Stat 8931 (Aster Models)
Lecture Slides Deck 2

Charles J. Geyer

School of Statistics
University of Minnesota

June 7, 2015

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".

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).

And similarly for other distributions ("the normal distribution" instead of *the family of normal distributions*, and so forth).

Statistical Models (cont.)

When you have a statistical model, 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.

Statistical Models (cont.)

In "master's level" theoretical statistics (5101–5102 or 8101–8102 in our department) specifying a statistical model is simple. If the data are discrete, then you just write down the probability mass function (PMF) for the data. This is a function

$$f_\theta(y)$$

of the data y. It also depends on the parameter vector θ, but we do not say θ is an argument of the PMF (parameters are not arguments).

And this is the joint PMF (y is all the data).

And in case the data are continuous, everything is the same except that $f_\theta(y)$ is the probability density function (PDF) for the data.

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 (and in fact all published examples AFAIK), this is not necessary.

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. But we can still write down a function $f_\theta(y)$ that has an obvious interpretation as a **probability mass-density function** (PMDF). (When calculating probabilities or expectations, you sum over the discrete components of y and integrate over the continuous components of y. As we shall see, we do not need to calculate expectations this way for aster models. So this is purely a theoretical quibble.)

Statistical Models (cont.)

But there is one more issue that makes the previous slide wrong (oversimplified to the point of being wrong).

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 $y_1 > 0$.

Statistical Models (cont.)

So to be technically correct, what two slides back should have said is that we integrate over the continuous components of the response, *except when their predecessors are zero, in which case they are discrete and we sum, but since in that case they are degenerate, the sum has only one term*.

As we shall see, all this pedantic quibble never gets in the way, because the theory "just works" and we do not have to fuss this way after we verify that it does the right thing in such cases.

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 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

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

I claim this is a valid factorization, with what purports to be conditional distributions actually being conditional distributions, but to see that we need to work through some issues.

First issue. In aster models, variables at initial nodes are treated as constants, so this is the joint distribution of the variables at non-initial nodes. The vector y on the left-hand side has components y_j for $j \in J$.

Aster Model PMDF (cont.)

Second issue. Every aster graph has at least one terminal node.

Proof. Start at any node. If it is terminal, we are done. Otherwise, follow any outgoing arrow (there is one by definition if the node is nonterminal). If the node this arrow goes to is terminal, we are done. Otherwise, follow any outgoing arrow from it. Since the graph is finite, eventually we either get to a nonterminal node or we get to a node previously visited, but the latter possibility is forbidden by the assumption that the graph is acyclic. QED

Aster Model PMDF (cont.)

Let k be any node index, and let $G = \{k, p(k), p(p(k)), \ldots\}$, where this notation indicates a finite set despite the "...". (We just don't know how many times we can apply the predecessor function before we get to an initial node.) $G \cap J$ is the set of non-initial nodes in G, and $G \setminus J$ is the set of initial nodes in G. The latter is a singleton set (there is exactly one initial node in G).

For any subset A of nodes of the graph, let y_A denote the "subvector" whose components are y_j for $j \in A$.

We wish to calculate the joint distribution of $y_{G \cap J}$ given $y_{G \setminus J}$. To do that, we first have to calculate the marginal distribution of $y_{G \cap J}$ by summing-integrating out (sum for discrete, integrate for continuous) all of the variables not in G.

Aster Model PMDF (cont.)

Third issue. If G is not the whole node set N of the graph, then there is a terminal node not in G.

Proof. Start at any node in $N \setminus G$. If it is terminal, we are done. Otherwise, follow any outgoing arrow. This must take us to a node j not in G because (proof by contradiction) if $j \in G$, then $p(j) \in G$ and we must have started in G contrary to assumption. Now j is terminal, or we can repeat the process. As in the "second issue" proof, we eventually get to a terminal node, and each step takes us to a node not in G by the same argument as above. QED

Aster Model PMDF (cont.)

Start summing-integrating out the variables in $J \setminus G$ by choosing a terminal node $t \in J \setminus G$ (if there is one).

Since t is terminal, the only term in

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

that contains y_t is the term $f_{t,\theta}(y_t | y_{p(t)})$, and since this is a valid conditional probability distribution, it sums or integrates (as the case may be) to one. Thus this term just disappears from the product and leaves us with the joint distribution for a new aster model that has the same graph as before except the node t and the arrow $p(t) \longrightarrow t$ have been deleted.

Note that if y_t is (partly) continuous, then we integrate if $y_{p(t)} > 0$ and sum if $y_{p(t)} = 0$, but the argument works in either case.

Aster Model PMDF (cont.)

By the "third issue" proof, we can keep repeating the process on the previous slide until there are no nodes left that are not in G and G is the whole remaining portion of the graph. Thus we have computed

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

Aster Model PMDF (cont.)

Now we wish to compute the conditional distribution of y_k given $y_{G \setminus \{k\}}$. This is conditional = joint/marginal, where "joint" is the distribution on the preceding slide, and "marginal" is the same with y_k summed-integrated out (summed or integrated, as the case may be).

Since k is a terminal node of the subgraph with node set G (again by the acyclicity property), when we sum-integrate out y_k we just get one and just delete the term $f_{k,\theta}(y_k | y_{p(k)})$ from the product. Then dividing this result into the product leaves just this term. Thus we have proved that, if

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

is the joint distribution of the aster model, then the conditional distribution of y_k given $y_{G \setminus \{k\}}$ is

$$f_{k,\theta}(y_k | y_{p(k)})$$

Aster Model PMDF (cont.)

In summary,

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

is a valid factorization, in that what purport to be conditional distributions on the right-hand side actually are conditional distributions.

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

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

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

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

Nodes have 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.

Acyclicity The graph is acyclic: a path that follows arrows in the direction they point never returns to a node.

