Key Ideas about Aster Models

Charles J. Geyer
School of Statistics
University of Minnesota

Ruth G. Shaw
Department of Ecology, Evolution, and Behavior
University of Minnesota

http://www.stat.umn.edu/geyer/aster/

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

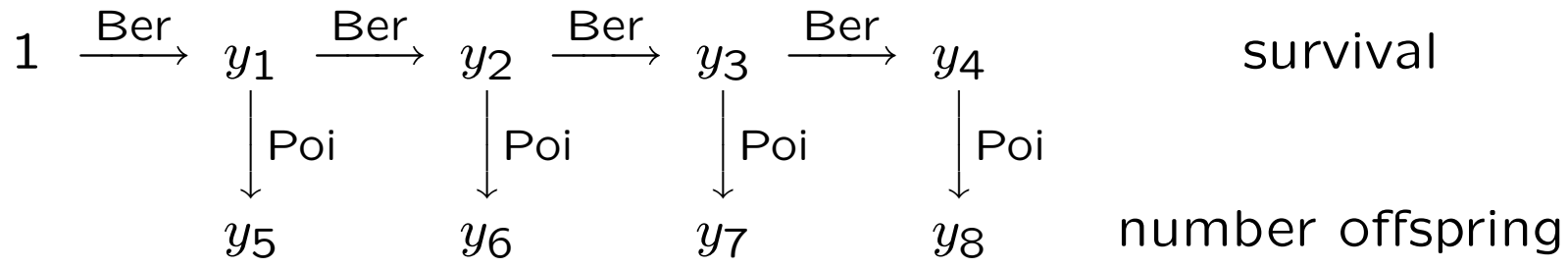
Shaw, R. G., Geyer, C. J., Wagenius, S., Hangelbroek, H. H.,
and Etterson, J. R. (2008).
Unifying life history analysis for inference of fitness and
population growth.
American Naturalist, **127** E35-–E47.

All details of all computations given in tech reports at
http://www.stat.umn.edu/geyer/aster/

Statistical Model Hierarchy

- linear models (multiple regression and ANOVA)
  - responses are independent from normal distribution
  - means are linear function of regression coefficients
- generalized linear models (logistic and Poisson regression)
  - responses are independent from same distribution
  - means are monotone function of regression coefficients
- aster models (life history analysis)
  - responses are dependent from different distributions
  - means are monotone function of regression coefficients

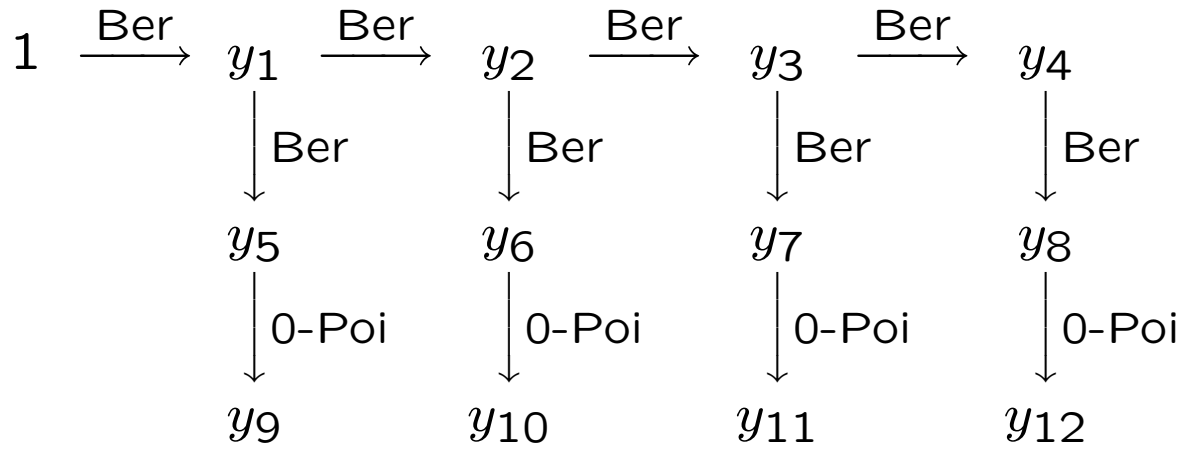
Graphical Model



y_j are components of response for one individual

Arrows indicate conditional distributions (Ber = Bernoulli and Poi = Poisson)

