STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Bowen Jia and Zheng Zhang

Lecture # 4

1 Sufficiencies

Recap: Neyman-Fisher Factorization criterion. T(X) is sufficient is sufficient iff $p(x;\theta) = g_\theta(T(x))h(x)$ prove for the discrete cases,$p(x|T)$ is independent of θ We will look at the proof for the continuous case (Ref. Keener 6.4).

To begin, suppose $p_\theta \in p$ and $\theta \in \Omega$

$$p(x;\theta) = g_\theta(T(x))h(x).$$

With respect to μ. Modifying h, we can assume without loss of generality that μ us a probability measure equivalent to the family $P = \{p_\theta : \theta \in \Omega\}$ [Equivalence referes to the situation where $\mu(N) = 0$ iff $p_\theta(N) = 0 \quad \forall\theta \in \Omega$].

Let E^* and P^* be the expectation and probability where $X \sim \mu$. Let G^* and G_θ denote marginal distribution for $T(x)$ where $X \sim \mu$ and $X \sim P_\theta$ respectively. Let Q be the conditional distribution for X given T where $X \sim \mu$.

To find the densities for T,

$$\begin{aligned} E_\theta f(T) &= \int f(T(x))g_\theta(T(x))h(x)d\mu(x) \\ &= E^*\{f(T)g_\theta(T)h(X)\} \\ &= \int\int f(t)g_\theta(T)h(x)dQ_t(x)dG^*(t) \\ &\triangleq \int f(t)g_\theta(t)\omega(t)dG^*(t), \end{aligned}$$

where $\omega(t) = \int h(x)dQ_t(x)$. If f is an indicator function this shows that G_θ has the density $g_\theta\omega(t)$ with respect to G^*. Next we define $\widetilde{Q}$ to have density $h/\omega(t)$ with respect to $Q(t)$, so that

$$\widetilde{Q}_t(B) = \int_B \frac{h(x)}{\omega(t)}dQ_t(x),$$

the conditional distribution of X given T under P_θ is independent of Q.

$$\begin{aligned} E_\theta \int (X,T) &= E^*\{f(X,T)g_\theta(T)h(x)\} \\ &= \iint f(x,t)g_\theta(t)h(x)dQ_t(x)dG^*(t) \\ &= \iint f(x,t)d\widetilde{Q}_t(x)dG_\theta(t) \end{aligned}$$

By the definition of conditional distribution, it shows that $\widetilde{Q}$ is a conditional distribution of X given under P_θ. Because $\widetilde{Q}$ does not depend on Q, it is sufficient statistic.
2nd part: T is sufficient statistic $\rightarrow$ factorization holds(tutorial)

2 Sufficiency

Data reduction $\rightarrow$ all information about θ is stored in Θ $\rightarrow$ improves data interpretability. (c.f. example

$$\begin{cases} \widetilde{X} = TU, \\ \widetilde{Y} = T(1-U), \end{cases} \tag{1}$$

where U is a uniform (0,1) independent of T.
Question: how much data compression/reduction can be achieved while the inference for θ is not impaired (in any sense)? what is the optimal data reduction strategy?

3 Exponential families

3.1 Basics

Definition: The model $\{P_\theta : \theta \in \Omega\}$ forms an s-dimensional exponential family if each P_θ has density of the form:

$$P(x_j, \theta) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right) h(x)$$

where $\eta_i(\theta) \in \mathbb{R}$ are called the natural parameters, $T_i(X) \in \mathbb{R}$ are its sufficient statistics, $B(\theta)$ is the log-partition function, which means that it is the logarithm of a normalising factor:

$$B(\theta) = \log\left(\int \exp\left\{\sum_{i=1}^{s} \eta_i(\theta) T_i(x)\right\} h(x) d_\mu(x)\right) \in \mathbb{R},$$

and $h(x) \in \mathbb{R}$ is the base measure (e.g. $I(x \in \mathbb{R})$ or $I(x) \geq 0$).

Remark: Many common distributions are exponential families. Examples include Normal, Binomial, Poisson distribution to name but a few. Exponential families are also closely related to the motions of sufficiency and optimal data reduction.

Example 1. *Exponential distribution $P = \{\exp(\theta) : \theta > 0\}$ the densities take the form:*

$$p(x; \theta) = \theta e^{-\theta x} = \exp(-\theta x + \log \theta) I_{(x \geq 0)},$$

which means that the family is a one-dimensional exp family with $\eta_i(\theta) = -\theta$, $T_i(x) = x$, $B(\theta) = -log(\theta)$ and $h(x) = I_{(x \geq 0)}$. It is noteworthy that the parameterization is not unique.

