Exponential Family Notes

Daniel J. Eck

Contents

Introduction

One of the main themes of this course will be developing regression models and demonstrating their use as a means to analyse data. We will see that data structure motivates theoretical and methodological development. Here data will often be collected with the purpose of answering some question that is of interest to a researcher. Examples of such questions include:

- Does adding in-person lectures to an online course improve learning outcomes for students in an introductory statistics course?
- Does a genetically modified genotype provide an improvement to the photosynthetic process for soybeans planted in the wild?
- Is there a racial component to police sentencing?
- What phenotypic traits of an organism are associated with increased ability to produce offspring?

Defensible answers to such questions can be provided by statistical regression models. In this course we are going to focus on statistical regression models that arise from exponential families. These models have been rigorously developed and can be applied to answer questions like those presented above. We will study the origins, fitting, and application of these models in detail, and we will study other statistical models when nuances in data and its analysis warrant different modeling strategies.

In my experience and in the experience of many I know, analyzing data to answer a question of interest to a researcher is very difficult. To do this often requires having extensive conversations with someone from a discipline that is not statistics. For these conversations to be effective one has to have a vast knowledge of statistics, has to be able to translate these concepts into spoken word understandable to a layman, and has to internally translate what they hear from a researcher into statistical terms. Misunderstandings are inevitable.

This course will not be a consulting course and we will not simulate such conversations directly. However, materials in this course will, to the best of my abilities, will be presented in a largely expository style with notation and symbols given secondary priority to stating concepts in words. This is meant to develop the student's ability to translate concepts. It is important to note that an expository writing style is not unique to this course. In fact, it is advocated as a style for writing mathematics by mathematicians who are interested in presenting their ideas clearly. The following passage is taken from an essay written by University of Illinois Urbana-Champaign graduate and well-known mathematician Paul Halmos:

> "The best notation is no notation; whenever it is possible to avoid the use of a complicated alphabetic apparatus, avoid it. A good attitude to the preparation of written mathematical exposition is to pretend that it is spoken. Pretend that you are explaining the subject to a friend on a long walk in the woods, with no paper available; fall back on symbolism only when it is really necessary."

Halmos's essay appeared in a book titled How to write mathematics. This book was the result of a committee authorized by the Council of the American Mathematical Society. Halmos wanted to resign from the committee almost immediately because he thought the project was too interesting to be leave to a committee who he felt would not be able to complete the task properly. His resignation was rejected by the chairman of the committee.

To say Halmos was passionate about mathematical writing would be an understatement. But this course is not just about mathematical writing. This course involves the writing of statistical concepts to be read by a generic researcher from some other discipline. It is important to distinguish mathematics from statistics. First of all, Mathematics and Statistics are separate disciplines. Their distinction is perhaps best articulated by John Nelder who, perhaps by coincidence, played a major role in developing the exponential family regression models that will be studied in this course.

Nelder often references the following Bertrand Russell quote:

> "Mathematics is a subject in which we do not know what we are talking about, nor care whether what we say is true."

One of Nelder's take on Russell's quote is given in his 1986 Presidential Address to the Royal Statistical Society [Nelder, 1986]:

> "A mathematical theory, such as group theory, constructs an edifice of theorems built on a well-defined set of axioms. The method of exposition (though not usually the method of discovery) is deductive, and some of the results are of enormous power and generality. But the theorems are totally abstract, as Russell's characteristic aphorism so aptly declares. That is, the theory stands on its own, without reference to possible interpretation in terms of objects in the world outside, their properties and behaviour. In statistics, by contrast, we ought to know what we are talking about, in the sense of relating our theory to external objects. We should also care about whether what we say is true, in the sense of our inferences and predictions being well supported by the data."

Nelder goes on to state:

> "When mathematicians construct theories they do not seem in general to think of themselves as constructing tools for others to use. That they frequently, and apparently inadvertently, do just that has often been remarked upon... If the applicability of mathematical theories as tools in statistics is indeed unplanned, then we should not be surprised if their application can be both liberating and constricting... We need both to take what is useful from a theory and to refuse to be constrained by it where it proves unsuitable for our purposes... The main danger, I believe, in allowing the ethos of mathematics to gain too much influence in statistics is that statisticians will be tempted into types of abstraction that they believe will be thought respectable by mathematicians rather than pursuing ideas of value to statistics... However, there is little doubt that this temptation ought to be resisted, for the two disciplines have very different objectives."

The objective of statistics according to Nelder is stated in the first sentence of the abstract of his Presidential Address:

> "**Statistics is seen as being primarily concerned with the theory and practice of the matching of theory to data by research worker.**"

As alluded to previously in this introduction, this course will primarily be concerned with the theory and practice of the matching of theory to data by research worker.

The matching of theory to data by research worker requires data obtained by research workers to exist and it requires collaboration between the statistician and the research worker. Thus the expository style of this course is required to go beyond Halmos's expository style for mathematics and will occasionally require plain speaking of aspects of data, statistical concepts, or both. Additionally, some homework problems in this course will be vague. A final goal will be stated in homework problems, but the specific model to be applied ot the specific covariates to use will not be explicitly stated. This will be uncomfortable but it is by design. Homework problems in this course will build experience with translating written words circling a question of interest into statistical terms, fitting models to answer the question of interest, back translating answers from statistical models back into vernacular understandable by a layman, and presenting results and analysis clearly.

Nelder [1999] collects his ideas in the following sentence:

> "Mathematics remains the source of our tools, but statistical science is not just a branch of mathematics; it is not a purely deductive system, because it is concerned with quantitative inferences from data obtained from the real world."

We now develop exponential families and explore their mathematical properties. Exponential families and regression models that arise from them are needed tools for making quantitative inferences from data obtained from the real world. Data of the form:

```r
set.seed(13)
n = 50

## Bernoulli
rbinom(n = n, size = 1, prob = 0.25)
```

```
##  [1] 0 0 0 0 1 0 0 1 1 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0
## [39] 1 0 0 0 1 0 0 1 0 0 0 0
```

```r
## Poisson
rpois(n = n, lambda = 10)
```

```
##  [1] 10 13  5 12  8  8  9  8  8  9  7 11  9  9 10  9 20 14 14  8 12 13 10 10 16
## [26]  9  7  7 12  6 10 17  7  7 13  9  6 11 13 11 11  8 10  9 11 13 13 12 12 12
```

```r
## Normal
rnorm(n = n)
```

```
##  [1]  1.45220302  0.23400474 -0.62822125 -2.88088757 -0.05461001 -0.30682025
##  [7] -1.93230970  1.72747690  0.82827281  0.28158880  2.61745473 -0.15096193
## [13] -1.89606166  1.32567044  0.25153188 -0.42020630  2.02578307  0.22481310
## [19]  0.51349255  0.97362537  2.42577100 -0.41792890 -2.29381013 -1.36004169
## [25]  0.05444450 -0.01681048 -1.53919240  0.75665139  0.38411449 -0.30143957
## [31] -0.67610539 -0.47362192  0.72946611  1.05485783 -0.86416775 -0.39363148
## [37] -0.74302218 -1.87596294 -0.39570349  1.20444672  0.12989528 -1.38555391
## [43]  0.67068362 -0.28299731 -2.27810871 -0.09873861  0.41139707  1.18896385
## [49] -0.87415590  0.46426986
```

```r
## Logistic regression
p = 3
beta = rep(1,p+1)
x = matrix(rnorm(n*p, sd = 0.5), nrow = n, ncol = p)
M = cbind(1, x)
y = rbinom(n = n, size = 1, prob = 1/(1 + exp(-M %*% beta)))
dat = data.frame(y = y,
                 x1 = x[, 1],
                 x2 = x[, 2],
                 x3 = x[, 3])
head(dat)
```

```
##   y          x1          x2          x3
## 1 1 -0.86729736  0.52288044  0.41335803
## 2 1  0.05994054 -0.05204069  0.30972733
## 3 1 -0.03327264 -1.20832770  0.51360990
## 4 1  0.22808504  1.15006522  0.08587834
## 5 1  0.29499420 -0.12562875 -0.36819712
## 6 0  0.68368826  0.25237883  0.04707178
```

```r
## Poisson regression
y = rpois(n = n, lambda = exp(M %*% beta))
dat = data.frame(y = y,
                 x1 = x[, 1],
                 x2 = x[, 2],
                 x3 = x[, 3])
head(dat)
```

```
##    y          x1          x2          x3
## 1  1 -0.86729736  0.52288044  0.41335803
## 2  6  0.05994054 -0.05204069  0.30972733
## 3  3 -0.03327264 -1.20832770  0.51360990
## 4 15  0.22808504  1.15006522  0.08587834
## 5  2  0.29499420 -0.12562875 -0.36819712
## 6 12  0.68368826  0.25237883  0.04707178
```

Definitions and properties of exponential families

Log likelihood

In this class we will define a member of an *exponential family of distributions* as a parametric statistical model having log likelihood

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

Here,

y is the canonical statistic,

θ is the canonical parameter,

$\langle y, \theta \rangle$ is the usual inner product,

$c(\theta)$ is the cumulant function.

We use the convention that terms that do not contain the parameter vector can be dropped from a log likelihood; otherwise additional terms also appear in (1). When the log likelihood can be expressed as (1) we say that y is the *canonical statistic* and θ is the *canonical parameter*. We will often refer to the log likelihood (1) as being in canonical form.

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.

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). 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.

Many widely used statistical distributions are exponential families that have log likelihoods that can be written in canonical form. This current presentation is simple and general, we will discuss support sets for y and parameter spaces for θ later.

Example (Binomial distribution): Done in class.

Example (Normal distribution): Done in class.

Densities

We will have some trouble writing down exponential family densities with our definition of a log likelihood (1). First y is not the data; rather it is a statistic, a function of the data. Let w represent the full data, then the densities have the form

$$f_\theta(w) = h(w) \exp\left(\langle Y(w), \theta \rangle - c(\theta) \right) \tag{2}$$

and the word *density* here can refer to a probability mass function (PMF) or a probability density function (PDF) or to a probability mass-density function (PMDF) if we are referring to a distribution that is partly discrete and partly continuous (either some components of the Y are discrete and some continuous or some components are a mixture of discrete and continuous) or to a density with respect to an arbitrary positive measure in the sense of probability theory.

The $h(w)$ arises from any term not containing the parameter that is dropped when writing the log likelihood (1). We saw this above in our Binomial distribution example. The function h has to be nonnegative, and any point w such that $h(w) = 0$ is not in the support of any distribution in the family.

Example (Binomial distribution): Done in class

Example (Normal distribution): Done in class.

