Stat 8931 (Aster Models)
Lecture Slides Deck 2

Basic Theory

Charles J. Geyer

School of Statistics
University of Minnesota

November 2, 2020

R and License

- The version of R used to make these slides is 4.0.3.
- The version of R package `aster` used to make these slides is 1.0.3.

-

Statistical Models

A **statistical model** is a family of probability distributions.

In many courses this concept is hidden behind sloppy terminology.

We often say "the binomial distribution" when we really mean *the family of binomial distributions* (each different parameter value gives a different binomial distribution).

We often say "the normal distribution" when we really mean *the family of normal distributions* (each different pair of parameter values gives a different normal distribution).

And similarly for other distributions.

Statistical Models (cont.)

When you have a statistical model for your data, all the techniques of mathematical statistics are available. Any question that can be phrased in terms of probabilities and expectations with respect to distributions in the model can be answered.

When you do not have a statistical model for your data, many of the techniques of mathematical statistics are not available. For example, you can use the method of moments, but you cannot use the method of maximum likelihood or the method of Bayesian inference. For another example, you can use the method of least squares, but when no model is assumed it comes with no hypothesis tests or confidence intervals.

Aster models are statistical models.

Statistical Models (cont.)

In “classical” or “master’s level” theoretical statistics (5101–5102 or 8101–8102 in our department) specifying a statistical model is simple. There are two kinds of models for two kinds of data: discrete and continuous.

If the data are discrete, then the model is specified by a **probability mass function (PMF)** and probabilities and expectations are calculated by sums.

If the data are continuous, then the model is specified by a **probability density function (PDF)** and probabilities and expectations are calculated by integrals.

Statistical Models (cont.)

In “master’s level” theoretical statistics, it may have been mentioned that there are probability models that are neither discrete or continuous, and in aster models we get them, but only in a rather trivial way.

Although our first example had all components of the response vector discrete, this is not necessary.

The predecessor-is-sample-size property requires all predecessor nodes to have nonnegative-integer-valued random variables. But terminal nodes can have continuous random variables.

The `aster` package has `fam.normal.location`, which specifies normal with unknown mean and known variance as a family that components of the response vector can have.

The `aster2` package has `fam.normal.location.scale`, which specifies normal with unknown mean and unknown variance as a family that components of the response vector can have.

Statistical Models (cont.)

So if we have some continuous and some discrete components of the response vector, then we do not have either a PMF or a PDF.

If the data have some components discrete and some continuous, then the model is specified by a **probability mass-density function (PMDF)** and probabilities and expectations are calculated by sums over the discrete components and integrals over the continuous components.

We do not need to calculate expectations this way for aster models. Instead we use moment generating functions (more on this later). So this is purely a theoretical quibble.

Statistical Models (cont.)

In aster models, even "continuous" families are partly discrete.

$$1 \xrightarrow{\text{Poi}} y_1 \xrightarrow{\text{Nor}} y_2$$

The sum of n IID Normal(μ, σ^2) random variables is Normal$(n\mu, n\sigma^2)$.

The conditional distribution of y_2 given y_1 is

- degenerate, concentrated at zero if $y_1 = 0$
- Normal$(y_1\mu, y_1\sigma^2)$, if $y_1 > 0$

So the conditional distribution of y_2 given y_1 is discrete when $y_1 = 0$ and continuous when $y_1 > 0$.

Exponential Families of Distributions

An **exponential family of distributions** is a **statistical model** having a **log likelihood** of the form

$$\langle y, \theta \rangle - c(\theta),$$

where y is a vector statistic, θ is a vector parameter of the same dimension (say d) and

$$\langle y, \theta \rangle = \sum_{i=1}^{d} y_i \theta_i.$$

A statistic y and a parameter θ that give a log likelihood of this form are called the **canonical statistic** and **canonical parameter**.

They are also called *natural parameter* and *natural statistic*, but, as elsewhere, we avoid terms of biological origin in aster model theory.

Exponential Families of Distributions (cont.)

The function c in

$$\langle y, \theta \rangle - c(\theta),$$

is called the **cumulant function** of the family. It has many important and amazing properties.

Exponential Families of Distributions (cont.)

We are using modern terminology about these models.

An older terminology would call the exponential family, the collection of all of what we are calling exponential families.

Old terminology: this statistical model is **in the** exponential family.

New terminology: this statistical model is **an** exponential family.

The old terminology has nothing to recommend it. It makes the primary term — "exponential family" — refer to a heterogeneous collection of statistical models of no interest in any application.

The new terminology describes a property that, if a statistical model has it, implies many other properties. It is a key concept of theoretical statistics.

Notational Variation

Those who insist that all vectors are really matrices (so-called column vectors and row vectors) would write the exponential family log likelihood as either

$$y^T \theta - c(\theta)$$

or

$$\theta^T y - c(\theta)$$

The $\langle \cdot , \cdot \rangle$ notation used here is more mathematical, treating vectors as vectors. It may come as a surprise to those who have not taken that many math courses, but most advanced math uses this notion rather than "vectors are really matrices".

Vectors

In aster model theory we use vectors whose indices take values in abstract sets. We do not insist the indices take values $1, \ldots, d$ for some d.

In set theory, the set of all functions $A \to B$ is denoted B^A.

Using this notation, we say our vectors are elements of $\mathbb{R}^J$ for some abstract set J rather than elements of $\mathbb{R}^d$ for some positive integer d.

With this definition comes the notion that there is no difference between vectors and functions except in notation.

A vector y in $\mathbb{R}^J$ is a function $J \to \mathbb{R}$ but we denote function evaluation y_j as is usual for components of vector rather than $y(j)$.

Exponential Families of Distributions (cont.)

When we are considering canonical statistic vectors and canonical parameter vectors as elements of $\mathbb{R}^J$ for some abstract set J, then

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

This angle brackets notation does not denote an inner product on $\mathbb{R}^J$ but rather a duality pairing placing the vector space where the canonical statistic vector lives and the vector space where the canonical parameter vector lives in duality.

The notation $\langle \,\cdot\,, \,\cdot\, \rangle$ must always have a canonical statistic vector as one argument and a canonical parameter vector as the other argument.

Aster Graph

In aster model theory and practice the word “graph” refers to two different things.

In practice, users think of the graph as the graph for a single “individual” (in scare quotes).

This the graph we specify with the arguments `pred` and `fam` to R function `aster`.

In theory, we think of the graph as the graph for all individuals, which we call the *full* aster graph for emphasis.

Aster Graph (cont.)

Usually, every “individual” has the “same” graph (more scare quotes). Pedantically “same” should be replaced by **isomorphic** since they have different variables at the nodes. They have the same shape but are not the same.

In this case, the full aster graph is just many copies of the graph for a single “individual” with the indices of the variables changed so they are all different.

But there is no requirement that “individuals” have isomorphic graphs.

R package `aster2` makes it easy to have this. R package `aster` makes it hard. To have this with R package `aster`, lie. Say there is only one “individual” which has a very large graph (the full aster graph).

Aster Graph (cont.)

There are two reasons why “individual” is in scare quotes.

Ideally the aster graph for a single “individual” goes one or more times around the life cycle ending at the same point where it started.

But it doesn’t have to start at any particular point. Thus the graph for a single “individual” may involve data on a single biological individual and its offspring and perhaps offspring of offspring (if the graph goes twice around the life cycle).

The constants at initial nodes do not have to be 1. They can be any positive integers.

When the constant at an initial node is n, then the data at the other nodes are sums of the data for n biological individuals.

Aster Graph (cont.)

For the following graph

$$\begin{tikzcd}
n \arrow[r, "\text{Ber}"] & y_1 \arrow[r, "\text{Ber}"] \arrow[d, "\text{Poi}"] & y_2 \arrow[r, "\text{Ber}"] \arrow[d, "\text{Poi}"] & y_3 \arrow[d, "\text{Poi}"] \\
 & y_4 & y_5 & y_6
\end{tikzcd}$$

where the first row (y_1, y_2, y_3) are survival indicators and the second row (y_4, y_5, y_6) are offspring counts, all of these variables are for n biological individuals.

- y_1 is the number of those n that survived year 1
- y_2 is the number of those y_1 that survived year 2
- y_3 is the number of those y_2 that survived year 3
- y_4 is the number of offspring those y_1 had in year 1
- y_5 is the number of offspring those y_2 had in year 2
- y_6 is the number of offspring those y_3 had in year 3

Dependence Groups

Aster models have a feature called **dependence groups** that is not implemented in R package `aster`. It is implemented in R package `aster2`, but that package is incomplete not having many of the features of R package `aster`.

Moreover, dependence groups have been used AFAIK in only one paper (Eck, Shaw, Geyer, and Kingsolver, *Evolution*, 2015).

Since aster models with dependence groups are harder to understand than those without, we start our discussion by omitting them.

Aster Model PMDF

In an aster model, we have a bunch of variables y_j, where $j \in N$, the index set N being the set of nodes of the (full!) graph. Since each node has at most one predecessor, we can specify the graph by a function, the **predecessor function**, that gives the predecessor for each node that has a predecessor.

Let J denote the set of non-initial nodes of the graph. Then the predecessor function is a function $p : J \to N$ such that $p(j)$ is the predecessor of j.

Aster Model PMDF (cont.)

In an aster model, the graph specifies the joint PMDF in factorized form, each arrow in the graph corresponds to a conditional distribution in the factorization

$$f_\theta(y) = \prod_{j \in J} f_{j,\theta}(y_j | y_{p(j)})$$

Note that only variables y_j for $j \in J$ appear “in front of the bar” in a conditional. So only those variables are treated as random. Variables y_j for $j \in N \setminus J$ are conditioned on, which is the same as being treated as constant.

In short, variables at non-initial nodes are random, those at initial nodes are constant.

I claim this is a valid factorization, with what purports to be conditional distributions actually being conditional distributions.

Factorization

In “classical” or “master’s level” probability theory, we factor joint distributions into products of marginal and conditional distributions

$$f_\theta(x, y) = f_\theta(y \mid x) f_\theta(x)$$

This also works when x and y are vectors.

When x is a vector, we can apply factorization again

$$f_\theta(y_1, y_2, y_3) = f_\theta(y_1 \mid y_2, y_3) f_\theta(y_2 \mid y_3) f_\theta(y_3)$$

and again

$$f_\theta(y_1, y_2, y_3, y_4) = f_\theta(y_1 \mid y_2, y_3, y_4) f_\theta(y_2 \mid y_3, y_4) f_\theta(y_3 \mid y_4) f_\theta(y_4)$$

and so forth. And all of the variables here can also be vectors.

Factorization (cont.)

In a factorization like

$$f_\theta(y_1, y_2, y_3, y_4) = f_\theta(y_1 \mid y_2, y_3, y_4) f_\theta(y_2 \mid y_3, y_4) f_\theta(y_3 \mid y_4) f_\theta(y_4)$$

it may be that some variables (or vectors) “behind the bar” can be omitted because the conditional distribution does not happen to actually depend on those variables. In aster models we only ever have one variable (the predecessor) “behind the bar”.

Theorem (Valid Factorization)

A factorization of a joint distribution as a product of marginals and conditionals is valid if and only if there exists a total ordering of the variables such that

- *every variable occurs at most once “in front of the bar” in a conditional, and*
- *every variable “behind the bar” in a conditional comes after (in this total ordering) every variable “in front of the bar” in that conditional.*

Aster Model PMDF (cont.)

So an aster factorization

$$f_\theta(y) = \prod_{j \in J} f_{j,\theta}(y_j | y_{p(j)})$$

is valid if and only if there is a total ordering such that predecessors come after successors.

In theory, the valid factorization theorem is satisfied if and only if the aster graph is acyclic.

There is an algorithm called **topological sort** (implemented in R package `pooh`) that either finds a total ordering compatible with the graph (predecessors before successors) or discovers that no such ordering exists (in which case the graph is not acyclic).

Aster Model PMDF (cont.)

In practice, R function `aster` does not use this algorithm but rather forces the user to find such a total order.

R function `aster` requires that its argument `pred` always has predecessors before successors. So the inverse total order always has predecessors after successors.

Exponential Families and IID

Suppose we have an exponential family with log likelihood

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

and we observe z_1, ..., z_n independent and identically distributed (IID) from this family.

Then, because of independence, the joint is the product of the marginals, and because log of product is sum of logs, the log likelihood is

$$\sum_{i=1}^{n} \bigl[\langle z_i, \theta \rangle - c(\theta) \bigr] = \left\langle \sum_{i=1}^{n} z_i, \theta \right\rangle - n c(\theta)$$

and we just get another exponential family with canonical statistic $\sum_{i=1}^{n} z_i$, canonical parameter θ, and cumulant function $\theta \mapsto n c(\theta)$.

Exponential Families and IID (cont.)

Many "addition rules" from math stats are a consequence.

Sum of n IID Bernoulli(p) random variables is binomial(n, p).

Sum of n IID Geometric(p) random variables is negative-binomial(n, p).

Sum of n IID Poisson(μ) random variables is Poisson($n\mu$).

Sum of n IID Normal(μ, σ^2) random variables is Normal($n\mu, n\sigma^2$).

Predecessor is Sample Size (cont.)

Recall from deck 1 the **predecessor is sample size property**

For any arrow

$$y_{p(j)} \longrightarrow y_j$$

y_j is the sum of y_i independent and identically distributed (IID) random variables having the distribution named by the arrow label (by convention, a sum with zero terms is zero).

Now we make another assumption, the **exponential family assumption**, that $y_j = z_1 + \cdots + z_{y_{p(j)}}$, where the z_i are IID realizations of the canonical statistic of the one-dimensional exponential family with cumulant function c_j and canonical parameter θ_j. (The random variable y_j is a random sum of random variables with $y_{p(j)}$ terms in the sum.)

Summary of Assumptions

Graphical Axioms

Directed Edges of the graph are directed (arrows).

Acyclic The graph is acyclic: a path that follows arrows never returns to a node.

At Most One Predecessor Each node of the graph has at most one predecessor. Initial nodes have none. Non-initial nodes have one. If j is non-initial, $p(j)$ is its predecessor.

Summary of Assumptions (cont.)

Statistical Axioms

Factorization

$$f_\theta(y) = \prod_{j \in J} f_{j,\theta}(y_j | y_{p(j)})$$

Predecessor is Sample Size If j is non-initial, then $y_j = z_1 + \cdots + z_{y_{p(j)}}$ (a random sum of random variables). By convention, a sum with zero terms is zero, so $y_{p(j)} = 0$ implies $y_j = 0$.

Exponential Family In $y_j = z_1 + \cdots + z_{p(j)}$ the distribution of the z_k is one-parameter exponential family with canonical statistic z_k, canonical parameter θ_j, and cumulant function c_j.

Aster Log Likelihood

This means — using the rule that the sum of IID random variables from an exponential family is another exponential family and the cumulant function for the latter is n times the cumulant function for the former, where n is the sample size — the conditional distribution of y_j given $y_{p(j)}$ is one-parameter exponential family with canonical statistic y_j, canonical parameter θ_j, and cumulant function $\theta_j \mapsto y_{p(j)} c_j(\theta_j)$.

In $y_{p(j)} c_j(\theta_j)$ the sample size is $y_{p(j)}$ (predecessor is sample size) and $c_j(\theta_j)$ is the cumulant function for "the former", that is, for each of the $y_{p(j)}$ IID random variables whose sum is y_j.

Aster Log Likelihood (cont.)

Hence the aster model log likelihood is

$$
\begin{aligned}
l(\theta) &= \log \left(\prod_{j \in J} f_{j,\theta}(y_j | y_{p(j)}) \right) - \text{constant} \\
&= \sum_{j \in J} \log f_{j,\theta}(y_j | y_{p(j)}) - \text{constant} \\
&= \sum_{j \in J} \bigl[y_j \theta_j - y_{p(j)} c_j(\theta_j) \bigr]
\end{aligned}
$$

where the "minus a constant" (that does not depend on the parameters) accounts for the fact that such constants can be dropped in going from log PMDF to log likelihood.

Aster Log Likelihood (cont.)

Do we need to do anything special to handle cases where the predecessor is zero (which implies the successor is also zero)?

$$l(\theta) = \sum_{j \in J} \bigl[y_j \theta_j - y_{p(j)} c_j(\theta_j) \bigr]$$

No. Such terms do contribute zero to the log likelihood. But that is exactly what they should do. The conditional distribution of y_j given $y_{p(j)} = 0$ is degenerate and concentrated at zero. That is

$$\Pr(y_j = 0 | y_{p(j)} = 0) = 1$$

and $\log(1) = 0$, so this arrow should contribute zero to the log likelihood.

