STAT5010 Advanced Statistical Inference **Fall 2023**

Lecture 2: Further Discussion on Exponential Family, Minimal Sufficiency, Ancillarity and Completeness

Lecturer: Tony Sit *Scribe: Jianting FENG and Yujia LIU*

2.1 Exponential Families

Definition 2.1 *The model $\{P_\theta : \theta \in \Theta\}$ forms an s-dimensional exponential family if each p_θ has the density of the form:*

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

where

- $\eta_i(\theta) \in \mathbb{R}$ *are called the natural parameters,*
- $T_i(x) \in \mathbb{R}$ *are its sufficient statistics (by NFFC),*
- $B(\theta)$ *is the log-partition function (normalization factor)*

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

- $h(x) \in \mathbb{R}$ *is the base measure.*

Example 2.2 *Beta distribution $P = \{Beta(\alpha, \beta) : \alpha, \beta > 0\}$, $\theta = (\alpha, \beta)$. The densities take the form*

$$\begin{aligned} p(x,\theta) &= x^{\alpha-1}(1-x)^{\beta-1} I\left(x \in (0,1)\right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\ &= \exp\left((\alpha-1)\log x + (\beta-1)\log(1-x) + \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)\right) I\left(x \in (0,1)\right) \\ &= \exp\left(\underbrace{\alpha \log x}_{\eta_1(\theta)=\alpha} + \underbrace{\beta \log(1-x)}_{\eta_2(\theta)=\beta} + \underbrace{\log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)}_{B(\theta)}\right) \underbrace{\frac{I(x \in (0,1))}{x(1-x)}}_{h(x)} \end{aligned}$$

Definition 2.3 *An exponential family is in canonical form when the density has the form*

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

Definition 2.4 *The set of all valid natural parameters Θ_{nat} is called the natural parameter space: for each $\eta \in \Theta_{nat}$, there exists a normalizing constant $A(\eta)$ such that $\int p(x;\eta) = 1$, or equivalently,*

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

2.1.1 Reducing the Dimension

"Superficial dimension" $\Leftarrow$ be a bit more careful.

Case I: The $T_i(x)$'s satisfy an affine equality constraint $\forall x \in \mathcal{X}$.

Example 2.5 *$X \sim Exp(\eta_1, \eta_2)$, $p(x;\eta_1,\eta_2) = \exp(-\eta_1 x - \eta_2 x + \log(\eta_1+\eta_2)) \times I(x \geq 0)$. Here $T_1(x) = T_2(x) = x$ (i.e. they are linearly dependent).*

We can rewrite it as

$$p(x;\eta_1,\eta_2) = \exp\left(-(\eta_1+\eta_2)x + \log(\eta_1+\eta_2)\right) I(x \geq 0)$$

Definition 2.6 *If $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, then θ is unidentifiable if for two parameters $\theta_1 \neq \theta_2$, $P_{\theta_1} = P_{\theta_2}$.*

E.g. In the above example,

$$p(x;\eta_1+\alpha,\eta_2-\alpha) = p(x;\eta_1,\eta_2)$$

for any $\alpha < \eta_2$.

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

Example 2.7

$$\begin{aligned} p(x;\eta) &\propto \exp(\eta_1 x + \eta_2 x^2) \textit{ for all } (\eta_1,\eta_2) \textit{ satisfying } \eta_1+\eta_2 = 1 \\ &= \exp(\eta_1(x-x^2)+x^2) \end{aligned}$$

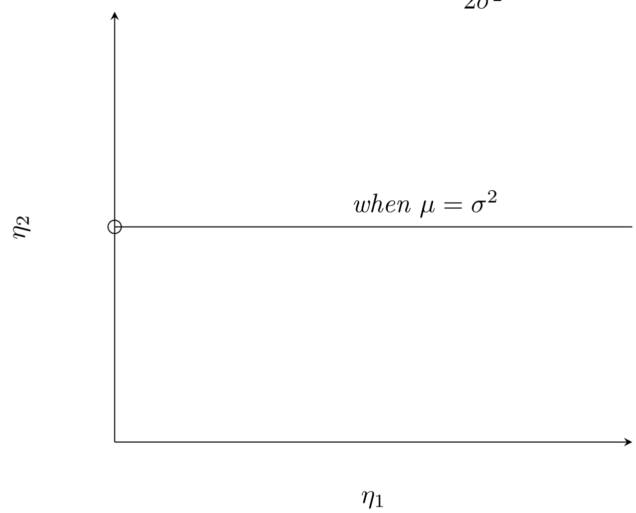
Definition 2.8 *A canonical exponential family $\mathcal{P} = \{P_\eta : \eta \in H\}$ is minimal if*

