Chapter 5. Maximum Likelihood Estimation

5.1 Definition

Let $\phi : R^k \to [0, \infty]$ be convex. Define $\ell : R^k \times R^k \to [-\infty, \infty]$ by

(1) $$\ell(\theta, x) = \ell_\phi(\theta, x) = \theta \cdot x - \phi(\theta) \quad .$$

For $S \subset N$ let

(2) $$\ell(S, x) = \sup \{\ell(\theta, x) : \theta \in S\}$$

and let

(3) $$\hat{\theta}_S(x) = \{\theta \in S : \ell(\theta, x) = \ell(S, x)\} \quad .$$

Note that according to this definition $\hat{\theta}_S$ is a subset of S. We will often abuse the notation slightly by letting $\hat{\theta}$ also denote an element of this set.

If $\phi = \psi$ is the cumulant generating function for an exponential family then

$$\ell_\psi(\theta, x) = \log p_\theta(x) \qquad \theta \in N$$

is the *log likelihood function* on N. (Of course, $\ell_\psi(\theta, x) \equiv -\infty$ for $\theta \notin N$ in accordance with the natural convention that $\psi(\theta) = \infty$ for $\theta \notin N$.) $\hat{\theta} \in \hat{\theta}_S(x)$ is then called a maximum *likelihood estimate* at x relative to $S \subset N$. A function $\delta : K \to \Theta$ for which $\delta(x) \in \hat{\theta}_\Theta(x)$ a.e.(ν) is called the (a) *maximum likelihood estimator*. This terminology is not always properly used in the literature; and we will also abuse it, at least to the extent of also referring to the set valued function $\hat{\theta}$ (·) as the maximum likelihood estimator.

5.2 Assumptions.

The main results of this section concern the existence and construction of maximum likelihood estimators, $\hat{\theta}$. The proofs of these results are based on the fact that ψ is a convex function satisfying certain additional properties, and not otherwise on the fact that ψ is a cumulant generating function. In Chapter 6 we will want to apply these same existence and construction results to convex functions, ϕ, which are not cumulant generating functions. To prepare for this application we now make explicit the conditions on ϕ which are needed in the proofs of the main results of this section.

Let $\phi : R^k \to (-\infty, \infty]$ be a lower semicontinuous convex function. Let $N = N_\phi = \{\theta : \phi(\theta) < \infty\}$. Such a function is called *regularly strictly convex* if it is strictly convex and twice differentiable on N_ϕ°, and

(1) $$D_2\phi \quad \text{is positive definite on} \quad N_\phi^\circ \quad .$$

In the following results we will assume ϕ is regularly strictly convex. In some of the following we also assume ϕ is steep. Note that if ψ is the cumulant generating function of a steep exponential family then it satisfies these assumptions.

Here are some useful facts.

Let $\ell = \ell_\phi$ be defined by 5.1(1), and let the mapping $\xi : N_\phi \to R^k$, be defined by $\xi(\theta) = \nabla\phi(\theta)$. Then, ξ is continuous and 1 - 1 since ϕ is strictly convex. (1) says that the Hessian of $\xi = \nabla\phi$ is positive definite. Hence $\xi(N^\circ)$ is an open set; call it R, or R_ϕ. $\xi^{-1}(\cdot)$ is continuous on R.

Theorem 3.6 establishes that

(2) $$R = K^\circ$$

when $\phi = \psi$ is the cumulant generating function of a minimal steep exponential family. In particular, in this case

(3) $$R \text{ is convex} \quad .$$

It will be shown in Proposition 6.7 that (3) is always valid under the above general assumptions on ϕ including steepness of ϕ.

As previously, let $\theta(\cdot) = \xi^{-1}(\cdot)$, i.e. $\xi(\theta(x)) = x$.

(The assumption above of the existence of second derivatives and of (1) is convenient, but can be dispensed with. The other assumptions are required for the following development.)

We emphasize again: the following results about ℓ_ϕ and maximum likelihood estimation concern the general situation where ϕ is as assumed above. These results therefore apply in particular to *maximum likelihood estimation from minimal steep standard exponential families*.

5.3 Lemma

Assume ϕ is regularly strictly convex. Then, $\ell(\cdot, x)$ is concave and upper semicontinuous on R^k for all $x \in R^k$. It is strictly concave on N .

If $\theta_0 \in N^\circ$ then

$$\nabla \ell(\cdot, x)_{|\theta_0} = x - \xi(\theta_0) \tag{1}$$

$$D_2 \ell(\cdot, x)_{|\theta_0} = -D_2\phi(\theta_0) = -\Sigma(\theta_0) \tag{2}$$

where $(\Sigma(\theta_0))_{ij} \quad \frac{\partial^2}{\partial\theta_i \partial\theta_j} \phi(\theta_0)$ is positive definite. If $x \in R$ (= K°) then

$$\lim_{||\theta||\to\infty} \ell(\theta, x) = -\infty \quad . \tag{3}$$

Proof. The first assertions are immediate from Assumption 5.2. Equations (1) and (2) are a direct calculation. The positive definiteness of $\Sigma(\theta_0)$ is a consequence of 5.2(1).

Assertion (3) has been proved in 3.6(4) for the case where $\phi = \psi$ is the cumulant generating function of a minimal steep exponential family. This proof was needed in order to show that $R = K^\circ$ in such a situation. However we now want a proof valid for arbitrary convex functions, ϕ, satisfying

5.2(1). This is easily supplied.

Assume $x \in R$, then $\theta(x) \in N^\circ$. Note using (1), (2) that $\nabla \ell(\theta(x), x) = 0$, and $D_2\ell(\theta(x), x)$ is negative definite. Hence for some $\delta > 0$, $\varepsilon > 0$

$$(4) \qquad \ell(\theta, x) = \ell(\theta(x), x) - (\theta - \theta(x))' \not{\Sigma}(\theta - \theta(x))/2 + o(||\theta - \theta(x)||^2)$$

$$< \ell(\theta(x), x) - \varepsilon \qquad \text{for} \qquad ||\theta - \theta(x)|| = \delta \quad .$$

It follows that when $||\theta - \theta(x)|| > \delta$

$$(5) \qquad \ell(\theta, x) \leq \ell(\theta(x), x) - \frac{||\theta - \theta_0||}{\delta} \varepsilon$$

by (4) since

$$\ell(\theta(x) + (\delta/(||\theta - \theta(x)||))(\theta - \theta(x))) \leq (1 - \delta/||\theta - \theta(x)||)\ell(\theta(x), x)$$

$$+ (\delta/||\theta - \theta(x)||)\ell(\theta, x)$$

by convexity. (5) implies (3). ||

(We note that the positive definiteness of $\not{\Sigma}$ is not really needed to establish (3). It is only necessary that the conclusion of (4) be valid -- i.e. for some $\delta > 0$, $\varepsilon > 0$

$$(4') \qquad \ell(\theta, x) < \ell(\theta(x), x) - \varepsilon \qquad \text{for} \qquad ||\theta - \theta(x)|| = \delta \quad .$$

This condition follows whenever $\ell(\cdot, x)$ is a strictly concave function which assumes its maximum at $\theta(x)$.)

It is useful to now prove the following lemma. This result is used in Theorem 5.5 to show that $\hat{\theta}_\Theta \subset N^\circ$ when Θ is convex.

5.4 Lemma

Assume ϕ is steep and regularly strictly convex. Let $\theta_1 \in N - N^\circ$, $\theta_0 \in N^\circ$. Let $\theta_\rho = \theta_0 + \rho(\theta_1 - \theta_0)$, $0 < \rho < 1$. Then

(1) $$\lim_{\rho \uparrow 1} \left(\frac{\partial}{\partial\rho} \ell(\theta_\rho, x)\right) = -\infty \quad .$$

Hence there is a $\rho' < 1$ such that

(2) $$\ell(\theta_{\rho'}, x) > \ell(\theta_1, x) \quad .$$

Proof. From 5.3(1)

$$\frac{\partial}{\partial\rho} \ell(\theta_\rho, x) = (\theta_0 - \theta_1) \cdot (x - \xi(\theta_\rho)) \to -\infty$$

as $\rho \uparrow 1$ because ψ is steep. This proves (1) from which (2) is immediate. (In case ψ is regular, i.e. $N = N^\circ$, then $\lim_{\rho\uparrow 1} \ell(\theta_\rho, x) = -\infty$ by upper semi-continuity, which can also be used to prove (2).) ||

FULL FAMILIES

Here is a fundamental result concerning maximum likelihood estimation. It follows easily from the above.

5.5 Theorem

Let ϕ be steep and regularly strictly convex. If $x \in R$ then

(1) $$\hat{\theta}_N(x) = \{\theta(x)\} \subset N^\circ \quad .$$

In other words, $\hat{\theta}_N(x)$ consists of the unique point $\hat{\theta} = \theta(x)$ satisfying

(1') $$\xi(\hat{\theta}) = x \in R \quad .$$

If $x \notin R$ then $\hat{\theta}_N(x)$ is empty. (Recall that if $\phi = \psi$ is the cumulant generating function of a steep canonical exponential family then $R = K^\circ$.)