Probability theory “just works”. We don’t have to do contortions to make it work.

Aster Log Likelihood (cont.)

Although each term in

$$l(\theta) = \sum_{j \in J} \bigl[y_j \theta_j - y_{p(j)} c_j(\theta_j) \bigr]$$

has exponential family form, the whole log likelihood does not because both the y_j and $y_{p(j)}$ in each term may be random.

However, because this is linear in the y's, this must be a joint exponential family with canonical statistic vector y. We just don't (yet) know the canonical parameter vector and cumulant function.

Aster Log Likelihood (cont.)

Let φ in $\mathbb{R}^J$ be the canonical parameter vector. Then the log likelihood for this parameterization has the form

$$l(\varphi) = \left[\sum_{j \in J} y_j \varphi_j \right] - c(\varphi)$$

where c is the cumulant function for the joint exponential family.

Aster Log Likelihood (cont.)

To identify the joint canonical parameters, we must rewrite the log likelihood collecting terms that multiply the same component of the canonical statistic

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

Aster Log Likelihood (cont.)

Thus an aster model is (jointly) an exponential family with canonical statistic vector y, canonical parameter vector φ having components

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

and cumulant function

$$c(\varphi) = \sum_{\substack{k \in J \\ p(k) \notin J}} y_{p(k)} c_k(\theta_k)$$

(note that all of the $p(k)$ in the later formula are initial nodes so all of the $y_{p(k)}$ in this formula are constants, so this does define a deterministic function rather than a random function).

The Aster Transform

I claim the change of parameter

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

is invertible. To invert it, just isolate θ_j obtaining

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

How is that an inversion? It still has thetas on the right-hand side!

Use $(*)$ in an order that calculates θ_j for successors before θ_j for predecessors. Then it works because when we use it to calculate θ_j we have already calculated all of the θ_k such that $p(k) = j$.

The Aster Transform (cont.)

Is there such an order? Yes there is by the acyclicity property (found by the topological sort algorithm).

Note that at terminal nodes we have $\theta_j = \varphi_j$. But we do not have this at non-terminal nodes.

We call this invertible change of parameter $\theta \longleftrightarrow \varphi$ the **aster transform** (pedantically, $\theta \longrightarrow \varphi$ is the aster transform and $\varphi \longrightarrow \theta$ is the inverse aster transform).

The Aster Transform (cont.)

Are you lost? If so, no surprise.

The aster transform makes mathematical-statistical-theoretical sense, but it doesn't make common sense. It is not intuitive at all.

To understand it we must apply Zen and not try to understand it.

If that doesn't make sense, wait a while. We hope you will eventually achieve enlightenment.

The technical report *A Philosophical Look at Aster Models* goes through one very simple example, but it only shows the algebraic formulas are a big mess that no one can understand intuitively. (The whole point of the example is to show you that you do not want to try to understand the aster transform by staring at the formulas.)

The Aster Transform (cont.)

A quote from my master's level theory notes

> Parameters are meaningless quantities. Only probabilities and expectations are meaningful.

Of course, some parameters are probabilities and expectations, but most exponential family canonical parameters are not.

A quote from *Alice in Wonderland*

> 'If there's no meaning in it,' said the King, 'that saves a world of trouble, you know, as we needn't try to find any.'

Realizing that canonical parameters are meaningless quantities "saves a world of trouble". We "needn't try to find any".

The Aster Transform (cont.)

How are we to distinguish θ and φ? They are both canonical parameters of a sort.

We call θ the **conditional canonical parameter vector** and φ the **unconditional canonical parameter vector**, despite this suggesting more parallelism than is really there.

Pedantically, θ is the vector having components θ_j that are the canonical parameters for the conditional distributions associated with the arrows $p(j) \longrightarrow j$ in the graph.

Pedantically, φ is the canonical parameter vector of the joint distribution of the aster model (which is an exponential family).

The Aster Transform (cont.)

Each θ_j is the canonical parameter of a one-parameter exponential family model (for one arrow). The vector θ is not a canonical parameter vector of a multivariate exponential family.

The vector φ is the canonical parameter vector of a multivariate exponential family. Each φ_j is not a canonical parameter of a one-parameter exponential family.

θ is componentwise canonical but not vectorwise canonical.

φ is vectorwise canonical but not componentwise canonical.

Dependence Groups (cont.)

We now want to redo everything we have done since the last mention of dependence groups, changing things to allow for dependence groups.

First we explain why dependence groups.

In an aster model without dependence groups, nodes are conditionally independent given their predecessors; this is inherent in the factorization. There are at least two good reasons to relax this assumption: multinomial families and two-parameter normal families.

Dependence Groups (cont.)

At terminal nodes we are allowed to have continuous random variables (at non-terminal nodes we must have nonnegative-integer-valued random variables).

But the rule that the conditional distribution be one-parameter exponential family means we can only have a one-parameter normal family. That is why R package `aster` only provides normal location families (the mean is an unknown parameter, the variance is known).

But in real life the variance is never known. So this is a problem.

We cannot fit two-parameter normal families into aster models without dependence groups.

One might think that introducing scale parameters the way GLM do is the right way to deal with two-parameter normal families, but that would us outside of exponential families and the aster transform would no longer work.

Dependence Groups (cont.)

The two-parameter normal distribution with its usual data and parameterization has log likelihood

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

and we see this has exponential family form with canonical statistic vector

$$y = (x, x^2)$$

and canonical parameter vector

$$\theta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right)$$

Of course, the dimension of the canonical statistic vector and canonical parameter vector must be the same.

Dependence Groups (cont.)

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

A multinomial distribution is degenerate. It is the distribution of the random vector of category counts for IID individuals classified into categories with the categories being mutually exclusive and exhaustive, so every individual goes in exactly one category, and the category counts sum to the sample size.

If the sample size is $n = 1$ and there are k categories, then a multinomial random vector serves as a k-way switch. Exactly one component is equal to one, and the rest are equal to zero. It indicates the category into which the individual is classified.

This k-way switch can be useful in life history with, for example, animals with life history stages, such as insect larva, pupa, and adult.

Dependence Groups (cont.)

In short, multinomial dependence groups allow aster models to incorporate capture-recapture as well as survival.

Dependence Groups (cont.)

In most of statistics we do not want degenerate random vectors. So we drop one of the components of a multinomial random vector to get a non-degenerate random vector.

With aster models we cannot do that. Random variables correspond to nodes of the graph. Dropping random variables changes the graph.

If we were to drop one category, a k-way switch would become $(k-1)$-way. Not what was needed.

Hence we have to deal with exponential families that have degenerate distributions of their canonical statistics and hence nonidentifiable canonical parameterizations. (More on this later.)

Vectors and Subvectors

Recall that we take vectors to be elements of $\mathbb{R}^J$ for some abstract set J (in aster models the set of non-initial nodes of the full aster graph).

We now need a way to refer not just to components of such vectors but to groups of components corresponding to dependence groups.

If $y \in \mathbb{R}^J$ and $G \subset J$, then y_G denotes the **subvector** having components y_j for $j \in G$.

Recall that $y \in \mathbb{R}^J$ means y is a function $J \to \mathbb{R}$ that maps $j \mapsto y_j$.

Similarly, y_G is a function $G \to \mathbb{R}$ that maps $j \mapsto y_j$. Hence $y_G \in \mathbb{R}^G$.

We have nothing to distinguish components y_j from subvectors y_G except for the convention that upper case letters denote sets and lower case letters denote elements of sets.

Aster Model PMDF (cont.)

Now our dependence groups can be subvectors.

In order for predecessor is sample size to hold for a dependence group the whole group must have the same predecessor.

Denote the predecessor of dependence group G by $q(G)$.

Dependence groups must be disjoint, hence they must form a partition $\mathcal{G}$ of the set J of non-initial nodes.

This predecessor function is a function $q : \mathcal{G} \to N$ that maps $G \mapsto q(G)$.

Aster Model PMDF (cont.)

The reason why we picked a new letter q for this predecessor function is that we also have the old predecessor function p. The relation between the two is

$$j \in G \in \mathcal{G} \text{ implies } p(j) = q(G)$$

q determines p because every $j \in J$ is in a unique $G \in \mathcal{G}$ because $\mathcal{G}$ is a partition of J.

But p obviously does not determine q because p doesn't know anything about dependence groups.

When we need words to distinguish our two predecessor functions we call

- p the **node-to-node** predecessor function and
- q the **set-to-node** predecessor function.

Aster Model PMDF (cont.)

Now we rewrite the aster model factorization as

$$f_\theta(y) = \prod_{G \in \mathcal{G}} f_{G,\theta}(y_G | y_{q(G)})$$

Everything works almost the same as before. The only conceptual difficulty is that our indices range over a family of sets (which, of course, is also a set). Once we get used to that, everything else is more or less the same as without dependence groups.

Note that the previous theory without dependence groups is the special case of our new theory where every element of $\mathcal{G}$ is a singleton set.

Aster Model PMDF (cont.)

So an aster factorization

$$f_\theta(y) = \prod_{G \in \mathcal{G}} f_{G,\theta}(y_G | y_{q(G)})$$

is valid (by the same valid factorization theorem as before) if there is a total ordering on $\mathcal{G}$ such that predecessors come after successors, that is if $q(G) \in G'$ then G' comes after G in this total ordering.

In theory, the topological sort algorithm can find such a total order or prove that none exists.

In practice, R package `aster2` forces the user to order the nodes of the graph so that $q(G)$ comes before any $j \in G$.

Aster Log Likelihood (cont.)

Now the aster model log likelihood is

$$
\begin{aligned}
l(\theta) &= \log \left(\prod_{G \in \mathcal{G}} f_{G,\theta}(y_G | y_{q(G)}) \right) - \text{constant} \\
&= \sum_{G \in \mathcal{G}} \log f_{G,\theta}(y_G | y_{q(G)}) - \text{constant} \\
&= \sum_{G \in \mathcal{G}} \left[\langle y_G, \theta_G \rangle - y_{q(G)} c_G(\theta_G) \right] \\
&= \left(\sum_{j \in J} y_j \theta_j \right) - \left(\sum_{G \in \mathcal{G}} y_{q(G)} c_G(\theta_G) \right)
\end{aligned}
$$

the last equality using $\mathcal{G}$ is a partition of J.

Aster Log Likelihood (cont.)

Again we observe that the joint distribution is an exponential family with canonical statistic y.

So to identify the joint canonical parameters, we must rewrite the log likelihood collecting terms that multiply the same component of the canonical statistic

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

Aster Log Likelihood (cont.)

Thus an aster model with dependence groups is (jointly) an exponential family with canonical statistic vector y, canonical parameter vector φ having components

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

and cumulant function

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

(note that all of the $q(G)$ in the later formula are initial nodes so all of the $y_{q(G)}$ in this formula are constants, so this does define a deterministic function rather than a random function).

The Aster Transform (cont.)

The **aster transform** is now given by

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

and again is invertible. To invert it, just isolate θ_j obtaining

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

Use $(**)$ in an order that calculates θ_j for successors before θ_j for predecessors. Then it works because when we use it to calculate θ_j we have already calculated all components of θ_G such that $q(G) = j$.

Aster Graph (cont.)

Now we want a set of axioms as before, but we have lost track of what the graph should be.

Fortunately, we don't have to invent a new theory of graphical models because we can find one already done for us in textbooks on graphical models like Lauritzen (1996) *Graphical Models*, Oxford University Press.

Section 2.1 of that book says that general graphs have two kinds of edges

- **directed edges**, also called **arrows**, and
- **undirected edges**, also called **lines**.

Aster Graph (cont.)

Section 3.2.3 of Lauritzen (1996) covers the most general factorization of a joint distribution into a product of marginals and conditionals, which he calls **chain graph models**. He calls **chain components** what we call **dependence groups**. His equation (3.23) is our

$$f_\theta(y) = \prod_{G \in \mathcal{G}} f_{G,\theta}(y_G | y_{q(G)})$$

So we only need to match up our theory with his. The result is that a general aster graph has

- an arrow $p(j) \longrightarrow j$ for every $j \in J$ (as before), and
- a line j — k for every pair of distinct nodes j and k in the same dependence group.

Exception

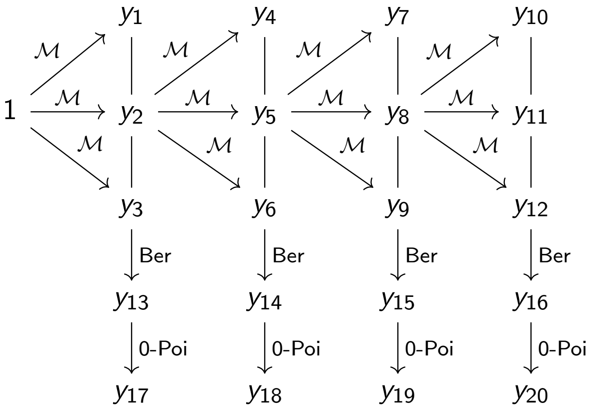
Except when we draw aster graphs with dependence groups we don't draw all the lines — only enough lines so each dependence group is connected.

This makes our graphs less cluttered and provides the same information.

In theory, the dependence groups can be found using Rem's algorithm for the recording of equivalence classes which done by R function `weak` in R package `pooh`.

In practice, R function `asterdata` in R package `aster2` makes the user specify the dependence groups so it does not need to use this algorithm.

A Graph With Dependence Groups



Here $\mathcal{M}$ stands for multinomial. The dependence groups are $\{1, 2, 3\}$, $\{4, 5, 6\}$, $\{7, 8, 9\}$, $\{10, 11, 12\}$, and every other node is a dependence group by itself.

A Graph With Dependence Groups (cont.)

In the graph on the preceding slide

$y_{\{1,2,3\}}$ is multinomial with sample size 1.

$y_{\{4,5,6\}}$ is conditionally multinomial with sample size y_2.

$y_{\{7,8,9\}}$ is conditionally multinomial with sample size y_5.

$y_{\{10,11,12\}}$ is conditionally multinomial with sample size y_8.

y_{13} is conditionally binomial with sample size y_3.

And so forth. All of the arrows that do not involve dependence groups with more than one node work just like arrows in a graph without dependence groups.

A Graph With Dependence Groups (cont.)

The intended application for this graph is life history of an insect.

The “columns” of the graph are for different days.

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 the mating success.

Nodes in the bottom “row” of this graph (y_{17} through y_{20}) count number of eggs laid.

A Graph With Dependence Groups (cont.)

In some ways this graph is very different from the way graphs without dependence groups work. Here death is a node. $y_1 = 1$ indicates death (or $y_4 = 1$, etc.). In the graph for Example One in Deck 1 $y_1 = 0$ indicates death (or $y_2 = 0$, etc.).

Here death is one category in a multinomial switch. The categories are dead, larva, adult.

The individual can move from larva at one day to any of these categories the next day. But it cannot move from death to anywhere. Nor can it move from adult to anywhere. The adults live a few days at most and either reproduce or not, and it is reproduction that is recorded (with zero inflation).

The Magic of Cumulant Functions

If

$$l(\varphi) = \langle y, \varphi \rangle - c(\varphi)$$

is the log likelihood of an exponential family, then we can write the ratio of the PMDF for φ and another parameter value φ^* as

$$e^{l(\varphi) - l(\varphi^*)}$$

because $l(\varphi)$ is the log of the PMDF for φ except, perhaps, some additive terms not containing φ that may have been dropped from the log likelihood. But, since any dropped terms do not depend on the parameter, they are the same for φ and φ^* and cancel in $l(\varphi) - l(\varphi^*)$.

The Magic of Cumulant Functions (cont.)

Thus

$$E_{\varphi^*}\left\{e^{l(\varphi)-l(\varphi^*)}\right\} = 1$$

(probabilities must sum-integrate to one). And this is

$$E_{\varphi^*}\left\{e^{\langle Y, \varphi - \varphi^* \rangle - c(\varphi) + c(\varphi^*)}\right\} = 1$$

or

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

If we think of φ as variable and φ^* as fixed, then this determines $c(\varphi)$ for all φ up to an unknown additive constant $c(\varphi^*)$, which can be dropped from log likelihoods.

The Magic of Cumulant Functions (cont.)

More precisely,

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

determines the cumulant function if the expectation exists. If the expectation does not exist, then we give $c(\varphi)$ the value ∞ so it is defined for all vectors φ.