Cumulant functions

Here we demonstrate that the cumulant function of an exponential family that is written in canonical form must also be written in a specific functional form. Being a density, (2) must sum, integrate, or sum-integrate to one. Hence,

$$\begin{aligned} 1 &= \int f_\theta(w)dw \\ &= \int h(w) \exp\left(\langle Y(w), \theta\rangle - c(\theta)\right) dw \\ &= \exp\left(-c(\theta)\right) \int \exp\left(\langle Y(w), \theta\rangle\right) h(w)dw. \end{aligned}$$

Rearranging the above implies that

$$c(\theta) = \log\left(\int \exp\left(\langle Y(w), \theta\rangle\right) h(w)dw\right).$$

Being the expectation of a strictly positive quantity, the expectation here must always be strictly positive, so the logarithm is well-defined. By convention, for θ such that the expectation does not exist, we say $c(\theta) = \infty$.

In probability theory the cumulant function is the log Laplace transformation corresponding to the *generating measure* of the exponential family which is given by $\lambda(dw) = h(w)dw$ when the random variable is continuous. Under this formulation

$$c(\theta) = \log\left(\int \exp\left(\langle Y(w), \theta\rangle\right) \lambda(dw)\right).$$

In our log likelihood based definition of the exponential family (1), the dropped terms which do not appear in the log likelihood are incorporated into the counting measure (discrete distributions) or Lebesgue measure (continuous distributions).

Ratios of densities

When we look at a ratio of two exponential family densities with canonical parameter vectors θ and ψ, the $h(w)$ term cancels, and

$$f_{\theta;\psi}(w) = \frac{f_\theta(w)}{f_\psi(w)} = e^{\langle Y(w), \theta-\psi\rangle - c(\theta) + c(\psi)} \tag{3}$$

is a density of the distribution with canonical parameter θ taken with respect to the distribution with canonical parameter ψ (a Radon-Nikodym derivative in probability theory). For any w such that $h(w) = 0$ (3) still makes sense because such w are not in the support of the distribution with parameter value ψ and hence do not not contribute to any probability or expectation calculation, so it does not matter how (3) is defined for such w. Now, since (3) is everywhere strictly positive, we see that every distribution in the family has the same support.

Full families

Our definition of a log likelihood for an exponential family did not specify a parameter space of allowable values for θ. We now revisit this. We will let

$$\Theta = \{\theta : c(\theta) < \infty\} \tag{4}$$

define a *full* exponential family. Many commonly used statistical models are full exponential families. There is literature about so-called *curved exponential families* and other non-full exponential families, but we will not discuss them. With parameter space (4), we now have a log likelihood (1) and density (2) for all $\theta \in \Theta$.

Example (Binomial distribution): Done in class

We now state a mathematical properties of cumulant functions that hold when an exponential family is either full or possesses a parameter space that is a subset of (4). First, some preliminary definitions.

Definition 1. *A function f on a metric space is lower semicontinuous (LSC) at x if*

$$\liminf_{n\to\infty} f(x_n) \geq f(x), \quad \textit{for all sequences } x_n \to x.$$

A function f is LSC if it is LSC at all points of its domain.

Definition 2. *For any function $f : S \to \bar{\mathbb{R}}$, where S is any set and $\bar{\mathbb{R}}$ is the extended real numbers ($\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$), the effective domain of f is*

$$dom f = \{x \in S : f(x) < \infty\}.$$

Definition 3. *A function f on a vector space is is convex if*

$$f(sx + (1-s)y) \leq sf(x) + (1-s)f(y), \quad x, y \in dom f \textit{ and } 0 < s < 1.$$

The above definitions of lower semicontinuity and convex functions are appropriate for functions defined, respectively, on metric and vector spaces. In this course functions relate to exponential families involving real-valued data and real-valued parameter spaces. Thus, the results above hold for our purposes. The above definition of effective domain was needed to define a convex function, but it is interesting to note a connection between effective domain and full exponential families when we take f to be a cumulant function. We now have

Theorem 1. *The cumulant function of an exponential family is a lower semicontinuous convex function.*

The proof of this Theorem follows from two measure theoretic results. LSC follows from Fatou's Lemma, and convexity follows from Hölder's inequality.

Moment and cumulant generating functions

We no longer fuss about $Y(w)$ and will suppress w when writing Y. We still mention the function h in (2) which is now derived with respect to Y instead of w. This distinction is under the hood and not that important. The moment generating function of the canonical statistic, if it exists, is given by

$$\begin{aligned}
M_\theta(t) &= \mathrm{E}_\theta\left(e^{\langle Y,t\rangle}\right) \\
&= \int e^{\langle y,t\rangle} h(y) e^{(\langle y,\theta\rangle - c(\theta))} dy \\
&= \int h(y) e^{(\langle y,t+\theta\rangle - c(\theta))} dy \\
&= \int h(y) e^{(\langle y,t+\theta\rangle - c(\theta) \pm c(\theta+t))} dy \\
&= e^{c(\theta+t)-c(\theta)}.
\end{aligned} \tag{5}$$

The moment generating function exists if it is finite on a neighborhood of zero, that is, if θ is an interior point of the full canonical parameter space (4). For other θ we say the moment generating function does not exist.

By the theory of moment generating functions, if the moment generating function exists, then moments of all orders exist and ordinary moments are given by the derivatives of $M_\theta(t)$ evaluated at zero. In particular,

$$\begin{aligned} \mathrm{E}_\theta(Y) &= \nabla M_\theta(0) = \nabla c(\theta) \\ \mathrm{E}_\theta(YY^T) &= \nabla^2 M_\theta(0) = \nabla^2 c(\theta) + [\nabla c(\theta)][\nabla c(\theta)]^T. \end{aligned}$$

A log moment generating function is called a *cumulant generating function* and its derivatives evaluated at zero are called the *cumulants* of the distribution. For θ in the interior of the full canonical parameter space Θ, the cumulant generating function corresponding to the canonical statistic is

$$k_\theta(t) = c(t + \theta) - c(\theta), \tag{6}$$

where $c(\theta)$ is the cumulant function corresponding to the exponential family in canonical form. The derivatives of $k_\theta(t)$ evaluated at 0 are the same as the cumulant function c evaluated at θ. The first and second cumulants of the canonical statistic are

$$\begin{aligned} \nabla c(\theta) &= \mathrm{E}_\theta(Y) \\ \nabla^2 c(\theta) &= \mathrm{E}_\theta(YY^T) - [\mathrm{E}_\theta(Y)]\,[\mathrm{E}_\theta(Y)]^T = \mathrm{Var}_\theta(Y). \end{aligned} \tag{7}$$

In short, the mean and variance of the natural statistic always exist when θ is in the interior of the full canonical parameter space Θ, and they are given by derivatives of the cumulant function.

Verify that (7) holds for the Binomial, Poisson, and Normal distriburions.

Regular exponential families

This property of having mean and variance of the canonical statistic given by derivatives of the cumulant function is so nice that families which have it for all θ are given a special name. An exponential family is *regular* if its full canonical parameter space (4) is an open set so that the moment and cumulant generating functions exist for all θ and the formulas in the preceding section hold for all θ. Nearly every exponential family that arises in applications is regular. We will not discuss non-regular exponential families. We break from our expository tone on exponential families to collect concepts and formally state the primary exponential families that we are working with in this course.

Definition 4. *A parametric statistical model is said to be a* ***full regular exponential family in canonical form*** *if it has log likelihood*

$$l(\theta) = \langle y, \theta \rangle - c(\theta).$$

Here, y is a vector statistic, θ is a canonical parameter vector, and $c(\theta)$ is the cumulant function where the parameter space $\Theta = \{\theta : c(\theta) < \infty\}$ is an open set. We use the convention that terms that do not contain the parameter vector can be dropped from a log likelihood.

Note that the log likelihood in the definition above is the same as (1) and Θ the definition above is denoted as Θ in (4).

Example (Binomial distribution): Done in class.

Identifiability and directions of constancy

In this section we will discuss geometric properties of exponential families as they concern identifiability. A statistical model is *identifiable* if any two distinct parameter values correspond to distinct distributions. An exponential family fails to be identifiable if there are two distinct canonical parameter values θ and ψ such that the density (2) of one with respect to the other is equal to one with probability one. This happens if $Y^T(\theta - \psi)$ is equal to a constant with probability one. And this says that the canonical statistic Y is concentrated on a hyperplane and the vector $\theta - \psi$ is perpendicular to this hyperplane.

Conversely, if the canonical statistic Y is concentrated on a hyperplane

$$H = \{y : y^T v = a\} \tag{8}$$

for some non-zero vector v, then for any scalar s

$$c(\theta + sv) = \log\left(\int e^{\langle y, \theta + sv\rangle} \lambda(dy)\right) = sa + \log\left(\int e^{\langle y, \theta\rangle} \lambda(dy)\right) = sa + c(\theta),$$

which immediately implies that

$$\begin{aligned} l(\theta + sv) &= \langle Y, \theta + sv\rangle - c(\theta + sv) \\ &= \langle Y, \theta\rangle + s\langle Y, v\rangle - (sa + c(\theta)) \\ &= \langle Y, \theta\rangle + sa - (sa + c(\theta)) \\ &= l(\theta). \end{aligned}$$

Therefore, we see that the canonical parameter vectors θ and $\theta + sv$ correspond to the same exponential family with probability equal to one for all $\theta \in \Theta$ when the canonical statistic is concentrated on a hyperplane (8). We summarize this as follows.

Theorem 2. *An exponential family fails to be identifiable if and only if the canonical statistic is concentrated on a hyperplane. If that hyperplane is given by* (8) *and the family is full, then θ and $\theta + sv$ are in the full canonical parameter space and correspond to the same distribution for every canonical parameter value θ and every scalar s.*

The direction sv along a vector v in the parameter space such that θ and $\theta + sv$ always correspond to the same distribution is called a *direction of constancy*. The theorem says that v is such a vector if and only if $Y^T v$ is constant with probability one. It is clear from this that the set of all such vectors is closed under vector addition and scalar multiplication, hence is a vector subspace. This subspace is called the *constancy space* of the family.

Note: It is always possible to choose the canonical statistic and parameter so the family is identifiable. Y being concentrated on a hyperplane means some components are affine functions of other components with probability one, and this relation can be used to eliminate components of the canonical statistic vector until one gets to an identifiable choice of canonical statistic and parameter. But this is not always advisable. Prematurely enforcing identifiability may complicate many theoretical issues.

