Lecture 2. Group families and exponential families

Last Thursday, we had an overview of this course. Basically, given a model, we will consider the optimal estimator in various settings.

In this class, we study two families of models, *group families and exponential families.* These two families cover many statistical models we used and will use. It is usually relatively easy to analyze those families. Many other families can be well approximated by them.

We first study **Group families**.
Location family: Let

$$X = \theta + U, \theta \in R,$$

where U is distributed according to some distribution F, for example, $N(0,1)$, or $C(0,1)$, for which

$$X \sim N(\theta, 1), \text{ or } C(\theta, 1), \theta \in R.$$

Local-scale family: Let U be a random variable with a distribution F. Define a set of transformations,

$$G = \{g : g(U) = \theta + \sigma U, \theta \in \mathbb{R}, \sigma > 0\}.$$

Let $X = g(U)$, then

$$P(X \leq x) = P(\theta + \sigma U \leq x) = P\left(U \leq \frac{x-\theta}{\sigma}\right) = F\left(\frac{x-\theta}{\sigma}\right).$$

The set of distributions above are said to form a location-scale family, for all $\theta \in \mathbb{R}$ and $\sigma > 0$.

Assume that U has a density function f. Then the density function of X is

$$f_X(x) = \frac{1}{\sigma} f\left(\frac{x-\theta}{\sigma}\right).$$

Concrete examples.

- Normal $N(\theta, \sigma^2)$:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}.$$

- Double exponential $DE(\theta, \sigma)$:

$$f_X(x) = \frac{1}{2\sigma} e^{-|x-\theta|/\sigma}.$$

.

- Cauchy $C(\theta, \sigma)$:

$$f_X(x) = \frac{\sigma}{\pi} \frac{1}{\sigma^2 + (x-\theta)^2}.$$

Group family.
The location-scale family

$$G = \{g : g(U) = \theta + \sigma U, \theta \in \mathbb{R}, \sigma > 0\}$$

has the following properties:

- Closed under composition: For $g_1(U) = \theta_1 + \sigma_1 U$ and $g_2(U) = \theta_2 + \sigma_2 U$, we have $(g_2 \circ g_1)(U) = g_2[g_1(U)] = \theta_2 + \sigma_2\theta_1 + \sigma_2\sigma_1 U$, which implies $g_2 \circ g_1 \in G$.
- Closed under inversion: For $g(U) = \theta + \sigma U$, we have $U = [g(U) - \theta]/\sigma$, i.e., $g^{-1}(U) = (U - \theta)/\sigma$, which implies $g^{-1} \in G$.

A class G of transformations is called a *transformation group* if it's closed under composition and inversion.

A *group family* of distributions is a family obtained by subjecting a random variable to a transformation group. Let G be a transformation group and U be a random variable. The set of distributions of $g(U)$ with $g \in G$ form a group family.

Linear model (location). Let $\mathbf{X} = A_{n\times p}\beta + \mathbf{U}$ ($A_{n\times p}$ is fixed), where $\beta = (\beta_1, \cdots, \beta_p)' \in \mathbb{R}^p$ is an unknown parameter. This is a group family. Define $G = \{g_\beta : g_\beta(\mathbf{Y}) = A_{n\times p}\beta + \mathbf{Y}\}$. Then

- Since $g_\beta(g_\alpha(\mathbf{Y})) = A_{n\times p}\beta + g_\alpha(\mathbf{Y}) = A_{n\times p}\beta + A_{n\times p}\alpha + \mathbf{Y} = A_{n\times p}(\beta + \alpha) + \mathbf{Y}$, then $g_\beta \circ g_\alpha \in G$.
- Since $g_\beta^{-1}(\mathbf{Y}) = A_{n\times p}(-\beta) + \mathbf{Y}$, then $g_\beta^{-1} \in G$.

