STA732
Statistical Inference

Lecture 02: Exponential families

Yuansi Chen

Spring 2023

Duke University

https://www2.stat.duke.edu/courses/Spring23/sta732.01/

Recap from Lecture 01

- Defined statistical inference problem

  Statistical experiement, data, statistical model, loss function, risk function

- Discussed how to argue for the optimal estimator

  statistical optimality, in addition to empirical success, fast computation, simplicity, etc.

Goal of Lecture 02

- Introduce exponential families
- Examples
- Differential identities (how to get moments and cumulants from exponential families?)

Chap. 2 in Keener or Chap. 1.5 in Lehmann and Casella

Exponential families

Exponential families

An s-parameter exponential family is a family $\mathscr{P} = \{P_\eta : \eta \in \Xi\}$ with densities p_η w.r.t. a common measure μ on $\mathcal{X}$ of the form

$$p_\eta(x) = \exp\left(\eta^\top T(x) - A(\eta)\right) h(x)$$

$T : \mathcal{X} \to \mathbb{R}^s$	sufficient statistics
$h : \mathcal{X} \to \mathbb{R}$	carrier/base density
$\eta \in \Xi \subseteq \mathbb{R}^s$	natural parameter
$A : \mathbb{R}^s \to \mathbb{R}$	cumulant-generating function (cgf)

Notes on $A(\eta)$

For any η, the cgf $A(\eta)$ is determined by h and T. Since $\int p_\eta d\mu = 1$ holds, we have

$$A(\eta) = \log\left[\int \exp\left(\eta^\top T(x)\right) h(x) d\mu(x)\right]$$

- We say p_η is normalizable if $A(\eta) < \infty$
- So $A(\eta)$ is also called the normalizing constant.

Example 2.1

Take μ to be Lebesgue measure on $\mathbb{R}$, $s = 1$, $h = \mathbf{1}_{(0,\infty)}$ and $T(x) = x$. Then we have

$$\begin{aligned} A(\eta) &= \log \int_0^\infty e^{\eta x} dx \\ &= \begin{cases} \log(-1/\eta), & \eta < 0 \\ \infty, & \eta \geq 0. \end{cases} \end{aligned}$$

What is the corresponding $p_\eta(x)$? What distribution? Is it in the usual form?

Notes on the natural parameter

The natural parameter space is the set of all normalizable η:

$$\Xi_1 = \{\eta : A(\eta) < \infty\}$$

We say $\mathscr{P}$ is in canonical form if $\Xi = \Xi_1$. Sometimes we could take $\Xi \subset \Xi_1$.

Other parameterization for an exponetial family

Take $\eta : \Omega \to \Xi$, define

$$p_\theta(x) = \exp\left[\eta(\theta)^\top T(x) - B(\theta)\right] h(x)$$
$$B(\theta) = A(\eta(\theta))$$

The family $\{p_\theta : \theta \in \Omega\}$ is also called an exponential family

Other parameterization for an exponetial family

Take $\eta : \Omega \to \Xi$, define

$$p_\theta(x) = \exp\left[\eta(\theta)^\top T(x) - B(\theta)\right] h(x)$$
$$B(\theta) = A(\eta(\theta))$$

The family $\{p_\theta : \theta \in \Omega\}$ is also called an exponential family (Many distribution belong to exponential families (see Wiki) but often some massaging is needed to realize)

Example 2.2: normal with unknown mean and variance

The normal distribution $\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0$ has density

$$
\begin{aligned}
p_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&= \exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right]
\end{aligned}
$$

Example 2.2: normal with unknown mean and variance

The normal distribution $\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0$ has density

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right] \end{aligned}$$

We identify

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \eta(\theta) = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}, T(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

$$h(x) = 1, \quad B(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\left(2\pi\sigma^2\right)$$

How to write in terms of natural parameters?

$$p_\eta(x) = \exp\left[\eta^\top \begin{pmatrix} x \\ x^2 \end{pmatrix} - A(\eta)\right]$$

where $\Xi = \{\eta \in \mathbb{R}^2 \mid \eta_2 < 0\}$ and

$$A(\eta) = \frac{-\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right)$$

$\{p_\eta : \eta \in \Xi\}$ lives inside a s-dimensional subspace

It is useful to think "$\log \{p_\eta : \eta \in \Xi\}$" is a subset of an s-dimensional subspace of the log-density space