- $\sum_{i=1}^{s} \lambda_i T_i(x) = \lambda_0, \forall x \in \mathcal{X} \Rightarrow \lambda_i = 0, \forall i \in \{0,\cdots,s\}$ *(no affine T_i equality constrains)*
- $\sum_{i=1}^{s} \lambda_i \eta_i = \lambda_0, \forall \eta \in H \Rightarrow \lambda_i = 0, \forall i \in \{0,\cdots,s\}$ *(no affine η_i equality constrains).*

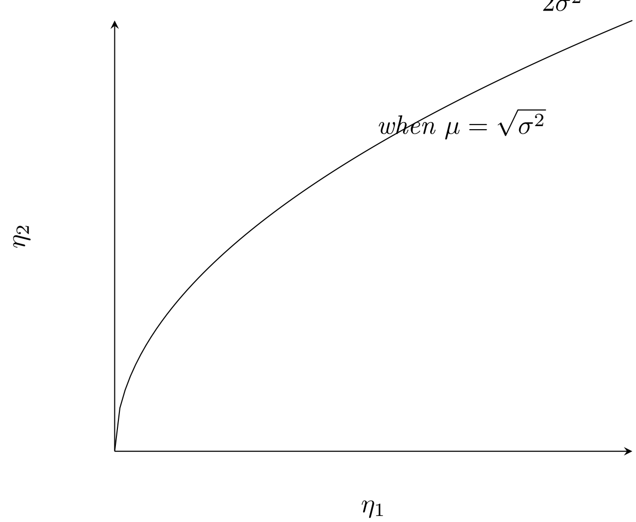
Definition 2.9 *Suppose $\mathcal{P} = \{P_\eta : \eta \in H\}$ is an s-dimensional minimal exponential family. If H contains an open s-dimensional rectangle, then $\mathcal{P}$ is called **full-rank** . Otherwise, $\mathcal{P}$ is **curved**. In curved exponential families, the η_i's are related in a non-linear way.*

Example 2.10 *Normal distribution $\mathcal{N}(\mu,\sigma^2)$, $\eta_1 = \frac{1}{2\sigma^2}, T_1(x) = -x^2, \eta_2 = \frac{\mu}{\sigma^2}, T_2(x) = x$*

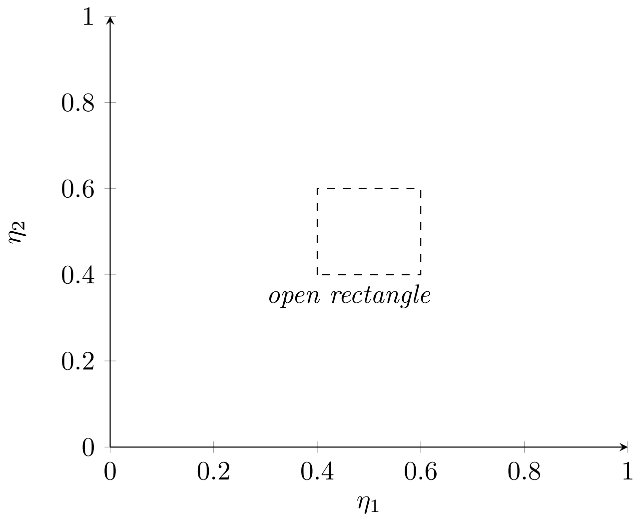
- *Non-minimal case: when* $\mu = \sigma^2$, $\eta_1 = \dfrac{1}{2\sigma^2}$, $\eta_2 = 1$



- *Minimal and curved: when* $\mu = \sqrt{\sigma^2}$, *so* $\eta_1 = \dfrac{1}{2\sigma^2}, \eta_2 = \dfrac{1}{\sqrt{\sigma^2}}, \eta_2^2 = 2\eta_1$



- *Minimal and full-rank (most common): when the natural parameter space is* $(0, +\infty) \times \mathbb{R}$



