Exponential Families II

MIT 18.655

Dr. Kempthorne

Spring 2016
Outline

1. Exponential Families II
   - Random Vectors
   - Properties of Exponential Families
Random Vectors: Expectation and Variance

\( \mathbf{U} \) \((k \times 1)\) and \( \mathbf{V} \) \((l \times 1)\) are random vectors

- If \( \mathbf{A} \) \((m \times k)\), \( \mathbf{B} \) \((m \times l)\) are nonrandom, and then

\[
E(\mathbf{A} \mathbf{U} + \mathbf{B} \mathbf{V}) = \mathbf{A} E(\mathbf{U}) + \mathbf{B} E(\mathbf{V})
\]

- If \( \mathbf{U} = \mathbf{c} \) with probability 1 \( E(\mathbf{U}) = \mathbf{c} \).

- For a random vector \( \mathbf{U} \), if \( E(|\mathbf{U}|^2) = \sum_{i=1}^{k} E(U_i^2) < \infty \), define the variance of \( \mathbf{U} \) by

\[
\text{Var}(\mathbf{U}) = E \left[ (\mathbf{U} - E(\mathbf{U}))(\mathbf{U} - E(\mathbf{U}))^T \right]
= \left\| \text{Cov}(U_i, U_j) \right\| \quad (k \times k)
\]

- For \( \mathbf{A} \) \((m \times k)\) as above:

\[
\text{Var}(\mathbf{A} \mathbf{U}) = \mathbf{A} \text{Var}(\mathbf{U}) \mathbf{A}^T \quad (m \times m)
\]

- For \( \mathbf{c} \) \((k \times 1)\) a constant vector

\[
\text{Var}(\mathbf{U} + \mathbf{c}) = \text{Var}(\mathbf{U})
\]

- For \( \mathbf{a} \) \((k \times 1)\) a constant vector,

\[
\text{Var}(\mathbf{a}^T \mathbf{U}) = \text{Var}(\sum_{j=1}^{k} a_j U_j)
= \mathbf{a}^T \text{Var}(\mathbf{U}) \mathbf{a} = \sum_{i,j} a_i a_j \text{Cov}(U_i, U_j)
\]
Proposition B.5.1 If $E[|U|^2] < \infty$ then

$Var(U)$ is positive definite

if and only if

$P[\mathbf{a}^T \mathbf{U} + b = 0] < 1,$

for every $\mathbf{a} \neq 0$, and $b \in \mathbb{R}$.

Proof. $Var(U)$ is not positive definite iff $\mathbf{A}^T Var(\mathbf{Y})\mathbf{a} = 0$ for some $\mathbf{a} \neq 0$ which is equivalent to $Var(\mathbf{a}^T \mathbf{U}) = 0$. 
**Definition:** For random vectors \( \mathbf{U} \) \((k \times 1)\) and \( \mathbf{V} \) \((l \times 1)\) define the *Covariance* of \( \mathbf{U} \) \((k \times 1)\) and \( \mathbf{V} \) \((l \times 1)\) by

\[
\text{Cov}(\mathbf{U}, \mathbf{V}) = E[(\mathbf{U} - E(\mathbf{U}))(\mathbf{V} - E(\mathbf{V}))^T] \quad (k \times l)
\]

(must assume: \( E|\mathbf{U}|^2 < \infty \) and \( E|\mathbf{V}|^2 < \infty \))

- If \( \mathbf{U} \) and \( \mathbf{V} \) are independent

  \[
  \text{Cov}(\mathbf{U}, \mathbf{V}) = 0.
  \]

- For nonrandom \( A, a, B, b, \)

  \[
  \text{Cov}(A\mathbf{U} + a, B\mathbf{V} + b) = ACov(\mathbf{U}, \mathbf{V})B^T
  \]