Another Graphical Model



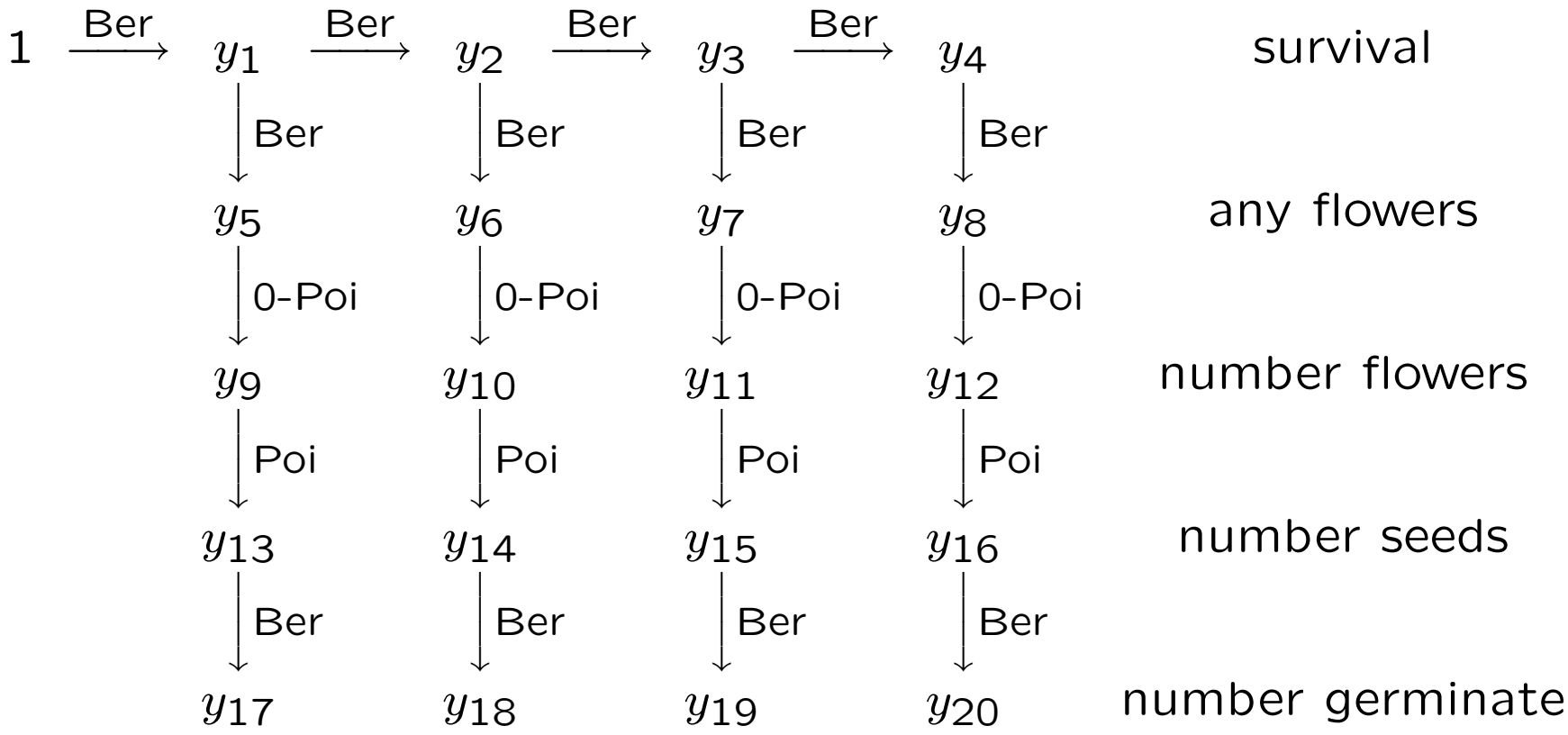
survival

any offspring

number offspring

0-Poi = zero-truncated Poisson

Yet Another Graphical Model



Predecessor is Sample Size

$$\underset{\text{predecessor}}{y_{p(j)}} \longrightarrow \underset{\text{successor}}{y_j}$$