Let

$$\Phi = \{ \varphi : c(\varphi) < \infty \}$$

We say the set Φ is the **canonical parameter space** of the **full** exponential family (containing the originally given exponential family if it was not full).

Any new distributions added to the family have ratios of their PMDF to the PMDF for parameter value φ^*

$$e^{\langle y, \varphi - \varphi^* \rangle - c(\varphi) + c(\varphi^*)}$$

just like the distributions in the originally given family.

Moment Generating Functions

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

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

(φ is the parameter vector for the distribution of Y) *provided that this expectation is finite for all t in some neighborhood of zero* (otherwise, we say Y does not have an MGF).

Moment Generating Functions (cont.)

The reason for the name is because ordinary moments can be computed by differentiating the MGF and evaluating the derivatives at $t = 0$

$$
\begin{aligned}
E_\varphi(Y_i) &= \left. \frac{\partial M_\varphi(t)}{\partial t_i} \right|_{t=0} \\
E_\varphi(Y_i Y_j) &= \left. \frac{\partial^2 M_\varphi(t)}{\partial t_i \partial t_j} \right|_{t=0} \\
E_\varphi(Y_i Y_j Y_k) &= \left. \frac{\partial^2 M_\varphi(t)}{\partial t_i \partial t_j \partial t_k} \right|_{t=0}
\end{aligned}
$$

and so forth.

Moment Generating Functions (cont.)

The reason why this works is "differentiation under the integral sign"

$$
\begin{aligned}
\frac{\partial M_\varphi(t)}{\partial t_i} &= \frac{\partial}{\partial t_i} E_\varphi \left\{ e^{\langle Y, t \rangle} \right\} \\
&= E_\varphi \left\{ \frac{\partial}{\partial t_i} e^{\langle Y, t \rangle} \right\} \\
&= E_\varphi \left\{ Y_i e^{\langle Y, t \rangle} \right\}
\end{aligned}
$$

(the middle equality being "differentiation under the integral sign" although, of course, the expectation may be a combination of summation and integration or even all summation). Setting $t = 0$ gives

$$
\left. \frac{\partial M_\varphi(t)}{\partial t_i} \right|_{t=0} = E_\varphi(Y_i)
$$

Moment Generating Functions (cont.)

Differentiation under the integral sign does not always work, but it is a theorem of MGF theory that it always does work for MGF (this is a theorem of measure-theoretic probability that uses the so-called dominated convergence theorem).

And now we see the reason for requirement that $M_\varphi(t)$ be finite for all t in some neighborhood of zero. We need it in order for partial derivatives at zero to exist. And we don't care about these partial derivatives existing at any other point.

Cumulant Generating Functions

The log of an MGF is called a **cumulant generating function** (CGF) and its partial derivatives evaluated at zero are called **cumulants**

$$
\begin{aligned}
\kappa_i &= \frac{\partial \log M_\varphi(t)}{\partial t_i}\bigg|_{t=0} \\
\kappa_{ij} &= \frac{\partial^2 \log M_\varphi(t)}{\partial t_i \partial t_j}\bigg|_{t=0} \\
\kappa_{ijk} &= \frac{\partial^2 \log M_\varphi(t)}{\partial t_i \partial t_j \partial t_k}\bigg|_{t=0}
\end{aligned}
$$

and so forth.

The cumulants of order m are polynomial functions of the ordinary moments up to order m and vice versa. The actual formulas can be found in comprehensive textbooks of mathematical statistics.

Cumulant Generating Functions (cont.)

We are only interested in the first two cumulants

$$\begin{aligned} E_\varphi(Y_i) &= \left. \frac{\partial \log M_\varphi(t)}{\partial t_i} \right|_{t=0} \\ \operatorname{cov}_\varphi(Y_i, Y_j) &= \left. \frac{\partial^2 \log M_\varphi(t)}{\partial t_i \partial t_j} \right|_{t=0} \end{aligned}$$

or, rewriting these as vector and matrix equations

$$\begin{aligned} E_\varphi(Y) &= \nabla \log M_\varphi(0) \\ \operatorname{var}_\varphi(Y) &= \nabla^2 \log M_\varphi(0) \end{aligned}$$

Cumulant Generating Functions (cont.)

In

$$E_\varphi(Y) = \nabla \log M_\varphi(0)$$

the left-hand side denotes the **mean vector**, which has components $E_\varphi(Y_i)$ and the right-hand side denotes the **gradient vector**, which has components $\partial \log M_\varphi(t) / \partial t_i$ evaluated at $t = 0$.

In

$$\operatorname{var}_\varphi(Y) = \nabla^2 \log M_\varphi(0)$$

the left-hand side denotes the **variance matrix**, which has components $\operatorname{cov}_\varphi(Y_i, Y_j)$ and the right-hand side denotes the **hessian matrix**, which has components $\partial^2 \log M_\varphi(t) / \partial t_i \partial t_j$ evaluated at $t = 0$.

The variance matrix is also called the **covariance matrix**, the **variance-covariance matrix**, or the **dispersion matrix**.

The Magic of Cumulant Functions (cont.)

What is the CGF of an exponential family?

The MGF is

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

provided this satisfies the condition to be an MGF, that is, provided that φ is an interior point of Φ.

The Magic of Cumulant Functions (cont.)

An exponential family is **regular** if its full canonical parameter space Φ is an open set. For a regular exponential family

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

is an MGF for all $\varphi \in \Phi$.

And the cumulant function is

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

The Magic of Cumulant Functions (cont.)

And the first two cumulants are

$$\begin{aligned} \nabla K_\varphi(0) &= \nabla c(\varphi + t)\big|_{t=0} = \nabla c(\varphi) \\ \nabla^2 K_\varphi(0) &= \nabla^2 c(\varphi + t)\big|_{t=0} = \nabla^2 c(\varphi) \end{aligned}$$

derivatives of the CGF evaluated at zero are derivatives of the cumulant function c evaluated at φ.

In short

$$\begin{aligned} E_\varphi(Y) &= \nabla c(\varphi) \\ \operatorname{var}_\varphi(Y) &= \nabla^2 c(\varphi) \end{aligned}$$

This is tremendously important with lots of consequences.

The Magic of Cumulant Functions (cont.)

Do aster models have this magic? The only requirement we needed is that the exponential family be full and regular. So the question becomes are aster models full and regular?

The answer is yes, provided all the exponential families for dependence groups are full and regular.

But the proof is somewhat complicated, so we have put it in a separate file in the "Course Notes" section of the course web site.

Convex Sets

A **convex set** of vectors is a set S having the property that for any two points x_1 and x_2 in the set, the entire line segment with these points as end points is also in the set, that is,

$$tx_1 + (1-t)x_2 \in S, \qquad 0 < t < 1$$

The Extended Real Number System

$\mathbb{R}$ denotes the real number system.

$\overline{\mathbb{R}}$ denotes the **extended real number system**.

As sets, $\overline{\mathbb{R}}$ is $\mathbb{R}$ with two points added, which are denoted $+\infty$ and $-\infty$.

To make a number system out of $\overline{\mathbb{R}}$, we need to specify

- its ordering,
- its arithmetic, and
- its topology.

The Extended Real Number System (cont.)

The ordering is obvious $-\infty < x < +\infty$ for $x \in \mathbb{R}$ and the usual ordering on $\mathbb{R}$.

Most of the arithmetic is obvious

$$
\begin{aligned}
x + \infty &= \infty, &\qquad& x \neq -\infty \\
x \cdot \infty &= \infty, && x > 0 \\
x \cdot \infty &= -\infty, && x < 0
\end{aligned}
$$

and so forth.

But there are no obvious definitions of $\infty - \infty$ or $0 \cdot \infty$.

People adopt different conventions in different contexts or just leave them undefined, like divide by zero in $\mathbb{R}$.

The Extended Real Number System (cont.)

The topology can be described by defining neighborhoods.

A set is a neighborhood of $x \in \mathbb{R}$ if it contains the interval $(x - \varepsilon, x + \varepsilon)$ for some $\varepsilon > 0$.

A set is a neighborhood of $+\infty$ if it contains the interval $(x, +\infty]$ for some $x \in \mathbb{R}$.

A set is a neighborhood of $-\infty$ if it contains the interval $[-\infty, x)$ for some $x \in \mathbb{R}$.

The Extended Real Number System (cont.)

This topology is metrizable. Take any bounded increasing function on $\mathbb{R}$, and extend it to $\overline{\mathbb{R}}$ by taking limits. The arc tangent function atan will do. $\operatorname{atan}(\pm\infty) = \pm\pi/2$. Now take the distance between x and y in $\overline{\mathbb{R}}$ to be $|\operatorname{atan}(x) - \operatorname{atan}(y)|$.

$\overline{\mathbb{R}}$ is a **compact** metrizable space. Every sequence has a convergent subsequence. If the sequence is bounded, then it has a subsequence that converges by the Bolzano-Weierstrass theorem. If the sequence is unbounded, then it has a subsequence that converges to $+\infty$ or to $-\infty$.

Convex Functions

An extended-real-valued function f on a vector space is **convex** if

$$f\bigl(tx + (1-t)y\bigr) \le tf(x) + (1-t)f(y),$$
$$\text{whenever } f(x) < \infty,\ f(y) < \infty, \text{ and } 0 < t < 1$$

The restrictions $f(x) < \infty$, $f(y) < \infty$, and $0 < t < 1$ avoid any possibility of $\infty - \infty$ or $0 \cdot \infty$.

If f is an extended-real-valued convex function, then the set

$$\operatorname{dom} f = \{\, x : f(x) < \infty \,\}$$

is called its **effective domain**. Of course, the *domain* of f is the whole vector space on which it is defined.

It follows immediately from the definition that $\operatorname{dom} f$ is a convex set.

Strictly Convex Functions, Proper Convex Functions

A convex extended-real-valued function f on a vector space is **strictly convex** if

$$f\bigl(tx + (1-t)y\bigr) < tf(x) + (1-t)f(y), \qquad \text{whenever } f(x) \in \mathbb{R},\ f(y) \in \mathbb{R}, \text{ and } 0 < t < 1$$

A convex extended-real-valued function f on a vector space is **proper** if

- it is not everywhere equal to $+\infty$, and
- it is not anywhere equal to $-\infty$.

The Magic of Cumulant Functions (cont.)

Theorem

The cumulant function of an exponential family is a proper convex extended-real-valued function. It is strictly convex unless the probability distributions of the canonical statistic are concentrated on a hyperplane.

Proof.

Hölder's inequality and the conditions for equality in Hölder's inequality. □

Concave Functions

A function f is **concave** if and only if $-f$ is convex.

Stand on your head and convex becomes concave and vice versa.

A function f is **strictly concave** if and only if $-f$ is strictly convex.

A concave function f is **proper** if and only if $-f$ is proper.

The **effective domain** of a concave function f is $\operatorname{dom}(-f)$.

The main virtue of convex functions is in minimization.

The main virtue of concave functions is in maximization.

Concave Functions (cont.)

The log likelihood of an exponential family is a proper concave extended-real-valued function. It is strictly concave unless the the probability distributions of the canonical statistic are concentrated on a hyperplane.

Local and Global Maximizers and Minimizers

A point x is a **global minimizer** of a function f if

$$f(x) \leq f(y), \qquad \text{for all } y$$

A point x is a **local minimizer** of a function f if

$$f(x) \leq f(y), \qquad \text{for all } y \text{ in some neighborhood of } x$$

Local and Global Maximizers and Minimizers (cont.)

Theorem

Every local minimizer of a convex function is a global minimizer.

Proof.

Let f be a convex function. If it is everywhere equal to $+\infty$, then every point is a global minimizer. If it is anywhere equal to $-\infty$, then that point is a global minimizer. Otherwise f is proper. So suppose $f(x)$ is finite. For any other point y such that $f(y) < f(x)$, we have

$$f\bigl(tx + (1-t)y\bigr) \le tf(x) + (1-t)f(y) < f(x)$$

for all $t \in (0,1)$. Thus if x is not a global minimizer is it not a local minimizer either. □

Local and Global Maximizers and Minimizers (cont.)

Theorem

A proper strictly convex function has at most one local minimizer.

If a local minimizer exists, then it is the unique global minimizer by this theorem and the preceding theorem.

Proof.

Assume to get a contradiction that x and y are distinct local minimizers. By the preceding theorem, they are global minimizers. So $f(x) = f(y)$. By definition of strict convexity, for $0 < t < 1$,

$$f\bigl(tx + (1-t)y\bigr) < tf(x) + (1-t)f(y) = f(x) = f(y)$$

but this contradicts x and y being local minimizers. It follows (proof by contradiction) that the assumption that two local minimizers exist is false. □

Local and Global Maximizers and Minimizers (cont.)

Theorem

A convex function f that is finite and differentiable on an open convex set O satisfies

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \qquad x, y \in O.$$

This is part (b) of Theorem 2.14 in Rockafellar and Wets *Variational Analysis*.

Corollary

For a proper convex function whose effective domain is an open set and which is differentiable on that open set, every point where the derivative is zero is a global minimizer.

Local and Global Maximizers and Minimizers (cont.)

Corollary

For a regular full exponential family, a necessary and sufficient condition that a parameter vector (globally) maximize the log likelihood is that the first derivative of the log likelihood is zero there.

If the probability distributions of the canonical statistic are not concentrated on a hyperplane (so the log likelihood is strictly concave) the maximizer is unique if it exists.

Local and Global Maximizers and Minimizers (cont.)

A local minimizer or maximizer need not exist.

Consider the convex function exp defined on $\mathbb{R}$ considered as a one-dimensional vector space. Its infimum is zero, but $\exp(x) > 0$ for all x.

The phenomenon does occur in aster models and other exponential family models. Much more on this later (Deck 9).

Mean Value Parameterizations

The map h defined by

$$h(\varphi) = \nabla c(\varphi) = E_\varphi(Y), \qquad \varphi \in \Phi$$

maps the canonical parameter vector φ of a regular full exponential family to the **mean value parameter vector** $\mu = h(\varphi)$.

Of course, we don't yet know that μ parameterizes the family. So calling it a parameter is premature.

Mean Value Parameterizations (cont.) I

Theorem

In a regular full exponential family no two distributions have the same mean vector.

Hence the mean value parameterization not only **is** a parameterization, it is an **identifiable** one.

A parameterization is **identifiable** if each distinct parameter (vector) value corresponds to a distinct distribution.

Proof

Suppose $\nabla c(\theta_1) = \nabla c(\theta_2) = \mu$, and consider the function l_μ defined by

$$l_\mu(\theta) = \langle \mu, \theta \rangle - c(\theta)$$

(which is just like the log likelihood except that the data vector y is replaced by a possible mean vector μ). Then l_μ is concave just

Mean Value Parameterizations (cont.) II

like the log likelihood. Since ∇l_μ is zero at θ_1 and θ_2, they are both global maximizers. By definition of concavity, every point in the interval (θ_1, θ_2) is also a global maximizer. Consequently

$$r(t) = l_\mu\bigl(t\theta_1 + (1-t)\theta_2\bigr)$$

is constant on $(0, 1)$, hence

$$r'(t) = \bigl\langle \mu - \nabla c\bigl(t\theta_1 + (1-t)\theta_2\bigr), \theta_1 - \theta_2 \bigr\rangle$$

and

$$r''(t) = -\bigl\langle \theta_1 - \theta_2, \nabla^2 c\bigl(t\theta_1 + (1-t)\theta_2\bigr)(\theta_1 - \theta_2) \bigr\rangle$$

are both zero when $0 < t < 1$. But we also know that

$$\begin{aligned} r''(t) &= -\bigl\langle \theta_1 - \theta_2, \operatorname{var}_{t\theta_1 + (1-t)\theta_2}(Y)(\theta_1 - \theta_2) \bigr\rangle \\ &= -\operatorname{var}_{t\theta_1 + (1-t)\theta_2}(\langle Y, \theta_1 - \theta_2 \rangle) \end{aligned}$$

Mean Value Parameterizations (cont.) III

so $\langle Y, \theta_1 - \theta_2 \rangle$ is almost surely constant, and this is true for all distributions in the family because all distributions in an exponential family have the same support.

