \in J, \tag{1.35}$$

and the cumulant function being the terms left over

$$c(\varphi) = \sum_{\substack{G \in \mathcal{G} \\ q(G) \notin J}} y_{q(G)} c_G(\theta_G) \tag{1.36}$$

(the fact that we have φ on one side of the equation and θ on the other will be explained presently). Note that all of the $y_{q(G)}$ appearing in (1.36) are constant random variables (because they are at initial nodes). Thus (1.36) does define a deterministic (non-random) function of the parameter, as the cumulant function of an unconditional distribution must be.

This notation means that φ is a vector with the same index set as the response vector y, that φ_G are subvectors of φ in the same way as y_G are subvectors of y, and that φ_j are components of φ in the same way as y_j are components of y.

The map $\theta \mapsto \varphi$ is called the *aster transform*. We claim this parameter transformation is invertible and both the transform and its inverse are infinitely differentiable. We invert the function very simply moving terms from the right-hand side to the left-hand side

$$\theta_j = \varphi_j + \sum_{\substack{G \in \mathcal{G} \\ q(G) = j}} c_G(\theta_G), \qquad j \in J. \tag{1.37}$$

How can this define θ in terms of φ when components of θ appear on both sides of the equation? Simply use the equations (1.37) in any order that visits successors before predecessors. Theorem 1.1 guarantees there is such an order (it guarantees an order among dependence groups, but we can order components within a dependence group arbitrarily). Then when we are using (1.37) to compute θ_j we will have already have computed θ_k for all k that are successors, successors of successors, and so forth of j, and, in particular, we will already have computed θ_G for all G such that $q(G) = j$.

The map $\varphi \mapsto \theta$ is called the *inverse aster transform*. The aster transform is clearly infinitely differentiable because cumulant functions are infinitely differentiable (when all of the families for the dependence groups

are regular full exponential families). The inverse function theorem of real analysis (Browder, 1996, Theorem 8.27) says the inverse function is differentiable as many times as the function it is the inverse of. Hence the inverse aster transform is also infinitely differentiable.

Now we see why it is OK for (1.36) to have φ on the left-hand side and θ on the right-hand side: θ is given as a function of φ by the inverse aster transform so the right-hand side of (1.36) can be considered a function of φ.

Now we can express (1.34) in exponential family form

$$l(\varphi) = \langle y, \varphi \rangle - c(\varphi) \tag{1.38}$$

where the cumulant function c of the joint distribution of the aster model is (1.36).

We call

- θ the *conditional canonical parameter vector* and
- φ the *unconditional canonical parameter vector*.

But this terminology makes these parameters more like two kinds of the same sort of thing than they really are.

- The subvectors θ_G are the canonical parameter vectors of the conditional distributions of the dependence groups, but, as we have seen (this is the whole point of the aster transform), θ is not the canonical parameter vector of the (unconditional, joint) distribution of the aster model.
- The vector φ is the canonical parameter vector of the (unconditional, joint) distribution of the aster model, but the subvectors φ_G or the components φ_j are not separately canonical for anything.

In short the subvectors of θ are groupwise canonical but θ is not vectorwise canonical, and, conversely, φ is vectorwise canonical, but its subvectors are not groupwise canonical.

1.19 More On Exponential Families

1.19.1 Directions of Constancy, Identifiability

A direction δ in the vector space where the canonical parameter vector of an exponential family takes values is called a *direction of constancy* if the

random variable $\langle Y, \delta \rangle$ is constant almost surely, where Y is the canonical statistic vector. Because all of the distributions in an exponential family have the same support (this follows from the log likelihood being finite for all all data values when the canonical parameter value is in the full canonical parameter space (1.27)) we don't have to say which distribution in the family almost surely refers to: if a property holds almost surely for one distribution in the exponential family, then it holds almost surely for all distributions in the exponential family.

Geyer (2009, Theorem 1) gives many equivalent characterizations of this concept. Here are some of them. In these we use the notation y is the canonical statistic vector, θ is the canonical parameter vector, δ is a vector in the vector space where θ takes values, l is the log likelihood (1.25), and Θ is the full canonical space (1.27).

(a) For some $\theta \in \Theta$, the function $s \mapsto l(\theta + s\delta)$ is finite on some open interval and is not strictly concave on that interval.

(b) For all $\theta \in \Theta$, the function $s \mapsto l(\theta + s\delta)$ is constant on the whole real line.

(c) The parameter vectors θ and $\theta + s\delta$ correspond to the same distribution for some $\theta \in \Theta$ and some $s \neq 0$.

(d) The parameter vectors θ and $\theta + s\delta$ correspond to the same distribution for all $\theta \in \Theta$ and all real s.

(e) The random variable $\langle Y, \delta \rangle$ is almost surely constant for some distribution in the exponential family.

(f) The random variable $\langle Y, \delta \rangle$ is almost surely constant for all distributions in the exponential family.

We arbitrarily picked (e) to serve as the definition, but, because they are all equivalent (Geyer, 2009, Theorem 1), we could have picked any of these to serve as the definition.

Condition (b) explains the terminology; these are directions of constancy of the log likelihood function as defined in Rockafellar (1970). They obviously imply, if $\hat{\theta}$ is a maximum likelihood estimate (MLE), then so is $\hat{\theta} + s\delta$ for all real s. Condition (a) gives a uniqueness condition for maximum likelihood. If δ is not a direction of constancy and $s \neq 0$, then θ and $\theta + s\delta$ cannot both be maximum likelihood estimates because (a) would imply the log likelihood is higher somewhere between these points. Thus the MLE for

a full exponential family (if it exists) is unique if and only if there are no nonzero directions of constancy.

It is clear from condition (e) or (f) that the set of all directions of constancy of an exponential family (with a particular canonical parameterization) is a vector subspace, which is called the *constancy space* of the family.

Conditions (c) and (d) connect this concept with identifiability. If there exist two distinct canonical parameter vectors θ and θ^* that correspond to the same distribution, then $\theta - \theta^*$ is a direction of constancy. Thus this concept characterizes the only form of nonidentifiability the canonical parameterization of an exponential family can have. Conversely, if there are no nonzero directions of constancy, then the canonical parameterization is identifiable.

However, we do not insist on identifiability for three reasons.

- As Geyer (2009) says, non-identifiability "is, at worst, merely a computational nuisance" so it is a big mistake to contort theory and applications to achieve identifiability too early. The computer can put it in at the end with no help from humans (for how see Section 1.20.3 below).

- Multinomial dependence groups in aster models require non-identifiable parameterization (Section D.9 in Appendix D).

- Limiting conditional models require non-identifiable parameterization if they are to use the same parameterization as the original model (Chapter 2).

Conditions (e) and (f) connect this concept with multivariate degeneracy. The vector δ is a direction of constancy if and only if the random variable $\langle Y, \delta \rangle$ is almost surely constant, which is the same thing as saying the canonical statistic vector Y is concentrated on a hyperplane. So the canonical parameterization of a full exponential family is identifiable if and only if Y is not concentrated on a hyperplane. A random variable that has finite variance is constant almost surely if and only if that variance is zero. Hence we can also characterize directions of constancy in terms of the variance matrix of the canonical statistic (1.32b). In a regular full exponential family, let

$$I(\theta) = \operatorname{var}_\theta(Y).$$

(we need a regular full exponential family to guarantee the variance exists). This is the Fisher information matrix as well as the variance matrix of

the canonical statistic because differentiating the log likelihood (1.25) twice gives

$$l''(\theta) = -c''(\theta) = -\operatorname{var}_\theta(Y).$$

Then clearly, δ is a direction of constancy if and only if

$$\operatorname{var}_\theta(\langle Y, \delta \rangle) = \delta^T I(\theta) \delta$$

is zero. And by the spectral theorem Halmos (1958, Section 79) this holds if and only if $I(\theta)\delta = 0$, in words, if and only if δ is a null eigenvector of the Fisher information matrix.

1.19.2 Mean Value Parameters

Theorem 1.4. *For a regular full exponential family, every distribution has a mean value (for the canonical statistic vector), and different distributions have different mean values. Hence mean values parameterize a regular full exponential family, and the mean value parameterization is always identifiable, whether or not the canonical parameterization is identifiable.*

Proof. We already know that mean values exist and are given by (1.32a). Let us temporarily adopt a notation for the map from canonical to mean value parameters: $f : \theta \mapsto c'(\theta)$. This map is differentiable with derivative matrix

$$f'(\theta) = c''(\theta) = I(\theta).$$

Consider distinct distributions in the exponential family having canonical parameter vectors θ and θ^*. Then $\theta - \theta^*$ is not a direction of constancy. Define a function g by

$$g(s) = f\big(s\theta + (1 - s)\theta^*\big).$$

Then the corresponding mean values are

$$\begin{aligned} \mu &= g(1) = f(\theta) \\ \mu^* &= g(0) = f(\theta^*) \end{aligned}$$

Differentiating g gives

$$\begin{aligned} g'(s) &= f'\big(s\theta + (1 - s)\theta^*\big)(\theta - \theta^*) \\ &= I\big(s\theta + (1 - s)\theta^*\big)(\theta - \theta^*) \end{aligned}$$

Since $I\big(s\theta+(1-s)\theta^*\big)$ is a variance matrix, it is positive semi-definite. Since $\theta - \theta^*$ is not a direction of constancy,

$$(\theta - \theta^*)^T I\big(s\theta + (1 - s)\theta^*\big)(\theta - \theta^*) \tag{1.39}$$

is strictly positive for all s in some open interval of the real line containing 0 and 1 (because θ and θ^* are interior points of the full canonical parameter space). We note that (1.39) is the derivative of the function h defined by

$$h(s) = \langle g(s), \theta - \theta^* \rangle$$

Since the derivative of this function is strictly positive it is strictly increasing on some open interval containing 0 and 1, hence

$$h(1) - h(0) = \langle g(1) - g(0), \theta - \theta^* \rangle = \langle \mu - \mu^*, \theta - \theta^* \rangle$$

is strictly positive. Hence $\mu \neq \mu^*$. □

The parameter vector $\mu = c'(\theta)$ is called the *mean value parameter* of the regular full exponential family. (A full exponential family that is not regular need not have a mean value for every distribution. It must have a mean value for every distribution whose canonical parameter vector θ in the interior of the full canonical parameter space Θ, but it need not have mean values for distributions whose θ is on the boundary of Θ.)

1.19.3 Calculating the Inverse Transformation

Discussed in the proof of the preceding theorem was the transformation $\theta \mapsto c'(\theta)$, which the theorem says maps from a not necessarily identifiable parameter (θ) to a necessarily identifiable parameter (μ). In this section we find out this parameter transformation is invertible if and only if the canonical parameterization is identifiable, and we also find out how to calculate the inverse if it exists.

Define a function $\tilde{l}$ that is just like the log likelihood of the exponential family (1.25) except that we replace the observed value of the canonical statistic vector y with a possible mean value

$$\tilde{l}(\theta) = \langle \mu, \theta \rangle - c(\theta), \tag{1.40}$$

where $\mu = E_\theta(Y)$ for some θ.

Lemma 1.5. *The cumulant function of a full exponential family is a convex function. It is strictly convex if and only if the exponential family has no nonzero directions of constancy.*

This is Theorem 7.1 in Barndorff-Nielsen (1978). An alternative proof for regular full exponential families can use the second derivative test in Theorem 2.14 in Rockafellar and Wets (1998). The function c is convex because $c''(\theta)$ is a variance matrix, hence positive semi-definite for all θ. When there are no nonzero directions of constancy, $c''(\theta)$ is positive definite for all θ, so c is strictly convex.

The derivatives of (1.40) are

$$\tilde{l}'(\theta) = \mu - c'(\theta) \tag{1.41a}$$

$$\tilde{l}''(\theta) = -c''(\theta) \tag{1.41b}$$

Applying the second derivative test in Theorem 2.14 in Rockafellar and Wets (1998) gives

Lemma 1.6. *The function $\tilde{l}$ defined by* (1.40) *is a concave function. It is strictly concave if and only if the exponential family has no nonzero directions of constancy.*

Lemma 1.7. *For a differentiable convex function, any point where the derivative is zero is a global minimizer. For a differentiable strictly convex function, any point where the derivative is zero is a unique global minimizer.*

For a differentiable concave function, any point where the derivative is zero is a global maximizer. For a differentiable strictly concave function, any point where the derivative is zero is a unique global maximizer.

This is part of Theorem 10.1 in Rockafellar and Wets (1998).

Theorem 1.8. *Global maximizers of the function $\tilde{l}$ exist, and any global maximizer θ satisfies $\mu = c'(\theta)$ and hence is a canonical parameter vector corresponding to the mean value parameter μ.*

In case the canonical parameterization is identifiable, which is when there are no nonzero directions of constancy, the function $\tilde{l}$ has a unique global maximizer.

Proof. By assumption $\mu = E_\theta(Y) = c'(\theta)$ for some θ, hence by Lemma 1.7 this θ is a global maximizer of $\tilde{l}$. If there are other global maximizers θ^*, then they also satisfy $\mu = c'(\theta^*)$. Hence, such θ and θ^* correspond to the same μ. Hence θ and θ^* correspond to the same distribution by Theorem 1.4. Hence, if the canonical parameterization is identifiable, then $\theta = \theta^*$. □

Theorem 1.9. *For a regular full exponential family with identifiable canonical parameterization, any algorithm that always goes uphill if it can on $\tilde{l}$*

defined by (1.40) *converges to the unique maximizer of* $\tilde{l}$ *which is the unique canonical parameter vector* θ *corresponding to mean value parameter vector* μ*.*

Proof. Any cumulant function is lower semicontinuous and convex (Barndorff-Nielsen, 1978, Theorem 7.1). Hence (1.40) defines an upper semicontinuous concave function. The rest of the theorem is Corollary 27.2.2 in (Rockafellar, 1970). □

Corollary 1.10. *For a regular full exponential family with identifiable canonical parameterization, any algorithm that always goes uphill if it can on the log likelihood defined by* (1.25) *converges to the unique MLE if the observed value of the canonical statistic vector* y *is a possible value of the mean value parameter vector.*

Proof. In case $y = \mu$ for some mean value parameter vector μ, this is a special case of Theorem 1.9. □

The point of the theorem and its corollary is that maximization of upper semicontinuous strictly concave functions is easy. Any algorithm, no matter what the starting point, and no matter how inefficient, can find the unique solution, so long as it is not so stupid as to be going downhill when it should be going uphill.

Of course we do not have to use inefficient algorithms, we can use algorithms as clever as we want, but we must be careful to not be so clever that we turn out to be stupid. No matter what algorithm we use it must have the property of always going uphill if it can. The algorithms used by R function `aster` in R package `aster` have this property. The algorithm used by R function `transformUnconditional` in R package `aster2` has this property.

Corollary 1.10 hides a big issue. There does not have to be a mean value parameter vector μ such that $\mu = y$, and in this case the MLE does not exist: there is no θ that maximizes (1.25) because we can always go uphill from any θ. This is the subject of Section 1.23.2 and Chapter 2.

1.20 Aster Mean Value Parameters

We have already met the mean value parameters of aster models in Section 1.13 above. They are

- the *unconditional mean value parameter vector* μ defined by (1.3) and

- the *conditional mean value parameter vector* ξ defined by (1.4) when the conditioning event in that equation makes sense and by the discussion following that equation otherwise.

In Section 1.13 above we were getting ahead of ourselves in two ways. We didn't yet know that these mean values existed, and we didn't yet know that they parameterized the model. Now we do. So we fill in the details on that.

1.20.1 Unconditional

Theorem C.1 in Appendix C says, if each family for a dependence group is a regular full exponential family, then the (unconditional, joint) distribution of the aster model is a regular full exponential family. All of the families for dependence groups (including dependence groups that are singleton sets) that have ever been implemented in aster software are regular full exponential families. Thus all aster models that have ever been implemented are regular full exponential families.

And hence we know how to map from aster model canonical parameter vectors to mean value parameter vectors and vice versa. The maps from canonical to mean value are

$$\mu = c'(\varphi) = E_\varphi(Y), \tag{1.42}$$

where c is the cumulant function of the (unconditional, joint) aster model (1.36), and Y is the response vector of the aster model, and

$$\xi_G = c'_G(\theta) = E_\theta(Y_G \mid Y_{q(G)} = 1), \qquad G \in \mathcal{G}, \tag{1.43}$$

where c_G is the cumulant function of dependence group G.

The inverse mappings to these mappings do not, in general, have closed form expressions but are described in Section 1.19.3 above. Apply Theorem 1.8 to the regular full exponential exponential family that is the saturated aster model in its unconditional canonical parameterization. Then (1.40) becomes in this special case

$$\tilde{l}(\varphi) = \langle \mu, \varphi \rangle - c(\varphi) \tag{1.44}$$

where μ is a possible value of the unconditional mean value parameter vector of the aster model and c is the cumulant function of the aster model (1.36). Then any point φ where the first derivative of $\tilde{l}$ is zero is a global maximizer of $\tilde{l}$ and is an unconditional canonical parameter vector φ that corresponds to

μ. Such a global maximizer always exists. If the canonical parameterization is identifiable (if there are no nonzero directions of constancy), then this global maximizer is unique. All of the above follows from Theorems 1.8 and C.1.

When the canonical parameterization is identifiable, this process allows us to calculate the map $\mu \mapsto \varphi$ that is the inverse to to the map $\varphi \mapsto \mu$ given by $\varphi \mapsto c'(\varphi)$. We know the forward transformation $\varphi \mapsto \mu$ is infinitely differentiable (because cumulant functions of regular full exponential families are infinitely differentiable). Hence it follows from the inverse function theorem of real analysis (Browder, 1996, Theorem 8.27) that the inverse function is infinitely differentiable if it exists (we are not using the inverse function theorem to prove the existence of a local inverse; we know a global inverse function exists if and only if the unconditional canonical parameterization is identifiable; we are using the assertion of the inverse function theorem about differentiability of this inverse function).

In summary, we know the inverse mapping $\mu \mapsto \varphi$ exists and is infinitely differentiable when φ is identifiable, but we generally have no closed-form expression for this mapping and must calculate it by maximizing $\tilde{l}$ defined by (1.44).

Another conclusion of the inverse function theorem is that the derivative of the inverse is the inverse of the derivative (this is not always stated in the theorem statement, but does appear in the proof). In our case, the derivative of the forward mapping $\varphi \to \mu$ is $c''(\varphi) = I(\varphi)$. We know this matrix is invertible if and only if φ is identifiable. And when it is invertible (when there are no nonzero directions of constancy, when Y is not concentrated on a hyperplane) the derivative of the inverse mapping $\mu \to \varphi$ is $c''(\varphi)^{-1}$. In more detail, if we temporarily give this inverse mapping a notation $f : \mu \to \varphi$ given by φ is the unique global maximizer of $\tilde{l}$ defined by (1.44) (assuming the inverse mapping exists, that is, assuming identifiablilty). Then

$$f'(\mu) = I(\varphi)^{-1}, \qquad \text{when } \varphi = f(\mu),$$

or

$$f'(\mu) = I\big(f(\mu)\big)^{-1}.$$

1.20.2 Conditional

Now we go over all of the preceding section again for conditional mean values and dependence groups. We fuss less about the details because they are the same *mutatis mutandis*. For dependence group G, define $\tilde{l}_G$ by

$$\tilde{l}_G(\theta_G) = \langle \xi_G, \theta_G \rangle - c_G(\theta_G) \tag{1.45}$$

where ξ_G is a possible value of the conditional mean value parameter vector for G and c_G is the cumulant function for G. Then any point θ_G where the first derivative of $\tilde{l}_G$ is zero is a global maximizer of $\tilde{l}_G$. Such a global maximizer always exists. If the canonical parameterization is identifiable (if there are no nonzero directions of constancy of the family for dependence group G), then this global maximizer is unique.

When we do this procedure for each dependence group G this defines a mapping $\xi \to \theta$ if each dependence group has identifiable canonical parameterization. This mapping is infinitely differentiable if it exists, and the derivative is given by inverting the derivative of the forward mapping. The forward mapping has block diagonal derivative with components

$$\begin{aligned} \frac{\partial \xi_j}{\partial \theta_k} &= \frac{\partial^2 c_G(\theta_G)}{\partial \theta_j \partial \theta_k}, && j, k \in G, \\ \frac{\partial \xi_j}{\partial \theta_k} &= 0, && \text{otherwise.} \end{aligned}$$

If we denote the matrix with these components by $I(\theta)$, then $I(\theta)^{-1}$ is the derivative of the inverse mapping at a point ξ that corresponds to θ.

1.20.3 Dealing with Non-Identifiability

We are used to the computer dealing with non-identifiability, also called collinearity, automatically. R functions `lm` and `glm` in core R (R Core Team, 2023), which fit linear and generalized linear models, handle non-identifiability automatically by dropping regressor vectors (columns of the model matrix) until identifiability is obtained. They signal this by reporting `NA` for coefficient estimates corresponding to these dropped regressor vectors.

R package `aster` does more or less the same thing as R functions `lm` and `glm` except that it simply does not mention regression coefficients for dropped regressors (it does say which regressors these are in the component `dropped` of the object returned by R function `aster`).

In aster models fit by R package `aster` (which cannot have dependence groups corresponding to more than one component of the response vector) the only way that nonidentifiability of the canonical parameterization can arise is if the model matrix does not have full column rank, which is getting ahead of ourselves because we have not discussed canonical affine submodels of aster models yet (next section).

In aster models fit by R package `aster2` the aster model can have non-identifiable canonical parameterizations (θ is nonidentifiable if and only if φ is nonidentifiable because the aster transform is one-to-one) when the model

has multinomial dependence groups or when we are analyzing a limiting conditional model of the aster model.

R package `aster2` does more or less the same thing as R functions `lm` and `glm` except that it returns zero for certain coefficients on the theory that constraining a parameter to be equal to zero is equivalent to dropping the corresponding regressor.

From a theoretical point of view there is just one issue: directions of constancy (Section 1.19.1 above). Suppose we have an exponential family in which the canonical parameter is β (this includes canonical affine submodels but also the saturated models they are submodels of). If δ is a direction of constancy, then β and $\beta + s\delta$ correspond to the same distribution for all s. If $\delta \neq 0$, then we do not have identifiability, and need to decide how to regain it. The way to do so that R users (at least) have been trained to expect is to constrain some β_j to be equal to zero.

If $\delta \neq 0$, then it has a nonzero component, say δ_i. Then we can constrain β_i to be equal to zero, because for any β, the parameter vector $\beta + s\delta$ with

$$\beta_i + s\delta_i = 0$$

or

$$s = -\beta_i / \delta_i$$

has i-th component zero and corresponds to the same distribution as β.

So this procedure of obtaining identifiability by constraining some parameter to be equal to zero works for any exponential family (and works for GLM that are not exponential families by them being close enough to exponential families in the theoretical aspects that are necessary for this to work).

This procedure can always be implemented by the computer provided it has a way of determining directions of constancy. As stated in Section 1.19.1 above, the set of all directions of constancy is a vector subspace of the vector space where the canonical parameter takes values, which is called the *constancy space*. Hence the set of all directions of constancy can be characterized by providing a basis for the constancy space.

So long as aster software has a function like R function `constancy` in R package `aster2` that provides a basis for the constancy space, the computer can do the rest of dealing with non-identifiability.

1.21 A Plethora of Parameters

We now have the parameterizations θ, φ, ξ, and μ for aster models. We now know that

- μ is always identifiable,
- ξ is always identifiable unless some $\mu_{q(G)} = 0$,
- φ is identifiable if and only if there are no nonzero directions of constancy, and
- θ is identifiable if and only if φ is identifiable.

The simple closed-form expressions we have for these parameter transformations are shown below.

$$\begin{array}{ccc} \theta & \underset{\text{inverse aster transform}}{\overset{\text{aster transform}}{\rightleftarrows}} & \varphi \\ {\scriptstyle \xi_G = c'_G(\theta_G)}\downarrow\uparrow & & \uparrow\downarrow{\scriptstyle \mu = c'(\varphi)} \\ \xi & \underset{\text{division}}{\overset{\text{multiplication}}{\rightleftarrows}} & \mu \end{array} \tag{1.46}$$

where (of course) "aster transform" is (1.35) above, "inverse aster transform" is (1.37) above, "multiplication" is (1.7) or (1.9) or (1.10) above, "division" is (1.11), and the equations labeling downwards arrows refer to (1.43) and (1.42) above.

Both upward arrows being unlabeled indicates no closed-form expressions for going from mean value parameters to canonical parameters. There is an algorithm (Section 1.19.3 above) but (in general) no closed-form expression. All of the arrows do, however, indicate smooth (infinitely differentiable) changes of parameter, if they correspond to a change of parameter at all.

The arrows that may not correspond to a change of parameter arrow are

- the arrow $\mu \mapsto \xi$ labeled "division" fails to be a map when there is division by zero, in which case ξ is not identifiable,
- the arrows $\xi \mapsto \theta$ and $\mu \mapsto \varphi$, which are unlabeled fail to be maps when the canonical parameterizations are not identifiable (and either both are identifiable or neither are).

As explained in Section 1.20.3 above, identifiability of either canonical parameterization can be recovered by constraining certain parameters to be equal to zero. When we do this we need to be careful to not mess up the other parameter transformations.

- If we constrain some components of φ to be equal to zero in order to make φ identifiable and the transformation $\mu \to \varphi$ a map, then this also makes θ identifiable, but since the inverse aster transform is nonlinear this does not correspond to setting any components of θ to be equal to zero.

- Conversely, if we constrain some components of θ to be equal to zero in order to make θ identifiable and the transformation $\xi \to \theta$ a map, then this also makes φ identifiable, but since the aster transform is nonlinear this does not correspond to setting any components of φ to be equal to zero.

- When some components of the response vector that are predecessors are equal to zero almost surely, ξ is not identifiable: if $\mu_{q(G)} = 0$, then ξ_G can be chosen arbitrarily, but then θ_G will be determined by this arbitrary choice, and then φ will be determined by θ. Since μ is always identifiable, it is unchanged when ξ_G is allowed to be chosen arbitrarily (the arbitrary choice is multiplied by zero in the conversion to μ).

All of these parameterizations are important. Any may be crucial in some applications and irrelevant to other applications. From this point forward we now call the statistical model we have been discussing the *saturated aster model*. No matter which parameterization we choose, the length of the parameter vector is the same as the length of the response vector (both have the same index set J).

The saturated aster model has too many parameters chasing too little data. More parsimonious models are used in applications.

The next few sections introduce these more parsimonious models and even more parameters, and Sections 1.22.2 and 1.22.4 below discuss their transformations similar to how this section does.

1.22 Aster Canonical Affine Submodels

It may come as a shock to the reader that all of the work done so far concerns models that are of no interest in applications.

1.22.1 Unconditional

In this section we straightforwardly apply the theory of Section 1.15.3 above, which describes canonical affine submodels of general exponential families, to aster models. In Section 1.15.3 the canonical parameter vector of

the exponential family was denoted θ. The canonical parameter vector of the (unconditional, joint) aster model is the *unconditional* canonical parameter vector φ. Thus we parameterize an aster *unconditional canonical affine submodel*

$$\varphi = a + M\beta, \tag{1.47}$$

where, as stated in Section 1.15.3 above, a is the offset vector and M is the model matrix. Either a or M may depend on covariate data since aster analyses, like all regression analyses, are done conditional on covariate data. Usually a does not depend on covariate data and M does depend on covariate data, and usually the i-th row of M depends only on data for case i, but this actually places no restriction on M because what is considered "data for case i" is arbitrary. This is the only place where covariate data enters an unconditional aster model (for short we refer to such models as *unconditional aster models* rather than *unconditional canonical affine submodels of aster models*).

From the theory in Section 1.15.3 above, an unconditional aster model is itself a regular full exponential family with

- canonical statistic vector $M^T y$,
- canonical parameter vector β,
- cumulant function c_{sub} defined by

  $$c_{\text{sub}}(\beta) = c(a + M\beta)$$

  where c is the cumulant function of the saturated aster model given by (1.36) above, and
- mean value parameter

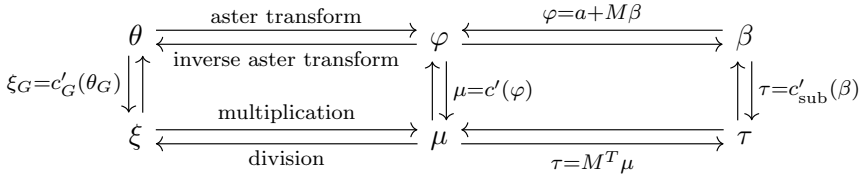
  $$\tau = c'_{\text{sub}}(\beta) = E_\beta(M^T Y) = M^T \mu,$$

  where μ is the saturated model mean value parameter.

Theorem C.2 in Appendix C says that every unconditional aster model is a regular full exponential family if the saturated model is a regular full exponential family. Theorem C.1 in Appendix C says the latter happens when every family for every dependence group is a regular full exponential family, which is the case for all families currently implemented in aster software.

1.22.2 A Plethora of Parameters Revisited

With unconditional canonical affine submodels (preceding section) we get two more parameters, so we can change our picture to



(1.48)

where (of course) all of the parameters and arrows for the "left square" are the same as in (1.46), where the three of the arrows for the "right square" give parameter transformations discussed in the preceding section, and the three unlabeled arrows are parameter transformations that have no closed-form expression.

The inverse mapping to the the mapping $\beta \mapsto \tau$ is calculated just like the other upward arrows in the diagram. Apply Theorem 1.8 to the regular full exponential exponential family that is the unconditional canonical affine submodel. Then (1.40) becomes in this special case

$$\tilde{l}_{\text{sub}}(\beta) = \langle \tau, \beta \rangle - c_{\text{sub}}(\beta) \tag{1.49}$$

as with the other $\tilde{l}$ functions, $\tilde{l}_{\text{sub}}$ is concave and strictly concave when the canonical parameter β of the unconditional canonical affine submodel is identifiable, which is when there does not exist a nonzero direction of constancy, a nonzero vector η such that $\langle M^T Y, \eta \rangle = \langle Y, M\eta \rangle$ is almost surely constant. Whenever τ is a possible value of the mean value parameter vector, that is, $\tau = E_\beta(M^T Y)$ for some β, then that β is a global maximizer of (1.49). That β is the unique maximizer if and only if β is identifiable (no nonzero directions of constancy exist).

As discussed in Section 1.20.3 above, we can always deal with non-uniqueness (non-identifiability) by constraining some components of β to be equal to zero.

The other unlabeled arrows on the "right square" in the diagram are the inverses of linear transformations, but these generally do not have unique solutions.

The mapping $\varphi = a + M\beta$ need be neither one-to-one nor onto. In applications, it is never onto. That is the whole point of submodels, to be a *proper* submodel. It will not be one-to-one if there is "collinearity" (if M does not have full column rank). Let B denote the full canonical parameter

space for the parameter β described in the Theorem C.2 in Appendix C and its proof. Temporarily, let f denote the mapping defined by $f(\beta) = a + M\beta$. As we just said, f is one-to-one if and only if M has full column rank, but f is never onto (in applications). Thus f is invertible (if one-to-one) only when it is considered a mapping $B \to f(B)$, that is, its codomain is its range. Then (assuming one-to-one) the inverse f^{-1} exists and is a mapping $f(B) \to B$.

But this is a triviality. It says that if $\varphi = f(\beta)$, then we know that φ is in the domain of the inverse mapping and $\beta = f^{-1}(\varphi)$. But otherwise, when just given a φ in the saturated model unconditional canonical parameter space, we don't know whether it is in $f(B)$ or not, since the only way we can know is if we know it is $f(\beta)$ for some β.

The mapping $\tau = M^T\mu$ need be neither one-to-one nor onto. In applications, it is never one-to-one. Again, that is the whole point of submodels. If M is not onto, then M^T is not one-to-one (considered as the linear transformations these matrices represent). Thus the equation $\tau = M^T\mu$ never has a unique solution for μ. The only way to find a μ that corresponds to a given τ is to go the other way around the square $\tau \mapsto \beta \mapsto \varphi \mapsto \mu$.

