A Philosophical Look at Aster Models

By

Charles J. Geyer

Technical Report No. 676
School of Statistics
February 2, 2010

1 Introduction

Aster models (Geyer, Wagenius and Shaw, 2007; Shaw, Geyer, Wagenius, Hangelbroek and Etterson, 2008) are statistical models designed especially for life history analysis of plants and animals.

Aster models are parametric statistical models. In some senses they are quite ordinary statistical models. The theory of aster models is just the theory of likelihood inference (Fisher, 1922; Severini, 2001) and the theory of exponential families of distributions (Barndorff-Nielsen, 1978; Brown, 1986; Geyer, 1990). Given an aster model and data, we use the method of maximum likelihood for parameter estimation; we use likelihood ratio tests for model comparison; and we use confidence intervals based on the asymptotic distribution of the maximum likelihood estimator (MLE). Details are given in Geyer et al. (2007); see also Section 2.11 below.

The theory mentioned in the preceding paragraph is known to all statisticians, covered, albeit not completely rigorously, in first-year graduate courses in theoretical statistics and covered again, this time completely rigorously, in second-year graduate courses in theoretical statistics. Many scientists in a wide variety of disciplines take a first-year graduate course in theoretical statistics, so this theory is very widely known.

Aster models are regression models. In some senses they are quite like linear models (LM), which include multiple linear regression and analysis of variance (explained in many textbooks, for example, Weisberg, 2005 and Oehlert, 2000), and generalized linear models (GLM) (McCullagh and Nelder, 1989). In regression models the objective is to model the conditional distribution of the response vector y given predictor data x (the structure of which does not matter). Aster models have the same objective. In LM and GLM, conditional on x the components of y are assumed to be stochastically independent and to have marginal distributions from the same family e. g., all normal, all Bernoulli, or all Poisson. In aster models, in contrast, conditional on x the components of y are allowed to be stochastically dependent and are allowed to have conditional distributions from different families e. g., some normal, some Bernoulli, and some Poisson.

LM and GLM are known to all statisticians, covered, albeit not at all rigorously, in undergraduate statistics courses and covered again, this time much more rigorously, in first-year graduate courses in applied statistics (proofs are left for theoretical statistics courses). Many scientists in a wide variety of disciplines take these courses, so these models are very widely known. Some courses teach ways of thinking about LM that do not generalize to GLM and ways of thinking about GLM that do not generalize to aster models.[1] Thus people who have had courses covering LM and GLM may have to unlearn certain things they learned about LM and GLM and learn more general concepts to understand aster models.

Aster models are graphical models (Lauritzen, 1996), but a very special case thereof. The only idea aster models take from graphical model theory is the fundamental idea of representing conditional distribution relationships between random variables by graphs. None of the mathematical theory of graphical models is used in aster model theory. Although there are a number of recent textbooks covering graphical models, many statistics departments do not teach courses using them. Thus the theory of graphical models is not widely known (few statisticians are expert in it), so it is fortunate aster models need so little of the theory of graphical models.

The "linear" in LM comes from the fact that means are modeled linearly[2]

$$\mu = M\beta, \tag{1}$$

where $\mu = E(y)$ is the mean of the response vector,[3] where M is a known matrix, called the *model matrix*, which may be an arbitrary function of predictor data x, where β is an unknown parameter to be estimated, and where $M\beta$ is multiplication of a matrix M and a vector β yielding a vector μ. Usually, the dimension of β is much less than the dimension of μ (which is the dimension of y). This "dimension reduction" in the jargon allows the explanation of something complicated involving many parameters (the components of μ) by something simple involving fewer parameters (the components of β), a "parsimonious model" in the jargon. It also contributes to efficient estimation (Section 2.11 below); the fewer the parameters, the better the estimation, in general.

[1] They may imply that the word "regression" only applies to a subset of LM (multiple regression but not analysis of variance) not to GLM or aster models. This, of course, cannot be because the terms "logistic regression" and "Poisson regression" for certain GLM are well established, as is the term "nonparametric regression" for models that are neither LM nor GLM. Most courses about LM teach "response is mean + error"

$$y = \mu + e$$

but this does not generalize even to GLM. See also Sections 2.8 and 2.10.

[2] The term "linear" is used here in the technical mathematical sense, as in "linear equations" and "linear algebra" and has no connection with the vague metaphorical sense popular in postmodernism, as in "linear thinking."

[3] Strictly speaking, we should write $\mu = E(y \mid x)$, a conditional expectation, instead of $\mu = E(y)$, an unconditional expectation, because we are conditioning on the predictor data x, but we follow common practice of not explicitly indicating this in the notation. The fact that μ depends on x is implicit in (1) because M is allowed to be an arbitrary function of x.

The "linear" in GLM comes from the fact that parameters other than means are modeled linearly

$$\eta = M\beta, \tag{2}$$

where η is a new parameter vector, called the *linear predictor*, and M and β are as in LM. The linear predictor is a componentwise monotone function of means

$$\eta_i = g(\mu_i),$$

where g is a monotone (strictly increasing) function called the *link function* and where η_i and μ_i denote components of η and μ. The reason for this monotone change-of-parameter is that modeling means linearly makes no sense when possible mean values are constrained. When y has Bernoulli (zero-or-one-valued) components, their means satisfy

$$0 \le \mu_i \le 1$$

but linear modeling (1) does not respect these constraints. The link function is often chosen so that the linear predictor is unconstrained. For example, in Bernoulli GLM the link function

$$\eta_i = \operatorname{logit}(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) \tag{3}$$

is the default (other link functions are allowed, but this is by far the most widely used), and one easily verifies that as μ_i goes from zero to one, η_i goes from minus infinity to plus infinity. GLM have a weak connection to exponential family theory in that the default link function is the one that makes the linear predictor the exponential family canonical parameter, for example, the logit link (3) for Bernoulli GLM and the log link

$$\eta_i = \log(\mu_i)$$

for Poisson GLM (again one easily verifies that as μ_i goes from zero to plus infinity, the allowed values for Poisson means, η_i goes from minus infinity to plus infinity).[4]

The "linear" in aster models comes from the fact that, just as in GLM, parameters other than means are modeled linearly (2).[5] In aster models,

[4] Not all exponential family canonical parameters are unrestricted, see footnote 21.

[5] Actually Geyer et al. (2007) describe a more general type of modeling in which (2) is generalized to

$$\eta = a + M\beta, \tag{4}$$

where a is a known vector, called the *offset vector*, and this formula is also used (rarely) with GLM. The offset term a is not needed for most GLM and aster models and we ignore it here. When (4) is used instead of (2) the proper term is "affine" rather than "linear" for this change-of-parameter.

unlike GLM, no choice of link function is allowed; the linear predictor vector η is always the exponential family canonical parameter vector. Since components of the response are allowed to be dependent in aster models, there is no simple relationship between marginal distributions of components of the response and their joint distribution (unlike LM and GLM where stochastic independence implies the joint is the product of the marginals). Thus the componentwise monotone relationship of linear predictor and means used in GLM makes less sense for aster models.[6] Instead, exponential family theory implies that the mean vector μ is a multivariate strictly monotone function of η (Barndorff-Nielsen, 1978, p. 121). The notion of a multivariate monotone function can be characterized in several different ways (Rockafellar and Wets, 2004, Chapter 12); here is one. Suppose η_1 and η_2 are two different allowed values of the linear predictor parameter vector and μ_1 and μ_2 are the corresponding values of the mean vector. Then

$$\langle \mu_1 - \mu_2, \eta_1 - \eta_2 \rangle > 0,$$

where $\langle \,\cdot\,, \,\cdot\, \rangle$ denotes the bilinear form defined by

$$\langle \mu, \eta \rangle = \sum_{i=1}^{p} \mu_i \eta_i, \qquad (5)$$

p being the dimension of μ and η. It may be thought that this property is too abstract to be of much use; it is certainly not taught in any of the theoretical or applied statistics courses mentioned above. But it is a very strong property that enables important arguments about aster modeling (Shaw and Geyer, submitted, appendix on monotonicity; see also Section 2.9 below).

2 Key Ideas behind Aster Models

2.1 Graphical Models

Prototypical data for an aster model can be pictured as shown in (6), which represents the data for one individual, which comprises eight mea-

[6] Emily Grosholz (personal communication) points out that the “thus” here is not justified. There is no logical reason (apparent to the author) why there could not, in principle, be an aster model competitor with componentwise monotone relationship between unconditional mean values and some other parameters. Such a competitor could not have all properties of aster models; see Section 3 below.

surements (components of the response vector y), denoted y_1, ..., y_8.

$$\begin{tikzcd}
1 \arrow[r, "\text{Ber}"] & y_1 \arrow[r, "\text{Ber}"] \arrow[d, "\text{Poi}"] & y_2 \arrow[r, "\text{Ber}"] \arrow[d, "\text{Poi}"] & y_3 \arrow[r, "\text{Ber}"] \arrow[d, "\text{Poi}"] & y_4 \arrow[d, "\text{Poi}"] \\
& y_5 & y_6 & y_7 & y_8
\end{tikzcd} \tag{6}$$

These data are measurements relating to four time periods. The variables y_1, ..., y_4 are Bernoulli and indicate survival; $y_i = 1$ indicates that the individual was alive at the beginning of the i-th time period, $y_i = 0$ indicates that the individual was dead at the beginning of this time period. The variables $y_5, \ldots, y_8$ are nonnegative-integer-valued and count offspring; y_i is the number of offspring the individual had in time period $i - 4$.

The arrows represent stochastic dependence (the idea taken from the theory of graphical models). The variable at the head of an arrow depends on the variable or constant at the tail of the arrow. For example, $y_1 = 0$ implies $y_2 = 0$ and $y_5 = 0$ (if the individual is dead at the beginning of the first time period, then it remains dead at the beginning of the second time period and cannot have offspring).

We say the variable at the head of an arrow is the *successor* of the variable or constant at the tail of the arrow (the leftmost arrow has the constant 1 at its tail). Conversely, we say the variable or constant at the tail of an arrow is the *predecessor* of the variable at the head of the arrow. For example, y_3 is the successor of y_2 and the predecessor of y_4 and y_7.

The graph represents stochastic dependence much more precisely than we have explained so far. It represents a factorization of the joint distribution of these random variables as a product of conditional distributions

$$f(y_1, \ldots, y_8) = \prod_{j=1}^{8} f(y_j \mid y_{p(j)}), \tag{7}$$

where p is the predecessor function: $y_{p(j)}$ denotes the variable or constant that is the predecessor of y_j (when $y_{p(j)}$ is a constant the conditional distribution of y_j given $y_{p(j)}$ is, in effect, unconditional).

Aster models are a very special case of graphical models in which each node of the graph (representing a random variable or a constant) has at most one predecessor and the graph is *acyclic*, meaning there is no path of arrows joined head to tail that goes around in a loop. Such a graph is called a *forest* in graph theory.

Geyer et al. (2007) formalize this notion as follows.[7] Let J and F be disjoint finite sets and let $p : J \to J \cup F$ be a map. For each $j \in J \cup F$ there is a random variable or a constant denoted y_j. Furthermore, y_j is a random variable when $j \in J$ and a constant when $j \in F$. Let y_J denote the random variables thought of as a single object and y_F the constants thought of as a single object. Then the joint probability distribution of the random variables factorizes as

$$f(y_J \mid y_F) = \prod_{j \in J} f(y_j \mid y_{p(j)}). \tag{8}$$

Again, these conditional distributions are, in effect, unconditional when what is "behind the bar" is constant. Thus on the left-hand side, because y_F is a constant vector, we have, in effect, an unconditional distribution. And on the right-hand side we have some truly conditional distributions, when $p(j) \in J$, and some in-effect unconditional distributions, when $p(j) \in F$.