Multivariate normal distribution:

$$\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} + B \begin{pmatrix} U_1 \\ \vdots \\ U_p \end{pmatrix}, U_i \sim N(0,1).$$

Let $\mathbf{X} = (X_1, \cdots, X_p)'$, and $\theta = (\theta_1, \cdots, \theta_p)'$. Assume that $B_{p\times p}$ is non-singular. The set of distributions of $\mathbf{X} = \mathbf{a} + B\mathbf{U}$ form a group family.

Let $\Sigma = B_{p\times p}B'_{p\times p}$. The density of $\mathbf{X}$ is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \frac{(\det\Sigma)^{-1/2}}{(2\pi)^{p/2}} \exp\left(-\frac{(\mathbf{x}-\theta)'\Sigma^{-1}(\mathbf{x}-\theta)}{2}\right) \\ &= \exp\left(-\frac{\mathbf{x}'\Sigma^{-1}\mathbf{x} - 2\theta'\Sigma^{-1}\mathbf{x}}{2} - \frac{\theta'\Sigma^{-1}\theta}{2} + \log\left(\frac{(\det\Sigma)^{-1/2}}{(2\pi)^{p/2}}\right)\right). \end{aligned}$$

Exponential families.

A family $\{P_\theta\}$ of distributions is called an s-dimensional *exponential family* if P_θ has a density of the following form

$$p_\theta(x) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right) h(x),\ h(x) \neq 0,$$

w.r.t. a measure μ. Note that $\int p_\theta(x) d\mu = 1$. Usually, it is more convenient to study the canonical form,

$$p_\eta(x) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x),\ h(x) \neq 0.$$

Remark 1. For an exponential family, the support of $p_\theta(x)$ (or $p_\eta(x)$) does not depend on θ, since

$$\{x :\ p_\theta(x) > 0\} = \{x :\ h(x) > 0\}.$$

Remark 2. Consider an exponential family of the following canonical form,

$$p_\eta(x) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x),\ h(x) \neq 0.$$

We have

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

i.e.,

$$\int \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x) dx = e^{A(\eta)}.$$

The equation above implies $\eta = (\eta_1, \ldots, \eta_s)$ should satisfy

$$\int \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x) d\mu(x) < +\infty.$$

The set Ξ of all $(\eta_1, \cdots, \eta_s)$ for which the equation above holds is called the *natural parameter space*, and $\eta = (\eta_1, \cdots, \eta_s)$ is called the natural parameter.

Remark 3. If $0 < \int \exp\left(\sum \eta_i T_i(x)\right) h(x) d\mu(x) < +\infty$ for some η, we may define

$$A(\eta) = \log\left(\int \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x)\right),$$

such that

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

to form an exponential family.

Remark 4. The representation

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

is called *minimal* if neither the T's or the η's satisfy a linear constraint. Suppose T's satisfies a linear constraint, i.e., there is $\mathbf{a} = (a_1, \cdots, a_s) \neq (0, \cdots, 0)$ s.t. $\sum_{i=1}^{s} a_i T_i(x) = c$ (for some c which is allowed to be nonzero) for all x. For example, for $T_1(x) = x$ and $T_2(x) = x + 1$, there does not exist any $(a_1, a_2) \neq (0, 0)$ s.t. $a_1 T_1(x) + a_2 T_2(x) = 0$, but $T_1(x) - T_2(x) = -1$. Without loss of generality, we assume $a_1 \neq 0$, $T_1(x) = c/a_1 - \sum_{i=2}^{s} a_i T_i(x)/a_1$, then

$$\begin{aligned}
p_\eta(x) &= \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x) \\
&= \exp\left\{\eta_1 \left(\frac{c}{a_1} - \sum_{i=2}^{s} \frac{a_i}{a_1} T_i(x)\right) + \sum_{i=2}^{s} \eta_i T_i(x) - A(\eta)\right\} h(x) \\
&= \exp\left\{\sum_{i=2}^{s} \left(\eta_i - \frac{a_i \eta_1}{a_1}\right) T_i(x) - \left(A(\eta) - \frac{c\eta_1}{a_1}\right)\right\} h(x).
\end{aligned}$$