Proof. For any x, $\{\hat{\theta}_N(x)\} \subset N^\circ$ by virtue of Lemma 5.4. Any maximum likelihood estimator must thus be a local maxima of $\ell(\cdot, x)$ and hence must satisfy

$$\nabla \ell(\cdot, x)_{|\hat{\theta}} = 0 \quad .$$

This implies (1') by 5.3(1). Furthermore, the solution to (1') is unique if it exists, and it exists if and only if $x \in R = \xi(N^\circ)$. ||

Remarks. Maximum likelihood estimation is defined in statistical theory for a general parametric family of densities $\{f_\theta : \theta \in \Theta\}$ by $\hat{\theta}(x) = \{\theta \in \Theta : f_\theta(x) = \sup_\alpha f_\alpha(x)\}$. Note that this definition is invariant under reparametrization. Thus, if $\xi = \xi(\theta)$ is a 1 - 1 map on Θ the maximum likelihood estimate of the parameter $\xi \in \xi(\Theta)$ is $\xi(\hat{\theta})$.

Accordingly, Theorem 5.5 says that for minimal steep exponential families $x = \xi(\theta(x))$ is the unique maximum likelihood estimator of the mean value parameter $\xi = \xi(\theta)$ at $x \in K^\circ$. To emphasize, in terms of the mean value parametrization the maximum likelihood estimator is determined by the trivial equation

$$(1'') \qquad \hat{\xi}(x) = x , \qquad x \in K^\circ \quad .$$

For the present, (1") is valid if and only if $x \in K^\circ$. This set of course contains almost every $x(\nu)$ if and only if

$$(2) \qquad \nu(K - K^\circ) = 0 \quad .$$

Note that (2) is satisfied if ν is absolutely continuous with respect to Lebesgue measure. It is never satisfied if ν has finite support or, more generally, has countable support and $K \neq R^k$. In the last part of Chapter 6 we expand such exponential families so that (1") usually remains valid for a.e.x (ν).

(Since $\xi = E_\theta(x)$ equation (1") also defines $\hat{\xi}(x) = x$ as the classical method-of-moments estimator. Thus for the mean value parametrization the maximum likelihood and method-of-moments estimators agree.)

Suppose that $X_1,\ldots,X_n$ are independent identically distributed random variables from the exponential family $\{p_\theta\}$. Then, as noted in 1.11(2), the distributions of the sufficient statistic $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ also form an exponential family with natural parameter $\alpha = n\theta$ and cumulant generating function $n\psi(\alpha/n)$. It follows that $\alpha(x) = n\theta(x)$. So, the maximum likelihood estimator of α based on $\bar{X}_n$ is $n\theta(\bar{X}_n)$ and the maximum likelihood estimator $\hat{\theta}_{(n)}$ of $\theta = \alpha/n$ based on $\bar{X}_n$ is

(3) $$\hat{\theta}_{(n)} = \hat{\alpha}/n = \theta(\bar{x}_n) \quad .$$

5.6 Examples (Beta Distribution)

For a variety of common full families the above remarks lead to easy calculation of the maximum likelihood estimator. These are situations such as those mentioned in 3.8 where the mean value parametrization has a convenient form. For example if $Y_1, Y_2, \ldots, Y_n$ are i.i.d. multivariate normal $(\mu, \mathcal{Z})$ random variables then the maximum likelihood estimators for μ and $\mu\mu' + \mathcal{Z}$ are, respectively, $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$ and $n^{-1} \sum_{i=1}^{n} Y_i Y_i'$. This leads to the conventional maximum likelihood estimates

(1) $$\begin{aligned} \hat{\mu} &= \bar{Y} \\ \hat{\mathcal{Z}} &= S = n^{-1} \Sigma(Y_i - \bar{Y})(Y_i - \bar{Y})' \quad . \end{aligned}$$

For the Fisher - Von Mises distributions the result of Theorem 5.5 is not so easy to implement. See 3.8. Another not so convenient, but important, family is the beta family, which will now be discussed.

Consider the family of densities

(2) $$f_{\alpha,\beta}(y) = B^{-1}(\alpha, \beta) y^{\alpha-1}(1 - y)^{\beta-1}, \qquad 0 < x < 1, \quad \alpha > 0, \quad \beta > 0 \quad .$$

realtive to Lebesgue measure on (0, 1), where $B = B(\alpha, \beta)$ denotes the beta function,

(3) $$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad '$$

This is a two parameter exponential family with canonical parameters $(\alpha, \beta) \in N = (0, \infty) \times (0, \infty)$. The corresponding canonical statistics are

$$(4) \qquad x_1 = \log y \qquad x_2 = \log (1 - y) \quad .$$

In this case the canonical parameters themselves have a convenient statistical interpretation since

$$(5) \qquad E(Y) = \alpha/(\alpha + \beta), \qquad E(1 - Y) = \beta/(\alpha + \beta)$$

$$\mathrm{Var}(Y) = \alpha\beta/(\alpha + \beta)^2(\alpha + \beta + 1) = \mathrm{Var}\,(1 - Y) \quad .$$

The mean value parameters are somewhat less convenient. One has

$$(6) \qquad \xi_2(\beta, \alpha) = \xi_1(\alpha, \beta) = B^{-1}(\alpha, \beta) \int_0^1 (\ln y) y^{\alpha-1}(1 - y)^{\beta-1} dy$$

$$= \frac{\partial}{\partial\alpha} (\ln B(\alpha, \beta)) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha + \beta)}{\Gamma(\alpha + \beta)}$$

$$= -\sum_{k=0}^{\infty} \left(\frac{1}{\alpha+k} - \frac{1}{\alpha+\beta+k}\right) = -\sum_{k=0}^{\infty} \frac{\beta}{(\alpha+k)(\alpha+\beta+k)} \quad ,$$

and

$$(7) \qquad \xi_1(\alpha, \beta) = -\sum_{k=0}^{\beta-1} \frac{1}{\alpha+k} \quad \text{if} \quad \beta = 1,2,\ldots \quad .$$

(See e.g. Courant and Hilbert (1953, p.499)).

Suppose $Y_1,\ldots,Y_n$ are i.i.d. beta variables, and X_{1i}, X_{2i} are defined from Y_i through (3), $i=1,\ldots,n$. Then the maximum likelihood estimates of (α, β) can be found numerically by solving

$$(8) \qquad \xi_j(\hat{\alpha}, \hat{\beta}) = \bar{X}_j \qquad j = 1, 2$$

from (6), where $\bar{X}_j = n^{-1} \sum_{i=1}^{n} X_{ji}$. An exact solution appears to be unavailable, except when $\hat{\alpha},\hat{\beta}$ turn out to be integers so that (7) applies.

According to Theorem 5.5, the solution to (8) exists if and only if $\bar{X} \in K^\circ$. Now,

$$K \;=\; \text{conhull}\,\{\ln y,\ \ln(1-y) : y \in (0,\infty)\} \quad .$$

Since $\{\ln y, \ln(1-y) : y \in (0,1)\}$ is strictly convex in R^2 this solution therefore exists if and only if $n \geq 2$ and $\sum_{i=1}^{n} (Y_i - \bar{Y})^2 > 0$. The event $\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 0$ occurs with zero probability when $n \geq 2$; hence the maximum likelihood estimate exists with probability one when $n \geq 2$.

NON-FULL FAMILIES

We now proceed to discuss the existence and construction of maximum likelihood estimators when $\Theta \subsetneq N$. Here is an existence theorem.

5.7 Theorem

Let ϕ be steep and regularly strictly convex. Let $\Theta \subset N$ be a non-empty relatively closed subset of N. Suppose $x \in R$. Then $\hat{\theta}_\Theta(x)$ is non-empty.

Suppose $x \in \bar{R} - R$. Suppose there are values $x_i \in R$, $i=1,\ldots,I$, and constants $\beta_i < \infty$ such that

$$(1) \qquad \Theta \subset \bigcup_{i=1}^{I} H^-((x - x_i), \beta_i) \quad .$$

Then $\hat{\theta}_\Theta(x)$ is non-empty.

Remark. See Exercises 5.7.1-2, 7.9.1-3, and Theorem 5.8 for more information about the theorem. In particular, (1) implies $x \notin (\xi(\Theta))^-$. See Figure 5.7(1) for an illustration of 5.7(1).

Proof. Let $x \in R$. $\ell(\cdot, x)$ is upper semi-continuous and satisfies 5.3(3). Hence $\ell(\cdot, x)$ assumes its supremum over $\bar{\Theta}$. But $\ell(\theta, x) = -\infty$ for $\theta \in (\bar{\Theta} - \Theta) \subset \bar{N} - N$. It follows that $\hat{\theta}_\Theta(x)$ is non-empty.

