Part 2

Multiparameter Exponential Families

2.1 *Natural parameters, sufficient statistics, cgf* (pp 35–36) Natural parameter space A; carriers

2.2 *Expectation and covariance* (pp 36–38) $\mu = \dot{\psi}$, $V = \ddot{\psi}$; relationship of μ and η; expectation space B

2.3 *Review of transformations* (pp 38–38) Scalar function $h(\eta) = H(\mu)$; $H'(\mu) = D\dot{h}(\eta)$; $H''(\hat{\mu}) = D\ddot{h}(\hat{\eta})D'$

2.4 *Repeated sampling* (pp 38–39) $\bar{y} = \sum y_i/n$; $g_\eta^{(n)}(\bar{y}) = e^{n[\eta^\top \bar{y} - \psi(\eta)]} g_0^{(n)}(\bar{y})$; $A^{(n)} = nA$, $B^{(n)} = B$

2.5 *Likelihoods, score functions, Cramér–Rao lower bounds* (pp 39–41) Fisher information i_η; Fisher information for functions of η or μ; when can CRLB be attained?

2.6 *Maximum likelihood estimation* (pp 41–45) $\hat{\mu} = \bar{y}$; mapping to $\hat{\eta}$ one-parameter subfamilies; $\dddot{\psi}$ and $\ddddot{\psi}$; Stein's least favorable family

2.7 *Deviance* (pp 45–46) Hoeffding's formula; relationship with Fisher information

2.8 *Examples of multiparameter exponential families* (pp 46–55) Beta; Dirichlet; univariate normal; multivariate normal; graph models; truncated data; conditional families; 2×2 tables; Gamma/Dirichlet; Wishart; the Poisson trick; rotation data

2.9 *The multinomial as exponential family* (pp 55–60) Categorical data, $\pi_l = \Pr\{\text{category } l\}$; count vector s_l, proportion $p_l = s_l/n$; symmetric parameterization

2.1 Natural parameters, sufficient statistics, cgf

One-parameter families are too restrictive for most real data analysis problems. It is easy, however, to extend exponential families to multiparametric situations. A *p-parameter exponential family* is a collection of probability densities

$$\mathcal{G} = \{g_\eta(y), \eta \in A, y \in \mathcal{Y}\}$$

of the form

$$g_\eta(y) = e^{\eta^\top y - \psi(\eta)} g_0(y). \tag{2.1}$$

- η is the $p \times 1$ *natural*, or *canonical*, parameter vector.
- y is the $p \times 1$ vector of *sufficient statistics*, range space $y \in \mathcal{Y} \subset \mathcal{R}^p$.
- $g_0(y)$ is the *carrying density*, defined with respect to some *carrying measure* $m(dy)$ on $\mathcal{Y}$.
- A is the *natural parameter space*: all η having $\int_{\mathcal{Y}} e^{\eta^\top y} g_0(y) m(dy) < \infty$.
- $\psi(\eta)$ is the *normalizing constant* or *cumulant generating function*.

For any point η_0 in A we can express $\mathcal{G}$ as

$$g_\eta(y) = e^{(\eta-\eta_0)^\top y - [\psi(\eta)-\psi(\eta_0)]} g_{\eta_0}(y).$$

$\mathcal{G}$ consists of exponential tilts of g_{η_0}. The log tilting functions are linear in the p sufficient statistics $y = (y(1), y(2), \ldots, y(p))^\top$.

Homework 2.1. Show that $x_1, x_2, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \Gamma)$ can be written in form (2.1) with $y = (\bar{x}, \bar{x^2})$.

In most applications, y is a function of a complete data set $\boldsymbol{x}$, usually much more complicated than a p-vector. Then y earns the name "sufficient vector". The mapping $y = t(\boldsymbol{x})$ from the full data to the sufficient vector is crucial to the statistical analysis. It says which parts of the problem are important, and which can be ignored.

2.2 Expectation and covariance

The *expectation vector* $\mu = E_\eta\{y\}$ is given by

$$\mu = E_\eta\{y\} = \underset{p\times 1}{\dot{\psi}}(\eta) = \begin{pmatrix} \vdots \\ \partial\psi(\eta)/\partial\eta_i \\ \vdots \end{pmatrix}$$

while the *covariance matrix* equals the second derivative matrix of ψ,

$$V = \mathrm{Cov}_\eta\{y\} = \underset{p\times p}{\ddot{\psi}}(\eta) = \begin{pmatrix} \vdots \\ \partial^2\psi(\eta)/\partial^2\eta_i\partial\eta_j \\ \vdots \end{pmatrix}.$$

Both μ and V are functions of η, usually suppressed in our notation.

Homework 2.2. Verify the expressions for μ and V.

Relationship of μ and η

Notice that

$$\underset{p\times p}{d\mu/d\eta} = (\partial\mu_i/\partial\eta_j) = \ddot{\psi} = V$$

(where $(\partial\mu_i/\partial\eta_j)$ stands for the $p \times p$ matrix having ijth element $\partial\mu_i/\partial\eta_j$), so that the matrix

$$d\eta/d\mu = (\partial\eta_j/\partial\mu_i) = V^{-1}.$$

Here we are assuming that the carrying density $g_0(y)$ in (2.1) is of full rank in the sense that it is not entirely supported on any lower-dimensional subspace of $\mathcal{R}^p$, in which case V will be positive definite for all choices of η.

Homework 2.3. Give an argument verifying this statement.

The vector μ is a 1:1 function of η, usually nonlinear, the local equations of transformation being

$$d\mu = V\,d\eta \quad \text{and} \quad d\eta = V^{-1}d\mu$$

(remembering that V changes with η or μ). The set A of all η vectors for which $\int_{\mathcal{Y}} e^{\eta^\top y} g_0(y)$ is finite is convex, while the set B of all μ vectors is connected but not necessarily convex (though counterexamples are hard to construct),

$$B = \{\mu = E_\eta\{y\}, \eta \in A\}\,.$$

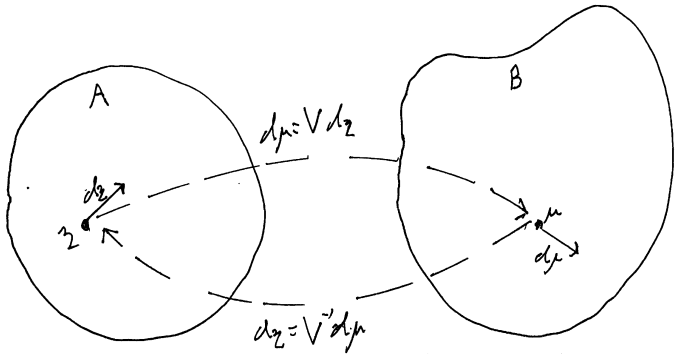


Figure 2.1

Note. $d\eta^\top d\mu = d\eta^\top V\,d\eta > 0$, so the angle between $d\eta$ and $d\mu$ is less than 90°; in this sense μ is an increasing function of η, and vice versa.

Homework 2.4. (a) Prove that A is convex. (b) If $\mathcal{Y}$ is the sample space of y, show that $B \subseteq$ convex hull of $\mathcal{Y}$. (c) Construct a one-parameter exponential family where the closure $\bar{B}$ is a proper subset of $\mathcal{Y}$.

REFERENCE Efron (1978), "Geometry of exponential families", *Ann. Statist.*, Section 2, provides an example of non-convex B.

2.3 Review of transformations

This review applies generally, though we'll be interested in (η, μ) as previously defined. Suppose η and μ are vectors in $\mathcal{R}^p$, smoothly related to each other,

$$\eta \overset{1:1}{\longleftrightarrow} \mu,$$

and that $h(\eta) = H(\mu)$ is some smooth real-valued function. Let D be the $p \times p$ derivative matrix (Hessian)

$$D = \begin{pmatrix} & \vdots & \\ \cdots & \partial\eta_j/\partial\mu_i & \cdots \\ & \vdots & \end{pmatrix}, \qquad i \downarrow \quad \text{and} \quad j \rightarrow .$$

($D = V^{-1}$ in our case.) Letting $\cdot$ indicate derivatives with respect to η, and $'$ indicate derivatives with respect to μ, we have

$$H'(\mu) = D\dot{h}(\eta)$$

where $\dot{h}(\eta) = (\cdots \partial h/\partial\eta_j \cdots)^\top$ and $H'(\mu) = (\cdots \partial H/\partial\mu_i \cdots)^\top$.

The second derivative matrices of $h(\eta)$ and $H(\mu)$ are related by

$$\begin{array}{ccccc} H''(\mu) & = & D\ddot{h}(\eta)D^\top & + & D_2\dot{h}(\eta) \\ \uparrow & & \uparrow & & \\ (\partial^2 H/\partial\mu_i\partial\mu_j) & & (\partial^2 h/\partial\eta_i\partial\eta_j) & & \end{array}$$

Here D_2 is the $p \times p \times p$ three-way array $(\partial^2\eta_k/\partial\mu_i\partial\mu_j)$.
Important: At a point where $\dot{h}(\eta) = 0$, $H''(\mu) = D\ddot{h}(\eta)D^\top$.