- If \( \mathbf{U} \) and \( \mathbf{W} \) are random \((k \times 1)\) vectors, then

  \[
  \text{Var}(\mathbf{U} + \mathbf{W}) = \text{Var}(\mathbf{U}) + \text{Cov}(\mathbf{U}, \mathbf{W}) + \text{Cov}(\mathbf{W}, \mathbf{U}) + \text{Var}(\mathbf{W})
  \]

  and if \( \mathbf{U} \) and \( \mathbf{W} \) are independent

  \[
  \text{Var}(\mathbf{U} + \mathbf{W}) = \text{Var}(\mathbf{U}) + \text{Var}(\mathbf{W})
  \]
Moment-Generating Function of a Random Vector

Let $\mathbf{T} = (T_1, T_2, \ldots, T_k)^T$ be a $(k \times 1)$ random vector.

- For $\mathbf{s} = (s_1, s_2, \ldots, s_k)^T \in \mathbb{R}^k$, define
  $$M(\mathbf{s}) \equiv E[e^{\mathbf{s}^T \mathbf{T}}]$$
- $M(\mathbf{s})$ is the moment-generating function (mgf) of $\mathbf{T}$.
- The mgf may not exist for a given $\mathbf{T}$. If it does exist, it is defined for $\mathbf{s}$ in some ball centered at $\mathbf{s} = \mathbf{0}$.

Define the characteristic function (cf) of $\mathbf{T}$:
$$\phi(\mathbf{s}) = E[e^{i\mathbf{s}^T \mathbf{T}}] = E[\cos(\mathbf{s}^T \mathbf{T})] + iE[\sin(\mathbf{s}^T \mathbf{T})]$$
- The cf always exists.
Theorem B.5.1 Let $S = \{ s : M(s) < \infty \}$. Then

- $S$ is convex.
- If $S$ has a nonempty interior $S^0$, (contains a sphere $S(0, \epsilon), \epsilon > 0$), then $M$ is analytic on $S^0$.
- If $S^0 \neq \emptyset$, and $E[|T|^p] < \infty$ for all $p$, then if $i_1 + i_2 + \cdots + i_k = p$,

$$\left. \frac{\partial^p M(s)}{\partial s_1^{i_1} \cdots \partial s_k^{i_k}} \right|_{s=0} = E[T_1^{i_1} \cdots T_k^{i_k}]$$

$$\left| \left| \frac{\partial M}{\partial s_j} (s = 0) \right| \right| = \left| \left| E(T_j) \right| \right| = E[T]$$

$$\left| \left| \frac{\partial^2 M}{\partial s_i \partial s_j} (s = 0) \right| \right| = \left| \left| E(T_i T_j) \right| \right| = E[T T^T]$$

- If $S^0$ is nonempty, then $M(s)$ determines the distribution of $U$ uniquely.
Definition: The *Cumulant Generating Function* of the random vector $\mathbf{T}$ with mgf $M_{\mathbf{T}}(s)$ is

$$K(s) = K_{\mathbf{T}}(s) = \log M_{\mathbf{T}}(s).$$

$$c_{i_1, \ldots, i_k} = c_{i_1, \ldots, i_k}(\mathbf{T}) = \left. \frac{\partial^p}{\partial s_{1}^{i_1} \cdots \partial s_{k}^{i_k}} K(s) \right|_{s=0}$$

- In the bivariate case ($k = 2$) where
  $$\mu = E[\mathbf{T}], \text{ and } \tau_{i,j} = E[(T_1 - \mu_1)^i (T_2 - \mu_2)^j]$$

$$
\begin{align*}
c_{10} &= \mu_1 \\
c_{01} &= \mu_2 \\
c_{20} &= \tau_{2,0} = \text{var}(T_1) \\
c_{02} &= \tau_{0,2} = \text{var}(T_2) \\
c_{11} &= \tau_{1,1} = \text{cov}(T_1, T_2) \\
c_{30} &= \tau_{3,0} = E[(T_1 - \mu_1)^3] \\
c_{03} &= \tau_{0,3} = E[(T_2 - \mu_2)^3] \\
c_{40} &= \tau_{4,0} - 3\tau_{2,0}^2 \\
c_{04} &= \tau_{0,4} - 3\tau_{0,2}^2
\end{align*}
$$
If \( \mathbf{U} \) and \( \mathbf{V} \) are independent \((k \times 1)\) random vectors, then

\[
\begin{align*}
M_{U+V}(s) &= M_U(s) \times M_V(s) \\
K_{U+V}(s) &= K_U(s) + K_V(s)
\end{align*}
\]
Multivariate Normal Distributions