Suppose $x \in \bar{R} - R$ and (1) is valid. Then for each $\theta \in \Theta$ there is an index i for which $\theta \in \bar{H}^-(x - x_i, \beta_i)$. For this index

$$(2) \qquad \ell(\theta, x) \;=\; \theta\cdot(x - x_i) + \theta\cdot x_i - \psi(\theta) \;\leq\; \beta_i + \theta\cdot x_i - \psi(\theta) \quad .$$

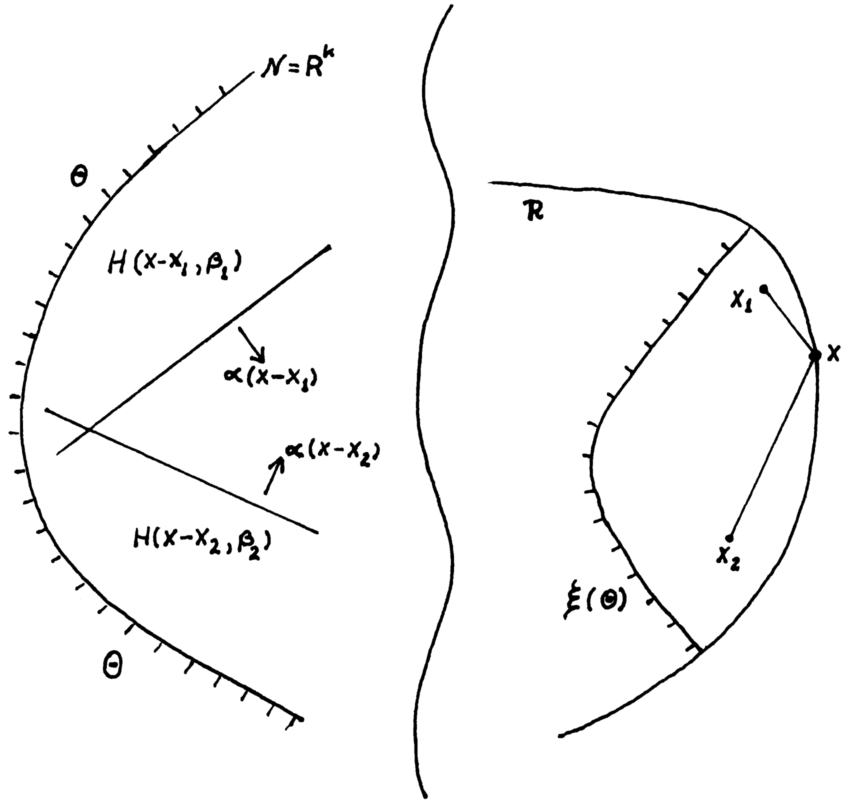

Figure 5.7(1):

An Illustration of 5.7(1) showing R, $x \in \bar{R} - R$, $\Theta \subset \bigcup_{i=1}^{2} H^-((x - x_i), \beta_i)$ and $\xi(\Theta)$.

It follows that

$$\ell(\theta, x) \leq \sup\{\beta_i + \theta \cdot x_i - \psi(\theta) : 1 \leq i \leq I\} \to -\infty \text{ as } ||\theta|| \to \infty ,$$

$$\theta \in \Theta ,$$

by 5.3(3). The second assertion of the theorem follows from (2) as did the first from 5.3(3). ||

CONVEX PARAMETER SPACE

When Θ is convex one gets a better result, including a fundamental equation defining the maximum likelihood estimator.

5.8 Theorem

Assume ϕ is as above. Suppose Θ is a relatively closed convex subset of N with $\Theta \cap N^\circ \neq \phi$. Then $\hat{\theta}_\Theta(x)$ is non-empty if and only if $x \in R$ $(= K^\circ)$ or $x \in \bar{R} - R$ and

$$(1) \qquad \Theta \subset H^-(x - x_1, \beta_1)$$

for some $x_1 \in R$, $\beta_1 \in \mathrm{R}$. ((1) is the same as 5.7(1) with $I = 1$.)

If $\hat{\theta}_\Theta(x)$ is non-empty then it consists of a single point. This is the unique point, $\hat{\theta} \in \Theta \cap N^\circ$ satisfying

$$(2) \qquad (x - \xi(\hat{\theta})) \cdot (\hat{\theta} - \theta) \geq 0 \qquad \forall\, \theta \in \Theta \qquad .$$

(An alternate form of (2) when $x - \xi(\hat{\theta}) \neq 0$ is

$$(2') \qquad \Theta \subset \bar{H}^-(x - \xi(\hat{\theta}),\ (x - \xi(\hat{\theta})) \cdot \hat{\theta}) \qquad .)$$

(Note that if $\theta(x) \in \Theta$ then, of course, $\hat{\theta}_\Theta(x) = \{\theta(x)\}$ and $\hat{\theta} = \theta(x)$ trivially satisfies (2). See 5.9 for illustrations of (2).)

Proof. $\ell(\cdot, x)$ is strictly concave on N and hence can assume its maximum at only one point of the convex set Θ. Furthermore, $\hat{\theta}_\Theta \subset N^\circ$ by Lemma 5.4.

Suppose (2) is satisfied. Then for $\theta \in \Theta$

$$\begin{aligned}(3) \qquad \ell(\hat{\theta}, x) - \ell(\theta, x) &= (\hat{\theta} - \theta) \cdot (x - \xi(\hat{\theta})) \\ &\quad + (\hat{\theta} - \theta) \cdot \xi(\hat{\theta}) - (\psi(\hat{\theta}) - \psi(\theta)) \\ &= (\hat{\theta} - \theta) \cdot (x - \xi(\hat{\theta})) + \ell(\hat{\theta}, \xi(\hat{\theta})) - \ell(\theta, \xi(\hat{\theta})) \\ &\geq 0 + 0 = 0 \ ,\end{aligned}$$

with equality if and only if $\theta = \hat{\theta}$. $(\ell(\hat{\theta}, \xi(\hat{\theta})) - \ell(\theta, \xi(\hat{\theta})) > 0$ when $\theta \neq \hat{\theta}$ since $\hat{\theta} = \theta(\xi(\hat{\theta}))$ is the unique maximum likelihood estimator over N corresponding to the observation $\xi(\hat{\theta})$.) Hence (2) implies that $\hat{\theta}_\Theta(x) = \{\hat{\theta}\}$.

On the other hand, suppose

$$(4) \qquad (x - \xi(\theta_0)) \cdot (\theta_0 - \theta_1) < 0 \qquad \text{for some } \theta_0, \ \theta_1 \in \Theta \quad .$$

Then

$$\theta_\rho = \theta_0 + \rho(\theta_1 - \theta_0) \in \Theta$$

for $0 \le \rho \le 1$ since Θ is convex. Then

$$\frac{d}{d\rho} \ell(\theta_\rho, x)_{|\rho=0} = (x - \xi(\theta_0)) \cdot (\theta_1 - \theta_0) > 0 \ .$$

Hence $\ell(\theta_\rho, x) > \ell(\theta_0, x)$ for $\rho > 0$ sufficiently small; and θ_0 cannot be the unique maximum likelihood estimator. It follows that the unique maximum likelihood estimator if it exists, must satisfy (2).

Finally, if $x \in R$ or (1) is satisfied then $\hat{\theta}_\Theta$ is non-empty by Theorem 5.7. Conversely, if $\hat{\theta} \in \hat{\theta}_\Theta$ is non-empty then $\hat{\theta}$ satisfies (2). Hence $\hat{\xi} = \xi(\hat{\theta}) \in R$ and

$$(x - \hat{\xi}) \cdot \theta \le (x - \hat{\xi}) \cdot \hat{\theta}$$

by (2) so that (1) is satisfied with $x_1 = \hat{\xi}$. ||

5.9 Construction

The criterion 5.8(1) is particularly easy to apply if $\Theta = (\theta_0 + L) \cap N$ for some linear subspace, L. This is because the vectors $\{(\hat{\theta} - \theta) : \theta \in L\}$ will then span L. Thus, by (1), in order to find $\hat{\theta}$ one need only search for the unique point $\theta^* \in \Theta$ for which $x - \xi(\theta^*) \perp L$. This process can be viewed from two slightly different perspectives. Because of its importance we illustrate both these perspectives in the simplest case where $\theta_0 + L$ is a hyperplane.

Thus, consider the case where $\Theta = H \cap N$ with H a hyperplane, say $H = H(a, \alpha)$. Let $x \in R$. (The same construction also works for $x \in \bar{R} - R$ if 5.7(1) is satisfied.) To find $\hat{\theta}_\Theta(x)$ one may proceed from $\theta(x)$ along the curve $\{\theta(x + \rho a) : \rho \in R\}$ until the unique point at which $\theta(x + \rho a) \in \Theta$. This point is $\hat{\theta}$. The process is illustrated in Figure 5.9(1).