2.1.2 Properties of Exponential Families

Property 1: If $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p(x;\theta) = \exp\left(\sum_{i=1}^s \eta_i(\theta) T_i(x) - B(\theta)\right) h(x)$, then

$$p(x_1, \cdots, x_n; \theta) = \exp\left(\sum_{i=1}^s \eta_i(\theta) \sum_{j=1}^n T_i(x_j) - nB(\theta)\right) \prod_{j=1}^n h(x_j)$$

By NFFC, $\left(\sum_{j=1}^n T_1(x_j), \cdots, \sum_{j=1}^n T_s(x_j)\right)$ is a sufficient statistic.
Property 2: If f is interable and $\eta \in \Theta_{\text{nat}}$, then $G(f,\eta) = \int f(x) \exp\left(\sum_{i=1}^s \eta_i T_i(x)\right) h(x) \mathrm{d}\mu(x)$ is infinitely differentiable with respect to η and the derivatives can be obtained by differentiate under the integral sign. (See Thm 2.7.1 of TSH).

Example 2.11 (Moments of T_i's) *Take $f(x) = 1$, then*

$$\begin{aligned}
G(f,\eta) &= \int \exp\left(\sum_{j=1}^s \eta_j T_j(x)\right) h(x) \mathrm{d}\mu(x) = \exp\left(A(\eta)\right) \\
\frac{\partial G(f,\eta)}{\partial \eta_i} &= \int T_i(x) \exp\left(\sum_{j=1}^s \eta_j T_j(x)\right) h(x) \mathrm{d}\mu(x) = \frac{\partial A(\eta)}{\partial \eta_i} \times \exp\left(A(\eta)\right) \\
\frac{\partial A(\eta)}{\partial \eta_i} &= \int T_i(x) \exp\left(\sum_{j=1}^s \eta_j T_j(x) - A(\eta)\right) h(x) \mathrm{d}\mu(x) \\
&= \mathbb{E}_\eta[T_i(X)] \\
\frac{\partial^2 A(\eta)}{\partial \eta_i \partial \eta_j} &= Cov_\eta\left(T_i(X), T_j(X)\right)
\end{aligned}$$

Theorem 2.12 (Pitman-Koopman-Darmois) *Amongst families of exponential distributions, whose* ***domain does not depend on/vary with the parameters being estimated****,only in exponential families is there a sufficient statistic whose dimension remains bounded as the sample size increases.*

($\Rightarrow$ Non-exponential families of distribution require non-parameteric statistic to fully capture the information in data)

2.2 Minimal Sufficiency

A notion of max achievable lossless data reduction.

Definition 2.13 (Minimal Sufficiency) *A sufficient statistic T is minimal if for every sufficient statistic T' and for any $x, y \in \mathcal{X}$, $T(x) = T(y)$ whenever $T'(x) = T'(y)$. In other words, T is a function of T'. (There exists f such that $T(x) = f(T'(x))$ for any $x \in \mathcal{X}$).*

Theorem 2.14 *Let $\{p(x;\theta), \theta \in \Theta\}$ be a family of densities w.r.t. same measure μ. Suppose that there exists a statistic T such that for every $x, y \in \mathcal{X}$:*

$$\frac{p(x;\theta)}{p(y;\theta)} = c_{x,y} \Leftrightarrow p(x;\theta) = c_{x,y}p(y;\theta) \Leftrightarrow T(x) = T(y)$$

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

Proof: We first prove that T is sufficient. Start with

$$\begin{aligned} T(\mathcal{X}) &= \{t : t = T(x) \text{ for some } x \in \mathcal{X}\} \\ &= \text{range of } T. \end{aligned}$$

For each $t \in T(\mathcal{X})$, we consider the preimage $A_t = \{x : T(x) = t\}$ and select an arbitrary representative x_t from each A_t. Then, for any $y \in \mathcal{X}$, we have $y \in A_{T(y)}$ and $X_{T(y)} \in A_{T(y)}$. By the definition of A_t, this implies that $T(y) = T(X_{T(y)})$. From the assumption of the theorem,

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

which yields sufficiency of T by the NFFC.