Example (Multinomial distribution): We will show that the multinomial distribution is an exponential family and the usual vector statistic is canonical. To see this, let canonical parameter value ψ correspond to the multinomial distribution with sample size n and usual parameter vector p, and we find the exponential family generated by this distribution. Let d denote the dimension of y and θ, let $\binom{n}{y}$ denote multinomial coefficients, and let S denote the sample space of the multinomial distribution (vectors having nonnegative integer components that sum to n).

In the same vein as (3), we obtain the identity

$$c(\theta) = c(\psi) + \log\left(\mathrm{E}_\psi\left(e^{\langle Y, \theta - \psi\rangle}\right)\right) \tag{9}$$

Then (9) gives

$$c(\theta) = c(\psi) + \log\left(\mathrm{E}_\psi\left(e^{\langle Y, \theta - \psi\rangle}\right)\right)$$

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

where the last equality follows from the multinomial theorem. Then (3) gives

$$
\begin{aligned}
f_\theta(y) &= f_\psi(y) e^{\langle y, \theta - \psi \rangle - c(\theta) + c(\psi)} \\
&= \binom{n}{y} \left(\prod_{i=1}^{d} \left[p_i e^{\theta_i - \psi_i}\right]^{y_i}\right) \left(\sum_{i=1}^{d} p_i e^{\theta_i - \psi_i}\right)^{-n} \\
&= \binom{n}{y} \prod_{i=1}^{d} \left(\frac{p_i e^{\theta_i - \psi_i}}{\sum_{j=1}^{d} p_j e^{\theta_j - \psi_j}}\right)^{y_i}.
\end{aligned}
$$

We simplify the above by choosing p to be the vector with all components $1/d$ and ψ to be the zero vector. We will also choose $c(\psi) = n \log(d)$, so that

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

Thus,

$$
f_\theta(y) = \binom{n}{y} \prod_{i=1}^{d} \left(\frac{e^{\theta_i}}{\sum_{j=1}^{d} e^{\theta_j}}\right)^{y_i}
$$

and this is the PMF of the multinomial distribution with sample size n and probability vector having components

$$
p_i(\theta) = \frac{e^{\theta_i}}{\sum_{j=1}^{d} e^{\theta_j}}.
$$

This, however, is not an identifiable parameterization. The components of y sum to n so Y is concentrated on a hyperplane to which the vector $(1, 1, \cdots, 1)^T$ is perpendicular, hence by Theorem 1 a direction of constancy of the family. Eliminating a component of Y to get an identifiability would destroy symmetry of formulas and make everything harder and messier. Best to wait until when (if ever) identifiability becomes absolutely necessary. □

The Right Way[1] (IMHO) to deal with nonidentifiability, which is also called collinearity in the regression context, is the way the `R` functions `lm` and `glm` deal with it. (We will have to see how linear and generalized linear models relate to exponential families before this becomes fully clear, but I assure you this is how what they do relates to a general exponential family). When you find you have a non-identifiable parameterization, you have $Y^T v$ constant with probability one. Pick any i such that $v_i \neq 0$ and fix $\theta_i = 0$ giving a submodel that (we claim) has all the distributions of the original one (we have to show this).

For any parameter vector θ in the original model (with θ_i free to vary) we know that $\theta + sv$ corresponds to the same distribution for all s. Choose s such that $\theta_i + sv_i = 0$, which is possible because $v_i \neq 0$, hence we see that this distribution is in the new family obtained by constraining θ_i to be zero (and the other components of θ vary freely).

[1] The Right Way is borrowed vernacular from Charles Geyer. The Right Way means anything that is not obviously the Wrong Way. There can be several Right Ways, and choosing among them can be subjective.

This new model obtained by setting θ_i equal to zero is another exponential family. Its canonical statistic and parameter are just those of the original family with the i-th component eliminated. Its cumulant function is just that of the original family with the i-th component of the parameter set to zero. This new model need not be identifiable, but if not there is another direction of constancy and the process can be repeated until identifiability is achieved (which it must because the dimension of the sample space and parameter space decreases in each step and cannot go below zero, and if it gets to zero the canonical statistic is concentrated at a single point, hence there is only one distribution in the family, and identifiability vacuously holds).

This is what `lm` and `glm` do. If there is non-identifiability (collinearity), they report `NA` for some regression coefficients. This means that the corresponding predictors have been "dropped" but this is equivalent to saying that the regression coefficients reported to be `NA` have actually been constrained to be equal to zero. The code below demonstrates this point with a simple linear regression model with perfect collinearity.

```r
# generate covariates
n = 500; p = 3
M = matrix(rnorm(n*p), nrow = n)

# generate responses
beta = rep(1, p)
Y = 1 + M %*% beta + rnorm(n)

# add perfect collinearity to the model matrix
M = cbind(M, 2*M[, 1] + M[, 2])

# fit linear regression model and produce model summary table
m1 = lm(Y ~ M)
summary(m1)
```

```
##
## Call:
## lm(formula = Y ~ M)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.06447 -0.56798  0.02027  0.54337  3.13953
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.98194    0.04277   22.96   <2e-16 ***
## M1           0.97563    0.04318   22.59   <2e-16 ***
## M2           0.97655    0.04303   22.69   <2e-16 ***
## M3           1.00563    0.04401   22.85   <2e-16 ***
## M4                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9553 on 496 degrees of freedom
## Multiple R-squared:  0.7408, Adjusted R-squared:  0.7392
## F-statistic: 472.5 on 3 and 496 DF,  p-value: < 2.2e-16
```

Mean value parameterization

The mean of the canonical statistic $\mathrm{E}_\theta(Y)$ is also a parameter. It is given as a function of the canonical parameter θ,

$$\mu = \mathrm{E}_\theta(Y) = \nabla c(\theta) = g(\theta). \tag{10}$$

We will refer to $g(\theta)$ as the change-of-parameter map (or change-of-parameter) from canonical parameter θ to mean value parameter μ. This change-of-parameter map is invertible when the model is identifiable (see below) so that (10) implies that $g^{-1}(\mu) = \theta$. This is very important for generalized linear models as we will soon see.

Theorem 3. *For a full regular exponential family, the change-of-parameter from canonical to mean value parameter is invertible if the model is identifiable. Moreover both the change-of-parameter and its inverse are infinitely differentiable.*

Note that some aspects of this proof are left to the reader. To prove this theorem we will let μ be a possible value of the mean value parameter (that is, $\mu = g(\theta)$ for some θ) and consider the function

$$h(\theta) = \langle \mu, \theta \rangle - c(\theta). \tag{11}$$

The second derivative of h is $-\nabla^2 c(\theta)$ which is equal to $-\mathrm{Var}_\theta(Y)$, and this is a negative definite matrix (**Why?**) Hence (11) is a strictly concave function by Theorem 2.14 in Rockafellar and Wets [1998], and this implies that the maximum of (11) is unique if it exists by Theorem 2.6 in Rockafellar and Wets [1998]. Moreover, we know a solution exists because the derivative of (11) is $\nabla h(\theta) = \mu - \nabla c(\theta)$, and we specified that $\mu = \nabla c(\theta)$ for some θ.

Show that cumulant functions are infinitely differentiable and are therefore continuously differentiable. Now we see that the Jacobian matrix for this change-of-parameters is

$$\nabla g(\theta) = \nabla^2 c(\theta)$$

which we (you) have already shown is nonsingular. The inverse function theorem thus says that g is locally invertible, and the local inverse must agree with the global inverse which we have already shown exists. The inverse function theorem goes on to state that the derivative of the inverse is the inverse of the derivative

$$\nabla g^{-1}(\theta) = \left[\nabla g(\theta)\right]^{-1}, \qquad \text{when } \mu = g(\theta) \text{ and } \theta = g^{-1}(\mu).$$

Now show that $g^{-1}(\theta)$ is infinitely differentiable.

Multivariate monotonicity

A mapping from $g : \mathbb{R}^d \to \mathbb{R}^d$ is multivariate monotone (Definition 12.1 in Rockafellar and Wets [1998]) if

$$\left[g(x_1) - g(x_2)\right]^T (x_1 - x_2) \geq 0, \qquad \text{for } x_1 \text{ and } x_2 \in \mathbb{R}^d, \tag{12}$$

and strictly multivariate monotone if (12) holds with strict inequality whenever $x_1 \neq x_2$. If g is differentiable, then by Proposition 12.3 in Rockafellar and Wets [1998] it is multivariate monotone if and only if the symmetric part of the Jacobian matrix ∇g is positive-semidefinite for each x. A sufficient but not necessary condition for g to be strictly multivariate monotone is that the symmetric part of ∇g be positive definite for each x.

Let g be the change-of-parameters mapping from canonical to mean value parameters (10) then we showed in the previous section that its Jacobian matrix is positive semidefinite in general and strictly positive definite when the model is identifiable. Thus this change-of-parameter is multivariate monotone in general and strictly multivariate monotone when the model is identifiable.

Thus, if μ_1 corresponds to θ_1 and μ_2 to θ_2, we have

$$(\mu_1 - \mu_2)^T (\theta_1 - \theta_2) > 0, \qquad \text{whenever;} \, \theta_1 \neq \theta_2. \tag{13}$$

In general, this is all we can say about the map from canonical to mean value parameters. However, there is a casual version of (13) which eases interpretation. If we rewrite (13) using subscripts

$$\sum_{i=1}^{d}(\mu_{1i} - \mu_{2i})(\theta_{1i} - \theta_{2i}) > 0$$

and consider θ_1 and θ_2 that differ in only one coordinate, say the kth, then we get

$$(\mu_{1k} - \mu_{2k})(\theta_{1k} - \theta_{2k}) > 0,$$

which says *if we increase one component of the canonical parameter vector, leaving the other components fixed, then the corresponding component of the mean value parameter vector also increases, and the other components can go any which way.* This is easier to explain than the full multivariate monotonicity property, but is not equivalent to it. The casual property is not enough to make some arguments about exponential families that are needed in applications (for example, see the Appendix in Shaw and Geyer [2010]).

Here is another rewrite of (13) that preserves its full force. Fix a vector $v \neq 0$. Write $\theta_2 = \theta$ and $\theta_1 = \theta + sv$, so multivariate monotonicity (12) becomes

$$[g(\theta + sv) - g(\theta)]^T v > 0, \qquad \text{for } s \neq 0.$$

Differentiate with respect to s and set $s = 0$, which gives the so-called directional derivative of g in the direction v at the point θ

$$\nabla g(\theta; v) = v^T \left[\nabla g(\theta)\right] v = v^T \left[\nabla^2 c(\theta)\right] v. \tag{14}$$

We know that $\nabla^2 c(\theta)$ is positive semi-definite in general and strictly positive definite when the model is identifiable. Hence we see (again) that the θ to μ mapping is multivariate monotone in general and strictly multivariate monotone when the model is identifiable.

