Stat 5421 Lecture Notes: Exponential Families

Charles J. Geyer

December 14, 2021

Contents

1 License

2 R

- The version of R used to make this document is 4.1.2.
- The version of the `rmarkdown` package used to make this document is 2.11.

3 Exponential Families

3.1 Definitions

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

$$l(\theta) = \langle y, \theta \rangle - c(\theta) \tag{1}$$

where

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

In equation (1) we have used the rule that additive terms in the log likelihood that do not contain the parameter may be dropped. Any such terms have been dropped in (1).

The angle brackets notation comes from functional analysis. It has the virtue of not treating y and θ as being vectors in the same space, but rather in dual vector spaces (a concept from linear algebra). Of course, we know that statistics and parameters are different, but it is nice to have our notation reflect that. Nevertheless, we will sometimes use other notation

$$\langle y, \theta \rangle = y^T \theta = \theta^T y = \sum_i y_i \theta_i$$

Often the canonical statistic and canonical parameter are not the usual statistic or usual parameter. One has to do a change of parameter or change of statistic to get the log likelihood into the form (1). More on this in Section 7 below.

3.2 Terminology

Many people use "natural" everywhere this document uses "canonical". In this we are following Barndorff-Nielsen (1978).

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

3.3 Some Examples

All of the models used in this course, Poisson, multinomial, product multinomial, univariate and multivariate normal, are exponential families. So are logistic regression, Poisson regression with log link, and multinomial response regression. So are hierarchical and graphical log-linear models. But not the chi-square distribution (because it has no parameters).

More about these examples in Sections 7 and 9.3 below.

Also many models outside of this course are exponential families. Linear regression is. So are many complicated models in spatial statistics (Geyer, 1990, 1999; Geyer and Møller, 1994) and biology (Geyer *et al.*, 1993, 2007; Geyer and Thompson, 1992).

The point of the theory of exponential families in general and of these notes in particular is to bring all of these disparate statistical models under one theoretical concept so we can learn general principles that apply to all of them.

3.4 Uniqueness

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

- any one-to-one affine function of a canonical statistic vector is another canonical statistic vector,

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

(see Section 9.3.2 below for the definition of affine function).

These possible changes of statistic, parameter, or cumulant function are not algebraically independent. Changes to one may require changes to the others to keep a log likelihood of the form (1) above.

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

3.5 Cumulant Functions

The cumulant function may not be defined by (1) above on the whole vector space where θ takes values. In that case it can be extended to this whole vector space by

$$c(\theta) = c(\psi) + \log\left\{E_\psi(e^{\langle Y, \theta - \psi\rangle})\right\} \tag{2}$$

where θ varies while ψ is fixed at a possible canonical parameter value, and the expectation and hence $c(\theta)$ are assigned the value ∞ for θ such that the expectation does not exist (our equation (2) is equation (5) in Geyer (2009)).

3.6 Full and Regular Exponential Families

An exponential family is *full* if its canonical parameter space is

$$\Theta = \{\, \theta : c(\theta) < \infty \,\} \tag{3}$$

(where the cumulant function is defined by (2)) and a full family is *regular* if its canonical parameter space is an open subset of the vector space where θ takes values.

Almost all exponential families used in real applications are full and regular. So-called *curved exponential families* (smooth non-affine submodels of full exponential families) are not full (an example is the ABO blood group example in our notes on likelihood computation). Constrained exponential families (Geyer, 1991) are not full. A few exponential families used in spatial statistics are full but not regular (Geyer, 1999; Geyer and Møller, 1994).

An example where the canonical parameter space of a full family (3) is not a whole vector space is the negative binomial distribution (Section 7.3 below).

3.7 Probability Mass-Density Functions

Our definition (Section 3.1 above) simplifies many arguments but it does not tell us exactly what the probability mass functions (PMF) or probability density functions (PDF) are.

We can get PMF or PDF as the case may be (we say PMDF to include both) by realizing that the log likelihood (1) is log PMDF plus perhaps a term that may depend on the data but does not contain parameters. Thus the PMDF of the distribution having canonical parameter vector θ has the form

$$f_\theta(w) = e^{\langle Y(w), \theta\rangle - c(\theta) + h(w)} \tag{4}$$

where w is the whole data (Y may be only a statistic, that is, a function of the whole data) and $h(w)$ is some term not containing the parameters that is dropped to get the log likelihood to have the form (1).

3.8 Non-Degeneracy

The PMDF (4) of an exponential family are never zero. This means all distributions in the family have the same support. There are no degenerate distributions having different support.

More on this in Section 7 below.

4 Moments and Cumulants

The reason why the cumulant function has the name it has is because it is related to the cumulant generating function (CGF), which is the logarithm of a moment generating function (MGF). Derivatives of an MGF evaluated at zero give moments (expectations of powers of a random variable). Derivatives of a CGF evaluated at zero give cumulants. Cumulants are polynomial functions of moments and vice versa.

Using (2), the MGF the distribution of the canonical (vector) statistic for canonical parameter value θ in an exponential family with log likelihood (1) is given by

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

provided this formula defines an MGF, which it does if and only if it is finite for t in a neighborhood of zero, which happens if and only if θ is in the interior of the full canonical parameter space (3).

So the cumulant generating function is

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

provided θ is in the interior of Θ.

It is easy to see that derivatives of K_θ evaluated at zero are derivatives of c evaluated at θ. So derivatives of c evaluated at θ are cumulants.

We will be only interested in the first two cumulants

$$E_\theta(Y) = \nabla c(\theta) \tag{5}$$
$$\operatorname{var}_\theta(Y) = \nabla^2 c(\theta) \tag{6}$$

(Barndorff-Nielsen, 1978, Theorem 8.1).

Equation (5) is a vector equation, on the left-hand side we have the mean vector of the canonical statistic vector, and on the right-hand side we have the first derivative vector (also called gradient vector) of the cumulant function. Its components are the partial derivatives $\partial c(\theta)/\partial\theta_i$.

Equation (6) is a matrix equation, on the left-hand side we have the variance matrix of the canonical statistic vector, and on the right-hand side we have the second derivative matrix (also called Hessian matrix) of the cumulant function. Its components are the second partial derivatives $\partial^2 c(\theta)/\partial\theta_i\partial\theta_j$.

Hence for any θ in the interior of Θ, the corresponding probability distribution has moments and cumulants of all orders. In particular, every distribution in a regular full exponential family has moments and cumulants of all orders and the mean and variance are given by the formulas above.

Conversely, any distribution whose canonical parameter value is on the boundary of the full canonical parameter space does not have a moment generating function or a cumulant generating function, and no moments or cumulants need exist.

5 Mean Value Parameters

5.1 Canonical to Mean Values

Define a mapping from the canonical parameter space Θ to the vector space where the canonical statistic vector and mean value parameter live by

$$h(\theta) = \nabla c(\theta) = E_\theta(Y) \tag{7}$$

using equation (5).

5.2 The Mean Value Parameterization

Theorem 5.1. *The mean values $\mu = E_\theta(Y)$ parameterize a regular full exponential family. Different distributions in the family have different mean vectors.*

The proof of this theorem (and all other theorems in these notes) is given in Appendix A.

We call μ as defined in the theorem and in equation (7) the *mean value parameter vector*.

In elementary applications mean value parameters are preferred. When we introduce the Poisson distribution we use mean the mean value parameter μ. When we introduce the binomial and multinomial distributions we use a parameter proportional to the mean value parameter: the usual parameter is π and the mean value parameter is $\mu = n\pi$.

It is only when we get to generalized linear models with Poisson, binomial, or multinomial response and log-linear models for contingency tables that we need canonical parameters. But for these models mean value parameters are also useful. We often reason using both.

More on this in Sections 9.3 and 10 and 13 below.

6 Identifiability

A parameterization of a statistical model is *identifiable* if no two distinct parameter values correspond to the same probability distribution.

6.1 Canonical Parameterization

The canonical parameterization of a regular full exponential family need not be identifiable. It is identifiable if and only if the canonical statistic vector is not concentrated on a hyperplane in the vector space where it takes values (Geyer, 2009, Theorem 1). It is always possible to choose an identifiable canonical parameterization but not always convenient (Geyer, 2009).

More on this in Section 7 below.

6.2 Mean Value Parameterization

The mean value parameterization of a regular full exponential family is always identifiable. No distribution can have two different mean vectors.

6.3 Maps Between the Two Parameterizations

Since cumulant functions are infinitely differentiable, the transformation from canonical to mean value parameters described above is infinitely differentiable.

If the canonical parameterization is identifiable, then the inverse transformation exists (because the transformation is then one-to-one and onto) and is also infinitely differentiable by the inverse function theorem of real analysis. But we often do not have a symbolic expression for this inverse transformation.

Theorem 6.1. *Suppose we are working with a regular full exponential family. Define the function*

$$q(\theta) = \langle \mu, \theta \rangle - c(\theta) \tag{8}$$

If μ is a possible value of the mean value parameter vector, then any θ that maximizes q is a corresponding value of the canonical parameter vector. Moreover the maximum always exists.

Observe that (8) is the same as the log likelihood (1) except that y in (1) is replaced by μ in (8), where $\mu = E_\theta(Y)$. That is, we replace the observed value of the canonical statistic vector by a possible mean value of it.

Theorem 6.1 allows us to compute θ in terms of μ but does not give us a symbolic expression for the transformation from one parameterization to the other (no such symbolic expression need exist).

The theorem does not assert that the maximizer of (8) is unique. It must be unique if the canonical parameterization is identifiable. It need not be unique if the canonical parameterization is not identifiable. In fact, any canonical parameter vector corresponding to the same distribution as μ maximizes (8).

Theorem 6.2. *Suppose we are working with a regular full exponential family. If the canonical parameterization is identifiable, any algorithm that always goes uphill if it can on q defined by (8) converges to the unique maximizer of q.*

The point of the theorem is that maximization of upper semicontinuous strictly concave functions having a unique maximizer is easy (which includes log likelihoods of regular full exponential families and also q defined by (8) when the canonical parameterization is identifiable). Any algorithm, no matter what the starting point, and no matter how inefficient, can find the unique solution, so long as it is not so stupid as to be going downhill when it should be going uphill.

Theorem 6.3. *Suppose we are working with a regular full exponential family. If the canonical parameterization is identifiable, any algorithm that always goes uphill if it can on the log likelihood defined by (1) converges to the unique MLE if the observed value of the canonical statistic vector y is a possible value of the mean value parameter vector.*

The case where the observed value of the canonical statistic vector y is not a possible value of the mean value parameter vector, and the MLE does not exist is discussed in other lecture notes.

More on this in Section 13 below.

7 More About Examples

7.1 The Binomial Distribution

The binomial distribution has log likelihood

$$l(\pi) = x \log(\pi) + (n - x) \log(1 - \pi)$$

where x is the "usual" data, π is the "usual" parameter, and n is neither data nor parameter but rather a constant. To get this into the form (1), assuming the family is one-dimensional (which having only one parameter suggests), we need to write this as a function of x (which will be the canonical statistic y) times a function of π (which will be the canonical parameter θ) minus a function of θ (the cumulant function). So isolate x

$$\begin{aligned} l(\pi) &= x\big[\log(\pi) - \log(1 - \pi)\big] + n \log(1 - \pi) \\ &= x \log\left(\frac{\pi}{1 - \pi}\right) + n \log(1 - \pi) \end{aligned}$$

and this tells us

- the canonical statistic is x,
- the canonical parameter is

$$\theta = \log\left(\frac{\pi}{1 - \pi}\right), \tag{9}$$

  and

- the cumulant function is

$$c(\theta) = -n \log(1 - \pi). \tag{10}$$

Of course (10) doesn't really make sense because it has θ on one side and π on the other. We need to solve (9) for π obtaining

$$\pi = \frac{e^\theta}{1 + e^\theta} \tag{11}$$

and plug this back into (10) obtaining

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

Since this formula is valid for all θ, we see the binomial family is a regular full exponential family whose canonical parameter space is the whole of Euclidean space (one-dimensional Euclidean space in this case).