- $e^{f_\eta(x)}$ is always proportional to a density if integrable
- For exponential family, we can write $f_\eta(x) = \log h(x) + \eta^\top T(x)$ (draw a picture)

The form of an exponential family is not unique

Operations to express the same family

1. Change the common measure so $h(x) = 1$:

$$\mu \rightsquigarrow \tilde{\mu} \text{ with } \frac{d\tilde{\mu}}{d\mu} = h$$

2. Reparameterize so $0 \in \Xi$: take $\eta_0 \in \Xi$

$$\begin{aligned} \eta &\rightsquigarrow \tilde{\eta} = \eta - \eta_0 \\ h &\rightsquigarrow \tilde{h} = p_{\eta_0}(x) \\ A &\rightsquigarrow \tilde{A}\left(\tilde{\eta}\right) = A\left(\eta_0 + \tilde{\eta}\right) - A(\eta_0) \end{aligned}$$

3. Reparameterize with an invertible map $\mathbb{R}^s \to \mathbb{R}^s$.

...

More examples

Example 2.3: joint density of n i.i.d. normal

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, the joint density is

$$\begin{aligned} p_\theta(x) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right] \\ &= \exp\left\{ \sum_{i=1}^{n} \left[\frac{\mu}{\sigma^2} x_i - \frac{1}{2\sigma^2} x_i^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log\left(2\pi\sigma^2\right) \right] \right\} \end{aligned}$$

Example 2.3: joint density of n i.i.d. normal

Given $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, the joint density is

$$\begin{aligned} p_\theta(x) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right] \\ &= \exp\left\{ \sum_{i=1}^{n} \left[\frac{\mu}{\sigma^2} x_i - \frac{1}{2\sigma^2} x_i^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right) \right] \right\} \end{aligned}$$

$$\eta(\theta) = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}, T(x) = \begin{pmatrix} \sum x_i \\ \sum x_i^2 \end{pmatrix}, B(\theta) = nB^{(1)}(\theta)$$

Ex: in general the joint density of n i.i.d. random variables from s-parameter Exp family is still an s-parameter Exp family with the same parameters

Example: binomial

For $X \sim$ Binomial(n, θ), X has probability mass function

$$\begin{aligned} p_\theta(x) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &= \left(\frac{\theta}{1-\theta}\right)^x (1-\theta)^n \binom{n}{x} \\ &= \exp\left[\log\left(\frac{\theta}{1-\theta}\right) x + n\log(1-\theta)\right] \binom{n}{x} \end{aligned}$$

This is a 1-parameter exponential family

$$T(x) = x, \quad \eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$$

Example: Poisson

For $X \sim$ Poisson(θ), X has probability mass function

$$
\begin{aligned}
p_\lambda(x) &= \frac{\lambda^x e^{-\lambda}}{x!} \\
&= \exp\left[\log(\lambda)x - \lambda\right] \frac{1}{x!}
\end{aligned}
$$

This is a 1-parameter exponential family

$$\eta(\lambda) = \log(\lambda)$$

Ex: try some on Wikipedia: Beta, Gamma, Dirichlet...

Differential Identities

Intuition for getting moments from cgf

Because the density integrates to 1, we always have

$$e^{A(\eta)} = \int e^{\eta^\top T(x)} h(x) d\mu(x)$$

Whenever a quantity is in the form of "integral of exponential tilt", we can obtain moments by differentiating on both sides

Intuition for getting moments from cgf

Because the density integrates to 1, we always have

$$e^{A(\eta)} = \int e^{\eta^\top T(x)} h(x) d\mu(x)$$

Whenever a quantity is in the form of "integral of exponential tilt", we can obtain moments by differentiating on both sides

Be careful: we need to be able to switch the order of derivative and integral!

Theorem 2.4 in Keener

Theorem 2.4

Let Ξ_f be the set of values for $\eta \in \mathbb{R}^s$ where

$$\int |f(x)| \exp\left[\eta^\top T(x)\right] h(x) d\mu(x) < \infty$$

Then the function

$$g(\eta) = \int f(x) \exp\left[\eta^\top T(x)\right] h(x) d\mu(x)$$

is continuous and has continuous partial derivatives of all orders for $\eta \in \Xi_f^o$.