Homework 2.5. Prove the preceding important statement.

2.4 Repeated sampling

Suppose we observe repeated samples from (2.1),

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \overset{\text{iid}}{\sim} g_\eta(y).$$

Let $\bar{y}$ denote the average of the vectors y_i,

$$\bar{y} = \sum_{i=1}^{n} y_i/n.$$

Then

$$g_\eta^{\boldsymbol{Y}}(\boldsymbol{y}) = e^{n[\eta^\top \bar{y} - \psi(\eta)]} g_0^{\boldsymbol{Y}}(\boldsymbol{y}), \tag{2.2}$$

as in (1.2), with $g_0^{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^n g_0(y_i)$. This is a p-dimensional exponential family (2.1), with:

- natural parameter $\eta^{(n)} = n\eta$
- sufficient vector $y^{(n)} = \bar{y}$
- expectation vector $\mu^{(n)} = \mu$
- variance matrix $V^{(n)} = V/n$ [since $\mathrm{Cov}(\bar{y}) = \mathrm{Cov}(y)/n$]
- cgf $\psi^{(n)}(\eta^{(n)}) = n\psi(\eta^{(n)}/n)$
- sample space = product space $\mathcal{Y}_1 \otimes \mathcal{Y}_2 \otimes \cdots \otimes \mathcal{Y}_n$

Since $\bar{y}$ is sufficient, we can consider it on its own sample space, say $\mathcal{Y}^{(n)}$, with densities

$$g_\eta^{(n)}(\bar{y}) = e^{n[\eta^\top \bar{y} - \psi(\eta)]} g_0^{(n)}(\bar{y}), \tag{2.3}$$

where $g_0^{(n)}(\bar{y})$ is the density of $\bar{y}$ for $\eta = 0$.

From $\eta^{(n)} = n\eta$ and $\mu^{(n)} = \mu$ we see that

$$A^{(n)} = nA \quad \text{and} \quad B^{(n)} = B.$$

In what follows, we will parameterize family (2.3) with η rather than $\eta^{(n)} = n\eta$. Then we can use Figure 2.1 relating A and B *exactly as drawn*.

What *does* change is the distance of the sufficient statistic $\bar{y}$ from its expectation μ: since $V^{(n)} = V/n$, the "error" $\bar{y} - \mu$ decreases at order $O_p(1/\sqrt{n})$. As $n \to \infty$, $\bar{y} \to \mu$. This makes asymptotics easy to deal with in exponential families.

Homework 2.6. Is $d\mu^{(n)} = V^{(n)} d\eta^{(n)}$ the same as $d\mu = V d\eta$?

2.5 Likelihoods, score functions, Cramér–Rao lower bounds

From (2.2), the log likelihood function of $\boldsymbol{y} = (y_1, \ldots, y_n)$ is

$$l_\eta(\boldsymbol{y}) = n\left[\eta^\top \bar{y} - \psi(\eta)\right].$$

The *score function* is defined to be the component-wise derivative with respect to η,

$$\dot{l}_\eta(\boldsymbol{y}) = (\partial l_\eta(\boldsymbol{y})/\partial \eta_j) = n(\bar{y} - \mu)$$

(remembering that $\dot{\psi}(\eta) = \mu$), and the second derivative matrix is

$$\ddot{l}_\eta(\boldsymbol{y}) = -nV$$

(since $\dot{\mu} = \ddot{\psi} = V$). The *Fisher information* for η in $\boldsymbol{y}$ is the outer product

$$i_\eta^{(n)} = E_\eta \left\{ \dot{l}_\eta(\boldsymbol{y}) \dot{l}_\eta(\boldsymbol{y})^\top \right\} = nV,$$

(also equaling $E\{-\ddot{l}_\eta(\boldsymbol{y})\}$).

We can also consider the score function with respect to μ,

$$\begin{aligned} \frac{\partial l_\eta(\boldsymbol{y})}{\partial \mu} &= \frac{\partial \eta}{\partial \mu} \frac{\partial l_\eta(\boldsymbol{y})}{\partial \eta} = V^{-1} \dot{l}_\eta(\boldsymbol{y}) \\ &= nV^{-1}(\bar{y} - \mu); \end{aligned}$$

the Fisher information for μ, denoted $i_\eta^{(n)}(\mu)$ ("at η, for μ, sample size n") is

$$\begin{aligned} i_\eta^{(n)}(\mu) &= E \left\{ \frac{\partial l_\eta(\boldsymbol{y})}{\partial \mu} \frac{\partial l_\eta(\boldsymbol{y})^\top}{\partial \mu} \right\} = n^2 V^{-1} \operatorname{Cov}(\bar{y}) V^{-1} \\ &= nV^{-1}. \end{aligned}$$

Cramér–Rao lower bound (CRLB)

Suppose we have a multiparameter family $f_\alpha(z)$, score vector $\dot{l}_\alpha(z) = (\partial \log f_\alpha(z)/\partial \eta_j)$, and Fisher information i_α the expected outer product $E\{\dot{l}_\alpha(z)\dot{l}_\alpha(z)^\top\}$. Then the CRLB for α is i_α^{-1}: if $\bar{\alpha}$ is any unbiased estimate of vector α,

$$\operatorname{Cov}\{\bar{\alpha}\} \geq i_\alpha^{-1}.$$

(This is equivalent to $\operatorname{Var}\{c^\top \bar{\alpha}\} \geq c^\top i_\alpha^{-1} c$ for estimating any linear combination $c^\top \alpha$.)

The CRLB for μ is seen to be

$$\operatorname{CRLB}(\mu) = i_\eta^{(n)}(\mu)^{-1} = \frac{V}{n}.$$

But $\operatorname{Cov}(\bar{y}) = V/n$, so the MLE $\hat{\mu} = \bar{y}$ attains the CRLB. This only happens for μ, or linear transformations of μ. So for example, the MLE $\hat{\eta}$ does not attain

$$\operatorname{CRLB}(\eta) = \frac{V^{-1}}{n}.$$

Homework 2.7. Let ζ be a scalar function of η or μ, say

$$\zeta = t(\eta) = s(\mu),$$

with gradient vector $\dot{t}(\eta) = (\cdots \partial t/\partial \eta_j \cdots)$, and likewise $s'(\mu) = (\cdots \partial s/\partial \mu_j \cdots)$. Having observed $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, show that the lower bound on the variance of an unbiased estimate of ζ is

$$\operatorname{CRLB}(\zeta) = \frac{\dot{t}(\eta)^\top V^{-1} \dot{t}(\eta)}{n} = \frac{s'(\mu)^\top V s'(\mu)}{n}.$$

Note. In applications, $\hat{\eta}$ is substituted for η, or $\hat{\mu}$ for μ, to get an approximate variance for the MLE $\hat{\zeta} = t(\hat{\eta}) = s(\hat{\mu})$. Even though ζ is not generally unbiased for ζ, the variance approximation — which is equivalent to using the delta method — is usually pretty good.

2.6 Maximum likelihood estimation

From the score function for η, $\dot{l}_\eta(\boldsymbol{y}) = n(\bar{y} - \mu)$, we see that the maximum likelihood estimate $\hat{\eta}$ for η must satisfy

$$\mu_{\hat{\eta}} = \bar{y}.$$

That is, $\hat{\eta}$ is the value of η that makes the theoretical expectation μ equal the observed value $\bar{y}$. Moreover, $\ddot{l}_\eta(\boldsymbol{y}) = -nV$ shows that the log likelihood $l_\eta(\boldsymbol{y})$ is a *concave* function of η (since V is positive definite for all η) so there are no other local maxima.

From $\partial/\partial_\mu l_\eta(\boldsymbol{y}) = nV^{-1}(\bar{y} - \mu)$, we see that the MLE of μ is $\hat{\mu} = \bar{y}$. This also follows from $\hat{\mu} = \mu(\hat{\eta})$ (that is, that MLE's map correctly) and $\mu_{\hat{\eta}} = \bar{y}$.

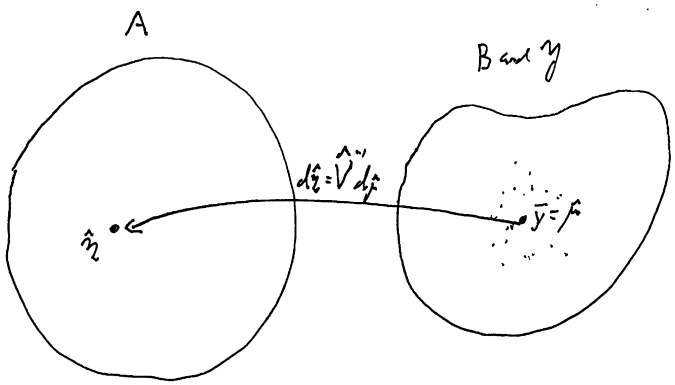


Figure 2.2

In Figure 2.2, the expectation space B and the sample space $\mathcal{Y}$ for individual y_i's are plotted over each other. The scattered dots represent the y_i's, with their average $\bar{y} = \hat{\mu}$ at the center. The nonlinear mapping $\hat{\eta} = \eta(\hat{\mu})$ has the local linear expression $d\hat{\eta} = \hat{V}^{-1} d\hat{\mu}$, where $\hat{V} = V_{\hat{\eta}}$, the variance matrix at the MLE point.