Consider another sufficient statistic T'. By NFFC,

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

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

$$\begin{aligned} p(x;\theta) &= \tilde{g}_\theta(T'(x))\tilde{h}(x) \\ &= \tilde{g}_\theta(T'(y))\tilde{h}(y) \cdot \frac{\tilde{h}(x)}{\tilde{h}(y)} \\ &= p(y;\theta)C_{x,y}. \end{aligned}$$

Hence, $T(x) = T(y)$ by the assumption of the theorem. So $T'(x) = T'(y)$ implies $T(x) = T(y)$ for any sufficient statistic T' and any x and y. As a result, T is a minimal sufficient statistic. ■

Remark 2.15 (Ex 3.12, Keener) *For any minimal s-dimensional exponential family the statistic $(\sum_i T_1(X_i), \cdots, \sum_i T_s(X_i))$ is a minimal sufficient statistic.*

Example 2.16 (Curved exponential family) *Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\sigma, \sigma^2)$, $\theta = \sigma > 0$.*

$$\begin{aligned}
\frac{p(x;\theta)}{p(y;\theta)} &= \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\sigma}{\sigma^2}\sum_{i=1}^n x_i - \frac{n\sigma^2}{2\sigma^2}\right)}{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 + \frac{\sigma}{\sigma^2}\sum_{i=1}^n y_i - \frac{n\sigma^2}{2\sigma^2}\right)} \\
&= \exp\left(-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2\right\} + \frac{1}{\sigma}\left\{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i\right\}\right)
\end{aligned}$$

Is $T(X) = (T_1(X), T_2(X)) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ minimal sufficient?
First, if $T(x) = T(y)$ for some $x, y \in \mathcal{X}$, then the ratio above is equal to 1, hence does not dependent on θ. Therefore, T is sufficient.
Second, if for some x, y, the ratio is independent of θ, notice that the ratio $\to 1$ as $\sigma \to \infty$. Therefore, $C_{x,y} = 1$ and $\log C_{x,y} = 0 = \log\left(\frac{p(x;\theta)}{p(y;\theta)}\right)$. This implies

$$\begin{aligned}
&\frac{1}{2\sigma^2}\left(T_1(y) - T_1(x)\right) + \frac{1}{\sigma}\left(T_2(x) - T_2(y)\right) = 0, \quad \forall\sigma \\
\Leftrightarrow \quad & T_1(y) - T_1(x) = 2\sigma\left(T_2(y) - T_1(y)\right) \quad \forall\sigma.
\end{aligned}$$

As $\sigma \to 0$, RHS $\to 0$. So, $T_2(y) = T_2(x)$. Consequently, T is a minimal sufficient statistic.

Example 2.17 *Let* $X_1, \cdots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$ *and* $T(x) = \max(X_1, \cdots, X_n)$. *In that case for* $x = (x_1, \cdots, x_n)$ *such that* $x_i > 0$, $i = 1, \cdots, n$,

$$p(x;\theta) = \prod_{i=1}^n \frac{1}{\theta} I(x_i < \theta) = \frac{1}{\theta^n} I\left(T(x) < \theta\right)$$

If $T(x) = T(y)$, *then* $p(x;\theta) = \underbrace{1}_{c_{x,y} \perp\!\!\!\perp \theta} \times p(y;\theta) \Rightarrow$ *sufficiency. Conversely, if* $x, y > 0$ *are supported by the same* θ*'s, then* $\{\theta$ *supporting* $x\} = (T(X), \infty) = (T(y), \infty) = \{\theta$ *supporting* $y\}$. *Therefore,* $T(x) = T(y)$ *and* T *is a minimal sufficient statistic.*

2.3 Ancillarity and Completeness

Sufficient statistic/minimal sufficient statistics don't achieve data reduction in a significant way.

Example 2.18 *Consider* $X_1, \cdots, X_n \overset{i.i.d.}{\sim}$ *CauchyLoc*(θ), *whose density is given by*

$$p(x;\theta) = \frac{1}{\pi}\frac{1}{1 + (x-\theta)^2} = f(x - \theta)$$

then $\left(X_{(1)}, \cdots, X_{(n)}\right)$ *is minimal sufficient. (See TPE 1.5).*

Another example is double exponential location model $p(x;\theta) \propto \exp\left(|x - \theta|\right)$.Amount of ancillary information present in its minimal sufficient statistics.

Definition 2.19 *A statistic* A *is ancillary for* $X \sim P_\theta \in \mathcal{P}$ *if the distribution of* $A(X)$ *does not depend on* θ.

