STATS 300A: Theory of Statistics I **Autumn 2016/17**

Lecture 2: September 29

Lecturer: Joseph Romano *Scribes: Kenneth Tay*

2.1 Exponential Families

Recall that we have data $X \sim P_\theta, \theta \in \Omega$. The family of distributions $\{P_\theta\}$ is an s-parameter exponential family if the P_θ's have densities of the form

$$\frac{dP_\theta}{d\mu}(x) = p_\theta(x) = \exp\left[\sum_{i=1}^s \eta_i(\theta) T_i(x) - B(\theta)\right] h(x).$$

Example: Suppose $X \sim \text{Poisson}(\lambda)$. This is an exponential family:

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

So $T(x) = x$. Let $\eta = \log \lambda$. Then $\lambda = e^\eta = A(\eta)$.

There is a natural parameterization $\theta \to \eta$, and we can write

$$p(x, \eta) = \exp\left[\sum_{i=1}^s \eta_i T_i(x) - A(\eta)\right] h(x).$$

Generally, we want to express the family in the most economical way. For example, a reduction in the number of terms is possible if the η's satisfy a linear constraint.

Example: $p_\theta(x) \propto \exp[\eta_1 x + \eta_2 x^2]$ but $\eta_1 + \eta_2 = 1$. Then we can rewrite this as $\exp[\eta_1(x + x^2) + x^2]$, hence is better expressed as a 1-parameter exponential family, rather than 2.

Suppose instead that $\eta_2 = \eta_1^2$. Then we can't make this sort of reduction.

We will always assume that the representation is minimal, i.e. neither the η's nor the T's satisfy a linear constraint.

Definition 2.1 *We say that an s-dimensional exponential family has* ***full rank*** *if the natural parameter space contains an s-dimensional rectangle.*

2.1.1 Properties of exponential families

Here are some basic properties of the exponential family:

1. The natural parameter space is convex. (Proof relies of Hölder's inequality.)

2. For any integrable function f and any η in the natural parameter space,

$$\int f(x) \exp\left[\sum \eta_i T_i(x) - A(\eta)\right] h(x)\mu(dx)$$

is infinitely differentiable w.r.t. η_i's, and the derivatives can be obtained by differentiating inside the integral.

3. If X comes from an exponential family with

$$p_\theta(x) \propto \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x),$$

then $T = (T_1, \ldots T_s)$ (as a function of X) is distributed according to an exponential family as well, with density of the form

$$\propto \exp\left[\sum \eta_i t_i - A(\eta)\right] k(t).$$

4. Suppose X_i's are iid and follow an exponential family model. Then $(X_1, \ldots X_n)$ has joint density

$$\prod_{j=1}^{n} \exp\left[\sum_{i=1}^{s} \eta_i T_i(X_j) - A(\eta)\right] h(x_j) = \exp\left[\sum_{i=1}^{s} \eta_i \sum_{j=1}^{n} T_i(X_j) - nA(\eta)\right] \prod_{j=1}^{n} h(x_j).$$

This is still an exponential family of the same "form".

Simple application of property 2: Let $f = 1$. Since

$$\int \exp\left[\sum \eta_i T_i(x) - A(\eta)\right] h(x)\mu(dx) = 1,$$

the derivative w.r.t. η_j must be equal to 0.

Let's try to obtain the derivative by differentiating inside the integral:

$$\int \exp\left[\sum \eta_i T_i(x) - A(\eta)\right] \left[T_j(x) - \frac{\partial}{\partial \eta_j} A(\eta)\right] h(x)\mu(dx) = \mathbb{E}_\eta T_j(x) - \frac{\partial}{\partial \eta_j} A(\eta),$$

so

$$\mathbb{E}_\eta T_j(x) = \frac{\partial}{\partial \eta_j} A(\eta).$$

In general, $\mathrm{Cov}(T_i, T_j) = \dfrac{\partial}{\partial \eta_i} \dfrac{\partial}{\partial \eta_j} A(\eta)$.

2.2 Sufficiency

Definition 2.2 *A statistic $T = T(X)$ is **sufficient** for X (or the family of distributions of X, or θ) if the conditional distribution of $X|T = t$ does not depend on θ, for all t.*

2.2.1 Examples of sufficiency

- $X = (X_1, \dots X_n)$ with X_i's iid, $X_i \sim \text{Bernoulli}(\theta)$. $T = \sum_{i=1}^{n} X_i$ is sufficient because the conditional distribution
$$(X_1, \dots X_n) \Big| \sum X_i = t$$
is uniform over the vectors each having t '1's and $n-t$ '0's, regardless of what θ is.
- $X_1, \dots X_n$ iid, uniformly distributed in $[0, \theta]$. Let $T = \max(X_1, \dots X_n)$. Then T is sufficient: the conditional distribution is with probability $\frac{1}{n}$, $X_i = t$, and the remaining $X_j \sim U(0, t)$.
- $X_1, \dots X_n$ iid, real-valued, from some continuous distribution. The order statistics $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ is sufficient. (Note: The k^{th} order statistic is the value of the k^{th} smallest value of the X_i's.)

2.2.2 Why is sufficiency important?

Say I want to estimate $g(\theta)$, and suppose $\delta(X)$ is an estimator of $g(\theta)$. I claim that based on the outcome of the sufficient statistic $T = t$, I can construct an estimator δ' such that δ' has the same distribution as δ for all θ, which means that their risk functions are the same.

Proof: The distribution of $\delta(X)$ can be thought of as being constructed in a two-stage way as follows:

1. Observe outcome $T = t$, then
2. Pick X' according to the distribution of $X | T = t$.

Then X' has the same distribution as X, and so $\delta(X)$ has the same distribution as $\delta(X')$.

$\blacksquare$

2.2.3 Fisher-Neyman Factorization Theorem

The main tool to determine sufficiency is the Fisher-Neyman Factorization Theorem:

Theorem 2.3 (Factorization Theorem) *Assume that P_θ is such that $\frac{dP_\theta}{d\mu} = p_\theta$. A necessary and sufficient condition for $T = T(X)$ to be sufficient is that there exist non-negative functions g_θ and h such that*
$$p_\theta(x) = g_\theta(T(x)) \cdot h(x).$$
with probability 1.

Intuitively, the idea is that the part which depends on θ only depends on X through $T(X)$.

Proof: We will only prove sufficiency for the discrete case. (For the continuous case, essentially the sums just get replaced by integrals.)

Let $p_\theta(x) = P_\theta(X = x)$. Suppose that we can write $p_\theta(x) = g_\theta(T(x)) \cdot h(x)$. We want to show that T is sufficient.

Let $T(x) = t$. Then

$$\begin{aligned} P_\theta(T = t) &= \sum_{x:T(x)=t} P(X = x), \\ P_\theta(X = x | T = t) &= \frac{P_\theta(X = x, T = t)}{P_\theta(T = t)} \\ &= \frac{p_\theta(x)}{\displaystyle\sum_{x:T(x')=t} p_\theta(x')} \\ &= \frac{g_\theta(t)h(x)}{\displaystyle\sum_{x':T(x')=t} g_\theta(t)h(x')} \\ &= \frac{h(x)}{\displaystyle\sum_{x':T(x')=t} h(x')} \end{aligned}$$

which does not depend on θ. ■

Example: $X_1, \ldots X_n$ iid, and

$$p_\eta(x) \propto \exp[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)]h(x).$$

By the Factorization Theorem, $(T_1(X), \ldots T_s(X))$ is sufficient based on one observation X. With n observations, $\left(\sum_j T_1(X_j), \ldots, \sum_j T_s(X_j)\right)$ is sufficient.