Homework 2.8. Use Figure 2.2 to get an approximation for $\mathrm{Cov}(\hat{\eta})$. How does it relate to $\mathrm{CRLB}(\hat{\eta})$?

Homework 2.9. Show that the $p \times p$ second derivative matrix with respect to μ is

$$\left(\frac{\partial^2 l_\eta(\boldsymbol{y})}{\partial \mu_i \partial \mu_j} \right) \Big|_{\hat{\mu}} = -n\hat{V}^{-1}.$$

From the central limit theorem,

$$\hat{\mu} = \bar{y} \doteq \mathcal{N}_p\left(\mu, {}^{V}/_{n}\right),$$

where the normality is approximate but the expectation and covariance are exact. Then $d\eta = V^{-1}d\mu$ suggests

$$\hat{\eta} \doteq \mathcal{N}_p\left(\eta, {}^{V^{-1}}/_{n}\right),$$

but now everything is approximate.

Table 2.1: Summary table

	η	μ
Score $\dot{l}$	$n(\bar{y}-\mu)$	$nV^{-1}(\bar{y}-\mu)$
Information $i^{(n)}$	nV	nV^{-1}
CRLB	V^{-1}/n	V/n
MLE	$\hat{\eta} = \dot{\psi}^{-1}(\bar{y})$	$\hat{\mu} = \bar{y}$
	$(E_\eta\{\overline{Y}\} = \bar{y})$	

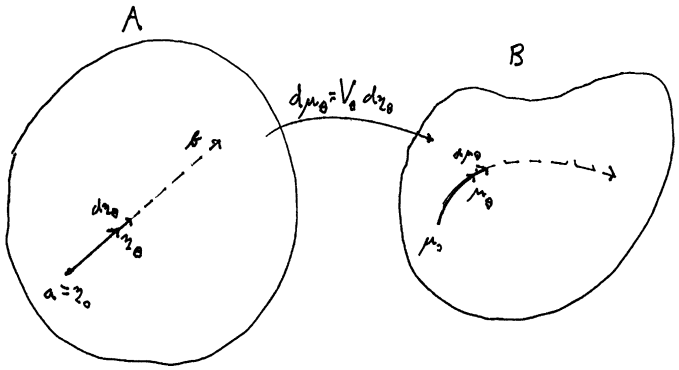


Figure 2.3

One-parameter subfamilies

Multiparameter exponential family calculations can sometimes be reduced to the one-parameter case by considering subfamilies of $g_\eta(y)$,

$$\eta_\theta = a + b\theta \qquad (\theta \in \mathcal{R}^1),$$

a and b fixed vectors in $\mathcal{R}^p$. This defines densities

$$f_\theta(y) = g_{\eta_\theta}(y) = e^{(a+b\theta)^\top y - \psi(a+b\theta)} g_0(y).$$

This is a one-parameter exponential family

$$f_\theta(y) = e^{\theta x - \phi(\theta)} f_0(y)$$

with

- natural parameter θ;
- sufficient statistic $x = b^\top y$;
- $\phi(\theta) = \psi(a + b\theta)$;
- $f_0(y) = e^{a^\top y} g_0(y)$.

For simple notation we write μ_θ for μ_{η_θ}, V_θ for V_{η_θ}, etc. As θ increases, η_θ moves in a straight line parallel to b, but μ_θ will usually follow a curve, the differential relationship being

$$d\mu_\theta = V_\theta d\eta_\theta.$$

Higher-order calculations are sometimes simplified by considering one-parameter subfamilies. For example we have

$$\begin{aligned} d\phi/d\theta &= E_\theta\{x\} = b^\top \mu_\theta = b^\top \dot\psi(\eta_\theta) \qquad \left[\dot\psi(\eta_\theta) \equiv \dot\psi(\eta)\Big|_{\eta=\eta_\theta}\right] \\ d^2\phi/d\theta^2 &= \mathrm{Var}_\theta\{x\} = b^\top V_\theta b = b^\top \ddot\psi(\eta_\theta) b \\ d^3\phi/d\theta^3 &= E_\theta\left\{x - E_\theta(x)\right\}^3 \\ &= \sum_i \sum_j \sum_k \dddot\psi_{ijk}(\eta_\theta) b_i b_j b_k \equiv \underset{p\times p\times p}{\dddot\psi(\eta_\theta)} \cdot b^3. \end{aligned}$$

This last implies that

$$\dddot\psi_{ijk} = E_\eta(y_i - \mu_i)(y_j - \mu_j)(y_k - \mu_k).$$

Homework 2.10. Verify the result above. What is $\ddddot\psi_{ijkl}$?

The subfamily $\mathcal{F} = \{f_\theta(y), \theta \in \Theta\}$ is defined for those values $\theta \in \Theta$ that keep η_θ within A, so

$$\mathcal{F} \subset \mathcal{G} = \{g_\eta(y), \eta \in A\}.$$

Suppose now that $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ is an iid sample from some member f_θ of $\mathcal{F}$.

Homework 2.11. (a) Show that the MLE $\hat\theta$ has $\mu_{\hat\theta}$ obtained by projection orthogonal to b, from $\bar y$ to $\{\mu_\theta, \theta \in \Theta\}$ as shown in Figure 2.4. (b) How would you estimate the standard deviation of $\hat\theta$?

Note. The set of $\bar y$ vectors having $\hat\theta$ as MLE is a $(p-1)$-dimensional hyperplane passing through $\mu_{\hat\theta}$ orthogonal to b. The hyperplanes are parallel for all values of $\hat\theta$.

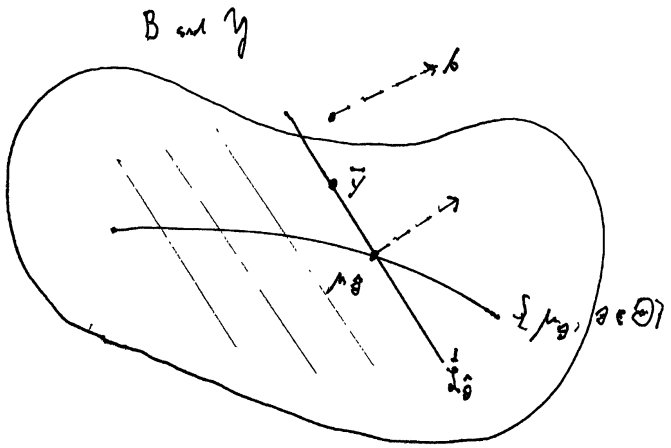


Figure 2.4

Stein's least favorable family (another use of one-parameter subfamilies)

In a p-parameter exponential family $\mathcal{G} = \{g_\eta(y), \eta \in A\}$, we wish to estimate a real-valued parameter ζ which can be evaluated as

$$\zeta = t(\eta) = s(\mu).$$

Let the true value of η be η_0 and $\mu_0 = \dot{\psi}(\eta_0)$ the true value of μ. Then the true value of ζ is $\zeta_0 = t(\eta_0) = s(\mu_0)$.

Gradients Let $\dot{t}_0$ be the gradient of $t(\eta)$ at $\eta = \eta_0$, and likewise s_0' for the gradient of $s(\mu)$ at $\mu = \mu_0$,

$$\dot{t}_0 = \dot{t}(\eta_0) = \begin{pmatrix} \vdots \\ \partial t(\eta)/\partial \eta_j \\ \vdots \end{pmatrix}_{\eta_0} \quad \text{and} \quad s_0' = s'(\mu_0) = \begin{pmatrix} \vdots \\ \partial s(\mu)/\partial \mu_j \\ \vdots \end{pmatrix}_{\mu_0},$$

both $\dot{t}_0$ and s_0' being p-vectors. As shown in Figure 2.5, $\dot{t}_0$ is orthogonal to the $(p-1)$-dimensional level surface of η vectors that give $t(\eta) = \zeta_0$, and s_0' is orthogonal to the level surface $s(\mu) = \zeta_0$.

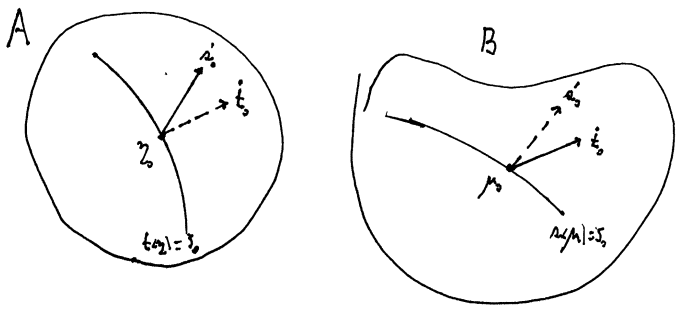


Figure 2.5

The *least favorable family* (LFF) is defined to be the one-parameter subfamily having

$$\eta_\theta = \eta_0 + s_0'\theta$$

(*not* $\eta_0 + \dot{t}_0\theta$) for θ in an open interval containing 0.