This change-of-parameter is so important that statisticians give it a name

$$\operatorname{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) \qquad 0 < \pi < 1. \tag{12}$$

(logit is pronounced "low-jit" with a soft "g").

Note that 0 and 1 are not possible arguments of the logit function. We get $\log(0)$ or division by zero, either of which is undefined. The logit function maps $(0, 1) \to (-\infty, +\infty)$. So the binomial family of distributions, considered as an exponential family, has "usual" parameter space $0 < \pi < 1$ and canonical parameter space $-\infty < \theta < +\infty$).

This is a little surprising because the usual "usual" parameter space is $0 \le \pi \le 1$ because otherwise we have no parameter values to be maximum likelihood estimates when we observe $x = 0$ or $x = n$. So exponential family theory causes some trouble here. Much more on this subject later.

If we think about it a bit more though, this is just the non-degeneracy property of exponential families (Section 3.8 above) at work. This is why the degenerate distributions concentrated at $x = 0$ and $x = n$ (for $\pi = 0$ and $\pi = 1$) cannot be included in the exponential family.

Theorem 7.1.

$$\begin{aligned} c'(\theta) &= n\pi \\ c''(\theta) &= n\pi(1 - \pi) \end{aligned}$$

Thus this theory gives us back the known mean and variance of the binomial distribution and the map from canonical parameter to mean value parameter and Fisher information.

Summary

For the binomial distribution

- the canonical statistic is x,
- the canonical parameter is $\theta = \operatorname{logit}(\pi)$,
- the mean value parameter is $\mu = n\pi$,
- the canonical parameter space of the full family is $-\infty < \theta < +\infty$,
- the mean value parameter space of the full family is $0 < \mu < n$,
- the cumulant function is
$$c(\theta) = n \log(1 + e^\theta), \tag{13}$$
and
- Fisher information for θ is $I(\theta) = n\pi(1 - \pi)$.

7.2 The Poisson Distribution

The Poisson distribution has log likelihood

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

where x is the "usual" data and μ is the "usual" parameter. To get this into the form (1), assuming the family is one-dimensional (which having only one parameter suggests), we need to write this as a function of x (which will be the canonical statistic y) times a function of μ (which will be the canonical parameter θ) minus a function of θ (the cumulant function). Obviously,

- the canonical statistic is x,
- the canonical parameter is
$$\theta = \log(\mu),$$
which has inverse function
$$\mu = e^\theta,$$
and
- the cumulant function is
$$c(\theta) = \mu = e^\theta.$$

Note that 0 is not a possible argument of the log function. The log function maps $(0, +\infty) \to (-\infty, +\infty)$. So the Poisson family of distributions, considered as an exponential family, has "usual" parameter space $0 < \mu < +\infty$ and canonical parameter space $-\infty < \theta < +\infty$).

The Poisson distribution for $\mu = 0$ is degenerate, concentrated at zero, so by the non-degeneracy property (Section 3.8 above) it cannot be included in the exponential family.

Theorem 7.2.

$$c'(\theta) = \mu$$
$$c''(\theta) = \mu$$

Summary

For the Poisson distribution

- the canonical statistic is x,
- the canonical parameter is $\theta = \log(\mu)$,
- the mean value parameter is μ (this is also the "usual" parameter),
- the canonical parameter space of the full family is $-\infty < \theta < +\infty$,
- the mean value parameter space of the full family is $0 < \mu < +\infty$, and
- the cumulant function is
$$c(\theta) = e^\theta, \tag{14}$$
and
- Fisher information for θ is $I(\theta) = \mu$.

7.3 The Negative Binomial Distribution

The negative binomial distribution has PMF

$$f(x) = \binom{r+x-1}{x} \pi^r (1-\pi)^x, \qquad x = 0, 1, \ldots \tag{15}$$

where x is the "usual" data and r and π are the "usual" parameters and

$$\binom{r}{x} = \frac{r \cdot (r-1) \cdots (r-x+1)}{x!}.$$

If we consider both r and π to be unknown parameters, this is not an exponential family. If we consider r to be known, so π is the only unknown parameter, then it is an exponential family. So we consider r as known.

Then the log likelihood is

$$l(\pi) = r \log(\pi) + x \log(1-\pi).$$

Obviously,

- the canonical statistic is x,
- the canonical parameter is
$$\theta = \log(1-\pi),$$
which has inverse function
$$\pi = 1 - e^\theta,$$
and
- the cumulant function is
$$c(\theta) = -r \log(\pi) = -r \log(1 - e^\theta). \tag{16}$$

The negative binomial distribution for $\pi = 0$ does not exist ($\pi = 0$ is not an allowed parameter value). The negative binomial distribution for $\pi = 1$ does exist but is degenerate, concentrated at zero. So this degenerate distribution cannot be included in the exponential family (by Section 3.8 above).

The log function maps $(0, 1) \to (-\infty, 0)$. So the negative binomial family of distributions, considered as an exponential family, has "usual" parameter space $0 < \pi < 1$ and canonical parameter space $-\infty < \theta < 0$).

We are tempted to use (2) to extend the family, but if we try, we find that (2) gives us the same cumulant function we already have, so the negative binomial family as we have described it is already a regular full exponential family.

Theorem 7.3.

$$c'(\theta) = \frac{r(1-\pi)}{\pi}$$
$$c''(\theta) = \frac{r(1-\pi)}{\pi^2}$$

Summary

For the negative binomial distribution

- the canonical statistic is x,
- the canonical parameter is $\theta = \log(1-\pi)$,
- the mean value parameter is $\mu = r(1-\pi)/\pi$,
- the canonical parameter space of the full family is $-\infty < \theta < 0$,

- the mean value parameter space of the full family is $0 < \mu < \infty$,
- the cumulant function is
$$c(\theta) = -r \log(1 - e^\theta), \tag{17}$$
and
- Fisher information for θ is $I(\theta) = r(1 - \pi)/\pi^2$.

Commentary

This provides an example of a regular full exponential family whose canonical parameter space (3) is not a whole Euclidean space (in this case one-dimensional Euclidean space).

7.4 The Multinomial Distribution

The multinomial PDF is
$$f(x) = \binom{n}{x} \prod_{i=1}^k \pi_i^{x_i}$$
where $x = (x_1, \ldots, x_k)$ is the vector of cell counts of a contingency table (the "usual" data vector), where $\pi = (\pi_1, \ldots, \pi_k)$ is the vector of cell probabilities (the "usual" parameter vector), where n is the sample size (the sum of the cell counts), and where $\binom{n}{x}$ is a multinomial coefficient. The log likelihood is
$$l(\pi) = \sum_{i=1}^k x_i \log(\pi_i)$$

7.4.1 Try I

It looks like we have exponential family form with the canonical statistic vector x, canonical parameter vector θ having components $\theta_i = \log(\pi_i)$, and cumulant function the zero function.

As we shall eventually see, this is correct but misleading. If we use (5) and (6) on the zero function we get that y has mean vector zero and variance matrix zero, which is not the mean vector and variance matrix of the multinomial distribution. But we are not allowed to use those formulas because the components of θ, as we have defined them in this section, are not separately variable: the components of π are constrained to sum to one. And that translates to a complicated nonlinear constraint on the canonical parameters
$$\sum_{i=1}^k e^{\theta_i} = 1. \tag{18}$$

That looks annoying. So we try something else.

7.4.2 Try II

If we eliminate one of the components, say π_k, we get log likelihood
$$l(\pi_1, \ldots, \pi_{k-1}) = x_k \log\left(1 - \sum_{i=1}^{k-1} \pi_i\right) + \sum_{i=1}^{k-1} x_i \log(\pi_i)$$

But since we are trying to put this in exponential family form in which the canonical statistic vector and canonical parameter vector must have the same dimension, we have to also eliminate the corresponding component of x. We can do that because the components of x are constrained to sum to n. If we use that

fact to eliminate x_k we get

$$\begin{aligned} l(\pi_1, \ldots, \pi_{k-1}) &= \left(n - \sum_{i=1}^{k-1} x_i \right) \log \left(1 - \sum_{i=1}^{k-1} \pi_i \right) + \sum_{i=1}^{k-1} x_i \log(\pi_i) \\ &= n \log \left(1 - \sum_{i=1}^{k-1} \pi_i \right) + \sum_{i=1}^{k-1} x_i \left[\log(\pi_i) - \log \left(1 - \sum_{i=1}^{k-1} \pi_i \right) \right] \\ &= n \log \left(1 - \sum_{i=1}^{k-1} \pi_i \right) + \sum_{i=1}^{k-1} x_i \log \left(\frac{\pi_i}{1 - \sum_{i=1}^{k-1} \pi_i} \right) \end{aligned}$$

and this has exponential family form with

- canonical statistic vector $(x_1, \ldots, x_{k-1})$,
- canonical parameter vector $(\theta_1, \ldots, \theta_{k-1})$, where

$$\theta_i = \log \left(\frac{\pi_i}{1 - \sum_{i=1}^{k-1} \pi_i} \right) \tag{19}$$

  and

- cumulant function

$$c(\theta_1, \ldots, \theta_{k-1}) = -n \log \left(1 - \sum_{i=1}^{k-1} \pi_i \right), \tag{20}$$

except that (as usual) this doesn't make sense with θ's on the left-hand side and π's on the right-hand side, so we have to solve for the π's in terms of the θ's and plug in. Reintroduce $\pi_k = 1 - \sum_{i=1}^{k-1} \pi_i$ and plug into (19) obtaining

$$\theta_i = \log \left(\frac{\pi_i}{\pi_k} \right)$$

or

$$e^{\theta_i} = \frac{\pi_i}{\pi_k}$$

or

$$\pi_i = \pi_k e^{\theta_i} \tag{21}$$

and this also holds for $i = k$ if we define $\theta_k = 0$ (which we may do because there was no definition of θ_k before). Since the π's must sum to 1, summing (21) gives

$$1 = \pi_k \sum_{i=1}^{k} e^{\theta_i} = \pi_k \left(1 + \sum_{i=1}^{k-1} e^{\theta_i} \right)$$

and

$$\pi_k = \frac{1}{1 + \sum_{i=1}^{k-1} e^{\theta_i}}$$

and (plugging this back into (21))

$$\pi_i = \frac{e^{\theta_i}}{1 + \sum_{i=1}^{k-1} e^{\theta_i}} \tag{22}$$

and (plugging this back into (20))

$$\begin{aligned} c(\theta_1, \ldots, \theta_{k-1}) &= -n \log \left(1 - \sum_{i=1}^{k-1} \frac{e^{\theta_i}}{1 + \sum_{i=1}^{k-1} e^{\theta_i}} \right) \\ &= n \log \left(1 + \sum_{i=1}^{k-1} e^{\theta_i} \right) \end{aligned}$$

I don't know what you think of this, but I think it is horribly messy. Also arbitrary. Clearly, which π_i we choose to eliminate is arbitrary. But we also know (Section 3.4 above) that there is a lot more arbitrariness than that in choosing canonical statistic, canonical parameter, and cumulant function.

Nevertheless, this messy parameterization is widely used. Agresti (2013) uses it in Chapter 8.