Predecessor is Sample Size If j is non-initial, then $y_j = z_1 + \cdots + z_{p(j)}$ (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 and canonical parameter θ_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

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

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 predecessor is also zero)?

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

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

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

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_J. We just don't (yet) know the canonical parameter vector and cumulant function.

Aster Log Likelihood (cont.)

Let φ_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

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

Aster Log Likelihood (cont.)

Thus an aster model is (jointly) an exponential family with canonical statistic vector y_J, canonical parameter vector φ_J 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

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

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, again by the acyclicity property (our "second issue" theorem says there are terminal nodes, and after we remove them from the graph, we have a new graph that again has terminal nodes. And so forth.)

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.

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. We say the set Φ of φ such that the expectation exists 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 $t = 0$

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

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

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

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**

$$\kappa_i = \left. \frac{\partial \log M_\varphi(t)}{\partial t_i} \right|_{t=0}$$
$$\kappa_{ij} = \left. \frac{\partial^2 \log M_\varphi(t)}{\partial t_i \partial t_j} \right|_{t=0}$$
$$\kappa_{ijk} = \left. \frac{\partial^2 \log M_\varphi(t)}{\partial t_i \partial t_j \partial t_k} \right|_{t=0}$$

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

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

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

or, rewriting these as vector and matrix equations

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

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

Cumulant Generating Functions (cont.)

In

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

the left-hand side denotes the **mean vector**, which has components $E(Y_i)$ 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

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

the left-hand side denotes the **variance matrix**, which has components $\text{cov}(Y_i, Y_j)$ 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

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

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

In short

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

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 regular. So the question becomes are aster models regular?

The answer is yes, provided all the one-parameter exponential families for the arrows are regular. But the proof is somewhat complicated.

For each j, let Θ_j be the full canonical parameter space for the one-parameter exponential family associated with the arrow $p(j) \to j$, and define

$$\Theta = \prod_{j \in J} \Theta_j$$

the set of all valid θ values. Then Θ is an open set if all of these families are regular (meaning each Θ_j is open).

The Magic of Cumulant Functions (cont.)

Let h denote the aster transform, and define

$$\Phi = h(\Theta) = \{\, h(\theta) : \theta \in \Theta \,\}.$$

Then Φ is also an open set because the aster transform is a diffeomorphism (both it and its inverse are differentiable) and hence maps open sets to open sets.

Then the question is whether Φ is the full canonical parameter space of the joint exponential family. This seems "obvious" and it is true. There is a theorem in the theory handout proving this (and the proof is of the "follow your nose" variety), but we won't put it on these slides.

The Magic of Cumulant Functions (cont.)

That finishes the proof that the aster model is regular, provided all of the one-parameter families associated with its arrows are regular.

And all of the one-parameter families that have been implemented in aster models are regular. In fact, it is hard to find an exponential family that is not regular.

Positive Definite Matrices

A symmetric matrix V is **positive semi-definite** if

$$\langle w, Vw \rangle = w^T Vw \geq 0, \qquad \text{for all vectors } w$$

A symmetric matrix V is **positive definite** if

$$\langle w, Vw \rangle = w^T Vw > 0, \qquad \text{for all nonzero vectors } w$$

Positive Definite Matrices (cont.)

Every variance matrix $V = \mathrm{var}(Y)$ is positive semi-definite because

$$\begin{aligned}
\mathrm{var}(\langle Y, w \rangle) &= \mathrm{var}\left\{ \sum_{i=1}^{d} Y_i w_i \right\} \\
&= \mathrm{cov}\left\{ \sum_{i=1}^{d} Y_i w_i, \sum_{j=1}^{d} Y_j w_j \right\} \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} w_i w_j \, \mathrm{cov}(Y_i, Y_j) \\
&= w^T V w
\end{aligned}$$

Positive Definite Matrices (cont.)

Every variance matrix $\operatorname{var}(Y)$ is positive definite unless there exists a nonzero vector w such that $\operatorname{var}(\langle Y, w \rangle) = 0$, which happens if and only if $\langle Y, w \rangle$ is almost surely constant, which happens if and only if Y is concentrated on the hyperplane

$$\{ y : \langle y, w \rangle = c \}$$

for some constant c.

For short we will say the distribution of Y is **degenerate** if it is concentrated on a hyperplane and **non-degenerate** otherwise.

Positive Definite Matrices (cont.)

In summary, every variance matrix is positive semi-definite, and a variance matrix is positive definite if and only if the corresponding distribution is non-degenerate.

Convex and Concave Functions

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,

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

A vector-to-scalar function f that is allowed to have the value $+\infty$ (like cumulant functions) such that the set

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

is open and convex and the restriction of f to $\operatorname{dom} f$ is twice differentiable is a **convex function** if the hessian matrix is positive semi-definite everywhere on $\operatorname{dom} f$. And f is **strictly convex** if the hessian matrix is positive definite everywhere on $\operatorname{dom} f$.

A Technical Quibble

This is the most general definition of "convex set".

The most general definition of "convex function" allows the value $-\infty$ as well as $+\infty$ and does not require differentiability, much less twice differentiability. (Example: the absolute value function is convex but not differentiable at zero.)

But we will not need the general definition.

Convex and Concave Functions (cont.)

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.

The main virtue of convex functions is in minimization.

The main virtue of concave functions is in maximization.

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

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

Convex and Concave Functions (cont.)

Theorem. If f is convex and $\operatorname{dom} f$ is not the empty set, then every local minimizer of f is a global minimizer of f.

Proof. Suppose x is a local minimizer and y is any other point. If $y \notin \operatorname{dom} f$, then $f(y) = \infty$ and there is nothing to prove.

Otherwise, by assumption the entire line segment between x and y lies in $\operatorname{dom} f$ and f is twice differentiable at every point of this line segment. Define

$$g(t) = f\big(ty + (1-t)x\big)$$

so

$$g'(t) = (y-x)^T \nabla f\big(ty + (1-t)x\big)$$
$$g''(t) = (y-x)^T \nabla^2 f\big(ty + (1-t)x\big)(y-x)$$

Convex and Concave Functions (cont.)