Now the PDMF of the distribution having parameter θ_1 with respect to the distribution having parameter θ_2 is

$$e^{l(\theta_1) - l(\theta_2)} = e^{\langle Y, \theta_1 - \theta_2 \rangle - c(\theta_1) + c(\theta_2)}$$

and we have just established that this is almost surely constant. In order for the PDMF to sum-integrate to one, the constant must be one. Hence θ_1 and θ_2 correspond to the same distribution. □

Mean Value Parameterizations (cont.)

The preceding proof almost proves the following theorem.

Theorem

In a regular full exponential family having canonical statistic Y, canonical parameter vectors θ_1 and θ_2 correspond to the same distribution if and only if $\langle Y, \theta_1 - \theta_2 \rangle$ is constant almost surely.

Proof.

If θ_1 and θ_2 correspond to the same distribution, then they have the same mean vector, and the preceding proof shows $\langle Y, \theta_1 - \theta_2 \rangle$ is constant almost surely.

Conversely, if $\langle Y, \theta_1 - \theta_2 \rangle$ is constant almost surely, then the preceding proof shows θ_1 and θ_2 correspond to the same distribution. □

Mean Value Parameterizations (cont.)

A vector η in the parameter space such that $\langle Y, \eta \rangle$ is constant almost surely, is called a **direction of constancy**.

The set of all directions of constancy is called the **constancy space**. Clearly, it is a vector subspace of the vector space containing the canonical parameter space.

If θ is an element of the full canonical parameter space, and η is a direction of constancy, then $\theta + \eta$ is in the full canonical parameter space, and θ and $\theta + \eta$ correspond to the same distribution.

The reason for the name is that the log likelihood is constant in that direction, that is, $l(\theta + t\eta)$ is a constant function of t.

Thus the canonical parameterization of a regular full exponential family is not necessarily identifiable.

Mean Value Parameterizations (cont.)

But all unconditional canonical parameterizations of aster models currently implemented are identifiable so long as the aster model has no multinomial dependence groups.

If the aster model has a multinomial dependence group G, then the indicator vector $\eta \in \mathbb{R}^J$ defined by

$$\eta_j = \begin{cases} 1, & j \in G \\ -1, & j = q(G) \\ 0, & \text{otherwise} \end{cases}$$

is a direction of constancy because

$$\langle Y, \eta \rangle = \begin{cases} -Y_{q(G)} + \sum_{j \in G} Y_j, & q(G) \in J \\ \sum_{j \in G} Y_j, & q(G) \notin J \end{cases}$$

and in either case is constant almost surely by definition of the multinomial distribution and the predecessor is sample size property.

Mean Value Parameterizations (cont.)

Because the aster transform is invertible, the same also applies to conditional canonical parameterizations.

All conditional canonical parameterizations of aster models currently implemented are identifiable so long as the aster model has no multinomial dependence groups.

Mean Value Parameterizations (cont.) I

Theorem

For a regular full exponential family, if the canonical parameterization is identifiable, then the mapping between the canonical parameter and the mean value parameter is invertible. Moreover, it is a C^∞ diffeomorphism. The matrix inverse of the first derivative is the first derivative matrix of the inverse mapping.

Proof

If the canonical parameterization is identifiable, then the mapping $\theta \longleftrightarrow \mu$ is one-to-one, hence invertible.

In the theorem about identifiability of canonical parameters, we learned that $\nabla^2 c(\theta)$ has a trivial null space if and only if the canonical parameterization is identifiable, hence it is invertible if and only if the canonical parameterization is identifiable, hence by the inverse function theorem (of real analysis) the mapping

Mean Value Parameterizations (cont.) II

$\theta \longleftrightarrow \mu$ locally invertible and differentiable both ways. Higher order derivatives of the inverse mapping are given by higher order derivatives of the cumulant function, the chain rule, and the rule for the derivative of the inverse of a matrix. □

In symbols, the last sentence of the theorem statement and the inverse function theorem say

$$\mu = h(\theta) = \nabla c(\theta)$$

implies

$$\nabla h(\theta) = \nabla^2 c(\theta)$$

and

$$\nabla h^{-1}(\mu) = \left[\nabla^2 c(\theta)\right]^{-1}$$

Mean Value Parameterizations (cont.)

The formula for the derivative of matrix inversion mentioned in the proof is easily derived from $A A^{-1} = I$, where I denotes the identity matrix. Differentiating gives

$$\frac{\partial A}{\partial t} A^{-1} + A \frac{\partial A^{-1}}{\partial t} = 0$$

and multiplying by A^{-1} on the left gives

$$\frac{\partial A^{-1}}{\partial t} = - A^{-1} \frac{\partial A}{\partial t} A^{-1}$$

Mean Value Parameterizations (cont.)

A similar analysis applied to the conditional exponential families associated with dependence groups gives the following.

The map h_G defined by

$$h_G(\theta_G) = \nabla c_G(\theta_G)$$

maps the canonical parameter vector θ_G of a regular full exponential family associated with dependence group G to $\xi_G = h_G(\theta_G)$.

The **conditional mean value parameter vector** is the vector ξ having subvectors ξ_G.

Mean Value Parameterizations (cont.)

So what expectations are the ξ_G?

Recall that $y_G = z_1 + \cdots + z_{y_{q(G)}}$, where the z_i are IID realizations of the canonical statistic of the exponential family with cumulant function c_G and canonical parameter θ_G (y_G is a random sum of random variables with $y_{q(G)}$ terms).

Thus

$$E(y_G | y_{q(G)}) = \sum_{i=1}^{y_{q(G)}} E(Z_i) = y_{q(G)} \xi_G$$

(because $E(Z_i) = \xi_G$ for all i). And

$$E(y_G | y_{q(G)} = 1) = \xi_G \tag{$*$}$$

assuming this makes sense. Equation $(*)$ does not make sense when the event $y_{q(G)} = 1$ has probability zero.

Mean Value Parameterizations (cont.)

When equation $(*)$ does not make sense, we cannot use it as a definition of ξ_G.

Then we have to use the circumlocution: ξ_G is the mean of each of the $y_{q(G)}$ IID random variables the sum of which is y_G. (This is the general definition that works in all cases.)

Mean Value Parameterizations (cont.)

Because the mean of a random vector is just the vector whose components are the means of the components of the random vector, we don't need to refer to dependence groups in the definitions just given.

For all $j \in J$

$$\xi_j = E(y_j | y_{p(j)} = 1)$$

when the conditioning event has nonzero probability.

Otherwise we use the circumlocution: ξ_j is the mean of each of the $y_{p(j)}$ IID random variables the sum of which is y_j.

This is where the node-to-node predecessor function p comes in handy even when the aster model has dependence groups.

A Confession

The first aster paper (Geyer, Wagenius, and Shaw, *Biometrika*, 2007) did not define conditional mean value parameters this way. They said

$$\xi_j = E(y_j | y_{p(j)}) = y_{p(j)} E(y_j | y_{p(j)} = 1)$$

rather than

$$\xi_j = E(y_j | y_{p(j)} = 1)$$

A referee said the former definition is dumb. It is a function of random variables $y_{p(j)}$ and parameters $E(y_j | y_{p(j)} = 1)$ and so shouldn't be called a parameter and shouldn't be denoted by a Greek letter. The R package `aster` uses the same dumb definition by default.

We didn't listen then. But now we agree with the referee. The R package `aster2` and recent papers and technical reports use the latter (non-dumb) definition (if they mention conditional mean value parameters at all).

A Confession (cont.)

Since version 1.0.2 of R package `aster`, the `aster` and `aster.formula` methods of R generic function `predict` have a new optional argument `is.always.parameter = FALSE` that controls which definition of ξ is used.

When this argument is `TRUE` it uses the new good definition

$$\xi_j = E(y_j | y_{p(j)} = 1)$$

When this argument is `FALSE` it uses the old dumb definition

$$\xi_j = E(y_j | y_{p(j)})$$

The default is `FALSE` for backwards compatibility. We do not want to break old code. Almost always you will want to add `is.always.parameter = TRUE` if you want conditional mean value parameters.

Mean Value Parameterizations (cont.)

It is useful to examine the direct change of parameter

$$\mu \longleftrightarrow \xi$$

rather than the long way round

$$\mu \longleftrightarrow \varphi \longleftrightarrow \theta \longleftrightarrow \xi$$

Applying the iterated expectation theorem to

$$E(y_j | y_{p(j)}) = y_{p(j)} \xi_j$$

gives

$$\mu_j = E(y_j) = E\{E(y_j | y_{p(j)})\} = E(y_{p(j)} \xi_j) = \xi_j E(y_{p(j)}) = \xi_j \mu_{p(j)}$$

Mean Value Parameterizations (cont.)

And iterating this gives

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

and so forth.

Keep going until the only μ is for an initial node, in which case, since the expectation of a constant is a constant,

$$\mu_{p(p(p(p(j))))} = y_{p(p(p(p(j))))}$$

(or perhaps with more p's, whatever it takes to get to an initial node).

Mean Value Parameterizations (cont.)

Here is a way to write μ in terms of ξ without saying "or perhaps with more p's, whatever it takes".

Let $\prec$ denote the **transitive closure of the node-to-node predecessor relation** defined by $j \prec k$ if and only if one of the following holds

$$\begin{aligned} j &= p(k) \\ j &= p(p(k)) \\ j &= p(p(p(k))) \\ &\vdots \end{aligned}$$

where the dots indicate arbitrarily many applications of p.

If we allowed ourselves to use the term "ancestor" like it is used in graph theory, this would be the "ancestor relation". But we avoid biological terminology for describing graphs and so have to use the more long-winded term in boldface above.

Mean Value Parameterizations (cont.)

But $\prec$, which is a strict partial order relation, is not as useful as $\preceq$, its corresponding partial order relation, defined by $j \preceq k$ if and only if $j \prec k$ or $j = k$.

$\preceq$ has the even more long-winded name: **reflexive transitive closure of the node-to-node predecessor relation**.

But it would have a clumsy name even if we used “ancestor”' like it is used in graph theory. What would you call it? Ancestor-or-self relation? Reflexive closure of the ancestor relation?

Whatever one calls them, we now have the two useful symbols $\prec$ and $\preceq$ for these relations.

Mean Value Parameterizations (cont.)

Using this new notation

$$\mu_k = \left(\prod_{\substack{j \in J \\ j \preceq k}} \xi_j \right) \left(\prod_{\substack{j \in N \setminus J \\ j \preceq k}} y_j \right)$$

(the second product always has exactly one term).

Mean Value Parameterizations (cont.)

Going the other way is even easier

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

assuming we do not have divide by zero. Since we already know that the mapping $\mu \longleftrightarrow \xi$ is invertible, it must be that we never have divide by zero.

A Plethora of Parameterizations

Now we have four different parameterizations. All are equally good, and any one can be mapped to any other.

$$\begin{tikzcd}
\theta \arrow[rrr, shift right, "\text{aster transform}"'] \arrow[d, shift right, "\nabla c_G"'] & & & \varphi \arrow[lll, shift right, "\text{inverse aster transform}"'] \arrow[d, shift left, "\nabla c"] \\
\xi \arrow[rrr, shift left, "\text{multiplication}"] \arrow[u, shift right, red] & & & \mu \arrow[lll, shift left, "\text{division}"] \arrow[u, shift left, red]
\end{tikzcd}$$

A Plethora of Parameterizations (cont.)

In the figure, the black arrows all have closed form expressions and all are infinitely differentiable.

The red arrows have no closed form expression and do not even exist if the canonical parameterizations are not identifiable (either both are identifiable or neither is).

In aster models currently implemented canonical parameterizations are identifiable unless there are multinomial dependence groups. Then the red arrows indicate infinitely differentiable functions.

We saw in the proof of identifiability of mean value parameters how to find φ given μ. Maximize the function l_μ defined by

$$l_\mu(\varphi) = \langle \mu, \varphi \rangle - c(\varphi)$$

A Plethora of Parameterizations (cont.)

Similarly, to find θ_G given ξ_G, maximize the function l_{ξ_G} defined by

$$l_{\xi_G}(\theta_G) = \langle \xi_G, \theta_G \rangle - c_G(\theta_G)$$

When there are multinomial dependence groups and the full canonical parameterizations are not identifiable, the canonical parameterizations can be made identifiable by imposing equality constraints on the canonical parameters. More on this later.

Some Distribution Theory

Let us do a little distribution theory to have some concrete examples.

Some Distribution Theory: Bernoulli

The PMF of the Bernoulli distribution is

$$f_p(x) = \begin{cases} 1 - p, & x = 0 \\ p, & x = 1 \end{cases}$$

where p is the "usual parameter" satisfying $0 < p < 1$. We can write this without case splitting

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

so the log likelihood is

$$\begin{aligned} l(p) &= x \log(p) + (1 - x) \log(1 - p) \\ &= x \bigl[\log(p) - \log(1 - p) \bigr] + \log(1 - p) \end{aligned}$$

Some Distribution Theory: Bernoulli (cont.)

From this we see that the usual statistic x is the canonical statistic. But the usual parameter is not the canonical parameter. The canonical parameter must be the term in square brackets

$$\theta = \log(p) - \log(1 - p) = \log\left(\frac{p}{1 - p}\right) = \operatorname{logit}(p)$$

We can solve for the usual parameter in terms of the canonical parameter

$$\begin{gathered} e^\theta = p/(1 - p) \\ (1 - p)e^\theta = p \\ e^\theta = p + pe^\theta \\ e^\theta = p + pe^\theta \\ p = e^\theta/(1 + e^\theta) \end{gathered}$$

Some Distribution Theory: Bernoulli (cont.)

Recall the log likelihood

$$l(p) = x \operatorname{logit}(p) + \log(1 - p)$$

and the change of parameter

$$p = \frac{e^\theta}{1 + e^\theta}$$

The term that does not contain x must be minus the cumulant function, that is,

$$c(\theta) = -\log(1 - p) = -\log\left(1 - \frac{e^\theta}{1 + e^\theta}\right) = -\log\left(\frac{1}{1 + e^\theta}\right)$$

or

$$c(\theta) = \log\left(1 + e^\theta\right)$$

Some Distribution Theory: Bernoulli (cont.)

And

$$
\begin{aligned}
c(\theta) &= \log\left(1 + e^{\theta}\right) \\
c'(\theta) &= \frac{e^{\theta}}{1 + e^{\theta}}
\end{aligned}
$$

Thus we see that the "usual" parameter p is also the mean value parameter ξ, so we will use that notation from now on.

And

$$
\begin{aligned}
c'(\theta) &= \frac{1}{e^{-\theta} + 1} \\
c''(\theta) &= \frac{e^{-\theta}}{[e^{-\theta} + 1]^2} = \frac{e^{\theta}}{[1 + e^{\theta}]^2} = \xi(1 - \xi)
\end{aligned}
$$

Some Distribution Theory: Bernoulli (cont.)

Thus we recover the usual theory of the Bernoulli distribution

$$\begin{aligned} E(X) &= \xi \\ \operatorname{var}(X) &= \xi(1 - \xi) \end{aligned}$$

But we obtain a lot more, everything we need to know to use Bernoulli arrows in aster models.

Some Distribution Theory: Poisson

The PMF of the Poisson distribution is

$$f_m(x) = \frac{m^x e^{-m}}{x!}$$

where m is the “usual parameter” satisfying $0 < m < \infty$. So the log likelihood is

$$l(m) = x \log(m) - m$$

(we drop the term $\log(x!)$ that does not contain the parameter).

Some Distribution Theory: Poisson (cont.)

From this we see that the usual statistic x is the canonical statistic. But the usual parameter is not the canonical parameter. The canonical parameter is what multiplies x in the log likelihood, that is,

$$\theta = \log(m)$$

which has inverse change of parameter

$$m = e^{\theta}$$

The term in the log likelihood that does not contain x must be minus the cumulant function, that is,

$$c(\theta) = m = e^{\theta}$$

Some Distribution Theory: Poisson (cont.)

And

$$
\begin{aligned}
c(\theta) &= e^{\theta} \\
c'(\theta) &= e^{\theta} \\
c''(\theta) &= e^{\theta}
\end{aligned}
$$

Thus we see that the "usual" parameter m is also the mean value parameter ξ, so we will use that notation from now on.

And we recover the usual theory of the Poisson distribution

$$
\begin{aligned}
E(X) &= \xi \\
\operatorname{var}(X) &= \xi
\end{aligned}
$$

But we obtain a lot more, everything we need to know to use Poisson arrows in aster models.