$y_{p(j)}$ is sample size for y_j. Only form of dependence allowed.

$$y_4 \xrightarrow{\text{Ber}} y_8 \xrightarrow{\text{0-Poi}} y_{12} \xrightarrow{\text{Poi}} y_{16} \xrightarrow{\text{Ber}} y_{20}$$

y_8 is successor of y_4 and predecessor of y_{12}
y_{12} is successor of y_8 and predecessor of y_{16}
etc.

Means, Conditional and Unconditional

"Predecessor is sample size" is the only form of dependence allowed in aster models. It has following important consequence. Define

$$\begin{aligned} \mu_j &= E(y_j) \\ \xi_j &= E(y_j \mid y_{p(j)} = 1) \end{aligned}$$

(ξ_j is the mean of one of the $y_{p(j)}$ independent and identically distributed random variables of which y_j is the sum). Then

$$\begin{aligned} E(y_j \mid y_{p(j)}) &= \xi_j y_{p(j)} \\ E(y_j) &= \xi_j \mu_{p(j)} \\ &= \xi_j \xi_{p(j)} \mu_{p(p(j))} \\ &= \xi_j \xi_{p(j)} \xi_{p(p(j))} \mu_{p(p(p(j)))} \end{aligned}$$

and so forth.

Means, Conditional and Unconditional (cont.)

If we go back to a root node, say $y_{p(p(p(p(p(j)))))} = 1$, then

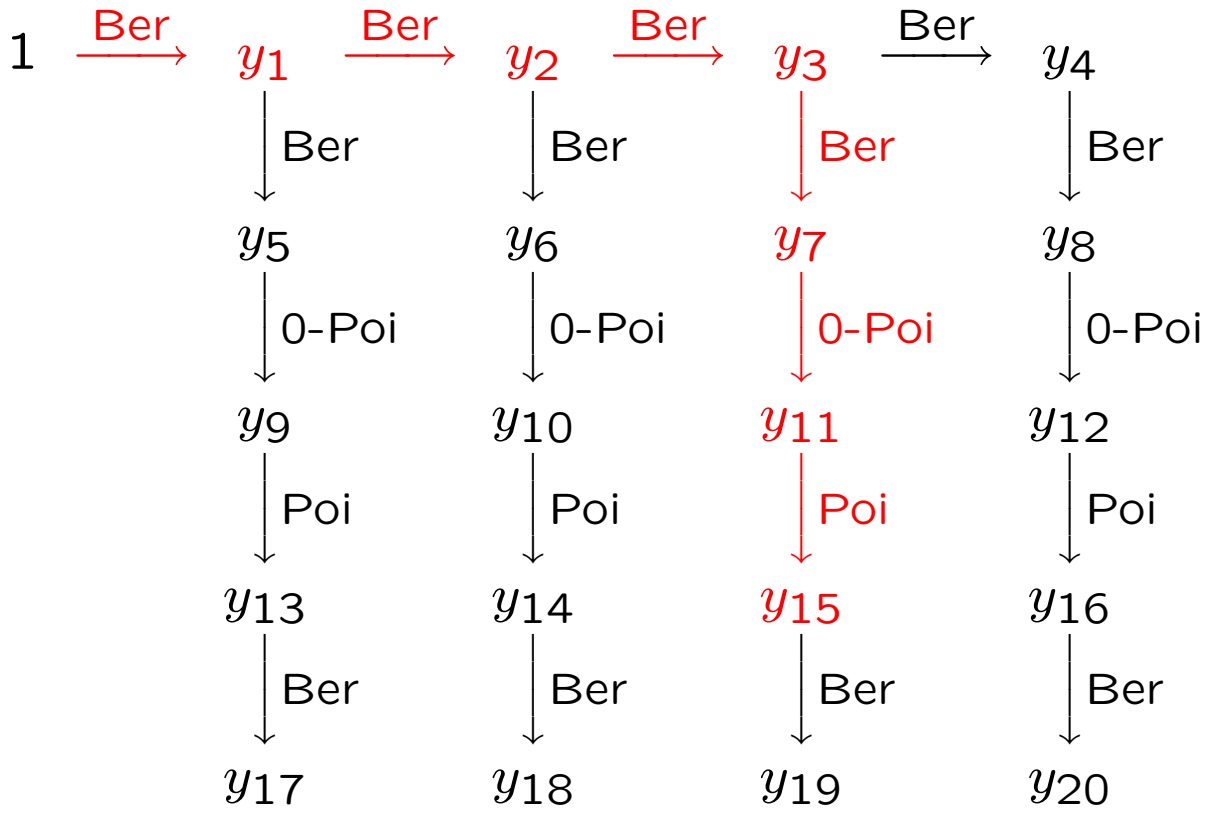
$$\mu_j = \xi_j \xi_{p(j)} \xi_{p(p(j))} \xi_{p(p(p(j)))} \xi_{p(p(p(p(j))))}$$

