STAT 512: Statistical Inference **Autumn 2022**

Quiz Session 7: Exponential Families

Yikun Zhang *November 23, 2022*

Some parts of this notes are based on Section 3.4 in Casella and Berger [2002], Section 11.4 in Perlman [2020], and the notes from Professor Peter Hoff[1].

Definition 1 (Exponential family)**.** A family of distributions is said to belong to an *exponential* family if the probability density function (PDF) for continuous distributions or probability mass function (PMF) for discrete distributions can be written as:

$$f(x|\boldsymbol{\theta}) = h(x)a(\boldsymbol{\theta})\exp\left(\sum_{i=1}^k \eta_i(\boldsymbol{\theta})T_i(x)\right), \tag{1}$$

where $h(x) \geq 0, T_i(x), i = 1, ..., k$ are real-valued functions of the observation x that do not depend on $\boldsymbol{\theta}$, and $a(\boldsymbol{\theta}), \eta_i(\boldsymbol{\theta}), i = 1, ..., n$ are real-valued functions of the possibly vector-valued parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_s)^T \in \Theta \subset \mathbb{R}^s$ that cannot depend on x. To make $f(x|\boldsymbol{\theta})$ identifiable, we also require that $(T_1(x), ..., T_k(x))$ *is a k-dimensional statistic that does not satisfy any linear constraint*[2].

Sometimes, we reparameterize the distribution (1) of an exponential family as:

$$f(x|\boldsymbol{\eta}) = h(x)a^*(\boldsymbol{\eta})\exp\left(\sum_{i=1}^k \eta_i T_i(x)\right), \tag{2}$$

where $h(x)$ and $T_i(x)$ are the same functions as in the original parameterization (1). The set

$$\mathcal{H} = \left\{\boldsymbol{\eta} = (\eta_1, ..., \eta_k) : \int_{-\infty}^{\infty} h(x)\exp\left(\sum_{i=1}^k \eta_i T_i(x)\right) dx < \infty\right\}$$

is called the *natural parameter space* for the family. (The integral is replaced by a sum over the values of x for which $h(x) > 0$ if X is discrete.) Since the original $f(x|\boldsymbol{\theta})$ in (1) is a PDF or PMF, the set $\{\eta = (\eta_1(\boldsymbol{\theta}), ..., \eta_k(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \Theta\}$ must be a subset of the natural parameter space.

Remark 1. *Another traditional definition of the exponential family is given by Brown [1986]: Let ν be a nonnegative function (or a σ-finite measure on the Borel subsets) of $\mathbb{R}^d$. Define the set*

$$\mathcal{N} = \mathcal{N}_\nu = \left\{\boldsymbol{\theta} \in \Theta : \int_{\mathcal{X}} e^{\boldsymbol{\theta}^T x}\nu(x)\, dx < \infty\right\},$$

where $\mathcal{X} \subset \mathbb{R}^d$. We let $\lambda(\boldsymbol{\theta}) = \int_{\mathcal{X}} e^{\boldsymbol{\theta}^T x}\nu(x)\, dx$ and the set of probability densities defined by

$$f(x|\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\theta}^T x}\nu(x)}{\lambda(\theta)}, \quad x \in \mathcal{X} \text{ and } \theta \in \mathcal{N}$$

is an exponential family.

[1] See http://www2.stat.duke.edu/~pdh10/Teaching/581/LectureNotes/expofam.pdf.

[2] If $(T_1(x), ..., T_k(x))$ does satisfy a linear constraint, the natural parameter space $\mathcal{H}$ will include points that correspond to the same probability distribution

Remark 2. *The exponential family is also known as the Darmois-Koopman-Pitman family/type*[3] *to credit their contributions to the sufficient statistics on exponential statistics [Andersen, 1970]. One sufficiency result related to exponential families is that, for a random sample* $\boldsymbol{X} = \{X_1, ..., X_n\}$ *from* $f(x|\boldsymbol{\theta})$ *in (1),*

$$T(\boldsymbol{X}) = \left(\sum_{j=1}^{n} T_1(X_j), ..., \sum_{j=1}^{n} T_k(X_j)\right)$$

is a sufficient statistic for $\boldsymbol{\theta}$ *(see Theorem 6.2.10 in Casella and Berger 2002) and is also minimal sufficient by Lehmann and Scheffé [1950] (see Theorem 6.2.13 in Casella and Berger 2002) provided that the parameter space* $\Theta \subset \mathbb{R}^s$ *affinely spans* $\mathbb{R}^s$. *Furthermore,* $T(\boldsymbol{X})$ *is a complete statistic if* $\{(\eta_1(\boldsymbol{\theta}), ..., \eta_k(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \Theta\}$ *contains an open set in* $\mathbb{R}^k$*; see Theorem 6.2.25 in Casella and Berger [2002].*

Remark 2 demonstrates that the concept of exponential families is very useful in deriving some statistical properties of a statistic. However, what drives statisticians to study exponential families in such a deep level is due to its ubiquity in probability distributions.