Partial derivatives are special cases of directional derivatives when the vector v points along a coordinate direction (only one component of v is nonzero). So the casual property only says that all the partial derivatives are nonzero and this corresponds to asserting (14) with v being along coordinate directions, and this is equivalent to asserting that the diagonal components of $\nabla^2 c(\theta)$ are positive. And now we clearly see how the casual property is indeed casual. It only asserts that the diagonal elements of $\nabla^2 c(\theta)$ are positive, which is far from implying that $\nabla^2 c(\theta)$ is a positive definite matrix.

Maximum likelihood estimation

We now provide an approach for obtaining maximum likelihood estimates for parameters in a full regular exponential family. In our context, the derivative of the log likelihood is

$$\nabla l(\theta) = y - \nabla c(\theta),$$

and the second derivative of the log likelihood is

$$\nabla^2 l(\theta) = -\nabla^2 c(\theta).$$

Hence observed Fisher information (the Hessian matrix of the log likelihood) and expected Fisher information for the canonical parameter vector θ are the same. We write Fisher information as

$$I(\theta) = \nabla^2 c(\theta). \tag{15}$$

Fisher information measures the expected curvature of the log likelihood around the true parameter value. If the likelihood is sharply curved around θ – the expected information $I(\theta)$ is large – then a small change in θ can lead to a drastic decrease in the likelihood. Conversely, if $I(\theta)$ is small then small changes in θ will not affect the likelihood that much. These heuristics are important when we cover separation and non-identifiability.

When the model is identifiable, the canonical statistic vector Y is not concentrated on a hyperplane, the second derivative is negative definite everywhere, hence the log likelihood is strictly concave, hence the maximum likelihood estimate is unique if it exists. Under this setup, $y = \nabla c(\hat{\theta})$ arises from setting the first derivative of the log likelihood to zero and rearranging terms. This implies that the maximum likelihood estimator (MLE) for θ is

$$\hat{\theta} = g^{-1}(y),$$

where g is the change-of-parameter from canonical to mean value parameters.

Derive the MLEs of the canonical parameters of the Binomial, Poisson, and normal distributions.

Nonexistence of the MLE

Unlike our proof of Theorem 3 where we assumed the existence of a solution, we cannot prove the maximum likelihood estimate (for the canonical parameter) exists. Consider the binomial distribution. The MLE for the usual parameterization is $\hat{p} = y/n$. The canonical parameter is $\theta = \text{logit}(p)$. But $\hat{\theta} = \text{logit}(\hat{p})$ does not exist when $\hat{p} = 0$ or $\hat{p} = 1$, which is when we observe zero successes or when we observe n successes in n trials. We will revisit this topic later in the course.

Observed equals expected

For a full regular exponential family, the MLE cannot be on the boundary of the canonical parameter space (regular means the boundary is empty), and the MLE, if it exists, must be a point where the first derivative is zero, that is, a θ value that satisfies

$$y = \nabla c(\theta) = \mathrm{E}_\theta(Y).$$

Thus the MLE is the (unique if the model is identifiable) parameter value that makes the observed value of the canonical statistic equal to its expected value. We call this the **observed equals expected** property of maximum likelihood in exponential families. This property is even simpler to express in terms of the mean value parameter. By invariance of maximum likelihood under change-of-parameter, the MLE for μ is

$$\hat{\mu} = \nabla c(\hat{\theta}).$$

The observed equals expected property therefore states that

$$y = \hat{\mu}. \tag{16}$$

Independent and identically distributed data

Suppose $y_1, \ldots, y_n$ are independent and identically distributed (iid) from some full regular exponential family (unlike our notation in the preceding section, y_i are not components of the canonical statistic vector but rather iid realizations of the canonical statistic vector, so each y_i is a vector). The log likelihood for sample size n is

$$l_n(\theta) = \sum_{i=1}^n \left[\langle y_i, \theta\rangle - c(\theta)\right] = \langle \sum_{i=1}^n y_i, \theta\rangle - nc(\theta), \tag{17}$$

and we see that the above log likelihood is an exponential family with canonical statistic $\sum_{i=1}^n y_i$, cumulant function $nc(\theta)$, canonical parameter θ, and full canonical parameter space Θ which is the same as the originally given family from which every observation is a member. Thus iid sampling gives us a new exponential family, but still an exponential family.

Asymptotics of maximum likelihood

We now discover an asymptotic distribution for the MLE of the canonical parameter vector in a full regular exponential family. Rewrite (17) as

$$l_n(\theta) = n\left[\langle \bar{y}_n, \theta\rangle - c(\theta)\right]$$

so that

$$\nabla l_n(\theta) = n \left[\bar{y}_n - \nabla c(\theta) \right].$$

From which we see that for an identifiable full regular exponential family where the MLE must be a point where the first derivative is zero, we can write

$$\nabla l_n(\theta) = n \left[\bar{y}_n - \nabla c(\theta) \right] = 0.$$

From here we see that $\bar{y}_n = \nabla c(\hat{\theta})$. Recall the change-of-parameters mapping $g : \theta \mapsto \mu$ given by (10) in the mean value parameters section. We can write

$$\hat{\theta}_n = g^{-1}(\bar{y}_n). \tag{18}$$

More precisely, (18) holds when the MLE exists (when the MLE does not exist, $\bar{y}_n$ is not in the domain of g^{-1}, which is in the range of g).

By the multivariate central limit theorem (CLT)

$$\sqrt{n} \left(\bar{y}_n - \mu \right) \to N \left(0, I(\theta) \right)$$

and we know that g^{-1} is differentiable (Theorem~3) with the derivative given by

$$\nabla g^{-1}(\theta) = \left[\nabla g(\theta) \right]^{-1}, \qquad \text{where } \mu = g(\theta) \text{ and } \theta = g^{-1}(\mu).$$

So the usual asymptotics of maximum likelihood

$$\sqrt{n} \left(\hat{\theta}_n - \theta \right) \to N \left(0, I(\theta)^{-1} \right) \tag{19}$$

is just the multivariate delta method applied to the multivariate CLT.

In summary, one “regularity condition” for (19) to hold is that we have an identifiable full regular exponential family. Of course, (19) holds for many non-exponential-family models, but the regularity conditions are so complicated that they are often hard to verify. In exponential families the verification is trivial: the usual asymptotics of maximum likelihood always works.

Example (Bernoulli distribution): Done in class

Finite sample concentration of MLE

The previous section is devoted to large sample properties of maximum likelihood estimation within the context of full regular exponential families. These properties are especially relevant for statistical inference. MLEs of parameters in full regular exponential families also possess desirable finite sample properties. We first motivate the concept of sub-Gaussian and sub-exponential random variables which represent classes of desirable tail behavior for statistical models. The following definitions come from Wainwright [2019]:

Definition 5. *A random variable Y with mean $\mu = \mathrm{E}(Y)$ is sub-Gaussian if there exists a positive number λ such that*

$$E\left(e^{\phi(Y-\mu)}\right) \leq e^{\lambda^2\phi^2/2} \qquad \textit{for all } \phi \in \mathbb{R}.$$

Definition 6. *A random variable Y with mean $\mu = \mathrm{E}(Y)$ is sub-exponential if there exist non-negative numbers (λ, b) such that*

$$E\left(e^{\phi(Y-\mu)}\right) \leq e^{\lambda^2\phi^2/2} \qquad \textit{for all } |\phi| < 1/b.$$

We will also need the following results taken from Wainwright [2019]:

Lemma 1. *Consider an independent sequence $\{Y_i\}_{i=1}^n$ of random variables with mean μ_i, such that each Y_i is sub-exponential with parameters (λ_i, b_i). Then $\sum_{i=1}^n (Y_i - \mu_i)$ is also sub-exponential with parameters (λ_*, b_*) where*

$$\lambda_* = \sqrt{\sum_{i=1}^n \lambda_i^2} \qquad \text{and} \qquad b_* = \max_{i=1,\ldots,n} b_i.$$

Lemma 2. *Consider an independent sequence $\{Y_i\}_{i=1}^n$ of random variables with mean μ_i, such that each Y_i is sub-exponential with parameters (λ_i, b_i). Then*

$$\mathbb{P}\left(n^{-1}\sum_{i=1}^n (Y_i - \mu_i) \geq t\right) \leq \begin{cases} \exp\left(-\frac{nt^2}{2(\lambda_*^2/n)}\right), & \text{for } 0 \leq t \leq \frac{\lambda_*^2}{nb_*}, \\ \exp\left(-\frac{nt}{2b_*}\right), & \text{for } t > \frac{\lambda_*^2}{nb_*}, \end{cases}$$

where (λ_*, b_*) are as defined in the previous lemma.

Our finiteness argument will be demonstrated in the case when Y is a scalar canonical statistic full regular exponential family with canonical parameter θ (although we will still use the ∇ to denote derivatives). **It is a problem for the reader to show that Y is a sub-exponential random variable**. Now let $\hat{\theta}$ be the MLE for the canonical parameter θ. We now show that the MLE of an exponential family obeys sub-exponential concentration. Consider a Taylor expansion of the score function of an exponential family evaluated at the MLE

$$\begin{aligned} 0 = \nabla l_n(\hat{\theta}) &= \nabla l_n(\theta) + \nabla^2 l_n(\theta)(\hat{\theta} - \theta) + R_n \\ &= \sum_{i=1}^n \{y_i - \nabla c(\theta)\} + \nabla^2 l_n(\theta)(\hat{\theta} - \theta) + R_n, \end{aligned}$$

where $\nabla^2 l_n(\theta) = -n\nabla^2 c(\theta) = -nI(\theta)$ and $R_n = o_P(n^{-1/2})$. Notice that $\sum_{i=1}^n \{y_i - \nabla c(\theta)\}$ is a sum of mean zero sub-exponential random variables, and is also sub-exponential by Lemma 1. Furthermore, scalar products of $\sum_{i=1}^n \{y_i - \nabla c(\theta)\}$ are also sub-exponential. After rearranging terms in the above displayed equation we see that

$$(\hat{\theta} - \theta) = n^{-1} I^{-1}(\theta) \sum_{i=1}^n \{y_i - \nabla c(\theta)\} + \widetilde{R}_n,$$

where $\widetilde{R}_n = n^{-1} I^{-1}(\theta) R_n$. Putting all of this together yields

$$\mathbb{P}\left((\hat{\theta} - \theta) \geq t\right) = \mathbb{P}\left(n^{-1} I^{-1}(\theta) \sum_{i=1}^n \{y_i - \nabla c(\theta)\} \geq t - \widetilde{R}_n\right),$$

where $t > 0$. There exists a number $a > 0$ such that, for n large,

$$\begin{aligned} &\mathbb{P}\left(n^{-1} I^{-1}(\theta) \sum_{i=1}^n \{y_i - \nabla c(\theta)\} \geq t - \widetilde{R}_n\right) \\ &\leq \mathbb{P}\left(n^{-1} \sum_{i=1}^n \{y_i - \nabla c(\theta)\} \geq aI(\theta)t\right). \end{aligned}$$