Example 2.20 *Consider again $X_1, \cdots, X_n \overset{i.i.d.}{\sim}$ CauchyLoc(θ). Then $A(X) = X_{(n)} - X_{(1)}$ is ancillary even though $(X_1, \cdots, X_n)$ is minimal sufficient. To see this, note that $X_i = Z_i + \theta$ for $Z_i \overset{i.i.d.}{\sim}$ CauchyLoc(θ), we can see that $X_{(i)} = Z_{(i)} + \theta$ and $A(X) = A(Z) \perp\!\!\!\perp \theta$.*

Definition 2.21 *A statistic A is first-order ancillary for $X \sim P_\theta \in \mathcal{P}$ if $\mathbb{E}_\theta[A(X)]$ does not dependent on θ.*

Definition 2.22 *A statistic T is comete for $X \sim P_\theta \in \mathcal{P}$ if no non-constant function of T is first-order ancillary. In other words, if $\mathbb{E}_\theta[f(T(X))] = 0$ for all θ, then $f(T(X))$ with probability 1 for all θ.*

Some properties:

1. If T is colete sufficient, then T is minimal sufficient. (Bahadur's theorem)
2. Complete sufficient static yields optimal unbiased estimators.

Example 2.23 *Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim}$ Bernoulli(θ), $\theta \in (0,1)$. Then $T(X) = \sum_{i=1}^n X_i$ is sufficient.*

Suppose $\mathbb{E}_\theta\left[f(T(X))\right] = 0$ for all $\theta \in (0,1)$,

$$\sum_{j=0}^{n} f(j)\binom{n}{j}\theta^j(1-\theta)^{n-j} = 0, \quad \forall \theta \in (0,1)$$

Dividing both sides by θ^n and reparameterizing $\beta = \dfrac{\theta}{1-\theta}$ we can rewrite it as

$$\sum_{j=d}^{n} f(j)\binom{n}{j}\beta^j = 0, \quad \forall \beta > 0$$

if f is non-zero, then LHS is a polynomial of degree at most n. However, an nth-degree polynomial has at most n roots. Hence, it is impossible for the LHS to be equal to 0 for every $\beta > 0$ unless $f = 0$. Therefore, T is complete.

Example 2.24 *Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$ with unknown θ and an known $\sigma^2 > 0$. Is $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ complete for this model?*

Consider the special case of $n = 1$ and $\sigma = 1$. $T(X) = X \sim \mathcal{N}(\theta, 1)$

$$\mathbb{E}_\theta\left[f(X)\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)\exp\left(-\frac{(x-\theta)^2}{2}\right)\mathrm{d}x = 0, \quad \forall \theta \in \mathbb{R}$$

Multiplying both sides by $\sqrt{2\pi}\mathrm{e}^{\dfrac{\theta^2}{2}}$, we have

$$\int_{-\infty}^{\infty} f(x)\exp\left(-\frac{x^2}{2}\right)\exp\left(\theta x\right)\mathrm{d}x = 0, \quad \forall \theta \in \mathbb{R}.$$

We decompose f into its positive and negative parts as $f(x) = f_+(x) - f_-(x)$, where $f_+(x) = \max(f(x), 0)$, and $f_-(x) = \max(-f(x), 0)$. Note that $f_+ \geq 0$ and $f_- \geq 0$. For all $x \in \mathbb{R}$ $f_+(x) = f_(x)$ if and only if $f_+(x) = f_-(x) = 0$.

Suppose f_+ and f_- have non-zero components, and we may write

$$\frac{\int_{-\infty}^{\infty} f_+(x) \mathrm{e}^{-\frac{x^2}{2}} \mathrm{e}^{\theta x} \mathrm{d}x}{\int_{-\infty}^{\infty} f_+(x) \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x} = \frac{\int_{-\infty}^{\infty} f_-(x) \mathrm{e}^{-\frac{x^2}{2}} \mathrm{e}^{\theta x} \mathrm{d}x}{\int_{-\infty}^{\infty} f_-(x) \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x}$$

Note that

$$\frac{f_+(x) \mathrm{e}^{-\frac{x^2}{2}}}{\int_{-\infty}^{\infty} f_+(x) \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x}$$

defines a probability density.
The equality of the mgfs implies equality of the densities, which in turn implies $f_+(x) = f_-(x)$ a.e..Then $f_+(x) = f_-(x) = 0$ a.e., or in other words, $f(x) = 0$ a.e.. Hence T is complete.