Example 2. *Beta distribution $P = \{Beta(\alpha, \beta) : \alpha, \beta > 0\}$, $\theta = (\alpha, \beta)$ the densities take the form*

$$\begin{aligned} p(x; \theta) &= x^{\alpha-1}(1-x)^{\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} I_{(0<x<1)} \\ &= \exp\left\{(\alpha-1)\log x + (\beta-1)\log(x-1) + \log\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right\} I(0<x<1) \end{aligned}$$

which means that the beta distribution belongs to a 2-dimensional exponential family with $\eta_i(\theta) = \alpha - 1$, $\eta_2(\theta) = \beta - 1$, $T = (T_1, T_2) = (\log x, \log(1-x))$, $B(\theta) = -\log(\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta)))$ *and* $h(x) = I(0 < x < 1)$. *One may also rewrite* $p(x;\theta)$ *as:*

$$p(x;\theta) = \exp\left\{\alpha \log x + p\log(1-x) + \log(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right\} \frac{I(0<x<1)}{x(1-x)}$$

which change the natural parameter from $\eta_1(\theta)$ *to* $\eta_1^*(\theta) = \alpha$ *and* $\eta_2(\theta)$ *to* $\theta_2^* = \beta$ *with* $h^*(x)$ *becomes* $I(0 < x < 1)/\{x(1-x)\}$.

Definition 1. *An exponential family is in canonical form when the density has the form*

$$p(x;\eta) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x). \tag{2}$$

This parameterises the densities in terms of the natural parameters η *instead of* θ.

Definition 2. *The set of all valid natural parameters* Θ *is called the natural parameter space: for each* $\eta \in \Theta$, *there exists a normalising constant* $A(\eta)$ *such that* $\int p(x;\eta)\,dx$ *= 1, Equivalently,*

$$\Theta = \left\{\eta : 0 < \int \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x) d\mu x < \infty\right\} \tag{3}$$

For any canonical exponential family $P = p_\eta : \eta \in H$, *we have* $H \in \Theta$. *One can show that* Θ *is convex. The differences between canonical and non-canonical one is that for the non-canonical one, there is other parametrisations.*

3.2 Dimension reduction

There are two cases when the superficial dimension of an s-dimensional exponential family $P = p_\eta : \eta \in H$ can be reduced.

3.2.1 Case 1

The $T_i(x)$'s satisfy an affine equality constraint for all $x \in X$. In other words, $\{T_i\}$ are linearly dependent and ~~we call η unidentifiable.~~

Definition 3. *If* $\mathcal{P} = \{p_\theta; \theta \in \Omega\}$, *then* θ *is unidentifiable if for two parameters* $\theta_1 \neq \theta_2$, $p_{\theta_1} = p_{\theta_2}$.

Example 3. *Let* $X \sim \exp(\eta_1, \eta_2)$ *with*

$$p(x;\eta_1,\eta_2) = \exp\{-\eta_1 x - \eta_2 x + \log(\eta_1+\eta_2)\} I(x \geq 0) \tag{4}$$

Here $T_1(x) = T_2(x) = x$ *(they are linearly dependent).We can actually combine* (η_1, η_2) *into* $\eta_1 + \eta_2$ *and write*

$$p(x;\eta_1,\eta_2) = \exp\{-(\eta_1+\eta_2)x + \log(\eta_1+\eta_2)\} I(x \geq 0) \tag{5}$$

Besides, η *is unidentifiable since* $p(x;\eta_1 + c, \eta_2 - c) = p(x;\eta_1,\eta_2)$ *for all* $c < \eta_2$.

3.2.2 Case 2

The η_i's satisfy an affine equality constraint for all $\eta \in H$.

Example 4. *Let* $p(x;\eta) = c(\eta_1,\eta_2)\exp(\eta_1 x + \eta_2 x^2)$ *for all* (η_1,η_2) *satisfying* $\eta_1 + \eta_2 = 1$. *Then we can rewrite*

$$p(x;\eta) = c(\eta_1,\eta_2)\exp(\eta_1(x - x^2) + x^2) \tag{6}$$

3.2.3 Minimal

When neither of the above two cases hold, we call the exponential family minimal.

Definition 4. *A canonical exponential family $P = p_\eta : \eta \in H$ is minimal if*
(1) $\sum_{i=1}^{s} \lambda_i T_i(x) = \lambda_0, \forall x \in X \implies \lambda_i = 0 \; \forall i \in \{0, ..., s\}$
(2) $\sum_{i=1}^{s} \lambda_i \eta_i = \lambda_0, \forall \eta \in H \implies \lambda_i = 0 \; \forall i \in \{0, ..., s\}$

Definition 5. *Suppose is $P = p_\eta : \eta \in H$ a s-dimensional exponential family. If H contains an open s-dimensional rectangle, then P is called full-rank, otherwise P is called curved, which means that the η_i's are related non-linearly.*

Example 5. *Consider $N(\mu, \sigma^2)$ where in this case $\eta_1 = 1/(2\sigma^2)$, $\eta_2 = \mu/\sigma^2$,$T_1(x) = -x^2$,$T_2(x) = x$.*