Lemma 2 implies that

$$\mathbb{P}\left(n^{-1} \sum_{i=1}^n \{y_i - \nabla c(\theta)\} \geq aI(\theta)t\right) \leq \begin{cases} \exp\left(-\frac{na^2 I^2(\theta) t^2}{2\lambda^2}\right), & \text{for } 0 \leq t \leq \frac{\lambda^2}{naI(\theta)b}, \\ \exp\left(-\frac{naI(\theta)t}{2b}\right), & \text{for } t > \frac{\lambda^2}{naI(\theta)b}. \end{cases}$$

We can therefore conclude that the MLE of θ exhibits sub-exponential concentration following the logic that $(\hat{\theta} - \theta)$ has the same tail bounds as a sub-exponential random variable. We can use these results to obtain the rate of convergence. Set $t = \sqrt{\log(n)/n}$ and observe that

$$\mathbb{P}\left((\hat{\theta} - \theta) \geq aI(\theta)\sqrt{\frac{\log(n)}{n}}\right) = O\left(n^{-\frac{a^2 I^2(\theta)}{2\lambda}}\right).$$

Introduction to generalized linear models

We now present generalized linear models (GLMs) within the context of exponential theory. We will not yet discuss specific examples of GLMs, such as logistic or Poisson regression, or extensions beyond canonical representations. These topics will be covered in the not-too-distant future. Before we begin our presentation on GLMs we note the difference between a saturated model and a submodel. A saturated model is a model that has one parameter per observation. It is a perfect fitting model that has little predictive power. On the other hand, a submodel has fewer parameters than the number of observations.

We first present canonical affine submodels of an exponential family. A canonical affine submodel of an exponential family is a submodel having parameterization

$$\theta = a + M\beta$$

where $\theta \in \mathbb{R}^n$, n being the sample size, is the canonical parameter vector corresponding to the original exponential family, $\beta \in \mathbb{R}^p$, $p < n$, is the canonical parameter vector for the submodel, a is known offset vector, and M is a known matrix. The matrix M is usually called the *model matrix* in the terminology used by the `R` functions `lm` and `glm`. The vector a is called the offset vector in the terminology used by the `R` functions `lm` and `glm`.

In most applications the offset vector is not used. Removal of the offset vector from consideration yields the familiar linear parameterization

$$\theta = M\beta.$$

Submodels with the above parametrization are refered to as *canonical linear submodels*. We will restrict attention to the canonical linear submodel in this course unless explicitly stated otherwise. The log likelihood for the canonical linear submodel arises directly from the saturated model log likelihood,

$$\begin{aligned} l(\theta) &= \langle y, \theta \rangle - c(\theta) \\ &= \langle y, M\beta \rangle - c(M\beta) \\ &= \langle M^T y, \beta \rangle - c(M\beta) \\ &= \langle M^T y, \beta \rangle - c_{\text{sub}}(\beta) \\ &= l(\beta). \end{aligned} \tag{20}$$

and we see that we again have an exponential family with

- canonical statistic $M^T y$,
- cumulant function $c_{\text{sub}}(\beta) = c(M\beta)$, and
- submodel canonical parameter vector β.

If θ varies freely (over a whole vector space), then β also varies freely (over a whole vector space of lower dimension, $p < n$). But if the originally given full canonical parameter space was Θ, then the full submodel canonical parameter space is

$$B = \{\beta : M\beta \in \Theta\}.$$

Thus a canonical linear submodel gives us a new exponential family, with lower-dimensional canonical parameter and statistic. The submodel exponential family is full if the original exponential family was full. Notice that Θ and B are defined through the column space of M, not the particular model matrix M, the particular β value does not determine the submodel uniquely.

To distinguish between the submodel and the originally given exponential family, we often call the latter the *saturated model*. Now we have four parameters: the saturated model canonical and mean value parameters, respectively, θ and μ, and the submodel canonical and mean value parameters, respectively, β and τ. Relations between these parameterizations are given in Figure 1.

The observed equals expected property for the submodel is

$$\hat{\tau} = M^T \hat{\mu} = M^T y. \tag{21}$$

$$\begin{tikzcd}
\theta \arrow[r, shift left] \arrow[d, shift left, "\mu = \nabla c(\theta)"] & \beta \arrow[l, shift left, "\theta = M\beta"] \arrow[d, shift left, "\tau = \nabla_\beta c_{\text{sub}}(\beta)"] \\
\mu \arrow[u, shift left] \arrow[r, shift left, "\tau = M^T \mu"] & \tau \arrow[l, shift left] \arrow[u, shift left]
\end{tikzcd}$$

Figure 1: A depiction of the transformations necessary to change between parameterizations. Arrows going in opposite directions specify transformations and their inverses. M is a known model matrix of full column rank, and c is the cumulant function for the exponential family model.

We cannot actually solve these equations for $\hat{\mu}$ because the mapping $\mu \rightarrow M^T\mu$ is usually not one-to-one (the $n > p$ case where $M \in \mathbb{R}^{n\times p}$ and is full column rank). Hence we cannot determine $\hat{\theta}$ and $\hat{\beta}$ from them either. The only way to determine the MLE is to maximize the log likelihood (20) for β to obtain $\hat{\beta}$ and then $\hat{\theta} = M\hat{\beta}$ and $\hat{\mu} = \nabla c(\hat{\theta})$ and $\hat{\tau} = M^T\hat{\mu}$. But the observed equals expected property is nevertheless very important.

Recall that the saturated model canonical parameter vector θ is "linked" to the saturated model mean value parameter vector through the change-of-parameter mappings $g(\theta)$. We can reparameterize $\theta = M\beta$ and write

$$\mu = \mathrm{E}_\theta(Y) = g(M\beta)$$

which implies that we can write

$$g^{-1}\left(\mathrm{E}_\theta(Y)\right) = M\beta.$$

Therefore, a linear function of the canonical submodel parameter vector is linked to the mean of the exponential family through the inverse change-of-parameter mapping g^{-1}. This is the basis of exponential family generalized linear models with link function g^{-1}. Note that most treatments of GLMs will present g as the link function. Instead we motivated g as a change of parameters mapping from canonical to mean value parameters.

A concise presentation of the various model parameterizations and how they are related to each other is given in Figure 1.

Asymptotics and inference

An asymptotic distribution for the MLE of the submodel canonical parameter vector follows directly from exponential family theory that we have already covered. The submodel log likelihood (20) takes the form

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

From invariance of maximum likelihood estimation derived in (19) or direct arguments, we have that the asymptotic distribution of the MLE $\hat{\beta}$ takes the form

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \overset{d}{\rightarrow} N(0, \Sigma^{-1}) \tag{22}$$

where $\Sigma = \mathrm{E}\left(-\nabla^2 l(\beta)\right)$ is the Fisher information matrix corresponding to the canonical linear submodel. The explicit form of $-\nabla^2 l(\beta)$ follows from the chain rule

$$-\nabla^2 l(\beta) = \nabla^2_\beta c_{\text{sub}}(\beta)$$

$$
\begin{aligned}
&= \nabla^2_\beta c(M\beta) \\
&= M^T \left(\nabla_\theta c(\theta)|_{\theta = M\beta} \right) M.
\end{aligned}
$$

An asymptotic distribution for $\hat\beta$ (22) allows for us to make asymptotic statistical inferences for β and individual components of β. First, let $\widehat\Sigma$ be estimated Σ using the MLE $\hat\beta$ in place of β. In particular, the jth element $\hat\beta_j$ of $\hat\beta$ is asymptotically normal with asymptotic variance being the jth diagonal element of $\widehat\Sigma$. We will define a Wald Z statistic for testing $H_o : \beta_j = \beta_{jo}$ as

$$
z_W = \frac{\hat\beta_j - \beta_{jo}}{\text{se}(\hat\beta_j)} \quad \overset{H_o}{\sim} \quad N(0,1),
$$

where $\text{se}(\hat\beta_j) = \sqrt{\widehat{\text{Var}}(\hat\beta_j)}$ and $\overset{H_o}{\sim}$ means distributed as assuming the null hypothesis. It is often of interest to test whether or not $\beta_j = 0$ in which the above Wald Z statistic reduces to

$$
\frac{\hat\beta_j}{\text{se}(\hat\beta_j)} \quad \overset{H_o}{\sim} \quad N(0,1).
$$

The above distributional result allows for one to statistically test whether or not a component of β differs from 0. Such a test can be performed using a p-value of the form

$$
2 - 2\Phi\left(\left| \frac{\hat\beta_j}{\text{se}(\hat\beta_j)} \right| \right)
$$

where Φ is the distribution function for the standard normal distribution. This p-value is then compared with a selected error tolerance. If the p-value is smaller than the error tolerance, then one may conclude that the modeling variable corresponding to the jth element of β is statistically significant at the level of the error tolerance. This is very important for applications, although one has to be careful in presenting conclusions. One needs to keep in mind the differences between statistical and practical significance, the differences between association and causation, and multiple testing issues.

Similar to the p-value above, we can also construct $(1-\alpha) \times 100\%$ Wald based confidence intervals of the form

$$
\hat\beta_j \pm z_{\alpha/2}\text{se}(\hat\beta_j)
$$

where $z_{\alpha/2}$ is a quantile of the standard normal distribution at $\alpha/2$ where $\alpha \in (0,1)$ is a chosen error tolerance.

Deviance, goodness of fit, and likelihood ratios

It is important for us to evaluate the fit of competing statistical models. In our exponential family regression model context, we will use likelihood ratio tests based on the deviance statistic to evaluate competing submodels. We will also compare submodels to the saturated model using likelihood ratio tests to determine if the submodel in question offers a dimension reduction that offers better fit than assuming one parameter per observation.