Let $\eta_i^* = \eta_i - a_i \eta_1 / a_1, 2 \leq i \leq s$, $A^*(\eta) = A(\eta) - c\eta_1/a_1$, then

$$p_\eta(x) = \exp\left(\sum_{i=2}^{s} \eta_i^* T_i(x) - A^*(\eta)\right) h(x),$$

where $A^*(\eta)$ a function of $\eta_2^* \cdots \eta_s^*$, since $\int \exp\left(\sum_{i=2}^{s} \eta_i^* T_i(x) - A^*(\eta)\right) h(x) d\mu(x) = 1$, which implies

$$\log\left\{\int \exp\left(\sum_{i=2}^{s} \eta_i^* T_i(x)\right) h(x) d\mu(x)\right\} = A^*(\eta),$$

thus the dimension of the exponential family representation can be reduced to $s - 1$. Similarly, if η's satisfies a linear constraint, we can do reduction as well.

If the representation of the canonical form is minimal and the natural parameter space contains an s-dimensional rectangle, then we say the s-dimensional exponential family is of *full rank*.

Example. Curved normal family $N(\theta, \theta^2)$:

$$\begin{aligned} p_\theta(x) &= \exp\left(-\frac{(x-\theta)^2}{2\theta^2}\right)\frac{1}{\sqrt{2\pi}\theta} \\ &= \exp\left(-\frac{x^2}{2\theta^2}+\frac{x}{\theta}-\frac{1}{2}\right)\exp\left(-\log(\sqrt{2\pi}\theta)\right) \\ &= \exp\left(-\frac{x^2}{2\theta^2}+\frac{x}{\theta}-\frac{1}{2}-\log(\sqrt{2\pi}\theta)\right). \end{aligned}$$

Let $T_1(X) = X^2, \eta_1 = -1/(2\theta^2), T_2(X) = X, \eta_2 = 1/\theta$, and write

$$A(\eta) = \frac{1}{2} + \log(\sqrt{2\pi}\theta) = \frac{1}{2} + \log(\sqrt{2\pi}) - \log(\eta_2).$$

For T_1 and T_2, we cannot find $(a_1, a_2) \neq (0,0)$ s.t. $a_1x^2 + a_2x = c$ for all x. It is also true for η_1 and η_2. The expression $p_\eta(x) = \exp[\eta_1 T_1 + \eta_2 T_2 - A(\eta)]$ is minimal, but it is not full rank, since $\eta_1 = -\eta_2^2/2$ is a curve in $\mathbb{R}^2$, i.e., Ξ does not contain any open rectangle.

Lecture 3. Moments and maximum entropy for exponential family
Let the canonical form of the density for an exponential family be as follows,

$$p_\eta(x) = \exp\left(\sum_{i=1}^{s}\eta_i T_i(x) - A(\eta)\right)h(x),\ h(x) \neq 0.$$

It is *minimal* if there is not any linear constraint for the T's or the η's. It is *full rank*, if it is minimal and the natural parameter space contains an s-dimensional rectangle.

Calculating moments.
Warning: In the following calculations we exchange the order of derivatives and integrations without justifications, but you can make them rigorous.)

First moment: $ET_1(X) = \int T_1(x)p_\eta(x)d\mu(x)$, i.e.,

$$ET_1(X) = \int T_1(x)\exp\left(\sum_{i=1}^{s}\eta_i T_i(x) - A(\eta)\right)h(x)d\mu(x).$$

Since $\int p_\eta(x)d\mu(x) = 1$, we have

$$\frac{\partial}{\partial\eta_1}\int p_\eta(x)d\mu(x) = \int\left(T_1(x) - \frac{\partial A(\eta)}{\partial\eta_1}\right)p_\eta(x)d\mu(x) = 0.$$

which implies

$$ET_1(X) = \int T_1(x)p_\eta(x)d\mu(x) = \frac{\partial A(\eta)}{\partial\eta_1}\int p_\eta(x)d\mu(x) = \frac{\partial A(\eta)}{\partial\eta_1}.$$