7.4.3 Try III

Let's try again, this time not eliminating parameters or statistics, and using (2) to define the cumulant function. Let ψ in (2) be the canonical parameter vector for the multinomial distribution having probability vector $p = (p_1, \ldots, p_k)$. Let S denote the sample space of the multinomial distribution: the set of all vectors x having nonnegative integer components that sum to n. Then (2) says

$$\begin{aligned}
c(\theta) &= c(\psi) + \log \sum_{x \in S} e^{\langle x, \theta - \psi \rangle} \cdot \binom{n}{x} \prod_{i=1}^{k} p_i^{x_i} \\
&= c(\psi) + \log \sum_{x \in S} e^{\sum_{i=1}^{k} x_i (\theta_i - \psi_i)} \cdot \binom{n}{x} \prod_{i=1}^{k} p_i^{x_i} \\
&= c(\psi) + \log \sum_{x \in S} \left[\prod_{i=1}^{k} e^{x_i (\theta_i - \psi_i)} \right] \cdot \binom{n}{x} \prod_{i=1}^{k} p_i^{x_i} \\
&= c(\psi) + \log \sum_{x \in S} \binom{n}{x} \prod_{i=1}^{k} p_i^{x_i} e^{x_i (\theta_i - \psi_i)} \\
&= c(\psi) + \log \sum_{x \in S} \binom{n}{x} \prod_{i=1}^{k} \left[p_i e^{\theta_i - \psi_i} \right]^{x_i} \\
&= c(\psi) + \log \left(\sum_{i=1}^{k} p_i e^{\theta_i - \psi_i} \right)^n \\
&= c(\psi) + n \log \left(\sum_{i=1}^{k} p_i e^{\theta_i - \psi_i} \right)
\end{aligned}$$

We use the multinomial theorem to evaluate the sum over the sample space. If we take ψ to be the vector with all components zero and p to be the vector with all components equal to $1/k$, which we are free to do because ψ and p were arbitrary, we get

$$\begin{aligned}
c(\theta) &= c(\psi) + n \log \left(\frac{1}{k} \sum_{i=1}^{k} e^{\theta_i} \right) \\
&= c(\psi) + n \log \left(\frac{1}{k} \sum_{i=1}^{k} e^{\theta_i} \right) \\
&= c(\psi) - n \log(k) + n \log \left(\sum_{i=1}^{k} e^{\theta_i} \right)
\end{aligned}$$

and, since we are free to choose $c(\psi)$, we can choose it to cancel the $-n \log(k)$, finally obtaining

$$c(\theta) = n \log \left(\sum_{i=1}^{k} e^{\theta_i} \right)$$

Theorem 7.4. *The map from canonical parameters to usual parameters is*

$$\pi_j = \frac{e^{\theta_j}}{\sum_{i=1}^{k} e^{\theta_i}} \tag{23}$$

and

$$\begin{aligned} \partial c(\theta)/\partial\theta_i &= n\pi_i \\ \partial^2 c(\theta)/\partial\theta_i^2 &= n\pi_i(1-\pi_i) \\ \partial^2 c(\theta)/\partial\theta_i\partial\theta_j &= -n\pi_i\pi_j, \qquad i \neq j \end{aligned}$$

So the Fisher information matrix for θ, denoted $I(\theta)$ has i,i components $n\pi(1-\pi)$ and i,j components $-n\pi\pi_j$ for $i \neq j$.

A multinomial distribution having usual parameter vector π with any component π_i equal to zero is partially degenerate, with the distribution of y_i concentrated at zero, so by the non-degeneracy property (Section 3.8 above) it cannot be included in the exponential family. In light of this it is not surprising that (23) is never zero. Similarly for the try II parameterization: (22) is never zero.

Summary

7.4.3.1 Try II

For the multinomial family of dimension k and sample size n

- the canonical statistic is the vector $(x_1, \ldots, x_{k-1})$ whose components are all but one of the category counts,
- the canonical parameter vector is $(\theta_1, \ldots, \theta_{k-1})$, where

$$\theta_i = \log\left(\frac{\pi_i}{\pi_k}\right), \tag{24}$$

- the mean value parameter vector is $(\mu_1, \ldots, \mu_{k-1})$, where $\mu_i = n\pi_i$,
- the canonical parameter space of the full family is the whole of $(k-1)$-dimensional Euclidean space,
- the mean value parameter space is the set of mean value parameter vectors that satisfy

$$\mu_i > 0, \qquad i = 1, \ldots k-1, \tag{25}$$

$$\sum_{i=1}^{k-1} \mu_i < n, \tag{26}$$

  and

- the cumulant function is

$$c(\theta) = n \log\left(1 + \sum_{i=1}^{k-1} e^{\theta_i}\right) \tag{27}$$

7.4.3.2 The Binomial Distribution

The binomial distribution is the $k = 2$ case of the multinomial distribution when the "try II" parameterization is used. If x_1 is the number of successes and x_2 is the number of failures, then we eliminate x_2 from the canonical statistic vector, just leaving the scalar variable x_1. Because the π's sum to one we have $\pi_2 = 1 - \pi_1$ and plugging this into (24) gives the analogous equation for the binomial distribution (9). Also (27) is the same as the analogous equation for the binomial distribution (13).

Try III

For the multinomial family of dimension k and sample size n

- the canonical statistic is the vector x of category counts

- the usual parameter is the vector π,
- the canonical parameter is the vector θ,
- the mean value parameter vector is $n\pi$, where the components of π are given by (23),
- the canonical parameter space of the full family is the whole of k-dimensional Euclidean space,
- the mean value parameter space of the full family is the set of all vectors μ whose components are strictly positive and sum to one,
- the cumulant function is

$$c(\theta) = n \log \left(\sum_{i=1}^{k} e^{\theta_i} \right) \tag{28}$$

- the canonical statistic vector satisfies the linear equality $x_1 + x_2 + \cdots + x_k = n$, hence (Section 6.1 above) the canonical parameterization is not identifiable, and
- if θ having components θ_i is a possible value of the canonical parameter vector, then so is $\theta + r$ having components $\theta_i + r$ for any real number r.

Commentary

This provides an example where we have three different ways to put the distribution in exponential family form and we can choose the one we like.

Try I seemed simple at first but is complicated by the nonlinear constraint on the parameters (18). Try II is ugly but simple to use. Try III is elegant but a little more complicated.

Try I is the special case of Try III where we impose the constraint (18). Try II is the special case of Try III where we impose the constraint $\theta_k = 0$ and then don't consider θ_k part of the canonical parameter vector and don't consider x_k part of the canonical statistic vector.

Try III is complicated by the fact that the mapping between canonical parameters and usual parameters (23) is not a one-to-one mapping (the "try III" parameterization is not identifiable). In order to get a one-to-one relationship we need to impose a constraint on the canonical parameters. We can choose the constraint to be anything convenient. We can use the ones already mentioned, (18) and $\theta_k = 0$. But we can also use others.

Try III is important because it keeps all the cell counts as components of the canonical statistic vector. This makes it much easier to reason about models. If we want to directly compare Poisson, multinomial, and product multinomial models, then we need to use the "try III" parameterization for the multinomial and product multinomial models in order for them to have the same the canonical parameter space as the Poisson model. In the complicated exponential family models that arise in aster models (Geyer, 2021; Geyer *et al.*, 2007), it is essential that the canonical statistic vector be the full vector of counts. Thus we have to use the "try III" parameterization.

We can explain the difference between try II and try III as follows. In try II we think that getting a one-to-one correspondence between usual parameters and canonical parameters is so important that we do it right away and make all the math and other reasoning about models very messy. In try III we think that getting a one-to-one correspondence is not important, and we do not allow it to mess up the math and other reasoning. We can impose a constraint to get one-to-one correspondence whenever we want.

7.5 The Normal Distribution

This section can be omitted. Its only purpose is to show an exponential family in which the "usual" statistic is not canonical.

The PDF of the univariate normal distribution is

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$$

from which we get log likelihood

$$\begin{aligned} l(\mu, \sigma) &= -\log(\sigma) - \frac{(x - \mu)^2}{2\sigma^2} \\ &= -\log(\sigma) - \frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} \end{aligned}$$

and we see this does have the exponential family form (1) with

- canonical statistic vector (x, x^2),
- canonical parameter vector θ with

$$\begin{aligned} \theta_1 &= \frac{\mu}{\sigma^2} \\ \theta_2 &= -\frac{1}{2\sigma^2} \end{aligned}$$

  and

- cumulant function

$$c(\theta) = \log(\sigma) + \frac{\mu^2}{2\sigma^2}$$

  where we still have to solve for μ and σ in terms of θ_1 and θ_2 to have this usable.

We won't bother to go any further because we are not going to use this theory of the normal distribution *as an exponential family*.

The only point was to have an example where the usual statistic x was not the canonical statistic vector (x, x^2). It is surprising that the usual statistic (a scalar) and the canonical statistic (a vector) don't even have the same dimension. Considered as an exponential family the univariate normal distribution isn't univariate. Its canonical statistic vector is two-dimensional. But then we see that it shouldn't be surprising because the canonical statistic vector and the canonical parameter vector always have the same dimension. So if we are going to have a two-dimensional canonical parameter vector (and we do need that because this is a two-parameter family of distributions), then we also must have a two-dimensional canonical statistic vector.

The multivariate normal distribution is also an exponential family, but the equations are even more complicated.

8 Sufficiency

8.1 Definition

Fisher (1922) invented the concept of sufficiency and sufficient statistics. A scalar or vector statistic (function of the data that does not depend on parameters) is *sufficient* if the the conditional distribution of the whole data given that statistic does not depend on parameters, in notation, if w is the whole data and y is the statistic, then y is sufficient if the conditional distribution of w given y does not depend on the parameters.

8.2 The Sufficiency Principle

Fisher argued that the sufficient statistic extracts all of the information in the data relevant to estimating the parameters. If y is sufficient and w is the whole data and we write joint equals marginal times conditional

$$f_\theta(w, y) = f(w \mid y) f_\theta(y),$$

only the marginal depends on the parameters. So Fisher argued that we should use that marginal to estimate parameters, because the conditional does not involve the parameters and hence has nothing to do with them.

Fisher's argument was later dubbed the *sufficiency principle:* statistical inference should only depend on the data through the sufficient statistics.

The likelihood may be written

$$L(\theta) = f_\theta(y)$$

(as usual we may drop multiplicative terms from the likelihood that do not contain parameters). Thus likelihood inference and Bayesian inference automatically obey the sufficiency principle. It is only ad hoc methods of estimation that may violate it.

8.3 The Neyman-Fisher Factorization Criterion

The definition of sufficient statistics is hard to apply (you have to factor the joint distribution into marginal times conditional), but the *Neyman-Fisher factorization criterion* (Fisher, 1922; Halmos and Savage, 1949; Neyman, 1935) is much easier to apply. This says a vector statistic is sufficient if and only if there is a version of the likelihood (possibly obtained by dropping some multiplicative factors that do not contain the parameters) or log likelihood (possibly obtained by dropping some additive factors that do not contain the parameters) that depends on the whole data only through that statistic.

From (1) and the Neyman-Fisher factorization criterion we see that the canonical statistic vector of an exponential family is always a sufficient statistic vector.

9 Sufficient Dimension Reduction

Nowadays, there is much interest in sufficient dimension reduction in regression that does not fit into the exponential family paradigm described in Section 9.3 below. But exponential families were there first.

9.1 Independent and Identically Distributed Sampling

Suppose y_1, y_2, ..., y_n are independent and identically distributed (IID) random variables from an exponential family with log likelihood for sample size one (1). Then the log likelihood for sample size n is

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

From this it follows that IID sampling converts one exponential family into another exponential family with

- canonical statistic vector $\sum_i y_i$, which is the sum of the canonical statistic vectors for the samples,
- canonical parameter vector θ, which is the same as the canonical parameter vector for the samples,
- cumulant function $nc(\,\cdot\,)$, which is n times the cumulant function for the samples,
- mean value parameter vector $n\mu$, which is n times the mean value parameter μ for the samples.

Many familiar "addition rules" for brand name distributions are special cases of this

- sum of IID binomial is binomial,
- sum of IID Poisson is Poisson,
- sum of IID negative binomial is negative binomial,
- sum of IID gamma is gamma,
- sum of IID multinomial is multinomial, and
- sum of IID multivariate normal is multivariate normal.