An alternative procedure is to map $\Theta \cap N^{\circ}$ into R as $\xi(\Theta \cap N^{\circ})$. Then proceed along the line $\{x + \rho a : \rho \in R\}$ until the unique point at which $x + \rho a \in \xi(\Theta \cap N^{\circ})$. This point is $\hat{x} = \xi(\hat{\theta})$. This process is illustrated in Figure 5.9(2).

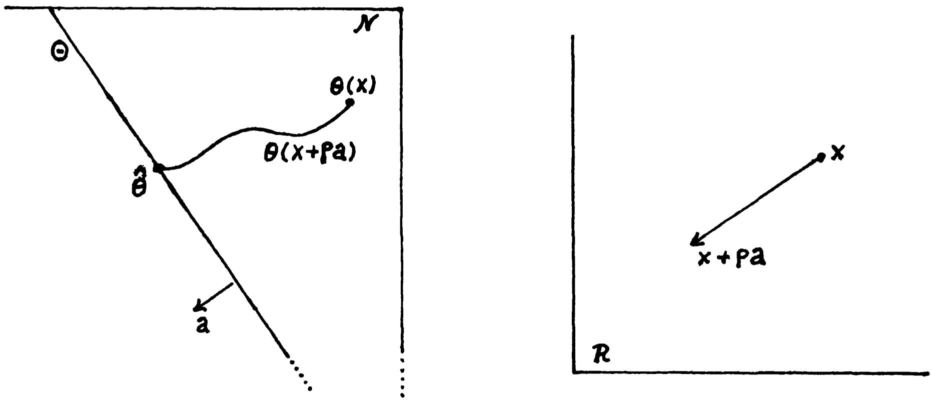

Figure 5.9(1): First construction of $\hat{\theta}$

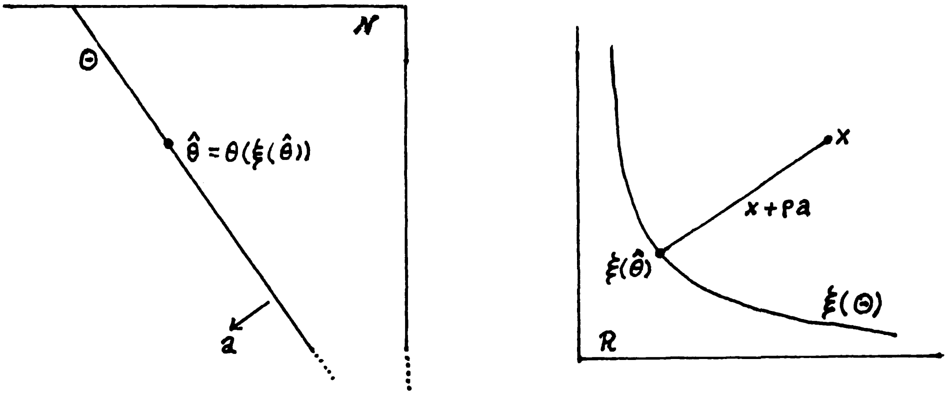

Figure 5.9(2): Second construction of $\hat{\theta}$

There are useful paradigms available also for the case where Θ is an arbitrary relatively closed convex set. These are described in 5.13.

The entire process illustrated above may also be viewed from a different perspective. Θ is contained in a proper linear subset of R^k. Hence the densities $\{p_\theta : \theta \in \Theta\}$ form an exponential family which is not minimal. This non-minimal family can be reduced by sufficiency and reparametrization to a minimal family of dimension $k' < k$. Let $(\phi_1,\ldots,\phi_{k'})$ and

$(y_1, \ldots, y_{k'})$ denote the natural parameters and corresponding observations in this family. (They are formed by projecting θ and x, respectively, onto $H(a, \alpha)$ or any translate $H(a, \beta)$.) This family will have log-Laplace transform $\psi^*(\phi) = \psi(\theta(\phi))$, and the m.l.e., $\hat{\phi}$, satisfies 5.5(1) -- i.e.

$$\hat{\phi}(y) = \phi(y)$$

where $\phi(y)$ is the inverse to $\xi^*(\phi) = \nabla\psi^*(\phi)$. Thus

$$\hat{\theta}(x) = \phi(y(x)) \quad .$$

These remarks can be used to yield a very simple proof of Theorem 5.8 in the special case where $\Theta = (\theta_0 + L) \cap N$. They also provide a method of easily constructing the maximum likelihood estimate in many such cases. Here are two examples.

5.10a Example

Consider the classical Hardy-Weinberg situation described in Example 1.8. (X_1, X_2, X_3) is multinomial (N, ξ) with expectation $\xi = N(p^2, 2pq, q^2)$, $0 < p = 1-q < 1$. This is a three-dimensional exponential family with two dimensional parameter space

$$\Theta = \{\theta: \; = \beta_1(1,1,1) + \beta_2(2,1,0) + (0, \ln 2, 0)\} = H((1,-2,1), -2 \ln 2).$$

(This family is not minimal. This fact affects but does not hinder the reasoning which follows.)

Reduction to a minimal exponential family yields a one-parameter exponential family with parameter $\phi = 2\theta_1 + \theta_2$ and natural observation $y = 2x_1 + x_2$. (Θ is two-dimensional but yields a family of only order one since the original family was not minimal.) Note that

$$E(Y) = N(2p^2 + 2pq) = 2pN \quad . \tag{1}$$

Hence

$$\hat{p} = \frac{y}{2N} = \frac{2x_1 + x_2}{2N} , \qquad 0 < y < 2N \quad . \tag{2}$$

Correspondingly, $\hat{\xi} = N(\hat{p}^2, 2\hat{p}\hat{q}, \hat{q}^2)$ and $\hat{\theta}$ can be defined from $\hat{\theta}_i = \beta_1 + \ln \hat{\xi}_i$, $\beta_1 \in R$. (Note that $\hat{\theta}$ is a line rather than a single point because the original representation of the multinomial family was not minimal.)

The simplicity of (1) is the special fact which enables the preceding construction to proceed so smoothly. Many other multinomial log-linear models behave similarly. Classes of such models are discussed in Darroch, Lauritzen, and Speed (1980) and in Haberman (1974). Here is a useful example.

5.10b Example

Consider a 2×2×2 contingency table. The observations will be denoted by y_{ijk}, $i,j,k = 0,1$. They are multinomial (N) variables with respective probabilities π_{ijk}. There are various useful log-linear models for such a table. The derivation of maximum likelihood estimates for such models provides a useful and illuminating application of the preceding theory. Here we consider the model in which responses in the first category (corresponding to index i) are conditionally independent of those in the third category given the level of response in the second category. This model illustrates several characteristic phenomena, and allows for direct and explicit maximum likelihood estimates of the parameters π_{ijk}.

In order to write the model in customary vector-matrix notation, let $z_\ell = y_{ijk}$ where $\ell = 1 + i + 2j + 4k$ $(1 \leq \ell \leq 8)$, and, similarly, $\pi_\ell = \pi_{ijk}$. Let $(\log \pi)$ denote the vector with coordinates $\log \pi_\ell$, $\ell=1,\ldots,8$. Let

$$D' = \begin{matrix} 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{matrix}$$

The log-linear model of interest here has

$$\theta^* = (\log \pi) = D\beta, \qquad \beta \in R^6 . \tag{2}$$

In order to normalize π one must choose β_6 so that

(3) $$\sum_{\ell=1}^{8} \pi_\ell = 1 \quad .$$

The resulting multinomial family is an 8-parameter exponential family. Its canonical statistic can be reduced via sufficiency. Let $x^* = D'z$ so that $\theta^{*\prime}z = \beta'D'z = \beta'x^*$. Furthermore $x^*_6 = N$ with probability one. Hence $x \in R^5$ with $(x_1,\ldots,x_5) = (x^*_1,\ldots,x^*_5)$ is a sufficient, canonical statistic. The corresponding canonical parameter is $\theta \in R^5$ with $(\theta_1,\ldots,\theta_5) = (\beta_1,\ldots,\beta_5)$. It can be checked thatthis log-linear family is characterized by the conditional independence of responses in categories 1 and 3 given level of response in category 2. The conditional independence can be checked by noting that if $i \neq i'$, $k \neq k'$ then (2) yields

$$\ln \pi_{i'jk'} + \ln \pi_{ijk} = \ln \pi_{i'jk} + \ln \pi_{ijk'} \quad .$$

From this it follows that $\pi_{+j+}\,\pi_{ijk} = \pi_{+jk}\,\pi_{ij+}$, which implies the desired conditional independence.