Because x is a local minimizer $\nabla f(x) = 0$. Hence $g'(0) = 0$.

Because f is convex, $\nabla^2 f(y)$ is positive semi-definite for all $y \in \operatorname{dom} f$. Hence $g''(t)$ is nonnegative for all t.

By the fundamental theorem of calculus

$$g'(t) = g'(0) + \int_0^t g''(s)\, ds$$

Since $g'(0) = 0$ and $g''(s) \geq 0$ for all s, we have $g'(t) \geq 0$ for all t.

Convex and Concave Functions (cont.)

By another application of the fundamental theorem of calculus

$$g(1) = g(0) + \int_0^1 g'(s)\, ds$$

And since $g'(s) \geq 0$ for all s, we have

$$g(1) \geq g(0)$$

but $g(1) = f(y)$ and $g(0) = f(x)$, so this proves

$$f(y) \geq f(x)$$

QED

Convex and Concave Functions (cont.)

Theorem. If f is strictly convex and $\operatorname{dom} f$ is not the empty set, then every local minimizer of f is the unique global minimizer of f.

Proof. Follow the proof of the other theorem. Everything is the same except for the following changes.

Because f is strictly convex, $\nabla^2 f(y)$ is positive definite for all $y \in \operatorname{dom} f$. Hence $g''(t)$ is strictly positive for all t.

Since the integral of a strictly positive function is strictly positive, we conclude $g'(t) > 0$ for all $t > 0$ in our first application of the fundamental theorem of calculus, and we conclude $g(1) > g(0)$ in our second application of the fundamental theorem of calculus.

That is, we conclude $f(y) > f(x)$. So there can be no global minimizer other than x. QED

Convex and Concave Functions (cont.)

Corollary. If f is strictly convex and $\operatorname{dom} f$ is not the empty set, then every local minimizer of f is the unique zero of ∇f.

Proof. Follow the proof of the preceding theorem. In the middle we conclude that $g'(t) > 0$ for $t > 0$, and in particular

$$0 < g'(1) = (y - x)^T \nabla f(y)$$

This makes it impossible to have $\nabla f(y) = 0$. QED

Convex and Concave Functions (cont.)

The proofs of the following are obvious. (Just stand the preceding ones on their heads.)

Theorem. If f is concave and $\operatorname{dom} f$ is not the empty set, then every local maximizer of f is a global maximizer of f.

Theorem. If f is strictly concave and $\operatorname{dom} f$ is not the empty set, then every local maximizer of f is the unique global maximizer of f.

Corollary. If f is strictly concave and $\operatorname{dom} f$ is not the empty set, then every local maximizer of f is the unique zero of ∇f.

Convex and Concave Functions (cont.)

The sum of convex functions is convex, and, if one is strictly convex, then the sum is strictly convex.

The sum of concave functions is concave, and, if one is strictly concave, then the sum is strictly concave.

Every linear function is both convex and concave.

The Magic of Cumulant Functions (cont.)

Theorem. Every cumulant function of a regular full exponential family is convex. It is strictly convex if and only if the distribution of the canonical statistic vector is non-degenerate.

Proof. This follows from

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

so this is a positive semidefinite matrix and is positive definite if and only if the distribution of Y is non-degenerate, if (a fact that remains to be proved) $\operatorname{dom} c$ is a convex set.

The Magic of Cumulant Functions (cont.)

So suppose φ^* and φ^{**} and φ^{***} are in $\operatorname{dom} c$ and we calculate for $0 < t < 1$

$$
\begin{aligned}
& c\big(t\varphi^{**} + (1-t)\varphi^{***}\big) \\
& \qquad = c(\varphi^*) + \log E_{\varphi^*} \left\{ e^{\langle Y, t\varphi^{**} + (1-t)\varphi^{***} - \varphi^* \rangle} \right\} \\
& \qquad = c(\varphi^*) + \log E_{\varphi^*} \left\{ e^{t\langle Y, \varphi^{**} - \varphi^* \rangle + (1-t)\langle Y, \varphi^{***} - \varphi^* \rangle} \right\} \\
& \qquad \le c(\varphi^*) + \log E_{\varphi^*} \left\{ e^{\langle Y, \varphi^{**} - \varphi^* \rangle} + e^{\langle Y, \varphi^{***} - \varphi^* \rangle} \right\} \\
& \qquad = c(\varphi^{**}) + c(\varphi^{***}) - c(\varphi^*)
\end{aligned}
$$

the inequality being the fact that the exponential function $x \mapsto e^x$ is increasing, so the maximum value of the integrand occurs at $t = 0$ or at $t = 1$. QED

The Magic of Cumulant Functions (cont.)

An aster model has a non-degenerate joint distribution (hence strictly convex cumulant function) if every one-parameter exponential family associated with an arrow is non-degenerate and no initial node has the constant zero. (There is a proof of this in the theory handout that we won't put on these slides.)

The log likelihood of an aster model with non-degenerate distribution is strictly concave. Hence the maximum likelihood estimator (MLE) of the unconditional canonical parameter is unique if it exists. (It need not exist. Much more on this later.)

Mean Value Parameterizations

The map h defined by

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

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

In an aster model we say μ is the **unconditional mean value parameter vector** because (1) it is an unconditional expectation and (2) there are also the mean value parameters of the one-parameter exponential families associated with the conditional distributions for the arrows.

Mean Value Parameterizations (cont.)

Theorem. Assuming the aster model is non-degenerate, the mapping $\varphi \longleftrightarrow \mu$ is invertible.