The point is that the dimension reduction from y_1, y_2, ..., y_n to $\sum_i y_i$ is a *sufficient dimension reduction*. It loses no information about the parameters assuming the model is correct.

9.2 Product Models

What about independent but not identically distributed? Models covered in this course having this property are Poisson sampling (the joint distribution is the product of Poisson distributions), product multinomial sampling, Poisson regression with log link (a special case of Poisson sampling), and logistic regression (a special case of product multinomial).

Suppose we have a product model in which the total data vector y has components y_A for A in some partition $\mathcal{A}$ of the index set of y. And the y_A, $A \in \mathcal{A}$ are independent random vectors. And for each $A \in \mathcal{A}$ the subvector y_A is the canonical statistic vector of an exponential family model with canonical parameter vector θ_A and cumulant function c_A. Then the log likelihood for the whole model is

$$\sum_{A \in \mathcal{A}} [\langle y_A, \theta_A \rangle - c_A(\theta_A)] = \langle y, \theta \rangle - c(\theta)$$

where

$$c(\theta) = \sum_{A \in \mathcal{A}} c_A(\theta_A)$$

So not only are the marginal distributions of the y_A exponential family, so is the joint distribution of y.

If each of the exponential families for the y_A are regular and full, then so is the exponential family for y.

There is no sufficient dimension reduction here. The sufficient statistic vector is y, which is sometimes the whole data. But we combine this idea with the idea in the following section.

9.3 Canonical Affine Submodels

9.3.1 Definition

Suppose we parameterize a submodel of our exponential family with parameter transformation

$$\theta = a + M\beta \tag{29}$$

where

- a is a known vector, usually called the *offset vector*,
- M is a known matrix, usually called the *model matrix* (also called the *design matrix*),
- θ is the canonical parameter vector of the original family, and
- β is the *canonical parameter vector* of the *canonical affine submodel* (also called *coefficients vector*).

The terms *offset vector*, *model matrix*, and *coefficients* are those used by R functions `lm` and `glm`. The term *design matrix* is widely used although it doesn't make much sense for data that do not come from a designed experiment (but language doesn't have to make sense and often doesn't). The terminology *canonical affine submodel* is from Geyer *et al.* (2007).

9.3.2 Linear versus Affine

For a linear model (fit by R function `lm`) θ is the mean vector $\theta = E_\theta(y)$. For a generalized linear model (fit by R function `glm`) θ is the so-called the *linear predictor*, and it not usually called a parameter, even though it is a parameter. The transformation (29) is usually called "linear", but Geyer *et al.* (2007) decided to call it "affine". The issue is that there are two meanings of linear

- in calculus and all mathematics below that level (including before college) a linear function is a function whose graph is a straight line, but

- in linear algebra and all mathematics above that level (including real analysis and functional analysis, which are just advanced calculus with another name) a linear function is a function that preserves vector addition and scalar multiplication, in particular, if f is a linear function, then $f(0) = 0$.

In linear algebra and all mathematics above that level, if we need to refer to the other notion of linear function we call it an *affine function*. An affine function is a linear function plus a constant function, where "linear" here means the notion from linear algebra and above.

All of this extends to arbitrary transformations between vector spaces. An affine function from one vector space to another is a linear function plus a constant function. Going the other way, a linear function is an affine function f that satisfies $f(0) = 0$.

So (29) is an *affine change of parameter* in the language of linear algebra and above and a *linear change of parameter* in the language of calculus and below. It is also a *linear change of parameter* in the language of linear algebra and above in the special case $a = 0$ (no offset). The fact that (29) is almost always used when $a = 0$ (offsets are very rarely used) may contribute to the tendency to call this parameter transformation "linear".

It is not just in applications that offsets rarely appear. Even theoreticians who pride themselves on their knowledge of advanced mathematics usually ignore offsets. The familiar formula $\hat{\beta} = (M^T M)^{-1} M^T y$ for the least squares estimator is missing an offset a.

Another reason for confusion between the two notions of "linear" is that for simple linear regression (R command `lm(y ~ x)`), the *regression function* is affine (linear in the lower-level notion). But this is applying "linear" in the wrong context.

> It's called "linear regression" because it's linear in the regression coefficients, not because it's linear in x.
>
> — Werner Stutzle

If we change the model to quadratic regression (R command `lm(y ~ x + I(x^2))`), then the regression function is quadratic (nonlinear) but the model is still a *linear model* fit by R function `lm`. Another way of saying this is some people think of simple linear regression as being linear in the lower-level sense because it has an intercept, but, as sophisticated statisticians, we know that having an intercept does not put an a in equation (29), it adds a column to M (all of whose components are equal to one).

Statisticians generally ignore this confusion in terminology. Most clients of statistics, including most scientists, do not take linear algebra and math classes beyond that, so we statisticians use "linear" in the lower-level sense when talking to clients. I myself use "linear" in the lower-level sense in Stat 5101 and 5102, a master's level service course in theoretical probability and statistics. Except we are inconsistent. When we say "linear model" we usually mean (29) with $a = 0$ so $(M^T M)^{-1} M^T y$ makes sense, and that is the higher-level sense of "linear".

Hence Geyer *et al.* (2007) decided to introduce the term *canonical affine submodel* for what was already familiar but either had no name or was named with confusing terminology.

9.3.3 Conditioning on Covariates

In the list following (29), "known" means nonrandom. In regression analysis we allow M to depend on covariate data, and saying M is nonrandom means we are treating covariate data as fixed. If the covariate data happen to be random, we say we are doing the analysis conditional on the observed values of the covariate data (which is the same as treating these data as fixed and nonrandom). In other words, the statistical model is for the conditional distribution of the response y given the covariate data, and the (marginal) distribution of the covariate data *is not modeled*. Thus to be fussily pedantic we should write

$$E_\theta(y \mid \text{the part of the covariate data that is random, if any})$$

everywhere instead of $E_\theta(y)$, and similarly for $\operatorname{var}_\theta(y)$ and so forth. But we are not going to do that, and almost no one does that. We can also allow a to depend on covariate data (but almost no one does that).

9.3.4 Sufficient Dimension Reduction

Now we come to the point of this section. The log likelihood for the canonical affine submodel is

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

and we may drop the term $\langle y, a \rangle$ from the log likelihood because it does not contain the parameter β giving

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

and because

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

we finally get log likelihood for the canonical affine submodel

$$l(\beta) = \langle M^T y, \beta \rangle - c_{\text{sub}}(\beta) \tag{30}$$

where

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

From this it follows that the change of parameter (29) converts one exponential family into another exponential family with

- canonical statistic vector $M^T y$,
- canonical parameter vector β,
- cumulant function c_{sub}, and
- mean value parameter $\tau = M^T \mu$, where μ is the mean value parameter of the original model.

If Θ is the canonical parameter space of the original model, then

$$B = \{ \beta : a + M\beta \in \Theta \}$$

is the canonical parameter space of the canonical affine submodel. If the original model is full, then so is the canonical affine submodel. If the original model is regular full, then so is the canonical affine submodel.

There are many points to this section. It is written the way it is because of aster models (Geyer *et al.*, 2007), but it applies to linear models, generalized linear models, and log-linear models for categorical data too, hence to the majority of applied statistics. But the point of this section that gets it put in its supersection is that the dimension reduction from y to $M^T y$ is a *sufficient dimension reduction*. It loses no information about β assuming this submodel is correct.

9.4 The Pitman–Koopman–Darmois Theorem

Nowadays, exponential families are so important in so many parts of theoretical statistics that their origin has been forgotten. They were invented to be the class of statistical models described by the Pitman–Koopman–Darmois theorem, which says (roughly) that the only statistical models having the sufficient dimension reduction property in IID sampling described in Section 9.1 above are exponential families. In effect, it turns that section from if to if and only if.

But the reason for the "roughly" is that as just stated with no conditions, the theorem is false. Here is a counterexample. For an IID sample from the Uniform$(0, \theta)$ distribution, the maximum likelihood estimator (MLE) is $\widehat{\theta}_n = \max\{y_1, \ldots, y_n\}$ and this is easily seen to be a sufficient statistic (the Neyman-Fisher factorization theorem again) but Uniform$(0, \theta)$ is not an exponential family.

So there have to be side conditions to make the theorem true. Pitman, Koopman, and Darmois independently (not as co-authors) in 1935 and 1936 published theorems that said under two side conditions,

- the distribution of the canonical statistic is continuous and

- the support of the distribution of the canonical statistic does not depend on the parameter,

then any statistical model with the sufficient dimension reduction property in IID sampling is an exponential family. (Obviously, Uniform$(0, \theta)$ violates the second side condition).

Later, other authors published theorems with more side conditions that covered the discrete case. But the side conditions for those are really messy so those theorems are not so interesting.

Nowadays exponential families are so important for so many reasons (not all of them mentioned in this document) that no one any longer cares about these theorems. We only want the if part of the if and only if, which is covered in Section 9.1 above.

10 Observed Equals Expected

The usual procedure for maximizing the log likelihood is to differentiate it, set it equal to zero, and solve for the parameter. The derivative of (1) is

$$\nabla l(\theta) = y - \nabla c(\theta) = y - E_\theta(Y) \tag{31}$$

(Here and throughout this section Y is a random vector having the distribution of the canonical statistic and y is the observed value of the canonical statistic.) So the MLE $\hat{\theta}$ is characterized by

$$y = E_{\hat{\theta}}(Y) \tag{32}$$

We can say a lot more than this. Cumulant functions are convex functions (Barndorff-Nielsen, 1978, Theorem 7.1). Hence log likelihoods of regular full exponential families are concave functions that are differentiable everywhere in the canonical parameter space. Hence (31) being equal to zero is a necessary and sufficient condition for θ to maximize (1) (Rockafellar and Wets, 1998, part (b) of Theorem 2.14). Hence (32) is a necessary and sufficient condition for $\hat{\theta}$ to be an MLE.

The MLE need not exist and need not be unique.

The MLE does not exist if the observed value of the canonical statistic is on the boundary of its support in the following sense, there exists a vector $\delta \neq 0$ such that $\langle Y - y, \delta \rangle \leq 0$ holds almost surely and $\langle Y - y, \delta \rangle = 0$ does not hold almost surely (Geyer, 2009, Theorems 1, 3, and 4). When the MLE does not exist in the classical sense, it may exist in an extended sense as a limit of distributions in the original family, but that is a story for another time (see Geyer (2009) and Geyer (2016a) for more on this subject).

The MLE is not unique if there exists a $\delta \neq 0$ such that $\langle Y - y, \delta \rangle = 0$ holds almost surely (Geyer, 2009). This is the same phenomenon described in Section 6.1 above. In theory, nonuniqueness of the MLE for a full exponential family is not a problem because every MLE corresponds to the same probability distribution (Geyer, 2009, Theorem 1 and Corollary 2), so the MLE canonical parameter vector is not unique (if any MLE exist), but the MLE probability distribution is unique (if it exists) and the MLE mean value parameter is unique (if it exists).

It is always possible to arrange for uniqueness of the MLE. Simply arrange that the distribution of the canonical statistic have full dimension (not be concentrated on a hyperplane).

But one does not want to do this too early in the data analysis process. In the lecture notes about Bayesian inference and Markov chain Monte Carlo (Geyer, 2020a) we used a nonidentifiable parameterization for the example (volleyball standings) for the frequentist analysis and this caused no problems because the computer (R function `glm`) knew how to deal with it. This same nonidentifiable parameterization when used in the Bayesian analysis allowed us to easily specify a prior distribution that did not favor any team.

This sort of data provides simple examples of when the MLE does not exist in the classical sense. In the example cited above one team did not win any matches, and the MLE for that team consequently does not exist. We set its MLE to be minus infinity, but that is a mathematical fiction. Minus infinity is not an allowed parameter value for an exponential family (not a real number).