For an unconditional canonical affine submodel

- τ is always identifiable,
- β may or may not be identifiable, if not, identifiability can be forced by constraining some components of β to be equal to zero,
- the set of allowed $\varphi = a + M\beta$ values is either an affine subspace of $\mathbb{R}^J$ or an open subset of such an affine subspace,
- the set of allowed μ values is a smooth manifold having the same dimension as β, but is a curved manifold since the mapping $\varphi \to \mu$ is nonlinear,
- similarly, the set of allowed θ values is a smooth manifold having the same dimension as β, and
- similarly, the set of allowed ξ values is a smooth manifold having the same dimension as β.

1.22.3 Conditional

Since all of the theory in the preceding two sections works so smoothly, it may be surprising that we have a competing view of how to do submodels.

We almost don't have a competing view. As far as I know, there are only two published examples of conditional aster analyses. Example 1 in Shaw et al. (2008b) was done as a conditional aster analysis but can easily be redone as an unconditional aster analysis (Geyer, 2018, Slide Deck 4, slides 47 ff.). The analysis in Shaw et al. (2015) titled "Relating Plant Fitness to Aphid-Load" was done as a conditional aster analysis and had to be so done because time-dependent covariates (in this case aphid load) do not mesh well with unconditional aster models.

So in rare cases (one so far published, as far as I know) conditional aster models are necessary. But in the vast majority of applications unconditional aster models (which are the default for aster software) are the ones used and the only ones users consider.

Nevertheless, if we have conditional aster models at all, then we need their theory. In *conditional canonical affine submodels* of aster models we do not use the submodel parameterization (1.47). Instead we use

$$\theta = a + M\beta, \tag{1.50}$$

This is undeniably a TTD (thing to do) but has no theoretical justification. It does not connect with exponential family theory except, as a smooth submodel of a regular full exponential family (the saturated model), this is a *curved exponential family*. As we shall see, there are a few good properties that are shared by conditional and unconditional aster models, but only a few. The theory of conditional aster models is impoverished compared to that of unconditional aster models.

1.22.4 A Plethora of Parameters Re-Revisited

The analog of (1.46) or (1.48) for conditional aster models is

$$\begin{array}{ccccc} \beta & \underset{}{\overset{\theta=a+M\beta}{\longrightarrow}} & \theta & \underset{\text{inverse aster transform}}{\overset{\text{aster transform}}{\rightleftarrows}} & \varphi \\ & & \Big\downarrow \xi_G = c_G'(\theta_G) \Big\uparrow & & \Big\uparrow \Big\downarrow \mu = c'(\varphi) \\ & & \xi & \underset{\text{division}}{\overset{\text{multiplication}}{\rightleftarrows}} & \mu \end{array} \tag{1.51}$$

There is no analog of the submodel mean value parameter τ for a conditional aster model (because a conditional aster model is not a regular full exponential family).

1.23 Maximum Likelihood

1.23.1 Exponential Families

Maximum likelihood estimation in a regular full exponential family is much like the problem of inverting the parameter transformation from canonical to mean value parameters discussed in Section 1.19.3 above. We simply maximize l given by (1.25) rather than $\tilde{l}$ given by (1.40). Both of these functions are always concave and strictly concave if and only if the canonical parameterization is identifiable.

A point is a global maximizer of l or $\tilde{l}$ if and only if the first derivative is equal to zero. Hence θ is an MLE if and only if it satisfies the *observed equals expected property*

$$y = E_\theta(Y) \tag{1.52}$$

where the left hand side is the observed value of the canonical statistic and the right-hand side is the expected value corresponding to parameter value θ. If the canonical parameterization is identifiable (if the canonical statistic is not concentrated on a hyperplane, if there are no nonzero directions of constancy), the solution θ of (1.52).

The only difference between the problem of this section and Section 1.19.3 above is that there we *assumed* there was a solution, that is, we assumed that the μ in (1.40) satisfied the expected equals expected condition (which is a triviality) $\mu = E_\theta(Y)$. Here we do not assume (1.52) has a solution. The data are what they are. We assume they are possible data under the statistical model, but we do not assume anything else.

All of the same arguments apply to canonical affine submodels, only the notation changes. Replace y by $M^T y$ and replace θ by β. Thus β is a global maximizer of the submodel log likelihood if and only if

$$M^T y = M^T E_\beta(Y) \tag{1.53}$$

1.23.2 Directions of Recession, Existence

A direction δ in the vector space where the canonical parameter vector of an exponential family takes values is called a *direction of recession* of the log likelihood if

$$\langle Y, \delta \rangle \le \langle y, \delta \rangle, \qquad \text{almost surely,} \tag{1.54}$$

where y is the observed value of the canonical statistic vector and Y is a random value of the canonical statistic vector.

Suppose δ is a nonzero direction of recession that is not a direction of constancy. By monotonicity of expectation we have

$$E_\theta\{\langle Y, \delta\rangle\} < \langle y, \delta\rangle$$

for all parameter values θ, because we know $\langle y - Y, \delta\rangle$ is nonnegative almost surely, so if it had zero expectation, it would have to be zero almost surely (which would imply that δ is a direction of constancy). Consequently, if δ is a nonzero direction of recession that is not a direction of constancy, then (1.52) has no solutions and maximum likelihood estimates for θ do not exist. It turns out this is the only way the MLE for a regular full exponential family can fail to exist.

Geyer (2009, Theorem 3) gives many equivalent characterizations of this concept. Here are some of them. In these we use the notation y is the observed value of the canonical statistic vector, Y is a random value of the canonical statistic vector, θ is the canonical parameter vector, δ is a vector in the vector space where θ takes values, l is the log likelihood (1.25), and Θ is the full canonical space (1.27).

(a) For some $\theta \in \Theta$,
$$\limsup_{s\to\infty} l(\theta + s\delta) > -\infty.$$

(b) For all $\theta \in \Theta$, the function $s \mapsto l(\theta + s\delta)$ is nondecreasing on the whole real line.

(c) $\langle Y - y, \delta\rangle \le 0$ almost surely for some distribution in the exponential family.

(d) $\langle Y - y, \delta\rangle \le 0$ almost surely for every distribution in the exponential family.

We arbitrarily picked (c) to serve as the definition, but, because they are all equivalent we could have picked any of these to serve as the definition of direction of recession. All of these equivalences are a consequence of Theorem 3 in Geyer (2009), except the implication that (a) implies the others uses Theorem 8.6 in Rockafellar (1970).

By Theorem 4 in Geyer (2009) the MLE exists if and only if every direction of recession is a direction of constancy. (It is clear from the definitions that every direction of constancy is a direction of recession.)

By Theorem 5 in Geyer (2009) if δ is a direction of recession that is not a direction of constancy, then the function

$$s \mapsto l(\theta + s\delta)$$

is strictly increasing on the interval where it is finite, and this is true for all $\theta \in \Theta$. This gives us another argument, besides the argument given above why a direction of recession that is not a direction of constancy implies nonexistence of an MLE. But we need Theorem 4 in Geyer (2009) for the reverse conclusion that nonexistence of the MLE cannot occur for any other reason.

All of the same arguments apply to unconditional canonical affine submodels, only the notation changes. Replace y by $M^T y$ and replace θ by β. Thus a direction δ in the vector space where the submodel canonical parameter β takes values is a direction of recession if and only if $\langle y - Y, M\delta \rangle \geq 0$ almost surely (for any one distribution in the submodel and hence for all distributions). And MLE for β exist if and only if $\langle y - Y, M\delta \rangle \geq 0$ almost surely implies $\langle y - Y, M\delta \rangle = 0$ almost surely. Moreover, β is an MLE if and only if (1.53) holds.

A lot more can be said and will be said on this subject (Chapter 2).

1.23.3 Unconditional Aster Models

Almost nothing needs to be said in this section that hasn't already been said. Unconditional canonical affine submodels of aster models are regular full exponential families (under the conditions of Theorems C.1 and C.2 in Appendix C). So the penultimate paragraph of the preceding section says it all.

1.23.4 Conditional Aster Models

Conditional aster models are not regular full exponential families (only curved exponential families). At least, they are not when considered statistically, probabilistically.

But if we play a trick, they are when considered numerically, algebraically. Rewrite (1.33) as

$$l(\theta) = \sum_{G \in \mathcal{G}} \left[\langle y_G, \theta_G \rangle - n_{q(G)} c_G(\theta_G) \right] \tag{1.55}$$

and pretend that all of the $n_{q(G)}$ are constants. This makes no sense statistically, probabilistically because those $n_{q(G)}$ are actually $y_{q(G)}$ and some of them are the same variables that appear as components of the y_G that also appear in this expression. But if we are just treating this log likelihood as a numerical, algebraic function to maximize to find MLE, then this doesn't matter.

We call this the *associated independence model* of the conditional aster model. In this model y factorizes as

$$\operatorname{pr}(y) = \prod_{G \in \mathcal{G}} \operatorname{pr}(y_G \mid n_{q(G)}) \tag{1.56}$$

where all of the $n_{q(G)}$ are considered constants and not components of y. Then of course, we have the parameter transformation (1.50) of the canonical affine submodel. This is a regular full exponential family (Theorem C.3 in Appendix C).

So now we can apply the theory of Geyer (2009) to the associated independence model.

Theorem 1.11. *If $y_{q(G)} = 0$ for any dependence group G, replace the family for this dependence group by the degenerate family concentrated at zero so c_G is the zero function. Then η is a direction of recession of the associated independence model if and only if, writing $\delta = M\eta$,*

$$E_\beta\{\langle y_G - Y_G, \delta_G \rangle \mid y_{q(G)}\} \geq 0, \qquad \text{for all } G \in \mathcal{G} \text{ such that } y_{q(G)} > 0,$$

and η is a direction of constancy of the associated independence model if and only if

$$E_\beta\{\langle y_G - Y_G, \delta_G \rangle \mid y_{q(G)}\} = 0, \qquad \text{for all } G \in \mathcal{G} \text{ such that } y_{q(G)} > 0.$$

Then the MLE for β exists if and only if every direction of recession is a direction of constancy, and the MLE is unique if and only if no nonzero directions of constancy exist.

Note that the theorem ignores δ_G for G such that $y_{q(G)} = 0$, hence unless the model matrix M is such that $\delta_G = 0$ for all such G there will be non-uniqueness whenever there are such G.

1.23.5 Fisher Information

Exponential Family

Differentiating twice the log likelihood (1.25) of an exponential family we get

$$l''(\theta) = -c''(\theta)$$

and minus this is observed Fisher information, and the expectation is expected Fisher information. Since the right-hand side is constant, observed and expected Fisher information are the same and are given by either side of (1.32b).

Unconditional

The Fisher information matrix for the saturated model unconditional canonical parameter φ is given by either side of (1.32b) with θ replaced by φ (because φ is the canonical parameter vector of the (unconditional, joint) aster model). Following Geyer et al. (2007), their equations (17) and (18) we choose to calculate Fisher information using probability theory (calculating covariances) rather than calculus (calculating derivatives).

This is done using the factorization and the iterated covariance theorem, which says for any random variables X, Y, and Z

$$\operatorname{cov}(X, Y) = E\{\operatorname{cov}(X, Y \mid Z)\} + \operatorname{cov}\{E(X \mid Z), E(Y \mid Z)\}$$

From this, for i and j in the same dependence group G

$$\begin{aligned} \operatorname{cov}_\varphi(Y_i, Y_j) &= E_\varphi\{\operatorname{cov}_\varphi(Y_i, Y_j \mid Y_{q(G)})\} \\ &\quad + \operatorname{cov}_\varphi\{E_\varphi(Y_i \mid Y_{q(G)}), E_\varphi(Y_j \mid Y_{q(G)})\} \\ &= E_\varphi\{Y_{q(G)}\gamma_{ij}\} + \operatorname{cov}_\varphi\{Y_{q(G)}\xi_i, Y_{q(G)}\xi_j\} \\ &= \mu_{q(G)}\gamma_{ij} + \operatorname{var}_\varphi(Y_{q(G)}\xi_i\xi_j \end{aligned} \tag{1.57}$$

where μ and ξ are the unconditional and conditional mean value parameter vectors, respectively, corresponding to φ, and we introduce the notation

$$\gamma_{ij} = \frac{\partial^2 c_G(\theta_G)}{\partial\theta_i \partial\theta_j}$$

which can also be written

$$\gamma_{ij} = \operatorname{cov}_\varphi(Y_i, Y_j \mid Y_{q(G)} = 1) \tag{1.58}$$

provided the conditioning event does not have probability zero (so the latter equation makes no sense) but we always have

$$\operatorname{cov}_\varphi(Y_i, Y_j \mid Y_{q(G)}) = Y_{q(G)}\gamma_{ij}$$

by the predecessor-is-sample-size property (this was used in deriving (1.57)).

For future use we also define $\gamma_{ij} = 0$ when i and j are not in the same dependence group. This makes sense because if $i \in G$ and $j \in H$ and $G \neq H$

$$\frac{\partial^2 c_G(\theta)}{\partial\theta_i \partial\theta_j} = \frac{\partial^2 c_H(\theta)}{\partial\theta_i \partial\theta_j} = 0$$

And, for i and j not in the same dependence group, say $i \in G$ and $j \in H$ and $G < H$ in the order on $\mathcal{G}$ asserted to exist by Theorem 1.1,

$$\begin{aligned}
\operatorname{cov}_\varphi(Y_i, Y_j) &= E_\varphi\{\operatorname{cov}_\varphi(Y_i, Y_j \mid Y_{q(G)})\} \\
&\quad + \operatorname{cov}_\varphi\{E_\varphi(Y_i \mid Y_{q(G)}), E_\varphi(Y_j \mid Y_{q(G)})\} \\
&= \operatorname{cov}_\varphi\{E_\varphi(Y_i \mid Y_{q(G)}), E_\varphi(Y_j \mid Y_{q(G)})\} \\
&= \operatorname{cov}_\varphi\{\xi_i Y_{q(G)}, E_\varphi(Y_j \mid Y_{q(G)})\} \\
&= \xi_i \operatorname{cov}_\varphi\{Y_{q(G)}, E_\varphi(Y_j \mid Y_{q(G)})\} \\
&= \xi_i E_\varphi\{[Y_{q(G)} - \mu_{q(G)}] E_\varphi(Y_j \mid Y_{q(G)})\} \\
&= \xi_i E_\varphi\{[Y_{q(G)} - \mu_{q(G)}] Y_j\} \\
&= \xi_i \operatorname{cov}_\varphi(Y_{q(G)}, Y_j) \\
&= \xi_i \operatorname{cov}_\varphi(Y_{p(i)}, Y_j)
\end{aligned} \tag{1.59}$$

where the second equality is the Markov property (Theorem B.2 in Appendix B): Y_i and Y_j are conditionally independent given $Y_{q(G)}$.

Because covariances are symmetric in their arguments, we do not need to do the case $G > H$.

It should be clear that we can calculate the whole Fisher information matrix using (1.57) and (1.59) traversing the full aster graph in any order that visits predecessors before successors (the inverse of the order in Theorem 1.1).

For computational efficiency, we should note that if i and j are nodes of the full aster graph for different "individuals" in scare quotes (defined in Section 1.9 above), then they are unconditionally independent by Corollary B.3 in Appendix B, so

$$\operatorname{cov}(Y_i, Y_j) = 0$$

in this case. The Fisher information matrix for φ is block diagonal with the blocks being for "individuals" (in scare quotes).

Now we revert to calculus to find the Fisher information matrix for the unconditional submodel canonical parameter β. The map $\beta \mapsto \varphi$ given by (1.47) has derivative M, the model matrix, that is, if m_{ij} are components of M, then

$$\frac{\partial \varphi_i}{\partial \beta_j} = m_{ij}.$$

It follows from the chain rule that

$$c''_{\text{sub}}(\beta) = M^T c''(\varphi) M$$

or

$$I(\beta) = M^T I(\varphi) M$$

where we abuse notation using I for both Fisher information for β and Fisher information for φ.

Conditional

Differentiating twice the log likelihood (1.33) for the conditional canonical parameter vector θ of an aster model we get

$$\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} = - \sum_{G \in \mathcal{G}} y_{q(G)} \frac{\partial^2 c_G(\theta_G)}{\partial \theta_i \partial \theta_j} = - \sum_{G \in \mathcal{G}} y_{q(G)} \gamma_{ij}$$

Since this is random, observed and expected Fisher information are not the same. Also noting that $\gamma_{ij} = 0$ unless $i, j \in G$ we have the following.

- The observed Fisher information matrix for θ is block diagonal with nonzero blocks corresponding to the dependence groups. The i, j component is $y_{q(G)} \gamma_{ij}$ for G such that $i, j \in G$ and zero otherwise.
- The expected Fisher information matrix for θ is similarly block diagonal. The i, j component is $\mu_{q(G)} \gamma_{ij}$ for G such that $i, j \in G$ and zero otherwise.

If in either case we denote this matrix by $I(\theta)$, then the Fisher information matrix for β is

$$I(\beta) = M^T I(\theta) M$$

for the same reason as in the preceding section (and I denotes the observed Fisher information matrix in both instances or the expected Fisher information matrix in both instances).

Chapter 2

Completion

In this chapter we deal with what to do when maximum likelihood estimates do not exist in the exponential family or aster model we are initially given. There may, and usually do, exist maximum likelihood estimates in the *completion* of the family. It is a bit unclear what we should call the statistical models studied in this chapter.

- Barndorff-Nielsen (1978, Sections 9.3 and 9.4) calls this concept *completion*.
- Brown (1986, Chapter 6) calls this concept an *aggregate exponential family* for reasons that will be explained presently.
- Geyer (1990, Chapters 2 and 4) calls this concept *closure*.
- Geyer (2009) calls this concept *Barndorff-Nielsen completion*.

Brown (personal communication) pointed out that the eponym chosen in Geyer (2009) was not quite correct, since Barndorff-Nielsen (1978) works under more restrictive regularity conditions than Brown (1986), and Brown (1986) works under more restrictive regularity conditions than Geyer (1990). The choice in Geyer (2009) follows Stigler's law of eponomy. At least in this case Barndorff-Nielsen had the concept first if not in the most generality. The reason why Geyer (1990) chose "closure" rather than "completion" is that when one works under the weakest regularity conditions, the topological space that is the statistical model being "completed" is not metrizable, hence "complete" (every Cauchy sequence converges) doesn't make any sense (the definition of Cauchy sequence requires a metric). Thus we have only the more general topological concept of closure. We won't fuss about any of this and will continue use Barndorff-Nielsen completion or just completion.

2.1 Binomial Example

For this simplest example of the phenomenon of interest, we consider the binomial distribution. We know from the discussion in Section 1.23.2 above that the MLE does not exist when the observed value of the canonical statistic, which for the binomial distribution is the number of successes, is an extreme value, either as small as it can be or as large as it can be, in this case either 0 or n, where n is the sample size.

Usually, we think the MLE for the usual parameter p, the success probability, does exist for all data and is $\hat{p} = x/n$. But when $x = 0$ or $x = n$, so $\hat{p}$ is zero or one, the MLE for the canonical parameter $\theta = \operatorname{logit}(p)$ does not exist because the domain of the logit function is the open interval $(0, 1)$ and does not include the endpoints. Since

$$\lim_{p \downarrow 0} \operatorname{logit}(p) = -\infty$$
$$\lim_{p \uparrow 0} \operatorname{logit}(p) = \infty$$

we could try to identify these endpoints with infinite values of the canonical parameter, but that is not the way exponential family theory works, and, as we shall see, it does not generalize to multiparameter problems.

So instead of trying to complete the parameter space, we try to complete the family of distributions. These distributions have PMF

$$f_p(x) = \binom{n}{x} p^x (1 - p)^{n - x}$$

and we have

$$\lim_{p \downarrow 0} f_p(x) = \begin{cases} 1, & x = 0 \\ 0, & x > 0 \end{cases}$$
$$\lim_{p \uparrow 1} f_p(x) = \begin{cases} 0, & x < n \\ 1, & x = n \end{cases}$$

so the completion contains the original exponential family we were given plus two new distributions, the degenerate distribution concentrated at zero and the degenerate distribution concentrated at n. And these new distributions are what are usually thought of as the binomial distributions for $p = 0$ and $p = 1$ (when $p = 0$ no successes are possible so $x = 0$ almost surely; when $p = 1$ no failures are possible so $x = n$ almost surely).

2.2 General Exponential Families

2.2.1 Support and Support Function

Let C denote the *closed convex support* of the exponential family under discussion. This is the smallest closed convex set that contains the canonical statistic vector with probability one. Hence it is a closed convex subset of the vector space where the canonical statistic takes values.

The closed convex support always exists because the intersection of closed sets is closed and the intersection of convex sets is convex and because finite-dimensional vector spaces are second countable. Define C to be the intersection of all closed convex sets that contain the canonical statistic vector Y with probability one under some distribution in the family, and hence for all distributions in the family (because all have the same support). Then event $Y \notin C$ is the union of a countable family of open sets having probability zero, hence $Y \in C$ almost surely.

Let σ_C denote the *support function* of C, defined by

$$\sigma_C(\delta) = \sup_{y \in C} \langle y, \delta \rangle \tag{2.1}$$

(Rockafellar and Wets, 1998, Section 8.E). The term "support" here is unfortunate in that it is unrelated to the term "support" in C being a support of the canonical statistic vector of the exponential family. But both terms are well established ("closed convex support" in exponential family theory and "support function" in convex analysis).

2.2.2 Probability Mass-Density Functions

If (1.25) is the log likelihood of an exponential family, the the PMDF of that family must be the exponential of the log likelihood. In order that we do not get extra terms that do not appear in the log likelihood and in order to get the right support of the family, we take the measure with respect to which we calculate densities to be a measure in the family, say the measure corresponding to canonical parameter vector ψ. Then the PMDF are

$$f_\theta(\omega) = e^{\langle Y(\omega), \theta - \psi \rangle - c(\theta) + c(\psi)} \tag{2.2}$$

where ω is the complete data (remember that Y is a statistic, not necessarily the complete data) (Geyer, 2009, Equation (4)).

2.2.3 Straight Line Limits

Theorem 2.1. *For a full exponential family having log likelihood* (1.25)*, densities* (2.2)*, canonical statistic vector Y, full canonical parameter space Θ, and closed convex support C, suppose δ is a direction in the vector space where the canonical parameter takes values and*

$$H_\delta = \{\, y \in \mathbb{R}^J : \langle y, \delta \rangle = \sigma_C(\delta) \,\}, \tag{2.3}$$

then for all $\theta \in \Theta$

$$\lim_{\theta + s\delta} f_{\theta + s\delta}(\omega) = \begin{cases} 0, & \langle Y(\omega), \delta \rangle < \sigma_C(\delta) \\ f_\theta(\omega) / \Pr_\theta(Y \in H_\delta), & \langle Y(\omega), \delta \rangle = \sigma_C(\delta) \\ \infty, & \langle Y(\omega), \delta \rangle > \sigma_C(\delta) \end{cases} \tag{2.4}$$

where the middle term is defined to be ∞ in case of divide by zero. If δ is not a direction of constancy and $\Pr_\theta(Y \in H_\delta) > 0$, then the function $s \mapsto \Pr_{\theta + s\delta}(Y \in H_\delta)$ is continuous, strictly increasing, and converges to one as $s \to \infty$.

In the two cases ruled out by the precondition of the last sentence the function $s \mapsto \Pr_{\theta + s\delta}(Y \in H_\delta)$ is a constant function. If δ is a direction of constancy, then $\Pr_\theta(Y \in H_\delta) = 1$ for all θ. If $\Pr_\theta(Y \in H_\delta) = 0$ for some θ, then $\Pr_\theta(Y \in H_\delta) = 0$ for all θ.

Proof. This is a complication of Theorem 6 in Geyer (2009) that is essentially Theorem 2.3 in Geyer (1990). However the proof of that Theorem 2.3 contains some errors, so a corrected proof is given in the appendix of Geyer (2009). Then the case $\Pr_\theta(Y \in H_\delta) > 0$ is Theorem 2.6 in Geyer (1990), and the case $\Pr_\theta(Y \in H_\delta) = 0$ follows from Theorem 2.2 in Geyer (1990). The last sentence of the theorem statement is in Corollary 5 and Theorem 6 in Geyer (2009). □

In case $\Pr_\theta(Y \in H_\delta) > 0$, we note three things about the limit in (2.4).

- It is a probability distribution because the set where it is infinite has measure zero under the dominating measure $\Pr_\psi$
- It is a conditional distribution of the original family, the conditional distribution of Y given the event $Y \in H_\delta$ for the parameter vector θ.
- It is a limit distribution of the original family, the limit of the distributions for parameter vectors $\theta + s\delta$ as $s \to \infty$. By Scheffé's lemma, convergence of PMDF implies convergence in total variation of the corresponding probability measures.

Note that the distribution under discussion is both a limit distribution and a conditional distribution. Thinking of it as just one or the other is missing something. It is both.

In case $\Pr_\theta(Y \in H_\delta) = 0$, we note one thing about the limit in (2.4).

- It is the zero measure because it is zero on the support of the dominating measure $\Pr_\psi$.

2.2.4 Limiting Conditional Models

We note one thing about the set of all limits in (2.4) in the case when $\Pr_\theta(Y \in H_\delta) > 0$.

- They form an exponential family of distributions. The log likelihood is
$$l_\delta(\theta) = \langle y, \theta \rangle - c(\theta) - \log \Pr_\theta(Y \in H_\delta) \tag{2.5}$$
and this is clearly an exponential family with
  - canonical statistic vector y,
  - canonical parameter vector θ, and
  - cumulant function given by
$$c_\delta(\theta) = c(\theta) + \log \Pr_\theta(Y \in H_\delta) \tag{2.6}$$

Geyer (2009) calls this family the *limiting conditional model* (LCM). Of course, there are many LCM, one in each direction, but as we shall presently see, there is usually only one LCM of interest in any particular data analysis.

Theorem 2.2. *Equation* (2.6) *gives the correct limit of the cumulant function to make* (2.5) *equal to the limit of* (1.25) *when limits are taken as in Theorem 2.1 in the case* $\langle y, \theta \rangle = \sigma_C(\delta)$.

Proof. In symbols, the assertion of the theorem is

$$\lim_{s \to \infty} l(\theta + s\delta) = l_\delta(\theta)$$

And

$$\begin{aligned}
\lim_{s \to \infty} l(\theta + s\delta) &= \lim_{s \to \infty} \bigl[\langle y, \theta + s\delta \rangle - c(\theta + s\delta) \bigr] \\
&= \langle y, \theta \rangle + \lim_{s \to \infty} \bigl[s \langle y, \delta \rangle - c(\theta + s\delta) \bigr] \\
&= \langle y, \theta \rangle + \lim_{s \to \infty} \bigl[s \sigma_C(\delta) - c(\theta + s\delta) \bigr] \\
&= \langle y, \theta \rangle - c(\theta) - \log \Pr_\theta(Y \in H_\delta) \\
&= \langle y, \theta \rangle - c_\delta(\theta)
\end{aligned}$$

where the fourth equality is Theorem 2.2 in Geyer (1990). □

We know, of course, that cumulant functions can be redefined by adding an arbitrary constant (the $c(\psi)$ in (1.26)). As mentioned in Section 1.15.1 above, we could even redefine the cumulant function by adding an arbitrary affine function if we were to accept a different choice of canonical statistic. But things would get very confusing if we made different arbitrary choices for the original exponential family and its limiting conditional models. Hence, however the cumulant function of the original exponential family was chosen, we will always use (2.6) to define cumulant functions for limiting conditional models.

2.2.5 Aggregate Exponential Family

Denote the LCM in the direction δ by $\mathcal{P}_\delta$. When $\Pr_\theta(Y \in H_\delta) = 0$ we say $\mathcal{P}_\delta$ is empty (there are no limit probability distributions, and we do not want to include the zero measure in our completion, at least not yet). Taking limits when $\delta = 0$ does nothing (because (2.1) says $\sigma_C(0) = 0$ for any C and this gives $H_\delta = \mathbb{R}^J$ in (2.3)). So $\mathcal{P}_0$ is the exponential family we started with, which we call the original model (OM) for short.

Then

$$\mathcal{P} = \bigcup_{\delta \in \mathbb{R}^J} \mathcal{P}_\delta$$

is a union of exponential families that contains all straight-line limits.

Under certain regularity conditions used by Barndorff-Nielsen (1978), Brown (1986), and Geyer (2009) this union is the completion. We do not get anything more by taking further straight-line limits in $\mathcal{P}_\delta$ for the various δ.

But in general (Geyer, 1990, Chapters 2 and 4) we may need to take further straight-line limits or general (not straight line) limits to arrive at the completion.

Anyway, one can see why Brown (1986) gave this idea the name aggregate exponential family. It is a union (or aggregate) of exponential families.

2.2.6 Support and Directions of Recession and Constancy

Theorem 2.3. *A vector δ is a direction of recession of the log likelihood of a full exponential family with closed convex support C and observed value of the canonical statistic vector y if and only if $\langle y, \delta \rangle \geq \sigma_C(\delta)$. If δ and $-\delta$ are both directions of recession, then δ is a direction of constancy. Conversely,*

if δ is a direction of constancy and $y \in C$, then δ and $-\delta$ are both directions of recession.

The condition $y \in C$ is measure-theoretic nonsense. We have to say $y \in C$ to be measure-theoretically correct, but if your data fail to satisfy $y \in C$, then something is wrong with your data.

Proof. The first sentence is Corollary 2.4.1 in Geyer (1990). If δ and $-\delta$ are both directions of recession then

$$\langle y, \delta \rangle \geq \sigma_C(\delta) = \sup_{x \in C} \langle x, \delta \rangle$$

and

$$\begin{aligned} -\langle y, \delta \rangle &= \langle y, -\delta \rangle \\ &\geq \sigma_C(-\delta) \\ &= \sup_{x \in C} \langle x, -\delta \rangle \\ &= -\inf_{x \in C} \langle x, \delta \rangle \end{aligned}$$

or

$$\langle y, \delta \rangle \leq \inf_{x \in C} \langle x, \delta \rangle$$

so

$$\langle y, \delta \rangle \leq \inf_{x \in C} \langle x, \delta \rangle \leq \sup_{x \in C} \langle x, \delta \rangle \leq \langle y, \delta \rangle \tag{2.7}$$

hence

$$\langle x, \delta \rangle = \langle y, \delta \rangle, \qquad x \in C \tag{2.8}$$

hence $\langle Y, \delta \rangle = \langle y, \delta \rangle$ almost surely, and δ is a direction of constancy.

Conversely, if δ is a direction of constancy, then $\langle Y, \delta \rangle$ is constant almost surely. And if $y \in C$, that constant must be $\langle y, \delta \rangle$. Hence (2.8) holds. Hence (2.7) holds. And we have already seen that (2.7) is equivalent to both δ and $-\delta$ being directions of recession. □

2.2.7 Curved Line Limits

Chapter 4 of Geyer (1990) covers completely general limits of sequences of distributions in an exponential family of distributions, these limits being in the sense of convergence of probability mass-density functions. Although the section title says "curved line limits," these limits are just limits of

sequences. We only get a line by connecting the dots, and that line does not have to be a smooth curve.

Theorems 4.1 through 4.5 in Geyer (1990) show that taking general limits gives no more limits that correspond to probability distributions than taking iterated straight line limits. General limits can produce limits that are subprobability distributions (Geyer, 1990, Examples 4.2 through 4.4), but these can never be maximum likelihood estimates for a full family, because iterated straight line limits produce the corresponding probability distribution, which must have higher likelihood. This shows that we do not need to consider curved line limits, so long as we limit our attention to full families.

2.3 Unconditional Aster Models

Unconditional aster models are regular full exponential families. Thus the theory of the preceding section applies to them.