Proof. Suppose μ^* is a possible value of the mean value parameter vector, that is, $\mu^* = h(\varphi^*)$ for some φ^*. Define

$$l(\varphi) = \langle \mu^*, \varphi \rangle - c(\varphi), \qquad \varphi \in \Phi,$$

(this would be a log likelihood if μ^* were a possible value y of the canonical statistic vector). Then

$$\nabla l(\varphi) = \mu^* - h(\varphi)$$

so $\nabla l(\varphi^*) = 0$. By assumption, l is a strictly concave function, hence φ^* is the unique point φ such that $h(\varphi) = \mu^*$. Thus h is one-to-one, hence invertible (considered as a function from its domain to its range). QED

Mean Value Parameterizations (cont.)

A similar analysis applied to the one-parameter exponential families associated with the arrows gives the following.

The map h_j defined by

$$h_j(\theta_j) = c_j'(\theta_j)$$

(where the prime denotes differentiation) maps the canonical parameter vector θ_j of a regular full exponential family associated with the j-th arrow to $\xi_j = h_j(\theta_j)$.

The **conditional mean value parameter vector** is the vector ξ having components ξ_j.

Mean Value Parameterizations (cont.)

So what expectations are the ξ_j?

Recall 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 (y_j is a random sum of random variables with $y_{p(j)}$ terms).

Thus

$$E(y_j | y_{p(j)}) = \sum_{i=1}^{y_{p(j)}} E(Z_i) = y_{p(j)} \xi_j$$

(because $E(Z_i) = \xi_j$). And

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

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

Mean Value Parameterizations (cont.)

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

Then we have to 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 the general definition that works in all cases.)

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. The R package `aster` uses the same dumb definition.

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).

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.)

To find μ_k in terms of ξ, follow the arrows going backwards (in the opposite direction the arrow points) back to the initial node, multiplying by the ξ_j for each arrow and the y_j for the initial node.

Recall our notation $G = \{k, p(k), p(p(k)), \ldots\}$. Using this

$$\mu_k = \left(\prod_{j \in G \cap J} \xi_j \right) \left(\prod_{j \in G \setminus J} y_j \right)$$

(the second product always has exactly one term, because $G \setminus J$ is always a singleton set).

Mean Value Parameterizations (cont.)

Here is another way to write μ in terms of ξ. Let $\succ$ denote the **transitive closure of the predecessor relation** defined by $j \succ k$ 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 $\succ$, which is a strict partial order relation, is not as useful as $\succeq$, its corresponding partial order relation, defined by $j \succeq k$ if and only if $j \succ k$ or $j = k$.

$\succeq$ has the even more long-winded name: **reflexive transitive closure of the 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 $\succ$ and $\succeq$ for these relations.

Mean Value Parameterizations (cont.)

Using this new notation

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

(as before, the second product always has exactly one term).

Mean Value Parameterizations (cont.)

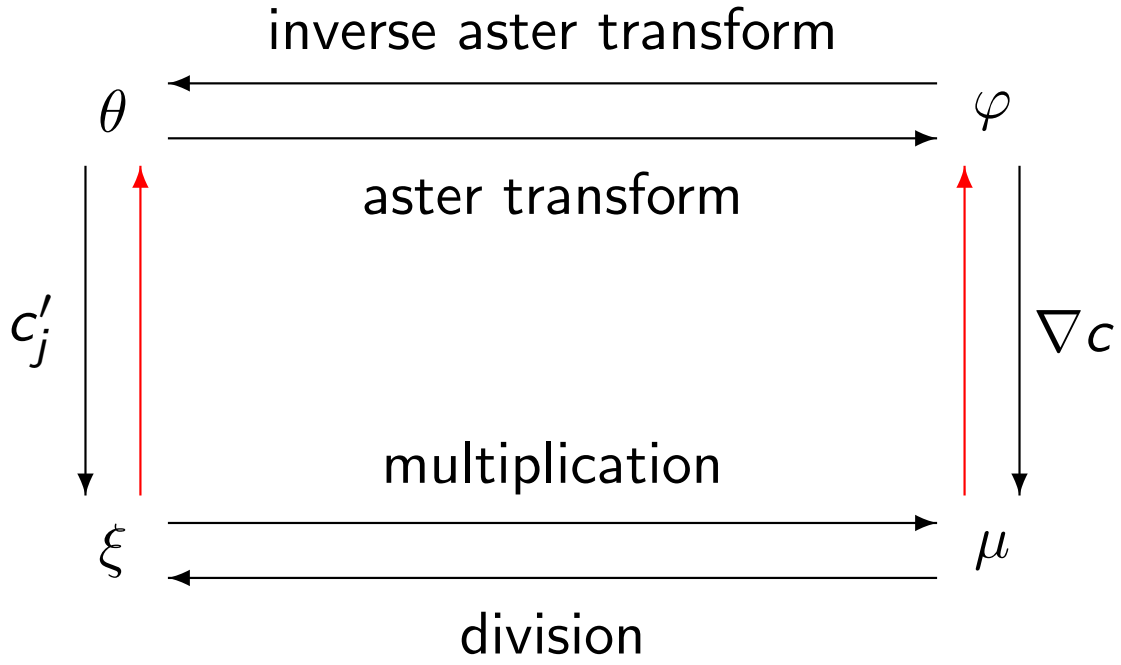
Going the other way is even easier

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

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 (this follows from the aster model distribution being non-degenerate).

A Plethora of Parameterizations

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



A Plethora of Parameterizations (cont.)

Because of the moment generating function argument, we know that cumulant functions are infinitely differentiable. This tells us that the mappings $\theta \longleftrightarrow \varphi$ and $\varphi \longrightarrow \mu$ and $\theta \longrightarrow \xi$ are infinitely differentiable. Of course, multiplication and division are infinitely differentiable, so the mappings $\xi \longleftrightarrow \mu$ are infinitely differentiable.

This leaves the two red arrows in the diagram, which have, in general, no closed form expression.

