Exponential Family Notes

Daniel J. Eck

Contents

Introduction

One of the main themes of this course will be developing regression models and demonstrating their utility as methods for data analysis. We will see that the structure of data 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 regression models. In this course we are going to primarily focus on 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, 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 alumnus 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 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 takes on the above Russell 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 workers.**"

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 or 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 analyses clearly.

Nonetheless, mathematics has an important role in this course, as perhaps best articulated by Nelder [1999]:

> "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:

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

```
## 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
```

```
## 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
```

```
## 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
```

```
## 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,

$$1 = \int f_\theta(w) dw$$

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

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

Example (Normal 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*

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

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*

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

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

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

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.

Example (Normal 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

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

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

$$f_\theta(y) = f_\psi(y) e^{\langle y, \theta - \psi \rangle - c(\theta) + c(\psi)}$$

$$= \binom{n}{y} \left(\prod_{i=1}^{d} \left[p_i e^{\theta_i - \psi_i} \right]^{y_i} \right) \left(\sum_{i=1}^{d} p_i e^{\theta_i - \psi_i} \right)^{-n}$$
$$= \binom{n}{y} \prod_{i=1}^{d} \left(\frac{p_i e^{\theta_i - \psi_i}}{\sum_{j=1}^{d} p_j e^{\theta_j - \psi_j}} \right)^{y_i}.$$

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).

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.

[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.

```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

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

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

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

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

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

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*

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

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*

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

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

Proposition 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*

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

Proposition 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 Proposition 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,

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

Proposition 2 implies that

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

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

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

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.

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.

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.