Maximum entropy:
For $X \sim P_\theta$ with a density p_θ w.r.t. a measure μ, we want to maximize

$$H(p) = -\int p \log p$$

subject to

$$\int p = 1$$

and

$$ET_i = a_i,\ i = 1, 2, \ldots s.$$

The solution can be shown to be of the form

$$p(x) \propto \exp\left(\sum_{i=1}^{s} \lambda_i T_i(x)\right).$$

To obtain the solution, we consider the following functional

$$F(p) = -\int p(x) \log p(x)\, d\mu + \lambda_0 \left(\int p(x)\, d\mu - 1\right) + \sum_{i=1}^{s} \lambda_i \left(\int T_i(x)\, p(x)\, d\mu - a_i\right).$$

Informally we see $p(x)$ as an infinite dimensional vector (p_x) and $\int p(x) \log p(x)$ as $\sum p_x \log p_x$, then the maximizer is attained at

$$-\log p(x) - 1 + \lambda_0 + \sum_{i=1}^{s} \lambda_i T_i(x) = 0,$$

which leads to the exponential family,

$$p(x) = \exp\left(-1 + \lambda_0 + \sum_{i=1}^{s} \lambda_i T_i(x)\right).$$

If there are $\lambda_i = \lambda_i^*$, $i = 0, 1, 2, \ldots, s$ such that the constraints are satisfied, we obetain

$$p^*(x) = \exp\left(-1 + \lambda_0^* + \sum_{i=1}^{s} \lambda_i^* T_i(x)\right).$$

A formal maximum entropy argument: For any p satisfying the con-

straints above, we have

$$\begin{aligned}
H(p) &= -\int p\log p \\
&= -\int p\left(\log\frac{p}{p^*}+\log p^*\right) \\
&= -\int p\log\frac{p}{p^*}-\int p\log p^* \ \longleftarrow \text{Due to the fact: } \int p\log\frac{p}{p^*}\geq 0 \\
&\leq -\int p\log p^* \\
&= -\int p\left(-1+\lambda_0^*+\sum_{i=1}^{s}\lambda_i^* T_i(x)\right) \\
&= -\int p^*\left(-1+\lambda_0^*+\sum_{i=1}^{s}\lambda_i^* T_i(x)\right) \longleftarrow \text{ Due to the constraints} \\
&= -\int p^*\log p^* = H(p^*).
\end{aligned}$$

Second moment:

$$\begin{aligned}
\frac{\partial^2}{\partial\eta_j\partial\eta_i}\int p_\eta(x)d\mu(x) &= \frac{\partial}{\partial\eta_j}\int\left(T_i(x)-\frac{\partial A}{\partial\eta_i}\right)p_\eta(x)d\mu(x) \\
&= -\frac{\partial^2 A}{\partial\eta_j\eta_i}+\int\left(T_i(x)-\frac{\partial A}{\partial\eta_i}\right)\frac{\partial}{\partial\eta_j}p_\eta(x)d\mu(x) \\
&= -\frac{\partial^2 A}{\partial\eta_j\eta_i}+\int\left(T_i(x)-\frac{\partial A}{\partial\eta_i}\right)\left(T_j(x)-\frac{\partial A}{\partial\eta_j}\right)p_\eta(x)d\mu(x) \\
&= 0,
\end{aligned}$$

thus

$$\operatorname{cov}(T_i,T_j)=\int\left(T_i(x)-\frac{\partial A}{\partial\eta_i}\right)\left(T_j(x)-\frac{\partial A}{\partial\eta_j}\right)p_\eta(x)d\mu(x)=\frac{\partial^2 A}{\partial\eta_j\eta_i}.$$

Remark. We should be able to calculate higher moments by calculating higher order of derivations. For $s=1$, we have $p_\eta(x)=e^{\eta T(x)-A(\eta)}h(x)$, and know that $ET=A'(\eta)$, $\operatorname{var}(T)=A''(\eta)$. It can be shown that $E(T-ET)^3=A^{(3)}(\eta)$, and $E(T-ET)^4=A^{(4)}+3(A'')^2$.