Theorem 2.4. *Suppose $\{j\}$ is a univariate dependence group in an aster graph, and the one-parameter exponential family of distributions for the arrow $y_{p(j)} \longrightarrow y_j$ has closed convex support that is an interval with endpoints a_j and b_j (either of which may be infinite and which satisfy $a_j \le b_j$ with equality possible, in which case this distribution is concentrated at one point). Let J be the set of non-initial nodes of the aster graph, and let Y denote the response vector and y its observed value.*

If $y_j = a_j y_{p(j)}$, then the vector η having index set J and coordinates

$$\eta_i = \begin{cases} -1, & i = j \\ a_j & i = p(j) \\ 0, & \text{otherwise} \end{cases} \tag{2.9}$$

is a direction of recession of the saturated aster model.

Taking the limit in the direction of recession (2.9) *gives the LCM that is the same as the OM except the arrow $y_{p(j)} \longrightarrow y_j$ has the degenerate family of distributions concentrated at a_j. This LCM is the OM conditioned on the event $Y_j = a_j Y_{p(j)}$.*

If $y_j = b_j y_{p(j)}$, then the vector η having index set J and coordinates

$$\eta_i = \begin{cases} 1, & i = j \\ -b_j & i = p(j) \\ 0, & \text{otherwise} \end{cases} \tag{2.10}$$

is a direction of recession of the saturated aster model.

Taking the limit in the direction of recession (2.10) *gives the LCM that is the same as the OM except the arrow* $y_{p(j)} \longrightarrow y_j$ *has the degenerate family of distributions concentrated at* b_j*. This LCM is the OM conditioned on the event* $Y_j = b_j Y_{p(j)}$*.*

In case $a_j = b_j$ *the vectors* (2.9) *and* (2.10) *are both directions of recession, and are negatives of each other, hence are both directions of constancy. But in this case we only need one direction of constancy since one is a scalar multiple of the other.*

In case $y_{p(j)} = 0$ *and* $-\infty < a_j < b_j < \infty$ *the vectors* (2.9) *and* (2.10) *are still both directions of recession, but are not directions of constancy unless* $Y_{p(j)} = 0$ *almost surely.*

Proof. We have a direction of recession if and only if $\langle Y - y, \eta \rangle \le 0$ almost surely. From the definition of a_j and b_j we know $a_j Y_{p(j)} \le Y_j \le b_j Y_{p(j)}$ almost surely.

In case $y_j = a_j y_{p(j)}$ and η is given by (2.9) we have two cases. If $p(j)$ is noninitial, then

$$\langle Y - y, \eta \rangle = -(Y_j - y_j) + a_j (Y_{p(j)} - y_{p(j)}) = -Y_j + a_j Y_{p(j)}$$

and this is indeed less than or equal to zero almost surely by definition of a_j. If $p(j)$ is initial, then

$$\langle Y - y, \eta \rangle = -(Y_j - y_j) = -Y_j + a_j y_{p(j)} = -Y_j + a_j Y_{p(j)}$$

the last equality being that $Y_{p(j)}$ is a constant random variable so $Y_{p(j)} = y_{p(j)}$ almost surely, and this is indeed less than or equal to zero almost surely by definition of a_j. Thus in either case we have (2.9) is a direction of recession and

$$\langle Y - y, \eta \rangle = -Y_j + a_j Y_{p(j)} \tag{2.11}$$

Taking limits in the direction (2.9) arrives at the limiting conditional model that conditions on (2.11) being equal to zero, that is, on the event $Y_j = a_j Y_{p(j)}$. By the predecessor-is-sample-size principle this is the same thing as saying the arrow $y_{p(j)} \longrightarrow y_j$ has the degenerate family of distributions concentrated at a_j.

The proofs of the assertions about (2.10) are similar.

That the conditions for (2.9) and (2.10) both hold when $a_j = b_j$ and that one is then the negative of the other is obvious. That η and $-\eta$ both being directions of recession implies either is a direction of constancy is Geyer (2009) Theorem 3 part (e) and Theorem 1 part (g).

In case $y_{p(j)} = 0$ and $-\infty < a_j < b_j < \infty$ we also have $y_j = 0$ so $y_j = a_j y_{p(j)} = b_j y_{p(j)}$ holds trivially. Thus we have already proved that both are directions of recession. In order for (2.9) to be a direction of constancy we need $Y_j = a_j Y_{p(j)}$ to hold almost surely, but this is false unless $Y_{p(j)} = 0$ almost surely. Similarly for (2.10). □

As we see in the proof, there are two cases. When $p(j)$ is noninitial the arrow $y_{p(j)} \longrightarrow y_j$ represents a conditional distribution and η has two nonzero components unless $a_j = 0$ and η is given by (2.9) or $b_j = 0$ and η is given by (2.10). When $p(j)$ is initial the arrow $y_{p(j)} \longrightarrow j$ represents, in effect, a marginal distribution and η has one nonzero component. But the formulas (2.9) and (2.10) work in either case because the middle case does not occur when $p(j) \notin J$.

The case $a_j = b_j$ cannot occur in aster models allowed by R package `aster`. But once we start taking limits, then they can. So they are allowed by R package `aster2`.

Degenerate distributions concentrated at a_j or b_j are further discussed in the appropriate section of Appendix D (the details depend on the family the degenerate distribution is derived from). R package `aster2` implements them.

In the case considered last in the theorem where $y_{p(j)} = 0$ and (2.9) and (2.10) are both directions of recession and point in different directions, any nonnegative combination (linear combination with nonnegative coefficients) of these two vectors is another direction of recession (any nonnegative combimation of directions of recession is another direction of recession, Geyer, 2009, Theorem 3). For example, the vector η whose only nonzero component is $\eta_{p(j)} = a_j - b_j$ is a direction of recession.

Theorem 2.5. *Suppose G is a multinomial dependence group in an aster graph. Let J be the set of non-initial nodes of the aster graph, and let Y denote the response vector and y its observed value.*

If $j \in G$ and $y_j = 0$, then the vector η having index set J and coordinates

$$\eta_i = \begin{cases} -1, & i = j \\ 0, & \text{otherwise} \end{cases} \tag{2.12}$$

is a direction of recession of the saturated aster model.

This vector is not a direction of constancy of the saturated aster model unless $Y_{q(G)} = 0$ almost surely.

The vector η having index set J and coordinates

$$\eta_i = \begin{cases} -1, & i \in G \\ +1, & i = q(G) \\ 0, & \text{otherwise} \end{cases} \tag{2.13}$$

is a direction of constancy of the saturated aster model.

Taking the limit in the direction of recession (2.12) *gives the LCM that is the same as the OM except the arrow $y_{p(j)} \longrightarrow y_j$ has the degenerate family of distributions concentrated at zero, and the conditional distribution for dependence group G becomes a partially degenerate multinomial distribution that has $Y_j = 0$ almost surely.*

Proof. For (2.12) we need to show that $\langle Y - y, \eta \rangle \le 0$ almost surely. This is obvious.

$$\langle Y - y, \eta \rangle = -(Y_j - y_j) = -Y_j \tag{2.14}$$

and this is indeed less than or equal to zero almost surely by definition of the multinomial distribution.

In order for (2.12) to be a direction of constancy we need (2.14) to be zero almost surely. But this is false unless $Y_{q(G)} = 0$ almost surely.

For η given by (2.13) to be a direction of constancy we need to show that $\langle Y - y, \eta \rangle = 0$ almost surely, where Y and y are as above. This too is obvious.

$$\langle Y - y, \eta \rangle = (Y_{q(G)} - y_{q(G)}) - \sum_{j \in G} (Y_j - y_j)$$

and this is indeed equal to zero almost surely, by definition of the multinomial distribution.

By Theorem 2.1, taking limits in the direction (2.12) results in an LCM that conditions the OM on the event $Y_j = 0$ almost surely, and this corresponds to the arrow $y_{p(j)} \longrightarrow y_j$ having the degenerate family concentrated at zero almost surely. And this makes the multinomial distribution of Y_G given $Y_{q(G)}$ partially degenerate. □

As mentioned after the preceeding theorem, any nonnegative combination of directions of recession is another direction of recession. This includes the direction of constancy (any direction of constancy is a direction of recession, and so is the negative of any direction of constancy). Hence a vector

η is a direction of recession described by this theorem if and only if

$$\eta_j < \max_{i \in G} \eta_i \quad \text{implies} \quad y_j = 0$$
$$q(G) \text{ noninitial} \quad \text{implies} \quad \max_{i \in G} \eta_i = -\eta_{q(G)}$$

In case all but one of the components of a multinomial Y_G are zero, we can apply the theorem repeatedly to get a completely degenerate multinomial family. If $j \in G$ and $y_k = 0$ for $k \in G \setminus \{j\}$, then repeated limits give us the degenerate multinomial family that conditions on $Y_k = 0$ for $k \in G \setminus \{j\}$, but then we must also have $Y_j = Y_{q(G)}$ almost surely by definition of the multinomial distribution.

These partially degenerate multinomial distributions are further discussed in Section D.9 in the appendix.

In case $y_{q(G)} = 0$ but $Y_{q(G)} = 0$ does not hold almost surely, the theorem says that every η whose only nonzero component is $\eta_j = -1$ is a direction of recession. Hence any nonnegative combination of these is a direction of recession. Hence considering also (2.13) gives a vector η having nonzero components η_G in case $q(G)$ is initial and $\eta_{G \cup \{q(G)\}}$ in case $q(G)$ is noninitial, is a direction of recession if and only if

$$q(G) \text{ noninitial} \quad \text{implies} \quad \max_{i \in G} \eta_i = -\eta_{q(G)}$$

and this direction of recession is a direction of constancy if and only if

$$\eta_i = \eta_j, \qquad i, j \in G$$

Theorem 2.6. *Suppose $G = \{j, k\}$ is a normal-location-scale dependence group in an aster graph. Let J be the set of non-initial nodes of the aster graph, and let Y denote the response vector and y its observed value.*

If $y_{q(G)} = 1$, then the vector η having index set J and coordinates

$$\eta_i = \begin{cases} 2y_j, & i = j \\ -1, & i = k \\ 0, & \text{otherwise} \end{cases} \tag{2.15}$$

is a direction of recession of the saturated aster model.

But this direction of recession does not produce a limiting conditional model because it corresponds to the case $\Pr_\theta(Y \in H_\delta) = 0$ in in Theorem 2.1.

If $y_{q(G)} \geq 2$, then almost surely there are no directions of recession for this family.

If $y_{q(G)} = 0$, then every vector of the form (2.15) *is a direction of recession (for all real numbers y_j). These directions of recession are not directions of constancy unless $Y_{q(G)} = 0$ almost surely.*

Proof. The closed curve consisting of points y_G such that $y_k = y_j^2$ supports the conditional distribution of y_G given $y_{q(G)} = 1$. The function $f : y_j \mapsto y_j^2$ is a convex function. By the gradient inequality (Rockafellar and Wets, 1998, Theorem 2.13 (b))

$$f(Y_j) - f(y_j) \geq f'(y_j)(Y_j - y_j)$$

but this can also be written

$$Y_k - y_k \geq 2y_j(Y_j - y_j)$$

because $f(y_j) = y_k$ and $f(Y_j) = Y_k$ and $f'(y_j) = 2y_j$. And it can also be written

$$-\eta_k(Y_k - y_k) \geq \eta_j(Y_j - y_j)$$

because of the definition of η in the theorem statement. And this says $\langle Y - y, \eta \rangle \leq 0$ so η is a direction of recession.

Since f is strictly convex (Rockafellar and Wets, 1998, Theorem 2.13 (a′)), we have strict inequality in the gradient inequality (Rockafellar and Wets, 1998, Theorem 2.13 (b′)) when $Y_j \neq y_j$. Hence we have $\langle Y - y, \eta \rangle < 0$ almost surely. But the latter is equivalent to $\Pr_\theta(Y \in H_\delta) = 0$ in in Theorem 2.1.

The (closed) convex support of the family for sample size one is the set

$$C = \{\, y_G : y_k \geq y_j^2 \,\}$$

The (closed) convex support of the family for sample size n is is the n-fold Minkowski sum of this set, which is nC (Rockafellar and Wets, 1998, Proposition 2.23). Because the distributions in this family are continuous, the interior of nC actually supports the family for $n \geq 2$. Hence (almost surely) it is not possible to observe data on the boundary of the convex support for $n \geq 2$.

In case $y_{q(G)} = 0$, the convex support of the conditional distribution of Y_G given $Y_{q(G)}$ is $0 \cdot C = \{0\}$. Now we have to consider the three-dimensional set of all possible vectors $Y_{\{j,k,l\}}$ with $l = q(G)$. Since limits of sequences of normal vectors are again normal vectors (Rockafellar and Wets, 1998, Proposition 6.6 and Theorem 6.9), normal vectors at the point $0 = (0, 0, 0)$ are those of the form (2.15) and any nonnegative combinations of such.

In order for such a vector η to be a direction of constancy, we must have $\eta_j Y_j + \eta_k Y_k = 0$ almost surely. But this is false unless $Y_{q(G)} = 0$ almost surely. □

If $y_{q(G)} \geq 1$, then the theorem gives at most one direction of recession, and it is not a direction of constancy. If $y_{q(G)} = 0$, then the theorem gives an infinite number of directions of recession pointing in different directions. Our intention with these theorems is to use them to discover generic directions of recession of (not saturated) unconditional aster models using repeated linear programming. But we cannot put an infinite number of vectors into a linear program.

Moreover, our use of this theorem has to be fundamentally different from our use of Theorems 2.4 and 2.5. From the latter we discover DOR that lead to LCM in which we find MLE. From the former we discover DOR that do not lead to LCM, and the MLE does not exist, and we have to rudely inform users that their models are no good: you cannot estimate the variance of a normal distribution from one observation.

Users can avoid these kind of error messages by assuming homoscedastic errors (just like in linear models). The way this is done in aster models is to have the variance node of each normal-location-scale family have the same parameter. And the way to do that is to have `+ foo` in the formula, where `foo` is the indicator vector of the variance nodes of all normal-location-scale dependence groups (the k in the theorem) and no other appearance of `foo` in the formula.

Users can have more complicated formulas in which variance differs among normal-location-scale dependence groups, but then it is the job of the computer to catch situations in which this leads to directions of recession described by the theorem.

Fortunately, we can separate these two kinds of problems. All normal dependence groups must be terminal (so not really "fortunately" because this follows from the predecessor-is-sample-size principle). Thus we can trim all of the normal dependence groups off of the aster graph and still have a possible aster model. Then we can apply our algorithm (still to be developed in what follows) to determine whether a GDOR exists and, if so, what the LCM is (for the model with normal dependence groups trimmed off).

Then we can put the normal dependence groups back, and look for more directions of recession. In this, the following theorem helps.

Theorem 2.7. *Suppose G is a dependence group in an aster graph, and $y_{q(G)} = 0$ where Y is the response vector and y its observed value, then*

the vector η whose only nonzero component is $\eta_{q(G)} = -1$ is a direction of recession of the saturated aster model.

Taking the limit in this direction gives the LCM that is the same as the OM except $Y_{q(G)} = 0$ almost surely, and this means the distributions of all successors, successors of successors, successors of successors of successors, etc. are not identifiable in this LCM.

This direction of recession is not a direction of constancy unless $Y_{q(G)} = 0$ almost surely.

Proof. We know $q(G)$ is not a terminal node. It is not an initial node either, because aster models are required to have nonzero data at initial nodes. We know from the predecessor-is-sample-size property that $Y_{q(G)} \geq 0$ is required. Since $y_{q(G)}$ is at the lower endpoint of the support of $Y_{q(G)}$, the vector described in the theorem statement is a direction of recession.

If we take limits in this direction, we get the OM conditioned on the event $Y_{q(G)} = 0$ almost surely, and this implies $Y_j = 0$ almost surely for all $j \prec q(G)$, where $\prec$ is the transitive closure of the successor relation. The cumulant function for this LCM does not depend on any of the variables φ_j for $j \preceq q(G)$. This can be seen by applying (1.26) to this model.

$$\begin{aligned} c(\varphi) &= c(\varphi^*) + \log\left\{ E_{\varphi^*}\left(e^{\langle Y, \varphi - \varphi^* \rangle} \right) \right\} \\ &= c(\varphi^*) + \log\left\{ E_{\varphi^*}\left(\prod_{j \in J} e^{Y_j(\varphi_j - \varphi_j^*)} \right) \right\} \\ &= c(\varphi^*) + \log\left\{ E_{\varphi^*}\left(\prod_{\substack{j \in J \\ j \not\preceq q(G)}} e^{Y_j(\varphi_j - \varphi_j^*)} \right) \right\} \end{aligned}$$

where φ varies and φ^* is fixed. And then (1.38) shows that the log likelihood for the LCM does not depend on any of these variables either.

This vector is a direction of constancy if and only if $Y_{q(G)} = 0$ almost surely. □

The vector this theorem finds to be a direction of recession is also found by theorems preceeding it, but the point of this theorem is that it applies to any aster model whatsoever, even those having families that have not been implemented yet. And the theorem also provides more information about LCM.

In particular, it tells us that for LCM we found by applying our (yet to be developed) GDOR algorithm to the model with normal dependence

groups trimmed off, any normal dependence groups having $Y_{q(G)} = 0$ in the LCM can still be ignored: their parameters will be non-identifiable in the LCM.

Theorem 2.8. *The set of all directions of recession is a closed convex cone. The set of all directions of constancy is a vector subspace. Every direction of constancy is also a direction of recession. A vector δ is a direction of constancy if and only if both δ and $-\delta$ are directions of recession.*

The set of all directions of recession of saturated aster models having families described by Theorems 2.4, 2.5, 2.6, and 2.7 is the smallest closed convex cone containing all of the directions of recession described by those theorems.

The set of all directions of constancy of such aster models is the smallest vector subspace containing all of the directions of constancy described by those theorems.

Proof. The assertions of the first paragraph are all in Theorems 1 and 3 of Geyer (2009) and the discussion surrounding them.

Sections 4.1 and 4.2 in Geyer (1990) characterize all possible limit distributions in an exponential family of distributions. Theorem 2.7 in Geyer (1990) says that all limit distributions can be obtained by taking iterated straight line limits. The limit of a product being the product of the limits, when we take a limit we get a limit in each term of the fundamental factorization of aster models (1.1). Thus when we have all possible limiting conditional models for each of the families for each of the dependence groups, we have also gotten all of the limits for the whole aster model.

Conversely, since the theorems mentioned describe all possible limits of distributions for dependence groups in the families described by those theorems, which includes all families currently implemented in R packages `aster` and `aster2`, we have discovered all possible limits. There are no other directions of recession. □

As the theorem statement only implicitly refers to but the the proof explicitly says, the theorem does not apply to aster models having multivariate dependence groups whose families have not been invented yet. We would need to add theorems about their directions of recession if we add them to aster models.

Now we need to figure out how to use these theorems when applied to general unconditional aster models (canonical affine submodels of the saturated aster model). The principle is simple. If M is the model matrix of

a canonical affine submodel, then δ is a direction of recession (resp. constancy) of that submodel if and only if $\eta = M\delta$ is a direction of recession (resp. constancy) of the saturated model.

So we revisit the theorems.

In Theorem 2.4 we have either (2.9) or (2.10) or both or neither is a direction of recession. If $a_j = b_j$, then we can take either to be a direction of constancy and ignore the other. If the predecessor is zero almost surely (this cannot happen unless the model we are considering is already an LCM) then any directions of recession are directions of constancy, but not otherwise.

In Theorem 2.5 when $y_{q(G)} > 0$ we have one direction of recession for each $j \in G$ such that $y_j = 0$ and we also have one direction of constancy for the whole dependence group. If the predecessor is zero almost surely (this cannot happen unless the model we are considering is already an LCM) then all directions of recession are directions of constancy, but not otherwise.

For now we ignore Theorem 2.6.

So that completes our list of directions of recession and constancy.

Theorem 2.9. *A vector δ is a direction of recession of an unconditional canonical affine submodel with normal dependence groups trimmed off and no arrows having degenerate families if and only if $\eta = M\delta$ has the form*

$$\eta = \sum_{j \in J_r} e_j \eta_j \tag{2.16}$$

where J_r is the index set for directions of recession of the saturated model discussed above, J_c is the subset of J_r indexing directions of constancy, and the e_j are real numbers satisfying $e_j \geq 0$ for $j \in J_r \setminus J_c$.

Proof. This just makes explicit what Theorem 2.8 already says. □

Consider the following linear program having variables e_j for $j \in J_r$ and δ_k for $k \in K$.

$$\begin{aligned} &\text{maximize} && \sum_{j \in J_r \setminus J_c} e_j \\ &\text{subject to} && 0 \leq e_j \leq 1, && j \in J_r \setminus J_c \\ & && M\delta = \sum_{j \in J_r} e_j \eta_j \end{aligned} \tag{2.17}$$

Theorem 2.10. *Linear program* (2.17) *always has a solution. Linear program* (2.17) *has optimal value zero if and only if there does not exist a*

direction of recession that is not a direction of constancy and the MLE exists in the originally given unconditional aster model with model matrix M. Otherwise, the optimal value is greater than or equal to one and the δ part of the solution is a direction of recession that is not a direction of constancy.

Proof. The feasible region is nonempty because it always contains the zero vector. Then solutions exist because the objective function is obviously bounded on the feasible region.

The optimal value is zero if and only if at the solution $e_j = 0$ for $j \in J_r \setminus J_c$, in which case the δ part of the solution is a direction of constancy of the submodel and $\eta = M\delta$ is a direction of constancy of the saturated model. The assertion about existence of MLE is then Theorem 4 in Geyer (2009).

If there exists any feasible point such that some e_j for $j \in J_r \setminus J_c$ is nonzero, then we can multiply all components of δ and all e_j by a strictly positive constant to make the largest e_j for $j \in J_r \setminus J_c$ equal to one, in which case the objective function is greater than or equal to one. Optimizing then only increases the objective function. The solution then is clearly a direction of recession that is not a direction of constancy by Theorem 2.9. □

We are not done yet because we haven't yet in the terminology of Geyer (2009) found a *generic* direction of recession (GDOR). That will be one that has the maximal number of nonzero e_j for for $j \in J_r \setminus J_c$. Since any nonnegative combination of directions of recession is another direction of recession, we can seek GDOR by modifying our linear program to find DOR with $e_j > 0$ that we haven't found so far.

Let J^* be any nonempty subset of $J_r \setminus J_c$, and consider the following linear program.

$$\begin{aligned} \text{maximize} \quad & \sum_{j \in J^*} e_j \\ \text{subject to} \quad & 0 \le e_j \le 1, && j \in J^* \\ & 0 \le e_j, && j \in (J_r \setminus J_c) \setminus J^* \\ & M\delta = \sum_{j \in J_r} e_j \eta_j \end{aligned} \tag{2.18}$$

Then we iterate (Algorithm 1, page 83).

Theorem 2.11. *Algorithm 1 always terminates, and γ is a generic direction of recession unless $\gamma = 0$, in which case the MLE exists in the originally given unconditional aster model.*

Algorithm 1 Find GDOR for Unconditional Aster Model

Set $J^* = J_r \setminus J_c$
Set $J^{**} = \varnothing$
Set $\gamma = 0$
repeat {
    Solve the linear program (2.18)
    if (linear program has no solution) **error**
    if (optimal value is zero) **break**
    Set δ to be the δ part of the solution of the linear program
    Set $\gamma = \gamma + \delta$
    Set e to be the e part of the solution of the linear program
    Set $J^{**} = J^{**} \cup \{ j \in J^* : e_j > 0 \}$
    Set $J^* = J^* \setminus J^{**}$
    if ($J^* = \varnothing$) **break**
}

Proof. The algorithm must terminate because J^* decreases in each iteration. So we terminate when $J^* = \varnothing$ if not before.

Since each δ found is a direction of recession that is not a direction of constancy, so is γ.

The termination condition of optimal value zero or $J^{**} = \varnothing$, proves that γ is such that $\eta = M\gamma$ has the most possible nonzero components η_j for $j \in J_r \setminus J_c$. Hence it is generic unless $\gamma = 0$ and the algorithm proves the MLE exists in the OM. □

From now on we only use the LCM corresponding to the GDOR found. This means every η_j for $j \in J^{**}$ found by the algorithm is a direction of constancy of this LCM. The other DOR remain the same.

Now we add back in all of the normal dependence groups. Normal-location arrows are covered by Theorem 2.4 with $a_j = -\infty$ and $b_j = +\infty$. They can never have directions of recession. So that leaves normal-location-scale dependence groups.

From Theorem 2.6 we know that those with $y_{q(G)} \geq 2$ have no directions of recession to add to our problem. From Theorem 2.7 we know that those with $y_{q(G)} = 0$ have non-identifiable parameters in the LCM. If $G = \{j, k\}$ then we add a vector whose only nonzero component is $\eta_j = 1$ to the list of directions of constancy and also a vector whose only nonzero component is $\eta_k = 1$ to the list of directions of constancy. Finally, from Theorem 2.6

we know that those with $y_{q(G)} = 1$ have exactly one direction of recession that is not a direction of constancy given by (2.15). There are no further directions of recession to add to our problem.

So we throw all of this back into linear program (2.17). It had better have optimal value zero. Otherwise users get rude error messages.

2.4 Conditional Aster Models

Saturated aster models are regular full exponential families. Unconditional canonical affine submodels of aster models are regular full exponential families. So the theory in this chapter up to now applies to them.

Conditional canonical affine submodels of aster models are not regular full exponential families. As smooth submodels of saturated aster models, they are what the jargon calls curved exponential families. But that does not tell us much about existence or non-existence of MLE. We know that all possible limits have to be limits of distributions in the saturated model (because it is a submodel of the saturated model). But when those limits are MLE is something for which there is no general theory for curved exponential families.

2.4.1 Associated Independence Models

A cheap trick, however, does crack the problem of conditional aster models. This is the notion of associated independence models (Section 1.23.4 above). For reference, we repeat (1.55) above

$$l(\theta) = \sum_{G \in \mathcal{G}} \left[\langle y_G, \theta_G \rangle - n_{q(G)} c_G(\theta_G) \right] \tag{2.19}$$

(so this equation now has two equation numbers, one here and one there).

The actual conditional model has (saturated model) log likelihood that is (2.19) with $n_{q(G)}$ replaced by $y_{q(G)}$. The associated independence model (AIM) has (saturated model) log likelihood that is (2.19) with $n_{q(G)}$ constant and y_j random.

As Section 1.23.4 above says, the AIM makes no sense when considered statistically, probabilistically, because it pretends that variables that are actually the same ($n_{q(G)}$ and $y_{q(G)}$) are different, and one is constant and the other random. But, as Section 1.23.4 above also says, the AIM makes perfect sense when considered numerically, algebraically when we are considering maximum likelihood estimation. Then $n_{q(G)}$ and $y_{q(G)}$ are just numbers,

fixed at their observed values, and if we use different notation for the same number in different parts of the expression, that is OK.

In Section 1.23.4 we used the AIM to reach the conclusion that the log likelihood of a conditional aster model is concave, something that is not generally true of curved exponential family models.

In this section, we will use the AIM to completely characterize existence and uniqueness of MLE for conditional aster models and directions of recession and constancy of (2.19) and for its canonical affine submodels (conditional aster models).

What puts the I (for independence) in AIM is that the AIM makes the Y_G for $G \in \mathcal{G}$ independent random vectors. This makes the AIM much easier to reason about than the actual conditional aster model.

So now we repeat the preceding section, *mutatis mutandis* reasoning about AIM rather than unconditional aster models.

Theorem 2.12. *Suppose $\{j\}$ is a univariate dependence group in an AIM, and the one-parameter exponential family of distributions for the arrow $n_{p(j)} \longrightarrow y_j$ has closed convex support that is an interval with endpoints a_j and b_j (either of which may be infinite and which satisfy $a_j \le b_j$ with equality possible, in which case this distribution is concentrated at one point). Let J be the set of non-initial nodes of the aster graph, let Y denote the response vector and y its observed value, and let n denote the vector of sample sizes whose components are n_j.*

If $n_{p(j)} = 0$ or $a_j = b_j$, then the vector η whose only nonzero component is $\eta_j = 1$ is a direction of constancy of (2.19).

If $n_{p(j)} > 0$ and $a_j < b_j$ and $y_j = a_j n_{p(j)}$, then the vector η whose only nonzero component is $\eta_j = -1$ is a direction of recession that is not a direction of constancy of (2.19).

Taking the limit in this direction of recession gives the LCM that is the same as the OM except the arrow $n_{p(j)} \longrightarrow y_j$ has the degenerate family of distributions concentrated at a_j. This LCM is the OM (of the AIM) conditioned on the event $Y_j = a_j n_{p(j)}$.

If $n_{p(j)} > 0$ and $a_j < b_j$ and $y_j = b_j n_{p(j)}$, then the vector η whose only nonzero component is $\eta_j = 1$ is a direction of recession that is not a direction of constancy of (2.19).

Taking the limit in this direction of recession gives the LCM that is the same as the OM except the arrow $n_{p(j)} \longrightarrow y_j$ has the degenerate family of distributions concentrated at b_j. This LCM is the OM (of the AIM) conditioned on the event $Y_j = b_j n_{p(j)}$.

Theorem 2.13. *Suppose G is a multinomial dependence group in an AIM. Let J be the set of non-initial nodes of the aster graph, let Y denote the response vector and y its observed value, and let n denote the vector of sample sizes whose components are n_j.*

If $n_{p(j)} = 0$ then the vector η whose only nonzero component is $\eta_j = 1$ is a direction of constancy of (2.19), and this is true for each $j \in G$.

If $n_{p(j)} > 0$ and $y_j = 0$, then the vector η whose only nonzero component is $\eta_j = -1$ is a direction of recession that is not a direction of constancy of (2.19), and this is true for each $j \in G$.

Taking the limit in any of these directions of recession, say the one having η_j nonzero, gives the LCM that is the same as the OM (of the AIM) except the arrow $n_{p(j)} \longrightarrow y_j$ has the degenerate family of distributions concentrated at zero.

The vector η having index set J and coordinates

$$\eta_i = \begin{cases} 1, & i \in G \\ 0, & \text{otherwise} \end{cases} \tag{2.20}$$

is a direction of constancy of of (2.19).

As in the discussion following Theorem 2.5 (which this theorem duplicates *mutatis mutandis*), we note that any nonnegative combination of directions of recession is another direction of recession. Hence the directions of recession described by this theorem are vectors η whose only nonzero components are in the subvector η_G and such a vector is a direction of recession of (2.19) if

$$\eta_j < \max_{i \in G} \eta_i \quad \text{implies} \quad y_j = 0$$

and such a vector is a direction of constancy of (2.19) if $n_{q(G)} = 0$ or if its nonzero components are all the same.

Theorem 2.14. *Suppose $G = \{j, k\}$ is a normal-location-scale dependence group in an AIM. Let J be the set of non-initial nodes of the aster graph, let Y denote the response vector and y its observed value, and let n denote the vector of sample sizes whose components are n_j.*

If $n_{q(G)} = 0$, then any vector η whose only nonzero components are η_j and η_k is a direction of constancy of of (2.19).

If $n_{q(G)} = 1$, then (2.15) is a direction of recession of (2.19) that is not a direction of constancy.

But this direction of recession does not produce a limiting conditional model because it corresponds to the case $\Pr_\theta(Y \in H_\delta) = 0$ in in Theorem 2.1.

If $n_{q(G)} \geq 2$, then, almost surely, there are no directions of recession for this family.

Theorem 2.15. *In addition to the general properties of directions of recession and constancy found in Theorem 2.8, the set of all directions of recession of AIM having families described by Theorems 2.12, 2.13, and 2.14 is the smallest closed convex cone containing all of the directions of recession described by those theorems.*

The set of all directions of constancy of such aster models is the smallest vector subspace containing all of the directions of constancy described by those theorems.

Theorem 2.16. *For an AIM with normal dependence groups trimmed off define*

$$J_{\text{up}} = \{\, j \in J : y_j = a_j n_{p(j)} \,\}$$
$$J_{\text{dn}} = \{\, j \in J : y_j = b_j n_{p(j)} \,\}$$

where these include multinomial dependence groups with the convention $a_j = 0$ and $b_j = \infty$ for them. Then a vector η is a direction of recession of this model if

$$\eta_j \leq 0, \qquad j \in J_{\text{dn}} \setminus J_{\text{up}}$$
$$\eta_j \geq 0, \qquad j \in J_{\text{up}} \setminus J_{\text{dn}}$$

Conversely, any direction of recession of this model satisfies these conditions if we modify it by subtracting $\max(\eta_G)$ from the elements of η_G for each multinomial dependence group G.

