Exponential families

"Algebraic & Geometric Methods in Statistics"

Sonja Petrović
Created for Math/Stat 561

Feb 1, 2023.

Objectives

- Wrap up introductory overview of maximum likelihood estimation
  - there will be more material on this from chapter 7 soon!
  - see an example with observed data for the model of 2×2 independence.
- Understand the setup of exponential families.
  - broad class of models
  - structure $\leftrightarrow$ sufficient statistics, MLE, etc.

Material:

Sourced from chapter 6 ("exponential families") of the textbook. Other resources provided in subsequent links.

Maximum likelihood estimation

I.i.d. sampling: $L(\theta|D) = \prod_{i=1}^{n} L(\theta|X^{(i)})$

- Likelihood function (discrete case): $L(\theta|D) = \prod_{i=1}^{n} p_\theta(X^{(i)})$
- Let $u \in \mathbb{N}^r$ be the vector of counts, i.e. $u_j = \#\{i : X^{(i)} = j\}$: $L(\theta|D) = \prod_{i=1}^{n} p_\theta(X^{(i)}) = \prod_{j=1}^{r} p_\theta(j)^{u_j}$
- Example for $\left\{ \left(\theta^2, 2\theta(1-\theta), (1-\theta)^2\right) : \theta \in [0,1] \right\}$: $L(\theta|D) = (\theta^2)^{u_0} \cdot (2\theta(1-\theta))^{u_1} \cdot ((1-\theta)^2)^{u_2}$
- Likelihood function (continuous case): $L(\theta|D) = \prod_{i=1}^{n} f_\theta(X^{(i)})$

Figure 1: Source: K.Kubjas

Log-likelihood function

- The log-likelihood function is

$$l(\theta \mid D) = \log L(\theta \mid D)$$

- I.i.d. data: turns a product into a sum
- Example:
  - $L(\theta \mid D) = (\theta^2)^{u_0} \cdot (2\theta(1-\theta))^{u_1} \cdot ((1-\theta)^2)^{u_2}$
  - $l(\theta \mid D) = u_0 \log(\theta^2) + u_1 \log(2\theta(1-\theta)) + u_2 \log((1-\theta)^2)$
- The likelihood and log-likelihood function have the same maximizer, because logarithm is a monotone function

Figure 2: Source: K.Kubjas

Score equations

Let $\Theta \subseteq \mathbb{R}^d$ be an open full-dimensional parameter set.

Def: The score equations or critical equations of the model $\mathscr{M}_\Theta$ are the equations obtained by setting the gradient of the log-likelihood function to zero:

$$\frac{\partial}{\partial \theta_i} l(\theta \mid D) = 0, \quad i = 1, \ldots, d.$$

Figure 3: Source: K.Kubjas

Example: the independence model 2×2

$$\mathcal{M}_{X \perp\!\!\!\perp Y} = \{p = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \in \Delta_3 : p_{ij} = \alpha_i \beta_j, (\alpha, \beta) \in \Delta_1 \times \Delta_1\} \text{ and } u = \begin{pmatrix} 19 & 141 \\ 17 & 149 \end{pmatrix}$$

Log-likelihood function: $l(\alpha, \beta \,|\, u) = 160 \log \alpha_1 + 166 \log \alpha_2 + 36 \log \beta_1 + 290 \log \beta_2$

$$= 160 \log \alpha_1 + 166 \log(1 - \alpha_1) + 36 \log \beta_1 + 290 \log(1 - \beta_1)$$

Score equations:

$$\frac{\partial l(\alpha, \beta \,|\, u)}{\partial \alpha_1} = \frac{160}{\alpha_1} - \frac{166}{1 - \alpha_1} = 0$$

$$\frac{\partial l(\alpha, \beta \,|\, u)}{\partial \beta_1} = \frac{36}{\beta_1} - \frac{290}{1 - \beta_1} = 0$$

Figure 4: Source: K.Kubjas

Exponential families

- An exponential family is a *parametric statistical model* with probability distributions of a *certain form*.
- General enough to include many of the most common families of probability distributions:
  - multivariate normal
  - exponential
  - Poisson
  - binomial (with fixed number of trials)
- Specific enough to have nice properties:
  - likelihood function is strictly concave [next lecture]
  - have conjugate priors.

Objectives

- What is an exponential family?
- How to find the polynomial ideal of an exponential family?
  - Discrete exponential models: Hypothesis testing [future lecture]
  - Gaussian exponential submodels: Conditional independence implications [past lecture]

Exponential families

Let X be a random variable taking values in a set $\mathcal{X}$.

