Chapter 2

METHODS OF ESTIMATION

2.1 BASIC HEURISTICS OF ESTIMATION

2.1.1 Minimum Contrast Estimates; Estimating Equations

Our basic framework is as before, that is, $X \in \mathcal{X}$, $X \sim P \in \mathcal{P}$, usually parametrized as $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. See Section 1.1.2. In this parametric case, how do we select reasonable estimates for $\boldsymbol{\theta}$ itself? That is, how do we find a function $\widehat{\boldsymbol{\theta}}(X)$ of the vector observation X that in some sense "is close" to the unknown $\boldsymbol{\theta}$? The fundamental heuristic is typically the following. Consider a function

$$\rho : \mathcal{X} \times \Theta \to R$$

and define

$$D(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \equiv E_{\boldsymbol{\theta}_0} \rho(X, \boldsymbol{\theta}).$$

Suppose that as a function of $\boldsymbol{\theta}$, $D(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ measures the (population) *discrepancy* between $\boldsymbol{\theta}$ and the true value $\boldsymbol{\theta}_0$ of the parameter in the sense that $D(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ is uniquely minimized for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. That is, if $P_{\boldsymbol{\theta}_0}$ were true and we knew $D(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, we could obtain $\boldsymbol{\theta}_0$ as the minimizer. Of course, we don't know the truth so this is inoperable, but $\rho(X, \boldsymbol{\theta})$ is the optimal MSPE predictor of $D(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ (Lemma 1.4.1). So it is natural to consider $\widehat{\boldsymbol{\theta}}(X)$ minimizing $\rho(X, \boldsymbol{\theta})$ as an estimate of $\boldsymbol{\theta}_0$. Under these assumptions we call $\rho(\cdot, \cdot)$ a *contrast function* and $\widehat{\boldsymbol{\theta}}(X)$ a *minimum contrast estimate*.

Now suppose Θ is Euclidean $\subset R^d$, the true $\boldsymbol{\theta}_0$ is an interior point of Θ, and $\boldsymbol{\theta} \to D(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ is smooth. Then we expect

$$\nabla_{\boldsymbol{\theta}} D(\boldsymbol{\theta}_0, \boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 0 \tag{2.1.1}$$

where ∇ denotes the *gradient*,

$$\nabla_{\boldsymbol{\theta}} = \left(\frac{\partial}{\partial \theta_1}, \ldots, \frac{\partial}{\partial \theta_d} \right)^T.$$

Arguing heuristically again we are led to estimates $\widehat{\boldsymbol{\theta}}$ that solve

$$\nabla_{\boldsymbol{\theta}} \rho(X, \boldsymbol{\theta}) = \mathbf{0}. \tag{2.1.2}$$

The equations (2.1.2) define a special form of *estimating equations*.

More generally, suppose we are given a function $\boldsymbol{\Psi} : \mathcal{X} \times R^d \to R^d$, $\boldsymbol{\Psi} \equiv (\psi_1, \ldots, \psi_d)^T$ and define

$$\mathbf{V}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0} \boldsymbol{\Psi}(X, \boldsymbol{\theta}). \tag{2.1.3}$$

Suppose $\mathbf{V}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \mathbf{0}$ has $\boldsymbol{\theta}_0$ as its unique solution for all $\boldsymbol{\theta}_0 \in \Theta$. Then we say $\widehat{\boldsymbol{\theta}}$ solving

$$\boldsymbol{\Psi}(X, \widehat{\boldsymbol{\theta}}) = \mathbf{0} \tag{2.1.4}$$

is an *estimating equation estimate*. Evidently, there is a substantial overlap between the two classes of estimates. Here is an example to be pursued later.

Example 2.1.1. *Least Squares.* Consider the parametric version of the regression model of Example 1.1.4 with $\mu(\mathbf{z}) = g(\boldsymbol{\beta}, \mathbf{z})$, $\boldsymbol{\beta} \in R^d$, where the function g is known. Here the data are $X = \{(\mathbf{z}_i, Y_i) : 1 \leq i \leq n\}$ where $Y_1, \ldots, Y_n$ are independent. A natural[(1)] function $\rho(X, \boldsymbol{\beta})$ to consider is the squared Euclidean distance between the vector $\mathbf{Y}$ of observed Y_i and the vector expectation of $\mathbf{Y}$, $\boldsymbol{\mu}(\mathbf{z}) \equiv (g(\boldsymbol{\beta}, \mathbf{z}_1), \ldots, g(\boldsymbol{\beta}, \mathbf{z}_n))^T$. That is, we take

$$\rho(X, \boldsymbol{\beta}) = |\mathbf{Y} - \boldsymbol{\mu}|^2 = \sum_{i=1}^{n} [Y_i - g(\boldsymbol{\beta}, \mathbf{z}_i)]^2. \tag{2.1.5}$$

Strictly speaking $\mathcal{P}$ is not fully defined here and this is a point we shall explore later. But, for convenience, suppose we postulate that the ϵ_i of Example 1.1.4 are i.i.d. $\mathcal{N}(0, \sigma_0^2)$. Then $\boldsymbol{\beta}$ parametrizes the model and we can compute (see Problem 2.1.16),

$$\begin{aligned} D(\boldsymbol{\beta}_0, \boldsymbol{\beta}) &= E_{\boldsymbol{\beta}_0} \rho(X, \boldsymbol{\beta}) \\ &= n\sigma_0^2 + \sum_{i=1}^{n} [g(\boldsymbol{\beta}_0, \mathbf{z}_i) - g(\boldsymbol{\beta}, \mathbf{z}_i)]^2, \end{aligned} \tag{2.1.6}$$

which is indeed minimized at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and uniquely so if and only if the parametrization is identifiable. An estimate $\widehat{\boldsymbol{\beta}}$ that minimizes $\rho(X, \boldsymbol{\beta})$ exists if $g(\boldsymbol{\beta}, \mathbf{z})$ is continuous in $\boldsymbol{\beta}$ and

$$\lim\{|g(\boldsymbol{\beta}, \mathbf{z})| : |\boldsymbol{\beta}| \to \infty\} = \infty$$

(Problem 2.1.10). The estimate $\widehat{\boldsymbol{\beta}}$ is called *the least squares estimate.*

If, further, $g(\boldsymbol{\beta}, \mathbf{z})$ is differentiable in $\boldsymbol{\beta}$, then $\widehat{\boldsymbol{\beta}}$ satisfies the equation (2.1.2) or equivalently the system of estimating equations,

$$\sum_{i=1}^{n} \frac{\partial g}{\partial \beta_j}(\widehat{\boldsymbol{\beta}}, \mathbf{z}_i) Y_i = \sum_{i=1}^{n} \frac{\partial g}{\partial \beta_j}(\widehat{\boldsymbol{\beta}}, \mathbf{z}_i) g(\widehat{\boldsymbol{\beta}}, \mathbf{z}_i),\ 1 \leq j \leq d. \tag{2.1.7}$$

In the important linear case,

$$g(\boldsymbol{\beta}, \mathbf{z}_i) = \sum_{j=1}^{d} z_{ij} \beta_j \text{ and } \mathbf{z}_i = (z_{i1}, \ldots, z_{id})^T$$

the system becomes

$$\sum_{i=1}^{n} z_{ij} Y_i = \sum_{k=1}^{d} \left(\sum_{i=1}^{n} z_{ij} z_{ik} \right) \widehat{\beta}_k, \tag{2.1.8}$$

the *normal equations.* These equations are commonly written in matrix form

$$\mathbf{Z}_D^T \mathbf{Y} = \mathbf{Z}_D^T \mathbf{Z}_D \boldsymbol{\beta} \tag{2.1.9}$$

where $\mathbf{Z}_D \equiv \|z_{ij}\|_{n \times d}$ is the *design matrix.* Least squares, thus, provides a first example of both minimum contrast and estimating equation methods.

We return to the remark that this estimating method is well defined even if the ϵ_i are not i.i.d. $\mathcal{N}(0, \sigma_0^2)$. In fact, once defined we have a method of computing a statistic $\widehat{\boldsymbol{\beta}}$ from the data $X = \{(\mathbf{z}_i, Y_i),\ 1 \leq i \leq n\}$, which can be judged on its merits whatever the true P governing X is. This very important example is pursued further in Section 2.2 and Chapter 6. □

Here is another basic estimating equation example.