The factorization (8) goes with a graph that has an arrow $y_{p(j)} \to y_j$ for each $j \in J$. It generalizes (7), which is specific to the graph (6), not only in allowing a general graph but also in allowing more constants

$$\begin{tikzcd}
y_1 \arrow[r] & y_2 \arrow[r] \arrow[d] & y_3 \arrow[r] \arrow[d] & y_4 \arrow[r] \arrow[d] & y_5 \arrow[d] \\
& y_6 & y_7 & y_8 & y_9 \\
y_{10} \arrow[r] & y_{11} \arrow[r] \arrow[d] & y_{12} \arrow[r] \arrow[d] & y_{13} \arrow[d] & \\
& y_{14} & y_{15} & y_{16} & \\
y_{17} \arrow[r] & y_{18} \arrow[r] \arrow[d] & y_{19} \arrow[d] & & \\
& y_{20} & y_{21} & &
\end{tikzcd} \tag{9}$$

In (9) y_1, y_{10}, and y_{17} are constants, because they have no predecessors. Each part of the graph connected by a chain of arrows to one of these

[7] Geyer et al. (2007) actually formalize a more general class of aster models in which groups of variables are allowed to have joint conditional distributions given their predecessors (which are still single variables). This generalization has not yet been implemented in software. We ignore it in this technical report.

constants is called a *maximal connected component* of the graph.[8]

Returning to biology, different maximal connected components represent data on different individuals. Variables within one maximal connected component represent different measurements on a single individual. As shown in (9) it is not necessary that all individuals have the same number of variables connected in the same graphical structure: in (9) there are three individuals (maximal connected components), and each is different from the others.

However, in all applications of aster models that have been done so far all individuals have the same graphical structure and the current version of the aster software (Geyer, 2009b) encourages users to think this way. In all papers about aster models that have been written so far the graphs shown refer to one individual, and it is assumed that all individuals have the same graphical structure (in the full graph, never shown, all maximal connected components are isomorphic subgraphs, and what is shown is just one of these subgraphs).

Aster model graphs can be much more complicated than those shown above. They can be arbitrarily complicated subject to the conditions described above that make the graph a forest (acyclic and each node has at most one predecessor). Here is an example taken from Shaw and Geyer (submitted)

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

We still have not explained the labels on the arrows in (6) and (10). That will have to wait until Section 2.4.

[8] Each maximal connected component of a forest graph is called a *tree* in graph theory (hence the name "forest" for a collection of trees), but we avoid this biological terminology, because aster models have biological applications, which may involve real biological trees.

2.2 Predecessor as Sample Size

The dependence in each conditional distribution of y_j given $y_{p(j)}$ in (8) is of a very specific form: the predecessor plays the role of sample size for the successor. This means that y_j is the sum of $y_{p(j)}$ independent and identically distributed (IID) random variables. By convention, a sum having zero terms is defined to be zero. Hence $y_{p(j)} = 0$ implies $y_j = 0$ with probability one.

The "predecessor is sample size" property has the following consequence. Define[9]

$$\mu_j = E(y_j) \tag{11a}$$

$$\xi_j = E(y_j \mid y_{p(j)} = 1) \tag{11b}$$

In words, μ_j is the unconditional mean value of y_j and ξ_j is the mean value of one of the IID random variables of which y_j is the sum. We call the vector μ with components μ_j the *unconditional mean value parameter vector* and the vector ξ with components ξ_j the *conditional mean value parameter vector*. Because the expectation of a sum is the sum of the expectations,

$$E(y_j \mid y_{p(j)}) = y_{p(j)} \xi_j, \tag{12}$$

and by the iterated expectation theorem

$$\mu_j = \mu_{p(j)} \xi_j. \tag{13}$$

Equation (13) is important. It relates conditional and unconditional mean values in aster models. It is peculiar to models having the "predecessor is sample size" property; it has no analog for general statistical models. Iterating it we obtain

$$\begin{aligned} \mu_j &= \xi_j \mu_{p(j)} \\ &= \xi_j \xi_{p(j)} \mu_{p(p(j))} \\ &= \xi_j \xi_{p(j)} \xi_{p(p(j))} \mu_{p(p(p(j)))} \end{aligned} \tag{14}$$

and so forth. To find the unconditional mean value for y_j we multiply the conditional mean values for all the arrows (each representing a conditional

[9] These definitions differ from those in Geyer et al. (2007) and the documentation for the R package `aster` (Geyer, 2009b). They used τ_j instead of μ_j and defined ξ_j to be (12) rather than (11b). A referee objected to defining ξ_j to be (12) because (12) does not define a parameter, being a function of both data and parameters, and in statistics greek letters are supposed to denote parameters. Our definitions here belatedly admit the referee was right.

distribution) going back to the predecessor of predecessor of predecessor etc. that is a constant. The expectation of a constant is that constant, so when $k = p(p(p(j)))$ or whatever (perhaps more p's) is such that y_k is a constant, then $\mu_k = y_k$, and this allows calculation of all the μ_j as a function of all the ξ_j.

Conversely, solving (13) for ξ_j gives

$$\xi_j = \frac{\mu_j}{\mu_{p(j)}} \tag{15}$$

except in the case, which never occurs in practice, where the denominator on the right-hand side is zero (division by zero is undefined), and this allows calculation of all the ξ_j as a function of all the μ_j.

Thus conditional mean values collectively determine unconditional mean values and vice versa via (14) and (15). For example, for the graph (6) we have

$$\begin{aligned}
\mu_1 &= \xi_1 \\
\mu_2 &= \xi_2\xi_1 \\
\mu_3 &= \xi_3\xi_2\xi_1 \\
\mu_4 &= \xi_4\xi_3\xi_2\xi_1 \\
\mu_5 &= \xi_5\xi_1 \\
\mu_6 &= \xi_6\xi_2\xi_1 \\
\mu_7 &= \xi_7\xi_3\xi_2\xi_1 \\
\mu_8 &= \xi_8\xi_4\xi_3\xi_2\xi_1
\end{aligned}$$

and

$$\begin{aligned}
\xi_1 &= \mu_1 \\
\xi_2 &= \mu_2/\mu_1 \\
\xi_3 &= \mu_3/\mu_2 \\
\xi_4 &= \mu_4/\mu_3 \\
\xi_5 &= \mu_5/\mu_1 \\
\xi_6 &= \mu_6/\mu_2 \\
\xi_7 &= \mu_7/\mu_3 \\
\xi_8 &= \mu_8/\mu_4
\end{aligned}$$

We are not finished with the implications of "predecessor as sample size" but the rest will have to wait until Section 2.4, when we will have developed more theory.

2.3 Exponential Families of Distributions

The likelihood function of a family of distributions is just the probability function thought of as a function of the parameters for one fixed data value rather than as a function of data for one fixed parameter value. Usually its logarithm (log likelihood) is used. For an aster model, the log likelihood is the logarithm of (8)

$$l(\theta) = \sum_{j \in J} \log f_\theta(y_j \mid y_{p(j)}), \tag{16}$$

where we have added explicit dependence on the parameter vector θ to the conditional probability functions. There is one term in the log likelihood for each arrow in the graph.

Because uses of the log likelihood do not depend on constants, one may add or subtract a term that does not depend on the parameters (but may depend on the data) without effect on statistical inference. This is usually done without comment; additive terms in the log likelihood that do not contain parameters are omitted.

A parametric family of distributions is said to be an *exponential family* if the log likelihood has the form

$$l(\theta) = \langle y, \theta \rangle - c(\theta), \tag{17}$$

where y is a vector statistic (a function of the data that does not depend on parameters), θ is a vector parameter, $\langle \cdot\,, \cdot \rangle$ is the bilinear form defined by (5), and c is a function of the parameter vector called the *cumulant function* of the family.

The vector statistic y and vector parameter θ in (17) need not be the originally given data and parameter; one may need to do either a change-of-variable or change-of-parameter (or both) to get the log likelihood into exponential family form (17). To express that y and θ are the special statistic and parameter vectors that appear in (17), they are called the *canonical statistic* and *canonical parameter*.

Exponential families have a simple relationship with IID (which, recall from Section 1, means independent and identically distributed).[10] Suppose, using somewhat confusing notation for this paragraph only, z_1, ..., z_n are IID random vectors (the confusion is that now subscripts denote different random vectors rather than components of one vector) all having the same exponential family distribution for which they are the canonical statistic vectors and for which θ is the canonical parameter vector. Stochastic independence means that the joint probability function is the product of the marginals

$$f_\theta(z_1, \ldots, z_n) = \prod_{i=1}^{n} f_\theta(z_i)$$

from which it follows from the log of a product being the sum of the logs that the log likelihood is

$$l(\theta) = \sum_{i=1}^{n} \log f_\theta(z_i)$$

and because of the exponential family assumption each term of the log likelihood is (17) with y replaced by z_i giving

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \bigl[\langle z_i, \theta \rangle - c(\theta) \bigr] \\ &= \langle z_1 + \cdots + z_n, \theta \rangle - n c(\theta) \end{aligned} \tag{18}$$

Thus we again get an exponential family. The canonical statistic vector for IID data is the sum $z_1 + \cdots + z_n$ of the canonical statistic vectors z_i for the IID constituents and the canonical parameter vector θ is the same for both. Moreover, the cumulant function for the IID family is $nc(\theta)$, just n times the cumulant function for the IID constituents. This relationship is peculiar to exponential families; it has no analog for general statistical models.

2.4 Exponential Families and Aster Models

In an aster model, the conditional distribution associated with each arrow in the graph and each term in (16) is required to be a one-parameter

[10] In aster models, components of the response vector are neither independent nor identically distributed (because of the graphical structure and because they have different parameters), and in LM and GLM they are not identically distributed (because they have different parameters). Nevertheless, the property of exponential family models discussed in the remainder of this section will be useful in constructing aster models (Section 2.4 below).

exponential family: y_j has the distribution of a sum of IID random variables in some exponential family, the number of terms in the sum being $y_{p(j)}$ (this is the "predecessor is sample size" property again). From the relationship between IID and exponential families in the previous section, this means the term of the log likelihood associated with the j-th arrow is

$$y_j \theta_j - y_{p(j)} c_j(\theta_j) \tag{19}$$

because we have a one-parameter family so the bilinear form becomes just $y_j \theta_j$ and what was n in (18) is $y_{p(j)}$ here because of "predecessor is sample size." Also we have subscripts on the parameter and cumulant function because each arrow in the graph may be associated with a different one-parameter exponential family, each having a different parameter θ_j and different cumulant function c_j.

It is a trivial but not entirely obvious fact that (19) works for the case where $y_{p(j)} = 0$ as well as for other cases. When $y_{p(j)} = 0$ the assumption is that y_j is the sum of zero IID random variables, and (as mentioned at the beginning of Section 2.2) a sum with zero terms is zero (by convention). Thus the conditional distribution of y_j given $y_{p(j)} = 0$ is the distribution concentrated at zero, that is, $y_j = 0$ with (conditional) probability one. Since $\log(1) = 0$, the term in the log likelihood for the conditional distribution of y_j given $y_{p(j)}$ should be zero when $y_{p(j)} = 0$, and that is what (19) gives, because $y_{p(j)} = 0$ implies $y_j = 0$. Thus, although $y_{p(j)} = 0$ is a special case, we do not need to deal with it specially, since the general formula (19) works for this special case too.[11]

Thus the log likelihood for the aster model is

$$l(\theta) = \sum_{j \in J} \bigl[y_j \theta_j - y_{p(j)} c_j(\theta_j) \bigr], \tag{20}$$

and this works for all cases, including when some of the y_j or $y_{p(j)}$ are zero.

Now we can explain the labels attached to the arrows in the graphs (6) and (10). We call the labels the *annotation* of the graph. Each denotes the distribution of one of the IID random variables of which y_j is the sum, that is, it denotes the distribution having cumulant function $c_j(\theta_j)$. In (6) and (10) "Ber" indicates a Bernoulli distribution and "Poi" indicates a Poisson distribution. In (10) "0-Poi" indicates a zero-truncated Poisson distribution

[11] This observation explains why what many think of as a missing data problem is not. What is the number of offspring y_j in a time period when the individual is dead so $y_{p(j)} = 0$? In aster models $y_{p(j)} = 0$ implies $y_j = 0$ so y_j is not missing but known to be zero. As we have just seen, the aster log likelihood does the right thing in this case.

(Poisson conditioned on being nonzero). It is perhaps somewhat confusing that the labels do not denote the conditional distribution of y_j given $y_{p(j)}$. For example, if the distribution having cumulant function $c_j(\theta_j)$ is Bernoulli, then the distribution having cumulant function $y_{p(j)} c_j(\theta_j)$ is binomial with sample size $y_{p(j)}$ and canonical parameter θ_j. It might seem more sensible to name these distributions binomial rather than Bernoulli, since the name binomial is more familiar to users, but this would not work in general. For some distributions used in aster models, we have a name for the distribution having cumulant function $c_j(\theta_j)$ but no name for the distribution having cumulant function $y_{p(j)} c_j(\theta_j)$; zero-truncated Poisson is an example.[12]

Since each arrow in the graph is allowed to have a different exponential family distribution, some may be discrete distributions (Bernoulli, Poisson, zero-truncated Poisson) and some may be continuous (normal). However, y_j that are predecessor variables must be nonnegative-integer-valued because they are sample sizes for their successors.[13] Thus only terminal nodes of the graph (those having no successors) can be associated with continuous distributions.