Some Distribution Theory: Zero-Truncated Poisson

The PMF of the zero-truncated Poisson distribution is

$$f_m(x) = \frac{m^x e^{-m}}{x!(1 - e^{-m})}$$

where m is the "usual parameter" satisfying $0 < m < \infty$. So the log likelihood is

$$l(m) = x \log(m) - m - \log(1 - e^{-m})$$

(we drop the term $\log(x!)$ that does not contain the parameter).

Some Distribution Theory: Zero-Truncated Poisson (cont.)

From this we see that the usual statistic x is the canonical statistic. But the usual parameter is not the canonical parameter. The canonical parameter is what multiplies x in the log likelihood, that is,

$$\theta = \log(m)$$

which has inverse change of parameter

$$m = e^{\theta}$$

The term in the log likelihood that does not contain x must be minus the cumulant function, that is,

$$c(\theta) = m + \log(1 - e^{-m}) = e^{\theta} + \log\left(1 - e^{-e^{\theta}}\right)$$

Some Distribution Theory: Zero-Truncated Poisson (cont.)

And

$$
\begin{aligned}
c(\theta) &= e^\theta + \log\left(1 - e^{-e^\theta}\right) \\
c'(\theta) &= e^\theta + \frac{e^\theta e^{-e^\theta}}{1 - e^{-e^\theta}} \\
&= m + \frac{m e^{-m}}{1 - e^{-m}} \\
&= \frac{m}{1 - e^{-m}}
\end{aligned}
$$

Thus we see that the “usual” parameter m is not the mean value parameter ξ either. In fact, m is the mean of the (untruncated) Poisson random variable that we truncate to get X.

Although for the Bernoulli and Poisson distributions, there was a closed form expression for the mapping $\xi \longrightarrow \theta$, for this distribution there is not.

Some Distribution Theory: Zero-Truncated Poisson (cont.)

And

$$
\begin{aligned}
c'(\theta) &= e^{\theta} + \frac{e^{\theta} e^{-e^{\theta}}}{1 - e^{-e^{\theta}}} \\
c''(\theta) &= e^{\theta} + \frac{e^{\theta} e^{-e^{\theta}}}{1 - e^{-e^{\theta}}} - \frac{e^{2\theta} e^{-e^{\theta}}}{1 - e^{-e^{\theta}}} - \frac{e^{2\theta} e^{-2e^{\theta}}}{(1 - e^{-e^{\theta}})^2} \\
&= e^{\theta} - \frac{e^{\theta}(e^{\theta} - 1) e^{-e^{\theta}}}{1 - e^{-e^{\theta}}} - \left[\frac{e^{\theta} e^{-e^{\theta}}}{1 - e^{-e^{\theta}}} \right]^2 \\
&= m - \frac{m(m-1) e^{-m}}{1 - e^{-m}} - \left[\frac{m e^{-m}}{1 - e^{-m}} \right]^2
\end{aligned}
$$

Some Distribution Theory: Zero-Truncated Poisson (cont.)

Thus we discover the theory of the zero-truncated Poisson distribution

$$
\begin{aligned}
E(X) &= \frac{m}{1 - e^{-m}} = \xi \\
\operatorname{var}(X) &= m - \frac{m(m - 1)e^{-m}}{1 - e^{-m}} - \left[\frac{m e^{-m}}{1 - e^{-m}} \right]^2 \\
&= \xi(1 - \xi e^{-m}) \\
&= \xi(1 + m - \xi)
\end{aligned}
$$

And we obtain a lot more, everything we need to know to use zero-truncated Poisson arrows in aster models.

A Plethora of Parameterizations (cont.)

But don't we need to know a lot more distribution theory than that?

No. We just need to teach the computer a bit about the basics of differentiation: the rules for derivative of a sum, derivative of a product, derivative of a quotient, and the chain rule.

Then the computer can combine cumulant functions for one-parameter conditional distributions to obtain the cumulant function for the whole aster model, the log likelihood, and the gradient vector and hessian matrix of the log likelihood. These are needed to do maximum likelihood estimation and likelihood-based inference, which uses the **Fisher information matrix** and the **delta method** (much more on these later).

The computer can also do all of the changes of parameter between θ, φ, ξ, and μ and all the derivatives (Jacobian matrices) for these changes of parameter, which are needed for the **delta method**.

Unconditional Canonical Affine Submodels

It may come as a shock, that all of this theory and all of these parameterizations do not give us any useful models. Too many parameters!

We call the models already presented **saturated aster models**. They have one parameter per arrow in the graph, which is one parameter per non-initial node of the graph, which is one parameter per component of the response vector.

Useful models have to be submodels of these models .

Unconditional Canonical Affine Submodels (cont.)

We already know how to specify submodels, just like in linear models (LM) and generalized linear models (GLM), we specify the saturated model parameters as linear functions of other parameters.

As we learn from GLM theory, we do not want to specify means as linear functions because linear functions do not respect constraints. If we are doing Bernoulli GLM, then we know $0 < \mu_i < 1$, but writing a linear function

$$\mu_i = \alpha + \beta x_i$$

gives means outside the allowed range. Logistic regression specifies the saturated model canonical parameter vector as a linear function

$$\theta_i = \operatorname{logit}(\mu_i) = \alpha + \beta x_i$$

And since the range of θ_i is $-\infty$ to $+\infty$, this works.

Unconditional Canonical Affine Submodels (cont.)

In order to get all **canonical affine submodels** at once, we adopt matrix notation

$$\varphi = a + M\beta$$

where

- φ is the saturated model unconditional canonical parameter,
- a is a known vector (not a function of unknown parameters) called the **offset vector**.
- M is a known matrix (not a function of unknown parameters, usually a function of covariate data) called the **model matrix**.
- β is an unknown parameter vector.

Unconditional Canonical Affine Submodels (cont.)

"Offset vector" and "model matrix" is the terminology of the R function `glm`.

The `aster` package says "origin" rather than "offset vector" (which it probably shouldn't).

Other people say "design matrix" rather than "model matrix" but this doesn't really make sense when some of the covariates are random rather than fixed by experimental design.

Unconditional Canonical Affine Submodels (cont.)

The offset vector is zero in most applications. This gives us **canonical linear submodels** specified by

$$\varphi = M\beta$$

This is what we have seen over and over again in books on LM and GLM.

R package `aster` puts an offset vector in every model by default (it probably shouldn't, and the `aster2` package does not, more bad design).

However, as long as `varb` is in the model, the offset vector only affects the betas for `varb`, and these are of no scientific interest. So it doesn't really matter (but is confusing).

Unconditional Canonical Affine Submodels (cont.)

Nevertheless, offset vectors are occasionally useful. How many knew about and have used the `offset` optional argument of the R function `glm`?

So we keep them.

Unconditional Canonical Affine Submodels (cont.)

When we plug $\varphi = a + M\beta$ into the aster model log likelihood we get

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

for the submodel log likelihood. We may drop the additive term that does not contain the parameter vector β obtaining

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

and now we revert to matrix notation to see

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

so

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

Unconditional Canonical Affine Submodels (cont.)

And we see that

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

has the form of an exponential family log likelihood with

- canonical statistic vector $M^T y$
- canonical parameter vector β
- cumulant function

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

This is important: unconditional canonical affine submodels are themselves regular full exponential families.

Unconditional Canonical Affine Submodels (cont.)

$$
\begin{aligned}
c_{\text{sub}}(\beta) &= c(a + M\beta) \\
\nabla c_{\text{sub}}(\beta) &= M^T \nabla c(a + M\beta) \\
\nabla^2 c_{\text{sub}}(\beta) &= M^T \nabla^2 c(a + M\beta) M
\end{aligned}
$$

Unconditional Canonical Affine Submodels (cont.)

To see these, use coordinates. The i-th component of $M\beta$ is

$$\sum_k m_{ik} \beta_k$$

so

$$\frac{\partial c_{\text{sub}}(\beta)}{\partial \beta_k} = \sum_i \frac{\partial c(\varphi)}{\partial \varphi_i} \frac{\partial \varphi_i}{\partial \beta_k} = \sum_i \frac{\partial c(\varphi)}{\partial \varphi_i} m_{ik}$$

and

$$\frac{\partial^2 c_{\text{sub}}(\beta)}{\partial \beta_k \partial \beta_l} = \sum_i \sum_j \frac{\partial^2 c(\varphi)}{\partial \varphi_i \partial \varphi_j} m_{ik} m_{jk}$$

Unconditional Canonical Affine Submodels (cont.)

This gives us everything we need for maximum likelihood estimation and likelihood inference for canonical affine submodels.

Because these submodels are regular full exponential families, maximum likelihood estimates (MLE), if they exist, can be found by any algorithm that goes uphill on the log likelihood and doesn't stop until it finds a point where the gradient vector is zero.

If the canonical affine submodel is identifiable, then MLE are unique if they exist.

Unconditional Canonical Affine Submodels (cont.)

By our theorem about identifiability of the canonical parameterization of a regular full exponential family, as applied to the canonical affine submodel, the submodel is identifiable if and only if the only vector η in the submodel parameter space such that

$$\langle M^T Y, \eta \rangle = \langle Y, M \eta \rangle \tag{$*$}$$

is almost surely constant is $\eta = 0$. The set of all vectors η such that $(*)$ is almost surely constant is the submodel constancy space.

If η is in the submodel constancy space, then $M\eta$ is in the saturated model constancy space and vice versa.

Unconditional Canonical Affine Submodels (cont.)

Identifiability can fail in two ways, one deterministic, one stochastic.

- If $M\eta = 0$ for some $\eta \neq 0$, then we have classic **collinearity**. R function `aster` solves this the same way R functions `lm` and `glm` solve this: drop one or more columns of the model matrix while maintaining the same column space (hence the same submodel).
- If $\langle Y, M\eta \rangle$ is almost surely constant for some $\eta \neq 0$, then we have a **direction of constancy**.
  R function `transformUnconditional` in R package `aster2` solves this by dropping one or more columns of the model matrix while maintaining the same intersection of the model matrix column space and the saturated model constancy space.

Unconditional Canonical Affine Submodels (cont.)

We can always do this: obtain identifiability while keeping the same model by dropping some columns of the model matrix.

That is what users are used to from R functions `lm` and `glm` so that is what we do in R packages `aster` and `aster2`.

Which columns get dropped is arbitrary. This is another aspect of canonical parameters are meaningless.

Because we have the same model after dropping columns of M, mean value parameters stay the same. This is another aspect of probabilities and expectations are meaningful.

Unconditional Canonical Affine Submodels (cont.)

By the theory of exponential families, the **submodel mean value parameter** is

$$\tau = \nabla c_{\text{sub}}(\beta) = E(M^T y) = M^T E(y) = M^T \mu$$

A Plethora of Parameterizations (cont.)

Now we have six parameterizations:

- saturated model conditional canonical parameter vector θ,
- saturated model unconditional canonical parameter vector φ,
- saturated model conditional mean value parameter vector ξ,
- saturated model unconditional mean value parameter vector μ,
- unconditional canonical affine submodel canonical parameter vector β,
- unconditional canonical affine submodel mean value parameter vector τ,

All six parameterizations are important.

All six parameterizations play roles in scientific inference (not all on stage at the same time).

A Plethora of Parameterizations (cont.)

$$\begin{tikzcd}
\theta \arrow[r, shift right=1.5, "\text{aster transform}"'] \arrow[d, shift right=1.5, "\nabla c_G"'] & \varphi \arrow[l, shift right=1.5, "\text{inverse aster transform}"'] \arrow[r, red, shift right=1.5] \arrow[d, shift left=1.5, "\nabla c"] & \beta \arrow[l, shift right=1.5, "\beta \mapsto a + M\beta"'] \arrow[d, shift left=1.5, "\nabla c_{\text{sub}}"] \\
\xi \arrow[u, red, shift right=1.5] \arrow[r, shift left=1.5, "\text{multiplication}"] & \mu \arrow[u, red, shift left=1.5] \arrow[l, shift left=1.5, "\text{division}"] \arrow[r, shift left=1.5, "\mu \mapsto M^T \mu"] & \tau \arrow[l, red, shift left=1.5] \arrow[u, red, shift left=1.5]
\end{tikzcd}$$

A Plethora of Parameterizations (cont.)

As before, the black arrows all have closed form expressions and all are infinitely differentiable.

As before, the vertical red arrows have no closed form expression and do not even exist if the canonical parameterizations are not identifiable.

Like the other vertical red arrows, the one $\tau \longrightarrow \beta$ is computed by maximization of

$$l_{\text{sub},\tau}(\beta) = \langle \tau, \beta \rangle - c_{\text{sub}}(\beta)$$

The horizontal red arrow $\mu \longleftarrow \tau$ can only be computed as the composition $\tau \longrightarrow \beta \longrightarrow \varphi \longrightarrow \mu$.

The horizontal red arrow $\varphi \longleftarrow \beta$ only makes sense when φ has the form $\varphi = M\beta$ for some β, in which case we already know which β corresponds to which φ.

A Plethora of Parameterizations (cont.)

Let I be the index set for β and τ. As before, J is the index set of θ, φ, ξ, and μ.

Let n be the cardinality of J, the length of θ, φ, ξ, μ, and y. Let r be the cardinality of I, the length of β and τ.

- The possible values of β form an open subset of $\mathbb{R}^I$, the full canonical parameter space of the submodel.
- Only φ such that $\varphi = a + M\beta$ for some β occur, so the possible values of φ form an r-dimensional affine subspace of $\mathbb{R}^J$ or a relatively open subset thereof.
- Since the inverse aster transform and derivatives of cumulant functions are nonlinear, the possible values of θ, ξ, and μ form r-dimensional curved submanifolds of $\mathbb{R}^J$.
- The possible values of τ form an open subset of $\mathbb{R}^I$.

A Plethora of Parameterizations (cont.)

In GLM because components of the response vector are independent (conditional on covariates), there is no distinction between conditional and unconditional so we have $\varphi = \theta$ and $\mu = \xi$ and thus only four parameterizations.

In LM because mean value parameters are canonical for normal location models, we have $\theta = \varphi = \mu = \xi$ and thus only three parameterizations

$$\begin{aligned} \mu &= M\beta \\ \tau &= M^T \mu \end{aligned}$$

A Plethora of Parameterizations (cont.)

That we still have multiple parameterizations for LM and GLM (though not so many as aster) is hidden by the usual way textbooks and teachers woof about them.

Policy in all statistics courses (not policy enforced by anybody, just part of the culture) says that we only call β a parameter vector.

The parameter vector μ we do not mention at all. Its estimates are denoted $\hat{y}$ in LM rather than $\hat{\mu}$ and are called "predicted values" even though they are "predicting" the expectation of data already observed rather than any future data. And $\hat{\tau} = M^T \hat{y}$ are not mentioned at all or computed by any R function (although you can of course compute this matrix multiplication yourself).

A Plethora of Parameterizations (cont.)

I guess (who can really say where bits of culture come from) that this policy is an attempt to not confuse students with multiple parameterizations. The betas are the parameters; that's all you need to know.

But then what is

$$\hat{y}_i \pm t \text{ critical value} \times \text{standard error of } \hat{y}_i$$

It is a confidence interval, but for what? A confidence interval is an interval estimate *of a parameter!* What parameter? The parameter who must not be named!

IMHO this causes as much confusion as it avoids.

A Plethora of Parameterizations (cont.)

GLM teachers and textbooks again say β is the only parameter vector. They call φ the “linear predictor”, a term not used in general statistical theory. And μ and $\hat{\mu}$ are not called anything.

But there is a function to compute them in R. If `gout` is the result of a call to the `glm` function, then $\hat{\varphi}$ is computed by

```
phi.hat <- predict(gout)
```

and $\hat{\mu}$ is computed by

```
mu.hat <- predict(gout, type = response)
```

A Plethora of Parameterizations (cont.)

If the `glm` function was called with optional argument `x = TRUE` so its result (`gout`) has a component `gout$x` which is the model matrix, then

```
tau.hat <- t(gout$x) %*% mu.hat
```

computes the submodel canonical statistic $\hat{\tau}$.

A Plethora of Parameterizations (cont.)

Whether or not you think these parameterizations must not be named, they exist and are important for scientific inference.

IMHO the names and the symbols help. It's hard to talk about something that must not be named.

And $\hat{y}$ is (again, just IMHO) silly. Nowhere else in statistics to we put a hat on a symbol for a statistic to symbolize a parameter estimate. That is confusing all by itself.

Invariance of Maximum Likelihood

Maximum likelihood estimates **transform by invariance**.