Example 2.1.2. *Method of Moments (MOM).* Suppose $X_1, \ldots, X_n$ are i.i.d. as $X \sim P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in R^d$ and $\boldsymbol{\theta}$ is identifiable. Suppose that $\mu_1(\boldsymbol{\theta}), \ldots, \mu_d(\boldsymbol{\theta})$ are the first d moments of the population we are sampling from. Thus, we assume the existence of

$$\mu_j(\boldsymbol{\theta}) = \mu_j = E_{\boldsymbol{\theta}}(X^j),\ 1 \leq j \leq d.$$

Define the jth *sample moment* $\widehat{\mu}_j$ by,

$$\widehat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j,\ 1 \leq j \leq d.$$

To apply the method of moments to the problem of estimating $\boldsymbol{\theta}$, we need to be able to express $\boldsymbol{\theta}$ as a continuous function g of the first d moments. Thus, suppose

$$\boldsymbol{\theta} \to (\mu_1(\boldsymbol{\theta}), \ldots, \mu_d(\boldsymbol{\theta}))$$

is $1 - 1$ from R^d to R^d. The *method of moments* prescribes that we estimate $\boldsymbol{\theta}$ by the solution of

$$\widehat{\mu}_j = \mu_j(\widehat{\boldsymbol{\theta}}),\ 1 \leq j \leq d$$

if it exists. The motivation of this simplest estimating equation example is the law of large numbers: For $X \sim P_\theta$, $\widehat{\mu}_j$ converges in probability to $\mu_j(\theta)$.

More generally, if we want to estimate a R^k-valued function $q(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$, we obtain a MOM estimate of $q(\boldsymbol{\theta})$ by expressing $q(\boldsymbol{\theta})$ as a function of any of the first d moments $\mu_1, \ldots, \mu_d$ of X, say $q(\boldsymbol{\theta}) = h(\mu_1, \ldots, \mu_d)$, $d \geq k$, and then using $h(\widehat{\mu}_1, \ldots, \widehat{\mu}_d)$ as the estimate of $q(\boldsymbol{\theta})$.

For instance, consider a study in which the survival time X is modeled to have a gamma distribution, $\Gamma(\alpha, \lambda)$, with density

$$[\lambda^\alpha/\Gamma(\alpha)]x^{\alpha-1}\exp\{-\lambda x\},\ x > 0;\ \alpha > 0,\ \lambda > 0.$$

In this case $\boldsymbol{\theta} = (\alpha, \lambda)$, $\mu_1 = E(X) = \alpha/\lambda$, and $\mu_2 = E(X^2) = \alpha(1+\alpha)/\lambda^2$. Solving for $\boldsymbol{\theta}$ gives

$$\begin{aligned} \alpha &= (\mu_1/\sigma)^2, \quad \widehat{\alpha} = (\bar{X}/\widehat{\sigma})^2; \\ \lambda &= \mu_1/\sigma^2, \qquad \widehat{\lambda} = \bar{X}/\widehat{\sigma}^2 \end{aligned}$$

where $\sigma^2 = \mu_2 - \mu_1^2$ and $\widehat{\sigma}^2 = n^{-1}\Sigma X_i^2 - \bar{X}^2$. In this example, the method of moment estimator is not unique. We can, for instance, express θ as a function of μ_1 and $\mu_3 = E(X^3)$ and obtain a method of moment estimator based on $\widehat{\mu}_1$ and $\widehat{\mu}_3$ (Problem 2.1.11). □

Algorithmic issues

We note that, in general, neither minimum contrast estimates nor estimating equation solutions can be obtained in closed form. There are many algorithms for optimization and root finding that can be employed. An algorithm for estimating equations frequently used when computation of $M(X,\cdot) \equiv D\boldsymbol{\Psi}(X,\cdot) \equiv \left\| \frac{\partial \psi_i}{\partial \theta_j}(X,\cdot) \right\|_{d\times d}$ is quick and M is nonsingular with high probability is the *Newton–Raphson algorithm*. It is defined by initializing with $\boldsymbol{\theta}_0$, then setting

$$\widehat{\boldsymbol{\theta}}_{j+1} = \widehat{\boldsymbol{\theta}}_j - [M(X, \widehat{\boldsymbol{\theta}}_j)]^{-1}\boldsymbol{\Psi}(X, \widehat{\boldsymbol{\theta}}_j). \tag{2.1.10}$$

This algorithm and others will be discussed more extensively in Section 2.4 and in Chapter 6, in particular Problem 6.6.10.

2.1.2 The Plug-In and Extension Principles

We can view the method of moments as an example of what we call the *plug-in* (or *substitution*) and *extension* principles, two other basic heuristics particularly applicable in the i.i.d. case. We introduce these principles in the context of multinomial trials and then abstract them and relate them to the method of moments.

Example 2.1.3. *Frequency Plug-in*$^{(2)}$ *and Extension.* Suppose we observe multinomial trials in which the values $v_1, \ldots, v_k$ of the population being sampled are known, but their respective probabilities $p_1, \ldots, p_k$ are completely unknown. If we let $X_1, \ldots, X_n$ be i.i.d. as X and

$$N_i \equiv \text{number of indices } j \text{ such that } X_j = v_i,$$

then the natural estimate of $p_i = P[X = v_i]$ suggested by the law of large numbers is N_i/n, the proportion of sample values equal to v_i. As an illustration consider a population of men whose occupations fall in one of five different job categories, 1, 2, 3, 4, or 5. Here $k = 5$, $v_i = i$, $i = 1, \ldots, 5$, p_i is the proportion of men in the population in the ith job category and N_i/n is the sample proportion in this category. Here is some job category data (Mosteller, 1968).

Job Category

i	1	2	3	4	5	
N_i	23	84	289	217	95	$n = \sum_{i=1}^{5} N_i = 708$
$\widehat{p}_i$	0.03	0.12	0.41	0.31	0.13	$\sum_{i=1}^{5} \widehat{p}_i = 1$

for Danish men whose fathers were in category 3, together with the estimates $\widehat{p}_i = N_i/n$.

Next consider the more general problem of estimating a continuous function $q(p_1, \ldots, p_k)$ of the population proportions. The *frequency plug-in principle* simply proposes to replace the unknown population frequencies $p_1, \ldots, p_k$ by the observable sample frequencies $N_1/n, \ldots, N_k/n$. That is, use

$$T(X_1, \ldots, X_n) = q\left(\frac{N_1}{n}, \ldots, \frac{N_k}{n}\right) \tag{2.1.11}$$

to estimate $q(p_1, \ldots, p_k)$. For instance, suppose that in the previous job category table, categories 4 and 5 correspond to blue-collar jobs, whereas categories 2 and 3 correspond to white-collar jobs. We would be interested in estimating

$$q(p_1, \ldots, p_5) = (p_4 + p_5) - (p_2 + p_3),$$

the difference in the proportions of blue-collar and white-collar workers. If we use the frequency substitution principle, the estimate is

$$T(X_1, \ldots, X_n) = \left(\frac{N_4}{n} + \frac{N_5}{n}\right) - \left(\frac{N_2}{n} + \frac{N_3}{n}\right),$$

which in our case is $0.44 - 0.53 = -0.09$.

Equivalently, let P denote $\mathbf{p} = (p_1, \ldots, p_k)$ with $p_i = P[X = v_i]$, $1 \leq i \leq k$, and think of this model as $\mathcal{P} = \{\text{all probability distributions } P \text{ on } \{v_1, \ldots, v_k\}\}$. Then $q(\mathbf{p})$ can be identified with a parameter $\nu : \mathcal{P} \to R$, that is, $\nu(P) = (p_4 + p_5) - (p_2 + p_3)$, and the frequency plug-in principle simply says to replace $P = (p_1, \ldots, p_k)$ in $\nu(P)$ by $\widehat{P} = \left(\frac{N_1}{n}, \ldots, \frac{N_k}{n}\right)$, the multinomial empirical distribution of $X_1, \ldots, X_n$. □

Now suppose that the proportions $p_1, \ldots, p_k$ do not vary freely but are continuous functions of some d-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ and that we want to estimate a component of $\boldsymbol{\theta}$ or more generally a function $q(\boldsymbol{\theta})$. Many of the models arising in the analysis of discrete data discussed in Chapter 6 are of this type.