So unconditional means are products of (a certain set of) conditional means.

Very special property of the aster model structure, consequence of predecessor is sample size. Not true of general dependent data.

Means, Conditional and Unconditional (cont.)

$$\mu_{15} = \xi_{15}\xi_{11}\xi_{7}\xi_{3}\xi_{2}\xi_{1}$$



survival

any flowers

number flowers

number seeds

number germinate

Log Likelihood

Also require each conditional distribution for each arrow in the graph is one-parameter exponential family. Then log likelihood has very simple form

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

where $\boldsymbol{\theta}$ the parameter vector, θ_j its components, J set of non-root nodes of the graph where the variables are random, and $c_j(\theta)$ different but known function for each different one-parameter exponential family distribution.

Cumulant Functions

The function c_j is called *cumulant function* for the j-th one-parameter exponential family. It satisfies

$$\begin{aligned} E_{\theta_j}(y_j \mid y_{p(j)}) &= y_{p(j)} c_j'(\theta_j) \\ &= y_{p(j)} \xi_j \\ \operatorname{var}_{\theta_j}(y_j \mid y_{p(j)}) &= y_{p(j)} c_j''(\theta_j) \end{aligned}$$

where primes indicate derivatives.

Since variances are positive $c_j''(\theta_j) > 0$ and this means the map $\xi_j = c'(\theta_j)$ is strictly increasing, hence one-to-one.

Thus ξ_j is just as good a parameter as θ_j. To distinguish them we call ξ_j the *conditional mean value parameter* and θ_j the *conditional canonical parameter*.

Change of Parameter

Put log likelihood in multiparameter exponential family form

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= \sum_{j \in J} \left[y_j \theta_j - y_{p(j)} c_j(\theta_j) \right] \\
&= \sum_{j \in J} y_j \left[\theta_j - \sum_{\substack{i \in J \\ p(i) = j}} c_i(\theta_i) \right] - \sum_{\substack{i \in J \\ p(i) \notin J}} y_{p(i)} c_i(\theta_i) \\
&= \sum_{j \in J} y_j \varphi_j - c(\boldsymbol{\varphi}) \\
&= \langle \mathbf{y}, \boldsymbol{\varphi} \rangle - c(\boldsymbol{\varphi})
\end{aligned}
$$

Change of Parameter (cont.)

New parameters

$$\varphi_j = \theta_j - \sum_{\substack{i \in J \\ p(i)=j}} c_i(\theta_i)$$

are one-to-one function of old parameters θ_j. Can solve these for θ_j in terms of φ_j by using one equation at a time in any order that does successors before predecessors.

Thus $\boldsymbol{\varphi}$ is just as good a parameter as $\boldsymbol{\theta}$. To distinguish them we call $\boldsymbol{\varphi}$ the *unconditional canonical parameter* vector and $\boldsymbol{\theta}$ the *conditional canonical parameter* vector.