**Definition B.6.1:** A random vector $\mathbf{U} \ (k \times 1)$ has a $k$-variate normal distribution iff $\mathbf{U}$ can be written as

$$ \mathbf{U} = \mu + \mathbf{A}\mathbf{Z} $$

where $\mu$, $\mathbf{A}$ are constant and $\mathbf{Z} = (Z_1, \ldots, Z_k)^T$; $Z_i$ iid $N(0, 1)$.

**Definition B.6.2:** A random vector $\mathbf{U} \ (k \times 1)$ has a $k$-variate normal distribution iff for every $(k \times 1)$ nonrandom $\mathbf{a}$:

$$ \mathbf{a}^T\mathbf{U} = \sum_{i=1}^k a_i U_i \text{ has a univariate normal distribution} $$

The moment generating function of $\mathbf{U}$ is

$$ M_{\mathbf{U}}(s) = \exp \left\{ s^T \mu + \frac{1}{2} s^T \Sigma s \right\} $$

where $\mu = E[\mathbf{U}]$, and $\Sigma = \text{Cov}(\mathbf{U}) = \mathbf{A}\mathbf{A}^T$. 


Outline

1. Exponential Families II
   - Random Vectors
   - Properties of Exponential Families
Properties of Exponential Families
Properties of Exponential Families

**Theorem 1.6.3** Let \( \mathcal{P} \) be a canonical \( k \)-parameter exponential family generated by \((T, h)\), with corresponding natural parameter space \( \mathcal{E} \) and function \( A(\eta) \). Then

- \( \mathcal{E} \) is convex
- \( A : \mathcal{E} \to \mathbb{R} \) is convex
- If \( \mathcal{E} \) has nonempty interior \( \mathcal{E}^0 \subset \mathbb{R}^k \), and \( \eta_0 \in \mathcal{E}^0 \), then \( T(X) \) has under \( \eta_0 \) a mgf given by
  \[
  M(s) = \exp \{ A(\eta_0 + s) - A(\eta_0) \}
  \]
  valid for all \( s \) such that \( \eta_0 + s \in \mathcal{E} \).
  (this set of \( s \) includes a ball about \( \eta_0 \))

**Corollary 1.6.1** Under the conditions of the theorem

- \( E_{\eta_0}[T(X)] = \dot{A}(\eta_0) \)
- \( \text{Var}_{\eta_0}[T(X)] = \ddot{A}(\eta_0) \)

where \( \dot{A}(\eta_0) = \left| \frac{\partial A}{\partial \eta_j}(\eta_0) \right| \) and \( \ddot{A}(\eta_0) = \left| \left| \frac{\partial^2 A}{\partial \eta_i \partial \eta_j}(\eta_0) \right| \right| \)
Example: Multinomial Distribution

Multinomial Distribution

\[ X = (X_1, X_2, \ldots, X_q) \sim \text{Multinomial}(n, \theta = (\theta_1, \theta_2, \ldots, \theta_q)) \]

\[ p(x \mid \theta) = \frac{n}{x_1! \cdots x_q!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_q^{x_q} \]

where

- \( q \) is a given positive integer,
- \( \theta = (\theta_1, \ldots, \theta_q) : \sum_{j=1}^{q} \theta_j = 1 \).
- \( n \) is a given positive integer
- \( \sum_{i=1}^{q} X_i = n \).
Example: Multinomial Distribution