Example 2.1.4. *Hardy–Weinberg Equilibrium.* Consider a sample from a population in genetic equilibrium with respect to a single gene with two alleles. If we assume the three different genotypes are identifiable, we are led to suppose that there are three types of individuals whose frequencies are given by the so-called *Hardy–Weinberg proportions*

$$p_1 = \theta^2,\ p_2 = 2\theta(1-\theta),\ p_3 = (1-\theta)^2,\ 0 < \theta < 1. \tag{2.1.12}$$

If N_i is the number of individuals of type i in the sample of size n, then (N_1, N_2, N_3) has a multinomial distribution with parameters (n, p_1, p_2, p_3) given by (2.1.12). Suppose

we want to estimate θ, the frequency of one of the alleles. Because $\theta = \sqrt{p_1}$, we can use the principle we have introduced and estimate by $\sqrt{N_1/n}$. Note, however, that we can also write $\theta = 1 - \sqrt{p_3}$ and, thus, $1 - \sqrt{N_3/n}$ is also a plausible estimate of θ. □

In general, suppose that we want to estimate a continuous R^l-valued function q of $\boldsymbol{\theta}$. If $p_1, \ldots, p_k$ are continuous functions of $\boldsymbol{\theta}$, we can usually express $q(\boldsymbol{\theta})$ as a continuous function of $p_1, \ldots, p_k$, that is,

$$q(\boldsymbol{\theta}) = h(p_1(\boldsymbol{\theta}), \ldots, p_k(\boldsymbol{\theta})), \tag{2.1.13}$$

with h defined and continuous on

$$\mathcal{P} = \left\{ (p_1, \ldots, p_k) : p_i \geq 0, \ \sum_{i=1}^{k} p_i = 1 \right\}.$$

Given h we can apply the *extension principle* to estimate $q(\boldsymbol{\theta})$ as,

$$T(X_1, \ldots, X_n) = h\left(\frac{N_1}{n}, \ldots, \frac{N_k}{n} \right). \tag{2.1.14}$$

As we saw in the Hardy–Weinberg case, the representation (2.1.13) and estimate (2.1.14) are not unique. We shall consider in Chapters 3 (Example 3.4.4) and 5 how to choose among such estimates.

We can think of the extension principle alternatively as follows. Let

$$\mathcal{P}_0 = \{P_{\boldsymbol{\theta}} = (p_1(\boldsymbol{\theta}), \ldots, p_k(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \Theta\}$$

be a submodel of $\mathcal{P}$. Now $q(\boldsymbol{\theta})$ can be identified if $\boldsymbol{\theta}$ is identifiable by a parameter $\nu : \mathcal{P}_0 \to R$ given by $\nu(P_{\boldsymbol{\theta}}) = q(\boldsymbol{\theta})$. Then (2.1.13) defines an extension of ν from $\mathcal{P}_0$ to $\mathcal{P}$ via $\bar{\nu} : \mathcal{P} \to R$ where $\bar{\nu}(P) \equiv h(\mathbf{p})$ and $\bar{\nu}(P) = \nu(P)$ for $P \in \mathcal{P}_0$.

The plug-in and extension principles can be abstractly stated as follows:

Plug-in principle. If we have an estimate $\widetilde{P}$ of $P \in \mathcal{P}$ *such that* $\widetilde{P} \in \mathcal{P}$ and $\nu : \mathcal{P} \to \mathcal{T}$ is a parameter, then $\nu(\widetilde{P})$ is the plug-in estimate of ν. In particular, in the i.i.d. case if $\mathcal{P}$ is the space of all distributions of X and $X_1, \ldots, X_n$ are i.i.d. as $X \sim P$, the *empirical distribution* $\widehat{P}$ of X given by

$$\widehat{P}[X \in A] = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \in A) \tag{2.1.15}$$

is, by the law of large numbers, a natural estimate of P and $\nu(\widehat{P})$ is a plug-in estimate of $\nu(P)$ in this nonparametric context. For instance, if X is real and $F(x) = P(X \leq x)$ is the distribution function (d.f.), consider $\nu_\alpha(P) = \frac{1}{2}[F^{-1}(\alpha) + F_U^{-1}(\alpha)]$, where $\alpha \in (0, 1)$ and

$$F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\}, \ F_U^{-1}(\alpha) = \sup\{x : F(x) \leq \alpha\}, \tag{2.1.16}$$

then $\nu_\alpha(P)$ is the αth *population quantile* x_α. Here $x_{\frac{1}{2}} = \nu_{\frac{1}{2}}(P)$ is called the *population median*. A natural estimate is the αth *sample quantile*

$$\widehat{x}_\alpha = \frac{1}{2}[\widehat{F}^{-1}(\alpha) + \widehat{F}_U^{-1}(\alpha)], \tag{2.1.17}$$

where $\widehat{F}$ is the empirical d.f. Here $\widehat{x}_{\frac{1}{2}}$ is called the *sample median*.

For a second example, if X is real and $\mathcal{P}$ is the class of distributions with $E|X|^j < \infty$, then the plug-in estimate of the jth moment $\nu(P) = \mu_j = E(X^j)$ in this nonparametric context is the jth sample moment $\nu(\widehat{P}) = \int x^j d\widehat{F}(x) = n^{-1}\sum_{i=1}^n X_i^j$.

Extension principle. Suppose $\mathcal{P}_0$ is a submodel of $\mathcal{P}$ and $\widehat{P}$ is an element of $\mathcal{P}$ but not necessarily $\mathcal{P}_0$ and suppose $\nu : \mathcal{P}_0 \to \mathcal{T}$ is a parameter. If $\bar{\nu} : \mathcal{P} \to \mathcal{T}$ is an extension of ν in the sense that $\bar{\nu}(P) = \nu(P)$ on $\mathcal{P}_0$, then $\bar{\nu}(\widehat{P})$ is an extension (and plug-in) estimate of $\nu(P)$.

With this general statement we can see precisely how method of moment estimates can be obtained as extension and frequency plug-in estimates for multinomial trials because

$$\mu_j(\boldsymbol{\theta}) = \sum_{i=1}^k v_i^j p_i(\boldsymbol{\theta}) = h(\mathbf{p}(\boldsymbol{\theta})) = \nu(P_\theta)$$

where

$$h(\mathbf{p}) = \sum_{i=1}^k v_i^j p_i = \bar{\nu}(P),$$

$$\widehat{\mu}_j = \frac{1}{n}\sum_{l=1}^n X_l^j = \sum_{i=1}^k v_i^j \frac{N_i}{n} = h\left(\frac{\mathbf{N}}{n}\right) = \bar{\nu}(\widehat{P})$$

and $\widehat{P}$ is the empirical distribution. This reasoning extends to the general i.i.d. case (Problem 2.1.12) and to more general method of moment estimates (Problem 2.1.13). As stated, these principles are general. However, they are mainly applied in the i.i.d. case—but see Problem 2.1.14.

Remark 2.1.1. The plug-in and extension principles are used when P_θ, ν, and $\bar{\nu}$ are continuous. For instance, in the multinomial examples 2.1.3 and 2.1.4, P_θ as given by the Hardy–Weinberg $\mathbf{p}(\theta)$, is a continuous map from $\Theta = [0,1]$ to $\mathcal{P}$, $\nu(P_{\boldsymbol{\theta}}) = q(\boldsymbol{\theta}) = h(\mathbf{p}(\boldsymbol{\theta}))$ is a continuous map from Θ to R and $\bar{\nu}(P) = h(\mathbf{p})$ is a continuous map from $\mathcal{P}$ to R.

Remark 2.1.2. The plug-in and extension principles must be calibrated with the target parameter. For instance, let $\mathcal{P}_0$ be the class of distributions of $X = \theta + \epsilon$ where $\theta \in R$ and the distribution of ϵ ranges over the class of symmetric distributions with mean zero. Let $\nu(P)$ be the mean of X and let $\mathcal{P}$ be the class of distributions of $X = \theta + \epsilon$ where $\theta \in R$ and the distribution of ϵ ranges over the class of distributions with mean zero. In this case both $\bar{\nu}_1(P) = E_P(X)$ and $\bar{\nu}_2(P) =$ "median of P" satisfy $\bar{\nu}(P) = \nu(P)$, $P \in \mathcal{P}_0$, but only $\bar{\nu}_1(\widehat{P}) = \bar{X}$ is a sensible estimate of $\nu(P)$, $P \notin \mathcal{P}_0$, because when P is not symmetric, the sample median $\bar{\nu}_2(\widehat{P})$ does not converge in probability to $E_P(X)$.

Here are three further simple examples illustrating reasonable and unreasonable MOM estimates.