Problem 1 (Exercises 3.28 and 3.29 in Casella and Berger 2002)**.** *Show that each of the following families is an exponential family and describe its natural parameter space:*

- *(a)* *The normal distribution family* $N(\mu, \sigma^2)$ *with either parameter* μ *or* σ *known or both unknown.*
- *(b)* *The Gamma distribution family* $\text{Gamma}(\alpha, \beta)$ *with either parameter* α *or* β *known or both unknown.*
- *(c)* *The Beta distribution family* $\text{Beta}(\alpha, \beta)$ *with either parameter* α *or* β *known or both unknown.*
- *(d)* *The Poisson distribution family* $\text{Poisson}(\lambda)$ *with parameter* $\lambda > 0$ *unknown.*
- *(e)* *The negative binomial distribution family* $\text{NegBinomial}(r, p)$ *with* r *known and parameter* $0 < p < 1$ *unknown.*
- *(f)* *The binomial distribution family* $\text{Binomial}(n, p)$ *with the number of trials* n *fixed.*
- *(g)* *The multinomial distribution family* $\text{Multinomial}(n; p_1, ..., p_m)$ *with the number of trials* n *fixed.*

Proof. (a) Note that the density of $N(\mu, \sigma^2)$ is given by

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right).$$

(i) When both μ and σ are unknown, we take

$$h(x) = 1, \quad a(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right), \quad T_1(x) = x^2, \quad \text{and} \quad T_2(x) = x$$

with the natural parameters as $(\eta_1, \eta_2) = (\eta_1(\mu, \sigma), \eta_2(\mu, \sigma)) = \left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right)$ in (1) so that they lie in the space $\{(\eta_1, \eta_2) \in \mathbb{R}^2 : \eta_1 < 0, -\infty < \eta_2 < \infty\}$.

(ii) When μ is known, we take

$$h(x) = 1, \quad a(\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \mathbb{1}_{(0,\infty)}(\sigma^2), \quad \text{and} \quad T(x) = (x-\mu)^2$$

with the natural parameter $\eta(\sigma^2) = -\frac{1}{2\sigma^2}$ in (1) so that it lies in the space $\{\eta \in \mathbb{R} : \eta < 0\}$.

[3] See https://en.wikipedia.org/wiki/Exponential_family.

(iii) When σ^2 is known, we take

$$h(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right), \quad a(\mu) = \exp\left(-\frac{\mu^2}{2\sigma^2}\right), \quad \text{and} \quad T(x) = \frac{x}{2\sigma^2}$$

with the natural parameter $\eta(\mu) = \mu$ in (1) so that it lies in the space $\mathbb{R}$.

(b) Notice that the density of Gamma(α, β) is given by

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}\cdot \mathbb{1}_{(0,\infty)}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}\exp\left[(\alpha-1)\log x - \beta x\right]\cdot \mathbb{1}_{(0,\infty)}(x),$$

where $\alpha, \beta > 0$.

(i) When both α and β are unknown, we take

$$h(x) = \mathbb{1}_{(0,\infty)}(x), \quad a(\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}, \quad T_1(x) = \log x, \quad \text{and} \quad T_2(x) = x$$

with the natural parameters as $(\eta_1,\eta_2) = (\eta_1(\alpha,\beta), \eta_2(\alpha,\beta)) = (\alpha-1, -\beta)$ in (1) so that they lie in the space $\{(\eta_1,\eta_2)\in\mathbb{R}^2 : \eta_1 > -1, \eta_2 < 0\}$.

(ii) When α is known, we take

$$h(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)}\mathbb{1}_{(0,\infty)}(x), \quad a(\beta) = \beta^\alpha, \quad \text{and} \quad T(x) = x$$

with the natural parameter $\eta(\beta) = -\beta$ in (1) so that it lies in the space $\{\eta\in\mathbb{R} : \eta < 0\}$.