2.5 Reparameterization

Now we come to what is the only completely new idea in Geyer et al. (2007). The log likelihood (20) does not have exponential family form (17). Each term has this form, but their sum does not, because of both y_j and $y_{p(j)}$ being random in some terms. Observe that (20) is linear in the components of the data vector y_j and $y_{p(j)}$ so the joint distribution is an exponential family with y_J as its canonical statistic. To identify the canonical parameters we rewrite the log likelihood collecting terms that multiply the same

[12] Not only do we have no name for the distribution of the sum of IID zero-truncated Poisson random variables, we have no explicit formula for its probability function, but we do not need one because of the relationship between IID and exponential families discussed in Section 2.3. We know the cumulant function will be $y_{p(j)} c_j(\theta_j)$, and that's all we need to know. This ultimately derives from the property that additive terms not involving the parameters can be dropped from log likelihoods. Here we are dropping terms that we do not know how to calculate but do know do not contain the parameters.

[13] Geyer et al. (2007) note that if the successor is infinitely divisible, for example, Poisson or normal, then the predecessor can be positive-real-valued, but this generalization has not been used in aster models and may be ignored.

component of y

$$l(\theta) = \sum_{j \in J} y_j \left[\theta_j - \sum_{\substack{m \in J \\ j = p(m)}} c_m(\theta_m) \right] - \sum_{\substack{m \in J \\ p(m) \in F}} y_{p(m)} c_m(\theta_m). \tag{21}$$

Recall that the y_j are random when $j \in J$ and constant when $j \in F$ so that in (21) the random variables are the y_j and the constants are the $y_{p(m)}$. Thus the canonical parameters for the joint exponential family distribution of the data are

$$\varphi_j = \theta_j - \sum_{\substack{m \in J \\ j = p(m)}} c_m(\theta_m), \qquad j \in J, \tag{22}$$

which are the terms in square brackets in (21) that are the coefficients of the y_j. Geyer et al. (2007) observe that (22) is an invertible change-of-parameter, because (22) can be solved for θ_j giving

$$\theta_j = \varphi_j + \sum_{\substack{m \in J \\ j = p(m)}} c_m(\theta_m), \tag{23}$$

which expresses θ_j as a function of φ_j and θ_m such that $j = p(m)$, that is, components of θ for successors of θ_j. At *terminal* nodes of the graph (those having no successors) we have $\theta_j = \varphi_j$. Now we can solve for θ_j whose successors are terminal nodes. After that we can solve for θ_j whose successors or successors of successors are terminal nodes, and so forth. In fact, we can solve in any order that finds components of θ for successors before predecessors, whatever is computationally convenient. Thus (23), somewhat implicitly, defines the vector θ having components θ_j as a function of the vector φ having components φ_j.

The cumulant function of the joint exponential family with canonical parameter vector φ must be given by

$$c(\varphi) = \sum_{\substack{m \in J \\ p(m) \in F}} y_{p(m)} c_m(\theta_m), \tag{24}$$

the right-hand side being the leftovers, terms in (21) that have not been used in (22). As the subscripts on the summation sign say, all of the $y_{p(m)}$ appearing in this cumulant function have $p(m) \in F$, hence these denote constants not random variables so (24) does define a non-random function.

Using the fact that the cumulant function for the Bernoulli family is

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

and the cumulant function for the Poisson family is

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

we get for the graph (6)

$$\begin{aligned}
\varphi_8 &= \theta_8 \\
\varphi_7 &= \theta_7 \\
\varphi_6 &= \theta_6 \\
\varphi_5 &= \theta_5 \\
\varphi_4 &= \theta_4 - e^{\theta_8} \\
\varphi_3 &= \theta_3 - e^{\theta_7} - \log(1 + e^{\theta_4}) \\
\varphi_2 &= \theta_2 - e^{\theta_6} - \log(1 + e^{\theta_3}) \\
\varphi_1 &= \theta_1 - e^{\theta_5} - \log(1 + e^{\theta_2})
\end{aligned} \tag{25}$$

and

$$\begin{aligned}
\theta_8 &= \varphi_8 \\
\theta_7 &= \varphi_7 \\
\theta_6 &= \varphi_6 \\
\theta_5 &= \varphi_5 \\
\theta_4 &= \varphi_4 + e^{\theta_8} \\
&= \varphi_4 + e^{\varphi_8} \\
\theta_3 &= \varphi_3 + e^{\theta_7} + \log(1 + e^{\theta_4}) \\
&= \varphi_3 + e^{\varphi_7} + \log(1 + \exp(\varphi_4 + e^{\varphi_8})) \\
\theta_2 &= \varphi_2 + e^{\theta_6} + \log(1 + e^{\theta_3}) \\
&= \varphi_2 + e^{\varphi_6} + \log(1 + \exp(\varphi_3 + e^{\varphi_7} + \log(1 + \exp(\varphi_4 + e^{\varphi_8})))) \\
\theta_1 &= \varphi_1 + e^{\theta_5} + \log(1 + e^{\theta_2}) \\
&= \varphi_1 + e^{\varphi_5} + \log(1 + \exp(\varphi_2 + e^{\varphi_6} + \log(1 + \exp(\varphi_3 + e^{\varphi_7} \\
&\qquad + \log(1 + \exp(\varphi_4 + e^{\varphi_8}))))))
\end{aligned} \tag{26}$$

and

$$\begin{aligned}
c(\varphi) &= \log(1 + e^{\theta_1}) \\
&= \log(1 + \exp(\varphi_1 + e^{\varphi_5} + \log(1 + \exp(\varphi_2 + e^{\varphi_6} \\
&\qquad + \log(1 + \exp(\varphi_3 + e^{\varphi_7} + \log(1 + \exp(\varphi_4 + e^{\varphi_8}))))))))
\end{aligned}$$

As one can see, this reparameterization is quite complicated and even more so for larger graphs; if we were to repeat the calculations above for the graph (10) the formula for the cumulant function would not fit on one page. However, this reparameterization is very systematic and easily handled by the computer program that does aster models (Geyer, 2009b). Users do not need to know these formulas; the computer deals with them. We show them only for readers who want a concrete example.

Now we have put the log likelihood of an aster model in the exponential family form

$$l(\varphi) = \langle y, \varphi \rangle - c(\varphi), \tag{27}$$

where the cumulant function is given by (24) and we have written the canonical statistic vector as y rather than y_J and the canonical parameter vector as φ rather than φ_J, which we will do from now on. We see that (27) is just like the log likelihood for a general exponential family (17) except that the parameter has changed from θ to φ.

O. K. We have found that the joint distribution of an aster model is an exponential family and found the canonical parameters and cumulant function of the joint distribution. What was the point of that?

2.6 Means, Variances, Convexity, and Concavity

The cumulant function of an exponential family determines means and variances of the canonical statistic

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

$$\operatorname{var}_\varphi(y) = \nabla^2 c(\varphi) \tag{28b}$$

where $\mu = E_\varphi(y)$ denotes the vector whose components are $\mu_j = E_\varphi(y_j)$ and $\operatorname{var}_\varphi(y)$ denotes the matrix whose components[14] are

$$\operatorname{cov}_\varphi(y_j, y_k) = E_\varphi\{(y_j - \mu_j)(y_k - \mu_k)\} \tag{30}$$

[14] The *variance* of a random variable y_j with mean μ_j is

$$\operatorname{var}_\varphi(y_j) = E_\varphi\{(y_j - \mu_j)^2\}. \tag{29}$$

The *covariance* of random variables y_j and y_k with means μ_j and μ_k is (30). Note that when $j = k$ (30) reduces to (29), so the covariance of a random variable with itself is the variance. For this reason there is no generally accepted terminology for the matrix with components (30). We call it the "variance matrix" because it plays the same role for a random vector y as variance does for a random variable. But other people call it the "covariance matrix" or the "variance-covariance matrix" or (annoyed at the confusion, inventing a new name) the "dispersion matrix."

and where $\nabla c(\varphi)$ denotes the vector whose components are $\mu_j = \partial c(\varphi) / \partial \varphi_j$ and $\nabla^2 c(\varphi)$ denotes the matrix whose components are

$$\partial^2 c(\varphi) / \partial \varphi_j \partial \varphi_k$$

(Barndorff-Nielsen, 1978, Theorem 8.1). The identities (28a) and (28b) have a number of important consequences. First (at least for aster models where the cumulant functions have known expressions), they allow probabilistic calculations about means and variances, which may be difficult to compute (involving sums and integrals) with derivatives, which are easier to compute. Second, they have a number of important theoretical consequences.

A variance matrix is positive semi-definite and positive definite unless the distribution of the random vector is concentrated on a hyperplane, in which case some components can be expressed as linear functions of other components and eliminated. Thus we can always arrange that the distribution of the canonical statistic vector y is non-degenerate (not concentrated on a hyperplane), in which case the variance matrix in (28b) is positive definite, which in turn implies that the cumulant function is strictly convex (Rockafellar and Wets, 2004, Theorem 2.14).

Consider the map h between canonical parameter values φ and mean values μ defined by

$$h(\varphi) = E_\varphi(y) = \nabla c(\varphi).$$

Fix a possible value μ^* of the mean vector $\mu = h(\varphi)$, and define the function

$$q(\varphi) = \langle \mu^*, \varphi \rangle - c(\varphi) \tag{31}$$

which is just the log likelihood (27) with y replaced by μ^*. Applying (28a) and (28b) gives

$$\begin{aligned} \nabla q(\varphi) &= \mu^* - \nabla c(\varphi) \\ &= \mu^* - h(\varphi) \end{aligned} \tag{32}$$

and

$$\nabla^2 q(\varphi) = -\nabla^2 c(\varphi). \tag{33}$$

Equation (33) implies q is a strictly concave function,[15] which in turn implies the maximizer φ^* of q is unique if it exists (Rockafellar and Wets, 2004, Theorem 2.6). From calculus we know that the first derivative is zero at the maximum, so by (32) φ^* is a solution of

$$\mu^* = h(\varphi).$$

[15] A function f is concave if $-f$ is convex, and strictly concave if $-f$ is strictly convex. Here q is strictly concave because $-q$ is strictly convex, because $\nabla^2 c(\varphi)$ is positive definite.

By assumption μ^* is a possible mean value, hence there exists a φ^* such that $\mu^* = h(\varphi^*)$ and the concavity argument shows that there is only one such φ^*. This proves h is an invertible mapping. Each mean value μ corresponds to a unique canonical parameter value φ.

The same argument applied to the one-parameter exponential families corresponding to arrows in the graphical model shows that the mapping between θ_j and ξ_j given by

$$\xi_j = E_{\theta_j}(y_j \mid y_{p(j)} = 1) = c'_j(\theta_j),$$

where the prime denotes differentiation $c'_j(\theta_j) = dc_j(\theta j)/d\theta_j$, is also an invertible mapping. Each conditional mean value ξ_j corresponds to a unique canonical parameter value θ_j of the corresponding one-parameter conditional exponential family.

2.7 Many Parameterizations

An aster model has four parameterizations that we have introduced so far: conditional and unconditional mean value parameter vectors ξ and μ and what Geyer et al. (2007) call conditional and unconditional canonical parameter vectors θ and φ.

In Section 2.2 we established that the change of parameter $\xi \to \mu$ is invertible. In Section 2.5 we established that the change of parameter $\theta \to \varphi$ is invertible. At the end of the preceding section we established that the change of parameter $\varphi \to \mu$ is invertible and that the change of parameter $\theta \to \xi$ is invertible.

Thus all four parameterizations are equally valid. Each is interconvertible with the others. The changes-of-parameter $\xi \to \mu$ and $\mu \to \xi$ are given by explicit formulas (14) and (15). The changes-of-parameter $\theta \to \varphi$ and $\varphi \to \theta$ are given by explicit formulas (22) and (23). The changes-of-parameter $\theta \to \xi$ and $\varphi \to \mu$ are given by explicit formulas

$$\begin{aligned} \xi_j &= c'_j(\theta_j) \\ \mu &= \nabla c(\varphi) \end{aligned}$$

The changes-of-parameter $\xi \to \theta$ and $\mu \to \varphi$ are different. We know that each possible unconditional mean value parameter vector μ^* corresponds to a unique unconditional canonical parameter vector φ^*, but we can only find that φ^* by maximizing the function q defined in (31). This optimization problem may not have a solution given by a formula as a function of μ^*. We know the function mapping $\mu \to \varphi$ exists. We even know it is infinitely differentiable by the inverse function theorem of real analysis. But we know nothing else about it other than that we can calculate its value at any point μ^* by running a computer optimization algorithm to find the φ^* where q achieves its maximum.