By Theorem 4.5 x is the maximum likelihood estimate of $E(X) = \xi(\theta)$. Thus, $(\log \pi) = D\beta(x)$ is the maximum likelihood estimate of $(\log \pi) = \theta^*$ with $\beta(x) = (\beta_1(x),\ldots,\beta_5(x),\ \beta_6(\beta))'$ where $\beta_6(\cdot)$ is determined by (3).

The relation between $\xi(\theta)$ and $\pi(\theta)$ is easy to determine via simple calculations such as $\xi_1 = E(X_1) = \Sigma(-1)^i\, E(y_{ijk})$, etc. These yield

(4) $$\begin{aligned} \xi_1 &= N\Sigma(-1)^i\pi_{ijk} & \xi_2 &= N\Sigma(-1)^j\pi_{ijk} \\ \xi_3 &= N\Sigma(-1)^k\pi_{ijk} & \xi_4 &= N\Sigma(-1)^{i+j}\pi_{ijk} \\ \xi_5 &= N\Sigma(-1)^{j+k}\pi_{ijk} \quad . \end{aligned}$$

Thus

(5) $$\Sigma(-1)^i y_{ijk} = x_1 = N\Sigma(-1)^i\hat{\pi}_{ijk} \ , \qquad \text{etc.}$$

From these relationships and the structure of D it is possible in this case to give explicit expressions for $\{\hat{\pi}_{ijk}\}$ in terms of $\{y_{ijk}\}$. Let a "+" replacing a subscript denote addition over that subscript. Thus, $\pi_{1++} = \sum_{j,k} \pi_{1jk}$. Simple manipulation based on (3) and (5) yields

(6)
$$N\hat{\pi}_{i++} = y_{i++} , \qquad N\hat{\pi}_{+j+} = y_{+j+} , \qquad N\hat{\pi}_{++k} = y_{++k}$$
$$N\hat{\pi}_{ij+} = y_{ij+} , \qquad N\hat{\pi}_{+jk} = y_{+jk} \quad .$$

The conditional independence properties yield

$$\pi_{ijk} = \pi_{+j+} \frac{\pi_{ij+}}{\pi_{+j+}} \frac{\pi_{+jk}}{\pi_{+j+}} = \pi_{ij+}\, \pi_{+jk}/\pi_{+j+} \quad .$$

Hence

(7)
$$N\hat{\pi}_{ijk} = y_{ij+}\, y_{+jk}/y_{+j+} \quad .$$

FUNDAMENTAL EQUATION

5.11 Definition

For $\theta_0 \in \Theta \subset R^k$ define $\nabla_\Theta(\theta_0)$, the *set of (outward) normals* to Θ at $\theta_0 \in \Theta$, to be the set of all $\delta \in R^k$ satisfying

(1)
$$\delta \cdot (\theta_0 - \theta) \geq 0 + o(||\theta_0 - \theta||) \quad \forall \ \theta \in \Theta \quad .$$

∇ is obviously a convex cone, and can easily be shown to be closed.

Note that if $\theta_0 \in \operatorname{int} \Theta$ then $\nabla_\Theta(\theta_0) = \{0\}$. If θ_0 is an isolated point of Θ then $\nabla_\Theta(\theta_0) = R^k$. If Θ is a differentiable manifold with tangent space T at θ_0 then $\nabla_\Theta(\theta_0)$ is the orthogonal complement of T -- i.e., $\nabla_\Theta(\theta_0) = \{\delta\colon \delta \cdot \tau = 0 \ \ \forall \tau \in T\}$. Here $\nabla_\Theta(\theta_0)$ is a linear subspace of R^k. If Θ is convex and $\theta_0 \in \operatorname{bd} \Theta$ then $\nabla_\Theta(\theta) = \{\delta : \Theta \subset \bar{H}^- (\delta, \delta \cdot \theta_0)\}$.

5.12 Theorem

Assume ϕ is steep and regularly strictly convex. Let Θ be a relatively closed subset of N. Then for any $\hat{\theta} \in \hat{\theta}_\Theta(x) \cap N^\circ$

$$(1) \qquad x - \xi(\hat{\theta}) \in \nabla_\Theta(\hat{\theta}) \quad .$$

Proof. Let $\hat{\theta} \in \hat{\theta}_\Theta(x) \subset N^\circ$. Note that

$$(2) \qquad \nabla_\theta(\ell(\theta, \xi(\hat{\theta}))_{|\theta=\hat{\theta}} = 0 \quad .$$

and $x - \xi(\theta) = 0$ when $\theta = \hat{\theta}$. Hence, as in 5.8(3)

$$(3) \quad 0 \leq \ell(\hat{\theta},x) - \ell(\theta,x) = (\hat{\theta} - \theta) \cdot (x - \xi(\hat{\theta})) + \ell(\hat{\theta}, \xi(\hat{\theta})) - \ell(\theta, \xi(\hat{\theta}))$$

$$= (\hat{\theta} - \theta) \cdot (x - \xi(\hat{\theta})) + o(||\hat{\theta} - \theta||) \quad .$$

Thus, by definition, (1) is satisfied. ||

Note that the theorem does not require $x \in R$ $(= K^\circ)$.

5.13 Construction

The fundamental equation, 5.8(1) or 5.12(1), can be used to picture the process of finding a maximum likelihood estimator, by an extension of the process pictured in 5.9.

Fix $x \in R^k$. Suppose it is desired to locate $\hat{\theta}_\Theta(x)$. If $\Theta \cap N^\circ \neq \phi$ one should first check to see whether $x \in \xi(\Theta \cap N^\circ)$. If so, then $\theta(x) = \hat{\theta}_\Theta(x)$. If not, then $\hat{\theta}_\Theta(x) \subset \text{bd } \Theta$. To see whether a given $\theta_0 \in \text{bd } \Theta \cap N^\circ$ can be an element of $\hat{\theta}$ first locate Θ, θ_0,x, and $x_0 = \xi(\theta_0)$ on their respective graphs. Then carry a vector δ pointing in the direction of $x - x_0$ over to θ_0 in order to check whether δ is an outward normal to Θ at θ_0. If so, then θ_0 is a candidate for $\hat{\theta}$. In fact, if Θ is convex $\{\theta_0\} = \hat{\theta}_\Theta(x)$. If Θ is not convex one must search over bd Θ for all such candidates, then examine $\ell(\theta, x)$ at each of them to eliminate those which are not global maxima. (If ϕ is not regular and Θ is not convex one needs also to search over

$\Theta \cap (N - N^\circ)$.) The process is illustrated in Figure 5.13(1).

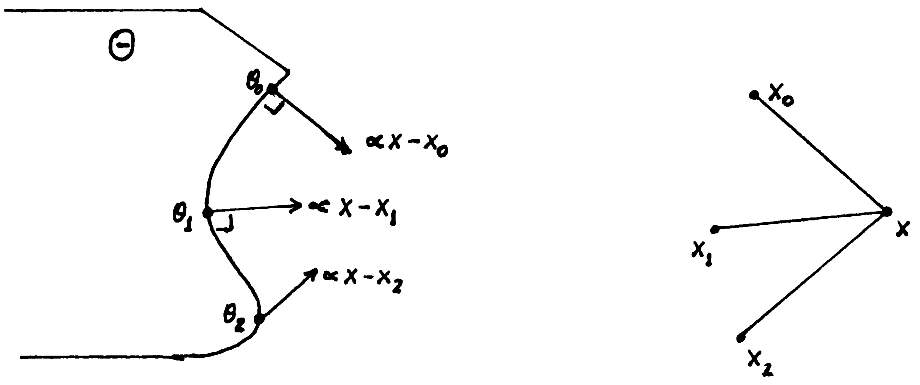

Figure 5.13(1): θ_0 and θ_1 are candidates for $\theta_\Theta(x)$. θ_2 is not.

If bd Θ is a curve as in Figure 5.13(1) then this process is relatively convenient. Otherwise, it is usually less convenient to search over all of bd Θ for the set of candidates.

An alternate picture can also be constructed. In this picture one constructs for each $\theta \in \Theta$ the collection of points in X space for which θ can possibly be the maximum likelihood estimator. In order to construct this picture one locates $\theta \in$ bd Θ and draws the unit outward normal(s), δ, to θ. One then maps θ to $\xi(\theta)$ and carries the vector(s) δ directly over to X space. The corresponding line or cone with vertex located at $\xi(\theta)$ is the locus of values of x for which $\theta \in \hat{\theta}_\Theta(x)$ is a possibility. Again, if x falls in more than one such locus then $\ell(\theta, x)$ must be separately examined at all such θ. This process is illustrated in Figure 5.13(2).