Suppose θ is a parameter vector (not necessarily having anything to do with aster models or even exponential families) and $\psi = h(\theta)$ is an invertible transformation $\theta = h^{-1}(\psi)$.

Theorem. If $\hat{\theta}$ is the MLE for θ, then $\hat{\psi} = h(\hat{\theta})$ is the MLE for ψ.

Proof. Think geometrically. The graph of the log likelihood is a hypersurface over the domain. The maximum occurs at one point (let us assume). θ and ψ are different coordinatizations of the domain. The point where the maximum occurs is called $\hat{\theta}$ in one coordinatization and $\hat{\psi}$ in the other. The relationship between the coordinatizations is $\psi = h(\theta)$. QED

Invariance of Maximum Likelihood (cont.)

So if we know the MLE for any parameter, we know the MLE for every parameter (transform by invariance).

Observed Equals Expected

The log likelihood for an exponential family (not necessarily an aster model) is

$$l(\varphi) = \langle y, \varphi \rangle - c(\varphi)$$

and the gradient vector is

$$\nabla l(\varphi) = y - \nabla c(\varphi)$$

Assuming the distribution of the canonical statistic y is not concentrated on a hyperplane so the MLE is unique if it exists, the unique MLE is determined by

$$y = \nabla c(\hat{\varphi})$$

but

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

is the change of parameter from canonical to mean value.

Observed Equals Expected (cont.)

So the relation between the MLE for φ and μ is

$$\hat{\mu} = h(\hat{\varphi}) \qquad \text{and} \qquad \hat{\varphi} = h^{-1}(\hat{\mu})$$

and

$$y = \hat{\mu}$$

This is called the **observed equals expected** property of maximum likelihood in a regular full exponential family: the observed value of the canonical statistic y is equal to the MLE of its expected value $\hat{\mu}$.

This is true for *any* regular full exponential family. It is a large part of the traditional woof about log-linear models for categorical data analysis. It is entirely absent from the traditional woof about GLM. There is no reason for this absence (other than tradition).

Observed Equals Expected (cont.)

When we apply the observed equals expected property to aster canonical affine submodels, we get

$$M^T y = \hat{\tau}$$

We cannot use this directly to find MLE of other parameters because we have no closed form expression for the transformation $\beta = h^{-1}(\tau)$ that gives

$$\hat{\beta} = h^{-1}(M^T y)$$

We have to find $\hat{\beta}$ using optimization software to maximize the log likelihood $l(\beta)$, and then use the transformations

$$\begin{aligned} \hat{\varphi} &= a + M\hat{\beta} \\ \hat{\mu} &= \nabla c(\hat{\varphi}) \\ \hat{\tau} &= M^T \hat{\mu} \end{aligned}$$

Observed Equals Expected (cont.)

Although

$$M^T y = \hat{\tau}$$

does not allow us to determine the MLE for any other parameterization except by doing maximum likelihood to find $\hat{\beta}$, it is extremely important because it is the only simple algebraic fact about maximum likelihood: **maximum likelihood in a regular full exponential family has the observed equals expected property**. This is an important part of interpretation of MLE.

Sufficient Dimension Reduction

Almost all statistical inference does **dimension reduction**. It replaces the whole of the data (dimension n) with a smaller vector of statistics (dimension r).

For example, when you reduce a vector of n numbers to its mean, $r = 1$. When you reduce it to its mean and variance, $r = 2$.

When you reduce the data to the MLE $\hat{\beta}$ for a statistical model, r is its dimension.

Sufficient Dimension Reduction (cont.)

Fisher (1922), the paper that introduced many of the ideas of mathematical statistics (statistical models, the idea that inference estimates parameters, maximum likelihood, Fisher information, asymptotics of maximum likelihood, efficiency, and sufficiency), asked and answered the question: how much information does a dimension reduction throw away?

A dimension reduction is **sufficient** if it throws away no information about the parameters. That is the ideal situation.

Sufficient Dimension Reduction (cont.)

A **statistic** (singular) is a random variable or random vector that is a function of the data and is not a function of the parameters of the statistical model. (This means it can actually be calculated even though the values of the parameters are unknown.)

A statistic is **sufficient** if the conditional distribution of the whole data given this statistic does not depend on the parameters of the statistical model.

Sufficient Dimension Reduction (cont.)

When we factorize the distribution of the data into marginal times conditional we get

$$\begin{aligned} & f_\theta(\text{whole data}) \\ & \qquad = f(\text{whole data} \mid \text{sufficient statistic}) f_\theta(\text{sufficient statistic}) \end{aligned}$$

and we can drop the multiplicative term that does not contain the parameter from the likelihood

$$L(\theta) = f_\theta(\text{sufficient statistic})$$

and log likelihood

$$l(\theta) = \log f_\theta(\text{sufficient statistic})$$

Sufficient Dimension Reduction (cont.)

Thus MLE depend on the whole data only through the sufficient statistic.

There is a converse to this. The Neyman-Fisher factorization criterion (which we do not prove) says that if the likelihood or log likelihood **depends on the whole data only through some statistic**, then **that statistic is sufficient**.

In particular, **the canonical statistic vector for an exponential family is always sufficient**.

Sufficient Dimension Reduction (cont.)

Some people (like my thesis adviser) always say **canonical sufficient statistic** rather than **canonical statistic** even though this is redundant (because the canonical statistic is always sufficient).

Just a reminder. Don't want anyone to forget how important sufficiency is.

Sufficient Dimension Reduction (cont.)

Any one-to-one function of a sufficient statistic is sufficient.

For an unconditional canonical affine submodel of an aster model, if $\tau = h(\beta)$ is the mapping from submodel canonical parameter to submodel mean value parameter, then

$$\hat{\beta} = h^{-1}(M^T y)$$

is a one-to-one function of the submodel canonical sufficient statistic vector $M^T y$, hence $\hat{\beta}$ is sufficient.

Since every other parameter is a one-to-one function of β, the MLE for all other parameters $\hat{\theta}$, $\hat{\varphi}$, $\hat{\xi}$, and $\hat{\mu}$ are also sufficient statistic vectors.

Sufficient Dimension Reduction (cont.)

In short, maximum likelihood for an unconditional aster model does sufficient dimension reduction.

(We haven't yet talked about so-called conditional aster models. They do not do sufficient dimension reduction.)

Maximum Entropy

Edwin Jaynes introduced the "maximum entropy formalism" that describes exponential families in terms of entropy.

Entropy comes from physics, in particular, from thermodynamics and statistical physics.

Negative entropy (also called negentropy) is also called Shannon information in information theory and Kullback-Leibler information in statistics.

Maximum Entropy (cont.)

The **second law of thermodynamics** says entropy increases in any isolated physical process.

A physical system that has maximum entropy is at thermodynamic equilibrium.

A glass of water with ice cubes in it is not at thermodynamic equilibrium. As the ice melts and the surrounding water becomes colder, entropy increases. After the ice melts and we have a glass of water at uniform temperature throughout, we are at thermodynamic equilibrium and at maximum entropy.

Maximum Entropy (cont.)

Ludwig Boltzmann and Josiah Willard Gibbs figured out the connection between entropy and probability and between the thermodynamic properties of bulk matter and the motions and interactions of atoms and molecules.

In this theory entropy is not certain to increase to its maximum possible value. It is only overwhelmingly probable to do so in any large system.

In a very small system, such as a cubic micrometer of air, it is less probable that entropy will be near its maximum value. In such a small system the statistical fluctuations are large.

This is the physical manifestation of the law of large numbers. The larger the sample size (the more molecules involved) the less stochastic variation.

Kullback-Leibler Divergence I

For any probability distributions P having density f with respect to another probability distribution Q, the **Kullback-Leibler divergence** between them is

$$D(P, Q) = -E_Q\{\log f(X)\}$$

Theorem

Kullback-Leibler divergence $D(P, Q)$ is strictly positive unless $P = Q$, in which case $D(P, Q) = 0$.

Proof

From $f(x + h) > f(x) + f'(x)h$ unless $h = 0$ for any strictly convex function f (this is part (b') of Theorem 2.13 in Rockafellar and Wets *Variational Analysis*), we get

$$-\log(x) > 1 - x, \qquad x \neq 1.$$

Kullback-Leibler Divergence II

It follows that

$$\begin{aligned} D(P, Q) &= -E_Q\{\log f(X)\} \\ &> E_Q\{1 - f(X)\} \\ &= 0 \end{aligned}$$

unless $f(x) = 1$ for all x, in which case $P = Q$. □

Maximum Entropy (cont.)

The entropy of a probability measure P having density f with respect to another probability measure Q is

$$H(P) = -E_Q\{f(X) \log f(X)\}$$

Theorem

The distribution that maximizes $H(P)$ subject to the constraint

$$E_P(Y) = \mu \tag{1}$$

is the distribution in the exponential family generated by Q and Y having mean value parameter μ, assuming the given μ is a possible value of the mean value parameter of this exponential family.

Maximum Entropy (cont.) I

Proof

Choose θ to be the canonical parameter corresponding to mean value parameter μ in the exponential family referred to in the theorem statement. And let f_θ denote the density with respect to Q of that distribution

$$f_\theta(x) = e^{\langle Y(x), \theta \rangle - c(\theta)}$$

where we give Q parameter value 0 in the exponential family, where c is the cumulant function for the family, and where we

Maximum Entropy (cont.) II

assume $c(0) = 0$. Then

$$\begin{aligned}
H(P) &= -E_Q\{f(X) \log f(X)\} \\
&= -E_Q\left\{f(X) \log\left(\frac{f(X)}{f_\theta(X)}\right)\right\} - E_Q\{f(X) \log f_\theta(X)\} \\
&= -E_Q\left\{f(X) \log\left(\frac{f(X)}{f_\theta(X)}\right)\right\} - E_Q\{f_\theta(X) \log f_\theta(X)\} \\
&= -E_Q\left\{f(X) \log\left(\frac{f(X)}{f_\theta(X)}\right)\right\} + H(P_\theta)
\end{aligned}$$

where P_θ is the distribution having density f_θ with respect to Q and where the third equality is the fact that $\log f_\theta(X) = \langle Y(X), \theta\rangle - c(\theta)$ has the same expectation with respect to P and P_θ, by the assumption that $E(Y) = \mu$ for both.

Maximum Entropy (cont.) III

Now we notice that

$$\begin{aligned}
-E_Q\left\{f(X)\log\left(\frac{f(X)}{f_\theta(X)}\right)\right\} &= -E_P\left\{\log\left(\frac{f(X)}{f_\theta(X)}\right)\right\} \\
&= E_P\left\{\log\left(\frac{f_\theta(X)}{f(X)}\right)\right\} \\
&= -D(P_\theta, P)
\end{aligned}$$

so

$$H(P) = -D(P_\theta, P) + H(P_\theta) < H(P_\theta), \qquad \text{unless } P = P_\theta,$$

so P_θ is the unique distribution maximizing entropy. □

Maximum Entropy (cont.)

To the extent that our statistics models real-world physics (and chemistry and biology), it should also maximize entropy.

This is a weak spot in the argument. How well do our models model? Should imperfect models, which leave out a lot of physics and chemistry and biology, still maximize their entropy?

Nevertheless, Jaynes considered maximizing entropy.

Maximum Entropy (cont.)

In the context of aster models, we choose the base measure Q to be any distribution in the saturated aster model. We choose the submodel canonical statistic vector $M^T y$ to be the statistic we control the means of.

Then the maximum entropy model is the canonical linear model with model matrix M.

If we want an offset vector, we can get that too by modifying the base measure.

Maximum Entropy (cont.)

So now the other shoe drops on interpretation of exponential families in general and aster models in particular.

Subject to being in the saturated aster model determined by the aster graph, the maximum entropy model that constrains the vector expectation

$$\tau = E(M^T y) \tag{$*$}$$

is the canonical linear model with model matrix M.

This submodel leaves all other aspects of the distribution of the response as random as possible (in the sense of maximum entropy) given $(*)$ holds.

Maximum Entropy (cont.)

The maximum entropy argument and the sufficient dimension reduction argument work together.

An unconditional aster model (or any exponential family model) has the sufficient dimension reduction property that makes the canonical affine submodel canonical statistic vector $M^T y$ and the MLE of all the parameters **sufficient statistics**.

Subject to having that property, every other aspect of the distributions in the model is as random as possible (maximizes entropy) subject to $M^T y$ having the expectation $\tau = E(M^T y)$ that it does, which is the submodel mean value parameter.

Maximum Entropy (cont.)

If you haven't seen it before, this is a new and different way to justify statistical models

Choose the "correct" submodel sufficient statistic vector $M^T y$, where "correct" means its components include the scientifically important and interpretable quantities.

Make the model the exponential family having $M^T y$ as the submodel canonical statistic vector.

Then we get the sufficient dimension reduction and maximum entropy properties.

Multivariate Monotonicity

A function h from a convex open subset Φ of a finite-dimensional vector space to the same finite-dimensional vector space is **multivariate monotone** if

$$\langle h(\varphi^{**}) - h(\varphi^{*}), \varphi^{**} - \varphi^{*} \rangle \geq 0, \qquad \text{for all } \varphi^{*} \text{ and } \varphi^{**} \text{ in } \Phi$$

and is **strictly multivariate monotone** if

$$\langle h(\varphi^{**}) - h(\varphi^{*}), \varphi^{**} - \varphi^{*} \rangle > 0, \qquad \text{whenever } \varphi^{*} \neq \varphi^{**}$$

Multivariate Monotonicity (cont.)

Multivariate monotonicity generalizes univariate monotonicity. If the space is one dimensional so φ^*, φ^{**}, $h(\varphi^*)$, and $h(\varphi^{**})$ are scalars, we have

$$\langle h(\varphi^{**}) - h(\varphi^*), \varphi^{**} - \varphi^* \rangle = \bigl[h(\varphi^{**}) - h(\varphi^*)\bigr] \cdot \bigl[\varphi^{**} - \varphi^*\bigr] \geq 0$$

and the only way this can hold is if

$$\varphi^* < \varphi^{**} \qquad \text{implies} \qquad h(\varphi^*) \leq h(\varphi^{**})$$

that is, h is **nondecreasing**.

Similarly, strict multivariate monotonicity of h and one-dimensional implies h is **increasing**.

Multivariate Monotonicity (cont.)

Theorem

The gradient function of a convex function is multivariate monotone. The gradient function of a strictly convex function is strictly multivariate monotone.

This is part (a) of Theorem 2.14 in Rockafellar and Wets *Variational Analysis* and the part about strict convexity.

Corollary

The mapping from canonical parameter vector to mean value parameter vector for a regular full exponential family is multivariate monotone. It is strictly multivariate monotone if the canonical parameterization is identifiable.

Multivariate Monotonicity (cont.)

Multivariate monotonicity is a hard concept to wrap your mind around, especially if you never heard of it before.

Here is a dumbed-down version. Suppose we increase one component of the unconditional canonical parameter vector φ, holding all other components of φ fixed. Then the corresponding component of the unconditional mean value parameter vector μ also increases (other components of μ can go any which way).

The dumbed-down version is not equivalent. It is implied by, but does not imply, strict multivariate monotonicity.

Multivariate Monotonicity (cont.)

Multivariate monotonicity is equivalent to the following. For every nonzero vector δ and every $\varphi \in \operatorname{dom} h$, the scalar function

$$g(t) = \langle h(\varphi + t\delta), \delta \rangle$$

is nondecreasing for t in any interval I where $\varphi + t\delta \in \operatorname{dom} h$.

Proof. Take t^* and t^{**} in I, with $t^* < t^{**}$. Then

$$g(t^{**}) - g(t^*) = \langle h(\varphi + t^{**}\delta) - h(\varphi + t^*\delta), \delta \rangle \tag{$*$}$$

and

$$(\varphi + t^{**}\delta) - (\varphi + t^*\delta) = (t^{**} - t^*)\delta$$

so $(*)$ is nonnegative if and only if $(**)$ is too.

$$\langle h(\varphi + t^{**}\delta) - h(\varphi + t^*\delta), (\varphi + t^{**}\delta) - (\varphi + t^*\delta) \rangle \tag{$**$}$$

QED

Multivariate Monotonicity (cont.)

Similarly, strict multivariate monotonicity is equivalent to the following. For every nonzero vector δ and every $\varphi \in \operatorname{dom} h$, the scalar function

$$g(t) = \langle h(\varphi + t\delta), \delta \rangle$$

is increasing for t in any interval where $\varphi + t\delta \in \operatorname{dom} h$.