An exponential family is the set of probability distributions whose probability mass function or density function **can be expressed as**

$$f_\theta(x) = h(x)e^{\eta(\theta)' T(x) - A(\theta)},$$

for a given statistic $T(x) : \mathcal{X} \to \mathbb{R}^k$, natural parameter $\eta(\theta) : \Theta \to \mathbb{R}^k$, and functions $h : \mathcal{X} \to \mathbb{R}_{\geq 0}$ and $A : \Theta \to \mathbb{R}$.

Three equivalent forms:

$$f_\theta(x) = h(x)e^{\eta(\theta)' T(x) - A(\theta)},$$
$$f_\theta(x) = h(x)g(\theta)e^{\eta(\theta)' T(x)},$$
$$f_\theta(x) = e^{\eta(\theta)' T(x) - A(\theta) + B(x)}.$$

Example: Bernoulli

The Bernoulli distribution

The probability mass function of a Bernoulli random variable X is given as follows:

THE MODEL AS YOU WRITE IT NATURALLY

EXPONENTIAL FAMILY FORM

$$
\begin{aligned}
p(x \mid \pi) &= \pi^x (1-\pi)^{1-x} && (8.4)\\
&= \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\}. && (8.5)
\end{aligned}
$$

where our trick, here and throughout the chapter, is to take the exponential of the logarithm of the original distribution. Thus we see that the Bernoulli distribution is an exponential family distribution with:

PIECES OF THE EXPONENTIAL FAMILY FORM

$$
\begin{aligned}
\eta &= \frac{\pi}{1-\pi} && (8.6)\\
T(x) &= x && (8.7)\\
A(\eta) &= -\log(1-\pi) = \log(1+e^{\eta}) && (8.8)\\
h(x) &= 1. && (8.9)
\end{aligned}
$$

Note moreover that the relationship between η and π is invertible. Solving Eq. (8.6) for π, we have:

$$\pi = \frac{1}{1+e^{-\eta}}, \qquad (8.10)$$

which is the logistic function.

Figure 5: Source: Michael I. Jordan's notes on exponenial families

Example: binomial distribution.

$X \sim Bin(r, \theta)$, $\mathcal{X} = \{0, \ldots, r\}$.

$$p(x) = \binom{r}{x} \theta^x (1-\theta)^{r-x} =$$

Example: binomial distribution.

$X \sim Bin(r, \theta)$, $\mathcal{X} = \{0, \ldots, r\}$.

$$p(x) = \binom{r}{x} \theta^x (1-\theta)^{r-x} =$$

$$= \binom{r}{x} \exp\left[\left(\log \frac{\theta}{1-\theta}\right) x + r \log(1-\theta)\right].$$

Example: binomial distribution.

$X \sim Bin(r, \theta)$, $\mathcal{X} = \{0, \dots, r\}$.

$$p(x) = \binom{r}{x} \theta^x (1-\theta)^{r-x} =$$

$$= \binom{r}{x} \exp\left[\left(\log \frac{\theta}{1-\theta}\right) x + r \log(1-\theta)\right].$$

Question: What are the k, T, η, h, A in this example?

1. $k = 1,\ T(x) = \log \frac{\theta}{1-\theta}, \eta = x, h = \binom{r}{x}, A = -r \log(1-\theta)$
2. $k = 1,\ T(x) = x, \eta = \log \frac{\theta}{1-\theta}, h = \binom{r}{x}, A = -r \log(1-\theta)$
3. $k = 2,\ T(x) = (x, r-x),\ \eta = (\theta, 1-\theta),\ h = \binom{r}{x},\ A = 0.$

Example: binomial distribution.

$X \sim Bin(r, \theta)$, $\mathcal{X} = \{0, \dots, r\}$.

$$p(x) = \binom{r}{x} \theta^x (1-\theta)^{r-x} =$$

$$= \binom{r}{x} \exp\left[\left(\log \frac{\theta}{1-\theta}\right) x + r \log(1-\theta)\right].$$

Question: What are the k, T, η, h, A in this example?

1. $k = 1$, $T(x) = \log \frac{\theta}{1-\theta}$,$\eta = x$,$h = \binom{r}{x}$,$A = -r\log(1-\theta)$. $\leftarrow$ wrong, because T should not depend on the parameters.

2. $k = 1$, $T(x) = x$,$\eta = \log \frac{\theta}{1-\theta}$,$h = \binom{r}{x}$,$A = -r\log(1-\theta)$ $\leftarrow$ correct.

3. $k = 2$, $T(x) = (x, r-x)$, $\eta = (\theta, 1-\theta)$, $h = \binom{r}{x}$, $A = 0$.$\leftarrow$ wrong.