The inverse function theorem from real analysis says the inverse of an infinitely differentiable function is also infinitely differentiable. Thus all of these changes of parameter are infinitely differentiable.

A Plethora of Parameterizations (cont.)

We have closed form expressions for six of the eight transformations represented by arrows in the picture. We do not have closed form expressions for the two transformations represented by red arrows in the picture. For these, in general, we can only evaluate them by optimization for any given argument.

Formulas for their derivatives are given by the inverse function theorem.

A Plethora of Parameterizations (cont.)

Recall that $\mu = h(\varphi)$, where $h(\varphi) = \nabla c(\theta)$, and we proved that if $\mu^* = h(\varphi^*)$, then φ^* is the unique global maximizer of the function

$$I(\varphi) = \langle \mu^*, \varphi \rangle - c(\varphi)$$

Hence given μ^* we can find φ^* by optimization and set

$$h^{-1}(\mu^*) = \varphi^*$$

Curiously, although we have no closed form expression for h^{-1} we do have a closed form expression for its derivative. The inverse function theorem says

$$\nabla h^{-1}(\mu^*) = \big(\nabla h(\varphi^*)\big)^{-1} = \big(\nabla^2 c(\varphi^*)\big)^{-1}$$

when $\mu^* = h(\varphi^*)$ and $\varphi^* = h^{-1}(\mu^*)$.

A Plethora of Parameterizations (cont.)

Similarly recall that $\xi_j = h_j(\theta_j)$, where $h_j(\theta_j) = c_j'(\theta_j)$, and we proved that if $\xi_j^* = h_j(\theta_j^*)$, then θ_j^* is the unique global maximizer of the function

$$I(\theta_j) = \xi_j^* \theta_j - c_j(\theta_j)$$

Hence given ξ_j^* we can find θ_j^* by optimization and set

$$h_j^{-1}(\xi_j^*) = \theta_j^*$$

Curiously, although we have no closed form expression for h_j^{-1} we do have a closed form expression for its derivative. The inverse function theorem says

$$\frac{d}{d\xi_j^*} h^{-1}(\xi_j^*) = \frac{1}{h_j'(\theta_j^*)} = \frac{1}{c_j''(\theta_j^*)}$$

when $\xi_j^* = h_j(\theta_j^*)$ and $\theta_j^* = h_j^{-1}(\xi_j^*)$.

A Plethora of Parameterizations (cont.)

Higher order derivatives can be done by differentiating the formulas for first derivatives that come from the inverse function theorem and the rules for differentiating inverses: for scalars

$$\frac{\partial}{\partial x}\frac{1}{a} = -\frac{1}{a^2}\frac{\partial a}{\partial x}$$

and for matrices

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

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

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

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

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

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

$$e^\theta = p/(1 - p)$$
$$(1 - p)e^\theta = p$$
$$e^\theta = p + pe^\theta$$
$$e^\theta = p + pe^\theta$$
$$p = e^\theta/(1 + e^\theta)$$

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

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

$$c'(\theta) = \frac{e^{\theta}}{1 + e^{\theta}}$$

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

$$c'(\theta) = \frac{1}{e^{-\theta} + 1}$$

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

Some Distribution Theory: Bernoulli (cont.)

Thus we recover the usual theory of the Bernoulli distribution

$$E(X) = \xi$$
$$\mathrm{var}(X) = \xi(1 - \xi)$$

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

$$c(\theta) = e^\theta$$
$$c'(\theta) = e^\theta$$
$$c''(\theta) = e^\theta$$

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

$$E(X) = \xi$$
$$\mathrm{var}(X) = \xi$$

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 simple 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

$$E(X) = \frac{m}{1 - e^{-m}}$$

$$\text{var}(X) = m - \frac{m(m-1)e^{-m}}{1 - e^{-m}} - \left[\frac{me^{-m}}{1 - e^{-m}} \right]^2$$

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

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

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 not "designed".

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.

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

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: canonical affine submodels are themselves regular full exponential families.

Unconditional Canonical Affine Submodels (cont.)

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

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 with non-degenerate distributions, maximum likelihood estimates are unique if they exist and 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.

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,

A Plethora of Parameterizations (cont.)

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.)

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

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

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 non-degenerate so the MLE is unique if it exists, the unique MLE is given 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 p).

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

When you reduce the data to the MLE $\hat{\beta}$ for a statistical model, p 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

$$f_\theta(\text{whole data}) = f(\text{whole data}|\text{sufficient statistic}) f_\theta(\text{sufficient statistic})$$

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 a 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.

Maximum Entropy (cont.)

For reasons that will become apparent later, suppose we have a probability model given by PMF f on a finite state space S.

The entropy of the whole system is the expectation

$$E\{-\log f(X)\} = -\sum_{x \in S} f(x) \log f(x)$$

This can be generalized to the case where S is a countably infinite set or to a continuous probability model where the sum is replaced by an integral, but the math becomes more complicated.

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.)

Thus our problem is

$$\begin{aligned} \text{maximize} \quad & -\sum_{x \in S} f(x) \log f(x) \\ \text{subject to} \quad & \sum_{x \in S} f(x) = 1 \end{aligned}$$

We might think we also need inequality constraints $f(x) \geq 0$, $x \in S$, but it turns out that the solution to the problem above satisfies them too.

It is hard to maximize with respect to a general function, but since we are assuming a finite state space, we can consider f a finite-dimensional vector with components $f(x)$, $x \in S$.

Maximum Entropy (cont.)

To solve this we use the method of Lagrange multipliers.

Multiply the constraint function by a new parameter (Lagrange multiplier) and add to the objective function. This gives the Lagrangian function

$$\begin{aligned}\mathcal{L}(f) &= -\sum_{x \in S} f(x) \log f(x) + \psi \sum_{x \in S} f(x) \\ &= -\sum_{x \in S} f(x) \big[\log f(x) - \psi\big]\end{aligned}$$

ψ is the Lagrange multiplier.