Chapter 3

Subsampling

3.1 Introduction

In aster models the ideal is to actually measure all components of fitness and make them nodes in the graphical model. Sometimes, however, it is just too much work to count all of some component of fitness, for example, all seeds produced by a plant.

For a part of an aster graph

$$\cdots \longrightarrow y_{\text{this}} \longrightarrow y_{\text{that}} \longrightarrow \cdots$$

suppose that y_{that} is typically too large to count (by available methods, in available time).

The conditional distribution of y_{that} given y_{this} is the sum of y_{this} independent and identically distributed (IID) random variables. This is the predecessor-is-sample-size property (Section 1.12 above). In short, y_{that} is a random "sample" from some "population" and the sample size is y_{this}, where the scare quotes are to indicate that "sample" and "population" don't refer to an actual sample and population but are just a way of discussing probability that is common in introductory statistics books, which take finite population sampling (actually taking a random sample from a known finite population) as an analogy for all applications of probability theory; any IID set of random variables is called a "sample" from a "population" whether or not that makes literal sense.

The obvious solution to our problem is to "subsample" the "sample." Take a random sample of the things y_{this} counts, and for that subsample count how many of whatever component of fitness y_{that} counts (this proposal will be generalized in Section 3.2.4 below). In this case, Shaw et al. (2008b,

p. E43) proposed to simply insert an extra arrow in the graph to represent the subsampling process. Taking a random subsample is a Bernoulli process (flip a "biased coin" to decide for each of the y_{this} things whether it goes in the subsample). So this is a Bernoulli arrow, but we mark it specially as a subsampling arrow

$$\cdots \longrightarrow y_{\text{this}} \xrightarrow{\text{samp}} y_{\text{this-sub}} \longrightarrow y_{\text{that-sub}} \longrightarrow \cdots .$$

Here y_{this} is the same variable it was before (the observed count for "this" fitness component), $y_{\text{this-sub}}$ is the subsample size (the subset of the y_{this} things that go in the subsample), and $y_{\text{that-sub}}$ is the observed count for "that" fitness component for the subsample). We no longer observe y_{that}, which is what we would have observed if we had not subsampled. Not having to count y_{that} was the whole point of the subsampling.

Shaw et al. (2008b) further proposed to treat aster models with subsampling just like any other aster model. This suggestion was backed up by Section 8 of a supporting technical report (Shaw et al., 2008a). That technical report, however, notes this suggestion is not quite the right thing. It says

> The somewhat odd thing about this proposal is that the parameter p [the subsampling probability] is *known* and is a *conditional* mean value parameter, but we intend to use an *unconditional* aster model and treat the [corresponding] unconditional canonical parameter as *unknown* [emphasis in the original].

and devotes the rest of its Section 8 to a simulation study that shows that, although, not quite the right thing, it does well enough.

Stanton-Geddes, Shaw, and Tiffin (2012a, Appendix S1) show how to, in effect, remove the effect of subsampling when producing point estimates and confidence intervals for expected fitness, completing something left undone by Shaw et al. (2008b) and the accompanying technical report.

Here we give a new proposal that does the right thing with subsampling arrows and hence supersedes all earlier proposals. We have to realize that aster models with subsampling are no longer regular full exponential families, so they no longer satisfy the original rationale for aster models.

With subsampling there are two models we have to consider: the model with subsampling, which reflects the experiment actually done, and the same model with subsampling removed, which reflects biology of the organisms being studied. We give both of these models the same parameterization (subsampling arrows in the aster graph have no unknown parameters because the subsampling probabilities are known). Once this fundamental

realization is made, everything else about aster models with subsampling follows from well known likelihood theory.

3.2 Subsampling

3.2.1 Graphs With and Without Subsampling

We already have one two-way classification of aster graphs: the full aster graph and graphs for "individuals" in scare quotes (Section 1.9 above). Now we introduce a different two-way classification: with and without subsampling. Together these give us a four-way classification.

When using aster models with subsampling, we are also interested in the graph and the corresponding aster model if subsampling had not been done. The graph and model with subsampling represent the experiment actually done. The graph and model without subsampling represent the biology. We must refer to both in our discussion. For example, we use the graph and model with subsampling to estimate parameters, but we use the graph and model without subsampling to predict biological properties of organisms from these estimates.

Understanding the correspondence between the two graphs is helped by referring to the following picture, which is part of an aster graph with subsampling.

$$\cdots \longrightarrow y_i \longrightarrow y_j \xrightarrow{\text{samp}} y_k \longrightarrow y_m \longrightarrow \cdots \tag{3.1}$$

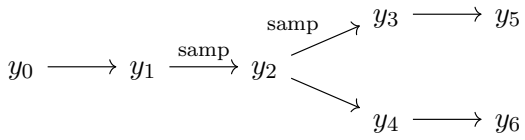
Only one subsampling arrow is shown (labeled "samp"). The other arrows are non-subsampling. Nodes at the head of subsampling arrows (here y_k) are called *subsampling nodes*. Thus y_k is the only subsampling node among the nodes shown. Here y_j is the count of some sort of thing (flowers, seeds, etc.) of actual individuals in the experiment, and y_k is the count of the same thing for a random subsample of those individuals who are carried forward to later stages of the experiment. Thus y_j and y_k are both measurements of the *same* component of fitness. The relationship between y_j and y_k is artificial (done by the experimenters) and has nothing to do with biology.

The corresponding graph without subsampling is formed by removing the subsampling arrow and subsampling node and pasting together the graph so no break is formed

$$\cdots \longrightarrow y_i \longrightarrow y_j \longrightarrow y_m \longrightarrow \cdots \tag{3.2}$$

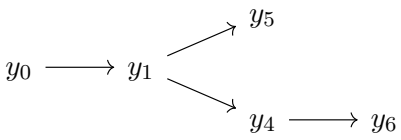
This graph corresponds to the experiment that would have been done if there were no subsampling.

Here is a more complicated example that illustrates that sometimes it may be necessary to have subsampling arrows following each other.



(3.3)

This is the graph for one "individual" assuming all "individuals" have isomorphic subgraphs. Here is the corresponding subgraph when we remove the subsampling arrows.



(3.4)

3.2.2 Notation for Graphs With and Without Subsampling

We are going to use mathematical notation that distinguishes analogous concepts for the full graphs with and without subsampling by decorating notation for the former with stars.

Sets of Nodes

The set of nodes of the full aster graph with subsampling is denoted N^*. The set of non-subsampling nodes in N^* is denoted N. This is the set of nodes of the full graph without subsampling.

The set of non-initial nodes in N^* is denoted J^*. The set of non-initial nodes in N is denoted J. ($J = N \cap J^*$.)

Families of Sets of Nodes

The set of dependence groups (Section 1.6 above) for the aster model with subsampling is denoted $\mathcal{G}^*$. The set of dependence groups for the aster model without subsampling is denoted $\mathcal{G}$.

Subsampling nodes are always dependence groups by themselves. No dependence group with more than one node can have any subsampling nodes. Thus the set of all subsampling nodes is $J^* \setminus J$.

Predecessor Functions

We already have two kinds of predecessor functions: set-to-index predecessor functions denoted q (Section 1.6 above) and index-to-index predecessor functions denoted p (Section 1.10 above). Now we have another two-way classification. We will have these functions with stars to denote with subsampling and without stars to denote without subsampling.

For example, in (3.1) we have $p^*(m) = k$ but $p(m) = j$. The latter agrees with (3.2), as it must.

For another example, in (3.3) we have

$$\begin{aligned} p^*(6) &= 4 \\ p^*(5) &= 3 \\ p^*(4) &= 2 \\ p^*(3) &= 2 \\ p^*(2) &= 1 \\ p^*(1) &= 0 \end{aligned}$$

but

$$\begin{aligned} p(6) &= 4 \\ p(5) &= 1 \\ p(4) &= 1 \\ p(1) &= 0 \end{aligned}$$

The latter agrees with (3.4), as it must.

We have not given examples of graphs with subsampling and dependence groups, but the set-to-index predecessor functions work the same way

$$\begin{aligned} p(j) &= q(G), \qquad j \in G \in \mathcal{G} \\ p^*(j) &= q^*(G), \qquad j \in G \in \mathcal{G}^* \end{aligned}$$

Partial Orders

In Section 1.11 above we introduced partial orders on the node set of the graph that tell us about predecessors, predecessors of predecessors, and so forth. Unlike the case with the other mathematical objects just discussed, we do not need starred and unstarred versions for these.

As in Section 1.11 above, let $\succ$ denote the transitive closure of the predecessor relation on N^*. Then the transitive closure of the predecessor relation

on N is just the same relation as on N^* but restricted to N, that is $j \succ k$ in N if and only $j \succ k$ when j and k are considered elements of N^*.

And similarly for $\succeq$, $\prec$, and $\preceq$.

Going from One to the Other

The user specifies the graph with subsampling. The computer should figure out the corresponding graph without subsampling (so no mistakes about that are made).

For this we use the notation p^k for k-fold composition of a function with itself from dynamical systems theory that is also explained in Section 1.11 above.

Suppose we are trying to determine $p(j)$ for $j \in J$. Define

$$m = \min\{\, k > 0 : (p^*)^k(j) \in N \,\}$$

Then $p(j) = (p^*)^m(j)$.

In words, we go back in the graph with subsampling looking at the predecessor, predecessor of predecessor, and so forth until we find one that is not a subsampling node, and that is the predecessor in the graph without subsampling.

3.2.3 Models With and Without Subsampling

Now that we have the relationship between the graph with and without subsampling, and hence the factorization (Section 1.6 above) with and without subsampling we need to consider statistical models that go with these factorizations. Models without subsampling are those already described (Chapter 1 above).

So we have to describe models with subsampling here. There are two key ideas.

- The conditional distributions for subsampling arrows are considered *known*. Thus they have *no* unknown parameter values to estimate.
- The conditional distributions for non-subsampling arrows should be the same for the models with and without subsampling. They should have the same statistical models parameterized in the same way (this statement is imprecise; it depends on which parameters we are talking about).

Data

First we introduce new notation for data (not used above) to distinguish the models with and without subsampling. Let y^* denote the response vector for the model with subsampling and y the response vector for the model without subsampling.

Factorization

Then the model with subsampling factorizes as

$$f^*(y^*) = \prod_{G \in \mathcal{G}^*} f_G^*(y_G^* \mid y_{q^*(G)}^*) \tag{3.5}$$

and the model without subsampling factorizes as

$$f(y) = \prod_{G \in \mathcal{G}} f_G(y_G \mid y_{q(G)}). \tag{3.6}$$

Now our first principle above says that $f_G^*(\cdot \mid \cdot)$ has no parameters when $G \notin \mathcal{G}$, and our second principle above says that

$$f_{G,\theta_G}^*(\cdot \mid \cdot) = f_{G,\theta_G}(\cdot \mid \cdot), \qquad G \in \mathcal{G}$$

that is, these are the same conditional distributions (the same functions of the variables on each side of the vertical bar) for the same parameter values.

Parameterization

This also makes clear what parameters we are talking about. The *conditional canonical parameter vector* θ (Section 1.18 above) is the same for the models with and without subsampling. But we also want the aster transform to be the same (also Section 1.18 above), so the *unconditional canonical parameter vector* φ will also be the same for the models with and without subsampling.

Mean Value Parameters

The *conditional mean value parameter vector* ξ will differ for the models with and without subsampling simply because the model with subsampling has more arrows. The components of ξ will be the same for the same arrows (Sections 1.13 and 1.20 above), that is, defining

$$\xi_j = E(y_j \mid y_{p(j)} = 1)$$
$$\xi_j^* = E(y_j^* \mid y_{p^*(j)}^* = 1)$$

we have

$$\xi_j = \xi_j^* = \frac{\partial c_G(\theta_G)}{\partial \theta_j}$$

whenever j is not a subsampling node and $j \in G \in \mathcal{G}$. But when j is a subsampling node there is no ξ_j since there is no node j in the model without subsampling, and ξ_j^* is not a function of the parameters of the model (θ and φ); it is not an unknown parameter but rather a known constant (this will be revisited in the next section).

Thus when we define

$$\mu = E(y)$$
$$\mu^* = E(y^*)$$

we will still have the relationship between μ and ξ discussed in Section 1.13.3 above. And we will have the analogous equations with stars for the relationship between μ^* and ξ^* but the relations between μ and μ^* will be very complicated and depend on the whole aster graph. But everything in this section depends on subsampling having the predecessor-is-sample-size property, which we drop in the next section. So nothing in this section holds for general aster models with subsampling.

3.2.4 Generalizing Our Notion of Subsampling

So far, we have been assuming the conditional distributions for the subsampling arrows obey the predecessor-is-sample-size principle. This means the conditional distribution for sample size one is Bernoulli, so subsampling arrows are Bernoulli arrows (but ones whose conditional canonical parameters are fixed and known rather than unknown parameters to be estimated). And it means the conditional distribution of y_j^* given $y_{p^*(j)}^*$ is binomial (when j is a subsampling node). Let ξ_j^* be the usual parameter for the binomial distribution, which is also the mean value parameter for the Bernoulli distribution. Then, defining $\mu^* = E(y^*)$, we have the analog of (1.7) with stars

$$\mu_j^* = \xi_j^* \mu_{p(j)}^*, \qquad j \in J^*. \tag{3.7}$$

and all of the consequences (1.7) found in Section 1.13.3 above, except with stars in the appropriate places.

But when we drop the predecessor-is-sample-size principle for subsampling arrows (3.7) no longer holds, and there is no longer any notion of ξ_j^* for such arrows analogous to non-subsampling arrows. We still do have

conditional mean values,

$$E\{y_j^* \mid y_{p^*(j)}^*\}$$

but these do not need to satisfy

$$E\{y_j^* \mid y_{p^*(j)}^*\} = \xi_j^* y_{p^*(j)}^*$$

for any constant ξ_j^* when j is a subsampling node that does not obey the predecessor-is-sample-size principle, but rather are arbitrary functions of $y_{p^*(j)}^*$.

This means that conditional mean value parameters for subsampling arrows make no sense for subsampling arrows that are not Bernoulli (do not obey the predecessor-is-sample-size principle). Since many biologists use forms of subsampling that are not Bernoulli (not simple random sample), we do not want to enforce the Bernoulli assumption.

All we assume about subsampling distributions is that they are known, having no unknown parameters to estimate. We make no other assumptions about them.

We do, of course, have an unconditional mean value parameter vector μ^* with components

$$\mu_j^* = E(Y_j^*), \qquad j \in J^* \tag{3.8}$$

but since the computer knows nothing about the subsampling distributions (they can be any distributions), the computer will be unable to compute them. If users want to use μ^* somehow, they will have to provide it themselves.

Thus ξ^* is undefined, in general, and μ^*, although defined, is no longer anything the computer can deal with.

Fortunately ξ^* and μ^* are unbiological. So users will mostly, perhaps always, only be interested in ξ and μ, which the computer can deal with.

3.2.5 Log Likelihood

When there is no subsampling, the saturated aster model log likelihood for θ is given by (1.33) which can be rewritten

$$l(\theta) = \sum_{j \in J} y_j \theta_j - \sum_{G \in \mathcal{G}} y_{q(G)} c_G(\theta_G). \tag{3.9}$$

Theorem 3.1. *The log likelihood for a model with subsampling is given by*

$$l(\theta) = \sum_{j \in J} y_j^* \theta_j - \sum_{G \in \mathcal{G}} y_{q^*(G)}^* c_G(\theta_G). \tag{3.10}$$

In (3.9) and (3.10) c_G is the cumulant function for the exponential family for dependence group G.

Proof. To be clear, we write down the log likelihood for all the data using the notation of Section 3.2.4 above

$$l(\theta^*) = \sum_{j \in J} y_j^* \theta_j - \sum_{G \in \mathcal{G}} y_{q^*(G)}^* c_G(\theta_G) + \sum_{j \in J^* \setminus J} \log f_j(y_j^* \mid y_{p^*(j)}^*)$$

and we are allowed to drop terms that do not contain unknown parameters from the log likelihood because this makes no difference to either frequentist or Bayesian inference. This gives (3.10). □

The cumulant function c_G satisfies

$$\begin{aligned} \frac{\partial c_G(\theta_G)}{\partial \theta_j} &= E_\theta(y_j \mid y_{q(G)} = 1) \\ &= E_\theta(y_j^* \mid y_{q^*(G)} = 1) \\ &= \xi_j \\ &= \xi_j^*, \qquad j \in G \in \mathcal{G}, \end{aligned} \tag{3.11}$$

and

$$\begin{aligned} \frac{\partial^2 c_G(\theta_G)}{\partial \theta_j \partial \theta_k} &= \operatorname{cov}_\theta(y_j, y_k \mid y_{q(G)} = 1) \\ &= \operatorname{cov}_\theta(y_j^*, y_k^* \mid y_{q^*(G)}^* = 1), \qquad j, k \in G \in \mathcal{G}, \end{aligned} \tag{3.12}$$

and these derivatives are zero if $j \notin G$ or (in the latter) $k \notin G$.

If the conditioning event in these equations has probability zero (which happens in actual aster models if the family in question is k-truncated with $k > 0$, Shaw et al., 2008b, have an example), then the conditional expectations are not well defined, but we still have

$$\frac{\partial c_G(\theta_G)}{\partial \theta_j} = \xi_j = \xi_j^* \tag{3.13}$$

with ξ_j being defined by the more long winded and careful definition given in Section 1.13.2 above when the conditioning event in (3.11) has probability zero.

The reason for the equality of starred and unstarred quantities is that we want the model without subsampling to be the same as the model with

subsampling when the subsampling arrows are removed as explained in Section 3.2.1 above and in this section. The distribution of y_G given $y_{q(G)} = n$ is the same as the distribution of y^*_G given $y^*_{q^*(G)} = n$.

There is a similar adjustment to be made for the conditional covariances above when their conditioning events have probability zero. We still have that y_G is the sum of $y_{q(G)}$ IID random vectors and $\partial^2 c_G(\theta_G)/\partial\theta_j\partial\theta_k$ is the unconditional covariance of the j and k components of one of those random vectors.

3.2.6 Aster Transform

We have the aster transform and inverse aster transform (Section 1.18) and these hold for models with subsampling because they are the same (we assume) as for models without subsampling. And these determine μ and ξ as discussed in the preceding section. And we are generally uninterested in μ^* and ξ^* as discussed in the preceding section. Thus we only have parameters without stars. We have aster graphs with stars and aster data with stars, but not parameters.

3.2.7 Canonical Affine Submodels

As is the situation without subsampling, we are interested in canonical affine submodels (Section 1.22 above) when we have subsampling. And we want them to be the same models with the same parameterizations with and without subsampling. Thus they are the same as in Section 1.22 above. Unconditional canonical affine submodels make φ an affine function of the submodel parameters β. Conditional canonical affine submodels make θ an affine function of the submodel parameters β.

Conditional

With subsampling, a conditional aster model still has a concave log likelihood for the reasons discussed in Section 1.23.4 above. The saturated model log likelihood (3.9) is a sum of linear and concave functions of θ. Therefore the submodel log likelihood is the composition of a concave function and an affine function, which is again a convex function. This means the MLE will be easily found by the computer (by any algorithm that always checks that it goes uphill on the likelihood in every iteration). It does not mean that conditional aster models have any other properties of regular full exponential families (but this is true with or without subsampling, as was discussed in Chapter 1).

Unconditional

With subsampling, an unconditional aster model is no longer an exponential family and does not have any full exponential family properties. It is, of course, still a curved exponential family, which gives it the usual asymptotics of maximum likelihood.

3.2.8 Differentiating the Aster Transform and Its Inverse

Here we follow Section A.2 of the technical report Geyer et al. (2005) which backs up the paper Geyer et al. (2007). Our notation is different from their notation, our notation here is what we have used for aster models since Geyer (2010).

Aster Transform

Let $\Delta\theta_j$ denote an infinitesimal increment of θ_j and similarly for $\Delta\varphi_j$. Then differentiating (1.35) and using (3.13) and the chain rule gives

$$\begin{aligned} \Delta\varphi_j &= \Delta\theta_j - \sum_{\substack{G \in \mathcal{G} \\ q(G) = j}} \sum_{k \in G} \xi_k \Delta\theta_k \\ &= \Delta\theta_j - \sum_{\substack{k \in J \\ p(k) = j}} \xi_k \Delta\theta_k \end{aligned} \tag{3.14}$$

In language that does not refer to infinitesimals and using the sophisticated view that derivatives are linear transformations (Browder, 1996, Definition 8.9; Lang, 1993, p. 334), the derivative of the aster transform is the linear transformation that maps the vector $\Delta\theta$ having components $\Delta\theta_j$ to the vector $\Delta\varphi$ having components $\Delta\varphi_j$.

We can think of this linear transformation as being represented by the matrix of partial derivatives, which can be read off (3.14),

$$\frac{\partial \varphi_j}{\partial \theta_k} = \begin{cases} 1, & j = k \\ -\xi_k, & j = p(k) \\ 0, & \text{otherwise} \end{cases} \tag{3.15}$$

Inverse Aster Transform

By the inverse function theorem, the derivative of the inverse is the inverse of the derivative (considered as a linear transformation), assuming it

exists, which it does if the derivative is invertible (Lang, 1993, p. 361–363). We will prove the derivative is invertible by inverting it.

As discussed in Section 1.18 above, the formula (1.37) gives an inductive definition that works when nodes of the graph are visited in any order that visits successors before predecessors.

The same is true of the derivative of the inverse aster transform. Moving a term from one side of (3.14) to the other gives

$$\Delta\theta_j = \Delta\varphi_j + \sum_{\substack{k \in J \\ p(k) = j}} \xi_k \Delta\theta_k. \tag{3.16}$$

When $\Delta\theta_j$ on the left-hand side is computed, all of the $\Delta\theta_k$ on the right-hand side will already have been computed (when we visit successors before predecessors).

As before, the derivative of the inverse aster transform is the linear transformation that maps the vector $\Delta\varphi$ having components $\Delta\varphi_j$ to the vector $\Delta\theta$ having components $\Delta\theta_j$.

Because of the nature of inductive definitions, the analog of (3.15) for the inverse transform is a bit more complicated. To help with it we introduce the following notation and conventions. Let $\mathbb{1}(\,\cdot\,)$ denote the function that maps logical formulas to numbers, mapping false formulas to zero and true formulas to one, and define empty sums (those having no terms) to be equal to zero (the identity for addition) and empty products (those having no terms) to be equal to one (the identity for multiplication).

Theorem 3.2. *Partial derivatives of the inverse aster transform are given by*

$$\frac{\partial \theta_j}{\partial \varphi_k} = \mathbb{1}(j \preceq k) \prod_{\substack{i \in J \\ j \prec i \preceq k}} \xi_i. \tag{3.17}$$

In (3.17) the product is empty when $j = k$. (The product is also empty when $j \succ k$ but then we have $\mathbb{1}(j \preceq k) = 0$ so it does not matter what the value of the product is.)

Proof. What is to be shown is that (3.17) agrees with (3.16) in the case $\Delta\varphi$ has only one nonzero component, say $\Delta\varphi_n$. In this case (3.16) says

$$\begin{aligned}
\Delta\theta_j &= 0, && j \not\preceq n \\
\Delta\theta_j &= \Delta\varphi_n, && j = n \\
\Delta\theta_j &= \sum_{\substack{k \in J \\ p(k) = j}} \xi_k \Delta\theta_k, && j \prec n
\end{aligned}$$

or

$$\begin{aligned}
\frac{\partial \theta_j}{\partial \varphi_n} &= 0, && j \not\preceq n \\
\frac{\partial \theta_j}{\partial \varphi_n} &= 1, && j = n \\
\frac{\partial \theta_j}{\partial \varphi_n} &= \sum_{\substack{k \in J \\ p(k) = j}} \xi_k \frac{\partial \theta_k}{\partial \varphi_n}, && j \prec n
\end{aligned}$$

Clearly the first two lines agree with (3.17). That leaves only the third line to check. We note in this line that the partial derivative on the right-hand side is zero unless $j \prec k \preceq n$ and that there is hence exactly one term in the sum. Thus the third line just above agrees with (3.17) by mathematical induction. □

3.2.9 Log Likelihood Derivatives

First Derivatives With Respect To θ

Applying (3.13) and (3.11) to (3.10), we obtain

$$\frac{\partial l(\theta)}{\partial \theta_j} = y_j^* - y_{p^*(j)}^* \xi_j, \qquad j \in J. \tag{3.18}$$

These are the first derivatives of the log likelihood for a saturated model with subsampling with respect to components of θ.

Notice that, as always, there is the curious mix of starred and unstarred thingummies. The data and the predecessor function have stars because this is for models with subsampling. The parameters do not have stars because we insist that models have the same parameters with and without subsampling. The index set for (3.18) is J because that is the index set for θ.

First Derivatives for CAM

In conditional aster models (CAM) with model equation (1.50) we have $\partial \theta_j / \partial \beta_k = m_{jk}$ where M has components m_{jk}. Thus, by (3.18) and the chain rule

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{j \in J} \frac{\partial l(\theta)}{\partial \theta_j} \frac{\partial \theta_j}{\partial \beta_k} = \sum_{j \in J} (y_j^* - y_{p^*(j)}^* \xi_j) m_{jk}$$

MLE are derived by setting these equal to zero, considering the ξ's as functions of β, and solving for β.

First Derivatives With Respect To φ

Theorem 3.3. *For $k \in J$ define*

$$n = \inf\{\, j \in J : j \preceq k \,\}. \tag{3.19}$$

Then

$$\frac{\partial l(\varphi)}{\partial \varphi_k} = y_k^* - y_{p^*(n)}^* \left(\prod_{\substack{i \in J \\ n \preceq i \preceq k}} \xi_i \right) + \sum_{\substack{m \in J \\ n \prec m \preceq k}} \left[y_{p(m)}^* - y_{p^*(m)}^* \right] \left(\prod_{\substack{i \in J \\ m \preceq i \preceq k}} \xi_i \right) \tag{3.20}$$

The infimum in (3.19) means $j \preceq k$ implies $n \preceq j$. The at-most-one-predecessor property (Section 3.2.2 above) makes the set $\{\, j \in J : j \preceq k \,\}$ totally ordered by $\preceq$. This set is also nonempty (it contains k) and finite (aster graphs are finite). Hence the infimum in (3.19) is always well defined.

The conventions that empty sums are zero and empty products are one (established in Section 3.2.8 above) are still in force.

If a node $p^*(m)$ is not a subsampling node, then $p(m) = p^*(m)$. Hence, if there is no subsampling, all terms $y_{p(m)}^* - y_{p^*(m)}^*$ in (3.20) are zero, and (3.20) reduces to

$$\frac{\partial l(\varphi)}{\partial \varphi_k} = y_k - y_{p(n)} \prod_{\substack{i \in J \\ n \preceq i \preceq k}} \xi_i = y_k - \mu_k$$

by (1.9), and this agrees with previous aster theory (Geyer et al., 2007, Section 3.2).

Node $p(n)$ is the unique initial node of the full aster graph satisfying $p(n) \prec k$. If $p^*(n) \neq p(n)$, then $p^*(n)$ is a subsampling node.

Proof.

$$
\begin{aligned}
\frac{\partial l(\varphi)}{\partial \varphi_k} &= \sum_{j \in J} \frac{\partial l(\theta)}{\partial \theta_j} \frac{\partial \theta_j}{\partial \varphi_k} \\
&= \sum_{\substack{j \in J \\ j \preceq k}} (y_j^* - y_{p^*(j)}^* \xi_j) \prod_{\substack{i \in J \\ j \prec i \preceq k}} \xi_i \\
&= \left(\sum_{\substack{j \in J \\ j \preceq k}} y_j^* \prod_{\substack{i \in J \\ j \prec i \preceq k}} \xi_i \right) - \left(\sum_{\substack{j \in J \\ j \preceq k}} y_{p^*(j)}^* \prod_{\substack{i \in J \\ j \preceq i \preceq k}} \xi_i \right) \\
&= y_k^* - y_{p^*(n)}^* \left(\prod_{\substack{i \in J \\ n \preceq i \preceq k}} \xi_i \right) \\
&\qquad + \left(\sum_{\substack{j \in J \\ n \preceq j \prec k}} y_j^* \prod_{\substack{i \in J \\ j \prec i \preceq k}} \xi_i \right) - \left(\sum_{\substack{m \in J \\ n \prec m \preceq k}} y_{p^*(m)}^* \prod_{\substack{i \in J \\ m \preceq i \preceq k}} \xi_i \right) \\
&= y_k^* - y_{p^*(n)}^* \left(\prod_{\substack{i \in J \\ n \preceq i \preceq k}} \xi_i \right) \\
&\qquad + \left(\sum_{\substack{m \in J \\ n \prec m \preceq k}} y_{p(m)}^* \prod_{\substack{i \in J \\ m \preceq i \preceq k}} \xi_i \right) - \left(\sum_{\substack{m \in J \\ n \prec m \preceq k}} y_{p^*(m)}^* \prod_{\substack{i \in J \\ m \preceq i \preceq k}} \xi_i \right) \\
&= y_k^* - y_{p^*(n)}^* \left(\prod_{\substack{i \in J \\ n \preceq i \preceq k}} \xi_i \right) + \sum_{\substack{m \in J \\ n \prec m \preceq k}} \left[y_{p(m)}^* - y_{p^*(m)}^* \right] \left(\prod_{\substack{i \in J \\ m \preceq i \preceq k}} \xi_i \right)
\end{aligned}
$$

where the first equality is the chain rule, the second equality is (3.17) and (3.18), and the rest are just algebra. □

First Derivatives for UAM

Hence, for unconditional aster models (UAM),

$$\frac{\partial l(\beta)}{\partial \beta_n} = \sum_{k \in J} \frac{\partial l(\varphi)}{\partial \varphi_k} m_{kn}.$$

Second Derivatives With Respect To θ

Because the R function `mlogl` in R package `aster` calculates *minus* the log likelihood and its first and second derivatives, we do the same. Negating and differentiating (3.18) gives

$$-\frac{\partial^2 l(\theta)}{\partial\theta_j\partial\theta_k} = y^*_{p^*(j)}\gamma_{jk}, \qquad j, k \in J, \tag{3.21}$$

where

$$\gamma_{jk} = \begin{cases} \partial^2 c_G(\theta_G)/\partial\theta_j\partial\theta_k, & j, k \in G \in \mathcal{G} \\ 0, & \text{otherwise} \end{cases} \tag{3.22}$$

See (3.12) above for more on γ_{jk}.

Hessian for CAM

Second derivative matrices are commonly called Hessian matrices in optimization theory. If the matrix having components (3.21) is denoted $H(\theta)$, then the Hessian for β is given by

$$H(\beta) = M^T H(\theta) M, \qquad \text{when } \theta = a + M\beta, \tag{3.23}$$

where, as usual, a is the offset vector and M is the model matrix. This is because M is the Jacobian matrix of the parameter transformation $\beta \to \theta$ in a CAM, and (3.23) is just the chain rule.

Hessian for UAM

If we follow the preceding section in denoting the Hessian for β by $H(\beta)$, the Hessian for θ by $H(\theta)$, and the Hessian for φ by $H(\varphi)$, then

$$H(\beta) = M^T H(\varphi) M, \qquad \text{when } \varphi = a + M\beta, \tag{3.24}$$

and the argument is exactly the same as in the preceding section.

3.2.10 Second Derivatives With Respect To φ

That leaves us with having to derive $H(\varphi)$, which is complicated. Before differentiating (3.20) again, it will simplify computations if we recognize that the terms in big round brackets are "parameterized common subexpressions" (one has n where the other has m). Using (3.17), we can give

one of these another notation

$$\frac{\partial \theta_{p(m)}}{\partial \varphi_k} = \prod_{\substack{i \in J \\ m \preceq i \preceq k}} \xi_i, \tag{3.25}$$

but when we change m to n in (3.25) to get a notation for the other term in big round brackets in (3.20) this does not work, because $p(n)$ is an initial node so there is no parameter $\theta_{p(n)}$. This would work, however, if we imagine that the aster graph we are working with is part of a larger aster graph in which $p(n)$ is not initial and not a subsampling node. If we work under this fiction, we should get correct mathematics.

