Statistical Inference I **Prof. Chin-Tsang Chiang**

Lecture Notes 16

November 16, 2015 **Scribe: Wei-Chang Lee, Chi-Ning Chou**

0.1 Exponential families

A family of distributions is a collection of pdf (pmf) sharing some common properties. The exponential families is a family of distributions in the following form:

$$f_X(x|\theta) = h(x)c(\theta)\exp(\sum_{i=1}^{k} w_i(\theta)t_i(x))$$

, where $h, c \geq 0$ and w_i, t_i are real-valued function.
Here, we can think of these functions as:

- $t_i(x)$: functionals of x, or empirical moment.
- $w_i(\theta)$: linear weight.
- $h(x), c(\theta)$: normalization term.

Remark:

1. When $dim(\theta) < k$, we call it a *curved* exponential family. Otherwise, we call it a *full* exponential family. Intuitively, parameters in a curved exponential family have some correlation, which can be geometrically considered as a curve.
2. In fact, we can **re-parametrize** an exponential family from parameter space θ to a *natural* parameter space η in the following sense:

$$f_X(x|\eta) = h(x)c^*(\eta)\exp(\sum_{i=1}^{k} \eta_i t_i(x))$$

, where $\mathcal{H} = \{\eta : \int_{-\infty}^{\infty} h(x)\exp(\sum_{i=1}^{k} \eta_i t_i(x))dx < \infty\}$ is the natural parameter space. Intuitively, parameters in η are independent to each others.

Now, you might wonder, what's the benefit we can get from exponential family? Why we want to discuss it? For now, we can benefit from the following theorem.

Theorem 1 (properties of exponential families)
Let X has the pdf (pmf) $f_X(x|\theta) = h(x)c(\theta)\exp(\sum_{i=1}^{k} w_i(\theta)t_i(x))$, then

- $\mathbb{E}[\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X)] = \frac{-\partial}{\partial \theta_j} \ln c(\theta)$

- $Var[\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X)] = \frac{-\partial^2}{\partial \theta_j^2} \ln c(\theta) - \mathbb{E}[\sum_{i=1}^{k} \frac{\partial^2 w_i(\theta)}{\partial \theta_j^2} t_i(X)]$

Proof: Start with the observation that

$$1 = \int_{-\infty}^{\infty} f_X(x|\theta) dx$$
$$0 = \frac{\partial}{\partial \theta_j} \int_{-\infty}^{\infty} f_X(x|\theta) dx$$

We have

$$\begin{aligned}
0 &= \frac{\partial}{\partial \theta_j} \int_{-\infty}^{\infty} f_X(x|\theta) dx \\
(\because \text{Fubini}) &= \int_{-\infty}^{\infty} \frac{\partial f_X(x|\theta)}{\partial \theta_j} dx = \int_{-\infty}^{\infty} \frac{\partial \ln f_X(x|\theta)}{\partial \theta_j} f_X(x|\theta) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{f_X(x|\theta)} [h(x) \frac{\partial c(\theta)}{\partial \theta_j} \exp(\sum_{i=1}^{k} w_i(\theta) t_i(x)) + f_X(x|\theta) \sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(x)] f_X(x|\theta) dx \\
&= \int_{-\infty}^{\infty} h(x) c(\theta) \exp(\sum_{i=1}^{k} w_i(\theta) t_i(x)) \frac{\partial \ln c(\theta)}{\partial \theta_j} dx + \mathbb{E}[\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(x)] \\
&= \mathbb{E}[\frac{\partial \ln c(\theta)}{\partial \theta_j}] + \mathbb{E}[\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(x)] = \frac{\partial \ln c(\theta)}{\partial \theta_j} + \mathbb{E}[\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(x)]
\end{aligned}$$

As a result,

$$\mathbb{E}[\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(x)] = -\frac{\partial \ln c(\theta)}{\partial \theta_j}$$

Similarly, we can show

$$Var[\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X)] = \frac{-\partial^2}{\partial \theta_j^2} \ln c(\theta) - \mathbb{E}[\sum_{i=1}^{k} \frac{\partial^2 w_i(\theta)}{\partial \theta_j^2} t_i(X)]$$

■

Remark: $\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(x)$ is the variation of the descriptive term w.r.t parameter θ_j.

0.1.1 Generalized linear model

Generalized linear model is composed of two parts: *random component* and *systematic component*. Basically, we have a response random variable Y and a set of explanatory random variables $\{Z_1, ..., Z_p\}$.

- Random component: it is from exponential family

$$f_Y(y|\theta, \phi) = \exp(\frac{y\theta - b(\theta)}{a(\phi)} + c(y; \phi))$$

  , where $\theta = \theta(Z_1, ..., Z_p)$ which is a functional of explanatory random variables and ϕ is the dispersion parameter.

- Systematic component: it describes the mean of the response random variable with a *link function* of a linear combination of explanatory random variables.

$$\mathbb{E}[Y|\theta(Z_1, ..., Z_p)] = h(\beta^T \mathbf{Z})$$

  , where h is the link function.

From Theorem 1, we have

- $\mathbb{E}[Y|\theta, \phi] = \frac{d}{d\theta} b(\theta)$
- $Var[Y|\theta, \phi] = a(\phi) \frac{d^2}{d\theta^2} b(\theta) = a(\phi) \frac{d}{d\theta} \mathbb{E}[Y|\theta, \phi]$

Proof: By Theorem 1, we have

$$\mathbb{E}[Y|\theta, \phi] = \mathbb{E}[\frac{\partial}{\partial\theta} \frac{y\theta}{a(\phi)}]a(\phi) = -a(\phi)\frac{\partial}{\partial\theta} \ln \exp(\frac{-b(\theta)}{a(\phi)}) = \frac{d}{d\theta} b(\theta)$$

$$Var[Y|\theta, \phi] = Var[\frac{\partial}{\partial\theta} \frac{y\theta}{a(\phi)}]a^2(\phi) = -a^2(\phi)[\frac{\partial^2}{\partial\theta^2} \ln \exp(\frac{-b(\theta)}{a(\phi)}) - 0] = a(\phi)\frac{\partial^2}{\partial\theta^2} b(\theta)$$

■