Example 2.1.5. Suppose that $X_1, \ldots, X_n$ is a $\mathcal{N}(\mu, \sigma^2)$ sample as in Example 1.1.2 with assumptions (1)–(4) holding. The method of moments estimates of μ and σ^2 are $\bar{X}$ and $\widehat{\sigma}^2$. □

Example 2.1.6. Suppose $X_1, \ldots, X_n$ are the indicators of a set of Bernoulli trials with probability of success θ. Because $\mu_1(\theta) = \theta$ the method of moments leads to the natural estimate of θ, $\bar{X}$, the frequency of successes. To estimate the population variance $\theta(1-\theta)$ we are led by the first moment to the estimate, $\bar{X}(1-\bar{X})$. Because we are dealing with (unrestricted) Bernoulli trials, these are *the* frequency plug-in (substitution) estimates (see Problem 2.1.1). □

Example 2.1.7. *Estimating the Size of a Population (continued).* In Example 1.5.3 where $X_1, \ldots, X_n$ are i.i.d. $\mathcal{U}\{1, 2, \ldots, \theta\}$, we find $\mu = E_\theta(X_i) = \frac{1}{2}(\theta + 1)$. Thus, $\theta = 2\mu - 1$ and $2\bar{X} - 1$ is a method of moments estimate of θ. This is clearly a foolish estimate if $X_{(n)} = \max X_i > 2\bar{X} - 1$ because in this model θ is always at least as large as $X_{(n)}$. □

As we have seen, there are often several method of moments estimates for the same $q(\boldsymbol{\theta})$. For example, if we are sampling from a Poisson population with parameter θ, then θ is both the population mean and the population variance. The method of moments can lead to either the sample mean or the sample variance. Moreover, because $p_0 = P(X = 0) = \exp\{-\theta\}$, a frequency plug-in estimate of θ is $-\log \widehat{p}_0$, where $\widehat{p}_0$ is $n^{-1}[\#X_i = 0]$. We will make a selection among such procedures in Chapter 3.

Remark 2.1.3. What are the good points of the method of moments and frequency plug-in?

(a) They generally lead to procedures that are easy to compute and are, therefore, valuable as preliminary estimates in algorithms that search for more efficient estimates. See Section 2.4.

(b) If the sample size is large, these estimates are likely to be close to the value estimated (consistency). This minimal property is discussed in Section 5.2.

It does turn out that there are "best" frequency plug-in estimates, those obtained by the method of maximum likelihood, a special type of minimum contrast and estimating equation method. Unfortunately, as we shall see in Section 2.4, they are often difficult to compute. Algorithms for their computation will be introduced in Section 2.4.

Discussion. When we consider optimality principles, we may arrive at different types of estimates than those discussed in this section. For instance, as we shall see in Chapter 3, estimation of θ real with quadratic loss and Bayes priors lead to procedures that are data weighted averages of θ values rather than minimizers of functions $\rho(\theta, X)$. Plug-in is not the optimal way to go for the Bayes, minimax, or uniformly minimum variance unbiased (UMVU) principles we discuss briefly in Chapter 3. However, a saving grace becomes apparent in Chapters 5 and 6. If the model fits, for large amounts of data, optimality principle solutions agree to first order with the best minimum contrast and estimating equation solutions, the plug-in principle is justified, and there are best extensions.

Summary. We consider principles that suggest how we can use the outcome X of an experiment to estimate unknown parameters.

For the model $\{P_\theta : \theta \in \Theta\}$ a *contrast* ρ is a function from $\mathcal{X} \times \Theta$ to R such that the *discrepancy*

$$D(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0} \rho(X, \boldsymbol{\theta}),\ \theta \in \Theta \subset R^d$$

is uniquely minimized at the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ of the parameter. A *minimum contrast estimator* is a minimizer of $\rho(\mathbf{X}, \boldsymbol{\theta})$, and the *contrast estimating equations* are $\nabla_{\boldsymbol{\theta}} \rho(X, \boldsymbol{\theta}) = 0$.

For data $\{(\mathbf{z}_i, Y_i) : 1 \leq i \leq n\}$ with Y_i independent and $E(Y_i) = g(\boldsymbol{\beta}, \mathbf{z}_i)$, $1 \leq i \leq n$, where g is a known function and $\boldsymbol{\beta} \in R^d$ is a vector of unknown regression coefficients, a *least squares estimate* of $\boldsymbol{\beta}$ is a minimizer of

$$\rho(X, \boldsymbol{\beta}) = \sum [Y_i - g(\boldsymbol{\beta}, \mathbf{z}_i)]^2.$$

For this contrast, when $g(\boldsymbol{\beta}, \mathbf{z}) = \mathbf{z}^T \boldsymbol{\beta}$, the associated estimating equations are called the *normal equations* and are given by $\mathbf{Z}_D^T \mathbf{Y} = \mathbf{Z}_D^T \mathbf{Z}_D \boldsymbol{\beta}$, where $\mathbf{Z}_D = \|z_{ij}\|_{n \times d}$ is called the *design matrix*. Suppose $X \sim P$. The *plug-in estimate* (PIE) for a vector parameter $\nu = \nu(P)$ is obtained by setting $\widehat{\nu} = \nu(\widehat{P})$ where $\widehat{P}$ is an estimate of P. When $\widehat{P}$ is the *empirical probability distribution* $\widehat{P}_E$ defined by $\widehat{P}_E(A) = n^{-1} \sum_{i=1}^n 1[X_i \in A]$, then $\widehat{\nu}$ is called the *empirical* PIE. If $P = P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$, is parametric and a vector $q(\boldsymbol{\theta})$ is to be estimated, we find a parameter ν such that $\nu(P_\theta) = q(\theta)$ and call $\nu(\widehat{P})$ a plug-in estimator of $q(\boldsymbol{\theta})$. Method of moment estimates are empirical PIEs based on $\nu(P) = (\mu_1, \ldots, \mu_d)^T$ where $\mu_j = E(X^j)$, $1 \leq j \leq d$. In the multinomial case the *frequency plug-in* estimators are empirical PIEs based on $\nu(P) = (p_1, \ldots, p_k)$, where p_j is the probability of the jth category, $1 \leq j \leq k$.

Let $\mathcal{P}_0$ and $\mathcal{P}$ be two statistical models for X with $\mathcal{P}_0 \subset \mathcal{P}$. An *extension* $\bar{\nu}$ of ν from $\mathcal{P}_0$ to $\mathcal{P}$ is a parameter satisfying $\bar{\nu}(P) = \nu(P)$, $P \in \mathcal{P}_0$. If $\widehat{P}$ is an estimate of P with $\widehat{P} \in \mathcal{P}$, $\bar{\nu}(\widehat{P})$ is called the *extension plug-in estimate* of $\nu(P)$. The general principles are shown to be related to each other.

2.2 MINIMUM CONTRAST ESTIMATES AND ESTIMATING EQUATIONS

2.2.1 Least Squares and Weighted Least Squares

Least squares[(1)] was advanced early in the nineteenth century by Gauss and Legendre for estimation in problems of astronomical measurement. It is of great importance in many areas of statistics such as the analysis of variance and regression theory. In this section we shall introduce the approach and give a few examples leaving detailed development to Chapter 6.

In Example 2.1.1 we considered the nonlinear (and linear) Gaussian model $\mathcal{P}_0$ given by

$$Y_i = g(\boldsymbol{\beta}, \mathbf{z}_i) + \epsilon_i,\ 1 \le i \le n \tag{2.2.1}$$

where ϵ_i are i.i.d. $\mathcal{N}(0, \sigma_0^2)$ and $\boldsymbol{\beta}$ ranges over R^d or an open subset. The contrast

$$\rho(X, \boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i - g(\boldsymbol{\beta}, \mathbf{z}_i)]^2$$

led to the least squares estimates (LSEs) $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Suppose that we enlarge $\mathcal{P}_0$ to $\mathcal{P}$ where we retain the independence of the Y_i but only require

$$\mu(\mathbf{z}_i) = E(Y_i) = g(\boldsymbol{\beta}, \mathbf{z}_i),\ E(\epsilon_i) = 0,\ \mathrm{Var}(\epsilon_i) < \infty. \tag{2.2.2}$$

Are least squares estimates still reasonable? For P in the semiparametric model $\mathcal{P}$, which satisfies (but is not fully specified by) (2.2.2), we can compute still

$$\begin{aligned} D_P(\boldsymbol{\beta}_0, \boldsymbol{\beta}) &= E_P \rho(X, \boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \mathrm{Var}_P(\epsilon_i) + \sum_{i=1}^{n} [g(\boldsymbol{\beta}_0, \mathbf{z}_i) - g(\boldsymbol{\beta}, \mathbf{z}_i)]^2, \end{aligned} \tag{2.2.3}$$

which is again minimized as a function of $\boldsymbol{\beta}$ by $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and uniquely so if the map $\boldsymbol{\beta} \to (g(\boldsymbol{\beta}, \mathbf{z}_1), \ldots, g(\boldsymbol{\beta}, \mathbf{z}_n))^T$ is $1-1$.