Lemma 3.4.

$$\frac{\partial^2 \theta_{p(m)}}{\partial \varphi_j \partial \varphi_k} = \sum_{\substack{r \in J \\ p(m) \prec r \preceq k}} \sum_{\substack{s \in J \\ p(m) \prec s \preceq j}} \gamma_{rs} \prod_{\substack{t \in J \\ p(m) \prec t \prec r}} \xi_t \prod_{\substack{u \in J \\ s \prec u \preceq j}} \xi_u \prod_{\substack{v \in J \\ r \prec v \preceq k}} \xi_v \tag{3.26}$$

When all dependence groups are singletons, as with R package `aster`*, this specializes to*

$$\frac{\partial^2 \theta_{p(m)}}{\partial \varphi_j \partial \varphi_k} = \sum_{\substack{r \in J \\ p(m) \prec r \preceq k \\ p(m) \prec r \preceq j}} \gamma_{rr} \prod_{\substack{t \in J \\ p(m) \prec t \prec r}} \xi_t \prod_{\substack{u \in J \\ r \prec u \preceq j}} \xi_u \prod_{\substack{v \in J \\ r \prec v \preceq k}} \xi_v \tag{3.27}$$

Lemma 3.5. *If $m \prec r$ and $\gamma_{rs} \neq 0$, then $m \prec s$.*

Proof. If $r = s$ the assertion is trivial. Otherwise, $\gamma_{rs} \neq 0$ implies that r and s are in the same dependence group, say G, and none of the arrows for this dependence group are subsampling arrows. It follows that $p(r) = p(s) = q(G)$. So $m \prec r$ implies $m \preceq q(G)$ implies $m \prec s$. □

Proof of Lemma 3.4.

$$\begin{aligned}
\frac{\partial^2 \theta_{p(m)}}{\partial \varphi_j \partial \varphi_k} &= \frac{\partial}{\partial \varphi_j} \prod_{\substack{i \in J \\ m \preceq i \preceq k}} \xi_i \\
&= \sum_{\substack{r \in J \\ m \preceq r \preceq k}} \frac{\partial \xi_r}{\partial \varphi_j} \prod_{\substack{i \in J \\ m \preceq i \preceq k \\ i \neq r}} \xi_i \\
&= \sum_{\substack{r \in J \\ m \preceq r \preceq k}} \sum_{s \in J} \frac{\partial \xi_r}{\partial \theta_s} \frac{\partial \theta_s}{\partial \varphi_j} \prod_{\substack{i \in J \\ m \preceq i \preceq k \\ i \neq r}} \xi_i \\
&= \sum_{\substack{r \in J \\ m \preceq r \preceq k}} \sum_{\substack{s \in J \\ s \preceq j}} \gamma_{rs} \prod_{\substack{u \in J \\ s \prec u \preceq j}} \xi_u \prod_{\substack{i \in J \\ m \preceq i \preceq k \\ i \neq r}} \xi_i \\
&= \sum_{\substack{r \in J \\ m \preceq r \preceq k}} \sum_{\substack{s \in J \\ s \preceq j}} \gamma_{rs} \prod_{\substack{u \in J \\ s \prec u \preceq j}} \xi_u \prod_{\substack{t \in J \\ m \preceq t \prec r}} \xi_t \prod_{\substack{v \in J \\ r \prec v \preceq k}} \xi_v \\
&= \sum_{\substack{r \in J \\ p(m) \prec r \preceq k}} \sum_{\substack{s \in J \\ s \preceq j}} \gamma_{rs} \prod_{\substack{u \in J \\ s \prec u \preceq j}} \xi_u \prod_{\substack{t \in J \\ p(m) \prec t \prec r}} \xi_t \prod_{\substack{v \in J \\ r \prec v \preceq k}} \xi_v
\end{aligned}$$

where the first equality is (3.25), the second equality is the product rule, the third equality is the chain rule, the fourth equality is (3.13), (3.22), and (3.17), the fifth equality just splits one product into two products. and the last equality is $m \preceq r$ if and only if $p(m) \prec r$ and similarly for t. Finally we use Lemma 3.5 to get $p(m) \prec r$ and $\gamma_{rs} \neq 0$ implies $p(m) \prec s$.

Going from (3.26) to (3.27) is just that, when all dependence groups are singletons, $\gamma_{rs} \neq 0$ implies $r = s$. □

Theorem 3.6. *Define n by* (3.19)*. For an unconditional aster model with subsampling, the j, k component of $H(\varphi)$ is*

$$-\frac{\partial^2 l(\varphi)}{\partial \varphi_j \partial \varphi_k} = y^*_{p^*(n)} \frac{\partial^2 \theta_{p(n)}}{\partial \varphi_j \partial \varphi_k} - \sum_{\substack{m \in J \\ n \prec m \preceq k}} \left[y^*_{p(m)} - y^*_{p^*(m)} \right] \frac{\partial^2 \theta_{p(m)}}{\partial \varphi_j \partial \varphi_k} \tag{3.28}$$

where the second partial derivatives of θ with respect to φ are given by (3.26) *or* (3.27) *and where these equations are to be used regardless of whether $p(n)$ actually indexes a parameter.*

Proof. Immediate from (3.20). □

As was remarked after Theorem 3.3, a node $p^*(m)$ is not a subsampling node if and only if $p(m) = p^*(m)$, in which case the term containing $y^*_{p(m)} - y^*_{p^*(m)}$ is exactly zero.

Since cumulant functions are infinitely differentiable, so is the aster transform, the inverse aster transform, and aster log likelihoods. Thus formulas (3.26), (3.27) and (3.28) must be equal when j and k are interchanged. We have written them in a form so that this is almost obvious. The only non-obvious spot is when we interchange j and k in (3.26) we (in effect) change $p(m) \prec t \prec r$ into $p(m) \prec t \prec s$, but this agrees with Lemma 3.5: $t \prec r$ and $\gamma_{rs} \neq 0$ implies $t \prec s$.

3.2.11 Fisher Information

Change of Parameter

We begin with a theorem about change of Fisher information under smooth change of parameter that is for general likelihood inference under the "usual regularity conditions." It has nothing in particular to do with exponential families. The "usual regularity conditions" hold for all regular exponential families including curved exponential families. So they hold for all models in this article. But they hold for many other models too.

The only regularity condition we need is the so-called Bartlett identities, which are usually derived by differentiating the integral of the probability densities twice with respect to the parameters

$$E_\theta\{\nabla l(\theta)\} = 0 \tag{3.29}$$

$$\operatorname{var}_\theta\{\nabla l(\theta)\} = -E_\theta\{\nabla^2 l(\theta)\} \tag{3.30}$$

where $\nabla l(\theta)$ denotes the vector of first partial derivatives of the log likelihood, $\nabla^2 l(\theta)$ denotes the matrix of second partial derivatives of the log likelihood, and var denotes the variance operator that produces a variance matrix (also called variance-covariance matrix, covariance matrix, and dispersion matrix).

Expected Fisher information is either side of (3.30). Observed Fisher information is $-\nabla^2 l(\theta)$.

The theorem is about what happens under a change of parameter $\theta = g(\psi)$. In short, the theorem says Fisher information transforms like a tensor. If expected Fisher information for θ is denoted $I(\theta)$, and observed Fisher information for θ is denoted $J(\theta)$ (note that the latter is a random quantity despite there being no explicit indication of this), and the Jacobian of the

transformation is $B(\psi) = \nabla g(\psi)$, then

$$I(\psi) = B(\psi)^T I(\theta) B(\psi), \qquad \text{when } \theta = g(\psi). \tag{3.31}$$

This is reminiscent of and ultimately derives from the formula for change of variance under a linear transformation. If $y = Bx$, where x and y are random vectors and B is a known matrix, then

$$\operatorname{var}(y) = B \operatorname{var}(x) B^T. \tag{3.32}$$

The analogous result for observed Fisher information is a bit trickier. Changing I to J in (3.31) gives a statement that is, in general, false. But it is true when MLE are plugged in

$$J(\hat{\psi}) = B(\hat{\psi})^T J(\hat{\theta}) B(\hat{\psi}), \qquad \text{when } \hat{\theta} = g(\hat{\psi}). \tag{3.33}$$

When $\hat{\theta} = g(\hat{\psi})$ and $\hat{\psi}$ is an MLE, then $\hat{\theta}$ is also an MLE by by invariance of MLE under parameter transformation.

Theorem 3.7. *Assume* (3.29) *and* (3.30)*. Then statement* (3.31) *is correct. If $\hat{\theta} = g(\hat{\psi})$ is a zero of the first derivative of the log likelihood for θ, then statement* (3.33) *is correct. The change of parameter function g must be injective and differentiable but need not be surjective.*

Proof. By the chain rule

$$\nabla l(\psi) = \nabla l(\theta) B(\psi), \qquad \text{when } \theta = g(\psi).$$

Take the variance of both sides and use (3.32) to obtain (3.31).

The situation is more complicated with second derivatives

$$-\frac{\partial^2 l(\psi)}{\partial \psi_i \partial \psi_j} = -\left(\sum_k \sum_m \frac{\partial^2 l(\theta)}{\partial \theta_k \partial \theta_m} \frac{\partial \theta_k}{\partial \psi_i} \frac{\partial \theta_m}{\partial \psi_j} \right) - \sum_k \frac{\partial l(\theta)}{\partial \theta_k} \frac{\partial^2 \theta_k}{\partial \psi_i \partial \psi_j}.$$

When MLE are plugged in,

$$\left. \frac{\partial l(\theta)}{\partial \theta_k} \right|_{\theta = \hat{\theta}} = 0$$

so this gives (3.33). □

Fisher Information for θ

Returning to aster models with subsampling, from (3.21) we get:

- observed Fisher information for θ is the matrix having components $y^*_{p^*(j)}\gamma_{jk}$ and
- expected Fisher information for θ is the matrix having components $\mu^*_{p^*(j)}\gamma_{jk}$.

Note that, since expected Fisher information depends on components of μ^*, which cannot, in general, be calculated by the computer, use of expected Fisher information depends on those subsampling probability distributions, whereas the log likelihood, its derivatives, and observed Fisher information do not depend on them.

Thus, in general, we can only use observed Fisher information.

Fisher Information for β for CAM

If $I(\theta)$ is expected Fisher information for θ derived in the preceding section, then expected Fisher information for β is

$$I(\beta) = M^T I(\theta) M, \qquad \text{when } \theta = a + M\beta,$$

because M is the derivative of θ with respect to β. We also have the analogous relationship for observed Fisher information when MLE are plugged in

$$J(\hat{\beta}) = M^T J(\hat{\theta}) M, \qquad \text{when } \hat{\theta} = a + M\hat{\beta}.$$

Fisher Information for β for UAM

As in the preceding section, let $I(\theta)$ and $J(\theta)$ be expected and observed Fisher information for θ for the saturated model with subsampling, which were derived in Section 3.2.11 above. Now let $a(\varphi)$ denote the inverse aster transform described in Section 1.18 above, and let $A(\varphi)$ denote its derivative, whose components are given by (3.17). Now the model equation is (1.47) so M is the derivative of φ with respect to β.

So now expected Fisher information for β is

$$I(\beta) = M^T A(\varphi)^T I(\theta) A(\varphi) M, \qquad \text{when } \varphi = a + M\beta \text{ and } \theta = a(\varphi),$$

and the analogous relationship for observed Fisher information when MLE are plugged in is

$$J(\hat{\beta}) = M^T A(\hat{\varphi})^T I(\hat{\theta}) A(\hat{\varphi}) M, \qquad \text{when } \hat{\varphi} = a + M\hat{\beta} \text{ and } \hat{\theta} = a(\hat{\varphi}).$$

3.2.12 Prediction

Prediction for all six parameterizations of UAM discussed in Section 1.22.2 above can be handled by the `aster` and `aster.formula` methods of the R generic function `predict` (that is, the functions `predict.aster` and `predict.aster.formula` considered as non-generic functions). One calls these functions by the name `predict` but must look up the help page with `help(predict.aster)`. These functions are in R package `aster`. (A referee for Geyer et al. (2007) complained that these are not predictions but rather parameter transformations and we agreed, but users expect to use the R generic function `predict` to do this job.)

Since version 1.0 of the package, these functions have a new optional argument `is.always.parameter = TRUE` which makes them "predict" ξ rather than the vector having components $E(y_j \mid y_{p(j)})$, which is not a parameter. If we always use this option, this function can make predictions for all parameters in the chain $\beta \to \varphi \to \theta \to \xi \to \mu$, and if we add $\tau = M^T \mu$, which we can do ourselves (given μ calculated by `predict`), we have all six parameters.

As parameters, none of these depend on the response vector, although they do depend on covariates because the model matrix depends on covariates. Thus if the R function `predict` is provided an object of class `"aster"` or `"aster.formula"` that has the MLE $\hat{\beta}$ for the aster model with subsampling as its `coefficients` component, and corresponding expected and observed Fisher information matrices as its `fisher` and `hessian` components but every other component as if subsampling had not been done, then `predict` will operate to make predictions without subsampling.

We also want the `deviance` component of our object of class `"aster"` or `"aster.formula"` containing the result of fitting the model with subsampling to be reflect the subsampling (be minus twice the maximized log likelihood) so that the R generic function `anova` does the right thing with these objects.

3.2.13 Parameter Transformation

The methods for the R generic function `predict` in R package `aster` cannot do all of the parameter transformations discussed in Sections 1.21 and 1.22.2 above. A function `astertransform` was added to this package, but it only does transformations from θ and φ to any of θ, φ, ξ, or μ.

R package `aster2` does do all of these parameter transformations. The function `transformSaturated` does any of the transformations from θ, φ,

ξ, or μ to any of these. The function `transformConditional` does any of the transformations for a CAM without subsampling from β to θ, φ, ξ, or μ (but not vice versa). The function `transformUnconditional` does any of the transformations from β or τ (the only parameters that are unconstrained) to β, θ, φ, ξ, μ, or τ.

Both `transformSaturated` and `transformUnconditional` are able to transform from mean value to canonical parameters.

Calculating the parameter transformations $\xi \to \theta$ or $\mu \to \varphi$ for saturated aster models without subsampling and $\tau \to \beta$ for UAM without subsampling is equivalent to doing maximum likelihood with data replaced by unconditional mean value parameters. Theorem 1.8 covers all of these cases.

Appendix A

The Factorization Theorem

Proof of Theorem 1.1. A valid factorization factors joint equals conditional times marginal

$$\operatorname{pr}(y) = \operatorname{pr}(y_{G_1} \mid y_{N \setminus G_1}) \operatorname{pr}(y_{N \setminus G_1})$$

The marginal on the right-hand side can then be considered a joint to be factored further

$$\operatorname{pr}(y) = \operatorname{pr}(y_{G_1} \mid y_{N \setminus G_1}) \operatorname{pr}(y_{G_2} \mid y_{N \setminus (G_1 \cup G_2)}) \operatorname{pr}(y_{N \setminus (G_1 \cup G_2)})$$

and again and again giving

$$\operatorname{pr}(y) = \operatorname{pr}(y_{N \setminus \bigcup_{j=1}^{k} G_j}) \prod_{i=1}^{k} \operatorname{pr}(y_{G_i} \mid y_{N \setminus \bigcup_{j=1}^{i} G_j}) \tag{A.1}$$

and the only condition that is required to make (A.1) valid is that the index sets G_i are disjoint. This is the only operation in classical (non-measure-theoretic) probability theory that factorizes probability distributions. A factorization is valid if and only if it has the form (A.1).

When we match up (1.1) and (A.1) we see that the G_i must be the elements of $\mathcal{G}$ so the two products are the same. For the conditional distributions to match up we must have $\operatorname{pr}(y_{G_i} \mid y_{N \setminus \bigcup_{j=1}^{i} G_j})$ in (A.1) can actually be written as $\operatorname{pr}(y_{G_i} \mid y_{q(G_i)})$, that is,

- this conditional distribution actually depends only on the single variable $y_{q(G_i)}$ not on the rest of the variables that are components of $y_{N \setminus \bigcup_{j=1}^{i} G_j}$ and
- $q(G_i) \in N \setminus \bigcup_{j=1}^{i} G_j$, that is, either $q(G_i) \in G_j$ for some $j > i$ or $q(G_i)$ is an initial node ($q(G_i) \notin G_j$ for any j). In either case, $q(G_i) \in G_j$

implies $i < j$. Thus we have the condition of the theorem: $G_i < G_j$ if and only if $i < j$.

Finally, we must match up the marginal term on the right-hand side of (A.1). It matches nothing in (1.1), which is the same as saying it must be equal to one, which is they same as saying $y_{N \setminus J}$ is a constant random vector, where $J = \bigcup \mathcal{G}$ as always. □

Appendix B

Markov Properties

Markov properties of graphical models are considered a fundamental part of the theory (Lauritzen, 1996, Chapter 3). They are much less important for aster models, so unimportant that the literature on aster models does not mention them. So perhaps most readers will want to skip this appendix. Nevertheless, perhaps these ideas might find some future use. So we do them.

A Markov property is a conditional independence relation derived from a graph (or for aster models from the fundamental factorization (1.1)). There are many more Markov properties than we bother to prove here.

Lemma B.1. *Let $\mathcal{H}$ be any subset of $\mathcal{G}$. Then the random vectors y_H, $H \in \mathcal{H}$ are conditionally independent given the random scalars $y_{q(H)}$, $H \in \mathcal{H}$.*

Note that some y_j can possibly appear among some $H \in \mathcal{H}$ and in some $y_{q(H)}$, $H \in \mathcal{H}$ so we have to say what that means. Conditioning on a random variable is the same as treating it as constant, and a constant random variable is independent of any random variables including itself. Thus for any sets A and B, we have

$$\operatorname{pr}(y_A \mid y_B) = \operatorname{pr}(y_{A \setminus B} \mid y_B). \tag{B.1}$$

In (B.1), the case $A \setminus B = \varnothing$ is possible, in which case $y_\varnothing$ is the constant random vector discussed in Section 1.4 above. Thus $\operatorname{pr}(y_\varnothing \mid y_B) = 1$ regardless of what y_B is.

This lemma does not say that the components of the random vectors y_H are conditionally independent. The components of y_G are dependent given $y_{q(G)}$ for any G. That is the whole point of dependence groups.

Proof. Use the total order on $\mathcal{G}$ guaranteed to exist by Theorem 1.1 to enumerate $\mathcal{G}$ as $G_1 < G_2 < \cdots < G_n$. Then $H_j = G_{i_j}$ for $j = 1, \ldots, m$, where $1 \le i_1 < i_2 < \cdots < i_m \le n$.

We integrate out y_G one at a time in order skipping when $G \in \mathcal{H}$ and also not integrating out any $y_{q(H)}$ for $H \in \mathcal{H}$.

We start by sum-integrating out y_{G_1} if $G_1 \ne H_1$ obtaining

$$\operatorname{pr}(y_{G_2 \cup \cdots \cup G_n}) = \prod_{i=2}^{n} \operatorname{pr}(y_{G_i} \mid y_{q(G_i)}).$$

and keep going repeating this again and again obtaining

$$\begin{aligned} \operatorname{pr}(y_{\bigcup\{\, G \in \mathcal{G} : G \ge H_1 \,\}}) &= \prod_{\substack{G \in \mathcal{G} \\ G \ge H_1}} \operatorname{pr}(y_G \mid y_{q(G)}) \\ &= \operatorname{pr}(y_{H_1} \mid y_{q(H_1)}) \prod_{\substack{G \in \mathcal{G} \\ G > H_1}} \operatorname{pr}(y_G \mid y_{q(G)}) \end{aligned}$$

(if $G_1 = H_1$ we haven't done anything yet and this is just the same factorization as (1.1) in different notation).

Now we have to be careful with our notation. Define

$$Q = \{\, q(H) : H \in \mathcal{H} \,\}$$

we need to not sum-integrate out any components of y_Q.

If $G_{i_1+1} \ne H_2$, then we want to sum-integrate out $y_{G_{i_1+1} \setminus Q}$ obtaining

$$\begin{aligned} &\operatorname{pr}(y_{H_1 \cup \{q(H_1)\} \cup \{\, G \in \mathcal{G} : G > G_{i_1+1} \,\}}) \\ &\qquad = \operatorname{pr}(y_{H_1} \mid y_{q(H_1)}) \operatorname{pr}(y_{G_{i_1+1} \cap Q} \mid y_{q(G_{i_1+1})}) \prod_{j=i_1+2}^{n} \operatorname{pr}(y_{G_i} \mid y_{q(G_i)}) \end{aligned}$$

(this uses the discussion of $y_\varnothing$ preceding this proof, since $G_{i_1+1} \cap Q$ may or may not be the empty set).

Continuing this process, we obtain

$$\begin{aligned} &\operatorname{pr}(y_{H_1 \cup \{q(H_1)\} \cup \{G_{i_2}, \ldots, G_n\}}) \\ &\qquad = \operatorname{pr}(y_{H_1} \mid y_{q(H_1)}) \prod_{j=i_1+1}^{i_2-1} \operatorname{pr}(y_{G_j \cap Q} \mid y_{q(G_j)}) \prod_{j=i_2}^{n} \operatorname{pr}(y_{G_i} \mid y_{q(G_i)}). \end{aligned}$$

And we can now see how this process continues

$$\operatorname{pr}(y_{H_1 \cup H_2 \cup \cdots \cup H_m \cup Q}) = \prod_{H \in \mathcal{H}} \operatorname{pr}(y_H \mid y_{q(H)}) \prod_{G \in \mathcal{G} \setminus \mathcal{H}} \operatorname{pr}(y_{G \cap Q} \mid y_{q(G)})$$

And now integrating out $y_{H \setminus Q}$ in order gives

$$\operatorname{pr}(y_Q) = \prod_{H \in \mathcal{H}} \operatorname{pr}(y_{H \cap Q} \mid y_{q(H)}) \prod_{G \in \mathcal{G} \setminus \mathcal{H}} \operatorname{pr}(y_{G \cap Q} \mid y_{q(G)})$$

(Note that every y_j for $j \in Q$ appears "in front of the bar" in exactly one of these conditional probabilities because $\mathcal{G}$ is a partition.) So

$$\begin{aligned} \operatorname{pr}(y_{H_1 \cup H_2 \cup \cdots \cup H_m} \mid y_Q) &= \prod_{H \in \mathcal{H}} \frac{\operatorname{pr}(y_H \mid y_{q(H)})}{\operatorname{pr}(y_{H \cap Q} \mid y_{q(H)})} \\ &= \prod_{H \in \mathcal{H}} \frac{\operatorname{pr}(y_{H \cup \{q(H)\}})}{\operatorname{pr}(y_{(H \cap Q) \cup \{q(H)\}})} \\ &= \prod_{H \in \mathcal{H}} \operatorname{pr}(y_{H \setminus Q} \mid y_Q) \\ &= \prod_{H \in \mathcal{H}} \operatorname{pr}(y_H \mid y_Q) \end{aligned}$$

the last step being (B.1). □

If $\mathcal{G}$ and $\mathcal{H}$ are partitions of a set J, then we say that $\mathcal{G}$ is *finer* than $\mathcal{H}$ if every element of $\mathcal{G}$ is contained in some element of $\mathcal{H}$. We also say that $\mathcal{H}$ is *coarser* than $\mathcal{G}$ to indicate the same concept.

Clearly, every element of $\mathcal{H}$ is the union of elements of $\mathcal{G}$ it contains (because $\mathcal{G}$ is a partition).

Theorem B.2. *Suppose $\mathcal{G}$ and q are as in* (1.1) *and Theorem 1.1, and suppose $\mathcal{H}$ is a coarser partition than $\mathcal{G}$. Define*

$$Q = \{\, q(G) : (G, H) \in \mathcal{G} \times \mathcal{H} \text{ and } G \subset H \text{ and } q(G) \notin H \,\} \tag{B.2}$$

then the random vectors y_H, $H \in \mathcal{H}$, are conditionally independent given the random vector y_Q.

Note that (B.1) is being used in this theorem too. Some y_j may appear in some y_H and also in y_Q.

Also we repeat the comment following the lemma. The theorem does not assert conditional independence of the components of y_H for any H. The components of y_G being dependent given $y_{q(G)}$ is the whole point of dependence groups.

Proof. We prove this by induction. The induction variable is the partition $\mathcal{H}$. We start with $\mathcal{H} = \mathcal{G}$. Then we change $\mathcal{H}$ to coarser and coarser partitions until we get to the $\mathcal{H}$ in the theorem statement.

The base of the induction is the case $\mathcal{H} = \mathcal{G}$ in which case the lemma and the theorem say the same thing. So that establishes the base of the induction.

In each induction step we decrease the cardinality of $\mathcal{H}$ by one. This means we take two elements H' and H'' of $\mathcal{H}$ and merge them to make one element of $\mathcal{H}$ after the induction step, and all other elements of $\mathcal{H}$ remain unchanged (in this particular induction step). We need to show that if the assertion of the theorem is true before the induction step, then it is true after the induction step, when $\mathcal{H}$ is changed as described.

Let $\mathcal{H}_{\text{before}}$ denote $\mathcal{H}$ before the induction step and $\mathcal{H}_{\text{after}}$ denote $\mathcal{H}$ after the induction step, so all elements of $\mathcal{H}_{\text{before}}$ and $\mathcal{H}_{\text{after}}$ are the same except

- $\mathcal{H}_{\text{before}}$ has elements H' and H'' which are not in $\mathcal{H}_{\text{after}}$ and
- $\mathcal{H}_{\text{after}}$ has the element $H' \cup H''$ which is not in $\mathcal{H}_{\text{before}}$.

Let Q_{before} denote Q before the induction step and Q_{after} denote Q after the induction step, so all of the elements of Q_{before} and Q_{after} are the same except

- $q(G)$, $G \in \mathcal{G}$ such that $G \subset H' \cup H''$ and $q(G) \in H' \cup H''$ are not in Q_{after}. (Some of these may not have been in Q_{before} either.)

In case $Q_{\text{before}} = Q_{\text{after}}$ there is nothing to prove. Conditional independence of y_H, $H \in \mathcal{H}_{\text{before}}$ given $y_{Q_{\text{before}}}$ clearly implies conditional independence of y_H, $H \in \mathcal{H}_{\text{after}}$ given $y_{Q_{\text{after}}}$ in this case where $Q_{\text{before}} = Q_{\text{after}}$. The latter statement just forgets part of the assertion of the former. (It forgets about conditional independence of $y_{H'}$ and $y_{H''}$.)

In case $Q_{\text{before}} \neq Q_{\text{after}}$ there is more work to be done. The induction hypothesis says

$$\text{pr}(y \mid y_{Q_{\text{before}}}) = \prod_{H \in \mathcal{H}_{\text{before}}} \text{pr}(y_H \mid y_{Q_{\text{before}}}).$$

First we notice

$$Q_{\text{before}} \setminus Q_{\text{after}} \subset H' \cup H'' \tag{B.3}$$

so by the induction hypothesis

$$\text{pr}(y_H \mid y_{Q_{\text{before}}}) = \text{pr}(y_H \mid y_{Q_{\text{after}}}), \qquad H \in \mathcal{H}_{\text{before}} \cap \mathcal{H}_{\text{after}}. \tag{B.4}$$

Now

$$\begin{aligned}\operatorname{pr}(y_{H'} \mid y_{Q_{\text{before}}}) \operatorname{pr}(y_{H''} \mid y_{Q_{\text{before}}}) \operatorname{pr}(y_{Q_{\text{before}}}) & \\ = \operatorname{pr}(y_{H' \cup H''} \mid y_{Q_{\text{before}}}) \operatorname{pr}(y_{Q_{\text{before}}}) & \\ = \operatorname{pr}(y_{H' \cup H'' \cup Q_{\text{before}}}) & \\ = \operatorname{pr}(y_{H' \cup H'' \cup Q_{\text{after}}}) &\end{aligned}$$

the first equality being the conditional independence asserted by the induction hypothesis and the last equality being (B.3). So

$$\begin{aligned}\operatorname{pr}(y_{H' \cup H''} \mid y_{Q_{\text{after}}}) &= \frac{\operatorname{pr}(y_{H' \cup H'' \cup Q_{\text{after}}})}{\operatorname{pr}(y_{Q_{\text{after}}})} \\ &= \operatorname{pr}(y_{H'} \mid y_{Q_{\text{before}}}) \operatorname{pr}(y_{H''} \mid y_{Q_{\text{before}}}) \frac{\operatorname{pr}(y_{Q_{\text{before}}})}{\operatorname{pr}(y_{Q_{\text{after}}})} \\ &= \operatorname{pr}(y_{H'} \mid y_{Q_{\text{before}}}) \operatorname{pr}(y_{H''} \mid y_{Q_{\text{before}}}) \operatorname{pr}(y_{Q_{\text{before}}} \mid y_{Q_{\text{after}}})\end{aligned}$$

By a similar argument we have

$$\begin{aligned}\operatorname{pr}(y \mid y_{Q_{\text{after}}}) &= \operatorname{pr}(y_{Q_{\text{before}}} \mid y_{Q_{\text{after}}}) \prod_{H \in \mathcal{H}_{\text{before}}} \operatorname{pr}(y_H \mid y_{Q_{\text{before}}}) \\ &= \operatorname{pr}(y_{H'} \mid y_{Q_{\text{before}}}) \operatorname{pr}(y_{H''} \mid y_{Q_{\text{before}}}) \operatorname{pr}(y_{Q_{\text{before}}} \mid y_{Q_{\text{after}}}) \\ &\quad \times \prod_{H \in \mathcal{H}_{\text{before}} \cap \mathcal{H}_{\text{after}}} \operatorname{pr}(y_H \mid y_{Q_{\text{before}}}) \\ &= \operatorname{pr}(y_{H' \cup H''} \mid y_{Q_{\text{after}}}) \prod_{H \in \mathcal{H}_{\text{before}} \cap \mathcal{H}_{\text{after}}} \operatorname{pr}(y_H \mid y_{Q_{\text{before}}}) \\ &= \operatorname{pr}(y_{H' \cup H''} \mid y_{Q_{\text{after}}}) \prod_{H \in \mathcal{H}_{\text{before}} \cap \mathcal{H}_{\text{after}}} \operatorname{pr}(y_H \mid y_{Q_{\text{after}}}) \\ &= \prod_{H \in \mathcal{H}_{\text{after}}} \operatorname{pr}(y_H \mid y_{Q_{\text{after}}})\end{aligned}$$

where the next-to-last step is (B.4). And reading from end to end gives the assertion that the induction step must prove. Hence we are done. □

Let

$$\mathcal{G}_{\text{initial}} = \{\, G \in \mathcal{G} : q(G) \notin J \,\}$$

(the notation is perhaps a bit misleading, this is the subset of $\mathcal{G}$ whose elements have predecessors that are initial nodes). Now for $G \in \mathcal{G}_{\text{initial}}$, let

$$H_G = \{\, j \in J : (\exists k \in G)(j \succeq k) \,\}$$

where, as usual $\succeq$ denotes the reflexive transitive closure of the predecessor relation. These H_G are the node sets for what are called aster graphs for "individuals" (in scare quotes) in Section 1.9 above. Let

$$\mathcal{H} = \{\, H_G : G \in \mathcal{G}_{\text{initial}} \,\}. \tag{B.5}$$

Corollary B.3. *The random vectors y_H, $H \in \mathcal{H}$, with $\mathcal{H}$ defined by* (B.5) *are (unconditionally) independent.*

Proof. Immediate from the theorem because the Q corresponding to this $\mathcal{H}$ consists of initial nodes only so y_Q is a constant random vector, and conditioning on a constant has no effect. Any things conditionally independent given y_Q are unconditionally independent given y_Q. □

Appendix C

Regularity

As mentioned in Section 1.16 where the exponential family assumption for aster models was introduced, the cumulant function for the degenerate family concentrated at zero is the zero function that is everywhere equal to zero. The family consisting of this distribution only is a regular full exponential family because c_G is everywhere finite. So the full canonical parameter space of this family is $\mathbb{R}^G$.