\[ p(x \mid \theta) = \frac{n}{x_1! \cdots x_q!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_q^{x_q} \]

\[ = \frac{n}{x_1! \cdots x_q!} \times \exp \{ \log(\theta_1)x_1 + \cdots + \log(\theta_{q-1})x_{q-1} \}
\]

\[ + \log(1 - \sum_1^{q-1} \theta_j)[n - \sum_1^{q-1} x_j] \}
\]

\[ = h(x) \exp \{ \sum_{j=1}^{q-1} \eta_j(\theta) T_j(x) - B(\theta) \}
\]

\[ = h(x) \exp \{ \sum_{j=1}^{q-1} \eta_j T_j(x) - A(\eta) \}
\]

where:

- \( h(x) = \frac{n}{x_1! \cdots x_q!} \)
- \( \eta(\theta) = (\eta_1(\theta), \eta_2(\theta), \ldots, \eta_{q-1}(\theta)) \)
  \[ \eta_j(\theta) = \log(\theta_j/(1 - \sum_1^{q-1} \theta_j)), j = 1, \ldots, q - 1 \]
- \( T(x) = (X_1, X_2, \ldots, X_{q-1}) = (T_1(x), T_2(x), \ldots, T_{q-1}(x)) \).
- \( B(\theta) = -n \log(1 - \sum_1^{q-1} \theta_j) \) and \( A(\eta) = +n \log(1 + \sum_{j=1}^{q-1} e^{\eta_j}) \)

\[ \dot{A}(\eta)_j = n \frac{e^{\eta_j}}{1 + \sum_{j=1}^{q-1} e^{\eta_j}} = n \frac{\theta_j/(1 - \sum_1^{q-1} \theta_k)}{1 + \sum_1^{q-1} \theta_k/(1 - \sum_1^{q-1} \theta_k)} = n \theta_j \]

\[ \ddot{A}(\eta)_{i,j} = -n \theta_i \theta_j, (i \neq j) \text{ and } \ddot{A}(\eta)_{i,i} = n \theta_i(1 - \theta_i), \]
Defining the Rank of an Exponential Family

- Every $k$-parameter exponential family is also a $k^*$-parameter exponential family for any $k^* > k$.
- The *minimal* value of $k$ defines the rank of the exponential family. Define *minimal* $k$ as the rank when the generating statistic $T(X)$ is $k$-dimensional, and the collection
  \[ \{1, T_1(X), T_2(X), \ldots, T_k(X)\} \]
  are linearly independent with positive probability, i.e.,
  \[
P[\sum_{j=1}^{k} a_j T_j(X) = a_{k+1} \mid \eta] < 1, \text{ unless all } a_j = 0.
\]

Note: the set of positive support on $\mathcal{X}$ does not depend on $\eta$. 
Theorem 1.6.4 Let \( P = \{ q(x \mid \eta), \eta \in \mathcal{E} \} \) be a canonical exponential family generated by \((T(X), h(X))\) with natural parameter space \( \mathcal{E} \) such that \( \mathcal{E} \) is open. Then the following statements are equivalent

- \( P \) is of rank \( k \).
- \( \eta \) is an identifiable parameter.
- \( \text{Var}(T \mid \eta) \) is positive definite
- \( \eta \rightarrow \dot{A}(\eta) \) is 1-to-1 on \( \mathcal{E} \).
- \( A(\eta) \) is strictly convex on \( \mathcal{E} \).

Note: \( \mathcal{E} \) open \( \implies \dot{A} \) defined on all \( \mathcal{E} \).

Corollary 1.6.2 If \( P \) is of rank \( k \) under Theorem 1.6.4, then

- \( P \) may be uniquely parametrized by
  \[ \mu(\eta) \equiv E[T(X) \mid \eta]. \]
- \( \log[q(x, \eta)] \) is strictly concave in \( \eta \) on \( \mathcal{E} \).
\section*{$p$-Variate Gaussian Family}

Let $\mathbf{Y}$ be a $(p \times 1)$ random vector with a $p$-variate Gaussian distribution

$$\mathbf{Y} \sim N_k(\mu, \Sigma)$$

where $\mu = E[\mathbf{Y}]$ and $\Sigma = \text{Var}(\mathbf{Y})$ is positive definite, rank $p$.