The estimates continue to be reasonable under the *Gauss–Markov* assumptions,

$$E(\epsilon_i) = 0,\ 1 \le i \le n, \tag{2.2.4}$$

$$\mathrm{Var}(\epsilon_i) = \sigma^2 > 0,\ 1 \le i \le n, \tag{2.2.5}$$

$$\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0,\ 1 \le i < j \le n \tag{2.2.6}$$

because (2.2.3) continues to be valid.

Note that the joint distribution H of $(\epsilon_1, \ldots, \epsilon_n)$ is any distribution satisfying the Gauss–Markov assumptions. That is, the model is semiparametric with $\boldsymbol{\beta}, \sigma^2$ and H unknown. The least squares method of estimation applies only to the parameters $\beta_1, \ldots, \beta_d$ and is often applied in situations in which specification of the model beyond (2.2.4)–(2.2.6) is difficult.

Sometimes $\mathbf{z}$ can be viewed as the realization of a population variable $\mathbf{Z}$, that is, $(\mathbf{Z}_i, Y_i)$, $1 \le i \le n$, is modeled as a sample from a joint distribution. This is frequently the case for studies in the social and biological sciences. For instance, Z could be educational level and Y income, or Z could be height and Y log weight. Then we can write the conditional model of Y given $\mathbf{Z}_j = \mathbf{z}_j$, $1 \le j \le n$, as in (a) of Example 1.1.4 with ϵ_j simply defined as $Y_j - E(Y_j \mid \mathbf{Z}_j = \mathbf{z}_j)$. If we consider this model, $(\mathbf{Z}_i, Y_i)$ i.i.d. as $(\mathbf{Z}, Y) \sim P \in \mathcal{P} = \{$All joint distributions of $(\mathbf{Z}, Y)$ such that $E(Y \mid \mathbf{Z} = \mathbf{z}) = g(\boldsymbol{\beta}, \mathbf{z})$, $\boldsymbol{\beta} \in R^d\}$, and $\boldsymbol{\beta} \to (g(\boldsymbol{\beta}, \mathbf{z}_1), \ldots, g(\boldsymbol{\beta}, \mathbf{z}_n))^T$ is $1-1$, then $\boldsymbol{\beta}$ has an interpretation as a parameter on $\mathcal{P}$, that is, $\boldsymbol{\beta} = \boldsymbol{\beta}(P)$ is the minimizer of $E(Y - g(\boldsymbol{\beta}, \mathbf{Z}))^2$. This follows

from Theorem 1.4.1. In this case we recognize the LSE $\widehat{\boldsymbol{\beta}}$ as simply being the usual plug-in estimate $\boldsymbol{\beta}(\widehat{P})$, where $\widehat{P}$ is the empirical distribution assigning mass n^{-1} to each of the n pairs $(\mathbf{Z}_i, Y_i)$.

As we noted in Example 2.1.1 the most commonly used g in these models is $g(\boldsymbol{\beta}, \mathbf{z}) = \mathbf{z}^T\boldsymbol{\beta}$, which, in conjunction with (2.2.1), (2.2.4), (2.2.5) and (2.2.6), is called the *linear* (multiple) *regression model.* For the data $\{(\mathbf{z}_i, Y_i);\ i = 1, \ldots, n\}$ we write this model in matrix form as

$$\mathbf{Y} = \mathbf{Z}_D\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{Z}_D = \|\mathbf{z}_{ij}\|$ is the design matrix. We continue our discussion for this important special case for which explicit formulae and theory have been derived. For nonlinear cases we can use numerical methods to solve the estimating equations (2.1.7). See also Problem 2.2.41, Sections 6.4.3 and 6.5, and Seber and Wild (1989). The linear model is often the default model for a number of reasons:

(1) If the range of the $\mathbf{z}$'s is relatively small and $\mu(\mathbf{z})$ is smooth, we can approximate $\mu(\mathbf{z})$ by

$$\mu(\mathbf{z}) = \mu(\mathbf{z}_0) + \sum_{j=1}^{d} \frac{\partial \mu}{\partial z_j}(\mathbf{z}_0)(z_j - z_{0j}),$$

for $\mathbf{z}_0 = (z_{01}, \ldots, z_{0d})^T$ an interior point of the domain. We can then treat $\mu(\mathbf{z}_0) - \sum_{j=1}^{d} \frac{\partial \mu}{\partial z_j}(\mathbf{z}_0) z_{0j}$ as an unknown β_0 and identify $\frac{\partial \mu}{\partial z_j}(\mathbf{z}_0)$ with β_j to give an approximate $(d+1)$-dimensional linear model with $z_0 \equiv 1$ and z_j as before and

$$\mu(\mathbf{z}) = \sum_{j=0}^{d} \beta_j z_j.$$

This type of approximation is the basis for nonlinear regression analysis based on local polynomials, see Ruppert and Wand (1994), Fan and Gijbels (1996), and Volume II.

(2) If as we discussed earlier, we are in a situation in which it is plausible to assume that $(\mathbf{Z}_i, Y_i)$ are a sample from a $(d+1)$-dimensional distribution and the covariates that are the coordinates of $\mathbf{Z}$ are continuous, a further modeling step is often taken and it is assumed that $(Z_1, \ldots, Z_d, Y)^T$ has a nondegenerate multivariate Gaussian distribution $\mathcal{N}_{d+1}(\boldsymbol{\mu}, \Sigma)$. In that case, as we have seen in Section 1.4, $E(Y \mid \mathbf{Z} = \mathbf{z})$ can be written as

$$\mu(\mathbf{z}) = \beta_0 + \sum_{j=1}^{d} \beta_j z_j \tag{2.2.7}$$

where

$$(\beta_1, \ldots, \beta_d)^T = \Sigma_{\mathbf{ZZ}}^{-1}\Sigma_{\mathbf{Z}Y} \tag{2.2.8}$$

$$\beta_0 = (-(\beta_1, \ldots, \beta_d), 1)\boldsymbol{\mu} = \mu_Y - \sum_{j=1}^{d} \beta_j \mu_j. \tag{2.2.9}$$

Furthermore,

$$\epsilon \equiv Y - \mu(\mathbf{Z})$$

is independent of $\mathbf{Z}$ and has a $\mathcal{N}(0, \sigma^2)$ distribution where

$$\sigma^2 = \sigma_{YY} - \Sigma_{Y\mathbf{Z}}\Sigma_{\mathbf{ZZ}}^{-1}\Sigma_{\mathbf{Z}Y}.$$

Therefore, given $\mathbf{Z}_i = \mathbf{z}_i$, $1 \leq i \leq n$, we have a Gaussian linear regression model for Y_i, $1 \leq i \leq n$.

Estimation of $\boldsymbol{\beta}$ in the linear regression model. We have already argued in Example 2.1.1 that, if the parametrization $\boldsymbol{\beta} \to \mathbf{Z}_D\boldsymbol{\beta}$ is identifiable, the least squares estimate, $\widehat{\boldsymbol{\beta}}$, exists and is unique and satisfies the normal equations (2.1.8). The parametrization is identifiable if and only if $\mathbf{Z}_D$ is of rank d or equivalently if $\mathbf{Z}_D^T\mathbf{Z}_D$ is of full rank d; see Problem 2.2.25. In that case, necessarily, the solution of the normal equations can be given "explicitly" by

$$\widehat{\boldsymbol{\beta}} = [\mathbf{Z}_D^T\mathbf{Z}_D]^{-1}\mathbf{Z}_D^T\mathbf{Y}. \tag{2.2.10}$$

Here are some examples.