(iii) When β is known, we take

$$h(x) = e^{-\beta x}\mathbb{1}_{(0,\infty)}(x), \quad a(\alpha) = \frac{\beta^\alpha}{\Gamma(\alpha)}, \quad \text{and} \quad T(x) = \log x$$

with the natural parameter $\eta(\alpha) = \alpha - 1$ in (1) so that it lies in the space $\{\eta\in\mathbb{R} : \eta > -1\}$.

(c) Recall that the density of Beta(α,β) is given by

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}_{(0,1)}(x) = \frac{1}{B(\alpha,\beta)}\exp\left[(\alpha-1)\log x + (\beta-1)\log(1-x)\right]\mathbb{1}_{(0,1)}(x),$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\alpha,\beta > 0$.

(i) When both α and β are unknown, we take

$$h(x) = \mathbb{1}_{(0,1)}(x), \quad a(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}, \quad T_1(x) = \log x, \quad \text{and} \quad T_2(x) = \log(1-x)$$

with the natural parameters as $(\eta_1,\eta_2) = (\eta_1(\alpha,\beta),\eta_2(\alpha,\beta)) = (\alpha-1,\beta-1)$ in (1) so that they lie in the space $\{(\eta_1,\eta_2)\in\mathbb{R}^2 : \eta_1 > -1, \eta_2 > -1\}$.

(ii) When α is known, we take

$$h(x) = x^{\alpha-1}\mathbb{1}_{(0,1)}(x), \quad a(\beta) = \frac{1}{B(\alpha,\beta)}, \quad \text{and} \quad T(x) = \log(1-x)$$

with the natural parameter $\eta(\beta) = \beta - 1$ in (1) so that it lies in the space $\{\eta\in\mathbb{R} : \eta > -1\}$.

(iii) When β is known, we take

$$h(x) = (1-x)^{\beta-1}\mathbb{1}_{(0,1)}(x), \quad a(\alpha) = \frac{1}{B(\alpha,\beta)}, \quad \text{and} \quad T(x) = \log x$$

with the natural parameter $\eta(\alpha) = \alpha - 1$ in (1) so that it lies in the space $\{\eta \in \mathbb{R} : \eta > -1\}$.

(d) The probability mass function of Poisson(λ) is

$$P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \cdot \mathbb{1}_{\{0,1,2,\dots\}}(x) = \frac{1}{x!}\exp\left[x\log\lambda - \lambda\right] \cdot \mathbb{1}_{\{0,1,2,\dots\}}(x),$$

where $\lambda > 0$. We take

$$h(x) = \frac{1}{x!} \cdot \mathbb{1}_{\{0,1,2,\dots\}}(x), \quad a(\lambda) = e^{-\lambda}, \quad \text{and} \quad T(x) = x$$

with the natural parameter $\eta(\lambda) = \log \lambda$ in (1) so that it lies in the space $\mathbb{R}$.

(e) The probability mass function of NegBinomial(r, p) with r known is given by

$$P(x|p) = \binom{r+x-1}{x} p^r (1-p)^x \cdot \mathbb{1}_{\{0,1,2,\dots\}}(x) = \binom{r+x-1}{x} p^r \exp\left[x\log(1-p)\right] \cdot \mathbb{1}_{\{0,1,2,\dots\}}(x),$$

where $0 < p < 1$. We take

$$h(x) = \binom{r+x-1}{x} \cdot \mathbb{1}_{\{0,1,2,\dots\}}(x), \quad a(p) = p^r, \quad \text{and} \quad T(x) = x$$

with the natural parameter $\eta(p) = \log(1-p)$ in (1) so that it lies in the space $\{\eta \in \mathbb{R} : \eta < 0\}$.

(f) The probability mass function of Binomial(n, p) is given by

$$P(x|p) = \binom{n}{p} p^x (1-p)^{n-x} \cdot \mathbb{1}_{\{0,1,\dots,n\}}(x) = \binom{n}{p}(1-p)^n \exp\left[x \log\left(\frac{p}{1-p}\right)\right] \cdot \mathbb{1}_{\{0,1,\dots,n\}}(x),$$

where $0 < p < 1$. We take

$$h(x) = \binom{n}{x} \cdot \mathbb{1}_{\{0,1,\dots,n\}}(x), \quad a(p) = (1-p)^n, \quad \text{and} \quad T(x) = x$$

with the natural parameter $\eta(p) = \log\left(\frac{p}{1-p}\right)$ in (1) so that it lies in the space $\mathbb{R}$.