Similarly, in general, we can only find the unique θ_j^* such that ξ_j^* is the corresponding conditional mean value parameter for the j-th arrow in the graph by running a computer optimization algorithm to find the θ_j^* where the function

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

achieves its maximum.

Most but not all of the one-parameter families implemented in the current version of the aster model do have expressions as formulas for the mapping $\xi_j \to \theta_j$. For example, for Bernoulli arrows

$$\begin{aligned} \xi_j &= \frac{e^{\theta_j}}{1 + e^{\theta_j}} \\ \theta_j &= \operatorname{logit}(\xi_j) \end{aligned}$$

and for Poisson arrows

$$\begin{aligned} \xi_j &= e^{\theta_j} \\ \theta_j &= \log(\xi_j) \end{aligned}$$

Thus for a graph like (6) having only such arrows we do have (very complicated) expressions as formulas for the map $\mu \to \varphi$ going by the route $\mu \to \xi \to \theta \to \varphi$.

In contrast, for the zero-truncated Poisson distribution

$$\xi_j = \frac{\exp(\theta_j)}{1 - \exp(- \exp(\theta_j))}$$

and this cannot be solved for θ_j to give a formula expressing θ_j as a function of ξ_j. Thus for a graph like (10) having zero-truncated Poisson arrows, there is no formula expressing the mapping $\xi \to \theta$ or the mapping $\mu \to \varphi$.

It must be admitted that the names "conditional canonical parameter" and "unconditional canonical parameter" for the vectors θ and φ are somewhat misleading because the two parameters are not as analogous as the parallel terminology makes them seem. Pedantically, θ is the vector whose components are the canonical parameters of the one-parameter exponential families associated with the arrows of the graph, whereas φ is the canonical parameter vector for the joint exponential family distribution of the aster model. Similarly, the names "conditional mean value parameter" and "unconditional mean value parameter" for the vectors ξ and μ are somewhat misleading because the two parameters are not as analogous as the parallel terminology makes them seem. Pedantically, ξ is the vector whose components ξ_j are the mean value parameters for sample size one of the one-parameter exponential families associated with the arrows of the graph,[16] whereas μ is the mean value parameter vector for the joint exponential family distribution of the aster model.

2.8 Canonical Linear Submodels and Maximum Likelihood

It may be disconcerting to the reader to be told the models described up to this point are uninteresting in themselves because they have too many parameters (one parameter per component of the response vector, no matter which parameterization φ, θ, μ, or ξ is used) to estimate well. As is the case with LM and GLM, only submodels with fewer parameters are interesting. We call the models described up to this point *saturated.* We are interested in *canonical linear submodels* having parameter vector β related to the other parameters by

$$\varphi = M\beta, \tag{34}$$

which is just like (2) except that in GLM there is no requirement that the linear predictor vector be the canonical parameter vector, and here we do require this.

Exponential families have a simple relationship with canonical linear submodels. The log likelihood for the new parameter vector β is (27) with $M\beta$ plugged in for φ giving

$$l(\beta) = \langle y, M\beta \rangle - c(M\beta). \tag{35}$$

Observe that

$$\langle y, M\beta \rangle = \sum_{j \in J} y_j \sum_{k=1}^{p} m_{jk} \beta_k$$

where p is the dimension of β and m_{jk} are the components of the matrix M. If we reverse the order of summation, thinking of the sum over k as the

[16] Recall that

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

one for an inner product and the one over j as the one for a matrix-vector multiplication we get

$$\langle M^T y, \beta \rangle = \sum_{k=1}^{p} \beta_k \sum_{j \in J} m_{jk} y_j$$

where M^T denotes the transpose of M (the matrix with j, k element m_{kj}).

This algebraically trivial transformation has very important statistical consequences. When we rewrite (35) as

$$l(\beta) = \langle M^T y, \beta \rangle - c(M\beta) \tag{36}$$

we see we have exponential family form. The submodel of the saturated model determined by the change-of-parameter $\varphi = M\beta$ is itself an exponential family with canonical statistic vector $M^T y$, canonical parameter vector β, and cumulant function $c_{\text{submodel}}(\beta) = c(M\beta)$.

Thus the log likelihood (36) is a strictly concave function and the MLE $\hat{\beta}$, the maximizer of (36) is unique if it exists (from exponential family theory developed in Section 2.6).[17] The principle of invariance of maximum likelihood gives MLE of all other parameters.[18]

The log likelihood is an extremely complicated function of β, and there is no formula expressing $\hat{\beta}$ as a function of y. We know nothing about the MLE $\hat{\beta}$ other than that we can calculate its value for any data y by running a computer optimization algorithm to find the $\hat{\beta}$ where l achieves its maximum.

We do know something more about the MLE $\hat{\mu}$ of the mean value parameter. Differentiating (36) gives

$$\begin{aligned} \nabla l(\beta) &= M^T y - M^T \nabla c(M\beta) \\ &= M^T y - M^T h(M\beta) \end{aligned}$$

(the M^T in the second term coming from the chain rule for differentiation), where h is the map $\mu = h(\varphi) = \nabla c(\varphi)$ introduced in Section 2.6. The

[17] The MLE need not exist. Conditions for when it does not were first described by Barndorff-Nielsen (1978) and generalized in the author's unpublished thesis (Geyer, 1990). They involve a fascinating interplay of the geometry of convex polyhedra, the geometry of exponential family log likelihood functions, linear programming, and general statistical theory. A recent paper (Geyer, 2009a) gives a complete computational solution for when the MLE does not exist and describes what to do when it does not. That paper is only about GLM, but does point the way toward corresponding theory for aster models.

[18] This principle says that if θ and ψ are two parameters related by an invertible change-of-parameter, say $\psi = g(\theta)$, then the MLE are related by $\hat{\psi} = g(\hat{\theta})$. Thus in the case under discussion $\hat{\varphi} = M\hat{\beta}$, and all other parameters follow, for example, $\hat{\mu} = \nabla c(\hat{\varphi})$.

derivative of the log likelihood is zero at the maximum, hence the MLE satisfies

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

or $M^T y = M^T \hat{\mu}$. This is a general property of exponential families not peculiar to aster models: the MLE of the *mean value parameter* satisfies a property

$$\text{observed} = \text{expected}$$

which is shorthand for observed value of the canonical statistic vector (for the submodel, here $M^T y$) equals expected value of the the canonical statistic vector with the MLE parameter value plugged in (here $M^T \hat{\mu}$).

"Observed = expected" is not useful for calculation, since the matrix M is not full rank, so if we simply seek solutions to the linear system of equations $M^T y = M^T \mu$, there are many solutions, only one of which is the MLE $\hat{\mu}$. The only way to find $\hat{\mu}$ is via the path already described, first find $\hat{\beta}$ by maximizing the log likelihood (36) and then use the transformation $\hat{\mu} = h(M\hat{\beta})$.

Thus "observed = expected" is useful only for interpretation. It is the only simple relationship between data y and the MLE $\hat{\mu}$ of some parameter vector. Note that we have to be careful in using "observed = expected" because it applies only to the canonical statistic vector of *the model being used*, which in this case is not the saturated model with canonical statistic vector y and canonical parameter vector φ but rather the canonical linear submodel with canonical statistic vector $M^T y$ and (submodel) canonical parameter vector β. So in this case "observed = expected" says the observed value of the submodel canonical statistic vector $M^T y$ is equal to its expected value $M^T \mu$ with the MLE $\hat{\mu}$ plugged in for the parameter vector μ.

We now have added a fifth parameter vector β to go with the four θ, φ, ξ, and μ we already had. This plethora of parameterizations is unique to aster models. Because there is no difference between conditional and unconditional in GLM (components of the response vector being independent), conditional and unconditional are the same so $\theta = \varphi$ and $\xi = \mu$. This leaves only three parameter vectors β, φ, and μ for GLM.[19] Because for the normal distribution canonical and mean value parameters are the same $\varphi = \mu$, this leaves only two parameter vectors β and μ for LM.

Multiple parameterizations are considered confusing to students in applied courses about GLM and LM, so they are disguised by changing the

[19] In GLM the linear predictor vector η, need not be the canonical parameter vector φ, but there are still only three parameter vectors β, η, and μ.

language. The components of β are called parameters, but φ is called the linear predictor vector and μ the mean vector, and that they are also parameters is not mentioned. Estimates $\hat{\varphi}$ and $\hat{\mu}$ are called "predicted values" (although there is no sense of future in this "prediction" and they are, in fact, MLE of the corresponding parameters). This is very anti-theoretical. It clouds the distinction between estimates $\hat{\mu}$ and parameters μ, which is the fundamental issue in statistics.[20]

Our view, in teaching aster models, is that this attempt at avoiding confusion is itself a source of confusion. In teaching aster models we say that all five parameter vectors β, θ, φ, ξ, and μ are parameters. That is, they are vectors that specify particular probability models within families of probability models. Each can be converted to any of the others by the changes-of-parameter discussed above. Tests of statistical hypotheses or confidence intervals can be performed for any of these parameters. If one doesn't call μ_j a parameter, then what does one call confidence intervals of the form $\hat{\mu}_j \pm$ margin of error? It is fundamental in statistics that confidence intervals are (interval) estimates of parameters, thus the thing being estimated μ_j must be a parameter. Avoiding calling μ a parameter vector can only lead to confusion about what confidence intervals are.

A fine point about the many parameterizations should be mentioned. For a saturated aster model there are four parameterizations θ, φ, μ, and ξ. All have the same dimension, the dimension n of the response vector. For a

[20] In teaching undergraduate introductory statistics, I try to say "the sample is not the population" at least once a day after these terms are introduced and also say "estimates are not parameters" at least once a day after these terms are introduced. This is related to the three fundamental errors of statistics.

1. The worst error is to not use data (anecdotes are not evidence).
2. The next worst error is to confuse sample and population or estimates and parameters: if a poll says Jones is ahead of Smith by 51 to 49 percent, then the erroneous conclusion is to say it's all over and the election might as well not be held, even though the margin of error of the poll is 3 percent.
3. The next worst error (and a very subtle one) is to fail to interpret statistical confidence intervals correctly. The slogan here is "it's called a 95% confidence interval because it misses 5% of the time." If a poll with 3% margin of error says Jones is ahead of Smith by 54% to 46%, then the erroneous conclusion is to say it's all over and the election might as well not be held. The true unknown population percentages could be Jones 49% and Smith 51% even though the poll says Jones 54% and Smith 46% and the margin of error is 3%. In fact the probability that the poll results differ from the (unknown) truth by more than the margin of error is 5% (that's the subtle point).

All of introductory statistics in one footnote!

canonical linear submodel, the parameter β has dimension $p < n$. The other parameterizations must also have dimension p. The set of allowed values of $\varphi = M\beta$ is a p-dimensional vector subspace of n-dimensional Euclidean space. Since the maps $\varphi \to \mu$, $\varphi \to \theta$, and $\theta \to \xi$ are nonlinear but smooth, the sets of allowed values of μ, θ, and ξ are different p-dimensional manifolds in n-dimensional Euclidean space.[21]

2.9 More on Monotonicity

Geyer et al. (2007) also consider what they call *conditional aster models* defined by $\theta = M\beta$ as opposed to *unconditional aster models* defined by $\varphi = M\beta$. Conditional aster models do not behave as well statistically as unconditional aster models and have been little used. Shaw et al. (2008) did provide one conditional aster model example, but they could have used

[21] This is true of most aster models. The canonical parameter space of most one-parameter exponential families used for aster models is the whole real line $\mathbb{R}$, but may be an open interval. The only example of the latter currently implemented is the negative binomial and truncated negative binomial families which have $-\infty < \theta_j < 0$.

Consider first the implications for *saturated aster models*. The allowed values of the conditional canonical parameter vector θ is usually all of n-dimensional Euclidean space $\mathbb{R}^n$ but may be only an open subset of $\mathbb{R}^n$. The mappings $\theta \to \varphi$, $\varphi \to \mu$, and $\theta \to \xi$ and their inverse mappings are all smooth and hence map open sets to open sets. It can be seen from (22) that when the set of allowed values of θ is all of $\mathbb{R}^n$, then so is the set of allowed values of φ. Typically, the sets of allowed values of the mean value parameter vectors μ and ξ are not all of $\mathbb{R}^n$ but only open subsets thereof.