Example 2.2.1. In the measurement model in which Y_i is the determination of a constant $\beta_1, d = 1, g(z, \beta_1) = \beta_1$, $\frac{\partial}{\partial\beta_1}g(z, \beta_1) = 1$ and the normal equation is $\sum_{i=1}^n (y_i - \beta_1) = 0$, whose solution is $\widehat{\beta}_1 = (1/n)\sum_{i=1}^n y_i = \bar{y}$. □

Example 2.2.2. We want to find out how increasing the amount z of a certain chemical or fertilizer in the soil increases the amount y of that chemical in the plants grown in that soil. For certain chemicals and plants, the relationship between z and y can be approximated well by a linear equation $y = \beta_1 + \beta_2 z$ provided z is restricted to a reasonably small interval. If we run several experiments with the same z using plants and soils that are as nearly identical as possible, we will find that the values of y will not be the same. For this reason, we assume that for a given z, Y is random with a distribution $P(y \mid z)$.

Following are the results of an experiment to which a regression model can be applied (Snedecor and Cochran, 1967, p. 139). Nine samples of soil were treated with different amounts z of phosphorus. Y is the amount of phosphorus found in corn plants grown for 38 days in the different samples of soil.

z_i	1	4	5	9	11	13	23	23	28
Y_i	64	71	54	81	76	93	77	95	109

The points (z_i, Y_i) and an estimate of the line $\beta_1 + \beta_2 z$ are plotted in Figure 2.2.1.

We want to estimate β_1 and β_2. The normal equations are

$$\sum_{i=1}^n (y_i - \beta_1 - \beta_2 z_i) = 0, \ \sum_{i=1}^n z_i(y_i - \beta_1 - \beta_2 z_i) = 0. \tag{2.2.11}$$

When the z_i's are not all equal, we get the solutions

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^n (z_i - \bar{z})y_i}{\sum_{i=1}^n (z_i - \bar{z})^2} = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})^2} \tag{2.2.12}$$

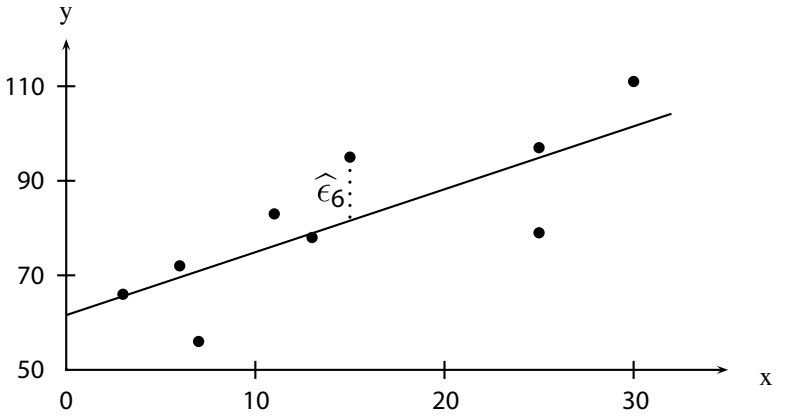

Figure 2.2.1. Scatter plot $\{(z_i, y_i);\ i = 1, \ldots, n\}$ and sample regression line for the phosphorus data. $\widehat{\epsilon}_6$ is the residual for (z_6, y_6).

and

$$\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2 \bar{z} \tag{2.2.13}$$

where $\bar{z} = (1/n)\sum_{i=1}^n z_i$, and $\bar{y} = (1/n)\sum_{i=1}^n y_i$.

The line $y = \widehat{\beta}_1 + \widehat{\beta}_2 z$ is known as the *sample regression line* or *line of best fit* of $y_1, \ldots, y_n$ on $z_1, \ldots, z_n$. Geometrically, if we measure the distance between a point (z_i, y_i) and a line $y = a + bz$ vertically by $d_i = |y_i - (a + bz_i)|$, then the regression line minimizes the sum of the squared distances to the n points $(z_1, y_1), \ldots, (z_n, y_n)$. The vertical distances $\widehat{\epsilon}_i = [y_i - (\widehat{\beta}_1 + \widehat{\beta}_2 z_i)]$ are called the *residuals* of the fit, $i = 1, \ldots, n$. The line $y = \widehat{\beta}_1 + \widehat{\beta}_2 z$ is an estimate of the best linear MSPE predictor $a_1 + b_1 z$ of Theorem 1.4.3. This connection to prediction explains the use of vertical distances in regression. The regression line for the phosphorus data is given in Figure 2.2.1. Here $\widehat{\beta}_1 = 61.58$ and $\widehat{\beta}_2 = 1.42$. □

Remark 2.2.1. The linear regression model is considerably more general than appears at first sight. For instance, suppose we select p real-valued functions of $\mathbf{z}$, $g_1, \ldots, g_p$, $p \geq d$ and postulate that $\mu(\mathbf{z})$ is a linear combination of $g_1(\mathbf{z}), \ldots, g_p(\mathbf{z})$; that is

$$\mu(\mathbf{z}) = \sum_{j=1}^p \theta_j g_j(\mathbf{z}).$$

Then we are still dealing with a linear regression model because we can define $\mathbf{w}_{p\times 1} =$

$(g_1(\mathbf{z}), \ldots, g_p(\mathbf{z}))^T$ as our covariate and consider the linear model

$$Y_i = \sum_{j=1}^{p} \theta_j w_{ij} + \epsilon_i,\ 1 \le i \le n$$

where

$$w_{ij} \equiv g_j(\mathbf{z}_i).$$

For instance, if $d = 1$ and we take $g_j(z) = z^j$, $0 \le j \le 2$, we arrive at quadratic regression, $Y_i = \theta_0 + \theta_1 z_i + \theta_2 z_i^2 + \epsilon_i$—see Problem 2.2.24 for more on *polynomial regression.*

Whether any linear model is appropriate in particular situations is a delicate matter partially explorable through further analysis of the data and knowledge of the subject matter. We return to this in Volume II.

Weighted least squares. In Example 2.2.2 and many similar situations it may not be reasonable to assume that the variances of the errors ϵ_i are the same for all levels z_i of the covariate variable. However, we may be able to characterize the dependence of $\mathrm{Var}(\epsilon_i)$ on z_i at least up to a multiplicative constant. That is, we can write

$$\mathrm{Var}(\epsilon_i) = w_i \sigma^2 \tag{2.2.14}$$

where σ^2 is unknown as before, but the w_i are known weights. Such models are called *heteroscedastic* (as opposed to the equal variance models that are *homoscedastic*). The method of least squares may not be appropriate because (2.2.5) fails. Note that the variables

$$\widetilde{Y}_i \equiv \frac{Y_i}{\sqrt{w_i}} = \frac{g(\boldsymbol{\beta}, \mathbf{z}_i)}{\sqrt{w_i}} + \frac{\epsilon_i}{\sqrt{w_i}},\ 1 \le i \le n,$$

are sufficient for the Y_i and that $\mathrm{Var}(\epsilon_i/\sqrt{w_i}) = w_i\sigma^2/w_i = \sigma^2$. Thus, if we set $\widetilde{g}(\boldsymbol{\beta}, \mathbf{z}_i) = g(\boldsymbol{\beta}, \mathbf{z}_i)/\sqrt{w_i}$, $\widetilde{\epsilon}_i = \epsilon_i/\sqrt{w_i}$, then

$$\widetilde{Y}_i = \widetilde{g}(\boldsymbol{\beta}, \mathbf{z}_i) + \widetilde{\epsilon}_i,\ 1 \le i \le n \tag{2.2.15}$$

and the $\widetilde{Y}_i$ satisfy the assumption (2.2.5). The *weighted least squares estimate* of $\boldsymbol{\beta}$ is now the value $\widehat{\boldsymbol{\beta}}$, which for given $\widetilde{y}_i = y_i/\sqrt{w_i}$ minimizes

$$\sum_{i=1}^{n} [\widetilde{y}_i - \widetilde{g}(\boldsymbol{\beta}, \mathbf{z}_i)]^2 = \sum_{i=1}^{n} \frac{1}{w_i} [y_i - g(\boldsymbol{\beta}, \mathbf{z}_i)]^2 \tag{2.2.16}$$

as a function of $\boldsymbol{\beta}$.