Theorem C.1. *If $Y_{q(G)} = 0$ almost surely for any dependence group G, replace the family for this dependence group by the degenerate family concentrated at zero so c_G is the zero function. Then, if families for every dependence group of the aster model are regular full exponential families, then so is the (joint) distribution of the aster model. The full (unconditional) canonical parameter space of the aster model is the range of the aster transform. The cumulant function of the aster model is given by* (1.36) *for parameter values where it is finite.*

Let Θ_G denote the full canonical parameter space of the exponential family for dependence group G, which is the set of points where c_G is finite. The the set of all θ values that correspond to possible distributions in the aster model is

$$\Theta = \prod_{G \in \mathcal{G}} \Theta_G \tag{C.1}$$

where the product denotes Cartesian product: this is the set of all θ such that $\theta_G \in \Theta_G$ for all G. If we temporarily give the aster transform a letter f, then the range of this function is denoted $\Phi = f(\Theta)$. This is the set of all vectors φ that correspond to vectors θ that parameterize distributions in the aster model.

Note that we don't have an explicit description of Φ. We don't even have a closed-form expression for f, only the recursive definition (1.35). But we know that f is a function having domain Θ and range Φ, and the inverse aster transform is a function having domain Φ and range Θ.

One assertion of the theorem is that when we calculate the cumulant function of the (joint) distribution of the aster model using (1.26) the result is finite if and only if $\varphi \in \Phi$ and when it is finite the result agrees with (1.36). Another assertion of the theorem is that Φ is an open subset of the vector space where φ takes values.

Proof. From (1.26)

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

or

$$e^{c(\varphi) - c(\varphi^*)} = E_{\varphi^*}\left(e^{\langle Y, \varphi - \varphi^* \rangle}\right) \tag{C.2}$$

(what were θ and ψ in (1.26) have become φ and φ^*, respectively, here because we want to emphasize that they are both possible values of the unconditional canonical parameter vector). Let θ and θ^* denote the conditional canonical parameter vectors corresponding to φ and φ^*, respectively.

We also note that (1.26) holds for each dependence group

$$e^{c_G(\theta_G) - c(\theta_G^*)} = E_{\varphi^*}\left(e^{\langle Y_G, \theta_G - \theta_G^* \rangle} \middle| Y_{q(G)} = 1\right) \tag{C.3}$$

and since the cumulant function for sample size n is n times the cumulant function for sample size one

$$e^{y_{q(G)}[c_G(\theta_G) - c(\theta_G^*)]} = E_{\varphi^*}\left(e^{\langle Y_G, \theta_G - \theta_G^* \rangle} \middle| y_{q(G)}\right) \tag{C.4}$$

Use the total order on $\mathcal{G}$ guaranteed to exist by Theorem 1.1 to enumerate $\mathcal{G}$ as $G_1 < G_2 < \cdots < G_n$, and for $k = 0, \ldots, n$ define

$$\mathcal{G}_k = \{ G_1, \ldots, G_k \}$$

where the notation is intended to mean that $\mathcal{G}_0$ is another notation for the empty set. We claim

$$\begin{aligned} &E_{\varphi^*}\left(e^{\langle Y, \varphi - \varphi^* \rangle}\right) \\ &\qquad = E_{\varphi^*}\left(\prod_{\substack{G \in \mathcal{G}_k \\ q(G) \notin \bigcup \mathcal{G}_k}} e^{Y_{q(G)}[c_G(\theta_G) - c_G(\theta_G^*)]} \prod_{G \in \mathcal{G} \setminus \mathcal{G}_k} e^{\langle Y_G, \varphi_G - \varphi_G^* \rangle} \right) \end{aligned} \tag{C.5}$$

hold for $k = 0, \ldots, n$ and we prove this by induction.

The base of the induction is the case $k = 0$ in which case the first product is empty and by convention equal to one. Then (C.5) is obviously equivalent to (C.2).

To prove the induction step we assume (C.5) and prove (C.5) with k replaced by $k + 1$. Note that (1.35) says

$$\theta_j - \theta_j^* = \varphi_j - \varphi_j^* + \sum_{\substack{G \in \mathcal{G} \\ q(G) = j}} [c_G(\theta_G) - c_G(\theta_G^*)] \tag{C.6}$$

so

$$E_{\varphi^*}\left(\prod_{\substack{G \in \mathcal{G}_k \\ q(G) \notin \bigcup \mathcal{G}_k}} e^{Y_{q(G)}[c_G(\theta_G) - c_G(\theta_G^*)]} \prod_{G \in \mathcal{G} \setminus \mathcal{G}_k} e^{\langle Y_G, \varphi_G - \varphi_G^* \rangle} \right) =$$

$$E_{\varphi^*}\left(\prod_{\substack{G \in \mathcal{G}_k \\ q(G) \notin \bigcup \mathcal{G}_{k+1}}} e^{Y_{q(G)}[c_G(\theta_G) - c_G(\theta_G^*)]} \prod_{G \in \mathcal{G} \setminus \mathcal{G}_{k+1}} e^{\langle Y_G, \varphi_G - \varphi_G^* \rangle} e^{\langle Y_{G_k}, \theta_{G_k} - \theta_{G_k}^* \rangle} \right)$$

$$= E_{\varphi^*}\left(\prod_{\substack{G \in \mathcal{G}_k \\ q(G) \notin \bigcup \mathcal{G}_{k+1}}} e^{Y_{q(G)}[c_G(\theta_G) - c_G(\theta_G^*)]} \prod_{G \in \mathcal{G} \setminus \mathcal{G}_{k+1}} e^{\langle Y_G, \varphi_G - \varphi_G^* \rangle} \right.$$

$$\left. \times E_{\varphi^*}\left\{ e^{\langle Y_{G_k}, \theta_{G_k} - \theta_{G_k}^* \rangle} \middle| Y_{\bigcup(\mathcal{G} \setminus \mathcal{G}_k)} \right\} \right)$$

and this is equal to (C.5) with k replaced by $k + 1$ by the Markov property

$$E_{\varphi^*}\left\{ e^{\langle Y_{G_k}, \theta_{G_k} - \theta_{G_k}^* \rangle} \middle| y_{\bigcup(\mathcal{G} \setminus \mathcal{G}_k)} \right\} = E_{\varphi^*}\left\{ e^{\langle Y_{G_k}, \theta_{G_k} - \theta_{G_k}^* \rangle} \middle| y_{q(G_k)} \right\}$$

and (C.4). That finishes the proof of the induction claim.

The $k = n$ case of (C.5) is

$$E_{\varphi^*}\left(e^{\langle Y, \varphi - \varphi^* \rangle}\right) = E_{\varphi^*}\left(\prod_{\substack{G \in \mathcal{G} \\ q(G) \notin J}} e^{Y_{q(G)}[c_G(\theta_G) - c_G(\theta_G^*)]} \right)$$

and because every $Y_{q(G)}$ appearing in the expectation is a constant random variable at an initial node, the expectation does nothing, so

$$E_{\varphi^*}\bigl(e^{\langle Y, \varphi - \varphi^* \rangle}\bigr) = \prod_{\substack{G \in \mathcal{G} \\ q(G) \notin J}} e^{y_{q(G)}[c_G(\theta_G) - c_G(\theta_G^*)]}$$

so by (C.2)

$$c(\varphi) = c(\varphi^*) + \sum_{\substack{G \in \mathcal{G} \\ q(G) \notin J}} y_{q(G)}[c_G(\theta_G) - c_G(\theta_G^*)]$$

Now (1.26) only determines the cumulant function up to an arbitrary constant, and here all of the starred parameters are constant, so this does agree with (1.36) up to an arbitrary constant (which is all it can do).

We have now established that (1.36) gives the cumulant function of the (unconditional, joint) distribution of the aster model when the parameter vectors in that formula are in the parameter space.

To prove the assertion of the theorem about when the cumulant function of the (unconditional, joint) distribution of the aster model is infinite, we need all the cases of (C.5) for $k = 1, \ldots, n$. Since θ^* and φ^* must be valid parameter vectors, $c_G(\theta_G^*)$ is always finite. In (C.5) it is unclear which G the first product runs over (it depends on the graph, or, alternatively, on the predecessor function), but we always know that $G_k \notin \bigcup \mathcal{G}_k$ because that is the way the total order on $\mathcal{G}$ works: $q(G_k)$ must either be an initial node or must be in some G_m with $k < m$. Thus case k of (C.5) tells us the expression is infinite if $c_{G_k}(\theta_{G_k}) = \infty$ and $Y_{q(G_k)}$ is not zero almost surely. But the former implies the latter cannot happen by the first sentence of the theorem statement: if $Y_{q(G_k)} = 0$ almost surely, then c_{G_k} is the zero function.

Putting these statements together for all k, we see that (C.5) is infinite whenever $\theta \notin \Theta$, where Θ is given by (C.1). Putting together everything we have proved so far, the cumulant function for the (unconditional, joint) distribution of the aster model is finite and given by (1.36) for θ in (C.1) and is infinite for θ not in (C.1).

Now letting f denote the aster transform as in the comments immediately preceding the theorem statement, we have also shown that the cumulant function for the (unconditional, joint) distribution of the aster model is finite and given by (1.36) for φ in $\Phi = f(\Theta)$ and is infinite for φ not in $\Phi = f(\Theta)$.

By assumption every Θ_G is an open set in the vector space containing it. Hence, a Cartesian product of open sets being an open set, Θ is an open set

in the vector space containing it. We know that the aster transform and its inverse are (infinitely) differentiable hence continuous. Hence for any $\varphi \in \Phi$ the point $\theta = f^{-1}(\varphi)$ is in the interior of Θ (because Θ is an open set), hence there is a neighborhood W of θ contained in Θ, but then $f(W)$ is a neighborhood of φ contained in Φ. Thus Φ is a neighborhood of each of its points, hence an open subset of the vector space containing it.

Thus the (unconditional, joint) distribution of the aster model is a regular full exponential family. □

The first sentence of the theorem statement is for limiting conditional models. An aster model in which no family is degenerate and no initial node is zero does not need this first sentence: we never have $Y_j = 0$ almost surely for any j. This is easily proved by induction, but we won't bother because limiting conditional models are a thing, so the theorem needs to be stated the way it is.

Theorem C.2. *If the saturated aster model is a regular full exponential family, then so is any unconditional canonical affine submodel.*

Conditions for the saturated model to be a regular full exponential family are the subject of Theorem C.1.

Proof. Applying (1.26) to the saturated model and a submodel give

$$\begin{aligned} c(\varphi) &= c(\varphi^*) + \log\left\{E_{\varphi^*}\left(e^{\langle Y, \varphi - \varphi^* \rangle}\right)\right\} \\ c_{\text{sub}}(\beta) &= c_{\text{sub}}(\beta^*) + \log\left\{E_{\beta^*}\left(e^{\langle M^T Y, \beta - \beta^* \rangle}\right)\right\} \end{aligned}$$

and

$$\begin{aligned} E_{\beta^*}\left(e^{\langle M^T Y, \beta - \beta^* \rangle}\right) &= E_{\beta^*}\left(e^{\langle Y, M(\beta - \beta^*) \rangle}\right) \\ &= E_{\beta^*}\left(e^{\langle Y, a + M\beta - (a + M\beta^*) \rangle}\right) \\ &= E_{\varphi^*}\left(e^{\langle Y, \varphi - \varphi^* \rangle}\right) \end{aligned}$$

where φ and φ^* are the saturated model unconditional canonical parameter vectors corresponding to β and β^*. Hence

$$c(\varphi) - c(\varphi^*) = c_{\text{sub}}(\beta) - c_{\text{sub}}(\beta^*)$$

so the two cumulant functions agree up to constants (which is all (1.26) can give us. In particular, they are both finite or both infinite for corresponding

arguments. Thus the full canonical parameter space for the unconditional canonical affine submodel is

$$B = \{ \beta \in \mathbb{R}^K : a + M\beta \in \Phi \},$$

where K is the index set of the parameter β and Φ is the full unconditional canonical parameter space of the saturated model characterized by Theorem C.1. If Φ is an open subset of $\mathbb{R}^J$, then B is an open subset of $\mathbb{R}^K$ because the pre-image of an open subset is open when the mapping is continuous, and every affine mapping is continuous. □

Theorem C.3. *If $n_{q(G)} = 0$ for any dependence group G, replace the family for this dependence group by the degenerate family concentrated at zero so c_G is the zero function. Then, if the saturated aster model is a regular full exponential family, then so is the associated independence model of a conditional aster model.*

Proof. This is just a combination of bits of earlier proofs. First, (1.55) has exponential family form (pretending the $n_{q(G)}$ therein are nonrandom). Denoting Θ_G and Θ as in (C.1), this is clearly the full canonical parameter space of this exponential family. And Θ is open because the Cartesian product of open sets is open.

We can also write the saturated associated independence model log likelihood (1.55) as

$$\begin{aligned} l(\theta) &= \left(\sum_{j \in J} y_j \theta_j \right) - \sum_{G \in \mathcal{G}} n_{q(G)} c_G(\theta_G) \\ &= \langle y, \theta \rangle - \tilde{c}(\theta) \end{aligned}$$

where

$$\tilde{c}(\theta) = \sum_{G \in \mathcal{G}} n_{q(G)} c_G(\theta_G)$$

so the submodel associated independence model log likelihood can be written

$$l(\beta) = \langle M^T y, \beta \rangle - \tilde{c}_{\text{sub}}(\beta)$$

where

$$\tilde{c}_{\text{sub}}(\beta) = \tilde{c}(a + M\beta)$$

so this (submodel) associated independence model is an exponential family and its full canonical parameter space is, by an argument similar to the preceding proof

$$B = \{ \beta \in \mathbb{R}^K : a + M\beta \in \Theta \}.$$

Since affine functions are continuous, B is an open subset of the vector space where β takes values. □

Appendix D

Families

D.1 Bernoulli

A random variable is *Bernoulli* if its possible values are zero and one. In other words, every Bernoulli random variable is zero-or-one-valued, and vice versa.

This is the *rationale* for the distribution, any dichotomous (two-valued) random variable can be coded as Bernoulli.

This is a *discrete* random variable.

This is a special case of the binomial distribution, which we do next.

D.2 Binomial

A random variable is *binomial* if it is the sum of IID Bernoulli random variables. Hence the Bernoulli distribution is the binomial distribution for sample size one (for one term in the sum).

The *probability mass function* is

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y = 0, 1, \ldots, n, \tag{D.1}$$

where p is the *usual parameter*, the probability that any of the n Bernoulli random variables in the sum is equal to one.

The *mean* and *variance* are

$$\begin{aligned} E(y) &= np \\ \operatorname{var}(y) &= np(1-p) \end{aligned}$$

This is an *exponential family*. From (D.1) the log likelihood is

$$l(\theta) = y \log(p) + (n - y) \log(1 - p) = y \cdot \log\left(\frac{p}{1 - p}\right) + n \log(1 - p)$$

from which we see that we have an exponential family with *canonical statistic* y and *canonical parameter*

$$\theta = \log\left(\frac{p}{1 - p}\right)$$

The right-hand side is so important that it is given a name. The logit function (pronounced low-jit) is given by

$$\operatorname{logit}(p) = \log\left(\frac{p}{1 - p}\right), \qquad 0 < p < 1.$$

Its inverse function is

$$\operatorname{logit}^{-1}(\theta) = \frac{e^\theta}{1 + e^\theta}, \qquad -\infty < \theta < \infty.$$

The *cumulant function* is

$$\begin{aligned} c(\theta) &= -n \log(1 - p) \\ &= -n \log\left(1 - \frac{e^\theta}{1 + e^\theta}\right) \\ &= -n \log\left(\frac{1}{1 + e^\theta}\right) \\ &= n \log\left(1 + e^\theta\right) \end{aligned}$$

Note that, as required for any sum of IID random variables, the cumulant function for sample size n is n times the cumulant function for sample size one (Section 1.15.2 above).

We check that this has the correct derivatives

$$c'(\theta) = \frac{ne^\theta}{1 + e^\theta} = np$$

and

$$c''(\theta) = \frac{ne^\theta}{1 + e^\theta} - \frac{ne^\theta e^\theta}{(1 + e^\theta)^2} = \frac{ne^\theta}{1 + e^\theta}\left[1 - \frac{e^\theta}{1 + e^\theta}\right] = np(1 - p)$$

The *mean value parameter* is $\xi = np$.

The *canonical parameter space* is the range of the logit function, which is the whole real line, $-\infty < \theta < \infty$.

The *mean value parameter space* is n times the domain of the logit function $0 < \xi < n$.

Theorem 2.1 says limiting conditional models are conditioned on the boundary of the closed convex support. The closed convex support is the closed interval $[0, n]$, and its boundary consists of two points 0 and n.

Thus there are two limiting conditional models, one of which contains only the distribution concentrated at zero and one of which contains only the distribution concentrated at n.

In one-dimensional space there are only two directions. Every positive vector points in the same direction and gives the same LCM. Every negative vector points in the same direction and gives the same LCM. (And, of course, the zero vector points in no direction and gives the original model back as the LCM corresponding to it.)

As discussed in Theorem 2.2 above and its following comments, it is important that we use (2.6) to determine the cumulant function for the LCM.

So

$$\begin{aligned} c_{-1}(\theta) &= c(\theta) + \log \Pr\nolimits_\theta(Y = 0) \\ &= n \log\left(1 + e^\theta\right) + n \log(1 - p) \\ &= n \log\left(1 + e^\theta\right) + n \log\left(\frac{1}{1 + e^\theta}\right) \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} c_{+1}(\theta) &= c(\theta) + \log \Pr\nolimits_\theta(Y = n) \\ &= n \log\left(1 + e^\theta\right) + n \log(p) \\ &= n \log\left(1 + e^\theta\right) + n \log\left(\frac{e^\theta}{1 + e^\theta}\right) \\ &= n\theta \end{aligned}$$

Addition rule: the sum of m independent and identically distributed binomial random variables with sample size n and usual parameter p has the binomial distribution with sample size mn and usual parameter p.

Hence if $y_{p(j)} \longrightarrow y_j$ is a binomial arrow for sample size n the conditional distribution of y_j given $y_{p(j)}$ is binomial for sample size $ny_{p(j)}$.

This family is not implemented in either R package `aster` or R package `aster2`. Only the $n = 1$ special case, the Bernoulli family is implemented.

D.3 Poisson

A random variable is *Poisson* if it has the *probability mass function*

$$f(y) = \frac{\xi^y}{y!} e^{-\xi}, \qquad y = 0, 1, 2, \ldots, \tag{D.2}$$

where ξ is the *usual parameter*, which turns out to be the mean and variance of the distribution, hence also the mean-value parameter.

This is a *discrete* random variable.

There are two rationales for this distribution, both so closely related that they are almost one rationale. First, the Poisson distribution is an approximation to the binomial(n, p) distribution when n is very large and p is very small and the mean np is moderate sized. An example is a lottery. Every week millions, sometimes hundreds of millions of tickets are sold (that's n), the probability of any one ticket winning is very small — for example, for the Powerball lottery, the probability is one over 292,201,338 (as we write this, the rules change from time to time) — (that's p), and np is moderate sized. In weeks where the jackpot is small and few tickets are sold, there are still tens of millions of tickets sold, so np is less than one but not very small. In weeks where the jackpot is large, there may be many hundreds of millions of tickets sold, so np is greater than one and multiple winners are expected (they split the jackpot among them). But regardless, the distribution of the number of winners is well approximated by the Poisson(np) distribution.

Before we can discuss the second rationale, we discuss the *addition rule:* the sum of independent Poisson random variables is again Poisson. It is not required that the independent Poisson random variables be identically distributed. Since the expectation of a sum is the sum of the expectations, the sum of independent Poisson random variables having means $\xi_1, \ldots, \xi_n$ has the Poisson$(\xi_1 + \cdots + \xi_n)$ distribution.

It follows (not obviously, but the derivation can be found in books about spatial point processes) that the sum of n independent Bernoulli random variables is well approximated by a Poisson distribution provided n is very large and the means of all of the Bernoulli random variables are very much smaller than the mean of the Poisson random variable. Again, if the means

of the Bernoulli random variables are $\xi_1, \ldots, \xi_n$, then the mean of the Poisson random variable is $\xi_1 + \cdots + \xi_n$. So we are assuming that each ξ_i is very much smaller than the sum. To return to our lottery example, it does not matter that each player is playing the same game. So long as the expectation of any one ticket winning is negligible compared to the expected number of winners (for all tickets), the distribution of the number of winners will be approximately Poisson.

Let's take a biological example. Suppose we are counting ants, and we have divided up the region in which we are counting ants with a very fine grid. If our grid is fine enough, the probability of counting more than one ant in a grid cell will be negligible, perhaps impossible (if our grid cells are so small that more than one ant could not fit). Then the number of ants in any one cell is a Bernoulli (zero-or-one-valued) random variable, and the number of ants in any region that contains a very large number of grid cells is very well approximated by the Poisson distribution. If we take the limit as the size of the grid cells goes to zero we get exact Poisson distributions. Except that we forgot to mention independence. This assumes the Bernoulli random variables are independent, that where one ant is has nothing whatsoever to do with where any other ant is. If we can accept this independence assumption, then the count of ants in any region of any size large enough to have a moderate sized expected number of ants can be assumed Poisson.

Now we abstract away from ants to be counting any things in regions of any dimension. The number of stars visible to the naked eye in a region of sky, the number of raisins in slice of carrot cake, the number of white blood cells in a drop of blood on a microscope slide, the number of ants in a square meter region of your back yard, the number of leaves on a tree, the number of calls arriving at a call center in a specified time interval, and many other things can be assumed Poisson.

The independence assumption is crucial. Pheromone trails and perhaps other phenomena may make our counts of ants noticeably non-Poisson. But if it can be plausibly asserted that the probability of any one thing being counted is independent of all the other things counted or not counted, then the distribution of the total count is Poisson.

And even if the distribution of a count random variable fails to be exactly Poisson due to some failure of the independence assumption, the Poisson distribution may still may be a pretty good approximation (or may fail badly if the independence assumption is grossly wrong).

As stated above, the *mean* and *variance* are

$$E(y) = \xi$$
$$\operatorname{var}(y) = \xi$$

This is an *exponential family*. From (D.2) the log likelihood is

$$l(\theta) = y \log(\xi) - \xi$$

(the term $\log(y!)$ can be dropped because it does not contain the parameter), from which we see that we have an exponential family with *canonical statistic* y and *canonical parameter*

$$\theta = \log(\xi),$$

so

$$\xi = e^\theta.$$

The *cumulant function* is

$$c(\theta) = \xi = e^\theta$$

We check that this has the correct derivatives (and this is trivial)

$$c'(\theta) = e^\theta = \xi$$

and

$$c''(\theta) = e^\theta = \xi$$

The *mean value parameter* is also the usual parameter ξ.

The *canonical parameter space* is the range of the log function, which is the whole real line, $-\infty < \theta < \infty$.

The *mean value parameter space* is the domain of the log function $0 < \xi < \infty$.

Thinning rule: in the following graph

$$y_1 \xrightarrow{\text{Poi}} y_2 \xrightarrow{\text{Ber}} y_3$$

the conditional distribution of y_3 given y_1 (both arrows combined) is Poisson($\xi_3\xi_2$). A thinned Poisson process is another Poisson process, where "thinning" means we take each "point" counted and accept or reject it independently with the same probability.

As discussed at the end of the preceding section, LCM are conditioned on the boundary of the closed convex support. The closed convex support is the closed interval $[0, \infty)$, and its boundary consists of the single point 0.

Thus there is one limiting conditional model, which contains only the distribution concentrated at zero.

Also as discussed at the end of the preceding section, it is important that we use (2.6) to determine the cumulant function for the LCM. So

$$\begin{aligned} c_{-1}(\theta) &= c(\theta) + \log \Pr{}_\theta(Y = 0) \\ &= e^\theta + \log(e^{-\xi}) \\ &= e^\theta - \xi \\ &= 0 \end{aligned}$$

So, again as in the preceding section, the cumulant function for the LCM concentrated at zero is the zero function.

As mentioned in Section 1.17 above, the Poisson distribution is infinitely divisible. This is easily verified from its cumulant function. For any positive real number r

$$rc(\theta) = re^\theta = e^{\theta + \log(r)}$$

is a cumulant function. In fact, it is a cumulant function for the Poisson family. One log likelihood for the Poisson family is

$$l(\theta) = y\theta - e^\theta$$

but if we make the substitution $\theta = \psi + \log(r)$ we get

$$l(\psi) = y\psi + y \log(r) - e^{\psi + \log(r)}$$

and we can drop the term that does not contain the new parameter ψ obtaining

$$l(\psi) = y\psi - e^{\psi + \log(r)}$$

and we see this has exponential family form with canonical statistic y, canonical parameter ψ, and cumulant function $c(\psi) = e^{\psi + \log(r)}$.

This is just a special case of the fact, noted without proof in Section 1.15.1, that adding a constant to a canonical parameter gives another canonical parameter.

Another way of thinking about this fact is that our new parameterization just puts an offset $\log(r)$ in the exponential family. But we know from Section 1.15.3 above that canonical affine submodels of full exponential families are again exponential families.

Note that in going from Section D.1 to Section D.2 we just went from the family having cumulant function $c(\theta) = 1 + e^\theta$ to the family having

cumulant function $nc(\theta)$, something we know from Section 1.15.2 above is always valid. So we might think that we would need another section to go from the family having cumulant function $c(\theta) = e^\theta$ to the family having cumulant function $rc(\theta)$, which is valid only when the family is infinitely divisible. But we have just found that that does not give us a new family, but rather the same old Poisson family (with an offset), so we do not need a new section for a new family.

D.4 Zero-Truncated Poisson

The *zero-truncated Poisson* distribution is the Poisson distribution conditioned on being nonzero.

The *rationale* is that it can be used to incorporate zero-inflated Poisson random variables into aster models.

This is a *discrete* distribution.

If f is the PMF of the Poisson distribution, then the PMF of the zero-truncated Poisson distribution is

$$g(y) = \frac{f(y)}{1 - f(0)}, \qquad y = 1, 2, \ldots, \tag{D.3}$$

that is, if m is the mean of the untruncated Poisson distribution, then the PDF of the zero-truncated Poisson distribution is

$$g(y) = \frac{m^y e^{-m}}{y!(1 - e^{-m})}, \qquad y = 1, 2, \ldots. \tag{D.4}$$

Since this is not a "brand name distribution" the mean and variance cannot just be looked up. In aid of this calculation we prove a rather trivial general theorem.

Theorem D.1. *Suppose X is a nonnegative-integer-valued random variable, and Y is the corresponding zero-truncated random variable. Then*

$$E(Y^k) = E(X^k)/\Pr(X > 0)$$

for any positive integer k.

Proof. For this proof let f denote the PMF of X and g the PMF of Y, so the relationship between the two is given by (D.3) even though we are no

longer assuming X is Poisson. Then

$$\begin{aligned} E(Y^k) &= \sum_{y=1}^{\infty} y^k g(y) \\ &= \frac{1}{1 - f(0)} \sum_{x=1}^{\infty} x^k f(x) \\ &= \frac{1}{1 - f(0)} \sum_{x=0}^{\infty} x^k f(x) \\ &= \frac{E(X)}{1 - f(0)} \\ &= \frac{E(X)}{\Pr(X > 0)} \end{aligned}$$

where the third equality is the fact that the $x = 0$ term in the sum is equal to zero. □

Together with

$$\begin{aligned} \operatorname{var}(Y) &= E(Y^2) - E(Y)^2 \\ E(Y^2) &= \operatorname{var}(Y) + E(Y)^2 \end{aligned}$$

which are well known from elementary probability theory, we can use the theorem to calculate the mean and variance of zero-truncated random variables.

For the Poisson distribution, we have $E(X) = \operatorname{var}(X) = m$ so $E(X^2) = m + m^2$, so

$$E(Y) = \frac{E(X)}{\Pr(X > 0)} = \frac{m}{1 - e^{-m}} \tag{D.5}$$

and

$$\begin{aligned} \operatorname{var}(Y) &= E(Y^2) - E(Y)^2 \\ &= \frac{E(X^2)}{\Pr(X > 0)} - \left(\frac{E(X)}{\Pr(X > 0)} \right)^2 \\ &= \frac{m + m^2}{1 - e^{-m}} - \left(\frac{m}{1 - e^{-m}} \right)^2 \end{aligned} \tag{D.6}$$

This is an *exponential family*. From (D.4) the log likelihood is

$$l(\theta) = y \log(m) - m - \log(1 - e^{-m})$$

(the term $\log(y!)$ can be dropped because it does not contain the parameter), from which we see that we have an exponential family with *canonical statistic* y and *canonical parameter*

$$\theta = \log(m),$$

so

$$m = e^\theta,$$

the relation between θ and m being the same as for the Poisson distribution.

But the *usual parameter* m is not the *mean value parameter*, which is

$$\xi = \frac{m}{1 - f(0)} = \frac{m}{1 - e^{-m}} = \frac{\exp(\theta)}{1 - \exp(-\exp(\theta))} \tag{D.7}$$

as we know from general exponential family theory, the mapping $\theta \mapsto \xi$ given by the formula above is strictly increasing and invertible and both it and its inverse mapping $\xi \to \theta$ are infinitely differentiable. But in this case the inverse mapping $\xi \to \theta$ seems to have no closed-form expression. The map $\xi \to \theta$ is what is called a *link function* in the terminology of generalized linear models (GLM). The failure of some families to have link functions in useful form is one reason why aster model theory and practice never mentions link functions. They make sense for some families but not others.

The *cumulant function* is

$$c(\theta) = m + \log(1 - e^{-m}) = e^\theta + \log(1 - \exp(-\exp(\theta))) \tag{D.8}$$

We check that this has the correct derivatives

$$\begin{aligned} c'(\theta) &= e^\theta + \frac{\exp(-\exp(\theta))\exp(\theta)}{1 - \exp(-\exp(\theta))} \\ &= m + \frac{me^{-m}}{1 - e^{-m}} \\ &= \frac{m}{1 - e^{-m}} \end{aligned}$$

and

$$\begin{aligned} c''(\theta) &= e^\theta + \frac{\exp(-\exp(\theta))\exp(\theta)}{1 - \exp(-\exp(\theta))} - \frac{\exp(-\exp(\theta))\exp(\theta)^2}{1 - \exp(-\exp(\theta))} \\ &\quad - \frac{\exp(-\exp(\theta))^2\exp(\theta)^2}{(1 - \exp(-\exp(\theta)))^2} \\ &= m + \frac{me^{-m}}{1 - e^{-m}} - \frac{m^2e^{-m}}{1 - e^{-m}} - \frac{m^2e^{-2m}}{(1 - e^{-m})^2} \end{aligned}$$

and this does simplify to be equal to our other expression for variance.

Two other formulas for the variance are also useful (Geyer, 2017c).

$$\begin{aligned} \operatorname{var}(y) &= \xi(1 + m - \xi) && \text{(D.9a)} \\ &= \xi(1 - \xi e^{-m}) && \text{(D.9b)} \end{aligned}$$

As $\theta \to -\infty$ and $m \to 0$ the mean value parameter ξ converges (using L'Hospital's rule) to

$$\lim_{m \to 0} \frac{m}{1 - e^{-m}} = \lim_{m \to 0} \frac{1}{e^m} = 1$$

and (D.9a) shows the variance converges to zero as $m \to 0$ and $\xi \to 1$. As $\theta \to \infty$ and $m \to \infty$ the mean value parameter ξ is approximately equal to m because $f(0) = e^{-m}$ is approximately zero. Then ξe^{-m} is small compared to one, and and (D.9b) shows the variance is also approximately equal to $\xi \approx m$.

As we said above, the *mean value parameter* ξ is not the usual parameter m.