Multiparameter Cumulant Function

Cumulant function for the multivariate, multiparameter, joint, unconditional exponential family for all the variables

$$c(\boldsymbol{\varphi}) = \sum_{\substack{i \in J \\ p(i) \notin J}} y_{p(i)} c_i(\theta_i)$$

satisfies

$$\begin{aligned} E_{\boldsymbol{\varphi}}(\mathbf{y}) &= \nabla c(\boldsymbol{\varphi}) \\ &= \boldsymbol{\mu} \\ \text{var}_{\boldsymbol{\varphi}}(\mathbf{y}) &= \nabla^2 c(\boldsymbol{\varphi}) \end{aligned}$$

where $E_{\boldsymbol{\varphi}}(\mathbf{y})$ and $\text{var}_{\boldsymbol{\varphi}}(\mathbf{y})$ denote the mean vector and variance-covariance matrix of the random vector $\mathbf{y}$ and $\nabla c(\boldsymbol{\varphi})$ and $\nabla^2 c(\boldsymbol{\varphi})$ denote the vector of first partial derivatives and the matrix of second partial derivatives of the cumulant function.

Multiparameter Cumulant Function (cont.)

Since variance-covariance matrices $\operatorname{var}_{\boldsymbol{\varphi}}(\mathbf{y}) = \nabla^2 c(\boldsymbol{\varphi})$ are positive definite, the function $c(\boldsymbol{\varphi})$ is strictly convex. This means the map $\boldsymbol{\mu} = \nabla c(\boldsymbol{\varphi})$, which can be solved for $\boldsymbol{\varphi}$ by maximizing

$$\langle \boldsymbol{\mu}, \boldsymbol{\varphi} \rangle - c(\boldsymbol{\varphi}),$$

is one-to-one (because strictly concave functions have unique maximizers).

Thus $\boldsymbol{\mu}$ is just as good a parameter as $\boldsymbol{\varphi}$. To distinguish them we call $\boldsymbol{\mu}$ the *unconditional mean value parameter* vector and $\boldsymbol{\varphi}$ the *conditional canonical parameter* vector.

A Plethora of Parameters

Too many parameterizations?

	conditional	unconditional
canonical	θ	φ
mean value	ξ	μ

In generalized linear models have only two, because $\varphi = \theta$ and $\mu = \xi$ when all components of response vector are independent.

In linear models have only one, because $\mu = \varphi$ when all components of response vector are normally distributed.

Not too many! Unavoidable consequence of generality of aster models.

Multivariate Monotonicity

Since map $\boldsymbol{\varphi} \mapsto \boldsymbol{\mu}$ is given by

$$\boldsymbol{\mu} = \nabla c(\boldsymbol{\varphi})$$

and $c(\boldsymbol{\varphi})$ is a strictly convex function, this map is *multivariate monotone:* if $\boldsymbol{\varphi} \neq \boldsymbol{\varphi}^*$ and

$$\boldsymbol{\mu} = \nabla c(\boldsymbol{\varphi})$$
$$\boldsymbol{\mu}^* = \nabla c(\boldsymbol{\varphi}^*)$$

then

$$(\boldsymbol{\mu} - \boldsymbol{\mu}^*)^T(\boldsymbol{\varphi} - \boldsymbol{\varphi}^*) > 0$$

Not as easy to visualize as univariate monotonicity, but just as useful.

Canonical Affine Submodels

Model discussed so far has too many parameters (now meaning *dimension* of parameter vector), one per response variable ($\boldsymbol{\varphi}$, $\boldsymbol{\theta}$, $\boldsymbol{\mu}$, $\boldsymbol{\xi}$ all same dimension as $\mathbf{y}$).

Both log likelihoods

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

$$l(\boldsymbol{\varphi}) = \langle \mathbf{y}, \boldsymbol{\varphi} \rangle - c(\boldsymbol{\varphi})$$

are strictly convex functions so maximizer is unique if it exists.

Canonical Affine Submodels (cont.)

To preserve strict convexity, use affine submodels,

$$\boldsymbol{\theta} = \mathbf{a} + \mathbf{M}\boldsymbol{\beta} \tag{1}$$

or

$$\boldsymbol{\varphi} = \mathbf{a} + \mathbf{M}\boldsymbol{\beta} \tag{2}$$

where $\mathbf{a}$ is known vector and $\mathbf{M}$ is known matrix, called *offset* and *model matrix*.