Homework 2.12. Show that the CRLB for estimating ζ in the LFF, evaluated at $\theta = 0$, is the same as the CRLB for estimating ζ in $\mathcal{G}$, evaluated at $\eta = \eta_0$; and that both equal $(s_0')^\top V_{\eta_0} s_0'$.

In other words, the reduction to the LFF does not make it any easier to estimate η. LFF operations are used in theoretical calculations where it is desired to extend a one-dimensional result to higher dimensions.

EXTRA CREDIT Any choice other than $b = s_0'$ in the family $\eta_\theta = \eta_0 + b\theta$ makes the one-parameter CRLB smaller.

2.7 Deviance

Everything said in Section 1.8 holds for deviance in multiparameter families:

$$\begin{aligned} D(\eta_1, \eta_2) &= 2E_{\eta_1}\left\{\log \frac{g_{\eta_1}(y)}{g_{\eta_2}(y)}\right\} \\ &= 2\left[(\eta_1 - \eta_2)^\top \mu_1 - \psi(\eta_1, \eta_2)\right] \geq 0. \end{aligned}$$

(Also denoted $D(\mu_1, \mu_2)$ or just $D(1, 2)$.) Under iid sampling, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$,

$$D_n(\eta_1, \eta_2) = nD(\eta_1, \eta_2).$$

Usually, $D(\eta_1, \eta_2) \neq D(\eta_2, \eta_1)$.

Hoeffding's formula

Now indexing with μ rather than η,

$$\frac{g_\mu(\boldsymbol{y})}{g_{\hat{\mu}}(\boldsymbol{y})} = e^{-nD(\hat{\mu},\mu)/2} = e^{-nD(\bar{y},\mu)/2}.$$

Relationship with Fisher information

$$D(\eta_1, \eta_2) = (\eta_2 - \eta_1)^\top V_{\eta_1}(\eta_2 - \eta_1) + O(\|\eta_2 - \eta_1\|^3)$$

(remembering that $i_{\eta_1}^{(1)} = V_{\eta_1}$).

The tangency picture on page 25 in Part 1 remains valid: now $\psi(\eta)$ is a convex surface over the convex set A in $\mathcal{R}^p$. The tangent line on page 25 is now the tangent hyperplane to the surface, passing through $(\eta_1, \psi(\eta_1))$ in $\mathcal{R}^{p+1}$; $D(\eta_1, \eta_2)$ is the distance from $(\eta_2, \psi(\eta_2))$ projected down to the tangent hyperplane.

Homework 2.13. Prove the previous statement.

EXTRA CREDIT Draw a schematic picture analogous to the illustration at the top of page 25.

Homework 2.14. Show that

$$(\eta_2 - \eta_1)^\top (\mu_2 - \mu_1) = {}^1\!/_2 \left[D(1,2) + D(2,1) \right].$$

Since the right-hand side is positive, this proves that the relationship between η and μ is "globally monotonic": the angle between $(\eta_2 - \eta_1)$ and $(\mu_2 - \mu_1)$ is always less than 90°.

2.8 Examples of multiparameter exponential families

There is a short list of named multiparameter exponential families that show up frequently in applications. Later we will consider some important examples that don't have familiar names.

Beta

- X is univariate with density on $[0,1]$,

$$\frac{x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}}{\mathrm{Be}(\alpha_1, \alpha_2)} \qquad \left[\mathrm{Be}(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)} \right],$$

  α_1 and α_2 positive. The density is defined with respect to Lebesgue measure on $[0,1]$. This is written in exponential family form as

$$g_\alpha(x) = e^{\alpha^\top y - \psi(\alpha)} g_0(x) \begin{cases} y = [\log(x), \log(1-x)]^\top \text{ and } \mathcal{Y} = \mathcal{R}^2 \\ \psi = \log\left[\mathrm{Be}(\alpha_1, \alpha_2)\right] \\ g_0(x) = 1/[x(1-x)]. \end{cases}$$

Homework 2.15. (a) Verify directly that x has

$$\text{mean } \frac{\alpha_1}{\alpha_1 + \alpha_2} \quad \text{and} \quad \text{variance } \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}.$$

(b) Find an expression for $E\{\log(x)\}$ in terms of the digamma function, digamma $(z) = \Gamma'(z)/\Gamma(z)$.

- Beta is *conjugate* to the binomial. Suppose

$$x \sim \mathrm{Bi}(n, \theta) \qquad \text{where } \theta \text{ has prior density } \mathrm{Be}(\alpha_1, \alpha_2).$$

Then the posterior density of θ given x is

$$\begin{aligned}\pi(\theta \mid x) &= c\pi(\theta)g_\theta(x) = c\theta^x(1-\theta)^{n-x}\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1} \\ &= c\theta^{x+\alpha_1-1}(1-\theta)^{n-x+\alpha_2-1} \sim \mathrm{Be}(x+\alpha_1, n-x+\alpha_2),\end{aligned}$$

with posterior expectation

$$E\{\theta \mid x\} = \frac{x+\alpha_1}{n+\alpha_2};$$

α_1 and α_2 act as prior observations of 1 and 0, respectively.

Example. $(\alpha_1, \alpha_2) = (2, 2)$, $n = 10$, $x = 8$, $n - x = 2$. Then $\hat\theta = x/n = 0.80$ but

$$E\{\theta \mid x\} = \frac{10}{14} = 0.71.$$

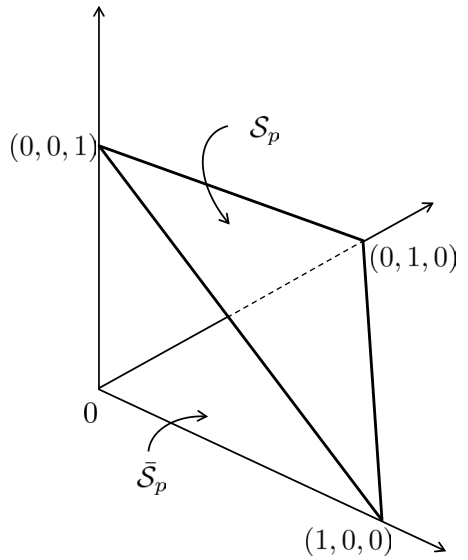
Dirichlet

- Let $\mathcal{S}_p$ indicate the p-dimensional simplex

$$\mathcal{S}_p = \left\{ x = (x_1, x_2, \ldots, x_p) : x_i \geq 0 \text{ and } \sum_1^p x_i = 1 \right\},$$

and $\bar{\mathcal{S}}_p$ the projected simplex

$$\bar{\mathcal{S}}_p = \left\{ x : x_p = 0, x_i \geq 0, \text{ and } \sum_1^{p-1} x_i \leq 1 \right\}.$$



- *An almost obvious fact*, $\bar{\mathcal{S}}_p$ has $(p-1)$-dimensional "area" of $1/(p-1)!$ (the case $p = 3$ *is* obvious).

- The Dirichlet is Beta's big brother for $p > 2$. Let

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p) \qquad \text{with all } \alpha_i > 0.$$

The corresponding Dirichlet density is

$$g_\alpha(x) = \prod_{i=1}^p x_i^{\alpha_i - 1} \Big/ \mathrm{Di}(\alpha) \qquad \left[\mathrm{Di}(\alpha) = \prod_1^p \Gamma(\alpha_i) \Big/ \Gamma\left(\sum \alpha_i\right) \right]$$

with respect to uniform measure on $\bar{\mathcal{S}}_p$, or uniform/$\sqrt{p}$ measure on $\mathcal{S}_p$. (Since $x_p = 1 - \sum_1^{p-1} x_i$, $g_\alpha(x)$ is actually $(p-1)$-dimensional; $\sqrt{p}$ is the area ratio of $\mathcal{S}_p$ to $\bar{\mathcal{S}}_p$.)

- $g_\alpha(x)$ can be written in exponential family form as

$$g_\alpha(x) = e^{\alpha^\top y - \psi(\alpha)} m(dx),$$

  where

  - α is the natural parameter vector ("η");
  - $y = \log(x) = [\log(x_1), \log(x_2), \ldots, \log(x_p)]$;
  - $\psi(\alpha) = \log[\mathrm{Di}(\alpha)]$; $m(dx) =$ uniform$/\sqrt{p}$ measure on $\mathcal{S}_p$.

Homework 2.16. What are μ and V for the Dirichlet? Compare with Homework 2.15. What is rank(V)?

Univariate normal

- $X \sim \mathcal{N}(\lambda, \Gamma)$, univariate, with

$$g(x) = \frac{1}{\sqrt{2\pi\Gamma}}\, e^{-\frac{1}{2\Gamma}(x-\lambda)^2} = \frac{1}{\sqrt{2\pi\Gamma}}\, e^{\left\{-\frac{x^2}{2\Gamma} + \frac{\lambda}{\Gamma}x - \frac{\lambda^2}{2\Gamma}\right\}}.$$

- In exponential family form,

$$g_\eta(x) = e^{\eta_1 y_1 + \eta_2 y_2 - \psi(\eta)} g_0(x),$$

  with

$$\eta = \begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = \begin{pmatrix}\lambda/\Gamma\\ -1/2\Gamma\end{pmatrix}, \qquad \mu = \begin{pmatrix}\lambda\\ \lambda^2+\Gamma\end{pmatrix};$$
$$y = \begin{pmatrix}y_1\\ y_2\end{pmatrix} = \begin{pmatrix}x\\ x^2\end{pmatrix}, \qquad \psi = \frac{1}{2}\left(\frac{\lambda^2}{\Gamma} + \log\Gamma\right);$$
$$g_0(x) = \frac{1}{\sqrt{2\pi}} \quad \text{with respect to uniform measure on } (-\infty, \infty).$$

Homework 2.17. Use $\dot\psi$ and $\ddot\psi$ to derive μ and V.