But all of this about nonexistence and nonuniqueness of the MLE is not the main point of this section. The main point is that (32) characterizes the MLE when it exists and whether or not it is unique.

- The MLE in an exponential family satisfies "observed equals expected". The MLE for the mean value parameter vector satisfies $\hat{\mu} = y$ or $y = E_{\hat{\theta}}(y)$.

More precisely, we should say the observed value of the *canonical statistic vector* equals its expected value under the MLE distribution (which is unique if it exists). It is not the observed value of anything whatsoever that equals its MLE expected value.

The observed-equals-expected property is one of the keys to interpreting MLE's for exponential families. Strangely, this is considered very important in some areas of statistics and not mentioned at all in other areas.

In categorical data analysis, it is considered key. The MLE for a hierarchical log-linear model for categorical data satisfies observed equals expected: the marginals of the table corresponding to the terms in the model are equal to their MLE expected values, and those marginals are the canonical sufficient statistics. So this gives a complete characterization of maximum likelihood for these models, and hence a complete understanding in a sense. (See also Section 12 below.)

In regression analysis, it is ignored. The most widely used linear and generalized linear models are exponential families: linear models, logistic regression, and Poisson regression with log link. Thus maximum likelihood satisfies the observed-equals-expected property.

- The MLE for a canonical affine submodel satisfies "observed equals expected". If y is the response vector and M is the model matrix, then the MLE for the submodel mean value parameter vector satisfies $\hat{\tau} = M^T y$ or $M^T y = M^T E_{\hat{\beta}}(y)$.

More precisely, we should say the observed value of the *submodel canonical statistic vector* equals its expected value under the MLE distribution (which is unique if it exists). It is not the observed value of anything whatsoever that equals its MLE expected value.

So this tells us that the submodel canonical statistic vector $M^T y$ is crucial to understanding linear and generalized linear models (and aster models, Geyer *et al.* (2007)) just like it is for hierarchical log-linear models for categorical data. But do regression books even mention this? Not that your humble author knows of (except for indirectly using it in the proof of Theorem 17.1 below, which some regression textbooks do mention).

Let's check this for linear models with no offset where we have original model mean value parameter

$$\mu = E(y) = M\beta$$

and MLE for the submodel canonical parameter β

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

and consequently

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

and by invariance of maximum likelihood (slides 100 ff. of deck 3 of Geyer (2016b))

$$\hat{\mu} = M\hat{\beta}$$

so

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

which we claim is the key to understanding linear models. But regression textbooks never mention it. So who is right, the authors of regression textbooks or the authors of categorical data analysis textbooks? Our answer is the latter. $M^T y$ is important.

Another way people sometimes say this is that every MLE in a regular full exponential family is a method of moments estimator, but not just any old method of moments estimator. It is the method of moments

estimator that sets the expectation of the *canonical statistic vector* equal to its observed value and solves for the parameter. For example, for linear models, the method of moments estimator we want sets

$$M^T M\beta = M^T y$$

and solves for β. And being precise, we need the method of moments estimator that sets the *canonical statistic vector for the model being analyzed* equal to its expected value. For a canonical affine submodel, that is the *submodel* canonical statistic vector $M^T y$.

But there is nothing special here about linear models except that they have a closed form expression for $\widehat{\beta}$. In general, we can only determine $\widehat{\beta}$ as a function of $M^T y$ by numerically maximizing the likelihood using a computer optimization function. But we always have "observed equals expected" up to the inaccuracy of computer arithmetic.

And usually "observed equals expected" is the only simple equality we know about maximum likelihood in regular full exponential families.

11 Maximum Likelihood Estimation

When the maximum likelihood estimate exists and is unique, Theorem 6.3 above says that any algorithm that always goes uphill on the log likelihood when it can will converge to the MLE. This includes a lot of algorithms including most of the optimization algorithms in core R.

We do not need to worry about multiple local maxima and so do not ever need to use the default method of R function `optim` or `method = "SANN"` of that function. Other methods of R function `optim` always go uphill if they can. So do R functions `nlm` and `nlminb` except that R function `nlm` has no options for doing maximization rather than minimization so with it one must give it negative log likelihood as the function to minimize.

The point of our Theorem 6.3 is that one does not need a good starting point for the optimization (as one does for maximum likelihood estimation in models that are not regular full exponential families). For regular full exponential families, any starting point works.

12 Maximum Entropy

Many scientists in the early part of the nineteenth century invented the science of thermodynamics, in which some of the key concepts are *energy* and *entropy*. Entropy was initially defined physically as

$$dS = \frac{dQ}{T}$$

where S is entropy and dS its differential, Q is heat and dQ its differential, and T is temperature, so to calculate entropy in most situations you have to do an integral (the details here don't matter — the point is that entropy defined this way is a physical quantity measured in physical ways).

The first law of thermodynamics says energy is conserved in any closed physical system. Energy can change form from motion to heat to chemical energy and to other forms. But the total is conserved.

The second law of thermodynamics says entropy is nondecreasing in any closed physical system. But there are many other equivalent formulations. One is that heat always flows spontaneously from hot to cold, never the reverse. Another is that there is a maximum efficiency of a heat engine or a refrigerator (a heat engine operated in reverse) that depends only on the temperature difference that powers it (or that the refrigerator produces).

So, somewhat facetiously, the first law says you can't win, and the second law says you can't break even.

Near the end of the nineteenth century and the beginning of the twentieth century, thermodynamics was extended to chemistry. And it was found that chemistry too obeys the laws of thermodynamics. Every

chemical reaction in your body is all the time obeying the laws of thermodynamics. No animal can convert all of the energy of food to useful work. There must be waste heat, and this is a consequence of the second law of thermodynamics.

Also near the end of the nineteenth century Ludwig Boltzmann discovered the relationship between entropy and probability. He was so pleased with this discovery that he had

$$S = k \log W$$

engraved on his tombstone. Here S is again entropy, k is a physical constant now known as Boltzmann's constant, and W is probability (*Wahrscheinlichkeit* in German). Of course, this is not probability in general, but probability in certain physical systems.

Along with this came the interpretation that entropy does not always increase. Physical systems necessarily spend more time in more probable states and less time in less probable states. Increase of entropy is just the inevitable move from less probable to more probable on average. At the microscopic level entropy fluctuates as the system moves through each state according to its probability.

In mid twentieth century Claude Shannon recognized the relation between entropy and information. The same formulas that define entropy statistically define information as negative entropy (so minus a constant times log probability). He used this to bound how much signal could be put through a noisy communications channel.

A little later Kullback and Leibler imported Shannon's idea into statistics, defining what we now call Kullback-Leibler information.

What does maximum likelihood try to do theoretically? It tries to maximize the expectation of the log likelihood function, which is the Kullback-Leibler information function, that maximum being the true unknown parameter value if the model is identifiable (Wald, 1949). There is also a connection between Kullback-Leibler information and Fisher information.

A little later Edwin Jaynes recognized the connection between entropy or negative Kullback-Leibler information and exponential families. Exponential families maximize entropy subject to constraints. Fix a probability distribution Q and a random vector Y on the probability space of that distribution. Then for each vector μ find the probability distribution P that maximizes entropy (minimizes Kullback-Leibler information) with respect to Q subject to $E_P(Y) = \mu$. If the maximum exists, call it P_μ. Then the collection of all such P_μ is a full exponential family having canonical statistic Y and mean value parameter vector μ (for a proof see Geyer (2018) Deck 2, Slides 176 ff.).

Jaynes is not popular among statisticians because his maximum entropy idea became linked with so-called maxent modeling which statisticians for the most part have ignored.

But in the context of exponential families, maximum entropy is powerful. It says you start with the canonical statistic. If you start with a canonical statistic that is an affine function of the original canonical statistic of an exponential family, then the canonical affine submodel maximizes entropy subject to the distributions in the canonical affine submodel having the mean value parameters they do. Every other aspect of those distributions is just randomness in the sense of maximum entropy or minimum Kullback-Leibler information. Thus the (submodel) mean value parameter tells you everything interesting about a canonical affine submodel.

When connected with observed equals expected (Section 10 above), this is a very powerful principle. Observed equals expected says maximum likelihood estimation matches exactly the submodel canonical statistic vector to its observed value. Maximum entropy says nothing else matters, everything important, all the *information* about the parameter is in the MLE. All else is randomness (in the sense of maximum entropy).

The sufficiency principle (Section 8.2 above) also says nothing else matters, all the information about the parameter is in the MLE, but the maximum entropy principle says more: the distributions of the maximum entropy submodel are completely determined by the maximum entropy principle.

Admittedly the one time I have made this argument in print (Geyer and Thompson, 1992) it was not warmly received. But it was a minor point of that article. Perhaps this section makes a better case.

This is the reason why the model for the MCMC and volleyball example is what it is. The official Big Ten conference standings only pay attention to team wins. Thus we use them as canonical statistics. It is true that we add one statistic the conference does not use, the total number of home wins, because we want home field advantage in the model because it obviously exists (it is highly statistically significant every year in every sport), and leaving out home field advantage would inflate the variance of estimates.

It is amazing (to me) that this procedure fully and correctly adjusts sports standings for strength of schedule. It is strange (to me) that only one sport, college ice hockey, uses this procedure, which in that context they call KRACH, and they do not use it alone, but as just one factor in a mess of procedures that have no statistical justification.

13 Multivariate Monotonicity

A link function, which goes componentwise from mean value parameters to canonical parameters for a generalized linear model that is an exponential family (linear models, logistic regression, Poisson regression with log link) is *univariate monotone*.

This does not generalize to exponential families with dependence among components of the response vector like aster models (Geyer *et al.*, 2007), Markov spatial point processes (Geyer and Møller (1994) and Geyer (2020c)), Markov spatial lattice processes (Geyer, 2020b) or even to log-linear models for contingency tables when multinomial or product multinomial sampling is assumed.

Instead we have *multivariate monotonicity*. This is not a concept statisticians are familiar with. It does not appear in real analysis, functional analysis, or probability theory. It comes from convex analysis. Rockafellar and Wets (1998) have a whole chapter on the subject (their Chapter 12). There are many equivalent characterizations. We will only discuss two of them.

A function f from one vector space to another is *multivariate monotone* if

$$\langle f(x) - f(y), x - y \rangle \geq 0, \qquad \text{for all } x \text{ and } y$$

and *strictly multivariate monotone* if

$$\langle f(x) - f(y), x - y \rangle > 0, \qquad \text{for all } x \text{ and } y \text{ such that } x \neq y$$

(Rockafellar and Wets (1998), Definition 12.1).

The reason this is important to us is that the gradient mapping of a convex function is multivariate monotone (Rockafellar and Wets, 1998, Theorem 12.17, indeed a differentiable function is convex if and only if its gradient mapping is multivariate monotone). We have differentiable convex functions in play: cumulant functions (Barndorff-Nielsen, 1978, Theorem 7.1) Also cumulant functions of regular full exponential families are differentiable everywhere in their canonical parameter spaces (3).

So define h by (7), the function that maps canonical parameter vector values to mean value parameter vector values. Then h is multivariate monotone. Hence if θ_1 and θ_2 are canonical parameter values and μ_1 and μ_2 are the corresponding mean value parameter values

$$\langle \mu_1 - \mu_2, \theta_1 - \theta_2 \rangle \geq 0$$

Moreover if the canonical parameterization is identifiable, h is *strictly multivariate monotone*

$$\langle \mu_1 - \mu_2, \theta_1 - \theta_2 \rangle > 0, \qquad \theta_1 \neq \theta_2$$

(Barndorff-Nielsen (1978), Theorem 7.1; Geyer (2009), Theorem 1; Rockafellar and Wets (1998), Theorem 12.17).