(g) The probability mass function of Multinomial($n; p_1, \dots, p_m$) is given by

$$\begin{aligned}
&P(x_1, \dots, x_m | p_1, \dots, p_m) \\
&= \frac{n!}{x_1! \cdots x_m!} \cdot p_1^{x_1} \cdots p_m^{x_m} \cdot \mathbb{1}_{\{\sum_{i=1}^m x_i = n\}} \cdot \mathbb{1}_{\{\sum_{i=1}^m p_i = 1\}} \\
&= \frac{n!}{x_1! \cdots x_m!} \cdot \exp\left(x_1 \log p_1 + \cdots x_m \log p_m\right) \cdot \mathbb{1}_{\{\sum_{i=1}^m x_i = n\}} \cdot \mathbb{1}_{\{\sum_{i=1}^m p_i = 1\}} \\
&= \frac{n!}{x_1! \cdots x_m!} \cdot e^{n\log(1-p_1-\cdots p_{m-1})} \exp\left[x_1 \log\left(\frac{p_1}{1-p_1-\cdots-p_{m-1}}\right) + \cdots + x_m \log\left(\frac{p_{m-1}}{1-p_1-\cdots-p_{m-1}}\right)\right].
\end{aligned}$$

Thus, we take

$$h(x) = \frac{n!}{x_1! \cdots x_m!}, \quad a(p_1, \dots, p_m) = e^{n\log(1-p_1-\cdots p_{m-1})} = e^{n \log p_m}, \quad \text{and} \quad T_i(x) = x_i \text{ for } i = 1, \dots, m-1$$

with the natural parameters $\boldsymbol{\eta} = (\eta_1, ..., \eta_{m-1}) = \left(\log\left(\frac{p_1}{1-p_1-\cdots-p_{m-1}}\right), ..., \log\left(\frac{p_{m-1}}{1-p_1-\cdots-p_{m-1}}\right)\right)$ so that they lie in the space $\mathbb{R}^{m-1}$. □

Definition 2. Within the exponential families, we can further classify them into two categories based on the configuration of the (natural) parameter space as:

- *Full exponential family*: If the (natural) parameter space for the distribution family (1) contains an k-dimensional open set, then it is called a full exponential family.
- *Curved exponential family*: If the (natural) parameter space for the distribution family (1) only contains an s-dimensional open set with $s < k$, then it is called a curved exponential family.

All the distribution families in Problem 1 so far are full exponential families. We present some examples of curved exponential families in Problem 2 below.

Problem 2. *Show that each of the following families is a curved exponential family and describe the curve on which the parameter vector $\boldsymbol{\theta}$ lies.*

(a) The normal distribution family $N(\mu, \mu^2)$ with $\mu > 0$.

(b) The Gamma distribution family $\text{Gamma}(\alpha, \alpha)$ with $\alpha > 0$.

(c) The distribution of (X, Y) with $X \sim \text{Binomial}(n, p), Y \sim \text{Binomial}(m, p^2)$ being independent, where m, n are fixed and $0 < p < 1$.

Proof. (a) The density of $N(\mu, \mu^2)$ is given by

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\mu^2}} \exp\left[-\frac{(x-\mu)^2}{2\mu^2}\right] = \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2} - \frac{x^2}{2\mu^2} + \frac{x}{\mu}\right),$$

where $\mu > 0$. We take

$$h(x) = e^{-1/2}, \quad a(\mu) = \frac{1}{\sqrt{2\pi\mu^2}}, \quad T_1(x) = x^2, \quad \text{and} \quad T_2(x) = x$$

with the natural parameters as $(\eta_1, \eta_2) = \left(-\frac{1}{2\mu^2}, \frac{1}{\mu}\right)$ in (1) so that they lie in the one-dimensional parabola $\left\{(\eta_1, \eta_2) \in \mathbb{R}^2 : \eta_1 = -\frac{1}{2}\eta_2^2\right\}$.

(b) The density of $\text{Gamma}(\alpha, \alpha)$ is given by

$$f(x|\alpha) = \frac{\alpha^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\alpha x} \cdot \mathbb{1}_{(0,\infty)}(x) = \frac{\alpha^\alpha}{\Gamma(\alpha)} \exp\left[(\alpha-1)\log x - \alpha x\right] \cdot \mathbb{1}_{(0,\infty)}(x),$$

where $\alpha > 0$. We take

$$h(x) = \mathbb{1}_{(0,\infty)}(x), \quad a(\alpha) = \frac{\alpha^\alpha}{\Gamma(\alpha)}, \quad T_1(x) = \log x, \quad \text{and} \quad T_2(x) = x$$

with the natural parameters as $(\eta_1, \eta_2) = (\alpha - 1, -\alpha)$ in (1) so that they lie in the one-dimensional straight line $\left\{(\eta_1, \eta_2) \in \mathbb{R}^2 : \eta_1 = -\eta_2 - 1\right\}$.

