Stat 5421 Lecture Notes: Exponential Families

Charles J. Geyer

December 02, 2020

Contents

1 License

2 R

- The version of R used to make this document is 4.0.3.
- The version of the `rmarkdown` package used to make this document is 2.5.

3 Exponential Families

We will use the following definition (Geyer, 2009). 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).

You may object to the angle brackets notation as unfamiliar and not what you saw in some other class and prefer some notation like $\sum_i y_i \theta_i$ or (y, θ) or $y \cdot \theta$ or $y^T \theta$ or $\theta^T y$ or one of the latter with little t or prime for transpose. In your humble author's opinion, the angle brackets are superior because they make it clear that $\langle y, y \rangle$ or $\langle \theta, \theta \rangle$ is always obviously wrong, whereas $y^T y$ or $\theta^T \theta$ or the same in any other notation is not obviously wrong. The angle brackets notation comes from functional analysis.

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 5.3 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.

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

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

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.

The family is *full* if its canonical parameter space is

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

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. Constrained exponential families (Geyer, 1991) are not full. A few exponential families used in spatial statistics are full but not regular (Geyer and Møller, 1994).

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. If a statistical model is a regular full exponential family, then it has all of the properties discussed here. If a statistical model is an exponential family (not necessarily full or regular), then it has many of the properties discussed here. Presumably, that is the reason for the newer terminology. In this we are again following Barndorff-Nielsen (1978).

4 Mean Value Parameters

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

Using (2), the MGF for an exponential family with log likelihood (1) is given by

$$M_\theta(t) = E_\theta(e^{ty}) = e^{c(\theta + t) - c(\theta)}$$

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

$$E_\theta(y) = \nabla c(\theta) \tag{4}$$

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

(Barndorff-Nielsen, 1978, Theorem 8.1).

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.

The canonical parameterization is not always 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).

The means of distributions in a regular full exponential family always constitute an identifiable parameterization (Geyer, 2020b, Theorem 2).

The mean value parameterization of a regular full exponential family is just as good as the canonical parameterization. Since cumulant functions are infinitely differentiable, the transformation from canonical to mean value parameters is infinitely differentiable. If the canonical parameterization is chosen so it is identifiable, then the inverse function theorem of real analysis says the inverse mapping is infinitely differentiable too. (See also Section 8 below.)

In elementary applications mean value parameters are preferred. When we introduce the binomial and Poisson distributions we use mean value parameters. It is only when we get to generalized linear models with binomial or Poisson response that we need canonical parameters. When we introduce the multinomial distribution we use mean value parameters. It is only when we get to hierarchical log-linear models for categorical data that we need canonical parameters.

5 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 5.3 below. But exponential families were there first.

For an introduction to the concept of sufficiency, see Section 4 of Geyer (2020b).

5.1 Canonical Statistics are Sufficient

Since the likelihood only depends on the data through the canonical statistic, the canonical statistic is always a (vector) *sufficient statistic*. This is one direction of the Neyman-Fisher factorization theorem.

5.2 Independent and Identically Distributed Sampling

Suppose $y_1, y_2, \ldots, y_n$ 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

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

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, $\ldots$, y_n to $\sum_i y_i$ is a *sufficient dimension reduction*. It loses no information about the parameters assuming the model is correct.

5.3 Canonical Affine Submodels

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

$$
\theta = a + M\beta \tag{6}
$$

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 original parameter, and
- β is the *canonical affine submodel canonical parameter* (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).

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 (6) 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.

So (6) 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 (6) 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 is 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 (6), it adds a column to M (all of whose components are ones).

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 (6) 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.

In the list following (6), "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 $\mathrm{var}_\theta(y)$ 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).

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

$$l(\beta) = \langle M^T y, \beta \rangle - c_{\text{sub}}(\beta) \tag{7}$$

where

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

From this it follows that the change of parameter (6) 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.