Maximum Entropy (cont.)

The method of Lagrange multipliers maximizes the Lagrangian.

The unknown in our problem is not x, it is the function f. However, in our setup (finite state space), we can think of f as just a vector specifying its values $f(x)$ for x in the finite set S.

Thus we differentiate the Lagrangian with respect to $f(x)$ not x.

$$\frac{\partial \mathcal{L}(f)}{\partial f(x)} = -\log f(x) + \psi - 1$$

setting this equal to zero and solving for $f(x)$ gives

$$f(x) = e^{\psi - 1}$$

Maximum Entropy (cont.)

Then we have to find the value of the Lagrange multiplier that makes the constraint satisfied.

Here we see that since $f(x)$ does not depend on x, it must be the uniform distribution on the state space.

But that isn't the problem we wanted to do.

That was just a warm-up exercise.

Maximum Entropy (cont.)

In order to get a non-trivial answer, we add more constraints.

Suppose we "know" the value of some expectations

$$\mu_j = E\{t_j(X)\} = \sum_{x \in S} t_j(x) f(x), \qquad j \in J$$

and we want f to maximize entropy subject to these constraints too.

Maximum Entropy (cont.)

Thus our problem is

$$\begin{aligned} \text{maximize} \quad & -\sum_{x \in S} f(x) \log f(x) \\ \text{subject to} \quad & \sum_{x \in S} t_j(x) f(x) = \mu_j, \qquad j \in J \\ & \sum_{x \in S} f(x) = 1 \end{aligned}$$

We might think we also need inequality constraints $f(x) \geq 0$, $x \in S$, but it turns out that the solution to the problem above satisfies them too (as we saw in the warm-up exercise).

Maximum Entropy (cont.)

To solve this we use the method of Lagrange multipliers. Multiply each constraint function by a new parameter (Lagrange multiplier) and add to the objective function. This gives the Lagrangian function

$$\mathcal{L}(f) = -\sum_{x \in S} f(x) \log f(x) + \sum_{j \in J} \varphi_j \sum_{x \in S} t_j(x) f(x) + \psi \sum_{x \in S} f(x)$$

$$= -\sum_{x \in S} f(x) \left[\log f(x) - \sum_{j \in J} \varphi_j t_j(x) - \psi \right]$$

φ_j, $j \in J$, and ψ are the Lagrange multipliers.

Maximum Entropy (cont.)

As before, we differentiate with respect to $f(x)$ not x.

$$\frac{\partial \mathcal{L}(f)}{\partial f(x)} = -\log f(x) + \sum_{j \in J} \varphi_j t_j(x) + \psi - 1$$

setting this equal to zero and solving for $f(x)$ gives

$$f(x) = \exp\left(\sum_{j \in J} \varphi_j t_j(x) + \psi - 1 \right)$$

Maximum Entropy (cont.)

Then we have to find the value of the Lagrange multipliers that make all the constraints satisfied. In aid of this, define φ to be the vector having components φ_j and $t(x)$ to be the vector having components $t_j(x)$, so we can write

$$f(x) = e^{\langle t(x),\varphi\rangle+\psi-1}$$

In order to satisfy the constraint that the probabilities sum to one we must have

$$e^{\psi-1}\sum_{x\in S} e^{\langle t(x),\varphi\rangle} = 1$$

or

$$1-\psi = \log\left(\sum_{x\in S} e^{\langle t(x),\varphi\rangle}\right)$$

Maximum Entropy (cont.)

Define

$$c(\varphi) = \log\left(\sum_{x \in S} e^{\langle t(x), \varphi \rangle}\right)$$

Then

$$f(x) = e^{\langle t(x), \varphi \rangle - c(\varphi)}$$

That looks familiar!

If we think of the Lagrange multipliers φ_j as unknown parameters rather than constants we still have to adjust, then we see that we have an exponential family with canonical statistic vector $t(x)$, canonical parameter vector φ, and cumulant function c.

Maximum Entropy (cont.)

Define μ to be the vector with components μ_j. Then we know from exponential family theory that

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

and h is a one-to-one function, so the Lagrange multiplier is

$$\varphi = h^{-1}(\mu)$$

and although we do not have a closed form expression for h^{-1} we can evaluate $h^{-1}(\mu)$ for any μ that is a possible value of the mean value parameter vector by doing an optimization.

Maximum Entropy (cont.)

But that isn't quite the problem we wanted to do.

We didn't get all exponential families on S.

For example if $S = \{0, 1, \ldots, n\}$. The binomial distribution has the form

$$f(x) = \binom{n}{x} e^{x\varphi - c(\varphi)}, \qquad x \in S,$$

and what we got from the maximum entropy formalism only has the second term but is missing the binomial coefficient.

So we need yet another complication.

Maximum Entropy (cont.)

Let m be a strictly positive function on S, which we think of as a positive measure (it does not have to be a PMF, it can be unnormalized). The relative entropy of f with respect to m is

$$-\sum_{x \in S} f(x) \log\left(\frac{f(x)}{m(x)}\right)$$

(it is the negative of this quantity that is Kullback-Leibler information of f with respect to m).

Maximum Entropy (cont.)

If we repeat the argument with this as the definition of entropy (for once we omit the details), we now get the solution

$$f(x) = m(x)e^{\langle t(x),\varphi\rangle - c(\varphi)}, \qquad x \in S,$$

where

$$c(\varphi) = \log\left(\sum_{x\in S} m(x)e^{\langle t(x),\varphi\rangle}\right)$$

Now we have all exponential families on S.

But the measure m seems rather arbitrary. Some people think of it as a Bayesian prior distribution. We call it the **base measure** of the exponential family.

Maximum Entropy (cont.)

The base measure can always be chosen to be a probability distribution in the exponential family, but doesn't have to be.

