Biostat 602 Winter 2017

Lecture Set 5

Principles of Data Reduction

Exponential Family of Distributions

Exponential Family

Reading: CB 6.2

Definition 3.4.1: The random variable X belongs to an exponential family of distributions, if its pdf/pmf can be written in the form

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right], \quad x \in A$$

where

- $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d),\ d \leq k,$
- $w_j(\theta),\ j \in \{1, \cdots, k\}$ and $c(\boldsymbol{\theta}) \geq 0$ are real valued functions of $\boldsymbol{\theta}$ alone,
- $t_j(x)$ and $h(x) \geq 0$ only involve data,
- Support of X, i.e. the set $A = \{x : f(x|\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.

Example 1: Show that a Poisson(λ) ($\lambda > 0$) belongs to the exponential family.

Proof: The pmf of X can be written as

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

Define $h(x) = 1/x!$, $c(\lambda) = e^{-\lambda}$, $w(\lambda) = \log \lambda$, and $t(x) = x$, then

$$f_X(x|\lambda) = h(x)c(\lambda)\exp\left[w(\lambda)t(x)\right]$$

Example 2: $\mathcal{N}(\mu, \sigma^2)$ belongs to an Exponential Family

Proof: The pdf of X is can be written as:

$$\begin{aligned} f_X(x|\boldsymbol{\theta} = (\mu, \sigma^2)) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{x^2}{2\sigma^2} + \frac{2\mu x}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right] \end{aligned}$$

Here $k = 2$. Define

$$w_1(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2}, \quad t_1(x) = x, \quad w_2(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}, \quad t_2(x) = x^2,$$

$$h(x) = 1, \quad c(\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\mu^2}{2\sigma^2}\right].$$

Then

$$f_X(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{2} w_j(\boldsymbol{\theta}) t_j(x)\right]$$

Example 3: Show that $Gamma(\alpha, \beta)$ belongs to an Exponential Family

Proof:

Example 4: $Unif(0, \theta)$ does not belong to an Exponential Family.

Alternative Parameterization of Exponential Families

An alternative parametrization of the exponential family of distributions in terms of "natural" or "canonical" parameters can be written as follows.

$$f_X(x|\boldsymbol{\eta}) \;=\; h(x)c^*(\boldsymbol{\eta}) \exp\left[\sum_{j=1}^{k} \eta_j t_j(x)\right]$$

The alternative parametrization can be achieved by defining $\eta_j = w_j(\boldsymbol{\theta})$ from the following equation,

$$f_X(x|\boldsymbol{\theta}) \;=\; h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(x)\right]$$

where $c^*(\boldsymbol{\eta}) = c \circ w(\boldsymbol{\theta})$. This alternative parametrization is most often used in the context of GLM (Generalized Linear Model).

Example 5: In $Bern(p)$ distribution, the canonical parameter is the logit function

$$\eta = \log\left(\frac{p}{1-p}\right),$$

since for $x = 0, 1$

$$\begin{aligned} f(x|p) &= p^x(1-p)^{1-x} \\ &= (1-p)\left(\frac{p}{1-p}\right)^x \\ &= (1+e^\eta)^{-1}\exp(\eta x) \end{aligned}$$

with $c(\eta) = (1+e^\eta)^{-1},\ k = 1,\ t(x) = x,\ h(x) = 1.$

Example 6 For $\mathcal{N}(\mu, \sigma^2)$ distribution,

$$
\begin{aligned}
f_X(x|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(\frac{\mu}{\sigma^2}x\right) \\
&= h(x)c(\boldsymbol{\eta}) \exp\left[\eta_1 t_1(x) + \eta_2 t_2(x)\right]
\end{aligned}
$$

where

$$
\begin{cases}
\boldsymbol{\eta} = (\eta_1, \eta_2) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right) \\
t_1(x) = x, \; t_2(x) = -x^2 \\
h(x) = 1/\sqrt{\pi} \\
c(\boldsymbol{\eta}) = \sqrt{\eta_2} \exp\left[-\eta_1^2/(2\eta_2)\right]
\end{cases}
$$

Sufficient Statistics and Exponential Families

Theorem 6.2.10: Let $X_1, \cdots, X_n$ *i.i.d.* with pdf $f_X(x|\boldsymbol{\theta})$ that belongs to an exponential family given by

$$f_X(x|\boldsymbol{\theta}) \;=\; h(x)c(\boldsymbol{\theta})\exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right]$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d),\ d \le k$. Then the following $T(\mathbf{X})$ is a sufficient statistic for $\boldsymbol{\theta}$.

$$T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \cdots, \sum_{j=1}^{n} t_k(X_j)\right)$$

Example 7: Let $X_1, \cdots, X_n$ *i.i.d.* $\mathcal{N}(\mu, \sigma^2)$, where both $\mu \in \mathbb{R}$, and $\sigma^2 > 0$ are unknown. Find sufficient statistics for μ and σ^2.