It seems like we have lost ground: our original univariate statistic x is now represented two-dimensionally by $y = (x, x^2)$. However, if we have an iid sample $x_1, x_2, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \Gamma)$, then

$$\bar y = \begin{pmatrix}\sum x_i/n\\ \sum x_i^2/n\end{pmatrix}$$

is still a two-dimensional sufficient statistic (though not the more usual form $(\bar x, \hat\sigma^2)$). Figure 2.6 presents diagrams of A and of B and $\mathcal{Y}$.

Homework 2.18. An iid sample $x_1, x_2, \ldots, x_n \sim \mathcal{N}(\lambda, \Gamma)$ gives $y_1, y_2, \ldots, y_n$, with $y_i = (x_i, x_i^2)$. Draw a schematic diagram of B and $\mathcal{Y}$ indicating the points y_i and the sufficient vector $\bar y$.

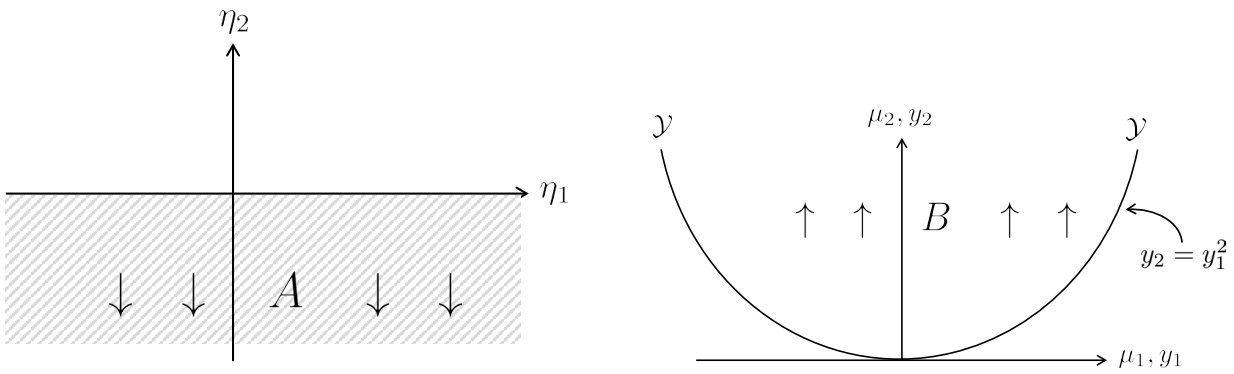


Figure 2.6

Ordinary least squares

- We observe the n-vector $\boldsymbol{x}$,

$$\boldsymbol{x} = \boldsymbol{M}\alpha + \boldsymbol{e} \qquad \left[\boldsymbol{e} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2, \boldsymbol{I})\right], \tag{2.4}$$

  that is, $e_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = 1, 2, \ldots, n$. Here $\boldsymbol{M}$ is a known $n \times p$ matrix, α an unknown p-dimensional parameter vector, and σ^2 an unknown positive scalar.

- The MLE of α is $\hat{\alpha} = (\boldsymbol{M}^\top \boldsymbol{M})^{-1}\boldsymbol{M}^\top \boldsymbol{x}$, while $\hat{\sigma}^2 = \|\boldsymbol{x} - \boldsymbol{M}\hat{\alpha}\|^2/n$ is the MLE (not the unbiased estimate) of σ^2.

Homework 2.19. Show that (2.4) is a $(p+1)$-parameter eponential family with

$$\eta = \begin{pmatrix} \alpha/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} (\boldsymbol{M}^\top \boldsymbol{M})^{-1}\hat{\alpha} \\ \|\boldsymbol{x}\|^2 \end{pmatrix}.$$

Hint: Let $\boldsymbol{r} = \boldsymbol{\Gamma}^\top \boldsymbol{x}$, where $\boldsymbol{\Gamma}$ is an $n \times (n-p)$ orthonormal matrix spanning the $(n-p)$-dimensional space orthogonal to the column space of $\boldsymbol{M}$, in which case

$$\boldsymbol{r} \sim \mathcal{N}_{n-p}(\boldsymbol{0}, \sigma^2 I) \text{ independent of } \hat{\alpha} \sim \mathcal{N}_p\left(\alpha, \sigma^2(\boldsymbol{M}^\top \boldsymbol{M})^{-1}\right).$$

Multivariate normal

- Another big brother case: we observe n independent observations from a d-dimensional normal distribution with expectation vector λ and $d \times d$ covariance matrix Γ,

$$x_1, x_2, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}_d(\lambda, \Gamma).$$

  (This is the generic beginning point for classical multivariate analysis.)

- We need some special notation: for H a $d\times d$ symmetric matrix, let $h = H^{(v)}$ be the $d(d+1)/2$ vector that strings out the on-or-above diagonal elements,

$$h = (H_{11}, H_{12}, \ldots, H_{1d}, H_{22}, H_{23}, \ldots, H_{2d}, H_{31}, \ldots, H_{dd})^\top,$$

  and let $h^{(m)}$ be the inverse mapping from h back to H.

- Also let $\mathrm{diag}(H)$ = matrix with on-diagonal elements those of H and off-diagonal elements 0.

- Straightforward but tedious (and easily bungled) calculations show that the density of $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ forms a

$$p = d(d+3)/2$$

  -dimensional exponential family

$$g_\eta(\boldsymbol{x}) = e^{n\left[\eta^\top y - \psi(\eta)\right]} g_0(\boldsymbol{x})$$

  described as follows:

$$y^\top = (y1^\top, y2^\top) \quad \text{where} \quad y1 - \bar{x} \quad \text{and} \quad y2 = \left(\sum_{i=1}^n x_i x_i^\top / n\right)^{(v)},$$

  dimensions d and $d(d+1)/2$, respectively;

$$\begin{aligned}
\eta^\top &= (\eta 1^\top, \eta 2^\top) \quad \text{where} \quad \eta 1 = n\Gamma^{-1}\lambda \quad \text{and} \quad \eta 2 = n\left[\mathrm{diag}(\Gamma^{-1})/2 - \Gamma^{-1}\right]^{(v)}; \\
\mu^\top &= (\mu 1^\top, \mu 2^\top) \quad \text{where} \quad \mu 1 = \lambda \quad \text{and} \quad \mu 2 = (\lambda\lambda^\top + \Gamma)^{(v)}; \\
\psi &= \frac{1}{2}\left[\lambda^\top \Gamma^{-1}\lambda + \log(\Gamma)\right].
\end{aligned}$$

- We also have

$$\begin{aligned}
\lambda = \mu 1 &= \frac{1}{n}\Gamma \cdot \eta 1; \\
\Gamma = \mu 2^{(m)} - \mu 1 \mu 1^\top &= -n\left[\mathrm{diag}(\eta 2^{(m)}) + \eta 2^{(m)}\right]^{-1}.
\end{aligned}$$

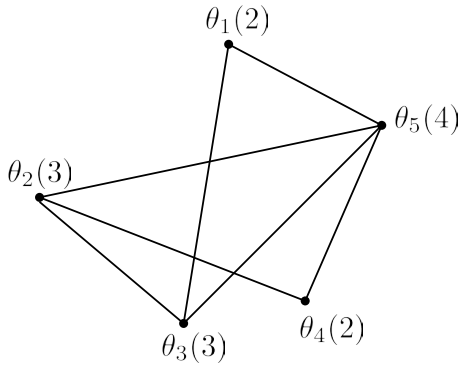
REFERENCE Efron and DiCiccio (1992), "More accurate confidence intervals in exponential families", *Biometrika*, Section 3.

Homework 2.20. Calculate the deviance $D[\mathcal{N}_p(\lambda_1, \Gamma_1), \mathcal{N}_p(\lambda_2, \Gamma_2)]$.

Graph models

Exponential families on graphs are now a booming topic in network theory. We have a graph with n nodes, each with its own known value θ_i. Let

$$x_{ij} = \begin{cases} 1 & \text{if an edge between } \theta_1 \text{ and } \theta_2, \\ 0 & \text{if no edge between } \theta_1 \text{ and } \theta_2. \end{cases}$$



A simple model assumes that the x_{ij} are independent Bernoulli variates, with probability $x_{ij} = 1$,

$$\pi_{ij} = \frac{e^{\theta_i+\theta_j}}{1+e^{\theta_i+\theta_j}}.$$

This is the "degree model".