As can be seen from that fact that (D.8) is finite for all θ, the *canonical parameter space* is the whole real line, $-\infty < \theta < \infty$.

As we saw when discussing variance formulas, $\xi \to 1$ as $m \to 0$. Thus the lower end of the mean value parameter space is one. And from m being the mean of a Poisson distribution so m has no upper bound, and from $m \approx \xi$ when either is large, we see that ξ also has no upper bound. Thus the *mean value parameter space* is $1 < \xi < \infty$.

As discussed at the end of the two preceding sections, LCM are conditioned on the boundary of the closed convex support. The closed convex support is the closed interval $[1, \infty)$, and its boundary consists of the single point 1.

Thus there is one limiting conditional model, which contains only the distribution concentrated at one.

Also as discussed at the end of the two preceding sections, it is important that we use (2.6) to determine the cumulant function for the LCM. So

$$\begin{aligned} c_{-1}(\theta) &= c(\theta) + \log \Pr\nolimits_\theta(Y = 1) \\ &= m + \log(1 - e^{-m}) + \log\left(\frac{m e^{-m}}{1 - e^{-m}}\right) \\ &= \log(m) \\ &= \theta \end{aligned}$$

So the cumulant function of the distribution concentrated at one is the identity function.

This just happens to agree with the $n = 1$ case for the binomial distribution (Section D.2 above), but it need not have. It all depends on how we defined the cumulant functions for these families in the first place. We could have added different arbitrary constants to the cumulant functions of these families and they would still be cumulant functions.

D.5 Normal Location

The univariate normal distribution has *probability density function* (PDF)

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\xi)^2}{2\sigma^2}}, \qquad -\infty < y < \infty. \tag{D.10}$$

This is a *continuous* random variable; except when incorporated into an aster model, it is a mixture of discrete and continuous. For a normal-location arrow, when the predecessor is zero the conditional distribution of the successor is the degenerate random variable concentrated at zero, which is discrete, and when the predecessor is greater than zero, the conditional distribution of the successor is continuous.

The *rationale* is the celebrated central limit theorem, or more precisely, theorems, because there are many variants. In non-technical terms these theorems say that a random variable that is the sum of a large number of random variables that are not too dependent, not too heavy tailed, and not too unequal in size will be well approximated by a normal distribution. (If the random variable in question is the sum of a large number of independent random variables, then Lindeberg's central limit theorem using Lindeberg's condition specifies what "not too heavy tailed, and not too unequal in size" means. If the random variable in question is the sum of the components of a dependent stochastic process, then various stationary process central limit theorems, Markov chain central limit theorems, and the martingale central limit theorem, give various notions of what "not too dependent" means.)

> Everybody believes in the law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact.
>
> — Lippman, quoted by Poincaré, quoted by Cramér (1951)

"The law of errors" is an old name for the normal distribution. It has also been named after de Moivre, Laplace, and Gauss. The term "normal

distribution" was popularized by K. Pearson in the early twentieth century. Like the term "law of errors" it builds into the name the idea that it is the main, principle, or only distribution for random data. Also note the Lippman quote is sarcastic. Justification for this belief was always known to be shaky. (Harald Cramér and Henri Poincaré are, of course, famous. It is unclear who the Monsieur Lippman was that Poincaré attributed this to.)

Since the nonparametrics revolution (Hollander et al., 2014), the exploratory data analysis revolution (Tukey, 1977), the bootstrap revolution (Efron and Tibshirani, 1993; Davison and Hinkley, 1997), and the robustness revolution (Huber and Ronchetti, 2009; Hampel et al., 1986) no user of statistics aware of these developments wants to blindly assume normality, especially when it can be demonstrated to be grossly incorrect using any of these tools. But the normal distribution may fit data well, so it continues to be used. It just is no longer considered the only distribution for data, as it was before 1950 (mostly, there was the chi-square test for contingency tables).

The other rationale for this distribution (which has nothing to do with aster models) is that the usual assumption of homoscedastic normal errors for linear models makes the distribution of point estimates exactly normal and the distribution of various test statistics exactly t or exactly F. This rationale is often attributed to Gauss and is why the normal distribution is sometimes called Gaussian, because Gauss independently co-invented the method of least squares and more-or-less gave this rationale (more-or-less because his discussion was Bayesian rather than frequentist), but of course this was a century before the t and F distributions were invented.

The *mean* and *variance* are

$$\begin{aligned} E(y) &= \xi \\ \operatorname{var}(y) &= \sigma^2 \end{aligned}$$

When used in R package `aster` every family must be a one-parameter exponential family of distributions, so when we consider this as such a family we must pick one parameter to be treated as unknown and the other parameter to be treated as known. Because the location parameter ξ is the mean value parameter, we pick this to be the unknown parameter.

With this understanding, the log likelihood is

$$l(\theta) = -\frac{(y - \xi)^2}{2\sigma^2}$$

(the term $\sqrt{2\pi}\sigma$ can be dropped because it does not contain the unknown

parameter ξ. If we expand the quadratic, we get

$$l(\theta) = -\frac{y^2}{2\sigma^2} + \frac{y\xi}{\sigma^2} - \frac{\xi^2}{2\sigma^2}$$

and can now drop another term not containing ξ obtaining

$$l(\theta) = \frac{y\xi}{\sigma^2} - \frac{\xi^2}{2\sigma^2}$$

from which we see that we have an *exponential family* with *canonical statistic* y and *canonical parameter*

$$\theta = \frac{\xi}{\sigma^2}$$

so the *mean value parameter* is

$$\xi = \sigma^2\theta$$

The *cumulant function* is

$$c(\theta) = \frac{\xi^2}{2\sigma^2} = \frac{\sigma^2\theta^2}{2}$$

We check that this has the correct derivatives

$$c'(\theta) = \sigma^2\theta = \xi$$

and

$$c''(\theta) = \sigma^2$$

Addition rule: the sum of n independent and identically distributed normal random variables with mean ξ and variance σ^2 has the normal distribution with mean $n\xi$ and variance $n\sigma^2$.

General Addition rule: any sum of independent normal random variables is again normal (identically distributed is not required), but this has no application in aster model theory.

There are no limit degenerate distributions. This is because the boundary of the closed convex support, which is the interval $(-\infty, +\infty)$ is empty. We can never observe data on the boundary.

D.6 Negative Binomial

D.6.1 Basics

According to the `help("NegBinomial")` in R, the negative binomial distribution has *probability mass function*

$$f(y) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\, y!} p^\alpha (1 - p)^y, \qquad y = 0, 1, 2, \ldots. \tag{D.11}$$

where $\alpha > 0$ is the shape parameter and $0 < p \leq 1$ is the usual parameter (success probability). The case $\alpha = 1$ is the geometric distribution.

The *first rationale* for this distribution is inverse sampling, and for this rationale α must be a positive integer. If one has an infinite sequence of IID Bernoulli random variables with usual parameter p, then the distribution of the number of observed zero outcomes before the α-th nonzero outcome is negative binomial with shape parameter α and usual parameter p, that is, if one observes y successes in n trials, then the distribution of y is binomial if n was fixed and the distribution of $n - y$ is negative binomial if y was fixed. But this rationale has nothing to do with aster models.

The *second rationale* for this distribution is overdispersed Poisson. This distribution arises as a mixture of Poisson distributions, as is discussed below (Section D.6.2). This is the reason it is implemented in R package `aster`. For this rationale α can be any positive real number.

The *mean* and *variance* in terms of these parameters are

$$E(y) = \frac{\alpha(1 - p)}{p}$$
$$\operatorname{var}(y) = \frac{\alpha(1 - p)}{p^2}$$

From (D.11) the log likelihood is

$$l(\theta) = \log \Gamma(\alpha + y) - \log \Gamma(\alpha) - \log(y!) + \alpha \log(p) + y \log(1 - p).$$

from which we can see that if α is considered an unknown parameter, this is *not* an exponential family, so we consider α known, which means we can drop terms not containing p obtaining

$$l(\theta) = y \log(1 - p) + \alpha \log(p)$$

from which we see that we have an *exponential family* with *canonical statistic* y and *canonical parameter*

$$\theta = \log(1 - p)$$

and solving for p gives

$$1 - p = e^\theta$$

and

$$p = 1 - e^\theta$$

The *cumulant function* is

$$c(\theta) = -\alpha \log(p) = -\alpha \log(1 - e^\theta)$$

As p goes from zero to one, θ goes from zero to $-\infty$ so the formula above does not define the cumulant function on the whole real line and equation (5) in Geyer (2009), which is (1.26) in this book, must be used

$$c(\theta) = c(\psi) + \log E_\psi\bigl(e^{y(\theta-\psi)}\bigr)$$

where ψ is a fixed canonical parameter value, θ varies over the whole real line, and the cumulant function has the value ∞ where the expectation does not exist.

Evaluating this we get, using the theorem associated with the negative binomial distribution (Geyer, 2019b),

$$\begin{aligned} c(\theta) &= c(\psi) + \log\left(\sum_{y=0}^{\infty} e^{y(\theta-\psi)} \cdot \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\, y!} p^\alpha (1-p)^y\right) \\ &= c(\psi) + \log\left(p^\alpha \sum_{y=0}^{\infty} \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\, y!} \left[(1-p)e^{\theta-\psi}\right]^y\right) \\ &= c(\psi) + \log\left(p^\alpha \left[1 - (1-p)e^{\theta-\psi}\right]^{-\alpha}\right) \end{aligned} \tag{D.12}$$

where p is the usual parameter value corresponding to canonical parameter value ψ, that is, $\psi = \log(1-p)$ and the formula is only valid when the infinite sequence converges, which it does if and only if $-1 < (1-p)e^{\theta-\psi} < +1$.

Now $1 - p = e^\psi$, so we can simplify $(1-p)e^{\theta-\psi} = e^\theta$. So the convergence criterion is $e^\theta < 1$ or $\theta < 0$, and the formula simplifies to

$$c(\theta) = c(\psi) + \alpha \log(p) - \alpha \log\left(1 - e^\theta\right)$$

but the formula only determines the cumulant function up to an arbitrary constant (which does not matter) so we can take the cumulant function to be

$$c(\theta) = \begin{cases} -\alpha \log(1 - e^\theta), & \theta < 0 \\ \infty, & \theta \geq 0 \end{cases} \tag{D.13}$$

So the full canonical parameter space is, as we guessed before,

$$\Theta = \{ \theta \in \mathbb{R} : \theta < 0 \} \tag{D.14}$$

and (D.13) agrees with what we derived just from looking at the log likelihood wherever the function is finite.

Let's check that this cumulant function gives the correct mean and variance.

$$\begin{aligned} c'(\theta) &= \frac{\alpha e^\theta}{1 - e^\theta} \\ &= \frac{\alpha(1-p)}{p} \\ c''(\theta) &= \frac{d}{d\theta} \frac{\alpha e^\theta}{1 - e^\theta} \\ &= \frac{\alpha e^\theta}{1 - e^\theta} - \frac{\alpha e^{2\theta}}{(1 - e^\theta)^2} \\ &= \frac{\alpha(1-p)}{p} \left(1 - \frac{1-p}{p} \right) \\ &= \frac{\alpha(1-p)}{p^2} \end{aligned}$$

as we had already been told but now have derived from exponential family theory.

The *mean value parameter*

$$\xi = \frac{\alpha(1-p)}{p} \tag{D.15}$$

is not the usual parameter p. Solving for p gives

$$p = \frac{\alpha}{\alpha + \xi} \tag{D.16}$$

The *canonical parameter space* is (D.14) which is not the whole real line.

The *mean value parameter space* is the range of the derivative of the cumulant function $0 < \xi < \infty$.

As discussed at the end of the Sections D.2, D.3, and D.4, LCM are conditioned on the boundary of the closed convex support. The closed convex support is the closed interval $[0, \infty)$, and its boundary consists of the single point 0.

Thus there is one limiting conditional model, which contains only the distribution concentrated at zero.

Also as discussed at the end of the those sections, it is important that we use (2.6) to determine the cumulant function for the LCM. So

$$\begin{aligned} c_{-1}(\theta) &= c(\theta) + \log \Pr\nolimits_\theta(Y = 0) \\ &= -\alpha \log(1 - e^\theta) + \log\left(p^\alpha\right) \\ &= 0 \end{aligned}$$

So the cumulant function of the family concentrated at zero is the zero function, as we also found in Sections D.2 and D.3 above. But as mentioned at the end of Section D.4 above, this agreement just happened because of arbitrary choices of arbitrary constants in cumulant functions.

D.6.2 Negative Binomial as Mixture of Poisson

As stated above, one rationale for the negative binomial distribution is that it is a mixture of Poisson distributions. Let the conditional distribution of Y given μ be Poisson with mean μ, and let the marginal distribution of μ be Gamma(α, λ). Then the marginal distribution of Y is given by

$$\begin{aligned} f(y) &= \int f(y \mid \mu) g(\mu)\, d\mu \\ &= \int_0^\infty \frac{\mu^y e^{-\mu}}{y!} \cdot \frac{\lambda^\alpha}{\Gamma(\alpha)} \mu^{\alpha - 1} e^{-\lambda\mu}\, d\mu \\ &= \frac{1}{y!} \cdot \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \mu^{y + \alpha - 1} e^{-(1 + \lambda)\mu}\, d\mu \\ &= \frac{1}{y!} \cdot \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(y + \alpha)}{(1 + \lambda)^{y + \alpha}} \end{aligned}$$

using the theorem associated with the gamma distribution (Geyer, 2019b).

For this to be equal to (D.11) we need

$$\frac{1}{y!} \cdot \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(y + \alpha)}{(1 + \lambda)^{y + \alpha}} = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\, y!} p^\alpha (1 - p)^y$$

that is

$$\frac{\lambda^\alpha}{(1 + \lambda)^{y + \alpha}} = p^\alpha (1 - p)^y$$

or

$$\left(\frac{\lambda}{1 + \lambda}\right)^\alpha \left(\frac{1}{1 + \lambda}\right)^y = p^\alpha (1 - p)^y$$

which happens if and only if $p = \lambda/(1 + \lambda)$ and $1 - p = 1/(1 + \lambda)$ so $\lambda = p/(1 - p)$.

D.6.3 Poisson as Limit of Negative Binomial

Reparameterize the negative binomial distribution so the parameters are α and ξ so the usual parameter is (D.16) and the PMF (D.11) becomes

$$\begin{aligned} f(y) &= \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\, y!} \left(\frac{\alpha}{\alpha + \xi} \right)^\alpha \left(\frac{\xi}{\alpha + \xi} \right)^y \\ &= \frac{1}{y!} \left(\frac{\alpha}{\alpha + \xi} \right)^\alpha \left(\frac{\xi}{\alpha + \xi} \right)^y \prod_{k=1}^{y} (\alpha + k - 1) \\ &= \frac{\xi^y}{y!} \left(1 - \frac{\xi}{\alpha + \xi} \right)^\alpha \prod_{k=1}^{y} \frac{\alpha + k - 1}{\alpha + \xi} \\ &\to \frac{\xi^y}{y!} e^{-\xi} \end{aligned}$$

so as $\alpha \to \infty$ with ξ fixed, we recover the Poisson distribution. But this does not happen if we let $\alpha \to \infty$ with some other parameter, such as p or θ, fixed. (The limit $\left(1 - \frac{\xi}{\alpha+\xi}\right)^\alpha \to e^{-\xi}$ as $\alpha \to \infty$, which was used in the derivation above, follows from $(1 + x/n)^n \to e^x$ as $n \to \infty$, which can be found in any calculus book.)

The upshot of this section is that if the shape parameter α of a negative binomial distribution is large, then it is well approximated by a Poisson distribution. One only needs the negative binomial family when the shape parameter is small.

The limit in this section is not like the limits in other sections of this appendix. In those other sections we took limits as the canonical parameter of the exponential family went to plus or minus infinity. Since the canonical parameter is considered unknown, this kind of limit can arise in the process of maximum likelihood. In this section we took a limit as the shape parameter α went to infinity. Since this parameter is considered known, this kind of limit cannot arise in the process of maximum likelihood.

D.7 Zero-Truncated Negative Binomial

This section is just like Section D.4 above *mutatis mutandis*.

The *zero-truncated negative binomial* distribution is the negative binomial distribution conditioned on being nonzero.

The *rationale* is that it can be used to incorporate zero-inflated negative binomial random variables into aster models.

This is a *discrete* distribution.

If f is the PMF of the negative binomial distribution, then the PMF of the zero-truncated negative binomial distribution is

$$g(y) = \frac{f(y)}{1 - f(0)}, \qquad y = 1, 2, \ldots, \tag{D.17}$$

that is, if α is the shape parameter and p is the usual parameter of the untruncated negative binomial distribution, then the PDF of the zero-truncated negative binomial distribution is

$$g(y) = \frac{\Gamma(\alpha + y) p^\alpha (1 - p)^y}{\Gamma(\alpha)\, y!\, (1 - p^\alpha)}, \qquad y = 1, 2, \ldots. \tag{D.18}$$

Since this is not a "brand name distribution" the mean and variance cannot just be looked up. We still use Theorem D.1 and the comment following it to calculate the mean and the variance but now

$$\begin{aligned} E(X) &= \frac{\alpha(1 - p)}{p} \\ \operatorname{var}(X) &= \frac{\alpha(1 - p)}{p^2} \\ E(X^2) &= \frac{\alpha(1 - p)}{p^2} + \left(\frac{\alpha(1 - p)}{p} \right)^2 \\ \Pr(X > 0) &= 1 - p^\alpha \end{aligned}$$

so

$$E(Y) = \frac{\alpha(1 - p)}{p(1 - p^\alpha)}$$

and

$$\begin{aligned} \operatorname{var}(Y) &= E(Y^2) - E(Y)^2 \\ &= \frac{E(X^2)}{\Pr(X > 0)} - \left(\frac{E(X)}{\Pr(X > 0)} \right)^2 \\ &= \frac{1}{1 - p^\alpha} \left[\frac{\alpha(1 - p)}{p^2} + \left(\frac{\alpha(1 - p)}{p} \right)^2 \right] - \left(\frac{\alpha(1 - p)}{p(1 - p^\alpha)} \right)^2 \end{aligned}$$

With the assumptions of Section D.6 above (α known and p unknown) this is an exponential family. From (D.18) the log likelihood is

$$l(\theta) = \log \Gamma(\alpha + y) + \alpha \log(p) + y \log(1 - p) - \log \Gamma(\alpha) - \log(y!) - \log(1 - p^\alpha)$$

and we may drop terms that do not contain the unknown parameter p obtaining

$$l(\theta) = y \log(1 - p) + \alpha \log(p) - \log(1 - p^\alpha)$$

from which we see that we have an *exponential family* with *canonical statistic* y and *canonical parameter*

$$\theta = \log(1 - p)$$

(the same as for the untruncated negative binomial distribution) and solving for p gives

$$p = 1 - e^\theta$$

(the same as for the untruncated negative binomial distribution).

The *cumulant function* is

$$c(\theta) = -\alpha \log(p) + \log(1 - p^\alpha) = -\alpha \log(1 - e^\theta) + \log(1 - (1 - e^\theta)^\alpha)$$

As in Section D.6 this does not define the cumulant function on the whole real line so we use

$$\begin{aligned}
c(\theta) &= c(\psi) + \log\left(\sum_{y=1}^{\infty} e^{y(\theta-\psi)} \cdot \frac{\Gamma(\alpha + y) p^\alpha (1 - p)^y}{\Gamma(\alpha)\, y!\, (1 - p^\alpha)}\right) \\
&= c(\psi) + \log\left(\frac{p^\alpha}{1 - p^\alpha} \sum_{y=1}^{\infty} \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\, y!} \left[(1 - p) e^{\theta-\psi}\right]^y\right) \\
&= c(\psi) + \log\left(\frac{p^\alpha}{1 - p^\alpha} \left\{-1 + \sum_{y=0}^{\infty} \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\, y!} \left[(1 - p) e^{\theta-\psi}\right]^y\right\}\right) \\
&= c(\psi) + \log\left(\frac{p^\alpha}{1 - p^\alpha} \left\{-1 + \left[1 - (1 - p) e^{\theta-\psi}\right]^{-\alpha}\right\}\right)
\end{aligned}$$

where the last equality is the theorem associated with the negative binomial distribution (Geyer, 2019b) and where, as in (D.12), p is the usual parameter that goes with ψ not θ so p is a known constant and $(1 - p) e^{\theta-\psi} = e^\theta$ and the infinite series converges if and only if $\theta < 0$. Thus our formula simplifies to

$$c(\theta) = c(\psi) + \log\left(\frac{p^\alpha}{1 - p^\alpha}\right) + \log\left(-1 + (1 - e^\theta)^{-\alpha}\right)$$

and we may drop the terms that do not contain θ obtaining

$$c(\theta) = \begin{cases} \log\left((1 - e^\theta)^{-\alpha} - 1\right), & \theta < 0 \\ \infty, & \theta \geq 0 \end{cases}$$

With some rearrangement, this agrees with what we deduced from looking at the log likelihood.

Let's check that this cumulant function gives the correct mean and variance.

$$\begin{aligned}
c'(\theta) &= \frac{-\alpha(1 - e^\theta)^{-\alpha-1}(-e^\theta)}{(1 - e^\theta)^{-\alpha} - 1} \\
&= \frac{\alpha(1 - e^\theta)^{-\alpha-1} e^\theta}{(1 - e^\theta)^{-\alpha} - 1} \\
&= \frac{\alpha e^\theta}{(1 - e^\theta)[1 - (1 - e^\theta)^\alpha]} \\
&= \frac{\alpha(1 - p)}{p(1 - p^\alpha)} \\
c''(\theta) &= \frac{d}{d\theta} \frac{\alpha e^\theta}{(1 - e^\theta)[1 - (1 - e^\theta)^\alpha]} \\
&= \frac{\alpha e^\theta}{(1 - e^\theta)[1 - (1 - e^\theta)^\alpha]} + \frac{\alpha e^\theta e^\theta}{(1 - e^\theta)^2[1 - (1 - e^\theta)^\alpha]} \\
&\quad - \frac{\alpha^2 e^\theta e^\theta (1 - e^\theta)^{\alpha-1}}{(1 - e^\theta)[1 - (1 - e^\theta)^\alpha]^2} \\
&= \frac{\alpha(1 - p)}{p(1 - p^\alpha)} + \frac{\alpha(1 - p)^2}{p^2(1 - p^\alpha)} - \frac{\alpha^2(1 - p)^2 p^{\alpha-2}}{(1 - p^\alpha)^2}
\end{aligned}$$

And after some rearrangement $c''(\theta)$ agrees with the variance calculated above.

The *mean value parameter*

$$\xi = \frac{\alpha(1 - p)}{p(1 - p^\alpha)} = \frac{\alpha e^\theta}{(1 - e^\theta)[1 - (1 - e^\theta)^\alpha]} \tag{D.19}$$

is not the usual parameter p. As with zero-truncated Poisson, we find that the inverse mapping $\xi \to \theta$ has no closed-form expression. As stated in Section D.4 above we know from general exponential family theory that this mapping $\xi \to \theta$ is strictly increasing, invertible, and infinitely differentiable. But in this case the inverse mapping $\xi \to \theta$ seems to have no closed-form expression. Also as stated in Section D.4 above, this means that this family does not have what GLM theory calls a link function in any useful form.

The full canonical parameter space is (D.14) as it was for the negative binomial.

Taking the limit in (D.19) as $\theta \to -\infty$ and $p \to 1$ we see that

$$\lim_{p \uparrow 1} \frac{\alpha(1-p)}{p(1-p^\alpha)} = \lim_{p \uparrow 1} \frac{-\alpha}{1 - p^\alpha - p\alpha p^{\alpha-1}} = 1$$

(using L'Hospital's rule). Taking the limit in (D.19) as $\theta \to \infty$ and $p \to 0$ we see that $\xi \to \infty$ in this case. And since we know from general exponential family theory that $c'(\theta)$ is a continuous increasing function, it follows that the *mean value parameter space* is $1 < \xi < \infty$.

As discussed at the end of the Sections D.2, D.3, D.4, and D.6.1, LCM are conditioned on the boundary of the closed convex support. The closed convex support is the closed interval $[1, \infty)$, and its boundary consists of the single point 1.

Thus there is one limiting conditional model, which contains only the distribution concentrated at one.

Also as discussed at the end of the those sections, it is important that we use (2.6) to determine the cumulant function for the LCM. So

$$\begin{aligned}
c_{-1}(\theta) &= c(\theta) + \log \Pr{}_\theta(Y = 1) \\
&= \log\left((1 - e^\theta)^{-\alpha} - 1\right) + \log\left(\frac{\alpha p^\alpha (1-p)}{1 - p^\alpha}\right) \\
&= \log\left(p^{-\alpha} - 1\right) + \log\left(\frac{\alpha p^\alpha (1-p)}{1 - p^\alpha}\right) \\
&= \log\left(\frac{1 - p^\alpha}{p^\alpha}\right) + \log\left(\frac{\alpha p^\alpha (1-p)}{1 - p^\alpha}\right) \\
&= \log(\alpha) + \log(1-p) \\
&= \log(\alpha) + \theta
\end{aligned}$$

Thus we see that unlike in Sections D.2 and D.4, the cumulant function for the family concentrated at one is not the identity function but rather the identity function plus a constant function.

D.8 Multivariate Bernoulli

A random vector is *multivariate Bernoulli* if it always has exactly one component equal to one and the rest of its components are equal to zero.

This is the *rationale* for the distribution. In categorical data analysis, where we have IID individuals that can take values in a finite number of categories, the result for each individual can be coded as multivariate Bernoulli.

The category corresponding to the component of the random vector that is equal to one says which category the individual is in.

This is a *discrete* random vector.

This is a special case of the multinomial distribution, which we do next.

This family is implemented in R package `aster2`. When incorporated in an aster model, this family is a *dependence group*. Because it is multivariate, it cannot be implemented in R package `aster`, which allows only univariate families. Although R package `aster2` calls this family "multinomial" created by the family function `fam.multinomial`, it is not the general multinomial described in the following section, but the multivariate Bernoulli described in this section, which is the $n = 1$ special case of the general multinomial.

D.9 Multinomial

This family is multi-dimensional and hence cannot be implemented in R package `aster`.

When IID individuals (a simple random sample) are classified into mutually exclusive and exhaustive categories (every individual falls in exactly one category), the vector of category counts has the *multinomial distribution*. This is one *rationale* for this distribution.

The other *rationale* is that when the sample size (predecessor) is equal to one, this family serves as a k-way switch if there are k categories (and this $n = 1$ special case is the multivariate Bernoulli family described in the preceding section).

This is the distribution of a random vector, not a random variable.

The Bernoulli and binomial distributions are related to the multinomial distribution, but the multinomial distribution with two categories is still a two-dimensional random vector, so it is not Bernoulli or binomial, which are one-dimensional distributions (of random variables). If Y is a Bernoulli random variable, then $(Y, 1 - Y)$ is a multivariate Bernoulli random vector. If Y is a binomial random variable with sample size n, then $(Y, n - Y)$ is a multinomial random vector with sample size n.

This is a *discrete* random vector.

The *probability mass function* is

$$f(y) = \frac{n!}{\prod_{i \in I} y_i!} \prod_{i \in I} p_i^{y_i}, \qquad y \in S, \tag{D.20}$$

where p is the *usual parameter vector*, which is a probability vector satisfying

$$p_i > 0$$
$$\sum_{i \in I} p_i = 1$$

where y is the the vector of counts (nonnegative integers) satisfying

$$y_i \geq 0 \tag{D.21a}$$
$$\sum_{i \in I} y_i = n \tag{D.21b}$$

where n is the sample size (the number of IID individuals classified), where S is the sample space for y, the set

$$S = \left\{ y \in \mathbb{N}^I : \sum_{i \in I} y_i = n \right\}, \tag{D.22}$$

where $\mathbb{N}$ denotes the *natural numbers* $\{0, 1, 2, 3, \ldots\}$, and where we have chosen the index set of y and θ to be an arbitrary finite set I rather than $\{1, \ldots, k\}$ for some k to fit in with the conventions of aster models (in an aster model I would be a subset of nodes of the aster graph comprising a multinomial dependence group, see Section 1.4 for vectors and subvectors).

The *mean vector* and *variance matrix* have components

$$E(Y_i) = np_i \tag{D.23a}$$
$$\operatorname{var}(Y_i) = np_i(1 - p_i) \tag{D.23b}$$
$$\operatorname{cov}(Y_i, Y_j) = -np_ip_j, \qquad i \neq j \tag{D.23c}$$

The mean of a random vector Y is the vector whose components are the means of the components of Y. Here $E(Y) = \xi$ and $\xi_i = np_i$. The variance of a random vector Y is the matrix whose components are the covariances of the components of Y. Here $\operatorname{var}(Y) = M$ has components m_{ij} which are given by (D.23b) when $i = j$ and by (D.23c) when $i \neq j$.

This is an *exponential family*. From (D.20) the log likelihood is

$$l(\theta) = \sum_{i \in I} y_i \log(p_i) \tag{D.24}$$

from which we see that we have an exponential family with *canonical statistic vector* y and *canonical parameter vector* θ having components

$$\theta_i = \log(p_i), \qquad i \in I. \tag{D.25}$$

Trying to read the cumulant function off of (D.24) seems to say $c(\theta)$ is the constant function everywhere equal to zero, and this is correct because the p_i must sum to one, and as we shall see the cumulant function does have the value zero when

$$\sum_{i \in I} e^{\theta_i} = 1. \tag{D.26}$$

But we want the cumulant function defined on the whole vector space where θ lives, so we must use equation (5) in Geyer (2009), which is (1.26) in this book,

$$\begin{aligned}
c(\theta) &= c(\psi) + \log E_\psi \left\{ e^{\sum_{i \in I} y_i (\theta_i - \psi_i)} \right\} \\
&= c(\psi) + \log \sum_{y \in S} e^{\sum_{i \in I} y_i (\theta_i - \psi_i)} \cdot \frac{n!}{\prod_{i \in I} y_i!} \prod_{i \in I} p_i^{y_i} \\
&= c(\psi) + \log \sum_{y \in S} \frac{n!}{\prod_{i \in I} y_i!} \prod_{i \in I} (p_i e^{\theta_i - \psi_i})^{y_i} \\
&= c(\psi) + \log \left(\sum_{i \in I} p_i e^{\theta_i - \psi_i} \right)^n \\
&= c(\psi) + n \log \left(\sum_{i \in I} p_i e^{\theta_i - \psi_i} \right)
\end{aligned}$$

where the last equality is the multinomial theorem also called the theorem associated with the multinomial distribution (Geyer, 2019b). Here ψ is a possible value of the canonical parameter vector (held fixed) and p is the usual parameter vector corresponding to it so $p_i = e^{\psi_i}$ and $p_i e^{\theta_i - \psi_i} = e^{\theta_i}$. So dropping $c(\psi)$, which is an arbitrary constant, we obtain

$$c(\theta) = n \log \left(\sum_{i \in I} e^{\theta_i} \right). \tag{D.27}$$

Since this is finite for all vectors θ, the full canonical parameter space is the whole vector space $\mathbb{R}^I$ where θ lives.

This gives a log likelihood valid on the whole vector space where θ lives

$$\begin{aligned}
l(\theta) &= \langle y, \theta \rangle - c(\theta) \\
&= \left(\sum_{i \in I} y_i \theta_i \right) - n \log \left(\sum_{i \in I} e^{\theta_i} \right)
\end{aligned} \tag{D.28}$$

We check that (D.27) has the correct derivatives

$$\frac{\partial c(\theta)}{\partial \theta_i} = \frac{n e^{\theta_i}}{\sum_{j \in I} e^{\theta_j}} \tag{D.29a}$$

$$\frac{\partial^2 c(\theta)}{\partial \theta_i^2} = \frac{n e^{\theta_i}}{\sum_{j \in I} e^{\theta_j}} - \frac{n e^{\theta_i} e^{\theta_i}}{\left(\sum_{j \in I} e^{\theta_j} \right)^2} \tag{D.29b}$$

$$\frac{\partial^2 c(\theta)}{\partial \theta_i \theta_j} = - \frac{n e^{\theta_i} e^{\theta_j}}{\left(\sum_{k \in I} e^{\theta_k} \right)^2} \tag{D.29c}$$

But here we have a problem that this does not even make sense with our previous notion of canonical parameters, which, recall, was defined by (D.25) but only subject to the constraint (D.26). So we do not have a notion of what the map $\theta \to p$ should be for values of θ that do not satisfy the constraint (D.26).