Now consider the implications for *canonical linear submodels*. If the set of allowed values of φ for the saturated model is all of $\mathbb{R}^n$, the set of allowed β values is all of $\mathbb{R}^p$, and the set of allowed values of φ for the submodel is a p-dimensional vector subspace of $\mathbb{R}^n$. If the set of allowed values of φ for the saturated model is not all of $\mathbb{R}^n$ but only an open subset thereof, the set of allowed values of β is the open subset of $\mathbb{R}^p$ consisting of β such that $\varphi = M\beta$ is allowed. The set of allowed values of φ is a flat p-dimensional submanifold (an open subset of a p-dimensional vector subspace of $\mathbb{R}^n$). The sets of allowed values of θ, μ, and ξ are each curved p-dimensional submanifolds of $\mathbb{R}^n$, and not much more can be said about them.

One of the motivations for modeling canonical parameters linearly, discussed at the end of the introduction, was to avoid restrictions, but we have seen that some aster models (those with negative binomial arrows) have restrictions. What does that do to the method of maximum likelihood? Not much. An exponential family is said to be *regular* if the set of allowed values of the canonical parameter is open. All exponential families used with aster models are regular. For a regular exponential family, the log likelihood goes to minus infinity as φ goes to the boundary of its set of allowed values, and this implies the MLE for β is still a solution of $\nabla l(\beta) = 0$, and the MLE for μ still satisfies the "observed = expected" property $M^T y = M^T \hat{\mu}$ (Barndorff-Nielsen, 1978, Corollary 9.6). Optimization software for finding $\hat{\beta}$ need not handle explicit constraints; it need only avoid steps outside the set of allowed β values by shortening such steps until they don't go outside.

an unconditional aster model instead. Their main reason for choosing a conditional aster model was just to provide an example of one.[22]

The choice between conditional and unconditional aster models is dictated by which kind of mean value parameter is more important. As discussed at the end of Section 1, there is a multivariate monotone relationship between the unconditional canonical and mean value parameter vectors of an aster model. If φ_1 and φ_2 are two distinct possible values of the unconditional canonical parameter vector and μ_1 and μ_2 are the corresponding values of the unconditional mean value parameter vector, then

$$\langle \mu_1 - \mu_2, \varphi_1 - \varphi_2 \rangle > 0.$$

Of course, the same holds true for the one-parameter conditional exponential families associated with each arrow of the graph. If θ_1 and θ_2 are two distinct possible values of the conditional canonical parameter vector and ξ_1 and ξ_2 are the corresponding values of the conditional mean value parameter vector, then

$$\langle \xi_1 - \xi_2, \theta_1 - \theta_2 \rangle > 0,$$

in fact, this holds componentwise

$$(\xi_{1j} - \xi_{2j})(\theta_{1j} - \theta_{2j}) > 0, \qquad j \in J,$$

where the subscripts j denote components. So the question is: which multivariate monotone relationship does one want? The choices are

$$\langle \mu_1 - \mu_2, M\beta_1 - M\beta_2 \rangle > 0 \tag{37a}$$

for unconditional aster models or

$$\langle \xi_1 - \xi_2, M\beta_1 - M\beta_2 \rangle > 0 \tag{37b}$$

for conditional aster models. In the vast majority of applications one wants (37a).

As a concrete example, consider the aster model with graph (6) and suppose the primary question of scientific interest is the relationship between

[22] This was not clearly stated in the paper because of the convention of scientific writing that every procedure be described as the unquestionably correct way to go even though in truth many ways are reasonable. Only by looking at papers by many different authors does one see that there are multiple reasonable ways to proceed in most situations. The first line of *Tao Te Ching*, literally translated as “tao that can tao is not constant tao,” can be more loosely translated as “methodology that is useful cannot be unchanging.” Science follows tao, but conventional scientific writing is about as untaoistic as it is possible to be.

Darwinian fitness, defined to be the total number of offspring produced, $y_5 + y_6 + y_7 + y_8$ in (6), and a quantitative covariate x, which is some phenotypic characteristic of the organism.[23] Also we should distinguish between observed fitness $y_5+y_6+y_7+y_8$ and expected fitness $\mu_5+\mu_6+\mu_7+\mu_8$. The primary question of scientific interest is more precisely stated as what is expected fitness expressed as a function of x? This function is called the *fitness landscape* or *adaptive landscape*.

We have been a bit sloppy in writing expected fitness as $\mu_5 + \mu_6 + \mu_7 + \mu_8$ forgetting that (6) is only the graph for one individual. Every other individual has an isomorphic graph (same shape but different numbers on the variables). To be precise about this, we introduce two discrete variables u_j and v_j which describe where in the graph a variable y_j is. First, v_j has values 1, …, 8 and says which node in (6) y_j corresponds to. Second, u_j has values 0 or 1 and says which kind of node in (6) y_j corresponds to: $u_j = 0$ for survival nodes ($v_j = 1$, 2, 3, 4) and $u_j = 1$ for offspring count nodes ($v_j = 5$, 6, 7, 8). Now we write our canonical linear model equations concretely as

$$\varphi_j = \beta_{v_j} + u_j(\beta_9 x_j + \beta_{10} x_j^2) \tag{38}$$

(abstractly, they are $\varphi = M\beta$), where x_j is the quantitative covariate value for the individual corresponding to the maximal connected component containing y_j. This requires β to be a vector of length 10 (since v_j goes from 1 to 8, the term β_{v_j} goes from β_1 to β_8).

Why (38)? Let us follow the logic of monotonicity and see the consequences. We play a trick commonly used with all forms of regression models (LM, GLM, and aster): imagine an arbitrary individual (not necessarily in the given data) having the same true unknown parameter vector β but arbitrary covariate value x. We can take this individual to have graph (6) in which case the variable indices are 1 to 8 and expected fitness is

$$\psi(x) = \sum_{j=5}^{8} \mu_j(x), \tag{39}$$

where we have written the mean value parameters as a function of the covariate $\mu_j(x)$. The function $\psi(x)$ is the fitness landscape (for this aster model); it gives the fitness of one (hypothetical) individual as a function of that individual's phenotypic covariate value x.

[23] Pedantically, one should say that $y_5 + y_6 + y_7 + y_8$ is only the best surrogate of Darwinian fitness available in this experiment or observational study, but we will not bother with that.

Even though we have an explicit formula for $\mu_j(x)$ as a function of x, it is so messy as to be useless for reasoning about the fitness landscape. So we use the multivariate monotonicity property (37a). For any two distinct values x and x^* of the covariate

$$
\begin{aligned}
0 &< \sum_{j=1}^{8} [\mu_j(x) - \mu_j(x^*)][\varphi_j(x) - \varphi_j(x^*)] \\
&= \sum_{j=5}^{8} [\mu_j(x) - \mu_j(x^*)][(\beta_9 x + \beta_{10} x^2) - (\beta_9 x^* + \beta_{10} (x^*)^2)] \\
&= [(\beta_9 x + \beta_{10} x^2) - (\beta_9 x^* + \beta_{10} (x^*)^2)] \sum_{j=5}^{8} [\mu_j(x) - \mu_j(x^*)] \\
&= [(\beta_9 x + \beta_{10} x^2) - (\beta_9 x^* + \beta_{10} (x^*)^2)][\psi(x) - \psi(x^*)]
\end{aligned}
$$

where the first equality comes from plugging in (38) and realizing that the β_{v_j} terms cancel because they do not involve x and that the $j = 1$, 2, 3, and 4 terms are zero because for them $u_j = 0$. This implies that both factors in the final expression must have the same sign, hence

$$
\beta_9 x + \beta_{10} x^2 < \beta_9 x^* + \beta_{10} (x^*)^2 \quad \text{if and only if} \quad \psi(x) < \psi(x^*) \tag{40}
$$

We express (40) in words by saying that fitness on the canonical parameter scale $\beta_9 x + \beta_{10} x^2$ is a monotone function of expected fitness $\psi(x)$ and vice versa. Although we have only given the argument for one concrete case, the same argument works when $\beta_9 x + \beta_{10} x^2$ is replaced by any function of any number of covariates (Shaw and Geyer, submitted, appendix on monotonicity).

This monotonicity relationship is the key to constructing scientifically interpretable aster models. We must model on the canonical parameter scale for many technical reasons: to avoid explicit restrictions on the parameters, to have concave log likelihood for which optimization is simple, to have the properties discussed in Sections 2.10 and 2.12 below. For scientific interpretability, however, we must be able to interpret our models on the mean value parameter scale. If there were no understandable relationship between the two scales — just an arbitrary function with no simple properties — then there would be no way to interpret our models.

There is a simple consequence of multivariate monotonicity that is sometimes presented in its place (Geyer et al., 2007, discussion section): if one canonical parameter is increased (the rest being held fixed), then the corresponding mean value parameter is increased (the other mean value parameters change too but can go any which way). This statement does not contain the full content of multivariate monotonicity and cannot replace it for mathematical derivations, but it can give some idea of its import without going into too much mathematics. For canonical linear submodels, this becomes: if β_k is increased (other betas being held fixed), then the expectation of the k-th submodel canonical statistic, the k-th component of $M^T y$ increases (expectations of other components of $M^T y$ change too but can go any which way). Thus even naive users should know about the submodel canonical statistic vector $M^T y$.

2.10 Sufficiency

Fisher (1922) introduced not only the method of maximum likelihood but also the notion of sufficiency. A statistic (function of the data that does not involve parameters) is *sufficient* if the conditional distribution of the data given the statistic does not depend on the parameter, that is, for data y a statistic $t(y)$ is sufficient if the conditional distribution

$$f_\theta\bigl(y \mid t(y)\bigr)$$

does not actually depend on the parameter θ. This means that the distribution of y given $t(y)$ tells nothing about the parameter. Hence any estimate of the parameter should only involve the data y through the sufficient statistic $t(y)$. This is the principle of sufficiency: any sensible estimate of the parameter must be a function of the sufficient statistic.

The definition of sufficiency is hard to apply, but there is a much simpler criterion that is much easier. If the log likelihood is a function of the data only through a statistic $t(y)$, then that statistic is sufficient. An exponential family only involves the data through the canonical statistic, hence the canonical statistic is always sufficient. Also the method of maximum likelihood automatically obeys the sufficiency principle, since the maximizer of the likelihood only involves the data through the log likelihood which in turn only involves the data through the sufficient statistic. Thus sometimes the canonical statistic of an exponential family is called the canonical *sufficient* statistic, although the “sufficient” is redundant (every canonical statistic is sufficient).

Sufficiency may seem a pedantic distinction, but before Fisher (1922), no one had identified this notion nor were there any methods of estimation in common use that obeyed the principle of sufficiency. Hence every methodology introduced before 1922 was either (in hindsight) wrong or was a special case of Fisher's methods (in particular, the method of least squares is a special case of the method of maximum likelihood if the error distribution is assumed to be homoscedastic normal).

The principle of sufficiency applied to aster models motivates unconditional canonical linear models ($\varphi = M\beta$). In such models the (submodel) canonical statistic vector $t(y) = M^T y$ is sufficient. Your humble author is of the opinion that in elementary statistics teaching too much is made of the linear change-of-parameter

$$\varphi = M\beta \tag{41a}$$

and too little is made of the linear change-of-variable

$$t(y) = M^T y \tag{41b}$$

from data to (submodel) sufficient statistic.[24] In fact, the latter is not mentioned at all in most courses about LM and GLM. Clearly each determines the other, because each determines and is determined by the model matrix M. The question is: which one does one choose to motivate the choice of M? Most courses about LM and GLM concentrate on (41a), teaching students how to interpret this equation.[25] Since (41b) is never mentioned in

[24] In the example in the preceding section, the concrete expression of (41a) is (38). The concrete expression of (41b) was not derived, and we do so now. The bilinear form in the log likelihood is

$$\langle y, M\beta \rangle = \sum_{j \in J} y_j \left[\beta_{v_j} + u_j(\beta_9 x_j + \beta_{10} x_j^2) \right]$$

and the coefficient of each beta is the corresponding canonical statistic. Thus the canonical statistics are

$$\begin{aligned} t_k(y) &= \sum_{\substack{j \in J \\ v_j = k}} y_j, \qquad k = 1, \ldots, 8 \\ t_9(y) &= \sum_{j \in J} y_j u_j x_j \\ t_{10}(y) &= \sum_{j \in J} y_j u_j x_j^2 \end{aligned}$$

These ten statistics are the components of the sufficient statistic vector. "Observed = expected" says the method of maximum likelihood matches the observed and expected values of these ten statistics. The sufficiency principle says all inference should depend on the data only through these ten quantities.