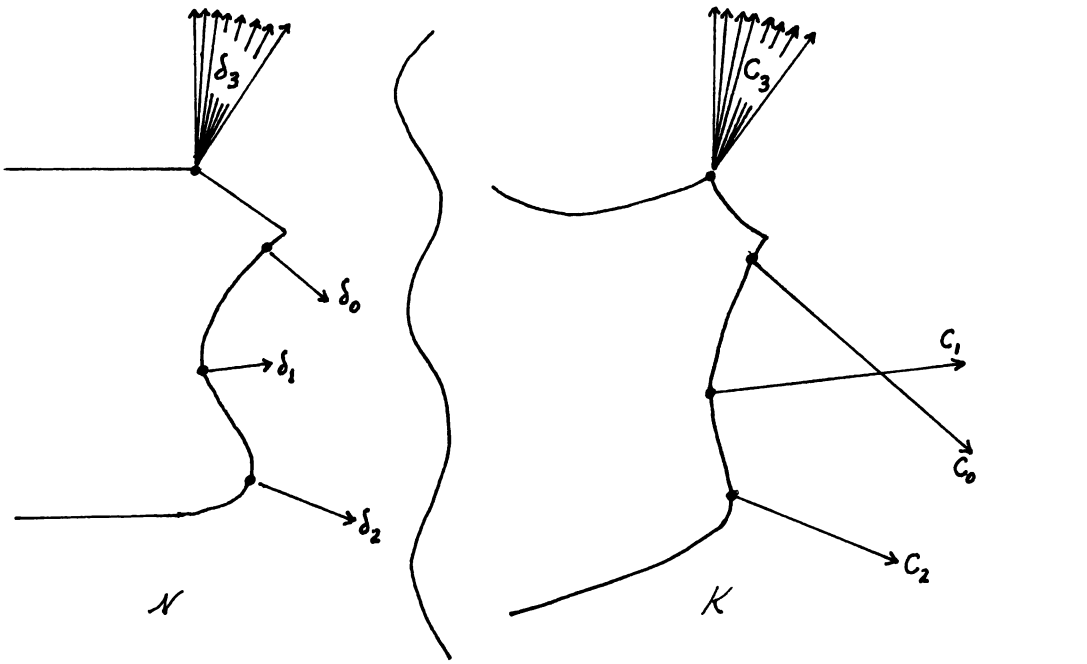

Figure 5.13(2): C_i is the locus of points, x, for which θ_i can possibly fall in $\hat{\theta}_\Theta(x)$.

5.14 Example

The curved exponential family described in Example 3.12 provides a particularly elegant instance of the above construction. The family is a two-parameter standard exponential family with $\theta(\lambda) = (-\lambda, -\ln \lambda)'$, and

$\Theta = \{\theta(\lambda): \lambda > 0\} \subset N = (-\infty, 0) \times R$, and $\psi(\theta) = \ln[(e^{\theta_1 T} - 1)/\theta_1 + e^{\theta_1 T + \theta_2}]$. $K = \text{conhull } \{(0, 0), (T, 0), (T, 1)\}$.

Then, $\xi(\theta(\lambda)) = (\frac{1 - e^{-\lambda T}}{\lambda}, e^{-\lambda T})$. Figure 5.14(1) shows both Θ and K and $\xi(\Theta)$ on a single plot. There is no overlap since $\Theta \subset \{(\theta_1, \theta_2): \theta_1 < 0\}$ and $K \subset \{(x_1, x_2): x_1 \geq 0\}$.

The tangent space to $\theta(\lambda)$ is spanned by $(-1, -1/\lambda)'$. Hence $\nabla_\Theta(\theta(\lambda))$ is the line $\{\rho(1, -\lambda): \rho \in R\}$. The locus, $C(\lambda)$, of points x for which $\theta(\lambda)$ can be the maximum likelihood estimator is the line

$$(2)\quad C(\lambda) = \{\xi(\theta(\lambda)) + \rho(1, -\lambda): \ \rho \in R\} = \{ (\frac{1 - e^{-\lambda T}}{\lambda} + \rho, \ e^{-\lambda T} - \lambda\rho): \rho \in R\}$$

$$= \{(0, 1) + \sigma(1, -\lambda): \ \sigma \in R\}$$

as can be seen by letting $\sigma = \frac{1 - e^{-\lambda T}}{\lambda} + \rho$. Formula (2) reveals that the loci $C(\lambda)$ are straight lines through the point (0, 1). Again, see Figure (1).

It can be seen from Theorem 5.7 that $\hat{\theta}(x) \neq \phi$ unless $x \in K$ is (0, 0) or (T, 1). (Applying 5.7(1) for points on the interior of the line joining (0, 0) to (T, 1) requires the choice I = 2. Of course, these points occur with probability zero, so it's not worth the effort!) Since the loci $C(\lambda)$ intersect only at $(0, 1) \notin K$ it follows from (2) that if $x \neq (0, 0)$ or (T, 1) then $\hat{\theta}_\Theta(x)$ is the single point, $\theta(\lambda)$, for which $x \in C(\lambda)$.

If $x = (0, 0)$ or (T, 1) then $\hat{\theta}(x) = \phi$ since neither of these points lies in $\bigcup_{\lambda \in R} C(\lambda)$. (That $\hat{\theta}(x) = \phi$ in this case can also be seen by applying the final part of Theorem 5.8 to the parameter set consisting of the convex hull of Θ).

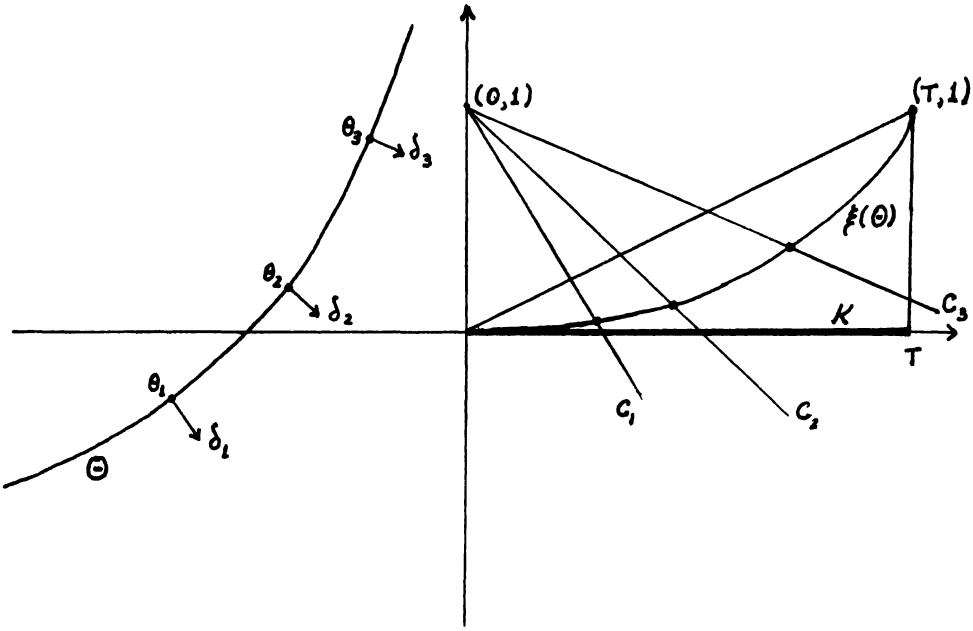

Figure 5.14(1): Illustrating the construction of $\hat{\theta}(x)$ via construction of loci C.

The original description of this example involves a single observation, X, which can take only values in $(0 \times [0, T]) \cup \{(T, 1)\}$. However, if one observes $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ where X_i are n i.i.d. variables each with the given distribution, then $\bar{X}_n$ can take values over more of K. This problem has natural parameter $\theta^* = n\theta$ and log Laplace transform $\psi^*(\theta^*) = n\psi(\theta^*/n)$. It follows that N, K *and* $\xi(\Theta)$ are as before. Θ undergoes a simple transformation. It is easy to check that the above picture applies equally well to this problem, for which various values of $x \in K°$ are possible. See also Proposition 5.15.

From (2) one sees that the maximum likelihood estimator of λ is

$$\hat{\lambda} = (1 - \bar{x}_2)/\bar{x}_1 \quad . \tag{3}$$

In terms of the original motivation for this problem the parameter $1/\lambda$ is the mean value (= mean lifetime) of the exponential variable Z. Thus,

$$\widehat{(1/\lambda)} = \bar{x}_1/(1 - \bar{x}_2) = \frac{n\bar{x}_1}{n(1 - \bar{x}_2)} \quad . \tag{3'}$$

In this problem $n\bar{x}_1 = \sum_{i=1}^{n} Y_i$ = "total time on test", and $n(1 - \bar{x}_2)$ = (number of observations < T) = "number of objects failing before truncation". This supplies the familiar expression for this problem:

$$\widehat{(1/\lambda)} = \frac{\text{total time on test}}{\text{number of objects failing before truncation}} \quad . \tag{3''}$$

Note that the value of T does not appear in (3"). This fact has been commented on and exploited by Cox (1975) and many others.