5.4 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 5.2 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 $\hat{\theta}_n = \max\{y_1, \ldots, y_n\}$ 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 5.2 above.

6 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

$$\nabla l(\theta) = y - \nabla c(\theta) = y - E_\theta(Y) \tag{8}$$

(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

$$y = E_{\hat{\theta}}(Y) \tag{9}$$

We can say a lot more than this. Cumulant functions are lower semicontinuous 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 (8) being equal to zero is a necessary and sufficient condition for θ to maximize (1) (Rockafellar and Wets (1998), Theorem 2.14). Hence (9) is a necessary and sufficient condition for $\hat{\theta}$ to be a 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). 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 (9) 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 7 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.

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 $\hat{\beta}$. In general, we can only determine $\hat{\beta}$ 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.

7 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 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 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 6 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).

Admittedly the one time I have made this argument in a published article (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 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.

8 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 (2020d)), Markov spatial lattice processes (Geyer, 2020c) 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 (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

$$h(\theta) = \nabla c(\theta), \qquad \theta \in \Theta$$

This maps the canonical parameter space into the space where the canonical statistic y and the mean value parameter μ take values, and defines the mapping from canonical to mean value parameter $\mu = h(\theta)$. 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 5.3 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.

9 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.

9.1 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
```

9.2 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. But they fit the same model

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

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

9.3 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
```

9.4 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 for 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, 2020c) (Geyer, 2020d); 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)).

10 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.

10.1 Observed Equals Expected

The simplest and most important property is the observed-equals-expected property (Section 6 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!

10.2 Sufficient Dimension Reduction

The next most important property is sufficient dimension reduction (Section 5.3 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 5.3 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.

10.3 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 7 above) and contains no information (sufficiency, just discussed).

10.4 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 9.4 above).

10.5 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 8 above).

10.6 The Model Equation

Most statistics courses that discuss the regression models teach students to woof about the *model equation* (6). 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.

11 Asymptotics

A story: when I was a first year graduate student I answered a 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.

12 More on Observed Equals Expected

12.1 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
```

12.2 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 $z^T y = z^T \hat{\mu}$ for each regressor z.

In particular, when z is the `(Intercept)` regressor (having all components equal to one), this says $\sum_i y_i = \sum_i \hat{\mu}_i$.

But we can also divide both sides by n and say $\bar{y} = \bar{\hat{\mu}}$.

So (when there is an "intercept") we can say either the sum or the average of the response vector equals the corresponding quantity for its MLE expected values.

Now for general z we have $\sum_i z_i y_i = \sum_i z_i \hat{\mu}_i$. But (when there is an "intercept") we also know $\bar{y} = \bar{\hat{\mu}}$. So we have $\sum_i z_i \bar{y} = \sum_i z_i \bar{\hat{\mu}}$. And subtracting this from our other equation gives

$$\sum_i z_i (y_i - \bar{y}) = \sum_i z_i (\mu_i - \bar{\hat{\mu}})$$

But we also have

$$\sum_i (y_i - \bar{y}) = 0$$
$$\sum_i (\mu_i - \bar{\hat{\mu}}) = 0$$

just by definition of averaging. So we also have

$$\frac{1}{n} \sum_i (z_i - \bar{z})(y_i - \bar{y}) = \frac{1}{n} \sum_i (z_i - \bar{z})(\mu_i - \bar{\hat{\mu}})$$

So this says the empirical (estimate of) the covariance of y and z equals its MLE expected value. And the same if we substitute "correlation" for "covariance".

This there are many ways to rephrase "observed equals expected". The fundamental meaning is $z^T y = z^T \hat{\mu}$ for each regressor z. But there many other statements that can be derived from that.

Bibliography

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

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

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.

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 5421 lecture notes: Exponential families, Part I. Available at: http://www.stat.

umn.edu/geyer/5421/notes/expfam.pdf.

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

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

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

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: https://doi.org/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.