To motivate the deviance of a statistical model we will revisit the mean value parameter vector μ and rewrite the log likelihood in the notation of this parameterization as $l(\mu; y)$. From the observed equals expected property we have that the unrestricted MLE of μ is y. Now consider a canonical submodel (GLM) of the form $\theta = M\beta$. Let $\hat\mu$ be the MLE of μ restricted to an identifiable canonical submodel ($\hat\mu = \nabla c(\hat\theta)$ where $\hat\theta = M\hat\beta$). It follows that

$$
l(y; y) \geq l(\hat\mu; y).
$$

We refer to the unrestricted case, in which each observation has its own mean ($\hat\mu = y$), is called the *saturated model*.

With all of this in mind, the *deviance* of the GLM is

$$
D(y; \hat\mu) = -2\left(l(\hat\mu; y) - l(y; y) \right).
$$

We see that the deviance statistic is a function of a ratio of two likelihoods, one corresponding to the canonical submodel and one corresponding to the saturated model. The deviance statistic has approximate distribution

$$D(y;\hat{\mu}) \overset{H_o}{\sim} \chi^2_{\text{df}}$$

where H_o is that the canonical submodel is correct and the alternative test is that the canonical submodel is incorrect but the saturated model is correct, df $= n - p$, n is the sample size, and p is the number of parameters in the canonical submodel. So we reject correctness of the canonical submodel when

$$D(y;\hat{\mu}) \; > \; \chi^2_{\text{df}}(\alpha).$$

We can use deviance based testing to nested models. Let $\mathcal{M}_0$ and $\mathcal{M}_1$ both be canonical submodels. We say that $\mathcal{M}_0$ is *nested* within $\mathcal{M}_1$ when every distribution in $\mathcal{M}_0$ is also in $\mathcal{M}_1$ but not vice-versa. That is, μ is more restricted under $\mathcal{M}_0$ than under $\mathcal{M}_1$. Let $\hat{\mu}_0$ be the MLE of μ under $\mathcal{M}_0$, and let $\hat{\mu}_1$ be the MLE of μ under $\mathcal{M}_1$. We can use this framework for testing

$$H_0 : \mathcal{M}_0 \text{ true} \qquad H_a : \mathcal{M}_1 \text{ true, but not } \mathcal{M}_0$$

using the likelihood ratio χ^2 statistic given by

$$\begin{aligned} -2\left[l(\hat{\mu}_0;y) - l(\hat{\mu}_1;y)\right] &= -2\left[l(\hat{\mu}_0;y) - l(y;y)\right] - \{-2\left[l(\hat{\mu}_1;y) - l(y;y)\right]\} \\ &= D(y;\hat{\mu}_0) - D(y;\hat{\mu}_1) \\ &\approx \chi^2_{\text{df}}, \end{aligned}$$

where df $= p_1 - p_0$, p_1 is the number of parameters in $\mathcal{M}_1$, and p_0 is the number of parameters in $\mathcal{M}_0$. (Note that the χ^2 approximation is often adequate here even it is not adequate for the saturated model provided that $\mathcal{M}_1$ is not too close to saturated.)

Information criteria

Most of the details in this section on information criteria can be found in Charles Geyer's lecture notes for his categorical data analysis class.

AIC

In the early 1970's Akaike proposed the first information criterion. Later many others were proposed, so Akaike's is now called the Akaike information criterion (AIC).

The "information" in AIC is Kullback-Leibler information, which is the concept of Shannon information imported into statistics, which in turn is the concept of entropy imported from statistical physics into communications theory. It is expected log likelihood, which is what the maximized value of the log likelihood is trying to estimate. See also the maximum entropy section of these notes on exponential families.

What Akaike discovered is that the maximized value of the log likelihood is a biased estimate of Kullback-Leibler information. It overestimates it by p, the dimension of the model (number of parameters). Or, what is the same thing, the likelihood ratio test (LRT) statistic, which is minus twice the (maximized value of the) log likelihood, underestimates its expectation by $2p$. So

$$\text{AIC} = -2l(\hat{\beta}) + 2p$$

All of this is a "large sample size" result based on the "usual" asymptotics of maximum likelihood, which is not valid for all statistical models. But it is always valid for exponential family models in general and categorical data analysis in particular (when sample size is "large").

AIC Versus Hypothesis Tests

The central dogma of hypothesis testing is “do only one test” or if you do more than one test, you must correct P-values to account for doing more than one test. So the theory of hypothesis tests is the Wrong Thing for comparing many models. And AIC is the Right Thing or at least *a* Right Thing.

Approaches based on hypothesis testing, such as that implemented by R function `step` come with no guarantees of doing anything correct. Also, hypothesis tests comparing fitted models like those above require that the models being compared be nested. AIC does not have any such requirement.

BIC

In the late 1970’s Schwarz proposed another information criterion, which is now usually called the Bayesian information criterion (BIC). Its formula is

$$\text{BIC} = -2l(\hat{\beta}) + \log(n) \cdot p$$

Since $\log(n) \geq 2$ for $n \geq 8$, BIC penalizes larger models more than AIC. BIC always selects smaller models than AIC.

The reason BIC is called “Bayesian” is that, if $\text{BIC}(m)$ denotes the BIC for model m and $g(m)$ denotes the prior probability for model m, then

$$\Pr(m \mid \text{data}) \approx \frac{\exp\bigl(-\frac{1}{2}\,\text{BIC}(m)\bigr)g(m)}{\sum_{m^* \in \mathcal{M}} \exp\bigl(-\frac{1}{2}\,\text{BIC}(m^*)\bigr)g(m^*)}$$

where $\mathcal{M}$ is the class of models under consideration.

This is a “large sample size” result based on the “usual” asymptotics of Bayesian inference (the Bernsteinâ€“von Mises theorem), which is not valid for all statistical models. But it is always valid for exponential family models in general and categorical data analysis in particular (when sample size is “large”).

When we use a flat prior (g is a constant function of m), the prior g cancels out of the formula, and we obtain

$$\text{BIC}(m) \approx -2 \log \Pr(m \mid \text{data}) + \text{a constant}$$

Clearly, BIC is defined the way it is to be comparable to AIC, not to produce the simplest Bayesian formulas.

AIC Versus BIC

In model selection AIC and BIC do two different jobs. No selection criterion can do both jobs [Yang, 2005]:

- BIC provides consistent model selection when the true unknown model is among the models under considexration.
- AIC is minimax-rate optimal (details of minimax-rate optimality are beyond the scope of this course) for estimating the regression function and other probabilities and expectations. It does not need the true unknown model to be among the models under consideration.

Assuming the true unknown model to be among the models under consideration, and Bayesians have to assume this - not among the models under consideration means prior probability zero and posterior probability zero - selecting the model with smallest BIC will select the true unknown model with probability that goes to one as n goes to infinity. Of course, that does not mean BIC is guaranteed to select the correct model at any finite sample size.

If we do not assume the true unknown model is among the models under consideration, then we only have AIC as an option. It generally does not do consistent model selection. However, it does give the best predictions of probabilities and expectations of random variables in the model. It is using the models under consideration to give the best predictions of probabilities and expectations under the true unknown model (which need not be among the models under consideration).

In short,

- use BIC when the true unknown model is assumed to be among the models under consideration, but
- use AIC when we do not want to assume this.

A shorthand often used is "sparsity". The *sparsity* assumption is that the true unknown model has only a few parameters and is one of the models under consideration. Under sparsity, BIC does consistent model selection. If you do not want to assume sparsity, then use AIC.

Optimization

In this section we discuss optimization routines for estimating parameters in canonical exponential family linear submodels. The goal will be to find

$$\text{argmax}_{\beta} l(\beta) \quad = \quad \text{argmax}_{\beta} \left[\langle M^T y, \beta \rangle - c(M\beta) \right] \tag{23}$$

in identifiable models. Note that we will blend our notation with the notation in Chapter 4 of Agresti [2013] when we define the Newton-Raphson, Fisher scoring, and iteratively reweighted least squares (IRLS) algorithms.

Newton-Raphson algorithm

A classic algorithm for handling iterative solutions of nonlinear systems of equations is the *Newton-Raphson algorithm*. This algorithm begins with an initial guess β_0 for the solution. It obtains a second guess by approximating the function to be maximized in a neighborhood of the initial guess by a second-degree polynomial and then finding the location of the polynomial's maximum value. This process is repeated iteratively until the discrepancy in successive evaluations of the objective function evaluate along the sequence of iterates is smaller than some convergence threshold. The sequence of iterates that this algorithm generates converge to a solution $\hat{\beta}$ when the optimization function is suitable (full rank properly conditioned Fisher Information matrix) and/or the initial guess is good.

We now explain the Newton-Raphson algorithm formally. Let $U(\beta) = \nabla l(\beta)$ be the score function corresponding to the log likelihood of a canonical linear exponential family submodel, and let $H(\beta) = \nabla^2 l(\beta)$ denote the Hessian matrix. At iteration k, consider the following second order Taylor series approximation of $l(\beta)$,

$$l(\beta) \approx l(\beta_k) + U(\beta_k)^T(\beta - \beta_k) + \frac{(\beta - \beta_k)^T H(\beta_k)(\beta - \beta_k)}{2}. \tag{24}$$

Now solving

$$U(\beta) \approx U(\beta_k) + H(\beta_k)(\beta - \beta_k) = 0$$

for β yields the next guess. That guess is

$$\beta_{k+1} = \beta_k - H(\beta_k)^{-1} U(\beta_k). \tag{25}$$

This algorithm is fast, exhibits quadratic convergence, provided that it converges. Convergence is likely in identifiable models where $H(\beta_0)$ is positive definite. However, the Newton-Raphson method can be quite sensitive to the choice of starting values β_0.

For many identifiable GLMs, including Poisson models with log link and binomial models with logit link, with full rank model matrices the Hessian is negative definite and the log likelihood is a strictly concave function. The maximum likelihood estimators of model parameters exist and are unique under quite general conditions [Wedderburn, 1976]. Thus Newton Raphson should exhibit solid performance for finding MLEs of the submodel canonical parameter vector.

Fisher scoring algorithm

The *Fisher scoring algorithm* is an alternative method for solving systems of equations. It resembles the Newton-Raphson algorithm, the distinction being with the Hessian matrix used in the Newton updates.

Fisher scoring uses the expected Fisher information matrix instead of the Hessian which is the observed Fisher information matrix.

We will let $\mathcal{H}$ be the expected information matrix so that $\mathcal{H}(\beta) = -\mathrm{E}\left\{\nabla^2 l(\beta)\right\}$. The Newton update step for the Fisher scoring method is

$$\beta_{k+1} = \beta_k + \{\mathcal{H}(\beta_k)\}^{-1} U(\beta_k). \tag{26}$$

For GLMs with canonical link (the entirety of these notes thus far), we have that the observed and expected information are the same so that Fisher scoring and Newton-Raphson are the same in our context. This is a consequence of the observed equals expected property $\hat{\mu} = y$ that is presented in our notes above. For noncanonical link models (which we see later), these quantities are not the same and Fisher scoring has advantages over Newton Raphson in that it produces the asymptotic covariance matrix as a by-product, the expected information is necessarily nonnegative definite, and as seen next, it is closely related to weighted least-squares methods for ordinary linear models. However, it need have quadratic convergence convergence, and for complex models the observed information is often easier to calculate.