Canonical form

$$f_\theta(x) = h(x)e^{\eta(\theta)' T(x) - A(\theta)}.$$

- If $\eta(\theta) = \theta$, then the exp.fam. is said to be in canonical form.
- By defining a transformed parameter $\eta = \eta(\theta)$, it is *always possible to convert an exponential family to canonical form*.
- The function A is determined by the other functions: It makes the pdf (pmf) to integrate (sum) to one. Thus it can be written as a function of η.
- The canonical form is $f_\eta(x) = h(x)e^{\eta' T(x) - A(\eta)}$.

Independence model in canonical form

Task

Write down the model of independence of two random variables, $\mathcal{M}_{1 \perp\!\!\!\perp 2}$, in exponential family form.

Here are the key steps:

- Starting point is the parametric description of the model as you know it: The probability of observing the data count table u is given by $\prod_{1 \leq i \leq r_1, 1 \leq j \leq r_2} (\alpha_i \beta_j)^{u_{ij}}$.
- The product can be written using the same log-exp trick:

$$\prod_{1 \leq i \leq r_1, 1 \leq j \leq r_2} (\alpha_i \beta_j)^{u_{ij}} = \exp\left(\sum_{ij} u_{ij} \log(\alpha_i \beta_j) \right) = \ldots$$

- This will be useful to you on the homework, and in several projects.
- Note: Compare your work to the outline that is on the next slide, right column!

Discrete exponential families

... There has got to be a general strategy so we don't have to work through all the algebra every time?

- Note that we can use the fact that $e^{a \cdot b} = e^a \cdot e^b$ to write the last quantity from the previous slide in product form.

Notation

- X a discrete random variable $X \in [r]$.
- $T(x) = a_x$, writing as a vector: $a_x = (a_{1x}, \ldots, a_{kx})^t$
- $h(x) = h_x$, so $h = (h_1, \ldots, h_r)$ is also a vector (of positive real numbers)
- $\eta = (\eta_1, \ldots, \eta_k)^t$ and $\theta_i = \exp \eta_i$.

$$p_\eta(x) = h(x) e^{\eta^t T(x) - A(\eta)} =$$

Discrete exponential families

... There has got to be a general strategy so we don't have to work through all the algebra every time?

- Note that we can use the fact that $e^{a \cdot b} = e^a \cdot e^b$ to write the last quantity from the previous slide in product form.

Notation

- X a discrete random variable $X \in [r]$.
- $T(x) = a_x$, writing as a vector: $a_x = (a_{1x}, \ldots, a_{kx})^t$
- $h(x) = h_x$, so $h = (h_1, \ldots, h_r)$ is also a vector (of positive real numbers)
- $\eta = (\eta_1, \ldots, \eta_k)^t$ and $\theta_i = \exp \eta_i$.

$$p_\eta(x) = h(x) e^{\eta^t T(x) - A(\eta)} =$$

$$= h_x e^{\sum_i \eta_i a_{ix} - A(\eta)} = h_x \prod_i e^{\eta_i a_{ix} - A(\eta)}$$

$$= h_x \prod_i (e_i^{\eta})^{a_{ix}} e^{-A(\eta)} = h_x \prod_i \theta_i^{a_{ix}} \frac{1}{Z(\theta)}$$

where $Z(\theta) = \sum_{x \in [r]} h_x \prod_j \theta_j^{a_{jx}}$.

$$p_\theta(x) \propto \frac{1}{Z(\theta)} h_x \prod_i \theta_i^{a_{ix}}.$$

- If a_{jx} are integers for all j and x, then the parametrizing functions are rational functions.
- The entires a_{jx} can be recorded in matrix: $\mathcal{A} = (a_{jx})_{j\in[k],x\in[r]} \in \mathbb{Z}^{k\times r}$.
- For each value x of X, the monomial $\prod_j \theta_j^{a_{jx}} \leftrightarrow$ a column of $\mathcal{A}$.

$$p_\theta(x) \propto \frac{1}{Z(\theta)} h_x \prod_i \theta_i^{a_{ix}}.$$

Example: binomial random variable with three trials

Let X be the result of one trial, and we are interested in counting the number of successes in three consecutive trials.
There are two parameters, $\theta = (\theta_1, \theta_2)$. Let
$\theta_1 = P(X = 0), \theta_2 = P(X = 1)$.
Take $h = (1, 3, 3, 1)$ and $\mathcal{A} = ?$

$$p_\theta(x) \propto \frac{1}{Z(\theta)} h_x \prod_i \theta_i^{a_{ix}}.$$

Example: binomial random variable with three trials

Let X be the result of one trial, and we are interested in counting the number of successes in three consecutive trials.
There are two parameters, $\theta = (\theta_1, \theta_2)$. Let
$\theta_1 = P(X = 0), \theta_2 = P(X = 1)$.
Take $h = (1, 3, 3, 1)$ and $\mathcal{A} =?$

Design matrix recipe

Columns of $\mathcal{A}$ are exponents of the parametrization of each given state.