It has been noted that the differentiable subfamily treated in this example is a stratum within the full two parameter family. It is really this fact which explains the elegance of the above construction and of Figure 5.14(1). See Exercise 5.14.1 - 5.14.3.

In general the maximum likelihood estimate for an i.i.d. sample is determined exactly as that from a single observation. The latter part of Example 5.14 mentions one special case of this. It is worthwhile to formally note this fact.

5.15 Proposition

Let $X_1,\ldots,X_n$ be i.i.d. random variables from a standard exponential family $\{p_\theta\colon \theta \in \Theta\}$. Let $\hat{\theta}_\Theta$ denote the set of maximum likelihood estimators of $\theta \in \Theta$ on the basis of a single observation.

The maximum likelihood estimator of $\theta \in \Theta$ based on the sample $X_1,\ldots,X_n$ is a function of the sufficient statistic, $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Let $\hat{\theta}_\Theta^{(n)}(\cdot)$ denote this function of $\bar{X}_n$. Then

$$(1) \qquad \hat{\theta}_\Theta^{(n)}(\bar{x}) = \hat{\theta}_\Theta(\bar{x}) \quad .$$

Proof. The cumulant generating function for the sufficient statistic $S_n = n\bar{X}_n$ is $n\psi(\theta)$. The proposition follows from the fact that

$$\begin{aligned} \ell_{n\psi}(\theta, s) &= \theta \cdot s - n\psi(\theta) \\ &= n(\theta \cdot s/n - \psi(\theta)) \\ &= n\ell_\psi(\theta, s/n) \quad , \end{aligned}$$

since this shows that $\ell_{n\psi}(\cdot, s)$ is maximized if and only if $\ell_\psi(\cdot, s/n)$ is maximized. ||

EXERCISES

5.6.1

Verify formula 5.6(6).

5.6.2

The multivariate generalization of the beta distribution is the *Dirichlet distribution*, $\mathcal{D}(\alpha)$, defined as follows: $k \geq 2$; $\theta_i > 0$, $i=1,\ldots,k$, $\theta_0 = \sum_{i=1}^{k} \theta_i$, $Y_i > 0$, $i=1,\ldots,k$; $Y_k = 1 - \sum_{i=1}^{k-1} Y_i$; the distribution has density with respect to Lebesgue measure over the allowable $\{(y_1,\ldots,y_{k-1})\}$

(1) $$f_\theta(y) = \frac{\Gamma(\theta_0)}{\prod_{i=1}^{k} \Gamma(\theta_i)} \prod_{i=1}^{k} y_i^{(\theta_i - 1)} .$$

This is a k-parameter exponential family with canonical statistic $X_i = \ln Y_i$.

(i) Describe K.

(ii) Verify the standard formulae:

(2) $$E(Y_i) = \theta_i/\theta_0 \qquad \mathrm{Var}(Y_i) = \frac{(\theta_0 - \theta_i)\theta_i}{\theta_0^2(\theta_0 + 1)}$$

$$\mathrm{Cov}\,(Y_i, Y_j) = -\frac{\theta_i\theta_j}{\theta_0^2(\theta_0 + 1)} , \qquad i \neq j \quad .$$

(iii) Derive formulae for $E(X_i)$ analogous to 5.6(6), (7).

(iv) Let $1 = s_0 < \ldots < s_\ell = k$ and define $z_j = \sum_{i=s_{j-1}+1}^{s_j} Y_i$, $j=1,\ldots,\ell$. Show that Z has a $\mathcal{D}(\theta')$ distribution, and describe θ' in terms of θ.

(v) Let $Y^{(i)}$, $i=1,\ldots,n$ be independent, k-dimensional $\mathcal{D}(\theta^{(i)})$ variables. Verify that the distribution of $n^{-1} \sum_{i=1}^{n} Y^{(i)}$ is $\mathcal{D}(\Sigma\theta^{(i)})$.

5.6.3

Let X_i, $i=1,\ldots,k$, be independent $\Gamma(\alpha_i, \beta)$ variables. Describe the conditional distribution of the variable $(X_1,\ldots,X_k)$ given $\sum_{i=1}^{k} X_i$ as a multiple of an appropriate Dirichlet variable. (Note the partial analogy between the situation here and that in Example 1.16. Note also that the situation here was described from another perspective in Exercise 2.15.1.)

5.6.4

The following is a valid statement: the k-dimensional Dirichlet distributions form the family of (proper) conjugate priors for the parameter $(p_1,\ldots,p_{k-1})$ of a k-dimensional multinomial distribution. Relate this statement to the general theory of Sections 4.18-4.20, and describe (in terms of the Dirichlet parameters) the posterior expectation of p given the multinomial observation. [Let $p_i = e^{\theta_i}/(1 + \sum_{j=1}^{k-1} e^{\theta_i})$, etc.]

(This conjugate relation between Dirichlet and multinomial distributions has an infinite dimensional generalization in which the Dirichlet distribution is replaced by a "Dirichlet process" and the multinomial distribution is replaced by a distribution over the family of cumulative distribution functions on [0, 1]. See Ferguson (1973) and Ghosh and Meeden (1984).)

5.7.1

(i) Show that 5.7(1) implies $x \notin (\xi(\Theta))^-$.

(ii) Show the converse is not valid by constructing an example in which $\phi = \psi$, $\mathcal{R} = \mathcal{K}^\circ$ is not strictly convex, $x \notin (\xi(\Theta))^-$, and 5.7(1) fails. (I believe no example exists when $\mathcal{R}$ is strictly convex. See Exercise 7.9.2 which shows that when $\mathcal{R} = \mathcal{K}^\circ$ is strictly convex and $x \notin (\xi(\Theta))^-$ then $\hat{\theta}(x) \neq \phi$.)

[(i) $x \notin (\xi(H^-((x - x_i), \beta_i)))^-$ for $x_i \in \mathcal{R}$, $x \in \bar{\mathcal{R}} - \mathcal{R}$.

(ii) Let ν give mass 1 to each of the four points $(\pm 1, \pm 1)$. Let $x = (1, 0)$

and $\Theta = \{(t, 2): t \in R\}$.]

5.7.2

Construct examples in which $\phi = \psi$ is steep, $R = K^\circ$, $x \in (\xi(\Theta))^-$, and (i) $\hat{\theta}(x) = \phi$, (ii) $\hat{\theta}(x) \neq \phi$. [For both examples let ν be the uniform distribution on the ball $\{x: (x_1 - 1)^2 + \sum_{i=2}^{k} x_i^2\}$ plus a point mass at 0. For (i) let $\Theta = \{\theta: \theta = (\alpha, 0,\ldots,0)\}$. For (ii) let $\Theta = \{\theta : \psi(\theta) = 3\}$. For every unit vector $v \neq e$, there is a unique $\eta(v) > 0$ such that $\psi(\eta(v)v) = 3$. As $v \to e_1$, $\eta(v) \to \infty$ and hence $\xi(\eta(v)v) \to 0$. Hence $0 \notin (\xi(\theta))^-$.]

5.8.1

Let $\{p_\theta: \theta \in \Theta\}$ be a standard one-parameter exponential family. Suppose $\xi(\Theta)$ is an unbounded interval -- i.e. $\xi(\Theta) \supset (\xi_0, \xi_1)$ with $\xi_0 = -\infty$ or $\xi_1 = +\infty$. For $\xi_0 < A < \xi_1$ suppose either

$$(1) \qquad \xi_0 = -\infty \text{ and } \int_{\xi_0}^{A} J(\xi)d\xi = \infty$$

or

$$\xi_1 = \infty \text{ and } \int_{A}^{\xi_1} J(\xi)d\xi = \infty$$

with $J^{-1}(\xi) = \theta'(\xi) = \mathrm{Var}_{\theta(\xi)}(X)$, so that J denotes the Fisher information for estimating ξ. Consider the problem of estimating ξ under the loss 4.6(1) -- i.e. $L(\xi, \delta) = J(\xi)(\delta - \xi)^2$. Show that: (i) the maximum likelihood estimator is minimax; and (ii) if $\Theta \subsetneq N$ then the maximum likelihood estimator is not admissible. (iii) Give examples when $\Theta = N$ and $\xi(\Theta)$ is unbounded in which the maximum likelihood estimator is not minimax, is minimax but not admissible, is both minimax and admissible. (iv) Can you generalize (i) to a k-parameter family?

[Let $\alpha_n \downarrow \xi_0$, $\beta_n \uparrow \xi_1$ and

$$(2) \qquad h_n^{1/2}(\xi) = \min\left(\int_{\alpha_n}^{\xi} J(t)dt,\ K_n,\ \int_{\xi}^{\beta_n} J(t)dt\right)$$

where K_n is chosen so that h_n is a probability density. Show $K_n \to 0$ because of (1). Then use 4.6(2). For (ii) use Theorem 4.24.]