The density of $\mathbf{Y}$ is

$$p(\mathbf{y}, \mu, \Sigma) = |\text{det}(\Sigma)|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} \exp\{-\frac{1}{2}(\mathbf{y} - \mu)^T \Sigma^{-1} (\mathbf{y} - \mu)\}$$

Taking logs:

$$\log[p(\mathbf{y}, \mu, \Sigma)] = -\frac{1}{2} \mathbf{y}^T \Sigma^{-1} \mathbf{y} + [\Sigma^{-1} \mu]^T \mathbf{y}$$

$$-\frac{1}{2} \mu^T \Sigma^{-1} \mu - \frac{1}{2} \log[\text{det}(\Sigma)] - \frac{p}{2} \log[2\pi]$$

Defining $\Sigma^{-1} = ||\sigma_{ij}||$, we can write the first 2 terms as

$$-\left[\sum_{i<j} \sigma_{ij} Y_i Y_j + \frac{1}{2} \sum_i \sigma_{ii} Y_i^2 \right] + \sum_{i=1}^p \left[\sum_{j=1}^p \sigma_{ij} \mu_j \right] Y_i$$

- The parameter space dimension is
  $$k = p + p(p + 1)/2 = p(p + 3)/2$$

- The sufficient statistics are
  $$\{(Y_1, \ldots, Y_p), \{Y_i Y_j, 1 \leq i \leq j \leq p\}\}.$$
Exponential Families II
Properties of Exponential Families

- \( h(Y) \equiv 1 \)
- \( \theta = (\mu, \Sigma) \)
- \( B(\theta) = \frac{1}{2} \left( \log[|\det(\Sigma)|] + \mu^T \Sigma^{-1} \mu \right) \)

For a sample \( Y_1, \ldots, Y_n \) of iid \( N_p(\mu, \Sigma) \) r.vectors, the data \( X = (Y_1, Y_2, \ldots, Y_n) \) follows the \( k = \frac{p(p+3)}{2} \) parameter exponential family with

\[ T = (\sum_i Y_i, \text{LowerTriangle}(\sum_i Y_i Y_i^T)) \]

(LowerTriangle(·) refers to matrix elements along and below the diagonal)
Conjugate Families of Prior Distributions

Let $X_1, \ldots, X_n$ be a sample from the $k$-parameter exponential family

$$p(x \mid \theta) = \left[\prod_{i=1}^n h(x_i)\right] \exp\left\{\sum_{j=1}^k \eta_j(\theta) \sum_{i=1}^n T_j(x_i) - nB(\theta)\right\}$$

where $\theta$ is $k$-dimensional.

- Treat $\theta$ as the variable of interest in $p(x \mid \theta)$
- Treat $n$ and $T_j$ as parameters in $p(x \mid \theta)$
- Find a normalizing function:

$$\omega(t) = \int \cdots \int \exp \left\{ \sum_{j=1}^k t_j \eta_j(\theta) - t_{k+1} B(\theta) \right\} d\theta_1 \cdots d\theta_k$$

and set

$$\Omega = \{(t_1, \ldots, t_{k+1}) : 0 < \omega(t_1, \ldots, t_{k+1}) < \infty\}$$

**Proposition 1.6.1** The $(k + 1)$-parameter exponential family given by

$$\pi_t(\theta) = \exp \left\{ \sum_{j=1}^k t_j \eta_j(\theta) - t_{k+1} B(\theta) - \log[\omega(t)] \right\}$$

where $t = (t_1, \ldots, t_{k+1}) \in \Omega$, is a conjugate prior to $p(x \mid \theta)$. 