From the exponential family representation on Page 3,
$T_1(\mathbf{X}) = \sum_{j=1}^{n} X_j,\ T_2(\mathbf{X}) = \sum_{j=1}^{n} X_j^2$ are sufficient statistics for μ, σ^2 by Theorem 6.2.10.

Example 8: Let $X_1, \cdots, X_n$ *i.i.d.* $Pois(\lambda)$, where $\lambda > 0$ is unknown. Find sufficient statistic for λ.

Digression

Definition: Open Set A set A is open in $\mathbb{R}^k$ if for every $x \in A$, there exists a ϵ-ball $B(x, \epsilon)$ around x such that $B(x, \epsilon) \subset A$. Here

$$B(x, \epsilon) = \{y : ||y - x|| < \epsilon,\ y \in \mathbb{R}^k\}$$

where $||$ denotes a distance measure in $\mathbb{R}^k$.

Examples

- $A = (-1, 1)$: A is open in $\mathbb{R}$
- $A = (-\infty, 0) \times \mathbb{R}$: A is open in $\mathbb{R}^2$
- $A = (-1, 1]$: A is not open in $\mathbb{R}$
- $A = (-\infty, 0] \times \mathbb{R}$: A is not open in $\mathbb{R}^2$
- $A = \{(x, y) : x^2 + y^2 < 1\}$: A is open in $\mathbb{R}^2$
- $A = \{(x, y) : x \in (-1, 1), y = 0\}$: A is not open in $\mathbb{R}^2$
- $A = \{(x, y) : x \in \mathbb{R}, y = x^2\}$: A is not open in $\mathbb{R}^2$
- $A = \{1, 2, 3, \cdots, \}$: A is not open in $\mathbb{R}$

This is the only additional concept one needs to connect exponential families to completeness.

Completeness and Exponential Families

Theorem 5.2.11 & 6.2.25: Suppose $X_1, \cdots, X_n$ is a random sample from pdf or pmf $f_X(x|\theta)$ where

$$f_X(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(x)\right]$$

is a member of an exponential family. Define a statistic $T(\mathbf{X})$ by

$$\mathbf{T}(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \cdots, \sum_{j=1}^{n} t_k(X_j)\right)$$

If the set $\boldsymbol{\Theta} = \{w_1(\boldsymbol{\theta}), \cdots, w_k(\boldsymbol{\theta}),\ \forall \boldsymbol{\theta} \in \boldsymbol{\Omega}\}$ contains an open subset of $\mathbb{R}^k$, then the following are true.

(a) The distribution of $\mathbf{T}(\mathbf{X})$ is an exponential family of the form

$$f_T(u_1, \cdots, u_k|\boldsymbol{\theta}) = H(u_1, \cdots, u_k)[c(\boldsymbol{\theta})]^n \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) u_i\right]$$

(b) The family of distributions for the statistic $T(\mathbf{X})$

$$\mathbf{T}(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \cdots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is complete.

Example 9: Let $X_1, X_2, \ldots, X_n$ be a i.i.d. random sample from the following distributions. Identify complete, sufficient statistics:

(a) $\mathcal{N}(\mu, \sigma^2)$

(b) $Poisson(\lambda)$

(c) $Bernoulli(p)$

(d) $Beta(\alpha, \beta)$

(e) $Inverse \;\; Gaussian(\mu, \lambda)$ with pdf

$$f(x|\mu, \lambda) = \left(\frac{\lambda}{2\pi x^2}\right)^{1/2} \exp\left[-\lambda(x-\mu)^2/2\mu^2 x\right], \quad 0 < x < \infty, \;\; \mu > 0, \;\; \lambda > 0.$$

(f) $Negative \;\; Binomial(r, p)$ with pmf

$$\binom{r+x-1}{x} p^r (1-p)^x, \quad x = 0, 1, 2, \cdots, ; 0 < p < 1; \;\; r \; known.$$

Curved and Full Exponential Families

For an exponential family, if $d = \dim(\boldsymbol{\theta}) < k$, then this exponential family is called *curved exponential family.* if $d = \dim(\boldsymbol{\theta}) = k$, then this exponential family is called *full exponential family.* The sufficiency and completeness results only hold for *full exponential families.*

- $\mathcal{N}(\mu, \mu^2), \mu \in \mathbb{R}$ is a curved exponential family

The parameter space no longer contains an open set in R^2.

Review

Basic Terminology

Model $\mathcal{P} = \{f_{\mathbf{X}}(\mathbf{x}|\theta), \theta \in \Omega\}$

Random Variables $\mathbf{X} = (X_1, \cdots, X_n)$ that can be generated from $f_{\mathbf{X}}(\mathbf{x}|\theta)$.

Data $\mathbf{x} = (x_1, \cdots, x_n)$ that is generated from $f_{\mathbf{X}}(\mathbf{x}|\theta)$.

Statistic A function of data or random variables $T(\mathbf{x})$ or $T(\mathbf{X})$.

Sample Space A set of possible values of random variables $\mathcal{X}$.