Homework 2.21. (a) Show that the degree model is an n-parameter exponential family with sufficient vector $\boldsymbol{y}$,

$$y_i = \#\{\text{edges entering node } i\} \qquad (\text{"degree } i\text{"}).$$

(b) Describe η, μ, and V.

If all θ_i's are the same, the degree model reduces to a one-parameter exponential family, density $e^{\eta E-\psi(\eta)}$ with E the total number of edges. There are more complicated models, for instance

$$e^{\eta_1 E+\eta_2 T-\psi(\eta_1,\eta_2)},$$

where T is the total number of triangles. Very large graphs — the kinds of greatest interest these days — make it difficult to compute the normalizing factor $\psi(\eta_1, \eta_2)$, or to directly sample from the family–giving rise to Markov chain Monte Carlo (MCMC) techniques.

Truncated data

Suppose $y \sim g_\eta(y) = e^{\eta^\top y-\psi(\eta)} g_0(y)$, but we only get to observe y if it falls into a given subset $\mathcal{Y}_0$ of the sample space $\mathcal{Y}$. (This is the usual case in astronomy, where only sufficiently bright objects can be observed.) The conditional density of y given that it is observed is then

$$g_\eta(y \mid \mathcal{Y}_0) = e^{\eta^\top y-\psi(\eta)} g_0(y)/G_\eta(\mathcal{Y}_0)$$

where

$$G_\eta(\mathcal{Y}_0) = \int_{\mathcal{Y}_0} g_\eta(y) m(dy).$$

But this is still an exponential family,

$$g_\eta(y \mid \mathcal{Y}_0) = e^{\eta^\top y - \psi(\eta) - \log G_\eta(\mathcal{Y}_0)} g_0(y).$$

Homework 2.22. $x \sim \text{Poi}(\mu)$ but we only observe x if it is > 0. (a) Describe $g_\mu(x)$ in exponential family form. (b) Differentiate $\psi(\eta)$ to get μ and V.

Conditional families

REFERENCE Lehmann and Romano (2005), *Testing Statistical Hypotheses*, 3rd edition, Section 2.7.

We have a p-parameter exponential family,

$$\mathcal{G} = \left\{ g_\eta(y) = e^{\eta^\top y - \psi(\eta)} dG_0(y), \eta \in A \right\},$$

where now we have represented the carrying density $g_0(y)m(dy)$ in Stiljes form "$dG_0(y)$". We partition η and y into two parts, of dimensions p_1 and p_2,

$$\eta = (\eta_1, \eta_2) \quad \text{and} \quad y = (y_1, y_2).$$

Lemma 2. *The conditional distributions of y_1 given y_2 form a p_1-parameter exponential family, with densities*

$$g_{\eta_1}(y_1 \mid y_2) = e^{\eta_1^\top y_1 - \psi(\eta_1 | y_2)} dG_0(y_1 \mid y_2), \tag{2.5}$$

natural parameter vector η_1, sufficient statistic y_1, and cgf $\psi(\eta_1 \mid y_2)$ that depends on y_2 but not *on η_2. Here $dG_0(y_1 \mid y_2)$ represents the conditional distribution of y_1 given y_2 when $\eta_1 = 0$.*

Less usefully, the *marginal* distributions of y_2 form a p_2-parameter exponential family that depends on η_1 but not on y_1,

$$g^{Y_2}_{\eta_1,\eta_2}(y_2) = e^{\eta_2^\top y_2 - \psi_{\eta_1}(\eta_2)} g_{\eta_1,0}(y_2).$$

Homework 2.23. Verify Lemma 2. (The fact that the carrier is $dG_0(y_1 \mid y_2)$ follows from a general probabilistic result, and may be assumed.)

The familiar uses of Lemma 2 often involve transformation of the original family $\mathcal{G}$: for M a $p \times p$ nonsingular matrix, let

$$\tilde{\eta} = M^{-1}\eta \quad \text{and} \quad \tilde{y} = My.$$

Since $\eta^\top \tilde{y} = \eta^\top y_1$, we see that the transformed densities $\tilde{g}_{\tilde{\eta}}(\tilde{y})$ also form an exponential family

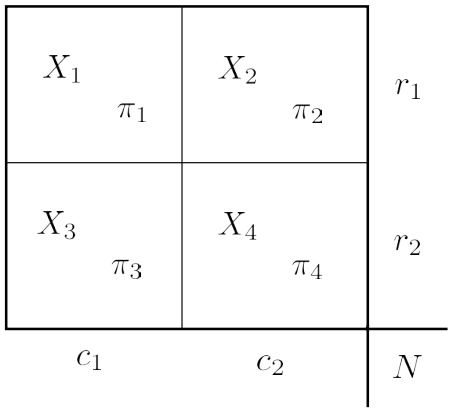
$$\tilde{\mathcal{G}} = \left\{ \tilde{g}_{\tilde{\eta}}(\tilde{y}) = e^{\tilde{\eta}^\top \tilde{y} - \psi(M\tilde{\eta})} d\tilde{G}_0(\tilde{y}), \tilde{\eta} \in M^{-1}A \right\},$$

to which Lemma 2 can be applied. What follows are four useful examples.

Example 1 (2×2 tables). We start with

$$(x_1, x_2, x_3, x_4) \sim \text{Mult}_4\,[N, (\pi_1, \pi_2, \pi_3, \pi_4)]$$

as in Section 1.4. The conditional distribution of $x_1 \mid x_2, x_3, x_4$ is identical to $x_1 \mid r_1, c_1, N$. According to Lemma 2, x_1 conditionally follows a one-parameter exponential family, which turns out to be the "tilted hypergeometric" family (1.3) (applying to Fisher's exact test). The reason why $\eta_1 = \log(\pi_1\pi_4/\pi_2\pi_3)$, the log odds ratio, is connected to the exponential family representation of the multinomial, as discussed in the next section.



Homework 2.24. What is the matrix M being used here?

Example 2 (Gamma/Dirichlet). We have p independent gamma variates of possibly different degrees of freedom,

$$s_i \overset{\text{ind}}{\sim} G_{\nu_i}, \qquad i = 1, 2, \ldots, p,$$

so $\boldsymbol{s} = (s_1, s_2, \ldots, s_p)^\top$ follows a p-parameter exponential family. Let M have first row $(1, 1, \ldots, 1)^\top$, so that $\tilde{s} = Ms$ has first element $\tilde{s}_1 = \sum_1^p s_i = s_+$. Define

$$z = \boldsymbol{s}/s_+,$$

z taking its values in the simplex $\mathcal{S}_p$. Then

$$z \mid s_+ \sim \text{Dirichlet}(\boldsymbol{\nu}) \qquad [\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_p)]\,,$$

a $(p-1)$-dimensional exponential family.

Example 3 (Wishart). Given a multivariate normal sample $x_1, x_2, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}_d(\lambda, \Gamma)$, let $y1 = \bar{x}$ and $y2 = \sum x_i x_i^\top/n$ as before. The conditional distribution of $y2 \mid y1$ is then a $p = d(d+1)/2$ exponential family. But the Wishart statistic

$$W = \sum (x_i - \bar{x})(x_i - \bar{x})^\top/n = \sum x_i x_i^\top/n - \bar{x}\bar{x}^\top$$

is, given $\bar{x}$, a function of $y2$. (In fact W is independent of $\bar{x}$.) This shows that W follows a $[d(d+1)/2]$-parameter exponential family.

Example 4 (The Poisson trick). Suppose that $\boldsymbol{s} = (s_1, s_2, \ldots, s_L)$ is a vector of L independent Poisson variates,

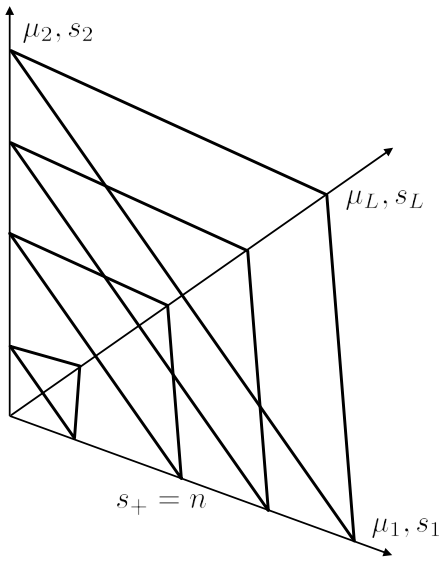
$$s_l \overset{\text{ind}}{\sim} \text{Poi}(\mu_l), \qquad l = 1, 2, \ldots, L.$$

Then the conditional distribution of $\boldsymbol{s} = (s_1, s_2, \ldots, s_L)$ — given that $\sum s_l$ equals n — is multinomial:

$$\boldsymbol{s} \mid n \sim \text{Mult}_L(n, \boldsymbol{\pi}), \tag{2.6}$$

the notation indicating an L category multinomial, n draws, true probabilities $\boldsymbol{\pi}$ where

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_L) \qquad \text{with } \pi_l = \mu_l \Big/ \sum_1^L \mu_i.$$



Homework 2.25. Directly verify (2.6).