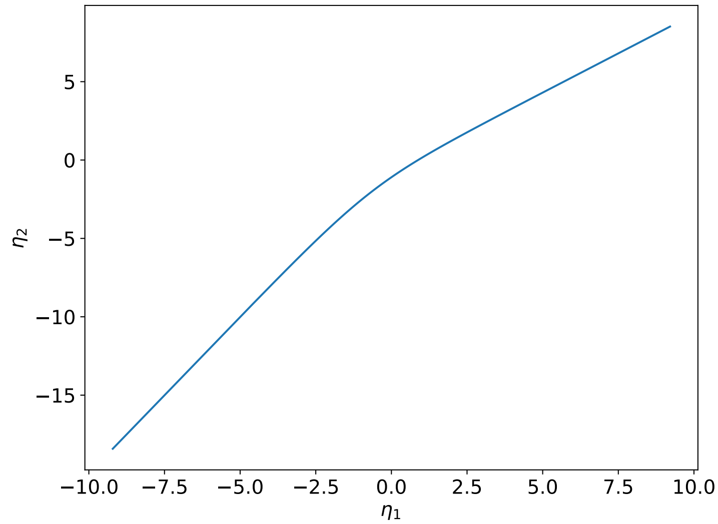

Figure 1: The plot of the parameterized curve $\left\{(\eta_1,\eta_2)=\left(\log\left(\frac{p}{1-p}\right),\log\left(\frac{p^2}{1-p^2}\right)\right)\in\mathbb{R}^2: 0<p<1\right\}$ on $\mathbb{R}^2$.

(c) The probability mass function of (X,Y) is given by

$$\begin{aligned}P(x,y|p) &= \binom{n}{x}p^x(1-p)^{n-x}\cdot\binom{m}{y}p^{2y}(1-p^2)^{m-y}\cdot\mathbb{1}_{\{0,1,\dots,n\}}(x)\cdot\mathbb{1}_{\{0,1,\dots,m\}}(y)\\ &= \binom{n}{x}\binom{m}{y}\cdot(1-p)^n(1-p^2)^m\exp\left[x\log\left(\frac{p}{1-p}\right)+y\log\left(\frac{p^2}{1-p^2}\right)\right]\cdot\mathbb{1}_{\{0,1,\dots,n\}}(x)\cdot\mathbb{1}_{\{0,1,\dots,m\}}(y),\end{aligned}$$

where $0<p<1$. We take

$$h(x)=\binom{n}{x}\binom{m}{y}\cdot\mathbb{1}_{\{0,1,\dots,n\}}(x)\cdot\mathbb{1}_{\{0,1,\dots,m\}}(y),\quad a(p)=(1-p)^n(1-p^2)^m,\quad T_1(x,y)=x,\quad\text{and}\quad T_2(x,y)=y$$

with the natural parameters as $(\eta_1,\eta_2)=\left(\log\left(\frac{p}{1-p}\right),\log\left(\frac{p^2}{1-p^2}\right)\right)$ in (1) so that they lie in the one-dimensional parameterized curve $\left\{(\eta_1,\eta_2)=\left(\log\left(\frac{p}{1-p}\right),\log\left(\frac{p^2}{1-p^2}\right)\right)\in\mathbb{R}^2: 0<p<1\right\}$; see Figure 1 for a graphical illustration. □

While the exponential families are relatively common in Statistics, they by no means cover all the possible distributions. Problem 3 below provides several counterexamples for exponential families.

Problem 3. *Show that each of the following families is not an exponential family:*

- *(a) The negative binomial distribution family* NegBinomial(r,p) *with both parameters r and $0<p<1$ unknown.*
- *(b) The set of probability density functions given by $f(x|\theta)=\frac{1}{\theta}\exp\left(1-\frac{x}{\theta}\right)$ with $0<\theta<x<\infty$.*

Proof. (a) Note that when both parameters r and p in NegBinomial(r,p) are also unknown, NegBinomial(r,p) no longer belongs to an exponential family, because the binomial coefficient $\binom{r+x-1}{x}$ cannot be factored into the products of $h(x)$ and $a(r, p)$ in which the function $h(x)$ does not depend on r, p while the function $a(r, p)$ is independent of the observation x.