Choose either (1) or (2). Never use both. That's why can use same notation on right-hand side. Never a conflict.

If (1) have *conditional aster model*. If (2) have *unconditional aster model*.

Canonical Affine Submodels (cont.)

So which one do you want?

Tempting, but wrong answer: use conditional or unconditional aster model to specify conditional or unconditional distributions, respectively.

No, because φ, θ, μ, and ξ are all one-to-one functions of each other. Specify one, specify all. Each determines all conditional and unconditional distributions and conditional and unconditional means, variances, and covariances.

Right answer: use conditional or unconditional aster model to establish a monone relationship to conditional or unconditional means, respectively.

Forget Conditional Aster Models

In most applications unconditional means are more interpretable than conditional means. Hence most applications need unconditional aster model.

Conditional aster models may be of some use in rare applications, so they are provided, but vast majority use unconditional.

Forget Conditional Aster Models (cont.)

Everything else in these slides about unconditional aster models with

$$\varphi = \mathbf{a} + \mathbf{M}\beta$$

Still also have parameters μ, ξ, and θ, since they are one-to-one functions of φ.

Prediction

A referee for the Biometrika paper pointed out that "prediction" is the wrong word here. What we are talking about is parameter estimates of

- any of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$, $\boldsymbol{\theta}$, $\boldsymbol{\mu}$, $\boldsymbol{\xi}$,

- or any differentiable functions of those parameter vectors $g(\boldsymbol{\beta})$, $g(\boldsymbol{\varphi})$, $g(\boldsymbol{\theta})$, $g(\boldsymbol{\mu})$, $g(\boldsymbol{\xi})$.

And we are talking about point estimates and standard errors.

Prediction (Cont.)

The reason it is called "prediction"

- In elementary statistics the mean value parameter estimates $\widehat{\mu}$ are called "predicted values" $\widehat{\mathbf{y}}$ to keep the kiddies from being confused (of course they are being confused by not calling a parameter estimate a parameter estimate, but they are unaware of their confusion).

- Consequently, the R function that does this job for linear and generalized linear models is called `predict`.

- The `predict` function is generic, we want it to work the same way for aster models as it does for LM and GLM, hence the name must be used.

Prediction (Cont.)

Function `predict.aster` doesn't actually do arbitrary functions $g(\boldsymbol{\mu})$ in one step. It does do arbitrary *linear* functions $\mathbf{A}\boldsymbol{\mu}$ in one step, where $\mathbf{A}$ is an arbitrary matrix.

But that is enough to do general functions by the delta method.

If have $\hat{\boldsymbol{\mu}}$, then $g(\hat{\boldsymbol{\mu}})$ is the MLE point estimate of $g(\boldsymbol{\mu})$ by invariance of maximum likelihood.

If $\mathbf{A} = \nabla g(\boldsymbol{\mu})$, then $\mathbf{A}\hat{\boldsymbol{\mu}}$ and $g(\hat{\boldsymbol{\mu}})$ have the same asymptotic variance and hence same standard errors by the delta method.

Multivariate Monotonicity (Cont.)

Details about consequences of

$$(\boldsymbol{\mu} - \boldsymbol{\mu}^*)^T(\boldsymbol{\varphi} - \boldsymbol{\varphi}^*) > 0$$

Fitness is sum over subset G of nodes of graph; observed fitness $\sum_{j \in G} y_j$, and expected fitness $\sum_{j \in G} \mu_j$.

Suppose want to estimate fitness landscape (fitness as a function of phenotypic covariate(s) $\mathbf{z}$. By tradition (Lande and Arnold, 1983) much literature is focused on quadratic function $q(\mathbf{z})$ but these don't make sense for means (same reason it usually doesn't make sense to model means directly in GLM and aster models).

Multivariate Monotonicity (Cont.)