Another way to say this: if

$$n \sim \text{Poi}(\mu_+) \quad \text{and} \quad \boldsymbol{s} \mid n \sim \text{Mult}_L(n, \boldsymbol{\pi}),$$

then

$$s_l \overset{\text{ind}}{\sim} \text{Poi}(\mu_+ \cdot \pi_l), \qquad l = 1, 2, \ldots, L. \tag{2.7}$$

If you are having trouble with a multinomial calculation, usually because of multinomial correlation among the s_l, then thinking of sample size n as Poisson instead of fixed makes the components independent. This "Poisson trick" is often used for quick multinomial approximations.

Homework 2.26. For the 2×2 table of Example 1, or of Section 1.4, use the Poisson trick to find the delta-method approximation of the standard error for the log odds ratio, $\log(\pi_1\pi_4/\pi_2\pi_3)$.

Rotation data

- Data consists of rotational speeds of 179 "early A" stars in the main sequence band. The dots in Figure 2.7 are binned counts, 45 bins having endpoints $0, 10, 20, \ldots, 450$. The counts are given as column N in Table 2.2.

- A theory based on statistical mechanics proposed that the rotational distributions should be mixtures of this density,

$$\text{func}_1(x) = 2xe^{-x^2}$$

(which is the square root of a standard e^{-x} exponential distribution).

- It was proposed that the observed data was a mixture of two such scaled densities,

$$f(x) = \frac{p_1 \,\mathrm{func}_1(x/c_1)}{c_1} + \frac{p_2 func_1(x/c_2)}{c_2} \qquad (p_2 = 1 - p_1).$$

- Usng the nonlinear maximizer `nlminb`, MLE values for $\theta = (p_1, c_1, c_2)$ were found: any trial choice of θ gave predicted bin counts $\widehat{N}$; `nlminb` minimized the sum of Poisson deviances between N and $\widehat{N}$. This was equivalent to maximizing the Poisson probabilities of the observed counts assuming that the Poisson parameters were proportional to $f(x)$. Columns 2 and 3 of Table 2.2 show the best fit parameters and their Poisson deviance residuals. The black curve in Figure 2.7 shows this best fit.

- The total deviance of the best fit was 52.8. This gives a chi-square p-value of

$$1 - \texttt{pchisq}(52.8, 42) = 0.123$$

  $(42 = 45 - 3$, three parameters having been fit).

- An alternate theory suggested using

$$\mathrm{func}_2(x) = 4\pi^{-1/2} x^2 e^{-x^2}$$

  in place of func_1 (which is the density of the square root of χ^2_3, called "maxwell" in the physics literature). The two "funcs" don't look all that different but func_2 didn't fit as well, as shown by the red curve in Figure 2.8, where total deviance was 78.3 and the p-value 0.0006. Columns 4 and 5 of Table 2.2 show these parameters and Poisson deviance residuals.

- I could have considered the data as a multinomial sample $\boldsymbol{p} \sim \mathrm{Mult}_{45}(179, \boldsymbol{\pi})$, with $\boldsymbol{p}$ the count proportions $p_1 = 16/179$, $p_2 = 14/179$, etc., and fit $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_2$ based on func_1 and func_2. This would give the same results as the Poisson fitting, as discussed in Part 3.

2.9 The multinomial as exponential family

Traditional multivariate analysis focused on the multivariate normal distribution. More recently there has been increased attention to categorical data and the multinomial distribution. The multinomial's exponential family structure is just a little tricky.

Table 2.2: The binned rotation data

bin	N	$N \times f_1$	devres_1	$N \times f_2$	devres_2
[1,]	16	13.01	.80	8.24	2.39
[2,]	14	17.72	−.92	23.46	−2.11
[3,]	7	6.94	.02	6.79	.08
[4,]	5	2.71	1.25	1.07	2.75
[5,]	3	2.63	.23	1.10	1.49
[6,]	5	3.11	.99	1.58	2.17
[7,]	1	3.58	−1.62	2.12	−.86
[8,]	6	4.01	.92	2.70	1.73
[9,]	7	4.40	1.14	3.29	1.77
[10,]	1	4.74	−2.09	3.89	−1.75
[11,]	2	5.03	−1.54	4.45	−1.31
[12,]	7	5.26	.72	4.98	.85
[13,]	5	5.44	−.19	5.45	−.20
[14,]	9	5.57	1.33	5.85	1.21
[15,]	9	5.65	1.30	6.17	1.06
[16,]	3	5.67	−1.23	6.41	−1.51
[17,]	3	5.65	−1.23	6.56	−1.56
[18,]	2	5.59	−1.75	6.62	−2.11
[19,]	8	5.48	1.00	6.60	.53
[20,]	2	5.34	−1.66	6.49	−2.07
[21,]	5	5.17	−.08	6.32	−.54
[22,]	6	4.97	.45	6.08	−.03
[23,]	5	4.75	.11	5.78	−.33
[24,]	3	4.51	−.76	5.45	−1.15
[25,]	6	4.26	.79	5.08	.40
[26,]	2	4.00	−1.11	4.69	−1.41
[27,]	1	3.73	−1.68	4.30	−1.92
[28,]	5	3.47	.77	3.89	.54
[29,]	1	3.20	−1.44	3.50	−1.58
[30,]	6	2.94	1.56	3.12	1.45
[31,]	5	2.68	1.26	2.75	1.22
[32,]	4	2.44	.91	2.41	.94
[33,]	3	2.20	.51	2.09	.59
[34,]	2	1.98	.01	1.80	.15
[35,]	3	1.77	.84	1.54	1.04
[36,]	2	1.58	.32	1.30	.57
[37,]	1	1.40	−.36	1.09	−.09
[38,]	0	1.23	−1.57	.91	−1.35
[39,]	1	1.08	−.08	.75	.27
[40,]	2	.95	.94	.62	1.39
[41,]	0	.82	−1.28	.50	−1.00
[42,]	0	.71	−1.19	.41	−.90
[43,]	0	.61	−1.11	.33	−.81
[44,]	0	.52	−1.02	.26	−.72
[45,]	1	.45	.71	.21	1.25
$\sum \text{devres}^2$			52.80		78.30

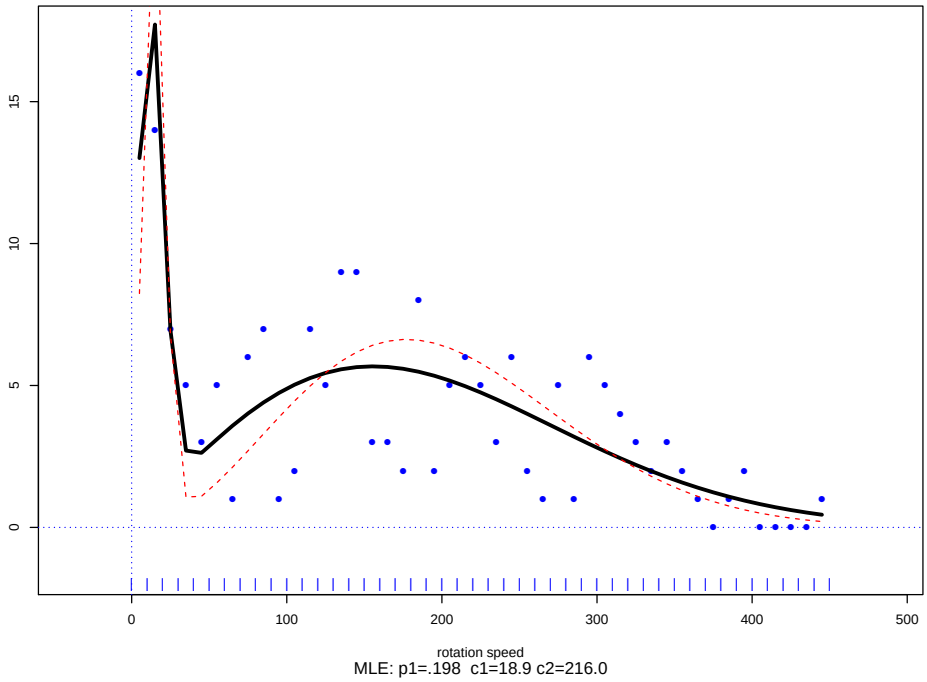


Figure 2.7: Rotation data. Binned counts (dots); func_1 fit (black), func_2 (red). Total deviation: func_1 52.8, χ^2_{42} p-value 0.123; func_2 78.3, p-value 0.0006.

We assume that n subjects have each been put into one of L categories. In the ulcer data example from Section 1.4, page 12, $n = 45$, $L = 4$, with categories *Treatment-Success*, *Treatment-Failure*, *Control-Success*, *Control-Failure*.