(b) Notice that $f(x|\theta) = \frac{1}{\theta}\exp\left(1 - \frac{x}{\theta}\right) \cdot \mathbb{1}_{(\theta,\infty)}(x)$. One can argue that $f(x|\theta)$ does not belong to an exponential family because the indicator function $\mathbb{1}_{(\theta,\infty)}(x)$ cannot be factored into the products of $h(x)$ and $a(r, p)$ as in (1).
Or, it can be seen from (1) that for any $\boldsymbol{\theta} \in \Theta$ with $a(\boldsymbol{\theta}) > 0$, we must have $\{x \in \mathcal{X} : f(x|\boldsymbol{\theta}) > 0\} = \{x \in \mathcal{X} : h(x) > 0\}$ and this set does not depend on $\boldsymbol{\theta}$. However,

$$f(x|\theta) = \frac{1}{\theta}\exp\left(1 - \frac{x}{\theta}\right) \cdot \mathbb{1}_{(\theta,\infty)}(x) > 0 \iff x > \theta,$$

which depends on the parameter θ. Hence, $f(x|\theta)$ is not an exponential family distribution.
(Notes: the distribution $f(x|\theta)$ here is an example of the one-dimensional truncation family; see Example 11.14 in Perlman 2020.) □

In the following problem, we develop general formulae for calculating the mean and variance of a distribution in exponential families. While the applications of these formulae in Problem 4 are of pedagogical purpose, they are useful when we want to compute the means and variances in generalized linear models [Nelder and Wedderburn, 1972].

Problem 4 (Exercises 3.31 and 3.32 in Casella and Berger 2002). *We first assume that the PDF of a random variable X is given by the exponential family form (1) as:*

$$f(x|\boldsymbol{\theta}) = h(x)a(\boldsymbol{\theta})\exp\left(\sum_{i=1}^{k}\eta_i(\boldsymbol{\theta})T_i(x)\right).$$

The similar arguments below apply to the PMF case.

(a) Starting from the equality

$$\int f(x|\boldsymbol{\theta})\,dx = \int h(x)a(\boldsymbol{\theta})\exp\left(\sum_{i=1}^{k}\eta_i(\boldsymbol{\theta})T_i(x)\right)dx = 1,$$

differentiate both sides, and then rearrange terms to establish

$$\mathbb{E}\left[\sum_{i=1}^{k}\frac{\partial\eta_i(\boldsymbol{\theta})}{\partial\theta_j}\cdot T_i(X)\right] = -\frac{\partial}{\partial\theta_j}\log a(\boldsymbol{\theta}). \tag{3}$$

(b) Differentiate the above equality a second time; then rearrange to establish

$$\mathrm{Var}\left[\sum_{i=1}^{k}\frac{\partial\eta_i(\boldsymbol{\theta})}{\partial\theta_j}\cdot T_i(X)\right] = -\frac{\partial^2}{\partial\theta_j^2}\log a(\boldsymbol{\theta}) - \mathbb{E}\left[\sum_{i=1}^{k}\frac{\partial^2\eta_i(\boldsymbol{\theta})}{\partial\theta_j^2}\cdot T_i(X)\right]. \tag{4}$$

(c) Use (3) and (4) to derive the mean and variance of $X \sim \mathrm{Binomial}(n, p)$.

In what follows, we assume that Y has its density function as in (2):

$$f(y|\boldsymbol{\eta}) = h(y)a^*(\boldsymbol{\eta})\exp\left(\sum_{i=1}^{k}\eta_i T_i(y)\right).$$

(d) *Show that the identities (3) and (4) can be simplified to*

$$\mathbb{E}\left(T_j(Y)\right) = -\frac{\partial}{\partial \eta_j} \log a^*(\boldsymbol{\eta}) \quad \text{and} \quad \operatorname{Var}\left(T_j(Y)\right) = -\frac{\partial^2}{\partial \eta_j^2} \log a^*(\boldsymbol{\eta}). \tag{5}$$

(e) *Use this identity to calculate the mean and variance of a* Gamma(α, β) *random variable.*