The dumbed-down version only considers direction vectors δ that point along coordinate axes. That is not enough for equivalence.

The first aster paper (Geyer, Wagenius, and Shaw, *Biometrika*, 2007) only presented the dumbed-down version (in the discussion). In a later paper (Shaw and Geyer, *Evolution*, 2010) we found we needed the real definition of multivariate monotonicity (in an appendix) to explain why the aster models under discussion worked.

Multivariate Monotonicity (cont.)

A more symmetric way to talk about multivariate monotonicity is the following. Let φ^* and φ^{**} be two distinct valid values of the saturated model unconditional canonical parameter vector. And let μ^* and μ^{**} be the corresponding values of the saturated model unconditional mean value parameter vector. Then

$$\langle \mu^{**} - \mu^*, \varphi^{**} - \varphi^* \rangle \geq 0$$

and this inequality is strict (> 0) if the aster model is non-degenerate.

This formulation makes it clear that the inverse of a multivariate monotone relationship is also multivariate monotone, and similarly with strictly multivariate monotone in both places.

Multivariate Monotonicity (cont.)

Since an unconditional canonical affine submodel of an aster model is itself a regular full exponential family, we have the same properties for its canonical and mean value parameters as for the saturated model.

Let β^* and β^{**} be two distinct valid values of an unconditional canonical affine model canonical parameter vector. And let τ^* and τ^{**} be the corresponding values of an unconditional canonical affine model mean value parameter vector. Then

$$\langle \tau^{**} - \tau^*, \beta^{**} - \beta^* \rangle \geq 0$$

and this inequality is strict (> 0) if the aster model is non-degenerate.

Multivariate Monotonicity (cont.)

Not only are the map $\varphi \longrightarrow \mu$ and its inverse strictly multivariate monotone, so are the map $\beta \longrightarrow \tau$ and its inverse.

Multivariate Monotonicity (cont.)

Applying what we know about monotonicity to the conditional distributions for dependence groups, we see that

$$\theta_G \mapsto \nabla c_G(\theta_G)$$

is multivariate monotone for each G.

In particular if $G = \{j\}$ is a singleton set, we have

$$\theta_j \mapsto c_j'(\theta_j)$$

is an increasing function.

If there are no dependence groups (pedantically, if every dependence group is a singleton) there is a componentwise univariate strictly monotone relationship between the saturated model conditional canonical vector θ and the saturated model conditional mean value parameter ξ.

Multivariate Monotonicity (cont.)

Let θ^* and θ^{**} be two distinct valid values of the saturated model conditional canonical parameter vector. And let ξ^* and ξ^{**} be the corresponding values of the saturated model conditional mean value parameter vector. Then

$$\langle \xi^{**} - \xi^*, \theta^{**} - \theta^* \rangle \geq 0$$

and this inequality is strict (> 0) if the aster model is non-degenerate.

If there are no dependence groups, more is true. Actually,

$$\bigl[\xi_j^{**} - \xi_j^* \bigr] \cdot \bigl[\theta_j^{**} - \theta_j^* \bigr] \geq 0, \qquad j \in J$$

Multivariate Monotonicity (cont.)

The map $\theta \longrightarrow \xi$ and its inverse are strictly multivariate monotone when the canonical parameterization is identifiable.

When there are no dependence groups, the map $\theta_j \longrightarrow \xi_j$ and its inverse are strictly univariate monotone, for each $j \in J$.

The Story So Far

We started off with our aster model assumptions (with or without dependence groups). These imply a **statistical model** with a valid factorization **joint = product of conditionals**.

The additional assumption **predecessor is sample size** yields the simple **transformation between conditional and unconditional mean value parameters** (multiplication and division).

The additional assumption **distributions for dependence groups are exponential family** yields **exponential family saturated model** and **aster transform**.

Then **unconditional canonical affine submodels** yield **exponential family submodels**.

The Story So Far (cont.)

Exponential families have many important properties.

- **Strictly concave log likelihood** assures **MLE are unique if they exist** and **well-behaved optimization**.
- **Derivatives of cumulant function give mean and variance** of canonical statistic makes statistical inference easy.
- **Observed = expected**.
- **Sufficient dimension reduction**.
- **Maximum entropy**.
- **Multivariate monotone relationship** between canonical and mean value parameters.

First two for the computer, the rest for people.

Interpretation of Aster Models

Observed Equals Expected Maximum likelihood matches the MLE of the submodel mean value parameter $\hat{\tau}$ to the observed value of the submodel canonical statistic $M^T y$. This determines MLE of all other parameters.

Sufficiency The submodel canonical statistic $M^T y$ and MLE of all parameters are sufficient statistic vectors.

Maximum Entropy Subject to having the expectations of $M^T y$ that they do and having the aster graph that they do, the distributions in the submodel are as random as possible (maximize entropy).

Multivariate Monotonicity To the extent that canonical parameters can be interpreted, their interpretation involves their multivariate monotone relationship with mean value parameters.

Interpretation of Aster Models (cont.)

When one first sees interpretation of regression-like models in intro statistics, one starts with “simple” linear regression. The data are independent (X_i, Y_i) pairs and the regression equation is

$$E(Y_i | X_i) = \alpha + \beta X_i$$

and this magically corresponds to the R formula mini-language formula `y ~ x`

One also learns to parrot that β is the slope of the regression line. Slope is rise over run, so β is the change in the (conditional) mean of the response Y corresponding to unit change in the predictor X.

Interpretation of Aster Models (cont.)

One may also learn that

> Correlation is not causation. And regression isn't either.

(because simple linear regression is just another view of correlation).

So the regression equation is only good for **prediction** for new data from the same population from which the (X_i, Y_i) pairs are a random sample. It is not good for **explanation**, and does not necessarily have anything to do with the **causal** relationship (if any) between the response and predictor.

Interpretation of Aster Models (cont.)

One may also learn that in a **designed experiment** with the levels of certain factors (call them **treatments**) controlled by the experimenters and **randomized** assignment of individuals to treatments, that one can make causal inferences about **treatment effects**.

But even in this setting any covariates that are not controlled by the experimenters are still subject to correlation is not causation.

Interpretation of Aster Models (cont.)

All of the this elementary material about model interpretation except for the interpretation of “slope” applies to any LM, GLM, or aster model (or other regression-like statistical models).

In a GLM the interpretation of regression coefficients gets more complicated. Even in the “simple” (just one predictor case) we have for logistic regression

$$\varphi_i = \alpha + \beta x_i$$

but there is a complicated nonlinear relationship between this and

$$\mu_i = E(Y_i | X_i)$$

which are canonical parameter and mean value parameter, respectively.

Interpretation of Aster Models (cont.)

$$\begin{aligned}
\frac{\partial \mu_i}{\partial \beta} &= \frac{\partial}{\partial \beta} \frac{1}{1 + e^{-\alpha - \beta x_i}} \\
&= \frac{x_i e^{-\alpha - \beta x_i}}{(1 + e^{-\alpha - \beta x_i})^2} \\
&= x_i \mu_i (1 - \mu_i)
\end{aligned}$$

And this changes as α and β change. So the simple “rise over run” interpretation does not transfer from LM to GLM.

Interpretation of Aster Models (cont.)

Some textbooks, wanting to keep the simple "rise over run" interpretation say it still holds but for

$$\varphi_i = \log\left(\frac{\mu_i}{1 - \mu_i}\right)$$

But why be interested that function of μ_i? The question cannot be answered without a lot of exponential family theory.

No matter which way you try to go, interpretation of GLM is not as simple as interpretation of LM.

Interpretation of Aster Models (cont.)

At least in GLM we have independence of components of the response vector (conditional on covariates).

This means the nonlinear relationship between canonical and mean value parameters is a componentwise univariate monotone relationship. So we only have to deal with univariate functions and univariate monotonicity.

Interpretation of Aster Models (cont.)

In aster models we have dependence of components of the response vector (conditional on covariates).

This means the nonlinear relationship between unconditional canonical and mean value (either for saturated models or for canonical affine submodels) parameters is an inherently multivariate monotone relationship.

We cannot escape or simplify multivariate monotonicity. We just have to deal with it.

Interpretation of Aster Models (cont.)

Somewhere after an intro statistics course — in a real regression or theory course — one gets introduced to multiple regression and model matrices.

There may be more than one predictor vector and the mean value parameter vector (for LM) or the canonical parameter vector (for GLM and aster) may be a function of any or all of the predictor vectors.

Furthermore, even if one is only given one predictor to start with say x, then one can make up other predictors, for example, x^2, x^3, $\ldots$ (polynomial regression) or $\sin(x)$, $\cos(x)$, $\sin(2x)$, $\cos(2x)$, $\ldots$ (trigonometric series regression, a. k. a., Fourier series regression).

There is always a potentially infinite number of predictor vectors, no matter how few were “given”.

Interpretation of Aster Models (cont.)

Nevertheless, one is still trained to write out the regression equation

$$\mu_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

where the x_{ij} are elements of the j-th predictor vector. These can be "given" or "made up". For example,

$$\mu_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k$$

(polynomial regression) or

$$\mu_i = \alpha + \beta_1 \sin(x_i) + \beta_2 \cos(x_i) + \cdots + \beta_{2k-1} \sin(k x_i) + \beta_{2k} \cos(k x_i)$$

(trigonometric series regression).

Interpretation of Aster Models (cont.)

Then one learns that there is no good reason to treat the “intercept” α specially. It is just a regression coefficient like the rest. The predictor vector it goes with is the constant predictor vector having all components equal to one, for example,

$$
\begin{aligned}
\mu_i &= \beta_1 \cdot 1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \cdots + \beta_{k+1} x_{ik} \\
\mu_i &= \beta_1 \cdot 1 + \beta_2 x_i + \beta_3 x_i^2 + \cdots + \beta_{k+1} x_i^k \\
\mu_i &= \beta_1 \cdot 1 + \beta_2 \sin(x_i) + \beta_3 \cos(x_i) + \cdots \\
&\qquad + \beta_{2k} \sin(k x_i) + \beta_{2k+1} \cos(k x_i)
\end{aligned}
$$

Interpretation of Aster Models (cont.)

Then one learns that the preceding slide still treated the intercept (now called β_1 specially). Just write

$$\mu_i = \sum_{j=1}^{p} x_{ij} \beta_j \tag{$*$}$$

so now we are writing x_{i1} instead of 1 and have bumped the indices of the other predictor vectors to correspond to their regression coefficients.

And we recognize $(*)$ as the matrix equation

$$\mu = M\beta$$

where M, the **model matrix**, is the matrix with components x_{ij}.

Interpretation of Aster Models (cont.)

The triumph of this matrix notation in LM theory is that we can write an explicit formula for the MLE

$$\hat{\beta} = (M^T M)^{-1} M^T y \tag{$*$}$$

Note that this goes together with what we know about parameterizations for LM; $\mu = M\beta$ and $\tau = M^T \mu$, so $\tau = M^T M \beta$. By the observed equals expected property, we have $\hat{\tau} = M^T y$. And by the invertibilty of the mapping $\beta \longrightarrow \tau$, we have

$$\hat{\beta} = (M^T M)^{-1} \hat{\tau}$$

which is the same as $(*)$.

Interpretation of Aster Models (cont.)

In GLM and aster model theory, we no longer have a closed-form expression for MLE as a function of data. All we can do is run optimization software to find out the value of $\hat{\beta}$ corresponding to each value of $M^T y$.

Also there is a difference between the unconditional mean value parameter vector μ and the unconditional canonical parameter vector φ and it is the latter that is linearly

$$\varphi = M\beta$$

or affinely

$$\varphi = a + M\beta$$

related to the regression coefficient parameter vector β.

Interpretation of Aster Models (cont.)

Still, both teachers and students are tempted by the carryover from LM theory to make regression equations like

$$\begin{aligned}
\varphi_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} \\
\varphi_i &= \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \cdots + \beta_p x_i^{p-1} \\
\varphi_i &= \beta_1 \cdot 1 + \beta_2 \sin(x_i) + \beta_3 \cos(x_i) + \cdots \\
&\qquad + \beta_{p-1} \sin\left(\tfrac{p-1}{2} x_i\right) + \beta_p \cos\left(\tfrac{p-1}{2} x_i\right)
\end{aligned}$$

and use them as the basis of one's “interpretation” of the model.

Interpretation of Aster Models (cont.)

I am here to tell you this is (IMHO) all wrong.

Remember that canonical parameters are meaningless quantities, and if there's no meaning in them, that saves a world of trouble as we needn't try to find any.

Consider the two linear transformations

$$\begin{aligned} \beta &\mapsto M\beta \\ \mu &\mapsto M^T\mu \end{aligned}$$

Since M determines M^T and vice versa, if you understand one of these transformations, then you also "understand" the other, but you "understand" it implicitly without clearly seeing it.

Interpretation of Aster Models (cont.)

Staring at $\varphi = M\beta$ written out with explicit sum and indices

$$\varphi_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_r x_{ir} \qquad (*)$$

doesn't tell you much about $\tau = M^T \mu$ written out with explicit sum and indices

$$\tau_i = \mu_1 x_{1i} + \mu_2 x_{2i} + \mu_3 x_{3i} + \cdots + \mu_n x_{ni} \qquad (**)$$

These sums do not have the same number of terms: r is the submodel dimension and n is the saturated model dimension. Moreover, $(*)$ contains x_{ij} in the i-th row of M and $(*)$ contains x_{ij} in the j-th column of M, the former covariate values pertaining to one node of the graph, the latter pertaining to one regression coefficient.

Interpretation of Aster Models (cont.)

The mapping

$$\varphi = M\beta$$

relates unconditional canonical parameter vectors (submodel to saturated model).

The mapping

$$\tau = M^T\mu$$

relates unconditional mean value parameter vectors (saturated model to submodel).

Remember which kind of parameters is meaningless and which kind is meaningful?

Interpretation of Aster Models (cont.)

The mapping

$$\varphi = M\beta$$

doesn't become meaningful without the very messy, highly nonlinear (but multivariate monotone) mapping

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

Interpretation of Aster Models (cont.)

The mapping

$$\tau = M^T \mu$$

is directly related to the observed equals expected property

$$\hat{\tau} = M^T y \tag{$*$}$$

Also $(*)$ is the sufficient dimension reduction from whole data y to sufficient statistic vector $\hat{\tau}$ (since all MLE are one-to-one functions of each other, all other MLE are one-to-one functions of $\hat{\tau}$, hence themselves sufficient statistic vectors).

Interpretation of Aster Models (cont.)

Thus (IMHO) the mapping $\mu \mapsto M^T \mu$ (which can also be written $y \mapsto M^T y$) is more important than the mapping $\beta \mapsto M\beta$ and deserves to be woofed about at least as much if not more when one is "interpreting" aster models (or GLM or LM).

The first submission of the first aster paper (Geyer, Wagenius, and Shaw, *Biometrika*, 2007) made an attempt in this direction only discussing models in terms of $y \mapsto M^T y$ and not at all in terms of $\beta \mapsto M\beta$. But the referees didn't get it, and we were forced to interpret both ways in the published version.

Interpretation of Aster Models (cont.)

This wasn't really our fault or the referees' fault. It's embedded in the culture.

The R generic function `summary` prints out the components of $\hat{\beta}$ and a lot of information about them.

No function prints out the submodel canonical sufficient statistic vector $\hat{\tau}$ or any information about it. At least, no generic function with a `glm` method will do this job. The `aster` and `aster.formula` methods of the generic function `predict` will do this job, as we shall presently see, but not in a user-friendly fashion.

Interpretation of Aster Models (cont.)

This wasn't really R's fault either. SAS or SPSS or Stata or whatever is no better. Nor are thousands of intro stats and regression and linear models textbooks any better.