In particular, taking $f = 1$, $A(\eta)$ has all partial derivatives

Proof sketch in 1-d (Chap. 2.3. in Keener)

We want to take derivative of $e^{A(\eta)} = \int \exp\left[\eta T(x)\right] h(x) d\mu(x)$ inside integral

- Sufficient to consider $\eta \in (-3\epsilon, 3\epsilon)$ and show the derivative at $\eta = 0$
- **Idea:** use dominated convergence theorem
- Construct a sequence that converges to the actual derivative

Proof:

What do we get by differentiating $A(\eta)$?

By differentiating once, show that

$$\nabla A(\eta) = \mathbb{E}_\eta[T(X)]$$

Because

$$\frac{\partial}{\partial \eta_j} e^{A(\eta)} = \frac{\partial}{\partial \eta_j} \int \exp\left[\eta^\top T(x)\right] h(x) d\mu(x)$$

Differentiating twice

By differentiating twice, show that

$$\nabla^2 A(\eta) = \mathrm{Var}_\eta[T(X)]$$

Example: Poisson

$$p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$T(x) = x, \eta(\lambda) = \log(\lambda), B(\lambda) = \lambda$$

Example: Poisson

$$p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$T(x) = x, \eta(\lambda) = \log(\lambda), B(\lambda) = \lambda$$

For the natural parameter η, $A(\eta) = e^\eta$, then

$$\mathbb{E}_\eta[X] = \frac{de^\eta}{d\eta} = e^\eta = \lambda$$
$$\mathrm{Var}_\eta[X] = \frac{d^2}{d\eta^2} e^\eta = e^\eta = \lambda$$

Moment-generating function

For T a random vector in $\mathbb{R}^s$, the moment generating function of T is

$$M_T(u) = \mathbb{E}\left[e^{u^\top T}\right]$$

The cumulant generating function is

$$K_T(u) = \log(M_T(u))$$

Useful properties of moment-generating function

1. If two random variables have the same moment-generating function, then the have the same distribution
2. Moments of T, denoted by

$$\mathbb{E}[T_1^{r_1} \times \cdots \times T_s^{r_s}]$$

can be found by differentiating M_T at $u = 0$

$$\frac{\partial^{r_1}}{\partial u_1^{r_1}} \cdots \frac{\partial^{r_s}}{\partial u_s^{r_s}} M_t(u) \bigg|_{u=0}$$

Moment-generating function of exponential family

$$
\begin{aligned}
M_\eta^{T(X)}(u) &= \mathbb{E}_\eta\left[e^{u^\top T(X)}\right] \\
&= \int e^{u^\top T} e^{\eta^\top T - A(\eta)} h d\mu \\
&= e^{A(\eta+u)-A(\eta)} \underbrace{\int e^{(\eta+u)^\top T - A(\eta+u)} h d\mu}_{=1} \\
&= e^{A(\eta+u)-A(\eta)}
\end{aligned}
$$

Moment-generating function of exponential family

$$
\begin{aligned}
M_\eta^{T(X)}(u) &= \mathbb{E}_\eta\left[e^{u^\top T(X)}\right] \\
&= \int e^{u^\top T} e^{\eta^\top T - A(\eta)} h d\mu \\
&= e^{A(\eta+u)-A(\eta)} \underbrace{\int e^{(\eta+u)^\top T - A(\eta+u)} h d\mu}_{=1} \\
&= e^{A(\eta+u)-A(\eta)}
\end{aligned}
$$

Hence, the cumulant generating function is

$$
K_T(u) = A(u+\eta) - A(\eta)
$$

Relationship between the moments and cumulants

For $s = 1$, from $M = e^K$, we get

$$M' = K'e^K \Rightarrow \mathbb{E}[T] = \kappa_1$$
$$M'' = (K'' + K'^2)e^K \Rightarrow \mathbb{E}[T^2] = \kappa_2 + \kappa_1^2$$
$$M''' = (K''' + 3K'K'' + K'^3)e^K \Rightarrow \mathbb{E}[T^3] = \kappa_3 + 3\kappa_1\kappa_2 + \kappa_1^3$$

Exampe 2.11: moments of normal

- Unknown μ, but known σ^2
- Unknown μ and σ^2

Proof:

Summary of useful properties of exponential families

$$p_\eta(x) = \exp\left(\eta^\top T(x) - A(\eta)\right) h(x)$$

1. The natural parameter space is convex
2. The joint density of n i.i.d. exponential family densities is still in an exponential family
3. Sufficient statistics $T(x)$
4. $A(\eta)$ infinitely differentiable (Theorem 2.4): easy to get moments

What is next?

- Sufficiency
- Factorization theorem
- Minimal sufficiency

Thank you