[25] More precisely, they teach students how to interpret concrete instances of this equation, (38) for example. The abstraction $\varphi = M\beta$ may never be mentioned in lower level courses in which students are not expected to understand matrices.

most courses, students do not learn to use it to motivate M.

This is wrongheaded. Data are much more concrete than parameters. Statistical models are very abstract, and parameters even more so, since parameters are a mere index for models (parameters can be changed without changing the model, the family of probability distributions indexed). We sometimes express this by the slogan

> Parameters are meaningless quantities. Only probabilities and expectations have scientific meaning.

Of course, some parameters are expectations. In aster models ξ and μ have scientific interpretation tied directly to data; they are expectations (conditional or unconditional) of components of the data (more precisely, the components of ξ have the relationship to conditional expectations recalled in footnote 16). Other parameters have no such meaning. In aster models β, θ, and φ are only weakly tied to data, through the monotonicity properties discussed at the end of the preceding section.

These considerations say that (41b) has a much more direct connection to data and hence to reality than (41a). It is a triumph of abstraction that (41a) is preferred to (41b) in elementary statistics teaching.

Aster models are more directly tied to reality when (41b) is used to motivate the choice of model (by choice of model matrix M): the principle is to choose M so that the components of the (submodel) canonical statistic vector $M^T y$ have sensible scientific interpretation. Then the aster model has sensible scientific interpretation. The "observed = expected" property assures that the MLE of the (submodel) mean value parameter vector $M^T \hat{\mu}$ also has sensible scientific interpretation, since it makes $M^T y = M^T \hat{\mu}$.

We call the map t defined by (41b) the *sufficient dimension reduction* map. It reduces the dimension of the data vector y to a linear function $M^T y$ of smaller dimension (the dimension of the parameter β) *without losing information* because $M^T y$ is sufficient.

Sufficiency is a very abstract property. It is a property of a statistic relative to a statistical model. Throwing away the all of the data except for the sufficient statistic vector loses no information *about that particular statistical model.* The point of sufficiency is purely philosophical. And it works both ways. Given a model, one wants to know the sufficient statistics. Given a set of scientifically important statistics, one wants to choose a model for which those statistics are sufficient. So in aster models, we want to choose canonical linear submodels in which the sufficient statistic vector $M^T y$ has components that are important scientifically interpretable quantities.

For a specific instance of this reasoning, let us return to the estimation of fitness landscapes introduced in the preceding section. In footnote 24 we derived the submodel canonical sufficient vector. An earlier methodology for estimating fitness landscapes, or, more precisely, the best quadratic approximation (BQA) to fitness landscapes, was introduced by Lande and Arnold (1983). This method uses LM rather than aster to estimate the fitness landscape. If fitness satisfied the assumptions for LM, that the conditional distribution for fitness given phenotypic trait covariate values is homoscedastic normal, then the sufficiency argument applied to that model would give the same sufficient dimension reduction to the same ten sufficient statistics $t_k(y)$, $k = 1, \ldots, 10$ that are the sufficient statistics for the aster model. This is because the two models agree about what fitness is (as a function of the data y), because they both use a quadratic model, and because both are exponential family models. So aster modelers agree completely with Lande and Arnold (1983) about the sufficient dimension reduction map. The only disagreement is about the statistical model: the LM used by Lande and Arnold (1983) is often grossly wrong for fitness landscapes whereas aster models may have no apparent wrongness.

2.11 Efficiency

Fisher (1922) introduced not only the method of maximum likelihood and the notion of sufficiency but also the notion of efficiency. The efficiency of an estimator $\hat{\varphi}$ of a scalar parameter φ is the expected squared error

$$E\{(\hat{\varphi} - \varphi)^2\}. \tag{42}$$

The smaller the expected squared error, the more efficient the estimator. Here $\hat{\varphi}$ is a function of the data, hence a random variable, hence it has a distribution,[26] and it is with respect to this distribution that the expectation (42) is defined. In contrast, the parameter φ is considered nonrandom, a property of the probabilistic law of nature that governs the generation of the data. The estimator $\hat{\varphi}$, being random, would be different if the experiment or observational study were repeated, so the amount of error $\hat{\varphi} - \varphi$ is a

[26] In all but the simplest cases we do not have an explicit formula for the distribution, but we know it exists because every random vector has a distribution. This distribution, called the *sampling distribution* of the estimator $\hat{\varphi}$ is a sticking point many people have with statistics. You have only one data set, and you have only one estimator $\hat{\varphi}$ calculated from it. It takes an act of imagination to visualize the distribution of possible other values of the estimator that would result from imaginary replications of the experiment or observational study and to think about properties of this distribution, such as efficiency.

random quantity, and we cannot say how large it would be, only how large it is on average. Hence the definition of efficiency. A more efficient estimator provides more accurate estimates, on average.

When the parameter is a vector φ, then so is the estimator $\hat{\varphi}$, and the definition of efficiency becomes more complicated, it is

$$E\{(\hat{\varphi} - \varphi)(\hat{\varphi} - \varphi)^T\}, \tag{43}$$

which denotes the matrix whose j, k component is $E\{(\hat{\varphi}_j - \varphi_j)(\hat{\varphi}_k - \varphi_k)\}$. Having a matrix measure of efficiency is more difficult to understand. If two estimators have efficiency matrices V_1 and V_2, then we say the first is more efficient than the second if $V_2 - V_1$ is a positive semi-definite matrix. It may be that neither $V_2 - V_1$ nor $V_1 - V_2$ is positive semi-definite, and this means that some but not all components of one estimator are more efficient estimates of the corresponding parameter than those of the other estimator (neither estimator is better for all components).

The exact efficiency of an estimator is usually impossible to calculate when the estimator is a complicated function of the data. Thus we make do with the so-called "asymptotic" distribution, which refers to the limit as the number of individuals in the data set goes to infinity. Under certain conditions the asymptotic distribution of the MLE $\hat{\varphi}$ is known to be normally distributed, and this (multivariate) normal distribution is centered at the true unknown parameter vector φ, and has variance matrix that is the inverse of the matrix

$$I(\varphi) = E_\varphi\{-\nabla^2 l(\varphi)\},$$

which is called the *Fisher information matrix*, because it too was introduced by Fisher (1922). For a general exponential family of distributions, and in particular for a saturated aster model, this simplifies to

$$I(\varphi) = \nabla^2 c(\varphi).$$

For a canonical linear submodel with parameterization $\varphi = M\beta$, this simplifies to

$$I(\beta) = M^T \nabla^2 c(M\beta) M \tag{44}$$

(Geyer et al., 2007, Section 3.2). So for exponential families of distributions in general and aster models in particular we know that the asymptotic distribution of the MLE $\hat{\beta}$ is normal with mean β (the true unknown parameter value) and variance $I(\beta)^{-1}$ (inverse Fisher information). The *asymptotic efficiency* is this asymptotic variance of the MLE $I(\beta)^{-1}$. The asymptotic normal distribution of other parameters can be derived via the delta method Geyer et al. (2007, Section 3.3).

It is a remarkable fact about statistical estimation that maximum likelihood estimates are as efficient as it is possible for an estimator to be. This fact was also first surmised by Fisher (1922). It is also referred to as the Cramér-Rao lower bound.[27]

This optimality property of maximum likelihood estimators (they are most efficient) is not quite correct without conditions. A large part of the mathematical statistics in the twentieth century was devoted to proving asymptotic normality of MLE under weaker and weaker conditions and to weakening the conditions under which MLE were efficient. Van der Vaart (2000, Sections 5.3, 5.5, 8.5, and 8.6) gives the current state of the theory (but is not easy to read).

Asymptotic efficiency is a very abstract property. It is a property of an estimator (such as the MLE) relative to all other estimators. It says that an asymptotically efficient estimator (such as the MLE) is at least as good as any other estimator for sufficiently large sample sizes (as n goes to infinity) when squared error (42) or (43) is the criterion. And why that criterion? Partly mathematical convenience — it is easy to work with and leads to interesting results — and partly the connection with asymptotic normality — for the multivariate normal distribution, the variance matrix is the natural criterion of error because it is a parameter of the distribution. The point of asymptotic efficiency is purely philosophical. It justifies using estimators (such as the MLE) that are efficient.

2.12 Entropy

Boltzmann made the connection between entropy and probability. He considered this his most important discovery, having $S = k \log W$ engraved on his tombstone (S is entropy, W is probability, and k is a physical constant, now known as Boltzmann's constant).

Exponential families have a simple relationship with entropy (Jaynes, 1978). The entropy of one probability distribution with probability function $f(y)$ with respect to another with probability function $g(y)$ is

$$E_f \left\{ \log \frac{f(y)}{g(y)} \right\} = \sum_y f(y) \log \frac{f(y)}{g(y)}, \tag{45}$$

[27] Rao was a student of Fisher. Cramér (1946) was the first recognizably modern textbook of theoretical statistics and included perhaps the first truly rigorous proof of the asymptotic normality of the MLE.

where the sum runs over allowed values of y (we consider here only the case where y is discrete, but the argument when some components of y are continuous is similar). Consider the following somewhat unmotivated problem (it is interesting that this setup leads to an interesting answer but the setup is not interesting in itself). We wish to maximize (45) subject to the constraints

$$\tau_k = E_f\{t_k(y)\} = \sum_y f(y) t_k(y), \qquad k \in K \tag{46}$$

where the τ_k are arbitrary real numbers and the t_k are arbitrary real-valued functions, and where f is allowed to vary over all possible real valued functions that specify probability distributions, hence satisfy the additional constraints $f(y) \geq 0$ for all y and $\sum_y f(y) = 1$. The solution to this problem has the following form

$$f(y) = g(y) \exp\left(\sum_{k \in K} \beta_k t_k(y) - c(\beta) \right),$$

where the β_k are Lagrange multipliers (real numbers chosen to make the constraints (46) hold, β is the vector having components β_k, and $c(\beta)$ is a function determined by the requirement that the probabilities sum to one. Introduce the vector $t(y)$ having components $t_k(y)$ and this solution can be rewritten

$$f(y) = g(y) e^{\langle t(y), \beta \rangle - c(\beta)}$$

a family of distributions (if we consider β a parameter vector) which is an exponential family because the log likelihood is

$$\log f(y) = \langle t(y), \beta \rangle - c(\beta)$$

(we can drop the term $\log g(y)$ from the log likelihood because it does not contain the parameter). The canonical statistic vector is $t(y)$ and the canonical parameter vector is β.

This discussion is a little odd. We started with the idea that the constraint equations (46) fixed the values of certain expectations, but now we are considering β a parameter, which gives different distributions $f(y)$ as β varies over its allowed values. We know from exponential family theory that the relationship between the canonical parameter vector β and the mean value parameter vector

$$\tau = E_\beta\{t(y)\}$$

is invertible (each vector β corresponds to a distinct vector τ). Thus there is at most one distribution in the family that solves the constraints (46). But having gotten to this point we now drop the idea that we were attempting to fix the mean vector τ at just one value. If we allow τ to vary over its allowed values we get the whole exponential family because (to repeat what was just said) the map $\beta \to \tau$ is invertible so either β or τ can serve as a parameter indexing the family.

Thus we have arrived at a conclusion that can be somewhat sloppily stated as saying that exponential families arise by maximizing entropy subject to fixing the means of a set of quantities, which turn out to be the canonical statistics of the exponential family that arises from entropy maximization. If one chooses a set of statistics, then the family of probability distributions that maximizes entropy is the exponential family having these chosen statistics as its canonical statistics.

We now emphasize that the distribution g which has been fixed throughout the argument plays a role too. Each different choice of g gives a different exponential family. In the context of aster models we let g be the probability function of any distribution in the saturated model, which is given by biological considerations that lead to a particular graphical model. The maximum entropy argument applies to submodels. If we choose statistics $t(y)$ of the form $t(y) = M^T y$, in which case the vector of means being fixed is $\tau = M^T \mu$, then we know this does not take us outside the original saturated model but rather leads to a canonical linear submodel. The maximum entropy argument then justifies these submodels saying that they are the models that maximize entropy subject to fixing the values of the means of the submodel canonical statistic vector $M^T y$.

Even more sloppily, we can say that the maximum entropy argument says that if we choose the submodel canonical statistic vector $M^T y$ to be scientifically interpretable, then the corresponding canonical linear submodel will also be scientifically interpretable. It is the model that maximizes entropy (randomness) subject to controlling the mean values of $M^T y$.