Iteratively reweighted least squares (IRLS)

We now derive the IRLS algorithm from Fisher scoring. Take the Newton update of the Fisher scoring algorithm (26) and right multiply by the Hessian matrix so that

$$\mathcal{H}(\beta_k)\beta_{k+1} = \mathcal{H}(\beta_k)\beta_k + U(\beta_k). \tag{27}$$

The chain rule and an appeal to Figure 1 allows us to write $\mathcal{H}(\beta) = M^T W(\beta) M$ where

$$W(\beta) = \left\{\nabla_\theta^2 c(\theta^*)\right\}|_{\theta^* = M\beta}.$$

The estimated asymptotic covariance matrix $\mathcal{H}^{-1}$ of $\hat{\beta}$ occurs as a by-product of this algorithm as $\{\mathcal{H}(\beta_k)\}^{-1}$ where k is an iteration number at which convergence is deemed to have occurred.

For both Fisher scoring and Newton-Raphson, the score function $U(\beta)$ can be written as

$$\begin{aligned} U(\beta) &= \nabla l(\beta) \\ &= M^T\left(Y - \nabla_\theta c(\theta^*)|_{\theta^*=M\beta}\right) \\ &= M^T\left(Y - \mu(\beta)\right), \end{aligned}$$

where $\mu(\beta)$ is the saturated model mean value parameter expressed as function of β. With this setup we can see that the right-hand side of (27) can be written as

$$\begin{aligned} \mathcal{H}(\beta_k)\beta_k + U(\beta_k) &= M^T W(\beta_k) M\beta_k + M^T(Y - \mu(\beta_k)) \\ &= M^T W(\beta_k)\left(M\beta_k + W(\beta_k)^{-1}(Y - \mu(\beta_k))\right), \end{aligned}$$

and the left-hand side of (27) can be written as

$$\mathcal{H}(\beta_k)\beta_{k+1} = M^T W(\beta_k) M\beta_{k+1}.$$

Putting this together and solving for the update β_{k+1} yields

$$\beta_{k+1} = \left(M^T W(\beta_k) M\right)^{-1} M^T W(\beta_k)\left(M\beta_k + W(\beta_k)^{-1}(Y - \mu(\beta_k))\right).$$

Now let

$$z(\beta_k) = M\beta_k + W(\beta_k)^{-1}(Y - \mu(\beta_k)).$$

Then we see that the update is given as

$$\beta_{k+1} = \left(M^T W(\beta_k) M\right)^{-1} M^T W(\beta_k) z(\beta_k).$$

This update step is the same as a weighted least squares regression of $z(\beta_k)$ on M with weights $W(\beta_k)$. Hence the name iteratively reweighted least squares.

The R programming language's fitting function for generalized linear models `glm.fit` implements an IRLS algorithm to compute $\hat{\beta}$. Although it should be noted that `glm.fit`'s actual IRLS algorithm implementation uses C code.

Other optimization methods

We will close this section on optimization by briefly discussing Quasi-Newton methods which are appropriate when the Hessian matrix used in a Newton update (25) are not available, and stochastic gradient descent. These optimization methods are important in practice when models are large. However, our primary focus will be on the above optimization methods in this course.

Quasi-Newton methods Quasi-Newton methods are modifications of (25), where the Hessian matrix is approximated (by the secant method for example), typically because an explicit formula for the Hessian matrix is not available or cumbersome to compute. One such Quasi-Newton method is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm for solving unconstrained nonlinear optimization problems. We present the steps for the BFGS algorithm in our optimization context (23) where l is a differentiable scalar function.

The algorithm begins at a user-specified initial value (initial estimate for the optimal value) $\beta_{k=0}$ and an initial guess B_0. Then proceed iteratively to get a better estimate of β at each stage. The steps are:

1. Obtain a direction x_k by solving $B_k x_k = \nabla_\beta l(\beta_k)$.
2. Perform a one-dimensional optimization (line search) to find an acceptable stepsize α_k to be made in the direction x_k, so that $x_k = \arg\min -l(\beta_k + \alpha x_k)$.
3. Set $s_k = \alpha_k x_k$ and update $\beta_{k+1} = \beta_k + s_k$.
4. Set $y_k = \nabla_\beta(\beta_k) - \nabla_\beta(\beta_{k+1})$.
5. Update
$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k^T}{s_k^T B_k s_k}.$$

The BFGS algorithm is a possible optimization method in R's optim function. There is also a variant of BFGS in R called L-BFGS-B. This algorithm extends a limited memory variant of BFGS to handle bound constraints in the domain space of $l(\beta)$. Bound constraints are of the form $l_i \leq \beta_i \leq u_i$ where β_i is one component of the canonical parameter vector β.

Stochastic Gradient Decent (SGD) The SGD algorithm with averaging [Polyak and Juditsky, 1992] is a computationally fast, and scalable optimization method that is applicable in large scale applications with applications in online learning. We will consider SGD in our context for estimating $\hat{\beta}$ in exponential family models. The SGD algorithm only requires use of one (or relatively few) data points at each iteration, which makes it computationally superior to other optimization routines. Conventional implementations average over all iterations to derive the final estimator. Consistency and asymptotic normality properties of SGD have been verified in Polyak and Juditsky [1992] for both parameter estimation. In addition to these results, Rakhlin et al. [2011] derived the optimal rates of $O(1/n)$ for the objective function under smoothness and strong convexity assumptions. These assumptions hold in our modeling context.

We will suppose that we have data $D_n = Z_1, \ldots, Z_n$ drawn as iid copies of $Z = (y, x)$. In this case we assume that the both the response and the predictors are random from some common generative process. We will suppose that the conditional distribution of interest $y|x$ can then be parameterized as an exponential family with log likelihood (20) where the matrix M has rows x_i^T, $i = 1, \ldots, n$. The SGD algorithm updates as follows:
$$\hat{\beta}_k = \hat{\beta}_{k-1} + \eta_k \nabla l(\hat{\beta}_{k-1}; Z_k)$$
for $k = 1, \ldots, n$ where η is called the learning rate. We set the learning rate to be $\eta_k = Ck^{-\alpha}$ for some $\alpha \in (0, 1)$. The final SGD estimators is then
$$\hat{\beta} = \frac{1}{n}\sum_{k=1}^{n} \hat{\beta}_k.$$

We will revisit optimization when we discuss specific generalized linear models.

Miscellaneous topics and concluding remarks

Sufficiency

A (possibly vector-valued) statistic is *sufficient* if the conditional distribution of the full data given this statistic does not depend on the parameter.

The interpretation is that the full data provides no information about the parameter that is not already provided by the sufficient statistic. The principle of sufficiency follows: all inference should depend on the data only through sufficient statistics. The Fisher-Neyman factorization criterion [Lehmann, 1959, Corollary 1 of Chapter 2] says that a statistic is sufficient if and only if the likelihood depends on the whole data only through that statistic.

Lemma 3. *The canonical statistic vector of an exponential family is a sufficient statistic.*

The proof of the above Lemma is left as an exercise for the reader.

Sufficient dimension reduction is a whole field of study. However, the *original* "sufficient dimension reduction" theory was about exponential families. The so-called Pitman-Koopman-Darmois theorem (proved independently by three different persons in 1935 and 1936) says that

> when we have iid sampling from a statistical model, all distributions in the model have the same support which does not depend on the parameter, and all distributions in the model are continuous, then there is a sufficient statistic whose dimension does not depend on the parameter if and only if the statistical model is an exponential family of distributions.

This theorem was responsible for the interest in exponential families early in the twentieth century. The condition of the Pitman-Koopman-Darmois theorem that the support does not depend on the parameter is essential. For IID sampling from the Uniform$(0, \theta)$ model the maximal order statistic $X_{(n)}$ is sufficient. Its dimension (one) does not depend on n. To show this note that the likelihood is

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} \frac{1}{\theta} I_{(0,\theta)}(X_i) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} I_{(0,\theta)}(X_i) \\ &= \frac{1}{\theta^n} I_{(0,\theta)}(X_{(n)}) \end{aligned}$$

because if $X_{(n)} < \theta$ then $X_i < \theta$ for all i,

The condition that the statistical model has to be continuous is ugly. Many of the most important applications of exponential family theory (logistic and Poisson regression, log-linear models for categorical data) are discrete, and the theorem does not say anything about them. But later theorems that did cover discrete distributions need extra conditions that seem just there so the theorem can be proved.

Interest in exponential families changed direction in the 1970's with the invention of generalized linear models [Nelder and Wedderburn, 1972, Wedderburn, 1974] and log-linear models for categorical data [Bishop et al., 2007] and with the publication of authoritative treatises [Barndorff-Nielsen, 2014, Brown, 1986] which used the mathematics of convex analysis [Rockafellar, 1970].

In that context the sufficient dimension reduction for canonical linear submodels (exponential family regression models) became more important than the Pitman-Koopman-Darmois property. This is the relation between the canonical sufficient statistic y of the saturated model and the canonical sufficient statistic $M^T y$ of a canonical linear submodel. The former has the row dimension of M and the latter has the column dimension of M, which is usually much smaller.

Maximum entropy

Entropy is a physical quantity involved in the second law of thermodynamics, which says that the the total entropy of an isolated physical system is nondecreasing in any physical process. It has to do with the maximum possible efficiency of a heat engine or refrigerator, with which chemical reactions proceed spontaneously, and with many other things.

Ludwig Boltzmann and Josiah Willard Gibbs figured out the connection between entropy and probability and between the thermodynamic properties of bulk matter and the motions and interactions of atoms and molecules.

In this theory entropy is not certain to increase to its maximum possible value. It is only overwhelmingly probable to do so in any large system. In a very small system, such as a cubic micrometer of air, it is less probable that entropy will be near its maximum value. In such a small system the statistical fluctuations are large. This is the physical manifestation of the law of large numbers. The larger the sample size (the more molecules involved) the less stochastic variation. Boltzmann thought this discovery so important that he had $S = k \log W$ inscribed on his tombstone (S is entropy, W is probability, and k is a constant of nature now known as Boltzmann's constant).