Example 2.2.3. *Weighted Linear Regression.* Consider the case in which $d = 2$, $z_{i1} = 1$, $z_{i2} = z_i$, and $g(\boldsymbol{\beta}, \mathbf{z}_i) = \beta_1 + \beta_2 z_i$, $i = 1, \ldots, n$. We need to find the values $\widehat{\beta}_1$ and $\widehat{\beta}_2$ of β_1 and β_2 that minimize

$$\sum_{i=1}^{n} v_i [y_i - (\beta_1 + \beta_2 z_i)]^2 \tag{2.2.17}$$

where $v_i = 1/w_i$. This problem may be solved by setting up analogues to the normal equations (2.1.8). We can also use the results on prediction in Section 1.4 as follows. Let (Z^*, Y^*) denote a pair of discrete random variables with possible values $(z_1, y_1), \ldots, (z_n, y_n)$ and probability distribution given by

$$P[(Z^*, Y^*) = (z_i, y_i)] = u_i,\ i = 1, \ldots, n$$

where

$$u_i = v_i / \sum_{i=1}^{n} v_i,\ i = 1, \ldots, n.$$

If $\mu_l(Z^*) = \beta_1 + \beta_2 Z^*$ denotes a linear predictor of Y^* based on Z^*, then its MSPE is given by

$$E[Y^* - \mu_l(Z^*)]^2 = \sum_{i=1}^{n} u_i [y_i - (\beta_1 + \beta_2 z_i)]^2.$$

It follows that the problem of minimizing (2.2.17) is equivalent to finding the best linear MSPE predictor of Y^*. Thus, using Theorem 1.4.3,

$$\widehat{\beta}_2 = \frac{\mathrm{Cov}(Z^*, Y^*)}{\mathrm{Var}(Z^*)} = \frac{\sum_{i=1}^n u_i z_i y_i - (\sum_{i=1}^n u_i y_i)(\sum_{i=1}^n u_i z_i)}{\sum_{i=1}^n u_i z_i^2 - (\sum_{i=1}^n u_i z_i)^2} \tag{2.2.18}$$

and

$$\widehat{\beta}_1 = E(Y^*) - \widehat{\beta}_2 E(Z^*) = \sum_{i=1}^{n} u_i y_i - \widehat{\beta}_2 \sum_{i=1}^{n} u_i z_i.$$

This computation suggests, as we make precise in Problem 2.2.26, that weighted least squares estimates are also plug-in estimates. □

Next consider finding the $\widehat{\boldsymbol{\beta}}$ that minimizes (2.2.16) for $g(\boldsymbol{\beta}, \mathbf{z}_i) = \mathbf{z}_i^T \boldsymbol{\beta}$ and for general d. By following the steps of Example 2.1.1 leading to (2.1.7), we find (Problem 2.2.27) that $\widehat{\boldsymbol{\beta}}$ satisfy the *weighted least squares normal equations*

$$\mathbf{Z}_D^T \mathbf{W}^{-1} \mathbf{Y} = (\mathbf{Z}_D^T \mathbf{W}^{-1} \mathbf{Z}_D) \widehat{\boldsymbol{\beta}} \tag{2.2.19}$$

where $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$ and $\mathbf{Z}_D = \|z_{ij}\|_{n \times d}$ is the design matrix. When $\mathbf{Z}_D$ has rank d and $w_i > 0$, $1 \leq i \leq n$, we can write

$$\widehat{\boldsymbol{\beta}} = (\mathbf{Z}_D^T \mathbf{W}^{-1} \mathbf{Z}_D)^{-1} \mathbf{Z}_D^T \mathbf{W}^{-1} \mathbf{Y}. \tag{2.2.20}$$

Remark 2.2.2. More generally, we may allow for correlation between the errors $\{\epsilon_i\}$. That is, suppose $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{W}$ for some invertible matrix $\mathbf{W}_{n \times n}$. Then it can be shown (Problem 2.2.28) that the model $\mathbf{Y} = \mathbf{Z}_D \boldsymbol{\beta} + \boldsymbol{\epsilon}$ can be transformed to one satisfying (2.2.1) and (2.2.4)–(2.2.6). Moreover, when $g(\boldsymbol{\beta}, \mathbf{z}) = \mathbf{z}^T \boldsymbol{\beta}$, the $\widehat{\boldsymbol{\beta}}$ minimizing the least squares contrast in this transformed model is given by (2.2.19) and (2.2.20).

Remark 2.2.3. Here are some applications of weighted least squares: When the ith response Y_i is an average of n_i equally variable observations, then $\text{Var}(Y_i) = \sigma^2/n_i$, and $w_i = n_i^{-1}$. If Y_i is the sum of n_i equally variable observations, then $w_i = n_i$. If the variance of Y_i is proportional to some covariate, say z_1, then $\text{Var}(Y_i) = z_{i1}\sigma^2$ and $w_i = z_{i1}$. In time series and repeated measures, a covariance structure is often specified for $\boldsymbol{\epsilon}$ (see Problems 2.2.29 and 2.2.42).

2.2.2 Maximum Likelihood

The method of maximum likelihood was first proposed by the German mathematician C. F. Gauss in 1821. However, the approach is usually credited to the English statistician R. A. Fisher (1922) who rediscovered the idea and first investigated the properties of the method. In the form we shall give, this approach *makes sense only in regular parametric models.* Suppose that $p(\mathbf{x}, \theta)$ is the frequency or density function of $\mathbf{X}$ if θ is true and that Θ is a subset of d-dimensional space.

Recall $L_{\mathbf{x}}(\theta)$, the likelihood function of θ, defined in Section 1.5, which is just $p(\mathbf{x}, \theta)$ considered as a function of θ for fixed $\mathbf{x}$. Thus, if $\mathbf{X}$ is discrete, then for each θ, $L_{\mathbf{x}}(\theta)$ gives the probability of observing $\mathbf{x}$. If Θ is finite and π is the uniform prior distribution on Θ, then the posterior probability that $\boldsymbol{\theta} = \theta$ given $\mathbf{X} = \mathbf{x}$ satisfies $\pi(\theta \mid \mathbf{x}) \propto L_{\mathbf{x}}(\theta)$, where the proportionality is up to a function of $\mathbf{x}$. Thus, we can think of $L_{\mathbf{x}}(\theta)$ as a measure of how "likely" θ is to have produced the observed $\mathbf{x}$. A similar interpretation applies to the continuous case (see A.7.10).

The *method* of *maximum likelihood* consists of finding that value $\widehat{\theta}(\mathbf{x})$ of the parameter that is "most likely" to have produced the data. That is, if $\mathbf{X} = \mathbf{x}$, we seek $\widehat{\theta}(\mathbf{x})$ that satisfies

$$L_{\mathbf{x}}(\widehat{\theta}(\mathbf{x})) = p(\mathbf{x}, \widehat{\theta}(\mathbf{x})) = \max\{p(\mathbf{x}, \theta) : \theta \in \Theta\} = \max\{L_{\mathbf{x}}(\theta) : \theta \in \Theta\}.$$

By our previous remarks, if Θ is finite and π is uniform, or, more generally, the prior density π on Θ is constant, such a $\widehat{\theta}(\mathbf{x})$ is a mode of the posterior distribution. If such a $\widehat{\theta}$ exists, we estimate any function $q(\theta)$ by $q(\widehat{\theta}(\mathbf{x}))$. The estimate $q(\widehat{\theta}(\mathbf{x}))$ is called the *maximum likelihood estimate* (MLE) of $q(\theta)$. This definition of $q(\widehat{\theta})$ is consistent. That is, suppose $\mathbf{q}$ is 1-1 from Θ to Ω; set $\omega = q(\theta)$ and write the density of $\mathbf{X}$ as $p_0(\mathbf{x}, \omega) = p(\mathbf{x}, q^{-1}(\omega))$. If $\widehat{\omega}$ maximizes $p_0(\mathbf{x}, \omega)$ then $\widehat{\omega} = q(\widehat{\theta})$ (Problem 2.2.16(a)). If q is not 1-1, the MLE of $\omega = q(\theta)$ is still $q(\widehat{\theta})$ (Problem 2.2.16(b)).

Here is a simple numerical example.

Suppose $\theta = 0$ or $\frac{1}{2}$ and $p(x, \theta)$ is given by the following table.

$x \backslash \theta$	0	$\frac{1}{2}$
1	0	0.10
2	1	0.90

Then $\widehat{\theta}(1) = \frac{1}{2}$, $\widehat{\theta}(2) = 0$. If $X = 1$ the only reasonable estimate of θ is $\frac{1}{2}$ because the value 1 could not have been produced when $\theta = 0$.

Maximum likelihood estimates need neither exist nor be unique (Problems 2.2.14 and 2.2.13). In the rest of this section we identify them as of minimum contrast and estimating equation type, relate them to the plug-in and extension principles and some notions in information theory, and compute them in some important special cases in which they exist, are unique, and expressible in closed form. In the rest of this chapter we study more detailed conditions for existence and uniqueness and algorithms for calculation of MLEs when closed forms are not available.