P(no 0)$=\theta_2^3$, P(one 0)$=\theta_1\theta_2^2$, P(two 0s)$=\theta_1^2\theta_2$, P(three 0s)$=\theta_1^3$.

So. . .

$$p_\theta(x) \propto \frac{1}{Z(\theta)} h_x \prod_i \theta_i^{a_{ix}}.$$

Example: binomial random variable with three trials

Let X be the result of one trial, and we are interested in counting the number of successes in three consecutive trials.
There are two parameters, $\theta = (\theta_1, \theta_2)$. Let
$\theta_1 = P(X = 0), \theta_2 = P(X = 1)$.
Take $h = (1, 3, 3, 1)$ and $\mathcal{A} =?$

Design matrix recipe

Columns of $\mathcal{A}$ are exponents of the parametrization of each given state.

P(no 0)$=\theta_2^3$, P(one 0)$=\theta_1\theta_2^2$, P(two 0s)$=\theta_1^2\theta_2$, P(three 0s)$=\theta_1^3$.

So... For the value $x = 0$, the corresopnding column a_{i0} should be

$$p_\theta(x) \propto \frac{1}{Z(\theta)} h_x \prod_i \theta_i^{a_{ix}}.$$

Example: binomial random variable with three trials

Let X be the result of one trial, and we are interested in counting the number of successes in three consecutive trials.
There are two parameters, $\theta = (\theta_1, \theta_2)$. Let
$\theta_1 = P(X = 0), \theta_2 = P(X = 1)$.
Take $h = (1, 3, 3, 1)$ and $\mathcal{A} =$?

Design matrix recipe

Columns of $\mathcal{A}$ are exponents of the parametrization of each given state.

P(no 0)$=\theta_2^3$, P(one 0)$=\theta_1\theta_2^2$, P(two 0s)$=\theta_1^2\theta_2$, P(three 0s)$=\theta_1^3$.

So... For the value $x = 0$, the corresopnding column a_{i0} should be $[0, 3]^t$.

$$p_\theta(x) \propto \frac{1}{Z(\theta)} h_x \prod_i \theta_i^{a_{ix}}.$$

Example: binomial random variable with three trials

Let X be the result of one trial, and we are interested in counting the number of successes in three consecutive trials.
There are two parameters, $\theta = (\theta_1, \theta_2)$. Let
$\theta_1 = P(X = 0), \theta_2 = P(X = 1)$.
Take $h = (1, 3, 3, 1)$ and $\mathcal{A} =$?

Design matrix recipe

Columns of $\mathcal{A}$ are exponents of the parametrization of each given state.

P(no 0)$=\theta_2^3$, P(one 0)$=\theta_1\theta_2^2$, P(two 0s)$=\theta_1^2\theta_2$, P(three 0s)$=\theta_1^3$.

So... For the value $x = 0$, the corresopnding column a_{i0} should be $[0, 3]^t$.

$$\mathcal{A} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{bmatrix}.$$

- The previous slide should be compared to slide 7 in this lecture.
- See also the example 6.2.5 in the book. “Twisted cubic”
- Finally, note that if

$$\mathcal{A} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{bmatrix}$$

and $h = [1, 1, 1, 1]$, then the parametric model equals:

$$p_\theta = \frac{1}{Z(\theta)}(\theta_2^3, \theta_1\theta_2^2, \theta_1^2\theta_2, \theta_1^3),$$

where $Z(\theta) = \theta_2^3 + \theta_1\theta_2^2 + \theta_1^2\theta_2 + \theta_1^3$.

- As I wrote on the board, the vector $(\theta_2^3, \theta_1\theta_2^2, \theta_1^2\theta_2, \theta_1^3)$ can be summarized by the matrix of exponents $\mathcal{A}$, and if you know which row corresponds to which parameter exactly, then you recover the full parametrization.

Coming up next

- log-affine models
- what to do with the h function in the parametrization of an exponential family model (nothing!)
- is there an "easy" way to compute the implicitization of all discrete exponential families.

Other reading, resources, and a task!

- Eliana Duarte's summer school lectures include these slides on [exponential families: an algebraic statistics perspective], see page 13-18. **link will be provided ASAP.**
- Michael I. Jordan's chapter on exponential families provides another resource equivalent to the background in Chapter 6.
- Martin Wainwright's notes on how to turn a multinomial model into an exponential family form are on page 6 of this document.
  - You should **try this on your own**. Fill out **all** the details of writing down the independence model $\mathcal{M}_{1 \perp\!\!\!\perp 2}$, for example, in exponential family form.

License

Major parts of this presentation are from Kaie Kubjas' course lectures, used with permission.

This document is created for Math/Stat 561, Spring 2023.