5.9.1

Consider the general linear model as defined in 1.14.1. (a) Verify that the usual least squares estimators of ξ are also the maximum likelihood estimators (i.e. $\hat{\mu} = B\hat{\xi}$). (b) What is the maximum likelihood estimator of σ^2? Is it unbiased? (Assume $m \geq r + 1$.) (c) Generalize the preceding questions to the situation where $Y \sim N(\mu, \sigma^2\Sigma)$ with $\mu = B\xi$ as in 1.14.1 and Σ a known positive matrix. [The maximum likelihood estimates are the usual generalized least squares estimates.]

5.9.2

Generalize 5.9.1 to the multivariate linear model defined in 1.14.3.

5.9.3

Let (X_1, X_2) be the canonical statistics from a normal sample with mean μ_1 and variance σ_1^2; and let (Z_1, Z_2) be from an independent normal sample with mean μ_2 and variance σ_2^2. Suppose $\mu_1 \leq \mu_2$, but the parameters are otherwise unrestricted. Show that $(\hat{\mu}_1, \hat{\mu}_2) = (x_1, z_1)$ if $x_1 \leq z_1$ and otherwise $\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}$ is the unique solution to

$$\frac{x_1}{x_2 - \hat{\mu}^2} + \frac{z_1}{z_2 - \hat{\mu}^2} = \left(\frac{1}{x_2 - \hat{\mu}^2} + \frac{1}{z_2 - \hat{\mu}^2}\right)\hat{\mu} \quad .$$

(Assume $x_2 > x_1^2$ and $z_2 > z_1^2$, which occurs with probability one.)

5.9.4

Let ξ be a normally distributed vector with mean 0 and covariance matrix Σ. Given ξ let Y be distributed according to the general linear model 1.14.1. (Assume $m \geq r + 1$.) Suppose B'B is diagonal and $\Sigma \subset \mathcal{D}$, a relatively closed convex subset of positive definite diagonal matrices.

(a) Show that the (marginal) distributions of Y form an exponential family with Θ a relatively closed proper convex subset of N. [$\hat{\xi}$ and $|(Y - B\hat{\xi})|^2$ are minimal sufficient statistics.] (b) When $\mathcal{D}$ is all positive definite diagonal matrices describe the maximum likelihood estimates of Σ, σ^2. (c) Extend (b) to include other suitable subsets, $\mathcal{D}$. (d) The preceding is a canonical form for a class of random effects models (see, e.g., Arnold (1981)). To see this convert the usual balanced one-way or two-way random effects models to a model of this form by applying suitable linear transformations to the usual parameters. [For the one-way model having $E(Y_{ij}) = \mu + \alpha_i$, $\mu \sim N(0, \sigma_\mu^2)$, $\alpha_i \sim N(0, \sigma_\alpha^2)$, $i=1,\ldots,I$, $j=1,\ldots,J$ let $\xi_1 = I\mu + \sum_{i=1}^{I} \alpha_i$ and $(\xi_2,\ldots,\xi_I) = (\alpha_1,\ldots,\alpha_I)M$ where M is a I × (I - 1) matrix whose columns are orthonormal and orthogonal to $\underset{\sim}{1}$.]

[The following three exercises concern the 2×2×2 contingency table.]

5.10.1

Consider the model under which the first category and third category are (marginally) independent (i.e., $\pi_{i+k} = \pi_{i++}\pi_{++k}$). Show this is a log-linear model and find an explicit expression for the maximum likelihood estimator.

5.10.2

Consider the log-linear model described by the restriction $0 = \phi_1 + \phi_4 + \phi_6 + \phi_7 - (\phi_2 + \phi_3 + \phi_5 + \phi_8)$. (This is the model described by the phrase, "no third-order interactions.") Write the equation(s) determining the maximum likelihood estimator. Determine that these equations do not have a closed form solution, such as 5.10(7). (See Darroch, Lauritzen, and Speed (1980). In such a case the likelihood equations must be solved numerically. The usual methods are the E-M algorithm or the Newton-Raphson algorithm. See Bishop, Feinberg, and Holland (1975) and Haberman (1974).)

5.12.1

Consider the model described by $\frac{1}{2} = \pi_{0++} = \pi_{1++} = \pi_{+0+} = \pi_{+1+}$.

Show this corresponds to a differentiable subfamily within the full exponential family, but is not a log-linear model. Find the maximum likelihood estimator for this differentiable subfamily [$\pi_{00+} = \pi_{11+}$.]

5.14.1

Let $\{p_\theta \colon \theta \in \Theta\}$ be a stratum of regular (or a steep) exponential family, as defined in Exercise 3.12.1. (a) Show that for $x \in R$ the maximum likelihood estimator exists and satisfies

$$\frac{\hat{\xi}_{(1)}}{\hat{\xi}_{(2)}} = \frac{x_{(1)}}{x_{(2)}} \quad . \tag{1}$$

(b) Discuss the situation when $x \in \bar{R} - R$. (c) Show (by example) that there can be two solutions to (1); but there can never be more than two. Is it possible for both of these solutions to be maximum likelihood estimators? [Suppose the family is defined by $\psi(\theta) = \psi_0$. Note that the set $\{\theta \colon \psi(\theta) \le \psi_0\}$ is convex and apply Theorem 4.8.]

5.14.2

Show how the result of Exercise 5.14.1 directly yields 5.14(3'). [Translate x_2.]

5.14.3

Apply 5.14.1 to describe the maximum likelihood estimator for the other examples discussed in 3.12.2.

5.15.1

Let $X_1,\ldots,X_n$ be i.i.d. with distribution p_θ from a canonical exponential family. Let $K \subset N^\circ$ be compact. Then $\bar{x}_n$ is uniformly asymptotically normal over $\theta \in K$ with mean $\xi(\theta)$ and covariance matrix $n^{-1}\mathcal{I}(\theta) = n^{-1}D_2\psi(\theta)$. [Apply Theorem 2.19.]

5.15.2

Consider the setting of 5.15.1: (a) The maximum likelihood estimators $\hat{\theta}_n$ and $\hat{\xi}_n$ exist with probability approaching 1 as $n \to \infty$ uniformly over $\theta \in K$. (b) They are asymptotically normal uniformly over $\theta \in K$ with means θ and ξ and covariances $n^{-1}\Sigma^{-1}(\theta)$ and $n^{-1}\Sigma(\theta)$, respectively. [(a) $P(\bar{x}_n \notin \mathcal{R})$ converges to 0(exponentially fast), uniformly on K. (b) if $g(t) = g(t_0) + (h(t_0))'(t - t_0) + o(||t - t_0||)$ then $g(\bar{x}_n)$ is asymptotically normal with mean $\xi(\theta)$ covariance $h'(\xi(\theta))\, \Sigma(\theta)h(\xi(\theta))$, uniformly for $\theta \in K$.]

5.15.3

Let $X_1,\ldots,X_n$ be i.i.d. with distribution p from a differentiable exponential subfamily $\{p_\theta: \theta \in \Theta\}$. Let $K \subset \Theta$ be compact. (a) Then $\hat{\theta}_n$ is uniformly asymptotically normal over $\theta \in K$ with mean θ. (b) For a curved exponential family with $\theta = \theta(t)$ the maximum likelihood estimator $\hat{t}_n$ of t is uniformly asymptotically normal at $\theta(t) \in K$ with mean t. (c) Write the asymptotic variance of $\hat{t}_n$ as a function of $\Sigma(\theta(t))$, $\theta'(t)$, and the statistical curvature at t of the curved exponential family. [See Theorem 5.12, the hint to 5.15.2(b), and Section 3.11. For (c), and for a geometric interpretation of (a) and (b) note that $\sqrt{n}||\hat{\xi}_n - \hat{\hat{\xi}}_n|| \to 0$ in probability where $\hat{\hat{\xi}}_n$ denotes the projection in the inner product $\langle s, t\rangle = s'\,\Sigma^{-1}(\theta)t$ of $\bar{x}_n$ on the tangent line at θ to Θ. If the problem is written in the canonical form of Section 3.11 the asymptotic variance is I.]

5.15.4

Let $\{p_\theta: \theta \in \Theta\}$ be a curved exponential family. Let $\theta' \in \mathcal{N}$ but $\theta' \notin \Theta$. Assume (w.l.o.g.) that the family has been written in the canonical form 3.11(1) - (4) with $0 = \hat{\xi}_\Theta(\theta') = \hat{\theta}_\Theta(\xi(\theta'))$. Show $\theta' = (0,\alpha,\ldots,0)$ with $\alpha \leq \rho$. Let $X_1,\ldots,X_n$ be i.i.d. observations under θ' from this family and let $\hat{t}$ be the maximum likelihood estimator of t. Show that if $\alpha<\rho$ then t is asymptotically normal about 0 with variance $\sigma_{11}(\theta')\rho^2/(\rho - \alpha)^2$. What happens when $\alpha = \rho$?