Claude Shannon imported entropy into information theory, using it to determine the maximum throughput of a noisy communication channel. Shannon information is negative entropy (minus log probability). Kullback and Leibler imported the same concept into statistics, where it is usually called *Kullback-Leibler information*. It is expected log likelihood and hence what likelihood attempts to estimate.

Edwin Jaynes, a physicist, introduced the "maximum entropy formalism" that describes exponential families in terms of entropy. To keep the derivation simple, we will do the finite sample space case. The same idea can be extended to the infinite discrete case or the continuous case, although the math is harder.

The *relative entropy* of a distribution with PMF f to a distribution with PMF m is defined to be

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

where S is the support of the distribution with PMF m. (It is the negative of this quantity that is Kullback-Leibler information of f with respect to m.) It is actually not necessary that m be a PMF; any positive function will do.

Suppose we "know" the value of some expectations

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

and we want f to maximize entropy subject to these constraints plus the constraints that f is nonnegative and sums to one. That is, we want to solve the following optimization problem

$$\begin{aligned} \text{maximize} \quad & - \sum_{x \in S} f(x) \log \left(\frac{f(x)}{m(x)} \right) \\ \text{subject to} \quad & \sum_{x \in S} t_j(x) f(x) = \mu_j, \qquad j \in J \\ & \sum_{x \in S} f(x) = 1 \\ & f(x) \geq 0, \qquad x \in S \end{aligned}$$

It turns out that the inequality constraints here are unnecessary. If we solve the problem without requiring f be nonnegative, the solution happens to be nonnegative. But we do need to enforce the equality constraints.

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

$$\begin{aligned}\mathcal{L}(f) &= -\sum_{x \in S} f(x) \log\left(\frac{f(x)}{m(x)}\right) + \sum_{j \in J} \theta_j \sum_{x \in S} t_j(x) f(x) + \psi \sum_{x \in S} f(x) \\ &= -\sum_{x \in S} f(x) \left[\log\left(\frac{f(x)}{m(x)}\right) - \sum_{j \in J} \theta_j t_j(x) - \psi\right],\end{aligned}$$

where θ_j and ψ are the Lagrange multipliers.

Because the domain of f is finite, we can think of it as a vector having components $f(x)$. The Lagrangian is maximized where its first derivative is zero, so we calculate the first partial derivatives as

$$\frac{\partial \mathcal{L}(f)}{\partial f(x)} = -\log\left(\frac{f(x)}{m(x)}\right) + \sum_{j \in J} \theta_j t_j(x) + \psi - 1,$$

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

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

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

$$f(x) = m(x) \exp\left(t(x)^T \theta + \psi - 1\right).$$

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

$$e^{\psi - 1} \sum_{x \in S} m(x) e^{t(x)^T \theta} = 1$$

or

$$1 - \psi = \log\left(\sum_{x \in S} m(x) e^{t(x)^T \theta}\right).$$

Now define

$$c(\theta) = \log\left(\sum_{x \in S} m(x) e^{t(x)^T \theta}\right).$$

Then,

$$f(x) = m(x) e^{t(x)^T \theta - c(\theta)},$$

and this is the density of an exponential family! If we think of the Lagrange multipliers θ_j as unknown parameters rather than constants we still have to adjust, then we see that we have an exponential family with canonical statistic vector $t(x)$, canonical parameter vector θ, and cumulant function $c(\theta)$.

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

$$\mu = \nabla c(\theta) = g(\theta)$$

and g is a one-to-one function (if the exponential family is identifiable, which happens if there are no redundant constraints), so the Lagrange multiplier vector is

$$\theta = g^{-1}(\mu)$$

and, although we do not have a closed form expression for g^{-1}, we can evaluate $g^{-1}(\mu)$ for any μ that is a possible mean value parameter vector found by optimization. Our use of the maximum entropy argument is a bit peculiar. First we said that we "knew" the expectations

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

and wanted to pick out one probability distribution that maximizes entropy and satisfies this constraint. Then we forgot about "knowing" this constraint and said as μ ranges over all possible values we get an exponential family of probability distributions. Also we have to choose a base measure.

Despite this rather odd logic, the maximum entropy argument does say something important about exponential families. Suppose we have a big exponential family (think a saturated model) and are interested in submodels. Examples are Bernoulli regression, Poisson regression, or categorical data analysis. The maximum entropy argument says the canonical linear submodels are the submodels that, *subject to constraining the means of their submodel canonical statistics, leave all other aspects of the data as random as possible*, where "as random as possible" means maximum entropy. Thus these models constrain the means of their canonical statistics and anti-constrain (leave as unconstrained as possible) everything else.

In choosing a particular canonical linear submodel parameterization $\theta = M\beta$ we are, in effect, modeling only the the distribution of the submodel canonical statistic $t(y) = M^T y$, leaving all other aspects of the distribution of y as random as possible given the control over the distribution of $t(y)$.

Overdispersion

A common explanation for large deviance (or poor fit) is the presence of a few outliers. When large number of points are identified as outliers, they become unexceptional, and it may be the case that the error distribution is misspecified. When more outliers are observed than expected by a model, the variance may need to be adjusted to accommodate the outliers. This is referred to as overdispersion and it is often encountered in Possion regression.

In the presence of misspecification in the form of overdispersion, the exponential family takes on a different functional form

$$f(y|\theta,\phi) = \exp\left(\frac{\langle y,\theta\rangle - c(\theta)}{a(\phi)} - b(y,\phi)\right), \tag{28}$$

where y and θ are as before, ϕ is a dispersion parameter, and $b(y,\phi)$ is a function of the data y and the dispersion parameter ϕ. From the perspective of the canonical exponential families that we have motivated throughout, the function $b(y,\phi)$ is similar to the base measure h that was dropped from consideration in log likelihood based arguments that focused on the parameters. Notice that the density (28) is a generalization of the exponential family density which specifies that $a(\phi) = 1$ and $b(y,\phi) = \log(h(y))$. Note that the dispersion parameter can be estimated using

$$\hat{\phi} = \frac{\sum_{i=1}^n (y - \hat{\mu}_i)^2/\hat{\mu}_i}{n-p}.$$

We will cover overdispersion in much more detail when we cover generalized linear models for count responses.

Concluding remarks

So now we can put all of this together to discuss interpretation of regular full exponential families and their canonical linear submodels.

The MLE is unique if it exists (from strict concavity). Existence is a complicated story, and non-existence results in complicated problems of interpretation, which we leave for now. We will revisit non-existence later.

The MLE satisfies the observed equals expected property, either (16) for a saturated model or (21) for a canonical linear submodel.

The sufficient dimension reduction property and maximum entropy property say that $M^T y$ is a sufficient statistic, hence captures all information about the parameter. All other aspects of the distribution of y are left as random as possible; the canonical linear submodel does not constrain them in any way other than its constraints on the expectation of $M^T y$.

The same model can be specified by different formulas or different model matrices, so that a particular canonical parameter value does not specify a model uniquely. A quote from Charlie Geyer:

> "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. Hence interpretations should focus on mean value parameters. This conclusion flies in the face the traditional way regression models are taught. In most courses, students are taught to "interpret" the equation $\theta = M\beta$, or, more precisely, since in lower level courses students aren't assumed to know about matrices, students are taught to interpret this with the matrix multiplication written out explicitly, interpreting equations like

$$\theta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2, \qquad \text{where } i \text{ runs over cases.}$$

The model matrix M determines two linear transformations

$$\begin{aligned} \beta &\mapsto M\beta \\ \mu &\mapsto M^T\mu \end{aligned}$$

The second equation, which takes saturated model canonical statistic to submodel canonical statistic and saturated model mean value parameter to submodel mean value parameter, is the more important of the two and should lead in interpretation, because the former is about canonical parameters (the meaningless ones) and the latter is about mean value parameters (the meaningful ones). This is especially so in light of the fact that $M^T y = M^T \hat{\mu}$ (observed equals expected) is the only algebraically simple property of maximum likelihood that users can hang an interpretation on.

When we do need to think about canonical parameters, the key concept is the multivariate monotone relationship (13) between canonical and mean value parameters. Note that this holds not only for saturated model parameters but also for canonical linear submodel parameters. If, as before, we let $\tau = M^T\mu$ denote the submodel mean value parameter, and τ_1 corresponds to β_1 and τ_2 to β_2, then

$$(\tau_1 - \tau_2)^T(\beta_1 - \beta_2) > 0, \qquad \text{whenever } \tau_1 \neq \tau_2.$$

By standard theory of maximum likelihood, MLEs of all parameters are consistent, efficient (have minimum asymptotic variance), and asymptotically normal, with easily calculated asymptotic variance (inverse Fisher information matrix). Fisher information is easily calculated, (15) is Fisher information for the saturated model canonical parameter θ;

$$\nabla^2_\beta c(M\beta) = M^T \left(\nabla c(M\beta)\right) M$$

is Fisher information for the submodel canonical parameter β.

The Delta method then gives asymptotic variance matrices for mean value parameters. If $\mu = h(\theta)$, then the asymptotic variance for $\hat{\mu}$ is

$$[\nabla h(\theta)] I(\theta)^{-1} [\nabla h(\theta)]^T = I(\theta) I(\theta)^{-1} I(\theta) = I(\theta)$$

and $M^T I(\theta) M$ is the asymptotic variance for $\hat{\tau}$. These can be used for hypothesis tests and confidence intervals about these other parameters.

Acknowledgments

These notes take materials from Charles Geyer's notes on exponential families, model selection, and other topics. We also borrow materials from Trevor Park's STAT 426 notes and Agresti [2013].

References

Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, 2013.

Ole Barndorff-Nielsen. *Information and exponential families: in statistical theory*. John Wiley & Sons, 2014.

Yvonne M Bishop, Stephen E Fienberg, and Paul W Holland. *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media, 2007.

Lawrence D Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986.

Erich L Lehmann. *Testing Statistical Hypotheses*. New York: John Wiley, 1959.

John A Nelder. Statistics, science and technology. *Journal of the Royal Statistical Society: Series A (General)*, 149(2):109–121, 1986.

John A Nelder. Statistics for the millennium: from statistics to statistical science. *Journal of the Royal Statistical Society Series D: The Statistician*, 48(2):257–269, 1999.

John A Nelder and Robert W M Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton University Press, 1970.

R Tyrrell Rockafellar and Roger J B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 1998. (The corrected printings contain extensive changes. We used the 3rd corrected printing, 2010.).

Ruth G Shaw and Charles J Geyer. Inferring fitness landscapes. *Evolution*, 64(9):2510–2520, 2010.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Robert W M Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447, 1974.

Robert W M Wedderburn. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63(3):27–32, 1976.

Yuhong Yang. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.