1. *Take $\mu = \sigma^2$, then $\eta_1 = 1/(2\sigma^2)$, $\eta_2 = 1$, then $1/(2\sigma^2)\eta_2 - \eta_1 = 0$. Therefore, the family is non-minimal in this case.*

2. *Take $\mu = \sqrt{\sigma^2}$, then $\eta_1 = 1/(2\sigma^2)$, $\eta_2 = 1/\sqrt{\sigma^2}$, then $\eta_2 = \sqrt{2\eta_1}$. Therefore, the family is minimal and curved in this case.*

3. *When there's no constraint on (μ, σ^2), H contains an open rectangle: $\mathbb{R} \times (0, \infty)$. Therefore, the family is minimal and full-rank in this case.*

3.3 Properties of exponential families

1. If $X_1, X_2, ..., X_n \overset{i.i.d}{\sim} p(x; \theta) = \exp\{\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\} h(x)$.
   Then by NFFC, $(\sum_{j=1}^{n} T_1(x), ..., \sum_{j=1}^{n} T_s(x))$ is a sufficient statistic. Hence the exponential family is exceptionally compressible.

2. If f is integrable and $\eta \in \Theta$, then

$$G(f, \eta) = \int f(x) \exp\left\{\sum_{i=1}^{s} \eta_i T_i(x)\right\} h(x) d\mu(x) \tag{7}$$

   is infinitely differentiable with respect to η and the derivatives can be obtained by differentiating under the integral sign.

3. The moments of T_i's can be directly calculated by taking $f(x) = 1$:

$$G(f, \eta) = \int \exp\left\{\sum_{i=1}^{s} \eta_i T_i(x)\right\} h(x) d\mu(x) = \exp(A(\eta)) \tag{8}$$

$$\frac{\partial G(f, \eta)}{\partial \eta_i} = \int T_i(x) \exp\left\{\sum_{i=1}^{s} \eta_i T_i(x)\right\} h(x) d\mu(x) = \frac{\partial A(\eta)}{\partial \eta_i} \exp(A(\eta)). \tag{9}$$

   Therefore,

$$\frac{\partial A(\eta)}{\partial \eta_i} = \int T_i(x) \exp\{\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\} h(x) d\mu(x) = E_\eta\{T_i(x)\} \tag{10}$$

   Besides, it can be shown that

$$\frac{\partial^2 A(\eta)}{\partial \eta_i \partial \eta_j} = \mathrm{Cov}_\eta(T_i(x), T_j(x)) \tag{11}$$

3.4 Minimal Sufficiency

Definition 6. *A sufficient statistic T is minimal if for every sufficient statistics T' and for every $x, y \in X$, $T(x) = T(y)$ when $T'(x) = T'(y)$. In other words, T is a function of T'. i.e. there exists a function f such that $T(x) = f(T'(x))$ for any $x \in X$.*

The following theorem allows us to verify whether a sufficient statistic is minimal or not.

Theorem 7. *Let $p(x;\theta) : \theta \in \Omega$ be a family of densities with respect to some measure μ(usually lebesgue measure for continuous distribution and counting measure for discrete distribution).Suppose that there exists a statistic T such that for every $x, y \in X$*

$$p(x;\theta) = c(x,y)p(y;\theta) \longleftrightarrow T(x) = T(y) \tag{12}$$

for every θ and some $c(x,y) \in \mathbb{R}$. Then T is a minimal sufficient statistic.

Proof. *First prove that T is sufficient and then T is minimal.*

1. *(T is sufficient) For all $t \in T(X)$(the image of T), consider the preimage $A_t = T^{-1}(t)$. For each A_t, we denote x_t as a representative. Then for any $y \in X$, we have $y \in A_{T(y)}$ and $x_{T(y)} \in A_{T(y)}$. From the assumption of T, we have*

$$p(y;\theta) = c(y, x_{T(y)})p(x_{T(y)};\theta) = h(y)g_\theta(T(y)) \tag{13}$$

Therefore, by NFFC, T is sufficient.

2. *(T is minimal) Consider another sufficient statistic T'. By NFFC,*

$$p(x;\theta) = \tilde{g}_\theta(T'(x))\tilde{h}(x) \tag{14}$$

Take any x and y such that $T'(x) = T'(y)$, then

$$p(x;\theta) = \tilde{g}_\theta(T'(x))\tilde{h}(x) = \tilde{g}_\theta(T'(y))\tilde{h}(y)\frac{\tilde{h}(x)}{\tilde{h}(y)} = p(y;\theta)C(x,y) \tag{15}$$

By the assumption of T, $T(x) = T(y)$. Therefore, we've proved that for any sufficient statistics T' and any x and y, $T'(x) = T'(y)$ implies $T(x) = T(y)$. T is minimal.