Proof. (a) We assume that interchanging the order of differentiation and integration is valid.[4] Then,

$$\begin{aligned} \log f(x|\boldsymbol{\theta}) &= \log h(x) + \log a(\boldsymbol{\theta}) + \sum_{i=1}^k \eta_i(\boldsymbol{\theta}) T_i(x), \\ \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j} \log a(\boldsymbol{\theta}) + \sum_{i=1}^k \frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \theta_j} T_i(x). \end{aligned} \tag{6}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^k \frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \theta_j} T_i(X)\right] &= \mathbb{E}\left[\frac{\partial \log f(X|\boldsymbol{\theta})}{\partial \theta_j}\right] - \frac{\partial}{\partial \theta_j} \log a(\boldsymbol{\theta}) \\ &= \int \frac{\frac{\partial}{\partial \theta_j} f(x|\boldsymbol{\theta})}{f(x|\boldsymbol{\theta})} \cdot f(x|\boldsymbol{\theta})\, dx - \frac{\partial}{\partial \theta_j} \log a(\boldsymbol{\theta}) \\ &= \frac{\partial}{\partial \theta_j} \underbrace{\int f(x|\boldsymbol{\theta})\, dx}_{=1} - \frac{\partial}{\partial \theta_j} \log a(\boldsymbol{\theta}) \\ &= -\frac{\partial}{\partial \theta_j} \log a(\boldsymbol{\theta}). \end{aligned}$$

(b) From (6), we know that

$$\begin{aligned} \operatorname{Var}\left[\sum_{i=1}^k \frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \theta_j} T_i(X)\right] &= \operatorname{Var}\left[\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j}\right] \\ &= \mathbb{E}\left[\left(\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j}\right)^2\right] - \underbrace{\left(\mathbb{E}\left[\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j}\right]\right)^2}_{=0} \quad \text{(by our calculation in (a))} \\ &= \mathbb{E}\left[\left(\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j}\right)^2\right]. \end{aligned}$$

(Indeed, one can show that

$$\operatorname{Cov}\left[\sum_{i=1}^k \frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \theta_j} T_i(X), \sum_{i=1}^k \frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \theta_m} T_i(X)\right] = \mathbb{E}\left[\left(\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j}\right)\left(\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_m}\right)\right],$$

which is exactly the (i, m)-entry of the *Fisher Information Matrix*.) Now, notice the fact that

$$\frac{\partial^2}{\partial \theta_j^2} \log f(x|\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_j}\left(\frac{\frac{\partial f(x|\boldsymbol{\theta})}{\partial \theta_j}}{f(x|\boldsymbol{\theta})}\right) = \left(\frac{\frac{\partial^2 f(x|\boldsymbol{\theta})}{\partial \theta_j^2}}{f(x|\boldsymbol{\theta})}\right) - \left(\frac{\frac{\partial f(x|\boldsymbol{\theta})}{\partial \theta_j}}{f(x|\boldsymbol{\theta})}\right)^2$$

[4] This is indeed guaranteed by the special form of the exponential family; see Theorem 18.2 in https://www.stat.purdue.edu/~dasgupta/expfamily.pdf.

$$= \left(\frac{\frac{\partial^2 f(x|\boldsymbol{\theta})}{\partial \theta_j^2}}{f(x|\boldsymbol{\theta})}\right) - \left(\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j}\right)^2.$$

Taking the expectation on both sides yields that

$$\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^k \frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \theta_j} T_i(X)\right] = \mathbb{E}\left[\left(\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j}\right)^2\right] &= \mathbb{E}\left[\left(\frac{\frac{\partial^2 f(x|\boldsymbol{\theta})}{\partial \theta_j^2}}{f(x|\boldsymbol{\theta})}\right)\right] - \mathbb{E}\left[\frac{\partial^2}{\partial \theta_j^2} \log f(x|\boldsymbol{\theta})\right] \\
&= \int \left(\frac{\frac{\partial^2 f(x|\boldsymbol{\theta})}{\partial \theta_j^2}}{f(x|\boldsymbol{\theta})}\right) f(x|\boldsymbol{\theta})\, dx - \mathbb{E}\left[\frac{\partial^2}{\partial \theta_j^2} \log f(x|\boldsymbol{\theta})\right] \\
&= \int \frac{\partial^2 f(x|\boldsymbol{\theta})}{\partial \theta_j^2}\, dx - \mathbb{E}\left[\frac{\partial^2}{\partial \theta_j^2} \log f(x|\boldsymbol{\theta})\right] \\
&= \frac{\partial^2}{\partial \theta_j^2} \underbrace{\int f(x|\boldsymbol{\theta})\, dx}_{=1} - \mathbb{E}\left[\frac{\partial^2}{\partial \theta_j^2} \log f(x|\boldsymbol{\theta})\right] \\
&= -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_j^2} \log f(x|\boldsymbol{\theta})\right].
\end{aligned}$$