Example One Revisited

```
                  Estimate Std. Error z value  Pr(>|z|)
(Intercept)     -1.0506435  0.1843320 -5.6997 1.200e-08
varbfl03        -0.3490958  0.2679185 -1.3030   0.19258
varbfl04        -0.3442222  0.2438992 -1.4113   0.15815
varbhdct02       1.3214136  0.2611741  5.0595 4.203e-07
varbhdct03       1.3433740  0.2146250  6.2592 3.870e-10
varbhdct04       1.8513276  0.1998528  9.2635 < 2.2e-16
varbld02        -0.0293022  0.3157033 -0.0928   0.92605
varbld03         1.7400507  0.3961890  4.3920 1.123e-05
varbld04         4.1885771  0.3342661 12.5307 < 2.2e-16
layerfl:nsloc    0.0701024  0.0146520  4.7845 1.714e-06
layerhdct:nsloc -0.0058043  0.0055499 -1.0458   0.29564
layerld:nsloc    0.0071652  0.0058667  1.2213   0.22196
layerfl:ewloc    0.0179769  0.0144128  1.2473   0.21229
layerhdct:ewloc  0.0076060  0.0055608  1.3678   0.17138
layerld:ewloc   -0.0047874  0.0059191 -0.8088   0.41863
fit:popAA        0.1292377  0.0891292  1.4500   0.14706
fit:popEriley   -0.0495612  0.0712789 -0.6953   0.48686
fit:popLf       -0.0332786  0.0795727 -0.4182   0.67579
fit:popNWLF      0.0210283  0.0635998  0.3306   0.74092
fit:popNessman  -0.1862690  0.1277869 -1.4577   0.14494
fit:popSPP       0.1491795  0.0677156  2.2030   0.02759
```

Example One Revisited (cont.)

So back to example one. It is actually easier to figure out the components of the unconditional canonical affine submodel canonical sufficient statistic vector $M^T y$ from looking at the names of the regression coefficients than from looking at the formula

```
> aout$formula

resp ~ varb + layer:(nsloc + ewloc) + fit:pop
```

For one thing, there is one component of the submodel canonical sufficient statistic vector for each regression coefficient. But there is no such correspondence with terms in the formula. There is some correspondence, but it is not one-to-one.

Let's go through the regression coefficient names one by one.

Example One Revisited (cont.)

A component of $M^T y$ has the form $x^T y$ where x is a column of M (a predictor vector, either “given” or “made-up”). So we need to figure out what the columns of the model matrix are.

Simplest first, the predictor vector named `"(Intercept)"`. All its components are equal to one, so the corresponding submodel canonical sufficient statistic is

$$x^T y = \sum_{i=1}^{n} y_i$$

This may not seem to make much sense, because the components of y are different kinds of variables, so this is like adding apples and oranges, but it will presently.

Example One Revisited (cont.)

Next come the predictor vectors with `"varb"` in the name: `varbfl03`, `varbfl04`, `varbhdct02`, `varbhdct03`, `varbhdct04`, `varbld02`, `varbld03`, and `varbld04`.

Recall that the variable `varb` in the data frame `redata` is a factor

```
> class(redata$varb)

[1] "factor"

> levels(redata$varb)

[1] "fl02"   "fl03"   "fl04"   "hdct02" "hdct03"
[6] "hdct04" "ld02"   "ld03"   "ld04"
```

Example One Revisited (cont.)

Recall that factors (categorical variables) get turned into dummy variables, which are zero-or-one-valued, zero indicating not in a particular category and one indicating in that category.

That gives nine dummy variables (for the nine levels of `varb` corresponding to the nine nodes of the aster graph). But these nine dummy variables add up to the `"(Intercept)"` dummy variable. So if we kept them all, we would not have a full rank model. R drops the first one in alphabetical order, which would have been named `varbfl02` if it hadn't been dropped.

Example One Revisited (cont.)

For a zero-or-one-valued predictor variable x the corresponding submodel canonical sufficient statistic is

$$x^T y = \sum_{i=1}^{n} x_i y_i = \sum_{\substack{i \in \{1, \ldots, n\} \\ x_i = 1}} y_i$$

Each of these submodel canonical sufficient statistics is a sum of the components of the response vector corresponding to a particular node of the aster graph.

Example One Revisited (cont.)

Thus we have one submodel canonical sufficient statistic for each node of the graph, except for the one (`"fl02"`) that R dropped.

But if we know the sum for all nodes (the `"(Intercept)"` statistic) and we know the sum for each node except `"fl02"` then we also know the sum for `"fl02"` (subtract the sums for each of the other nodes from the total).

In short, if we replaced the `"(Intercept)"` component of the submodel canonical sufficient statistic vector with the `"varbfl02"` component (what that component would have been if it hadn't been dropped) we would still have a sufficient statistic vector.

R would actually do this for us if we specified no intercept by putting `0 +` at the beginning of the formula.

Example One Revisited (cont.)

Next come the predictor vectors with `"nsloc"` or `"ewloc"` in the name: `layerfl:nsloc`, `layerhdct:nsloc`, `layerld:nsloc`, `layerfl:ewloc`, `layerhdct:ewloc`, and `layerld:ewloc`.

Recall that the variable `layer` is a factor and the variables `nsloc` and `ewloc` are quantitative

```
> sapply(redata, class)

        pop       ewloc       nsloc        varb
   "factor"   "integer"   "integer"    "factor"
       resp          id        root       layer
  "integer"   "integer"   "numeric" "character"
        fit
  "numeric"

> levels(redata$layer)

NULL
```

Example One Revisited (cont.)

The factor gets turned into three dummy variables (one for each of its levels). Nothing gets done to the quantitative variables.

Then the “interaction” operator (`:`) says take each of the former and multiply it componentwise by each of the latter making $3 \times 2 = 6$ new predictor variables. The colon in the regression coefficient name shows the corresponding predictor variable arose this way and also shows what variables were multiplied to make it.

Example One Revisited (cont.)

Now we have predictor vectors having components

$$x_i = d_i z_i$$

where d_i is the corresponding component of a dummy (zero-or-one-valued) variable (named `layerfl`, `layerhdct`, or `layerld`) and z_i is the corresponding component of a quantitative variable (`nsloc` or `ewloc`).

The corresponding submodel canonical sufficient statistic is

$$x^T y = \sum_{i=1}^{n} x_i y_i = \sum_{\substack{i \in \{1, \ldots, n\} \\ d_i = 1}} y_i z_i$$

Example One Revisited (cont.)

In short, this set of components of the sufficient statistic vector is sums of products of components of the response vector and corresponding components of a quantitative variable (`nsloc` or `ewloc`), the sums running over each "layer" of the graph (either the three `"ld"` nodes or the three `"fl"` nodes, or the three `"hdct"` nodes).

Why would we want something like that? Does that have a clear scientific interpretation?

Example One Revisited (cont.)

Again recall that any one-to-one function of a sufficient statistic vector is another sufficient statistic vector.

This means we can combine these sufficient statistics with others we already know about to make new sufficient statistics.

Example One Revisited (cont.)

Here we know

$$\sum_{\substack{i \in \{1, \ldots, n\} \\ d_i = 1}} y_i z_i \qquad \text{and} \qquad \sum_{\substack{i \in \{1, \ldots, n\} \\ d_i = 1}} y_i$$

are functions of the submodel canonical statistic, the former we just calculated and the latter is a sum of components with names containing `varb`, for example the sum over the `"ld"` layer is the sum of the sums over the `"ld02"`, `"ld03"`, and `"ld04"` nodes.

We also "know"

$$\sum_{\substack{i \in \{1, \ldots, n\} \\ d_i = 1}} z_i$$

because z (either `nsloc` or `ewloc`) is not considered random (it is a predictor, not the response).

Example One Revisited (cont.)

Any sums like these can be considered as n times expectations with respect to the conditional distribution

$$\widehat{E}(YZ|\texttt{layer}) = \frac{1}{n} \sum_{\substack{i \in \{1,\ldots,n\} \\ d_i = 1}} y_i z_i$$

$$\widehat{E}(Y|\texttt{layer}) = \frac{1}{n} \sum_{\substack{i \in \{1,\ldots,n\} \\ d_i = 1}} y_i$$

$$\widehat{E}(Z|\texttt{layer}) = \frac{1}{n} \sum_{\substack{i \in \{1,\ldots,n\} \\ d_i = 1}} z_i$$

where d_i are the components of the dummy variable for one of the levels of the factor `layer`.

Example One Revisited (cont.)

For any random variables Y, Z, and L in any probability model (not necessarily having anything to do with aster or even regression) the identity

$$\operatorname{cov}(Y, Z | L) = E(YZ | L) - E(Y | L) E(Z | L)$$

holds. And this holds, in particular, for empirical distributions

$$\widehat{\operatorname{cov}}(Y, Z | L) = \widehat{E}(YZ | L) - \widehat{E}(Y | L) \widehat{E}(Z | L)$$

holds.

Example One Revisited (cont.)

And this means components of $M^T y$ having the form

$$n \cdot \widehat{E}(YZ | \texttt{layer})$$

can be replaced by

$$n \cdot \widehat{\text{cov}}(Y, Z | \texttt{layer})$$

and we get another sufficient statistic vector.

The latter seem to have more obvious scientific significance.

Example One Revisited (cont.)

Finally come the predictor vectors with `"fit"` in the name: `fit:popAA`, `fit:popEriley`, `fit:popLf`, `fit:popNessman`, `fit:popNWLF`, and `fit:popSPP`.

The variable `fit` is numeric and zero-or-one-valued and the variable pop is a factor.

```
> class(redata$pop)

[1] "factor"

> class(redata$fit)

[1] "numeric"

> unique(redata$fit)

[1] 0 1
```

Example One Revisited (cont.)

So pop, being categorical, gets turned into 7 dummy variables one for each level of the factor

```
> levels(redata$pop)

[1] "AA"        "Eriley"  "Lf"        "NWLF"      "Nessman"
[6] "SPP"       "Stevens"
```

Then each of these dummy variables are multiplied componentwise by `fit` because that is what the “interaction” (`:`) operator indicates.

Example One Revisited (cont.)

We seem to have lost one. That makes 7 dummy variable times `fit` combinations, but we only got six. Where did the other one go?

```
> aout$dropped
```

```
[1] "fit:popStevens"
```

It was dropped because, if it hadn't been, then the model matrix wouldn't have been full rank. Why is that?

Example One Revisited (cont.)

Recall the definition of `fit`. It indicates the "layer" of nodes of the graph having `hdct` in their names.

```
> identical(redata$fit == 1, grepl("hdct", redata$varb))

[1] TRUE
```

If we kept `fit:popStevens`, then all of these components of the submodel canonical sufficient statistic would add up to `fit` (because every individual is in exactly one ancestral population). And `fit` is the sum of the dummy variables for `varbhdct02`, `varbhdct03`, and `varbhdct04`. So that is the collinearity that `fit:popStevens` was dropped to avoid.

Example One Revisited (cont.)

In short, the last set of components of the sufficient statistic vector is sums of components of the response vector for each ancestral population over the "fitness layer" of the graph (nodes with `hdct` in their names).

Example One Revisited (cont.)

That was exhausting. Does interpretation of aster models have to be that hard?

But notice that it was only hard because (1) it is unfamiliar (have you done anything like this before?) and (2) there is no computer support, nothing like the R function `summary` that prints out a lot of stuff you think you understand (even though we argue it is really "meaningless").

And it was only hard because we (being unfamiliar with the ideas) had to go through everything in gory detail.

The summary is not that complicated.

Example One Revisited (cont.)

The components of the unconditional canonical affine submodel canonical sufficient statistic are

- sums of response over each node of the graph,
- sums of response-location crossproducts over each layer of the graph, and
- sums of response over the fitness layer of the graph for each population.

These are what the observed equals expected property matches (observed values to MLE expected values).

The last group of sufficient statistics are scientifically crucial. They are observed fitness for each population.

Example One Revisited (cont.)

So what maximum likelihood is really doing in this model is what the preceding slide described: making MLE expected values of components of the submodel canonical sufficient statistic equal to their observed values.

And the maximum entropy property says every other aspect of the maximum likelihood model is as random as possible (maximizes entropy) subject to the constraints that the components of the submodel canonical sufficient statistic have the MLE expectations that they do and subject to the model having the structure described by the aster graphical model.

Notice this description of what maximum likelihood is really doing does not even mention the regression coefficients (betas).

Example One Revisited (cont.)

This is why we claim that understanding an aster model means understanding the submodel canonical sufficient vector $M^T y$.

If its components determine all scientifically important quantities, then the model has straightforward scientific interpretation. Otherwise it doesn't.

Interpretation Revisited

Did you notice that the word “interaction” only appeared in our interpretation in scare quotes as a name for the colon (`:`) operator?

Do you now see why the word “interaction” is not really helpful in interpreting aster models?

You may think that is because we are using the R formula mini-language in tricky ways, not as it was intended to be used. But it was never designed to be used with aster models or any models with dependence among components of the response vector. So we have to be “tricky” if we are going to use formulas at all.

A Technical Quibble

We have been saying *the* canonical statistic, *the* canonical parameter, and *the* cumulant function, but this is technically incorrect.

Suppose we have a general full exponential family (not necessarily an aster model) with log likelihood

$$l(\varphi) = \langle y, \varphi \rangle - c(\varphi)$$

and we do a one-to-one change of statistic

$$y = a + Mz$$

where a is a known vector and M is a known matrix (not an offset vector and model matrix, despite using the same letters — those names are reserved for submodel changes of parameter).

A Technical Quibble (cont.)

Then

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

and we see we again have the exponential family form with

- canonical statistic vector z,
- canonical parameter vector $M^T \varphi$, and
- cumulant function

$$c_{\text{new}}(\varphi) = c(\varphi) - \langle a, \varphi \rangle$$

A Technical Quibble (cont.)

Or suppose we do a one-to-one change of parameter

$$\varphi = a + M\beta$$

where a is a known vector and M is a known matrix (still not an offset vector and model matrix, despite using the same letters — those names are reserved for submodel changes of parameter — and this isn't a submodel because the mapping is one-to-one, and M is full rank).

Then

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

and we can drop the term $\langle y, a \rangle$ that does not contain the parameter.

A Technical Quibble (cont.)

Then

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

and we see we again have the exponential family form with

- canonical statistic vector $M^T y$,
- canonical parameter vector β, and
- cumulant function

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

(this is the same as we had for canonical affine submodels of aster models).

A Technical Quibble (cont.)

Finally, as we saw when we were discussing affine change of canonical statistic, the new cumulant function in a change can be the old cumulant function plus an arbitrary real-valued affine function.

A Technical Quibble (cont.)

Summary.

- Any one-to-one affine function of a canonical statistic is another canonical statistic.
- Any one-to-one affine function of a canonical parameter is another canonical parameter.
- An arbitrary real-valued affine function can be added to a cumulant function.

These changes are not independent, changing one requires changes in the others as shown in previous slides.

Meaningless Quantities Revisited

In aster models, we have little interest in changing the saturated model canonical statistic vector. We want its components to be the components of the response vector for the nodes of the aster graphical model.

But we do change parameters in going from saturated models to submodels.

And there is no one right offset vector and model matrix that determine a submodel. Let V denote the affine subspace of the saturated model canonical parameter space that corresponds to the submodel

$$V = \{ a + M\beta : \beta \in \mathbb{R}^p \}$$

If the saturated model unconditional canonical parameter space Φ is a full vector space, then $V \cap \Phi$ is the set of submodel values of φ.

Meaningless Quantities Revisited (cont.)

If the offset vector a and the model matrix M change but the set $V \cap \Phi$ does not. Then the submodel does not change.

Nor do the sets of allowed values of μ and ξ because these are defined by unconditional and conditional expectations of the saturated model canonical sufficient statistic, which has not been changed.

In short, the canonical "meaningless" parameters can change while the mean value "meaningful" parameters do not.

And the statistical model (the family of probability distributions) has not changed either.

Meaningless Quantities Revisited (cont.)

In practice, you get arbitrariness of the model matrix when you decide (or R decides) which dummy variables to drop to obtain full rank.

Does this arbitrariness matter? No! It is still the same statistical model, and it still has the same sets of saturated model mean value parameters.

In practice, you get arbitrariness of the model matrix when you decide (or R decides) how to parameterize polynomial functions of predictors (this comes up in the aster model competitor to Lande-Arnold analysis).

Does this arbitrariness matter? No! It is still the same statistical model, and it still has the same sets of saturated model mean value parameters.

Meaningless Quantities Revisited (cont.)

There may be other kinds of arbitrariness that arise in practice but I can't think of right now.

Would that arbitrariness matter? No! It would still be the same statistical model, and it would still have the same sets of saturated model mean value parameters.

A Technical Quibble (cont.)

In practice we don't quibble about arbitrariness of canonical statistics, canonical parameters, and cumulant functions. We keep to the definitions of the saturated model parameters (all four parameterizations) presented above.

And we recognize the arbitrariness of model matrices but don't fuss about it. Any choice of model matrix that results in the desired model is o. k. It doesn't matter to us that some other model matrix would do the same job.

We just have to be aware of the issue in case someone asks, why not some other model matrix?