For example, we can use

$$m(x) = \binom{n}{x}$$

to get the binomial distribution for sample size n.

Maximum Entropy (cont.)

Our use of the maximum entropy argument is a bit peculiar.

First we said that we "knew" the expectations

$$\mu = E\{t(X)\}$$

and wanted to pick out one probability distribution that maximizes entropy and satisfies this constraint.

Then we forgot about "knowing" this constraint and said as μ ranges over all possible values we get an exponential family of probability distributions.

Also we have to choose a base measure.

Maximum Entropy (cont.)

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

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

$$\tau = E(M^T y) \qquad (*)$$

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

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

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.

Proof. Let c be a convex function and h its gradient function, which is defined by

$$h(\varphi) = \nabla c(\varphi), \qquad \varphi \in \operatorname{dom} f$$

Suppose φ^* and φ^{**} are in $\operatorname{dom} c$, and define a univariate function

$$g(t) = \langle h\big(t\varphi^{**} + (1-t)\varphi^*\big) - h(\varphi^*), \varphi^{**} - \varphi^* \rangle, \qquad 0 \le t \le 1.$$

Its derivative is

$$g'(t) = (\varphi^{**} - \varphi^*)^T \nabla h\big(t\varphi^{**} + (1-t)\varphi^*\big)(\varphi^{**} - \varphi^*)$$

Multivariate Monotonicity (cont.)

Because

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

is positive semi-definite by convexity, we have $g'(t) \geq 0$ for all t and by the fundamental theorem of calculus

$$g(1) = g(0) + \int_0^1 g'(t)\, dt \geq g(0)$$

Now observe that $g(0) = 0$ and

$$g(1) = \langle h(\varphi^{**}) - h(\varphi^*), \varphi^{**} - \varphi^* \rangle$$

QED

Multivariate Monotonicity (cont.)

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

(The proof is almost the same.)

Corollary. The mapping from canonical parameter vector to mean value parameter vector for a non-degenerate exponential family is strictly multivariate monotone.

Corollary. The mapping from unconditional canonical parameter vector to unconditional mean value parameter vector for a non-degenerate aster model is strictly multivariate monotone.

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

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

and

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

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

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

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 one-parameter aster models for arrows of the graph, we see that

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

is an increasing function for each j.

Thus 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.

But more is true. Actually,

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

Multivariate Monotonicity (cont.)

Not only are the map $\theta \longrightarrow \xi$ and its inverse strictly multivariate monotone, but also 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 two assumptions: **acyclic graph** and nodes have **at most one predecessor**. This implies 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 arrows are one-parameter 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, ... (polynomial regression) or $\sin(x)$, $\cos(x)$, $\sin(2x)$, $\cos(2x)$, ... (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

$$\mu_i = \sum_{j=1}^{p} x_{ij} \beta_j \qquad (*)$$

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

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

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

$$\beta \mapsto M\beta$$
$$\mu \mapsto M^T \mu$$

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

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

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: p 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

$$\hat{\tau} = M^T y \qquad (*)$$

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:popNessman	-0.1862690	0.1277869	-1.4577	0.14494
fit:popNWLF	0.0210283	0.0635998	0.3306	0.74092
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 `varbf102` 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      resp
 "factor" "integer" "integer"  "factor" "integer"
       id      root     layer       fit
"integer" "numeric"  "factor" "numeric"

> levels(redata$layer)