Finally, by the form of $f(x|\boldsymbol{\theta})$ in (1), we obtain that

$$\mathbb{E}\left[\frac{\partial^2}{\partial \theta_j^2} \log f(x|\boldsymbol{\theta})\right] = \frac{\partial^2}{\partial \theta_j^2} \log a(\boldsymbol{\theta}) + \mathbb{E}\left[\sum_{i=1}^k \frac{\partial^2 \eta_i(\boldsymbol{\theta})}{\partial \theta_j^2} \cdot T_i(X)\right].$$

The result thus follows.

(c) Recall from (f) of Problem 1 that $a(p) = (1-p)^n$, $T(X) = X$, and $\eta(p) = \log\left(\frac{p}{1-p}\right)$. By (3) and (4), we know that

$$\mathbb{E}\left[\frac{d\eta(p)}{dp} \cdot T(X)\right] = -\frac{d}{dp} \log a(p) \iff \mathbb{E}\left[\frac{X}{p(1-p)}\right] = \frac{n}{1-p},$$
$$\mathrm{Var}\left[\frac{d\eta(p)}{dp} \cdot T(X)\right] = -\frac{d^2}{dp^2} \log a(p) - \mathbb{E}\left[\frac{d^2\eta(p)}{dp^2} \cdot T(X)\right] \iff \mathrm{Var}\left[\frac{X}{p(1-p)}\right] = \frac{n}{(1-p)^2} - \mathbb{E}\left[\frac{(2p-1)X}{p^2(1-p)^2}\right].$$

Hence, $\mathbb{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$.

(d) The results follow easily from (3) and (4) by noting that

$$\frac{\partial \eta_i}{\partial \eta_j} = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Indeed, one can even show that $\mathrm{Cov}\left(T_i(X), T_j(X)\right) = -\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log a^*(\boldsymbol{\eta})$.

(e) Recall from (b) in Problem 1 that the density of Gamma(α, β) is given by

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \cdot \mathbb{1}_{(0,\infty)}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp\left[(\alpha-1)\log x - \beta x\right] \cdot \mathbb{1}_{(0,\infty)}(x),$$

where

$$h(x) = \mathbb{1}_{(0,\infty)}(x), \quad a(\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}, \quad T_1(x) = \log x, \quad \text{and} \quad T_2(x) = x$$

with the natural parameters as $(\eta_1, \eta_2) = (\eta_1(\alpha,\beta), \eta_2(\alpha,\beta)) = (\alpha - 1, -\beta)$. Hence,

$$\mathbb{E}(X) = \mathbb{E}\left[T_2(X)\right] = -\frac{\partial}{\partial \eta_2} \log\left(\frac{(-\eta_2)^{\eta_1+1}}{\Gamma(\eta_1+1)}\right) = -\frac{\eta_1+1}{\eta_2} = \frac{\alpha}{\beta}$$

and

$$\mathrm{Var}(X) = \mathrm{Var}\left[T_2(X)\right] = -\frac{\partial^2}{\partial \eta_2^2} \log\left(\frac{(-\eta_2)^{\eta_1+1}}{\Gamma(\eta_1+1)}\right) = \frac{\eta_1+1}{\eta_2^2} = \frac{\alpha}{\beta^2}.$$

□

Our calculation in (a) and (b) of Problem 4 will also appear when we derive the Fisher Information matrix and Cramér-Rao lower bound in STAT 513; see Section 13.1 in Perlman [2020] and 7.3.2 in Casella and Berger [2002]. In addition, the exponential families have many interesting properties and connections to other results in Statistics that are covered here. The interested readers can refer to `https://www.stat.purdue.edu/~dasgupta/expfamily.pdf` for further reading.

References

E. B. Andersen. Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, 65(331):1248–1255, 1970.

L. Brown. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. IMS Lecture Notes. Institute of Mathematical Statistics, 1986.

G. Casella and R. Berger. *Statistical Inference*. Duxbury advanced series. Thomson Learning, 2nd ed. edition, 2002.

E. L. Lehmann and H. Scheffé. Completeness, similar regions, and unbiased estimation: Part i. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 10(4):305–340, 1950.

J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

M. Perlman. Probability and Mathematical Statistics II (STAT 513 Lecture Notes), 2020. URL `https://sites.stat.washington.edu/people/mdperlma/STAT%20513%20MDP%20Notes.pdf`.