We can see from the way the canonical and mean value parameters enter symmetrically, that when the canonical parameterization is identifiable so h is invertible (Geyer, 2013a, Lemma 9) the inverse h^{-1} is also *strictly multivariate monotone*

$$\langle \mu_1 - \mu_2, \theta_1 - \theta_2 \rangle > 0, \qquad \mu_1 \neq \mu_2$$

One final characterization: a differentiable function is strictly multivariate monotone if and only if the restriction to every line segment in the domain is strictly univariate monotone (obvious from the way the definitions above only deal with two points in the domain at a time).

Thus we have a "dumbed down" version of strict multivariate monotonicity: increasing one component of the canonical parameter vector increases the corresponding component of the mean value parameter vector, if the canonical parameterization is identifiable. The other components of μ also change but can go any which way.

When specialized to canonical affine submodels (Section 9.3 above) strict multivariate monotonicity becomes

$$\langle \tau_1 - \tau_2, \beta_1 - \beta_2 \rangle > 0, \qquad \beta_1 \neq \beta_2$$

where τ_1 and τ_2 are the submodel mean value parameters corresponding to the submodel canonical parameters β_1 and β_2. When "dumbed down" this becomes: increasing one component of the submodel canonical parameter vector β increases the corresponding component of the submodel mean value parameter vector $\tau = M^T \mu$, if the submodel canonical parameterization is identifiable. The other components of τ and components of μ also change but can go any which way.

Again we see the key importance of the sufficient dimension reduction map $y \mapsto M^T \mu$ and the corresponding original model to canonical affine submodel mean value parameter mapping $\mu \mapsto M^T \mu$, that is, the importance of thinking of M^T as (the matrix representing) a linear transformation.

These "dumbed down" characterizations say that strict multivariate monotonicity implies strict univariate monotonicity of the restrictions of the function h to line segments in the domain *parallel to the coordinate axes* (so only one component of the vector changes).

Compare this with our last (not dumbed down) characterization: strict multivariate monotonicity holds *if and only if* all restrictions of the function h to line segments in the domain are strictly univariate monotone (not just line segments parallel to the coordinate axes, *all* line segments).

So the "dumbed down" version only varies one component of the canonical parameter at a time, whereas the non-dumbed-down version varies all components. The "dumbed down" version can be useful when talking to people who have never heard of multivariate monotonicity. But sometimes the non-dumbed-down concept is needed (Shaw and Geyer, 2010, Appendix A). There is no substitute for understanding this concept. It should be in the toolbox of every statistician.

14 Regression Coefficients are Meaningless

The title of this section comes from my Stat 5102 lecture notes (Geyer, 2016b), deck 5, slide 19. It is stated the way it is for shock value. All of the students in that class have previously taken courses where they were told how to interpret regression coefficients. So this phrase is intended to shock them into thinking they have been mistaught!

Although shocking, it refers to something everyone knows. Even in the context of linear models (which those 5102 notes are) the same model can be specified by different formulas or different model matrices.

14.1 Example: Polynomial Regression

For example

```
foo <- read.table("http://www.stat.umn.edu/geyer/5102/data/ex5-1.txt",
    header = TRUE)
lout1 <- lm(y ~ poly(x, 2), data = foo)
summary(lout1)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 2), data = foo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2677  -2.7246   0.4333   3.6335   9.0588
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.762      0.846   24.54  < 2e-16 ***
## poly(x, 2)1   74.620      5.351   13.95 2.62e-16 ***
## poly(x, 2)2   -7.065      5.351   -1.32    0.195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.351 on 37 degrees of freedom
## Multiple R-squared:  0.8414, Adjusted R-squared:  0.8328
## F-statistic: 98.11 on 2 and 37 DF,  p-value: 1.613e-15
```

and

```
lout2 <- lm(y ~ poly(x, 2, raw = TRUE), data = foo)
summary(lout2)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 2, raw = TRUE), data = foo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2677  -2.7246   0.4333   3.6335   9.0588
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.883215   2.670473  -1.080    0.287
## poly(x, 2, raw = TRUE)1   1.406719   0.300391   4.683 3.74e-05 ***
## poly(x, 2, raw = TRUE)2 -0.009381   0.007105  -1.320    0.195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.351 on 37 degrees of freedom
## Multiple R-squared:  0.8414, Adjusted R-squared:  0.8328
## F-statistic: 98.11 on 2 and 37 DF,  p-value: 1.613e-15
```

have different fitted regression coefficients. But they fit the same model

```
all.equal(fitted(lout1), fitted(lout2))
```

```
## [1] TRUE
```

14.2 Example: Categorical Predictors

For another example, when there are categorical predictors we must "drop" one category from each predictor to get an identifiable model and which one we drop is arbitrary. Thus

```
bar <- read.table("http://www.stat.umn.edu/geyer/5102/data/ex5-4.txt",
    header = TRUE, stringsAsFactors = TRUE)
levels(bar$color)
```

```
## [1] "blue"  "green" "red"
```

```
lout1 <- lm(y ~ x + color, data = bar)
summary(lout1)
```

```
##
## Call:
## lm(formula = y ~ x + color, data = bar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2398  -2.9939   0.1725   3.5555  11.9747
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.16989    1.01710  12.948  < 2e-16 ***
## x            1.00344    0.02848  35.227  < 2e-16 ***
## colorgreen   2.12586    1.00688   2.111   0.0364 *
## colorred     6.60586    1.00688   6.561  8.7e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.034 on 146 degrees of freedom
## Multiple R-squared:  0.898,  Adjusted R-squared:  0.8959
## F-statistic: 428.6 on 3 and 146 DF,  p-value: < 2.2e-16
```

and

```
bar <- transform(bar, color = relevel(color, ref = "red"))
lout2 <- lm(y ~ x + color, data = bar)
summary(lout2)
```

```
##
## Call:
## lm(formula = y ~ x + color, data = bar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2398  -2.9939   0.1725   3.5555  11.9747
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.77575    1.01710  19.443  < 2e-16 ***
## x            1.00344    0.02848  35.227  < 2e-16 ***
## colorblue   -6.60586    1.00688  -6.561  8.7e-10 ***
## colorgreen  -4.48000    1.00688  -4.449  1.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 5.034 on 146 degrees of freedom
## Multiple R-squared:  0.898,  Adjusted R-squared:  0.8959
## F-statistic: 428.6 on 3 and 146 DF,  p-value: < 2.2e-16
```

have different fitted regression coefficients (even the names of the regression coefficients are different). But they fit the same model

```
all.equal(fitted(lout1), fitted(lout2))
```

```
## [1] TRUE
```

14.3 Example: Collinearity

Even in the presence of collinearity, where some coefficients must be dropped to obtain identifiability (and which one(s) are dropped is arbitrary) the mean values are unique, hence the fitted model is unique.

```
baz <- read.table("http://www.stat.umn.edu/geyer/5102/data/ex5-3.txt",
    header = TRUE, stringsAsFactors = TRUE)
x3 <-  with(baz, x1 + x2)
lout1 <- lm(y ~ x1 + x2 + x3, data = baz)
summary(lout1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = baz)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89036 -0.67352 -0.05958  0.69110  2.06976
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4342     0.9094   0.477 0.635232
## x1            1.4179     0.1719   8.247 1.09e-10 ***
## x2            0.6743     0.1688   3.993 0.000227 ***
## x3                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9722 on 47 degrees of freedom
## Multiple R-squared:  0.7388, Adjusted R-squared:  0.7277
## F-statistic: 66.48 on 2 and 47 DF,  p-value: 1.987e-14
```

and

```
lout2 <- lm(y ~ x3 + x2 + x1, data = baz)
summary(lout2)
```

```
##
## Call:
## lm(formula = y ~ x3 + x2 + x1, data = baz)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89036 -0.67352 -0.05958  0.69110  2.06976
##
```

```
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4342     0.9094   0.477   0.6352
## x3            1.4179     0.1719   8.247 1.09e-10 ***
## x2           -0.7436     0.2859  -2.601   0.0124 *
## x1                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9722 on 47 degrees of freedom
## Multiple R-squared:  0.7388, Adjusted R-squared:  0.7277
## F-statistic: 66.48 on 2 and 47 DF,  p-value: 1.987e-14
```

have different fitted regression coefficients. But they fit the same model

```
all.equal(fitted(lout1), fitted(lout2))
```

```
## [1] TRUE
```

14.4 Alice in Wonderland

After several iterations, this shocking advice became the following (Geyer (2018), deck 2, slide 41)

> A quote from my master's level theory notes.
>
> > Parameters are meaningless quantities. Only probabilities and expectations are meaningful.
>
> Of course, some parameters are probabilities and expectations, but most exponential family canonical parameters are not.
>
> A quote from *Alice in Wonderland*
>
> > 'If there's no meaning in it,' said the King, 'that saves a world of trouble, you know, as we needn't try to find any.'
>
> Realizing that canonical parameters are meaningless quantities "saves a world of trouble". We "needn't try to find any".

Thinking sophisticatedly and theoretically, of course parameters are meaningless. A statistical model is a family $\mathcal{P}$ of probability distributions. How this family is parameterized (indexed) is meaningless. If

$$\mathcal{P} = \{ P_\theta : \theta \in \Theta \} = \{ P_\beta : \beta \in B \} = \{ P_\varphi : \varphi \in \Phi \}$$

are three different parameterizations for the same model, then they are all the same model (duh!). The fact that parameter estimates in one parameterization tell us nothing about estimates in another parameterization tells us nothing.

But probabilities and expectations are meaningful. For $P \in \mathcal{P}$, both $P(A)$ and $E_P\{g(Y)\}$ depend only on P not what parameter value is deemed to index it. And this does not depend on what P means, whether we specify P with a probability mass function, a probability density function, a distribution function, or a probability measure, the same holds: probabilities and expectations only depend on the distribution, not how it is described.

Even if we limit the discussion to regular full exponential families, any one-to-one affine function of a canonical parameter vector is another canonical parameter vector (copied from Section 3 above). That's a lot of parameterizations, and which one you choose (or the computer chooses for you) is meaningless.

Hence we agree with the King of Hearts in *Alice in Wonderland*. It "saves a world of trouble" if we don't try to interpret canonical parameters.

It doesn't help those wanting to interpret canonical parameters that sometimes the map from canonical to mean value parameters has no closed-form expression (this happens in spatial point and lattice processes (Geyer, 2020b) (Geyer, 2020c); the log likelihood and its derivatives can only be approximated by MCMC using the scheme in Geyer and Thompson (1992) and Geyer (1994) or has a closed-form expression, but it is so fiendishly complicated that people have no clue what is going on, although the computer chugs through the calculation effortlessly (this happens with aster models, Geyer *et al.* (2007)).

15 Interpreting Exponential Family Model Fits

We take up the points made above in turn, stressing their impact on how users can interpret exponential family model fits.

15.1 Observed Equals Expected

The simplest and most important property is the observed-equals-expected property (Section 10 above).

The MLE for the submodel mean value parameter vector $\hat{\tau} = M^T \hat{\mu}$ is exactly equal to the submodel canonical statistic vector $M^T y$. That's what maximum likelihood in a regular full exponential family *does*.

So understanding $M^T y$ is the most important thing in understanding the model. If $M^T y$ is scientifically (business analytically, sports analytically, whatever) interpretable, then the model is interpretable. Otherwise, not!

15.2 Sufficient Dimension Reduction

The next most important property is sufficient dimension reduction (Section 9.3 above).

The submodel canonical statistic vector $M^T y$ is *sufficient*. It contains all the information about the parameters that there is in the data, assuming the model is correct.

Since $M^T y$ determines the MLE for the coefficients vector $\hat{\beta}$ (Section 9.3 above, assuming β is identifiable), and the MLE for every other parameter vector is a one-to-one function of $\hat{\beta}$, the MLE's for all parameter vectors ($\hat{\beta}$, $\hat{\theta}$, $\hat{\mu}$, and $\hat{\tau}$) are sufficient statistic vectors. The MLE for each parameter vector contains all the information about parameters that there is in the data, assuming the model is correct.