Maximum entropy is a very abstract property. It is a property of a family of statistical models (each a family of probability distributions) relative to a statistic vector. It says that the one family in the family of families that is an exponential family having the chosen statistic vector as its canonical statistic vector maximizes entropy (within the family of families).

So why maximize entropy? All physical processes maximize entropy. Thus we expect maximum entropy models with canonical statistics that are scientifically important quantities to model the biological situation well. The point of maximum entropy is purely philosophical. It justifies using exponential family statistical models and focuses attention on their submodels that are themselves exponential families (the canonical linear submodels).

3 Discussion

> All models are wrong, but some are useful.
>
> — George Box

The important philosophical point about all of this is how it all hangs together. Many different bits of mathematical statistics play a role. All are part of aster models being a practical solution to biological problems.

Start with the fact that aster models are statistical models. One does not need a statistical model to do some simple forms of statistical inference, but when one has a model — a joint probability distribution for all the data expressed as a function of parameters of the model — the full panoply of statistical methodology is available. If one has a statistical model, then *any question whatsoever* that can be phrased in terms of probabilities and expectations can be, in principle, answered by calculations based on the model. In practice, exact calculation may not be possible but approximate calculation based on simulation of the model is always possible.

In life history analysis, as recounted in the introduction of Shaw et al. (2008), before aster models there was no way to construct a statistical model for complete life history data (at least no ideas were proposed before aster models). In the absence of statistical models for complete data, many approaches were devised. One approach was to separately analyze parts of the data called "components of fitness" to which available models could be applied. But there was no way to combine these analyses to make inference about fitness itself. Other approaches analyzed fitness itself without a valid model. Both demographic approaches using Leslie matrices and Lande-Arnold analysis (see Shaw et al., 2008, for references) tacitly assume a normal distribution for fitness, which is often grossly and obviously wrong. All these approaches were recognized to be unsatisfactory and criticized in the literature reviewed by Shaw et al. (2008), but they were used despite their known problems because nothing else was available before aster came along.

The title of Shaw et al. (2008) is "*Unifying* life history analysis . . ." (our emphasis). Previously existing approaches to life history analysis could be applied, even unsatisfactorily, to only a small fraction of the analyses to which aster models can be applied. None could do more than one of the three examples of aster analysis in Shaw et al. (2008), and none could do as well as aster.

So the first virtue of aster models is to provide a unified theory of defensible statistical models for complete life history data. The description "defensible" is used advisedly. All statistical models are wrong. None can incorporate all the biology relevant to the data. But a model that is not obviously grossly wrong — one that passes statistical tests of goodness of fit — can be useful, as the Box aphorism says.

Having a defensible statistical model may be enough for sensible statistical inference in simple situations, but aster models are too complicated for that. Having a joint statistical model for the complete data does not, by itself, avoid the "components of fitness" problem. One must still use a model complicated enough so that all components of fitness are modeled well. In an aster model, such a model will typically have too many parameters to estimate well if the original (conditional canonical) parameterization is used. And even if there are not too many parameters, those parameters have no simple relationship with unconditional means, so it is difficult to choose submodels and difficult to interpret any deemed to fit.

It is the change-of-parameter given by equations (22) from conditional to unconditional canonical parameters that is the second virtue of aster models. Because of the monotonicity, sufficiency, and maximum entropy properties of exponential families, these parameters, when modeled linearly (34) lead to models that have canonical sufficient statistics that are scientifically interpretable and and have mean value parameters connected by multivariate monotone relationship (Section 2.9) with the canonical parameters that are modeled. It is this change-of-parameter that allows complicated data to be adequately modeled with few enough parameters to be estimated well, and it is this change-of-parameter that allows scientific interpretation of these models.

The change-of-parameter from conditional to unconditional canonical parameters has a very abstract origin. For each concrete instance of an aster model — a saturated model determined by a graph with annotation and a submodel of it determined by a model matrix — this change-of-parameter could be given by explicit formulas like (25), which whether explicitly written out or not are "understood" and used by the aster software, but no one would invent such a change-of-parameter except from theoretical considerations. The change of parameter (22) arises from the recognition that an aster model is an exponential family and from the process of finding its canonical parameter vector. Without the notion of an abstract exponential family, this change-of-parameter and all the properties of aster models that flow from it would never occur to anyone.

The change-of-parameter from conditional to unconditional canonical parameters causes some conceptual difficulties to those trained to think about components of fitness. The unconditional aster model directly addresses overall fitness but does not directly address the separate components (the conditional aster model does that). When one wants to have one's cake and eat it too (study both overall fitness and the separate components), unconditional aster models can do the job, but only in an indirect way that is difficult to wrap one's mind around. Example 2 of Shaw et al. (2008) presents data on fitness of *Echinacea angustifolia*. The experiment was done in two parts. Plants started life indoors in a growth chamber. Those that survived this initial period were replanted outdoors in an experimental field plot. Of interest were differences in fitness for different "cross types" representing different amounts of inbreeding. The unconditional aster model found there were statistically significant differences among cross types.

So far so good. The unconditional aster model made the analysis of overall fitness simple. Confusion creeps in when the question is asked whether the fitness differences are due to differential survival in the growth chamber or differential survival and growth (reproduction was not measured in these data) in the field plot. If one adds additional terms to the aster model modeling fitness differences among cross types both in the growth chamber and in the field plot, the growth chamber terms are not statistically significant ($P = 0.34$). So that says differential survival in the growth chamber doesn't affect fitness? No! In an unconditional aster model terms for later components of fitness propagate back through the equations to earlier ones. This is seen in (26) where the expression for θ_1 contains all components of φ, the expression for θ_2 contains all components of φ except φ_1 and φ_5, and so forth. Thus the comparison between submodels with and without explicit growth chamber terms says that differential survival in the growth chamber has no effect on fitness over and above what the maximum entropy principle automatically builds into the submodel that has only terms for overall fitness. Confidence intervals for survival in the growth chamber (Shaw et al., 2008, Figure 2A) show clear differences among cross types, whether or not the model has terms for fitness differences among cross types in the growth chamber.

In short, trying to have one's cake and eat it too — trying to analyze both overall fitness and separate components of fitness — causes confusion. Biologists know that "components of fitness" only act together not separately, so analyzing the effects of the components of fitness separately must be problematic. Aster models treat this problematicity correctly: only overall fitness is modeled simply. Once one has found a defensible aster model that passes tests for goodness of fit (in Example 2 of Shaw et al., 2008, the model with only terms for overall fitness differences among cross types), one can look at parameter estimates for various conditional or unconditional mean values to get some idea of the contributions of the separate components of fitness but one cannot completely disentangle them.

Aster models are very complicated compared to LM and GLM and very simple compared to other graphical models. Why are aster models defined they way they are? Why not more simple or more complicated? Aster models are complicated enough to encompass most but not all life history data. The "predecessor is sample size" property is the key to aster model theory. Without it, the log likelihood (20) would not be linear in y (both the y_j and $y_{p(j)}$ appearing linearly), the joint distribution of y would not be an exponential family, there would be no unconditional canonical parameterization, and there would be no multivariate monotone relationship of parameters with unconditional mean values. Nor would there be strictly concave log likelihood, which makes MLE unique and easy to find by computer optimization algorithms. So the "predecessor is sample size" property does a lot.

But "predecessor is sample size" is very restrictive. Many forms of dependence that can be imagined for life history data do not fit into this paradigm. So why stop where aster models stop? The strong theoretical properties of aster models mentioned in the preceding paragraph do many things but have two main effects: they allow *simple scientific interpretation* and *efficient computer implementation*. The simplicity of interpretation comes from the multivariate monotonicity relationship of canonical and mean value parameter values and the sufficiency and maximum entropy properties. The efficiency of implementation comes from simple (for the computer) expressions for the changes-of-parameter $\theta \leftrightarrow \varphi$, $\xi \leftrightarrow \mu$, $\theta \to \xi$, and $\varphi \to \mu$, through concavity properties of log likelihood that guarantee uniqueness of MLE and allow MLE and changes of parameter $\xi \to \theta$ and $\mu \to \varphi$ to be computed by straightforward computer optimization, and through the relation (44) that gives Fisher information as a simple (for the computer) expression and through it the asymptotic distribution of the MLE and resulting hypothesis tests and confidence intervals. Any generalization of aster models other than those already considered in Geyer et al. (2007) and mentioned in various footnotes (5, 7, 13, 17, 21) in this technical report would lose (as far as we can see) at least one of the the properties that give aster models their simple scientific interpretation and efficient computer implementation.

- Having defensible statistical models allows likelihood inference.

- The efficiency property of maximum likelihood estimates guarantees they are good. Without the efficiency of MLE, one would have no idea how to proceed with complicated life history data.

- The relationship between conditional and unconditional mean value parameter vectors (Section 2.2 above) is peculiar to "predecessor is sample size" models.

- The multivariate monotone relationship between canonical and mean value parameter vectors (Sections 1 and 2.9 above) is peculiar to exponential family models.

- Cumulant functions (Sections 2.3, 2.4, 2.5, 2.6, 2.8, and 2.11) are peculiar to exponential family models. They allow means, variances, and covariances of components of the canonical statistic vector, hence also the Fisher information matrix and the asymptotic distribution of the MLE to be calculated from derivatives of cumulant functions, which are simple for the computer to calculate.

- The relationship between IID and exponential families (Section 2.3) is peculiar to exponential family models. This makes the "predecessor is sample size" property lead to exponential family joint distributions (Section 2.4).

- The sufficient dimension reduction argument (Section 2.10 is unique to exponential family models.

- The maximum entropy argument (Section 2.12) is unique to exponential family models.

- That "structural zeros" (footnote 11) are handled automatically and correctly is unique to maximum likelihood estimation based on a statistical model.

All of these points are involved in making aster models work the way they do.

4 Statistical Models and Their Interpretation

We close with a "post-discussion discussion" of an important point that often arises in the context of aster models but really has nothing to do with them, since it applies to all statistical models. This issue is reification of parameters.

Probability models are already very abstract. The nature of probability and expectation has been argued about by scientists and philosophers for hundreds of years and no agreement has been reached (Hájek, 2007). Mathematicians cut through the confusion by axiomatizing probability and expectation, giving them purely mathematical definitions cut loose from debates about what features of the real world might or might not correspond to them. Kolmogorov (1933) gave a system of axioms for probability theory and since then all research-level probability theory been based on them. Their level of abstraction is very high, requiring the notion of abstract integration over so-called measurable spaces (abstract sets equipped with sigma-algebras). It is so high that no one attempts to teach this theory to undergraduates or master's students. This abstract integration theory is only taught in Ph. D. level courses in probability, theoretical statistics, and real and functional analysis (from textbooks like Fristedt and Gray, 1996 and Rudin, 1986). Consequently, pre-1933 notions of probability and expectation are still taught in all universities in courses designed for undergraduates and master's students (from textbooks like Casella and Berger, 2001). Most scientists who use probability and statistics are exposed to them in such courses. Call this theory "master's level probability theory" to give it a name (all of the probability theory in this technical report is "master's level").

Probability models in "master's level probability theory" are still very abstract (if not quite as abstract as in research level probability theory). Probability and expectation are still given purely mathematical definitions cut loose from debates about what features of the real world might or might not correspond to them. Discrete probability models are defined by giving a probability mass function (PMF), which is a real-valued function whose domain is an arbitrary (abstract) set S and satisfies two properties

$$f(y) \geq 0, \qquad y \in S \tag{47a}$$

(nonnegativity) and

$$\sum_{y \in S} f(y) = 1 \tag{47b}$$

(sums to one). Facetiously, one can say that (master's level) probability theory is the study of functions that are nonnegative and sum to one. The domain S of the PMF is called the *sample space.* A real-valued function on the sample space is called a *random variable*, and the *expectation* of a random variable $g(y)$ is given by

$$E\{g(y)\} = \sum_{y \in S} g(y) f(y) \tag{48}$$

if the sum exists (which it may not if the sample space is infinite). Probability is a special case of expectation; it is expectation of Bernoulli (zero-or-one-valued) random variables. If g is Bernoulli, define

$$A = \{\, y \in S : g(y) = 1 \,\}.$$

In this case the expectation (48) becomes

$$P(A) = \sum_{y \in A} f(y) \tag{49}$$

because $g(y) = 1$ for $y \in A$ and $g(y) = 0$ for $y \notin A$. The sum in (49) always exists because it is less than the sum in (47b) which exists by definition of PMF. Any subset of the sample space is called an *event* and (49) is called the *probability of the event A*.