[1] "fl"   "hdct" "ld"
```

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"      "Nessman" "NWLF"
[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, in addition to the changes in cumulant functions that accompany changes of canonical statistic or canonical parameter, a cumulant function is only determined up to an unknown additive constant, so we can always change the cumulant function to

$$c_{\text{new}}(\varphi) = a + c(\varphi)$$

where a is a known scalar, without changing the canonical statistic or canonical parameter.

A Technical Quibble (cont.)

Summary.

- Any one-to-one linear function of a canonical statistic is another canonical statistic (this also changes the canonical parameter and cumulant function).
- Any one-to-one linear function of a canonical parameter is another canonical parameter (this also changes the canonical statistic and cumulant function).
- An arbitrary constant can be added to a cumulant function (this does not change the canonical statistic or canonical parameter).

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 data 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?

Predecessor is Sample Size Revisited

Recall from deck 1, for the graph,

$$1 \xrightarrow{\text{Ber}} y_1 \xrightarrow{\text{Ber}} y_2$$

the conditional distribution of y_2 given y_1 is

- degenerate, concentrated at zero if $y_1 = 0$
- Bernoulli(ξ_2) if $y_1 = 1$

Recognizing that y_2 is zero-or-one-valued and any zero-or-one-valued random variable is Bernoulli, the unconditional distribution of y_2 is Bernoulli(μ_2), which denotes Bernoulli with mean μ_2.

Predecessor is Sample Size Revisited (cont.)

And, for this graph,

$$1 \xrightarrow{\text{Poi}} y_1 \xrightarrow{\text{Ber}} y_2$$

the conditional distribution of y_2 given y_1 is

- degenerate, concentrated at zero if $y_1 = 0$
- Binomial(y_1, ξ_2), which denotes binomial with sample size y_1 and mean μ_2, if $y_1 > 0$

Those that know that a "thinned Poisson" is again Poisson recognize that the unconditional distribution of y_2 is Poisson(μ_2), which denotes Poisson with mean $\mu_2 = \xi_2 \xi_1$.

Predecessor is Sample Size Revisited (cont.)

And, for this graph,

$$1 \xrightarrow{\text{Ber}} y_1 \xrightarrow{\text{Ber}} y_2 \xrightarrow{\text{0-Poi}} y_3$$

the conditional distribution of y_3 given y_1 (two arrows) is

- degenerate, concentrated at zero if $y_1 = 0$
- zero-inflated Poisson with mean $\xi_3 \xi_2$, if $y_1 = 1$

Predecessor is Sample Size Revisited (cont.)

And, for this graph,

$$1 \xrightarrow{\text{Poi}} y_1 \xrightarrow{\text{Ber}} y_2 \xrightarrow{\text{0-Poi}} y_3$$

the conditional distribution of y_3 given y_1 (two arrows) is

- degenerate, concentrated at zero if $y_1 = 0$
- zero-inflated Poisson with mean $\xi_3 \xi_2$, if $y_1 = 1$
- not a brand-name distribution if $y_1 > 1$, but it can be described as the sum of y_1 IID zero-inflated Poisson random variables with mean $\xi_3 \xi_2$.

Predecessor is Sample Size Revisited (cont.)

Although the usual value of the constant at an initial node i is $y_i = 1$, this is not necessary.

This value plays the role of sample size, so it must be a positive integer, but it can be any positive integer.

If $y_i = 2$, then the graph describes 2 individuals rather than 1, and all components of the response are the total for these two individuals.

If we look at our preceding examples, everything is almost the same if we take what was y_1 to be an initial node.

Predecessor is Sample Size Revisited (cont.)

For this graph,

$$n \xrightarrow{\text{Ber}} y_1$$

The unconditional distribution of y_1 is Binomial(n, ξ_1), and $\mu_1 = n\xi_1$.

Predecessor is Sample Size Revisited (cont.)

And, for this graph,

$$n \xrightarrow{\text{Poi}} y_1$$

Recalling that the sum of IID Poisson is Poisson, the unconditional distribution of y_1 is Poisson$(n\xi_1)$, and $\mu_1 = n\xi_1$.

Predecessor is Sample Size Revisited (cont.)

And, for this graph,

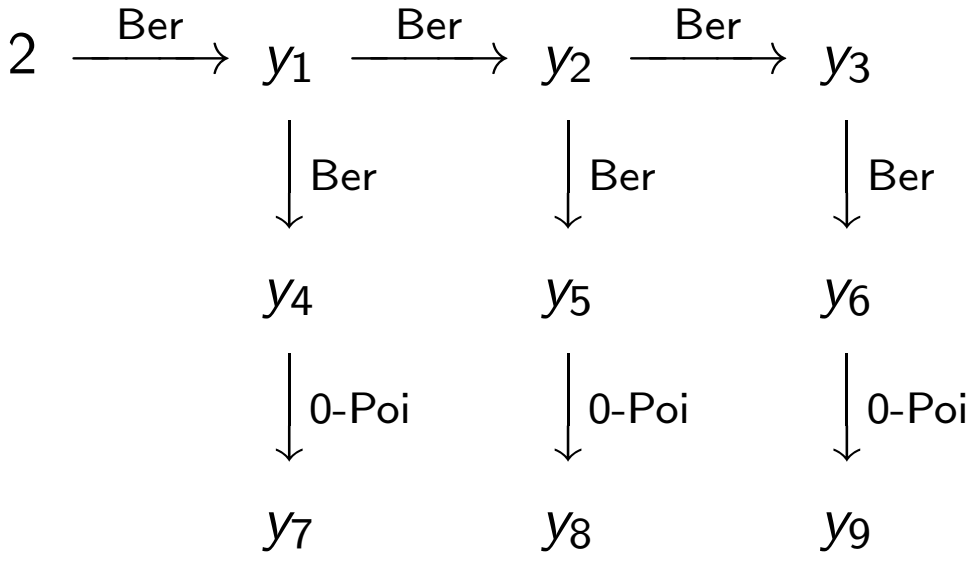
$$n \xrightarrow{\text{Ber}} y_1 \xrightarrow{\text{0-Poi}} y_2$$

the unconditional distribution of y_2 (two arrows) is

- zero-inflated Poisson with mean $\xi_3 \xi_2$, if $n = 1$
- not a brand-name distribution if $n > 1$, but it can be described as the sum of n IID zero-inflated Poisson random variables with mean $\xi_2 \xi_1$, and $\mu_2 = \xi_2 \xi_1 n$.

Predecessor is Sample Size Revisited (cont.)

And, for this graph,



which is just like the graph for example 1 except the initial node has the value 2 rather than 1, suppose we have the same interpretation of all the variables as in example 1 except for the differences entailed by the change of value at the initial node.

Predecessor is Sample Size Revisited (cont.)

y_1 is survival in the first year of the experiment. It has the value zero, one, or two. It is the number of the original two individuals that survive.

y_2 is survival in the second year of the experiment. It has integer values ranging from zero to y_1. It is the number of the original two individuals that survive to this point.

y_3 is survival in the third year of the experiment. It has integer values ranging from zero to y_2. It is the number of the original two individuals that survive to this point.

Predecessor is Sample Size Revisited (cont.)

y_4 is the number of individuals that flowered in the first year of the experiment. It has integer values ranging from zero to y_1.

Similarly, y_5 is the number of individuals that flowered in the second year. It has integer values ranging from zero to y_2.

And so forth.

Predecessor is Sample Size Revisited (cont.)

y_7 is the (compound) flower count of the y_1 individuals were surviving in the first year. It is the total number of flowers on up to two individuals (depending on how many of the 2 individuals at the initial node were still surviving when censused in the first year).

The conditional distribution of y_7 given y_4 is the sum of y_4 zero-truncated Poisson random variables, and $E(y_7|y_4) = \xi_7 y_4$.

The conditional distribution of y_7 given y_1 is the sum of y_1 zero-inflated Poisson random variables, and $E(y_7|y_1) = \xi_7 \xi_4 y_1$.

And so forth.

Predecessor is Sample Size Revisited (cont.)

Setting the initial node to more than one isn't used very much, but it has been used (researchers had experimental designs where they only collected data on groups of individuals rather than single individuals).

It's a nice feature of aster models, that they "just work" in this case too.