Partition $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\} \subseteq \mathcal{X}$.

Data Reduction Partition of sample space in terms of particular statistic.

Sufficient Statistic

Concept The statistic contains all information about θ

Definition 6.2.1 $f_{\mathbf{X}}(\mathbf{x}|T(\mathbf{X}))$ does not depend on θ

Theorem 6.2.2 $f_{\mathbf{X}}(\mathbf{x}|\theta)/q(T(\mathbf{X})|\theta)$ does not depend on θ $\Longrightarrow T(\mathbf{X})$ is sufficient.

Theorem 6.2.6 (Factorization) $f_{\mathbf{X}}(\mathbf{x}|\theta) = h(\mathbf{x})g(T(\mathbf{X})|\theta)$ $\Longleftrightarrow T(\mathbf{X})$ is sufficient.

Theorem 6.2.10 (Exponential Family) $(\sum_{i=1}^n t_1(X_i), \cdots, \sum_{i=1}^n t_k(X_i))$ is sufficient

Minimal Sufficient Statistic

Concept Sufficient statistic that achieves the maximum data reduction, or coarsest partition of the sample space.

Definition 6.2.11 T is sufficient and it is a function of every sufficient statistic.

Theorem 6.2.13 $T(\mathbf{X})$ is minimal sufficient if the following is true:
$f_{\mathbf{X}}(\mathbf{x}|\theta)/f_{\mathbf{X}}(\mathbf{y}|\theta)$ is constant as a function of θ
$\iff T(\mathbf{x}) = T(\mathbf{y})$

Non-unique MSS Any one-to-one function of MSS is also a MSS (i.e. MSS is not unique).

Unique partition The partition created by any minimal sufficient statistic is unique.

Theorem 6.2.28 Any complete sufficient statistic is also minimal sufficient.

Ancillary Statistic

Concept A statistic that does not have any information about θ.

Definition 6.2.16 Its distribution is constant to θ.

Location Family For location family of θ, $\{f(x-\theta) : \theta \in R\}$, range statistic is an ancillary statistic of θ.

Scale Family For scale family of θ, $\left\{\frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) : \sigma > 0\right\}$, median$(X)/\overline{X}$ or $X_{(1)}/X_{(n)}$ is an ancillary statistics for θ.

Theorem 6.2.24 (Basu) A complete sufficient statistic is independent of every ancillary statistic.

Complete Statistic

Concept Any non-zero function of the statistics cannot be ancillary (there is no unnecessary part).

Defined on family This family has to contain "many" distributions in order to be complete.

Definition 6.2.21 $E[g(T)|\theta] = 0 \Longrightarrow g(T) = 0$ almost surely across all θ.

Theorem 6.2.24 (Basu) A complete and (minimal) sufficient statistic is independent of every ancillary statistics.

Theorem 6.2.25 (Exponential Family) If the parameter space $\Theta = \{(w_1(\theta), \ldots, w_k(\theta) : \theta \in \Omega\}$ contains an open subset of $\mathbb{R}^k$, then $(\sum_{i=1}^n t_1(X_i), \cdots, \sum_{i=1}^n t_k(X_i))$ is complete.

Theorem 6.2.28 Any complete sufficient statistic is also minimal sufficient.

Example 10: Let $X_1, \cdots, X_n$ be *i.i.d.* random sample from the following pdf

$$f_X(x|\theta) = e^{-(x-\theta)} \exp(-e^{-(x-\theta)}), \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$$

1. Does the distribution belong to an exponential family?
2. What are canonical parameters for the distribution?
3. What is a sufficient statistic from the distribution?
4. Is the sufficient statistic also complete and/or minimal sufficient?

Representing into an exponential distribution

$$\begin{aligned} f_X(x|\theta) &= e^{-(x-\theta)} \exp(-e^{-(x-\theta)}) \\ &= e^{-x} e^{\theta} \exp(-e^{-x} e^{\theta}) \\ &= h(x) c(\theta) \exp[w(\theta) t(x)] \qquad \text{if} \end{aligned}$$

$$h(x) = e^{-x,} \qquad c(\theta) = e^{\theta}, \qquad w(\theta) = -e^{\theta}, \qquad t(x) = e^{-x}$$

Representing into a canonical form

$$f_X(x|\theta) \quad = \quad h(x)c^*(\eta)\exp[\eta t(x)] \qquad \text{if}$$

$$h(x) = e^{-x,} \qquad c^*(\eta) = -\eta, \qquad t(x) = e^{-x}, \qquad (\eta < 0)$$

Sufficiency

$$T(\mathbf{X}) = \sum_{i=1}^{n} t(X_i) = \sum_{i=1}^{n} e^{-X_i}$$

Completeness and minimal sufficiency

The parameter space $\mathbf{\Theta} = \{w(\theta) = -e^{\theta} : \theta \in \mathbb{R}\}$ contains an open set in $\mathbb{R}$, so it is complete by Theorem 6.2.25, and minimal sufficient by Theorem 6.2.28.