We solve this problem by using the fundamental relationship between cumulant functions and means and variances (Section 1.15.5 above), which says the derivatives above have to give means, variances, and covariances. Thus (D.29a) must give the correct mean values

$$\frac{n e^{\theta_i}}{\sum_{j \in I} e^{\theta_j}} = n p_i$$

so

$$p_i = \frac{e^{\theta_i}}{\sum_{j \in I} e^{\theta_j}}, \qquad i \in I, \tag{D.30}$$

gives the correct mapping between our new canonical parameters (now θ can be any vector in $\mathbb{R}^I$) and the usual parameters.

This function is not invertible. If one adds the same constant to all of the θ_i, then the value of p_i does not change

$$\frac{e^{\theta_i + c}}{\sum_{j \in I} e^{\theta_j + c}} = \frac{e^c e^{\theta_i}}{e^c \sum_{j \in I} e^{\theta_j}} = \frac{e^{\theta_i}}{\sum_{j \in I} e^{\theta_j}}$$

This illustrates another thing wrong with the concept of the link function: it forces the canonical parameterization to be identifiable even when this is inadvisable. Using our choice of parameterization here there can be no link function because the map $\theta \to \xi$ is not one-to-one, so its inverse mapping does not exist (an that inverse mapping is supposed to be the "link" function).

We also clear up a mystery left hanging and check that (D.27) does indeed evaluate to zero when the constraint (D.26) holds.

Having made the identification (D.30) we see that (D.29b) and (D.29c) do give the correct variances and covariances

$$\begin{aligned}\frac{\partial^2 c(\theta)}{\partial \theta_i^2} &= \operatorname{var}(y_i) = np_i(1 - p_i) \\ \frac{\partial^2 c(\theta)}{\partial \theta_i \partial \theta_j} &= \operatorname{cov}(y_i, y_j) = -np_i p_j\end{aligned}$$

The canonical parameter space is the whole vector space where θ lives.

The mean value parameter vector, as stated above, is the vector ξ having components $\xi_i = np_i$. In case $n = 1$, the mean value parameter vector is the usual parameter vector. Otherwise, not.

The mean value parameter space is the relative interior of the convex hull of S

$$\Xi = \left\{ \xi \in \mathbb{R}^I : \xi_i > 0, i \in I \text{ and } \sum_{j \in I} \xi_j = n \right\} \tag{D.31}$$

To see this we note that $\theta_i = \log(\xi_i)$ always defines a point in the canonical sample space so long as the ξ_i are strictly positive. And that point maps via (D.30) to

$$p_i = \frac{e^{\theta_i}}{\sum_{j \in I} e^{\theta_j}} = \frac{\xi_i}{\sum_{j \in I} \xi_j} = \frac{\xi_i}{n}$$

which makes the p_i sum to one. So every point in Ξ is a mean value parameter vector value, and, conversely, every point $\theta \in \mathbb{R}^I$ maps to a value ξ that satisfies $\xi_i > 0$ for all i and $\sum_{i \in I} \xi_i = n$.

Theorem D.2. *The closed convex support of the multinomial family is*

$$C = \left\{ \xi \in \mathbb{R}^I : \xi_i \geq 0, i \in I \text{ and } \sum_{j \in I} \xi_j = n \right\}$$

the closure (D.31). *The support function is given by*

$$\sigma_C(\delta) = n \max(\delta), \tag{D.32}$$

where

$$\max(\delta) = \max_{i \in I} \delta_i.$$

Define the hyperplane

$$H_\delta = \{\, y \in \mathbb{R}^I : \langle y, \delta \rangle = \sigma_C(\delta) \,\},$$

which is (2.3) *with the index set* I *instead of* J*. Then the cumulant function for the LCM conditioned on* H_δ *is*

$$c_\delta(\theta) = n \log \left(\sum_{\substack{i \in I \\ \delta_i = \max(\delta)}} e^{\theta_i} \right). \tag{D.33}$$

If we define

$$G = \{\, i \in I : \delta_i = \max(\delta) \,\},$$

then under the LCM conditioned on H_δ*, the family of distributions for* y_G *is multinomial with sample size* n *and* $y_{I \setminus G} = 0$ *almost surely.*

Proof. For any $i \in I$ define the vector $v^{(i)}$ (we use the temporary notation of superscript in parenthesis to denote a sequence of vectors) having $v_i^{(i)} = n$ and $v_j^{(i)} = 0$ for $j \in I \setminus \{i\}$. Clearly, each such $v^{(i)}$ is in the sample space (D.22) and has probability $p_i^n > 0$. Thus it must be in any support. Now the formula

$$C = \left\{ \sum_{i \in I} p_i v^{(i)} : p_i \ge 0, i \in I \text{ and } \sum_{j \in I} p_j = 1 \right\}$$

shows that C is the convex hull of the vectors $v^{(i)}$. Thus C must be contained in the closed convex support. On the other hand, C contains S, so C is the smallest closed convex set containing S. Hence C is the closed convex support.

We now claim that any vector v having $v_j = 0$ when $\delta_j < \max(\delta)$ maximizes $\langle \, \cdot \, , \delta \rangle$ over C. We have

$$\langle v, \delta \rangle = \sum_{\substack{i \in I \\ \delta_i = \max(\delta)}} v_i \delta_i = \max(\delta) \sum_{\substack{i \in I \\ \delta_i = \max(\delta)}} v_i = \max(\delta) \sum_{i \in I} v_i = n \max(\delta)$$

and for any $x \in C$ we have

$$\langle v, \delta \rangle = \sum_{i \in I} v_i \delta_i \le \max(\delta) \sum_{i \in I} v_i = n \max(\delta)$$

Thus (D.32) is correct.

Now if $y \in C$ such that

$$\langle y, \delta \rangle = \sigma_C(\delta) = n \max(\delta) \tag{D.34}$$

we must have, as we have already seen, $y_j = 0$ if $\delta_j < \max(\delta)$. Conversely, if $y \in C$ and $y_j = 0$ if $\delta_j < \max(\delta)$, then (D.34) holds. Hence conditioning the original model on $Y \in H_\delta$ is the same as conditioning on $Y_{I \setminus G} = 0$. Then the usual formulas for conditionals for multinomials show that Y_G is multinomial with sample size n, as asserted.

It only remains to calculate the cumulant function for the LCM using (2.6).

$$\begin{aligned}
c_\delta(\theta) &= c(\theta) + \log \Pr\nolimits_\theta(Y \in H_\delta) \\
&= n \log \left(\sum_{i \in I} e^{\theta_i} \right) + \log \left(\sum_{i \in G} p_i \right)^n \\
&= n \log \left(\sum_{i \in I} e^{\theta_i} \right) + n \log \left(\sum_{i \in G} p_i \right) \\
&= n \log \left(\sum_{i \in I} e^{\theta_i} \right) + n \log \left(\frac{\sum_{i \in G} e^{\theta_i}}{\sum_{i \in I} e^{\theta_i}} \right) \\
&= n \log \left(\sum_{i \in G} e^{\theta_i} \right)
\end{aligned}$$

□

D.10 Normal Location-Scale

The univariate normal distribution is curious in that it remains an exponential family even if we consider both parameters unknown, but then the dimensions of the canonical statistic vector and the canonical parameter vector must match. So if the canonical parameter vector is going to be two-dimensional so must be the canonical statistic vector.

Let's see how that happens. We already have the probability density function (D.10) of the normal distribution. Now if we write down the log likelihood not dropping any terms that contain either parameter we get

$$l(\theta) = -\log(\sigma) - \frac{(x - \nu)^2}{2\sigma^2} = -\log(\sigma) - \frac{x^2}{2\sigma^2} + \frac{x\nu}{\sigma^2} - \frac{\nu^2}{2\sigma^2} \tag{D.35}$$

where for reasons to be discussed presently we have changed the notation for the random variable from y to x and the notation for the mean from ξ to ν.

As was discussed in Section 1.15.1 in the main text, there is some freedom in choosing the canonical statistic vector and the canonical parameter vector we must have the terms containing both data and parameters in exponential family form, that is,

$$-\frac{x^2}{2\sigma^2} + \frac{x\nu}{\sigma^2} = y_1\theta_1 + y_2\theta_2$$

but that still allows lots of choices. We could, for example, choose $y_1 = x$ or $y_1 = x^2$ or $y_1 = -x^2/2$. And each such choice forces a different choice of the corresponding canonical parameter.

The choice made in the implementation in R package `aster2` is

$$\begin{aligned} y_1 &= x, \\ y_2 &= x^2, \\ \theta_1 &= \frac{\nu}{\sigma^2} \\ \theta_2 &= -\frac{1}{2\sigma^2} \end{aligned}$$

We have had many examples where the usual parameters are not the canonical parameters. Here is our first example where the usual statistics are not the canonical statistics (the usual statistic is one canonical statistic but not the other).

In terms of the canonical parameters, the usual parameters are

$$\begin{aligned} \sigma^2 &= -\frac{1}{2\theta_2} \\ \nu &= \theta_1\sigma^2 = -\frac{\theta_1}{2\theta_2} \end{aligned}$$

The first canonical parameter is unrestricted $-\infty < \theta_1 < \infty$ but the second canonical parameter is restricted $-\infty < \theta_2 < 0$. Thus our guess at the cumulant function from looking at the log likelihood

$$\begin{aligned} c(\theta) &= \log(\sigma) + \frac{\nu^2}{2\sigma^2} \\ &= \frac{1}{2}\log\left(-\frac{1}{2\theta_2}\right) + \left(-\frac{\theta_1}{2\theta_2}\right)^2 \cdot \frac{1}{2} \cdot (-2\theta_2) \\ &= \frac{1}{2}\log\left(-\frac{1}{2\theta_2}\right) - \frac{\theta_1^2}{4\theta_2} \end{aligned} \tag{D.36}$$

(this agrees with the cumulant function for this family in R package `aster2`).

Notice that in going from the PDF to the log likelihood (D.35), on a factor of $1/\sqrt{2\pi}$ was dropped, so the PDF of the family (with respect to Lebesgue measure) can now be written

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{x\theta_1 + x^2\theta_2 - c(\theta)}$$

and the fact that PDF integrate to one gives

$$\int e^{x\theta_1 + x^2\theta_2} \, dx = \sqrt{2\pi} e^{c(\theta)}$$

Because our guess at the cumulant function is not defined on a whole vector space we use equation (5) in Geyer (2009), which is (1.26) in this book, as we have done several times before in this appendix

$$\begin{aligned} c(\theta) &= c(\psi) + \log E_\psi\bigl(e^{y_1(\theta_1 - \psi_1) + y_2(\theta_2 - \psi_2)}\bigr) \\ &= c(\psi) + \log E_\psi\bigl(e^{y_1(\theta_1 - \psi_1) + y_2(\theta_2 - \psi_2)}\bigr) \\ &= c(\psi) + \log \left(\frac{1}{\sqrt{2\pi}} \int e^{x(\theta_1 - \psi_1) + x^2(\theta_2 - \psi_2)} e^{x\psi_1 + x^2\psi_2 - c(\psi)} \, dx \right) \\ &= \log \left(\frac{1}{\sqrt{2\pi}} \int e^{x\theta_1 + x^2\theta_2} \, dx \right) \end{aligned}$$

and, as we have already seen, the last expression is equal to $c(\theta)$ for θ such that $\theta_2 < 0$. In case $\theta_2 \geq 0$ the integrand is either constant or goes to infinity as $x \to \infty$ or $x \to -\infty$ or both. In any case the integral does not exist. Thus the canonical parameter space is indeed

$$\Theta = \{\, \theta \in \mathbb{R}^2 : \theta_2 < 0 \,\}. \tag{D.37}$$

The canonical sample space for sample size one is

$$S = \{\, y \in \mathbb{R}^2 : y_1^2 = y_2 \,\}$$

which is a curve. Thus any intersection with a tangent line consists of a single point. And, because this is a continuous distribution, single points have probability zero. Thus in the limiting conditional model theorem (Theorem 2.1 above) we are in the case where the probability of $Y \in H_\delta$ is equal to zero (regardless of what δ is). So there are no LCM for this family.

For sample size one, the data are on the curve S with probability one. So the MLE does not exist with probability one. This is no surprise. One cannot estimate two parameters from one variable x.

For sample size greater than one, the data are in the interior of the convex hull of the curve S with probability one. So the MLE exists with probability one. This is no surprise either. One can estimate the two parameters of the normal distribution from data x_1, x_2, ..., x_n for $n \geq 2$.

So we don't need the theory of limiting conditional models for this family. That theory is only needed for discrete families.

We forgot to check that (D.36) has the correct derivatives. Let's do that now.

$$\begin{aligned}
\frac{\partial c(\theta)}{\partial \theta_1} &= -\frac{\theta_1}{2\theta_2} \\
\frac{\partial c(\theta)}{\partial \theta_2} &= -\frac{1}{2\theta_2} + \frac{\theta_1^2}{4\theta_2^2} \\
\frac{\partial^2 c(\theta)}{\partial \theta_1^2} &= -\frac{1}{2\theta_2} \\
\frac{\partial^2 c(\theta)}{\partial \theta_1 \partial \theta_2} &= \frac{\theta_1}{2\theta_2^2} \\
\frac{\partial^2 c(\theta)}{\partial \theta_2^2} &= \frac{1}{2\theta_2^2} - \frac{\theta_1^2}{2\theta_2^3}
\end{aligned}$$

Translating these back to functions of x and the original parameters, they say

$$\begin{aligned}
E(X) &= \nu \\
E(X^2) &= \sigma^2 + \nu^2 \\
\operatorname{var}(X) &= \sigma^2 \\
\operatorname{cov}(X, X^2) &= 2\nu\sigma^2 \\
\operatorname{var}(X^2) &= 4\nu^2\sigma^2 + 2\sigma^4
\end{aligned}$$

The first three of these are well known. The other two do check (in Mathematica).

D.11 Gamma Rate

The gamma distribution has PDF

$$f(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}, \qquad 0 < y < \infty, \tag{D.38}$$

where $\alpha > 0$ and $\lambda > 0$ are parameters.

This is a *continuous* random variable; except when incorporated into an aster model, it is a mixture of discrete and continuous. For a gamma-rate arrow, when the predecessor is zero the conditional distribution of the successor is the degenerate random variable concentrated at zero, which is discrete, and when the predecessor is greater than zero, the conditional distribution of the successor is continuous.

This is the most well known continuous distribution after the normal distribution, and it has many rationales, but these rationales do not seem to justify its inclusion in aster models, which is why R packages `aster` and `aster2` do not include it. This family includes the exponential distribution (case $\alpha = 1$) and the chi-square distributions (case $\alpha = n/2$ and $\lambda = 1/2$) as special cases.

The main rationale is vague. This is the most well-known distribution with support $(0, \infty)$. So it is a TTD (thing to do) when one wants such.

The *mean* and *variance* are

$$E(y) = \frac{\alpha}{\lambda} \tag{D.39a}$$

$$\operatorname{var}(y) = \frac{\alpha}{\lambda^2} \tag{D.39b}$$

In this section, we treat this as a one-parameter family with α known and λ unknown.

With this understanding, the log likelihood is

$$l(\theta) = \alpha \log(\lambda) - \lambda y$$

(we have dropped terms that do not contain λ). From this we see that we have an *exponential family* with *canonical statistic* y, *canonical parameter* $\theta = -\lambda$, and *cumulant function*

$$c(\theta) = -\alpha \log(\lambda) = -\alpha \log(-\theta). \tag{D.40}$$

We check that this has the correct derivatives

$$c'(\theta) = \frac{\alpha}{-\theta} = \frac{\alpha}{\lambda}$$

and

$$c''(\theta) = \frac{\alpha}{\theta^2} = \frac{\alpha}{\lambda^2}.$$

Because our guess at the cumulant function is not defined on a whole vector space we use equation (5) in Geyer (2009), which is (1.26) in this

book, as we have done several times before in this appendix

$$\begin{aligned} c(\theta) &= c(\psi) + \log E_\psi\bigl(e^{y(\theta-\psi)}\bigr) \\ &= c(\psi) + \log \int e^{y(\theta-\psi)} f_\psi(y)\, dy \\ &= c(\psi) + \log \int e^{y(\theta-\psi)} \frac{\lambda_\psi^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda_\psi y}\, dy \end{aligned}$$

where $\lambda_\psi = -\psi$ so

$$c(\theta) = \log \int \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{y\theta}\, dy$$

and this integral clearly exists if and only if $\theta < 0$ in which case, by (D.38) integrating to one, we get (D.40), as we must.

Hence the full canonical parameter space of this family is

$$\Theta = \{\, \theta \in \mathbb{R} : \theta < 0 \,\}.$$

The *mean value parameter* is

$$\xi = \frac{\alpha}{-\theta}$$

Addition rule: the sum of n independent and identically distributed gamma random variables with shape parameter α and rate parameter λ has the gamma distribution with shape parameter $n\alpha$ and rate parameter λ.

General addition rule: the sum of n independent gamma random variables with shape parameters $\alpha_1, \ldots, \alpha_n$ and rate parameter λ (same for all) has the gamma distribution with shape parameter $\alpha_1 + \cdots + \alpha_n$ and rate parameter λ.

There are no limit degenerate distributions. This is because the boundary of the closed convex support, which is the interval $[0, \infty)$ is the point zero. We can never observe data on the boundary, because continuous distributions put probability zero at any point.

D.12 Gamma Shape-Rate

If we take the gamma distribution with PDF (D.38) to have both parameters unknown, then the log likelihood is

$$l(\theta) = \alpha \log(\lambda) - \log \Gamma(\alpha) + \alpha \log(x) - \lambda x$$

(we have dropped terms that do not contain α or λ and have changed the data variable from y to x). This has the form of a two-dimensional exponential family with *canonical statistics* $y_1 = x$ and $y_2 = \log x$, *canonical parameters* $\theta_1 = -\lambda$ and $\theta_2 = \alpha$, and *cumulant function*

$$c(\theta) = \log \Gamma(\alpha) - \alpha \log(\lambda) = \log \Gamma(\theta_2) - \theta_2 \log(-\theta_1) \tag{D.41}$$

With this definition of canonical parameters and statistics, we can rewrite the PDF as

$$\begin{aligned} f_\theta(y) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \\ &= x^{-1} e^{\langle y, \theta \rangle - c(\theta)} \\ &= y_1^{-1} e^{\langle y, \theta \rangle - c(\theta)} \end{aligned} \tag{D.42}$$

Because our guess at the cumulant function is not defined on a whole vector space we use equation (5) in Geyer (2009), which is (1.26) in this book, as we have done several times before in this appendix

$$\begin{aligned} c(\theta) &= c(\psi) + \log E_\psi\bigl(e^{\langle y, \theta - \psi \rangle}\bigr) \\ &= c(\psi) + \log \int e^{\langle y, \theta - \psi \rangle} f_\psi(y) \, dy \\ &= c(\psi) + \log \int e^{\langle y, \theta - \psi \rangle} y_1^{-1} e^{\langle y, \psi \rangle - c(\psi)} \, dx \\ &= \log \int y_1^{-1} e^{\langle y, \theta \rangle} \, dx \end{aligned}$$

We know that (D.42) integrates if and only if $\alpha > 0$ and $\lambda > 0$ (Geyer, 2019a, Slides 27–30). It follows that the full canonical parameter space is

$$\Theta = \{\, \theta \in \mathbb{R}^2 : \theta_1 < 0 \text{ and } \theta_2 > 0 \,\}$$

and in order that (D.42) integrate to one when the integral exists, the last integral above must be $c(\theta)$.

Let us check that (D.41) has the correct derivatives

$$\begin{aligned}
\frac{\partial c(\theta)}{\partial \theta_1} &= -\frac{\theta_2}{\theta_1} \\
\frac{\partial c(\theta)}{\partial \theta_2} &= \operatorname{digamma}(\theta_2) - \log(-\theta_1) \\
\frac{\partial^2 c(\theta)}{\partial \theta_1^2} &= \frac{\theta_2}{\theta_1^2} \\
\frac{\partial^2 c(\theta)}{\partial \theta_1 \partial \theta_2} &= -\frac{1}{\theta_1} \\
\frac{\partial^2 c(\theta)}{\partial \theta_2^2} &= \operatorname{trigamma}(\theta_2)
\end{aligned}$$

Translating these back to functions of x and the original parameters, they say

$$\begin{aligned}
E(X) &= \frac{\alpha}{\lambda} \\
E\{\log(X)\} &= \operatorname{digamma}(\alpha) - \log(\lambda) \\
\operatorname{var}(X) &= \frac{\alpha}{\lambda^2} \\
\operatorname{cov}\{X, \log(X)\} &= \frac{1}{\lambda} \\
\operatorname{var}\{\log(X)\} &= \operatorname{trigamma}(\alpha)
\end{aligned}$$

where the digamma function is the first derivative of $\log \Gamma(\,\cdot\,)$, and the trigamma function is the second derivative of this function.

Two of these agree with the known formulas for the mean and variance of a gamma random variable (D.39a) and (D.39b). The others involve integrals we don't know how to do other than by the method of cumulant functions (what we just did). (But Mathematica knows how to do these integrals.)

The mean value parameter is the two-dimensional vector ξ having components

$$\begin{aligned}
\xi_1 &= -\frac{\theta_2}{\theta_1} \\
\xi_2 &= \operatorname{digamma}(\theta_2) - \log(-\theta_1)
\end{aligned}$$

The addition rules for the gamma distribution were given in the preceding section (they do not depend on which parameters are considered known or unknown).

The canonical sample space for sample size one is

$$S = \{ y \in \mathbb{R}^2 : y_1 = e^{y_2} \}$$

which is a curve. Thus any intersection with a tangent line consists of a single point. And, because this is a continuous distribution, single points have probability zero. Thus in the limiting conditional model theorem (Theorem 2.1 above) we are in the case where the probability of $Y \in H_\delta$ is equal to zero (regardless of what δ is). So there are no LCM for this family.

For sample size one, the data are on the curve S with probability one. So the MLE does not exist with probability one. This is no surprise. One cannot estimate two parameters from one variable x.

For sample size greater than one, the data are in the interior of the convex hull of the curve S with probability one. So the MLE exists with probability one. This is no surprise either. One can estimate the two parameters of the normal distribution from data x_1, x_2, ..., x_n for $n \geq 2$.

So we don't need the theory of limiting conditional models for this family. That theory is only needed for discrete families.

The last three paragraphs almost exactly repeat what was said about the normal-location-scale family. The reasoning applies to any continuous distribution. The details distinguishing one of these families from the other do not matter in this argument.

D.13 K-Truncated Families

R package `aster` implements k-truncated Poisson and k-truncated negative binomial for any nonnegative integer k. Example 3 in Shaw et al. (2008b) used two-truncated negative binomial. We no longer think using this family is the best way to do this example. Moreover we know of no other examples in life history analysis that need these families (except for the zero-truncated ones already discussed). Therefore we omit further discussion of them.

Bibliography

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, England.

Browder, A. (1996). *Mathematical Analysis: An Introduction*. Springer-Verlag, New York.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayward, CA.

Caswell, H. (2001). *Matrix Population Models: Construction, Analysis, and Interpretation*. Sinauer Associates, Sunderland, MA, second edition.

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

Cuppens, R. (1975). *Decomposition of Multivariate Probability*. Academic Press, New York.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, New York.

Eck, D. J., Shaw, R. G., Geyer, C. J., and Kingsolver, J. G. (2015). An integrated analysis of phenotypic selection on insect body size and development time. *Evolution*, **69**, 2525–2532. doi:10.1111/evo.12744.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, England.

Geyer, C. J. (1990). *Likelihood and Exponential Families*. Ph.D. thesis, University of Washington. URL `http://hdl.handle.net/11299/56330`.

Geyer, C. J. (1991). Constrained maximum likelihood exemplified by isotonic convex logistic regression. *Journal of the American Statistical Association*, **86**, 717–724. doi:10.1080/01621459.1991.10475100.

Geyer, C. J. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, **3**, 259–289. doi: 10.1214/08-EJS349.

Geyer, C. J. (2010). A philosophical look at aster models. Technical Report 676, School of Statistics, University of Minnesota. URL `http://hdl.handle.net/11299/57163`.

Geyer, C. J. (2017a). *R package `aster2`: Aster Models, version 0.3*. URL `http://cran.r-project.org/package=aster2`.

Geyer, C. J. (2017b). *R package `pooh`: Partial Orders and Relations, version 0.3-2*. URL `http://cran.r-project.org/package=pooh`.

Geyer, C. J. (2017c). Stat 3701 lecture notes: Zero-truncated poisson distribution. URL `http://www.stat.umn.edu/geyer/3701/notes/zero.html`.

Geyer, C. J. (2018). Statistics 8931 (Geyer, fall 2018). URL `http://www.stat.umn.edu/geyer/8931aster/`.

Geyer, C. J. (2019a). Stat 5101 lecture slides: Deck 6: Existence of integrals and infinite sums, countable additivity and monotone convergence, existence of moments, correlation. URL `http://www.stat.umn.edu/geyer/5101/slides/s6.pdf`.

Geyer, C. J. (2019b). Stat 5101 notes: Brand name distributions. URL `http://www.stat.umn.edu/geyer/5101/notes/brand.pdf`.

Geyer, C. J. (2021). *R package `aster`: Aster Models, version 1.1-2*. URL `http://cran.r-project.org/package=aster`.

Geyer, C. J. and Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, **21**, 359–373. URL `http://www.jstor.org/stable/4616323`.

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. doi:10.1214/13-AOAS653.

Geyer, C. J., Wagenius, S., and Shaw, R. G. (2005). Aster models for life history analysis. Technical Report 644, School of Statistics, University of Minnesota. URL `http://hdl.handle.net/11299/212317`.

Geyer, C. J., Wagenius, S., and Shaw, R. G. (2007). Aster models for life history analysis. *Biometrika*, **94**, 415–426. doi:10.1093/biomet/asm030.

Goodman, L. A. (1968). An elementary approach to the population projection-matrix, to the population reproductive value, and to related topics in the mathematical theory of population growth. *Demography*, **5**, 382–409. doi:10.1007/BF03208583.

Halmos, P. R. (1958). *Finite-Dimensional Vector Spaces*. Van Nostrand, Princeton, NJ, second edition.

Halmos, P. R. (1960). *Naive Set Theory*. Van Nostrand, Princeton, NJ.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York.

Hollander, M., Wolfe, D. A., and Chicken, E. (2014). *Nonparametric Statistical Methods*. John Wiley & Sons, Hoboken, NJ, third edition.

Huber, P. J. and Ronchetti, E. (2009). *Robust Statistics*. John Wiley & Sons, Hoboken, NJ, second edition.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14. doi:10.1080/00401706.1992.10485228.

Lande, R. and Arnold, S. J. (1983). The measurement of selection on correlated characters. *Evolution*, **37**, 1210–1226. doi:10.2307/2408842.

Lang, S. (1993). *Real and Functional Analysis*. Springer-Verlag, New York, third edition.

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

Lenski, R. E. and Service, P. M. (1982). The statistical analysis of population growth rates calculated from schedules of survivorship and fecundity. *Ecology*, **63**, 655–662. doi:10.2307/1936785.

May, G., Shaw, R. G., Geyer, C. J., and Eck, D. J. (2022). Do interactions among microbial symbionts cause selection for greater pathogen virulence? *American Naturalist*, **199**, 252–265. doi:10.1086/717679.

Mitchell-Olds, T. and Shaw, R. G. (1987). Regression analysis of natural selection: Statistical inference and biological interpretation. *Evolution*, **41**, 1149–1161. doi:10.1111/j.1558-5646.1987.tb02457.x.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, NJ.

Rockafellar, R. T. (1988). *Network Flows and Monotropic Optimization*. Athena Scientific, Belmont MA. Originally published by John Wiley & Sons, 1984.

Rockafellar, R. T. and Wets, R. J.-B. (1998). *Variational Analysis*. Springer-Verlag, Berlin. Corrected printings contain extensive changes. We used the third corrected printing, 2010.

Rudin, W. (1991). *Functional Analysis*. McGraw-Hill, New York, second edition.

Shaw, R. G. and Geyer, C. J. (2010). Inferring fitness landscapes. *Evolution*, **64**, 2510–2520. doi:10.1111/j.1558-5646.2010.01010.x.

Shaw, R. G., Geyer, C. J., Wagenius, S., Hangelbroek, H. H., and Etterson, J. R. (2008a). More supporting data analysis for "Unifying life history analysis for inference of fitness and population growth". Technical Report 661, School of Statistics, University of Minnesota. URL `https://www.stat.umn.edu/geyer/aster/tr661.pdf`.

Shaw, R. G., Geyer, C. J., Wagenius, S., Hangelbroek, H. H., and Etterson, J. R. (2008b). Unifying life history analysis for inference of fitness and population growth. *American Naturalist*, **172**, E35–E47. doi:10.1086/588063.

Shaw, R. G., Wagenius, S., and Geyer, C. J. (2015). The susceptibility of *Echinacea angustifolia* to a specialist aphid: eco-evolutionary perspective on genotypic variation and demographic consequences. *Journal of Ecology*, **103**, 809–818. doi:10.1111/1365-2745.12422.

Stanton-Geddes, J., Shaw, R. G., and Tiffin, P. (2012a). Interactions between soil habitat and geographic range affect plant fitness. *PLoS One*, **7**, e36015.

Stanton-Geddes, J., Tiffin, P., and Shaw, R. G. (2012b). Role of climate and competitors in limiting fitness across range edges of an annual plant. *Ecology*, **93**, 1604–1613. doi:10.1890/11-1701.1. Supplemental material, *Ecological Archives*, E093-142-A1, `http://esapubs.org/archive/ecol/E093/142/appendix-A.htm`.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Warwell, M. V. and Shaw, R. G. (2017). Climate-related genetic variation in a threatened tree species, *Pinus albicaulis*. *American Journal of Botany*, **104**, 1205–1218.

Index of Notation

$\mathbb{R}$	the real number system
$\mathbb{N}$	the natural number system, which starts at zero
N	the set of nodes of the full aster graph, including initial nodes
J	the set of non-initial nodes of the full aster graph
$\mathcal{G}$	the family of dependence groups, a partition of J
q	the set-to-index predecessor function, which maps $\mathcal{G} \to N$
p	the index-to-index predecessor function, which maps $J \to N$
$\mathbb{R}^A$	the set of all functions $A \to \mathbb{R}$ considered as a vector space
y_j	a component of the vector y: if $y \in \mathbb{R}^A$ and $j \in A$, then y_j is the value of the function y at argument j
y_A	a subvector of the vector y: if $y \in \mathbb{R}^B$ and $A \subset B$, then y_A is the restriction of the function y to the set A
$\operatorname{pr}(y)$	unconditional distribution of the random vector y described somehow (Section 1.6)
$\operatorname{pr}(y_A \mid y_B)$	conditional distribution of the random vector y_A described somehow (Section 1.6 and equation (B.1) and the surrounding discussion)
θ	conditional canonical parameter vector of the saturated model
φ	unconditional canonical parameter vector of the saturated model
ξ	conditional mean value parameter vector of the saturated model
μ	unconditional mean value parameter vector of the saturated model
β	unconditional canonical parameter vector of a canonical affine submodel
τ	unconditional mean value parameter vector of a canonical affine submodel
a	offset vector
M	model matrix

f^n	the function f composed with itself n times; f^0 being the identity function, and $f^1 = f$
$\succ$	transitive closure of the predecessor relation
$\succeq$	reflexive transitive closure of the predecessor relation
$\prec$	transitive closure of the successor relation
$\preceq$	reflexive transitive closure of the successor relation
$\langle \cdot, \cdot \rangle$	bilinear form placing dual vector spaces in duality

Index