15.3 Maximum Entropy

And nothing else matters for interpretation.

Everything else about the model other than what the MLE's say is as random as possible (maximum entropy, Section 12 above) and contains no information (sufficiency, just discussed).

15.4 Regression Coefficients are Meaningless

In particular it "saves a world of trouble" if we realize "we needn't try to find any" meaning in coefficients vector $\hat{\beta}$ (Section 14.4 above).

15.5 Multivariate Monotonicity

But if we do have to say something about the coefficients vector $\hat{\beta}$ we do have the multivariate-monotonicity property available (Section 13 above).

15.6 The Model Equation

Most statistics courses that discuss the regression models teach students to woof about the *model equation* (29). In lower-level courses where students are not expected to understand matrices, students are taught to

woof about the same thing in other terms

$$y_i = \beta_1 + \beta_2 x_i + \text{error}$$

and the like. That is, they are taught to think about the model matrix as a linear operator $\beta \mapsto M\beta$ or the same thing in other terms. And another way of saying this is that they are taught to focus on the *rows* of M.

The view taken here is that this woof is all meaningless because it is about meaningless parameters (β and θ). The important linear operator to understand is the sufficient dimension reduction operator $y \mapsto M^T y$ or, what is the same thing described in other language, the original model to submodel mean value transformation operator $\mu \mapsto M^T \mu$. And another way of saying this is that we should focus on the *columns* of M.

It is not when you woof about $M\beta$ that you understand and explain the model, it is when you woof about $M^T y$ that you understand and explain the model.

16 Asymptotics

A story: when I was a first year graduate student I answered a homework question about asymptotics with "because it's an exponential family" but the teacher didn't think that was quite enough explanation. It is enough, but textbooks and courses don't emphasize this.

The "usual" asymptotics of maximum likelihood (asymptotically normal, variance inverse Fisher information) hold for every regular full exponential family, no other regularity conditions are necessary (all other conditions are implied by regular full exponential family). The "usual" asymptotics also hold for all curved exponential families by smoothness. For a proof of this using the usual IID sampling and n goes to infinity story see Geyer (2013a). In fact, these same "usual" asymptotics hold when there is complicated dependence and no IID in sight and whatever n goes to infinity story can be concocted either makes no sense or yields an intractable problem. For that see Sections 1.4 and 1.5 of Geyer (2013b).

And these justify all the hypothesis tests and confidence intervals based on these "usual" asymptotics, for example, those for generalized linear models that are exponential families and for log-linear models for categorical data and for aster models.

17 More on Observed Equals Expected

17.1 Contingency Tables

By "contingency tables" we mean data in which all variables are categorical. For an example of this let us revisit the seat belt use data.

```
# clean up R global environment
rm(list = ls())

count <- c(7287, 11587, 3246, 6134, 10381, 10969, 6123, 6693,
           996,   759,  973,  757,   812,   380, 1084,  513)
injury <- gl(2, 8, 16, labels = c("No", "Yes"))
gender <- gl(2, 4, 16, labels = c("Female", "Male"))
location <- gl(2, 2, 16, labels = c("Urban", "Rural"))
seat.belt <- gl(2, 1, 16, labels = c("No", "Yes"))

library(glmbb)
out <- glmbb(count ~ seat.belt * injury * location * gender,
    family = "poisson")
summary(out)
```

```
##
## Results of search for hierarchical models with lowest AIC.
## Search was for all models with AIC no larger than min(AIC) + 10
## These are shown below.
##
##   criterion  weight   formula
##   182.8      0.24105  count ~ seat.belt*injury*location + seat.belt*location*gender + injury*locatio
##   183.1      0.21546  count ~ injury*gender + seat.belt*injury*location + seat.belt*location*gender
##   184.0      0.13742  count ~ seat.belt*injury + seat.belt*location*gender + injury*location*gender
##   184.8      0.09055  count ~ seat.belt*injury*location + seat.belt*injury*gender + seat.belt*locati
##   184.9      0.08446  count ~ seat.belt*injury + injury*location + injury*gender + seat.belt*locatio
##   185.0      0.08042  count ~ seat.belt*injury*location + seat.belt*injury*gender + seat.belt*locati
##   185.5      0.06462  count ~ seat.belt*injury*location*gender
##   185.8      0.05365  count ~ seat.belt*injury*gender + seat.belt*location*gender + injury*location*g
##   186.8      0.03237  count ~ injury*location + seat.belt*injury*gender + seat.belt*location*gender
```

We will just look at the best model according to AIC having formula

```
f <- summary(out)$results$formula[1]
f
```

```
## [1] "count ~ seat.belt*injury*location + seat.belt*location*gender + injury*location*gender"
```

With four variables there are four possible three-way interactions, but we only have three of the four in this model.

Refit the model.

```
out.best <- glm(as.formula(f), family = poisson)

observed <- xtabs(count ~ seat.belt + injury + location + gender)
expected <- predict(out.best, type = "response")
expected <- xtabs(expected ~ seat.belt + injury + location + gender)
observed.minus.expected <- observed - expected
names(dimnames(observed))
```

```
## [1] "seat.belt" "injury"    "location"  "gender"
```

```
apply(observed.minus.expected, 1:3, sum)
```

```
## , , location = Urban
##
##          injury
## seat.belt            No           Yes
##       No  -3.365130e-11 -2.160050e-12
##       Yes -1.818989e-12 -1.421085e-12
##
## , , location = Rural
##
##          injury
## seat.belt            No           Yes
##       No  -9.549694e-12 -3.410605e-13
##       Yes  9.094947e-13 -1.136868e-12
```

```
apply(observed.minus.expected, c(1, 2, 4), sum)
```

```
## , , gender = Female
##
##          injury
```

```
## seat.belt        No       Yes
##       No   3.646043 -3.646043
##       Yes -3.646043  3.646043
##
## , , gender = Male
##
##          injury
## seat.belt        No       Yes
##       No  -3.646043  3.646043
##       Yes  3.646043 -3.646043
```

```
apply(observed.minus.expected, c(1, 3, 4), sum)
```

```
## , , gender = Female
##
##          location
## seat.belt         Urban         Rural
##       No  -1.955414e-11 -3.979039e-12
##       Yes -9.663381e-12  7.730705e-12
##
## , , gender = Male
##
##          location
## seat.belt         Urban         Rural
##       No  -1.625722e-11 -5.911716e-12
##       Yes  6.423306e-12 -7.958079e-12
```

```
apply(observed.minus.expected, 2:4, sum)
```

```
## , , gender = Female
##
##       location
## injury         Urban         Rural
##    No  -2.819434e-11  4.092726e-12
##    Yes -1.023182e-12 -3.410605e-13
##
## , , gender = Male
##
##       location
## injury         Urban         Rural
##    No  -7.275958e-12 -1.273293e-11
##    Yes -2.557954e-12 -1.136868e-12
```

Indeed we have observed equals expected (up to the accuracy of computer arithmetic) for three of the four three-way margins. Of course, we also have observed equals expected for lower order margins of these margins.

```
apply(observed.minus.expected, 1:2, sum)
```

```
##          injury
## seat.belt            No           Yes
##       No  -4.320100e-11 -2.501110e-12
##       Yes -9.094947e-13 -2.557954e-12
```

```
apply(observed.minus.expected, c(1, 3), sum)
```

```
##          location
```

```
## seat.belt         Urban         Rural
##       No  -3.581135e-11 -9.890755e-12
##       Yes -3.240075e-12 -2.273737e-13
```

```
apply(observed.minus.expected, c(1, 4), sum)
```

```
##          gender
## seat.belt        Female          Male
##       No  -2.353318e-11 -2.216893e-11
##       Yes -1.932676e-12 -1.534772e-12
```

```
apply(observed.minus.expected, 2:3, sum)
```

```
##       location
## injury         Urban         Rural
##    No  -3.547029e-11 -8.640200e-12
##    Yes -3.581135e-12 -1.477929e-12
```

```
apply(observed.minus.expected, c(2, 4), sum)
```

```
##       gender
## injury        Female          Male
##    No  -2.410161e-11 -2.000888e-11
##    Yes -1.364242e-12 -3.694822e-12
```

```
apply(observed.minus.expected, 3:4, sum)
```

```
##         gender
## location        Female          Male
##    Urban -2.921752e-11 -9.833911e-12
##    Rural  3.751666e-12 -1.386979e-11
```

```
apply(observed.minus.expected, 1, sum)
```

```
##            No           Yes
## -4.570211e-11 -3.467449e-12
```

```
apply(observed.minus.expected, 2, sum)
```

```
##            No           Yes
## -4.411049e-11 -5.059064e-12
```

```
apply(observed.minus.expected, 3, sum)
```

```
##         Urban         Rural
## -3.905143e-11 -1.011813e-11
```

```
apply(observed.minus.expected, 4, sum)
```

```
##        Female          Male
## -2.546585e-11 -2.370371e-11
```

17.2 Categorical Response But Quantitative Predictors

For an example of observed equals expected in a more general context, we revisit the time of day data

```
# clean up R global environment
rm(list = ls())

foo <- read.table("http://www.stat.umn.edu/geyer/5102/data/ex6-4.txt",
```

```
    header = TRUE)
count <- foo$count
w <- foo$hour / 24 * 2 * pi
out <- glm(count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w)) + I(cos(2 * w)),
    family = poisson, x = TRUE)
summary(out)
```

```
##
## Call:
## glm(formula = count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w)) +
##     I(cos(2 * w)), family = poisson, x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2043  -0.7431  -0.0905   0.6129   3.2662
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.65917    0.02494  66.516  < 2e-16 ***
## I(sin(w))     -0.13916    0.03128  -4.448 8.66e-06 ***
## I(cos(w))     -0.28510    0.03661  -7.787 6.86e-15 ***
## I(sin(2 * w)) -0.42974    0.03385 -12.696  < 2e-16 ***
## I(cos(2 * w)) -0.30846    0.03346  -9.219  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 704.27  on 335  degrees of freedom
## Residual deviance: 399.58  on 331  degrees of freedom
## AIC: 1535.7
##
## Number of Fisher Scoring iterations: 5
```

Check observed equals expected.

```
m <- out$x
e <- predict(out, type = "response")
t(m) %*% (count - e)
```

```
##                         [,1]
## (Intercept)    2.913225e-13
## I(sin(w))     -4.048600e-14
## I(cos(w))     -1.980638e-13
## I(sin(2 * w)) -4.794310e-13
## I(cos(2 * w)) -1.323386e-13
```

OK. It checks. But what does it mean?

Let z denote any regressor (column of the model matrix). Let y denote the response vector ("observed") and $\hat{\mu}$ its MLE prediction ("expected").

- For the `(Intercept)` column, z has all components equal to one.
- For the `I(sin(w))` column, z has i-th component $\sin(w_i)$.
- And so forth.

Now observed equals expected says

$$z^T y = z^T \hat{\mu}, \qquad \text{for each regressor } z. \tag{33}$$

Theorem 17.1. *Suppose we are working with a canonical affine submodel of a regular full exponential family, and this submodel has an "intercept". Then the residual vector $y - \hat{\mu}$ is empirically uncorrelated with each regressor (column of the model matrix).*

The term "empirically uncorrelated" in the theorem statement means

$$\sum_i (y_i - \hat{\mu}_i)(z_i - \bar{z}_i) = 0 \tag{34}$$

where $\bar{z}_i$ is the mean of the regressor vector z_i.

This theorem is a simple consequence of the observed-equals-expected property that is often discussed in the context of linear models because it expresses an important fact about residuals-versus-predictor plots. But the theorem is not much use for generalized linear models and log-linear models for categorical data because residual analysis is not usually discussed (residual analysis is very complicated when the response vector is not normally distributed, we will not mention it again in this course).

