Exponential Family Notes

Daniel J. Eck

Contents

Introduction **1**

Definitions and properties of exponential families **4**
- Log likelihood . . . 4
- Densities . . . 5
- Cumulant functions . . . 5
- Ratios of densities . . . 6
- Full families . . . 6
- Moment and cumulant generating functions . . . 7
- Regular exponential families . . . 8
- Identifiability and directions of constancy . . . 8

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:

```
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)
```

```r
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*

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

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.

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.

References

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.