"Master's level probability theory" also studies continuous probability models in which the sums above are replaced by integrals, but we need not discuss them here, because all aster models that have been used in published papers are discrete, so that is all the probability theory we need for this discussion.

Notice that, as we said above, these definitions have not the faintest trace of the philosophical arguments about what probability and expectation "really are." Mathematicians have simply declared that anyone's notions of probability and expectation that do not agree with these formal definitions (48) and (49) will not be entertained, and conversely any notions that do agree with these formal definitions are fine with them. Mathematicians do not care about the philosophical arguments; they define expectation and probability mathematically and get on with doing mathematics.

A *statistical model* is a family of probability models. The problem of *statistical inference* is, given data that is supposed to be from one probability model in a statistical model, to say something about which one it is.[28] This brings us to deep philosophical waters. Statistical inference is the statistician's answer to the philosopher's questions of induction and epistemology. When in a totalizing mood, statisticians can think that all learning about

[28] This "saying something about" can take the form of point estimates, such as maximum likelihood estimates, confidence intervals, or tests of statistical hypotheses.

the world is statistical inference.[29] But leaving that aside, what theoretical statistics is all about is statistical inference, so here we take the idea for granted.

Statistical models are often specified by *parameterization*.[30] Each probability model in the statistical model is specified by a PMF that is a function of another variable (or variables) called the *parameter* (or *parameters*). When there are multiple parameters, these are thought of as a single mathematical object called the *parameter vector*. The parameterization is indicated by subscripts f_θ on the PMF, where θ is the parameter or parameter vector (it is a convention to use roman letters for ordinary variables and random variables and greek letters for parameters). A subscript indicating the parameter is also added to the notation for expectation

$$E_\theta\{g(y)\} = \sum_{y \in S} g(y) f_\theta(y) \tag{50}$$

and probability

$$P_\theta(A) = \sum_{y \in A} f_\theta(y). \tag{51}$$

Statistical models are also very abstract. Firstly, they are abstract because probability models are abstract and each statistical model is an abstract set whose elements are probability models. Secondly, there is not only the abstraction of probability models (ignoring what probability and expectation "really mean") but also the abstraction of statistical inference. Statisticians talk about statistical models by saying one particular parameter value among the abstract set of allowed parameter values is the *truth*

[29] This is obvious nonsense because people learned things long before anything was known of statistics, but there is a grain of truth in it. Except for logical tautologies, all knowledge is uncertain, and all uncertainty can be modeled by probability theory — whether this gives a satisfactory and complete description of uncertainty is debatable, but such modeling is possible — from which it follows that all learning is statistical if the uncertainty of knowledge is to be properly accounted for. This is understood in the branch of philosophy called "Bayesian epistemology" (Talbott, 2008); our point here is that non-Bayesian statistical inference also answers epistemological questions. Most people most of the time feel no uncertainty about most of their knowledge; the customs of their tribe, sect, and clique are not questioned. This goes for intellectuals and academics too. Even scientists use statistics only when the uncertainty of inference is so glaring that any attempt at avoiding statistics must be forlorn. They use it only when they have to. The fact that in real life not all learning is statistical is a deep mystery. Most people most of the time are quite certain about many very questionable things, and this seems to work. But why?

[30] Statistical models that are too big to be specified by a finite set of parameters exist and are called "nonparametric."

or *true unknown parameter value*, the latter to emphasize that which is the truth is unknown and the problem of statistical inference is to say something about which parameter value it is (footnote 28). It is "assumed" that if θ is the truth, then the PDF f_θ corresponding to it is the probability model "for" the observed data, meaning that, whatever philosophical attitude we have about the correspondence of probability models to the real world, we have concluded that this model describes the natural process generating these data.[31] Piled on top of the probabilist's abstraction is the statistician's abstraction: learning the truth (about nature) is saying something about a parameter in a statistical model. All learning can be formalized this way, and statisticians do so.

It is an observable fact about statistics teaching that the concept of statistical models is considered problematic and completely avoided in introductory courses. The terms *statistic* and *parameter* are much discussed (see footnote 20), but the abstract notion of a family of probability distributions indexed by the parameter (so each particular value of the parameter vector specifies one probability model) is absent. The distinction is clearly drawn between a parameter θ and a statistic $\hat{\theta}$ used to estimate it. But what is left unmentioned is the fact that θ specifies the probability model through its PMF f_θ — most introductory statistics textbooks do not develop enough probability theory to make this clear — and hence the estimator $\hat{\theta}$ also specifies a probability model through its PMF $f_{\hat{\theta}}$.

The two probability models are different: statistics are not parameters so $\hat{\theta}$ is not θ, hence $f_{\hat{\theta}}$ is not f_θ. Nevertheless, θ being unknown, our only guide to what f_θ says about the data is $f_{\hat{\theta}}$. So we use $f_{\hat{\theta}}$ in place of f_θ, not blindly or stupidly but cautiously and sophisticatedly, worrying about and quantifying the differences between our calculations based on $f_{\hat{\theta}}$ and what those same calculations would be if based on f_θ (which they ought to be but cannot be because θ is unknown).

This caution gives rise to an infinite regress: when we worry about the differences between f_θ and $f_{\hat{\theta}}$ we realize those differences are random because $\hat{\theta}$ is random (being a function of the data, which are random). Thus we must do probabilistic calculations based on a PMF, which ought to be f_θ, but we must use $f_{\hat{\theta}}$ because θ is unknown. So we have higher-order worries that our calculations about the errors made in using $f_{\hat{\theta}}$ instead of f_θ are themselves wrong because of the same issue. So we repeat the process, attempting to

[31] This "assumed" is for the sake of argument, to get on with the mathematics of statistical inference. The assumption may later be questioned based on the inferences made. In their example 3, Shaw et al. (2008) first "assume" an aster model for the data but later check its validity using residual analysis, their Figure 4.

quantify our higher-order errors, and are faced with yet higher-order errors, and so forth ad infinitum. Fortunately, the errors get smaller as both order and sample size increase and after some point can be ignored. We raise this infinite regress to show that statistical inference has some very deep philosophical issues.

But that infinite regress is not the main subject of this section, the main subject is that most users of statistics, including most scientists, have only the foggiest ideas about statistical models. Perhaps this is the fault of bad statistics teaching or of human discomfort about randomness, probability, and statistics. Most scientists do not think of parameters θ as mere indices that specify a probability model through its PMF f_θ and do nothing more. Rather they reify parameters as "facts of nature" asking what the parameters "really mean" not understanding our dictum (page 30) that parameters are meaningless in themselves and only have meaning through f_θ and probabilities and expectations calculated using f_θ. If I had a nickel for every time I've tried to explain that fitting statistical models to data does not directly discover truth about nature,[32] and that the parameters estimated can only be interpreted through the statistical model and its properties, I'd be rich. I can make the same explanation over and over to the same person and get the response over and over "yes, I know that" but "these sorts of explanations" (reifying parameters) "are what's expected." That reifying parameters is just wrong is never directly confronted.

Returning to aster models, Geyer et al. (2007), Shaw et al. (2008), and Shaw and Geyer (submitted) are careful to avoid reifying parameters, so much so that a referee of Shaw and Geyer (submitted) complained that more explanation of "what the parameters mean" was needed. When one understands that the parameters don't "mean" anything and that interpretations should refer to the probability models indexed by parameters rather than the parameters themselves, how does one provide valid scientific interpretations of statistical results? One way is hypothesis tests of model comparison (done by the function `anova` for aster model fits). Models have scientific interpretations, hypothesis tests compare two models, one a submodel of the

[32] Am I not being contradictory? Does statistical inference discover truth or not? When I am wearing my statistician hat, yes it does, "truth" being *defined* as the parameter value that indexes the correct probability model, which is "assumed" to be a member of the statistical model being used. When I am wearing my scientist hat, no it does not. All models are wrong; none capture all aspects the phenomena being studied. Moreover, to the extent that they capture any aspects, they do so through probabilities and expectations calculated using the model. Hence our emphasis on mean value parameters corresponding to scientifically interesting random variables. Those have direct scientific interpretation. Parameters, in general, do not.

other, and show either that the submodel fits as well as the supermodel (so the supermodel adds nothing useful and scientific interpretation should be based on the submodel) or it does not (so only the supermodel fits the data and scientific interpretation should be based on it). Often, this is all that needs to be done: report which model fits the data best. When one wants more, an explanation of what this "best" model does and does not say, one must be very careful, as we saw in the discussion of Example 2 of Shaw et al. (2008) (Section 3 above). What Shaw et al. (2008) did in that example was use confidence intervals for mean value parameters as the basis of their scientific interpretation. This was also done in the example in Geyer et al. (2007). In Example 1 of Shaw et al. (2008) the parameter of scientific interest was the population growth rate λ which is a function of the mean value parameter vector μ through the "stable age equation" so the principle of invariance of maximum likelihood estimates (footnote 18) can be used to find the MLE of λ and the delta method can be used to find a confidence interval for λ. In Example 3 of Shaw et al. (2008) and the examples of Shaw and Geyer (submitted) the object of scientific interest is the fitness landscape, like the function ψ defined by (39). This is an infinite-dimensional parameter (a parameter vector with an infinite number of components), being a whole function and the components being the values of the function for each possible argument value. So it can only be "reported" by plotting a graph of the function, which these papers do. Confidence regions for infinite-dimensional parameter vectors are possible, but too complicated to be of much use; none are reported in these papers. Instead, confidence regions are reported for specific aspects of this function such as the point where it achieves its maximum (the fitness optimum). Seen in this light, scientific interpretation of statistical results is tricky. The scientific interpretation of a statistical model is by no means obvious. One must decide which probabilities of what events, expectations of what statistics, hypothesis tests comparing what models, or confidence intervals for what parameters should be computed to provide a clear and convincing scientific interpretation. The job is only started when the "best" statistical model is found.

Acknowledgments

The impetus for this technical report was helping a philosopher, Emily Grosholz, get up to speed on aster models in aid of writing a paper on them (Grosholz, submitted). But the process went both ways. Questions and comments from Emily led to large improvements in this technical report.

References

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

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, Hayward, CA.

Casella, G. and Berger, R. L. (2001). *Statistical Inference*, 2nd ed. Pacific Grove, CA: Duxbury.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, **222**, 309–368.

Fristedt, B. E. and Gray, L. F. (1996). *A Modern Approach to Probability Theory.* Boston: Birkhäuser.

Geyer, C. J. (1990). Likelihood and exponential families. Ph.D. thesis, University of Washington. `http://purl.umn.edu/56330`.

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

Geyer, C. J. (2009b). R package aster: Aster models, version 0.7-7. If the R statistical computing environment (R Development Core Team, 2009) has already been installed, then this package is installed by the R command `install.packages("aster")` on any computer.

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

Grosholz, E. (submitted). Studying populations without molecular biology: Aster models and a new argument against reductionism.

Hájek, A. (2007). Interpretations of probability. *Stanford Encyclopedia of Philosophy.* `http://plato.stanford.edu/`.

Jaynes, E. T. (1978). Where do we stand on maximum entropy? In *The Maximum Entropy Formalism*, Ed. R. D. Levine and M. Tribus, pp. 15-118. MIT Press, Cambridge, MA.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer. English translation (1950): *Foundations of the theory of probability.* Chelsea.

Lande, R., and Arnold, S. J. 1983. The measurement of selection on correlated characters. *Evolution*, **37**, 1210–1226.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

Oehlert, G. W. (2000). *A First Course in Design and Analysis of Experiments.* New York: W. H. Freeman.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`.

Rockafellar, R. T. and Wets, R. J.-B. (2004). *Variational Analysis*, corr. 2nd printing. Berlin: Springer-Verlag.

Rudin, W. (1986). *Real and Complex Analysis*, 3rd ed. New York: McGraw-Hill.

Severini, T. A. (2001). *Likelihood Methods in Statistics.* Oxford: Oxford University Press.

Shaw, R. G., Geyer, C. J., Wagenius, S., Hangelbroek, H. H., and Etterson, J. R. (2007) Unifying life history analysis for inference of fitness and population growth. *American Naturalist*, **172**, E35–E47.

Shaw, R. G. and Geyer, C. J. (submitted). Inferring fitness landscapes. Submitted to *Evolution.*

Talbott, W. (2008). Bayesian Epistemology. *Stanford Encyclopedia of Philosophy.* `http://plato.stanford.edu/`.

van der Vaart, A. W. (2000). *Asymptotic Statistics* Cambridge University Press.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd ed. Wiley, New York.