So model unconditional canonical parameter $\boldsymbol{\varphi}$ quadratically. Let $\mathbf{x}$ be all non-phenotypic covariates and

$$\varphi_j(\mathbf{x}, \mathbf{z}) = \begin{cases} a_j(\mathbf{x}) + q(\mathbf{z}), & j \in G \\ a_j(\mathbf{x}), & \text{otherwise} \end{cases}$$

Now consider two individuals with same value $\mathbf{x}$ of non-phenotypic covariates and values $\mathbf{z}$ and $\mathbf{z}^*$ of phenotypic ones.

$$\varphi_j(\mathbf{x}, \mathbf{z}) - \varphi_j(\mathbf{x}, \mathbf{z}') = \begin{cases} q(\mathbf{z}) - q(\mathbf{z}'), & j \in G \\ 0, & \text{otherwise} \end{cases}$$

The $a_j(\mathbf{x})$ terms drop out.

Multivariate Monotonicity (Cont.)

The multivariate monotonicity inequality written out in coordinates is

$$\begin{aligned} 0 &< \sum_{j \in J} \Big(\mu_j(\mathbf{x}, \mathbf{z}) - \mu_j(\mathbf{x}, \mathbf{z}^*) \Big) \Big(\varphi_j(\mathbf{x}, \mathbf{z}) - \varphi_j(\mathbf{x}, \mathbf{z}^*) \Big) \\ &= \sum_{j \in G} \Big(\mu_j(\mathbf{x}, \mathbf{z}) - \mu_j(\mathbf{x}, \mathbf{z}^*) \Big) \Big(q(\mathbf{z}) - q(\mathbf{z}^*) \Big) \\ &= \Big(q(\mathbf{z}) - q(\mathbf{z}^*) \Big) \sum_{j \in G} \Big(\mu_j(\mathbf{x}, \mathbf{z}) - \mu_j(\mathbf{x}, \mathbf{z}^*) \Big) \end{aligned}$$

Since a product is positive if and only if both terms are positive

$$q(\mathbf{z}) < q(\mathbf{z}^*) \quad \text{if and only if} \quad \sum_{j \in G} \mu_j(\mathbf{x}, \mathbf{z}) < \sum_{j \in G} \mu_j(\mathbf{x}, \mathbf{z}*)$$

In short, there is monotone relationship between fitness on the canonical parameter scale on on the mean value parameter scale.

Sufficiency

R. A. Fisher invented many statistical concepts including sufficiency and exponential families. A statistic is sufficient if it contains all the information in the data about the parameter.

In exponential families, the canonical statistic is minimal sufficient.

In a canonical affine submodel, the canonical statistic is $\mathbf{M}^T \mathbf{y}$

The likelihood equations are "observed equals expected"

$$\mathbf{M}^T \mathbf{y} = \mathbf{M}^T \hat{\boldsymbol{\mu}}$$

So maximum likelihood sets the observed value of the minimal sufficient statistic to its expected value and *doesn't pay any attention to anything else*.

Sufficiency (Cont.)

If the canonical statistics make scientific sense, then so does the statistical model. And vice versa.

In our fitness landscape example, the aster models we use have the same canonical statistics as the Lande-Arnold analysis. We are doing what the scientistists have long identified as important, the only difference is that we are now using a believable statistical model.

Entropy

A rather eccentric physicist Edwin Jaynes invented the notion of "maximum entropy" modeling and estimation. Widely used in several areas. Saw two posters using it in GIS models.

For our purposes exponential families are "maximum entropy".

A canonical affine submodel with model matrix $\mathbf{M}$ can be characterized as the family of distributions that maximizes entropy with respect to any distribution in the full family subject to the constraint

$$E(\mathbf{M}^T\mathbf{y}) = \boldsymbol{\mu}$$

for various values of $\boldsymbol{\mu}$.

Entropy (Cont.)

Thus an unconditional aster canonical affine submodel can be justified by a maximal entropy argument. It is the model that models the mean value of the sufficient statistic $\mathbf{M}^T \mathbf{y}$ and leaves everything else as random as possible.

Entropy (Cont.)

When the canonical statistic is fitness, this makes some scientific sense. Only fitness is visible to selection (by definition). So everything else just drifts and should maximize entropy.

Two caveats

- fitness isn't really fitness (usually only surrogate) and
- statistical entropy isn't real physico-chemical entropy because the statistical model doesn't model everything.

So argument isn't tight, but isn't worthless either.