There are many ways to rephrase "observed equals expected". The fundamental meaning (33), but many other statements can be derived from that.

A Proofs

of Theorem 5.1. Consider two canonical parameter values θ_1 and θ_2 and the corresponding mean value parameter values μ_1 and μ_2 (so $\mu_i = \nabla c(\theta_i)$). We assume θ_1 and θ_2 do not correspond to the same distribution. Then Theorem 1 part (a) in Geyer (2009) says that

$$g(s) = l(\theta_1 + s(\theta_2 - \theta_1)) = \langle y, \theta_1 \rangle + s\langle y, \theta_2 - \theta_1 \rangle - c(\theta_1 + s(\theta_2 - \theta_1))$$

is a strictly concave function. on the interval on which it is finite, which has 0 and 1 as interior points because of the assumption of regular full exponential family. Hence by Theorem 2.13 part (a) in Rockafellar and Wets (1998) and the chain rule from calculus its derivative

$$\begin{aligned} g'(s) &= \frac{d}{ds} l(\theta_1 + s(\theta_2 - \theta_1)) \\ &= \langle y, \theta_2 - \theta_1 \rangle - \langle \nabla c(\theta_1 + s(\theta_2 - \theta_1)), \theta_2 - \theta_1 \rangle \\ &= \langle y, \theta_2 - \theta_1 \rangle - \langle E_{\theta_1 + s(\theta_2 - \theta_1)}(Y), \theta_2 - \theta_1 \rangle \end{aligned}$$

is a strictly decreasing function. Thus

$$g'(0) = \langle y - \mu_1, \theta_2 - \theta_1 \rangle > g'(1) = \langle y - \mu_2, \theta_2 - \theta_1 \rangle$$

Hence

$$g'(0) - g'(1) = \langle \mu_2 - \mu_1, \theta_2 - \theta_1 \rangle$$

is strictly greater than zero. Hence $\mu_1 \neq \mu_2$. □

of Theorem 6.1. We try to maximize q given by (8) by setting the first derivative equal to zero and solving for θ. The first derivative is

$$q'(\theta) = \mu - \nabla c(\theta) = \mu - E_\theta(Y)$$

By assumption, such a θ does exist because μ is assumed to be equal to $E_\theta(Y)$ for some θ. By Theorem 2.14 part (b) in Rockafellar and Wets (1998), this solution is a global maximizer of q. □

of Theorem 6.2. Any cumulant function is lower semicontinuous and convex (Barndorff-Nielsen, 1978, Theorem 7.1). Hence (8) is an upper semicontinuous concave function. The rest of the theorem is Corollary 27.2.2 in Rockafellar (1970). □

of Theorem 6.3. In case $y = \mu$ for some mean value parameter vector μ, this is a special case of Theorem 6.2. □

of Theorem 7.1.

$$\begin{aligned} c'(\theta) &= \frac{d}{d\theta} n \log(1 + e^\theta) \\ &= \frac{n e^\theta}{1 + e^\theta} \\ &= n\pi \\ c''(\theta) &= \frac{d}{d\theta} \frac{n e^\theta}{1 + e^\theta} \\ &= \frac{n e^\theta}{1 + e^\theta} - \frac{n (e^\theta)^2}{(1 + e^\theta)^2} \\ &= n \frac{e^\theta}{1 + e^\theta} (1 - \frac{e^\theta}{1 + e^\theta}) \\ &= n\pi(1 - \pi) \end{aligned}$$

□

of Theorem 7.2.

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

□

of Theorem 7.3.

$$\begin{aligned}
c'(\theta) &= \frac{d}{d\theta}\left(-r\log(1-e^{\theta})\right) \\
&= \frac{re^{\theta}}{1-e^{\theta}} \\
&= \frac{r(1-\pi)}{\pi} \\
c''(\theta) &= \frac{d}{d\theta}\frac{re^{\theta}}{1-e^{\theta}} \\
&= \frac{re^{\theta}}{1-e^{\theta}} + \frac{r(e^{\theta})^2}{(1-e^{\theta})^2} \\
&= \frac{re^{\theta}}{1-e^{\theta}}\left(1+\frac{e^{\theta}}{1-e^{\theta}}\right) \\
&= \frac{re^{\theta}}{1-e^{\theta}}\frac{1}{1-e^{\theta}} \\
&= \frac{r(1-\pi)}{\pi^2}
\end{aligned}$$

□

of Theorem 7.4. To simplify notation, let I denote the index set $\{1, 2, \ldots, k\}$

$$\begin{aligned}
\frac{\partial c(\theta)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i} n\log\left(\sum_{j\in I} e^{\theta_j}\right) \\
&= \frac{ne^{\theta_i}}{\sum_{j\in I} e^{\theta_j}}
\end{aligned}$$

and from (6) we know this must be equal to the mean of the multinomial $n\pi_i$. And cancelling the n on both sides gives (23). And

$$\begin{aligned}
\frac{\partial^2 c(\theta)}{\partial \theta_i^2} &= \frac{\partial}{\partial \theta_i}\frac{ne^{\theta_i}}{\sum_{j\in I} e^{\theta_j}} \\
&= \frac{ne^{\theta_i}}{\sum_{j\in I} e^{\theta_j}} - \frac{n(e^{\theta_i})^2}{\left(\sum_{j\in I} e^{\theta_j}\right)^2} \\
&= \frac{ne^{\theta_i}}{\sum_{j\in I} e^{\theta_j}}\left(1-\frac{e^{\theta_i}}{\sum_{j\in I} e^{\theta_j}}\right) \\
&= n\pi_i(1-\pi_i)
\end{aligned}$$

and for $i \neq j$

$$\begin{aligned}
\frac{\partial^2 c(\theta)}{\partial \theta_i \partial \theta_j} &= \frac{\partial}{\partial \theta_j}\frac{ne^{\theta_i}}{\sum_{k\in I} e^{\theta_k}} \\
&= -\frac{ne^{\theta_i}e^{\theta_j}}{\left(\sum_{k\in I} e^{\theta_k}\right)^2} \\
&= -n\pi_i\pi_j
\end{aligned}$$

□

of Theorem 17.1. The special case of (33) with z the "intercept" regressor (all of whose components are equal to one) says $\sum_i y_i = \sum_i \hat{\mu}_i$ or

$$\sum_i (y_i - \hat{\mu}_i) = 0 \tag{35}$$

Rearranging (33) gives

$$\sum_i (y_i - \hat{\mu}_i) z_i = 0 \tag{36}$$

for any regressor z. Multiplying (35) by a constant gives

$$\sum_i (y_i - \hat{\mu}_i) \bar{z}_i = 0 \tag{37}$$

and subtracting (37) from (36) gives (34). □

B Concepts

- exponential family
- log likelihood (equation (1))
- canonical (vector) statistic
  - of exponential family
  - of canonical affine submodel of exponential family
- canonical parameter vector
  - of exponential family
  - of canonical affine submodel of exponential family
- mean value parameter vector
  - of exponential family
  - of canonical affine submodel of exponential family
- cumulant function, see also equation (2) and Sections 9.1 and 9.3.4.
- $\langle \cdot, \cdot \rangle$
- full exponential family
- regular full exponential family
- probability mass function (PMF)
- probability density function (PDF)
- probability mass-density function (PMDF)
- degenerate probability distribution
- cumulants
- mean is first derivative of cumulant function
- variance is second derivative of cumulant function
- identifiable parameterization
- binomial distribution as exponential family
- Poisson distribution as exponential family

- negative binomial distribution as exponential family
- multinomial distribution as exponential family
- univariate normal distribution as exponential family
- sufficient statistic
- sufficiency principle
- sufficient dimension reduction
  - in independent and identically distributed data
  - in-canonical-affine-submodels
- canonical affine submodel
  - offset vector
  - model matrix
  - design matrix
  - coefficients vector also called canonical affine submodel canonical parameter vector
- linear function
- affine function, see also Section 3.4
- observed equals expected property, see also Section 17
- maximum entropy property
- multivariate monotonicity property
- regression coefficients are meaningless
- asymptotics

Bibliography

Agresti, A. (2013) *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons.

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

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

Geyer, C. J. (1990) *Likelihood and exponential families*. PhD thesis. University of Washington. Available at: http://purl.umn.edu/56330.

Geyer, C. J. (1991) Constrained maximum likelihood exemplified by isotonic convex logistic regression. *Journal of the American Statistical Association*, **86**, 717–724.

Geyer, C. J. (1994) On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B*, **56**, 261–274. DOI: 10.1111/j.2517-6161.1994.tb01976.x.

Geyer, C. J. (1999) Likelihood inference for spatial point processes. In *Stochastic Geometry: Likelihood and Computation* (eds O. E. Barndorff-Nielsen, W. S. Kendall, and M. N. M. van Lieshout), pp. 79–140. Boca Raton, FL: Chapman & Hall/CRC.

Geyer, C. J. (2009) Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, **3**, 259–289. DOI: 10.1214/08-EJS349.

Geyer, C. J. (2013a) Asymptotics of exponential families. Available at: http://www.stat.umn.edu/geyer/8112/notes/expfam.pdf.

Geyer, C. J. (2013b) Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton* (eds G. L. Jones and X. Shen), pp. 1–24. Hayward, CA: Institute of Mathematical Statistics. DOI: 10.1214/12-IMSCOLL1001.

Geyer, C. J. (2016a) Stat 5421 lecture notes: Exponential families, Part II. Available at: http://www.stat.umn.edu/geyer/5421/notes/infinity.pdf.

Geyer, C. J. (2016b) Statistics 5102 (geyer, fall 2016) slides. Available at: http://www.stat.umn.edu/geyer/5102/slides/.

Geyer, C. J. (2018) Statistics 8931 (geyer, fall 2018) slides. Available at: http://www.stat.umn.edu/geyer/8931aster/slides/.

Geyer, C. J. (2020a) Stat 3701 lecture notes: Bayesian inference via Markov chain Monte Carlo (MCMC). Available at: http://www.stat.umn.edu/geyer/3701/notes/mcmc-bayes.html.

Geyer, C. J. (2020b) Stat 8501 lecture notes: Spatial lattice processes. Available at: http://www.stat.umn.edu/geyer/8501/lattice.pdf.

Geyer, C. J. (2020c) Stat 8501 lecture notes: Spatial point processes. Available at: http://www.stat.umn.edu/geyer/8501/points.pdf.

Geyer, C. J. (2021) *R Package* `aster`*: Aster Models, Version 1.1-2.* Available at: http://cran.r-project.org/package=aster.

Geyer, C. J. and Møller, J. (1994) Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, **21**, 359–373.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society, Series B*, **54**, 657–699. DOI: 10.1111/j.2517-6161.1992.tb01443.x.

Geyer, C. J., Ryder, O. A., Chemnick, L. G., et al. (1993) Analysis of relatedness in the California condors from dna fingerprints. *Molecular Biology and Evolution*, **10**, 571–589.

Geyer, C. J., Wagenius, S. and Shaw, R. G. (2007) Aster models for life history analysis. *Biometrika*, **94**, 415–426. DOI: 10.1093/biomet/asm030.

Halmos, P. R. and Savage, L. J. (1949) Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Annals of Mathematical Statistics*, **20**, 225–241.

Neyman, J. (1935) Sur un teorema concernente le cosiddette statistiche sufficienti. *Giornale dell'Istituto Italiano degli Attuari*, **6**, 320–334.

Rockafellar, R. T. (1970) *Convex Analysis*. Princeton: Princeton University Press.

Rockafellar, R. T. and Wets, R. J.-B. (1998) *Variational Analysis*. Berlin: Springer-Verlag.

Shaw, R. G. and Geyer, C. J. (2010) Inferring fitness landscapes. *Evolution*, **64**, 2510–2520. DOI: 10.1111/j.1558-5646.2010.01010.x.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595–601. Available at: http://www.jstor.org/stable/2236315.