	Success	*Failure*	
Treatment	**9**	**12**	**21**
Control	**7**	**17**	**24**
	16	**29**	**45**

It is convenient to code the L categories with indicator vectors,

$$e_l = (0, 0, \ldots, 0, 1, 0, \ldots, 0)^\top \qquad (1 \text{ in the } l\text{th place}),$$

indicating category l. The data can be written as

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top,$$

where

$$y_i = e_{l_i}, \qquad l_i \text{ the } i\text{th subject's category.}$$

The multinomial probability model says that each subject's category is chosen independently

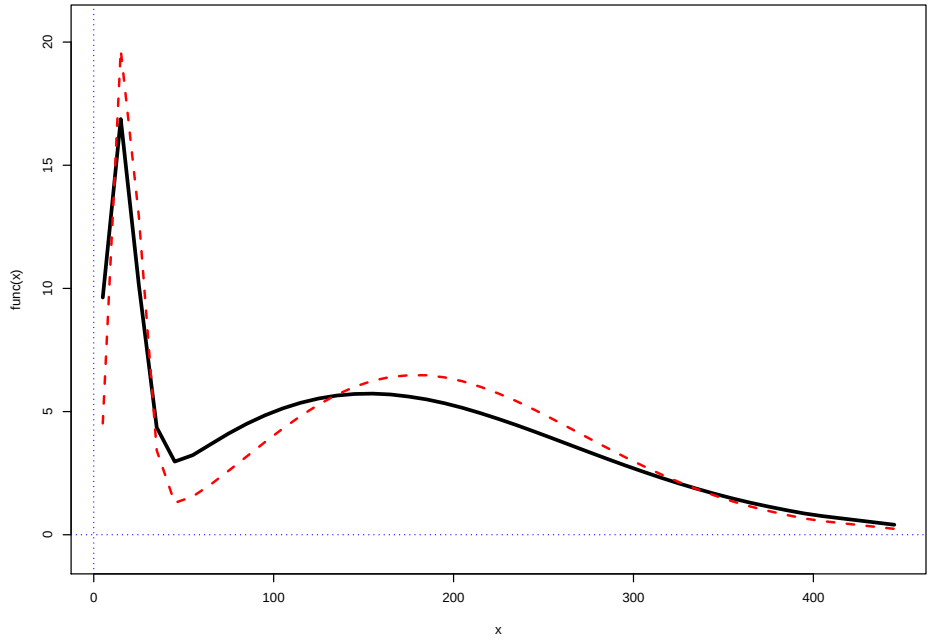


Figure 2.8: Rotation data: the two functions.

according to probability distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_L)^\top$,

$$y_i \sim \{e_l \text{ with probability } \pi_l,\ l = 1, 2, \ldots, L\}$$

independently for $i = 1, 2, \ldots, n$. The probability density function is

$$g_{\boldsymbol{\pi}}(\boldsymbol{y}) = \prod_{l=1}^{L} \pi_l^{s_l} = e^{\eta^\top s},$$

where $s_l = \#\{y_i = e_l\}$, the count for category l, and

$$\eta_l = \log \pi_l \qquad \left[\eta = (\eta_1, \eta_2, \ldots, \eta_L)^\top\right].$$

However, this isn't in exponential family form since $\eta = (\eta_1, \eta_2, \ldots, \eta_L)^\top$ is constrained to lie in a *nonlinear* subset of $\mathcal{R}^L$,

$$\sum_{l=1}^{L} e^{\eta_l} = \sum_{1}^{L} \pi_l = 1.$$

To circumvent this difficulty we let η be *any* vector in $\mathcal{R}^L$, and define

$$\pi_l = e^{\eta_l} \Big/ \sum_{j=1}^{L} e^{\eta_j} \qquad \text{for } l = 1, 2, \ldots, L, \tag{2.8}$$

so

$$\log \pi_l = \eta_l - \log \sum_{j=1}^{L} e^{\eta_j}.$$

Now the multinomial density can be written in genuine exponential family form,

$$g_\eta(\boldsymbol{y}) = e^{\eta^\top s - n\psi(\eta)} \qquad \left[\psi(\eta) = \log \sum_{l=1}^{L} e^{\eta_l}\right].$$

The count vector $s = (s_1, s_2, \ldots, s_L)$ is a sufficient statistic; the sample space $\mathcal{Y}$ for $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ has L^n points, all with $g_0(\boldsymbol{y}) = 1$.

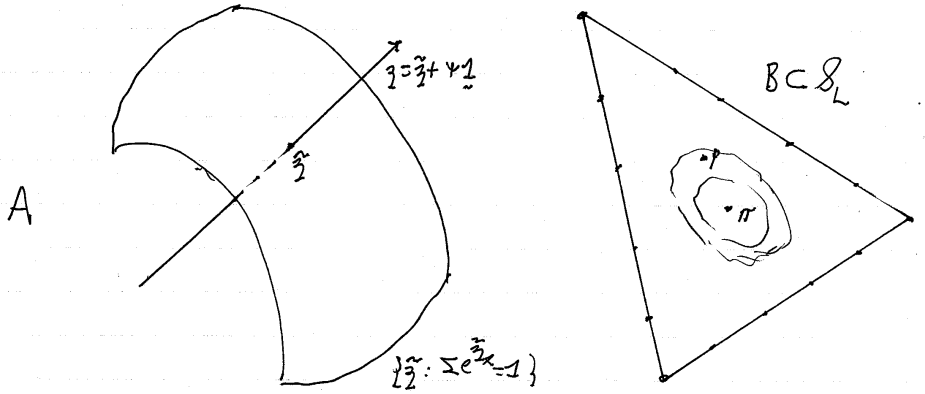


Figure 2.9

More familiarly, the multinomial is expressed in terms of the vector of observed proportions p,

$$p_l = \frac{s_l}{n} \qquad \left[p = (p_1, p_2, \ldots, p_n)^\top\right],$$

as

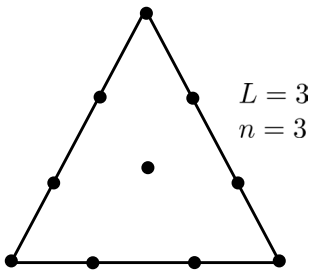
$$g_\eta(p) = e^{n\left[\eta^\top p - \psi(\eta)\right]} g_0(p). \tag{2.9}$$

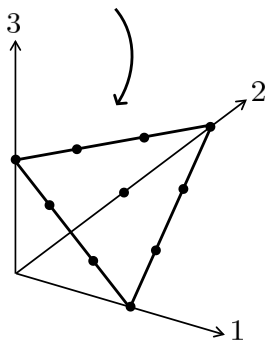
Now $g_0(p)$ is the multinomial coefficient

$$g_0(p) = \binom{n}{np_1, np_2, \ldots, np_L} = n! \Big/ \prod_{l=1}^{L} s_l!$$

defined with respect to counting measure on the lattice of points $(s_1, s_2, \ldots, s_L)^\top/n$, the s_l being non-negative integers summing to n. This will be denoted

$$p \sim \text{Mult}_L(n, \pi)/n. \tag{2.10}$$





To summarize, p is the vector of observed proportions, π the vector of category probabilities, parameterized in terms of the vector η at (2.8), and obeying density (2.9).

Homework 2.27. Re-express $g_\eta(p)$ so that $\eta = 0$ corresponds to $p \sim \text{Mult}(n, \pi^0)/n$, where $\pi^0 = (1, 1, \ldots, 1)^\top/L$ (the centerpoint of the simplex $\mathcal{S}$).

Homework 2.28. Why is η_1 the log odds ratio $\log(\pi_1\pi_4/\pi_2\pi_3)$ in the 2×2 case of Example 1 of Section 2.8? Hint: Homework 2.24.

Homework 2.29. Differentiate $\psi(\eta)$ to show that p has expectation vector π ("$= \mu$" in our original notation) and covariance matrix $V = D_\pi - \pi\pi^\top$, where $D = \text{diag}(\pi)$, the diagonal matrix with elements π_l, so that $V_{ij} = \pi_i(1 - \pi_i)$ if $i = j$, and $-\pi_i\pi_j$ otherwise.

As n gets large, p becomes approximately normal,

$$p \doteq \mathcal{N}_L(\pi, V),$$

but its L-dimensional distribution is confined to the $(L-1)$-dimensional flat set $\mathcal{S}_L$. The preceding diagram is a schematic depiction of the case $n = 3$ and $L = 3$.

In our parameterization, all η vectors on $\mathcal{R}^L$ of the form $\eta = \tilde{\eta} + \psi \cdot \mathbf{1}$, with ψ any number, $\mathbf{1} = (1, 1, \ldots, 1)^\top$, and $\sum e^{\eta_l} = 1$, map into the same expectation vector $\pi = (\cdots e^{\eta_l} \cdots)^\top$. As a result, the Hessian matrix $d\mu/d\eta = V$ is singular,

$$V \cdot \mathbf{1} = \pi - \pi = 0.$$

We can take the inverse matrix $d\eta/d\mu = V^{-1}$ to be any pseudoinverse of $D_\pi - \pi\pi^\top$, in particular

$$V^{-1} = \text{diag}(1/\pi),$$

the diagonal matrix with entries $1/\pi_l$; but we won't need to do so here. The multinomial can be expressed in standard form by taking $\eta = (\log \pi_1, \log \pi_2, \ldots, \log \pi_{L-1})^\top$ and $y = (p_1, p_2, \ldots, p_{L-1})^\top$, but this often causes problems because of asymmetry.

Homework 2.30. In what way is the Poisson trick related to V^{-1} above?

There is one more thing to say about the multinomial family: it represents *all* distributions supported on exactly L points, while all other exponential families are subfamilies of more general probability models.