When θ is real, MLEs can often be obtained by inspection as we see in a pair of important examples.

Example 2.2.4. *The Normal Distribution with Known Variance.* Suppose $X \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 is known, and let φ denote the standard normal density. Then the likelihood function

$$L_x(\theta) = \frac{1}{\sigma}\varphi\left(\frac{\theta - x}{\sigma}\right)$$

is a normal density with mean x and variance σ^2. The maximum is, therefore, achieved uniquely for

$$\widehat{\theta}(x) = x.$$

Suppose more generally that $X_1, \ldots, X_n$ is a sample from a $\mathcal{N}(\theta, \sigma^2)$ population. It is a consequence of Problem 2.2.15 that the MLE of θ based on $X_1, \ldots, X_n$ is the same as that based on the sufficient statistic $\bar{X}$, which has a $\mathcal{N}(\theta, \sigma^2/n)$ distribution. In view of our result for $n = 1$ we can conclude that

$$\widehat{\theta}(X_1, \ldots, X_n) = \bar{X}$$

is the MLE of θ. □

Example 2.2.5. *Estimating the Size of a Population (continued).* Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathcal{U}\{1, 2, \ldots, \theta\}$ with θ an integer ≥ 1. We have seen in Example 2.1.7 that the method of moments leads to the unreasonable estimate $2\bar{X} - 1$ for the size θ of the population. What is the maximum likelihood estimate of θ? From (1.5.10) we see that $L_{\mathbf{x}}(\theta)$ is 0 for $\theta = 1, \ldots, \max(x_1, \ldots, x_n) - 1$, then jumps to $[\max(x_1, \ldots, x_n)]^{-n}$ and equals the monotone decreasing function θ^{-n} from then on. Figure 2.2.2 illustrates the situation. Clearly, $\max(X_1, \ldots, X_n)$ is the MLE of θ. □

Maximum likelihood as a minimum contrast and estimating equation method. Define

$$l_x(\theta) = \log L_x(\theta) = \log p(x, \theta).$$

By definition the MLE $\widehat{\theta}(X)$ if it exists minimizes $-\log p$ because $-l_x(\theta)$ is a strictly decreasing function of p. Log p turns out to be the best monotone function of p to consider for many reasons. A prototypical one is that if the X_i are independent with densities or frequency function $f(x, \theta)$ for $i = 1, \ldots, n$, then, with $\mathbf{X} = (X_1, \ldots, X_n)$

$$l_{\mathbf{X}}(\theta) = \log p(\mathbf{X}, \theta) = \log \prod_{i=1}^{n} f(X_i, \theta) = \sum_{i=1}^{n} \log f(X_i, \theta) \tag{2.2.21}$$

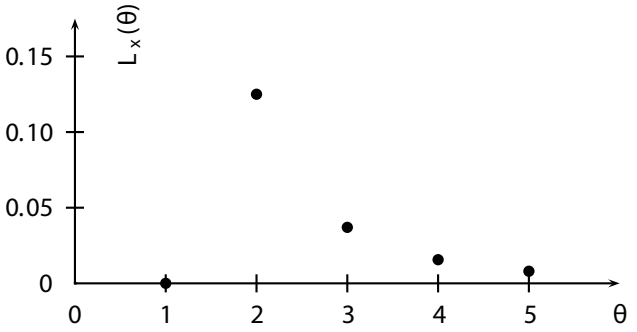

Figure 2.2.2. The likelihood function for Example 2.2.5. Here $n = 3$ and $\mathbf{x} = (x_1, x_2, x_3) = (1, 2, 1)$.

which when considered as a random quantity is a sum of independent variables, the most studied object of probability theory.

Another reason for turning to $\log p$ is that we can, by making a connection to information theory, identify $\rho(x, \theta) \equiv -l_x(\theta)$ as a contrast and $\widehat{\theta}$ as a minimum contrast estimate. In this case,

$$D(\theta_0, \theta) = -E_{\theta_0} \log p(X, \theta) \tag{2.2.22}$$

and to say that $D(\theta_0, \theta)$ is minimized uniquely when $P_\theta = P_{\theta_0}$ is equivalent to

$$\begin{aligned} D(\theta_0, \theta) - D(\theta_0, \theta_0) &= -(E_{\theta_0} \log p(X, \theta) - E_{\theta_0} \log p(X, \theta_0)) \\ &= -E_{\theta_0} \log \frac{p(X, \theta)}{p(X, \theta_0)} > 0 \end{aligned} \tag{2.2.23}$$

unless $\theta = \theta_0$. Here $D(\theta_0, \theta_0)$ is called the *entropy* of X. See Problem 1.6.28.

Let X have frequency or density function p_0 or p_1 and let E_0 denote expectation with respect to p_0. Define the *mutual entropy* or *Kullback–Leibler information divergence* $K(p_0, p_1)$ between p_0 and p_1 by

$$K(p_0, p_1) = -E_0 \log \frac{p_1}{p_0}(X) = -\sum_x p_0(x) \log \frac{p_1}{p_0}(x) \tag{2.2.24}$$

for X discrete and replacing sums by integrals for X continuous. By convention $\frac{0}{0} = 0$ and $0 \times \infty = 0$ so that if $p_0(x) = 0$, $p_0(x) \log(p_1(x)/p_0(x)) = 0$ and if $p_1(x) = 0$ also, then $p_1(x)/p_0(x) = 0$. Then (2.2.23) is equivalent to

Lemma 2.2.1 (Shannon, 1948)[2] *The mutual entropy $K(p_0, p_1)$ is always well defined and $K(p_0, p_1) \geq 0$ with equality if and only if $\{x : p_0(x) = p_1(x)\}$ has probability 1 under both P_0 and P_1.*

Proof. We give the proof for the discrete case. Kolmogorov (1956) gave the proof in the general case. Recall Jensen's inequality (B.9.3): If $g : R \to R$ is strictly convex and Z is any random variable such that $E(Z)$ is well defined, then $Eg(Z)$ is well defined and $Eg(Z) \geq g(E(Z))$ with = iff $P[Z = c_0] = 1$ where c_0 is a constant. Given X distributed according to P_0, let $Z = p_1(X)/p_0(X)$ and $g(z) = -\log z$. Because $g''(z) = z^{-2} > 0$, g is strictly convex. Thus, with our convention about 0 and ∞,

$$\begin{aligned} Eg(Z) &= \sum_x p_0(x) \left(-\log \frac{p_1}{p_0}(x) \right) \geq g(E(Z)) \\ &= -\log \left(\sum_x \frac{p_1}{p_0}(x) p_0(x) \right). \end{aligned} \tag{2.2.25}$$

Because $\frac{p_1}{p_0}(x)p_0(x) \leq p_1(x)$ we must have $g(E(Z)) \geq -\log \sum_x p_1(x) = 0$. Equality holds iff $p_1(x) > 0$ implies $p_0(x) > 0$ and $p_1(x)/p_0(x) = c$ if $p_0(x) > 0$. Then $1 = \sum_x p_1(x) = c \sum_x p_0(x) = c$, and we conclude that $p_0(x) = p_1(x)$ for all x. □

Lemma 2.2.1 shows that in the case of $X_1, \ldots, X_n$ i.i.d.

$$\rho(X, \theta) = -\frac{1}{n} \sum \log p(X_i, \theta)$$

satisfies the condition of being a contrast function, and we have shown that the MLE is a minimum contrast estimate.

Next let $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta\}$ and define $\nu : \mathcal{P}_0 \to \Theta$ by

$$\nu(P_{\theta_0}) = \arg\min\{-E_{\theta_0} \log p(X, \theta) : \theta \in \Theta\}.$$

The extension $\bar{\nu}$ of ν to $\mathcal{P} =$ all probabilities on $\mathcal{X}$ is

$$\bar{\nu}(P) = \arg\min\{-E_P \log p(X, \theta) : \theta \in \Theta\}.$$

Now the MLE is $\bar{\nu}(\widehat{P})$. That is, the MLE is the value of θ that minimizes the Kullback–Leibler divergence between the empirical probability $\widehat{P}$ and P_θ.

Likelihood equations

If Θ is open, $l_X(\theta)$ is differentiable in θ and $\widehat{\theta}$ exists then $\widehat{\theta}$ must satisfy the estimating equation

$$\nabla_\theta l_X(\theta) = 0. \tag{2.2.26}$$

This is known as the *likelihood equation.* If the X_i are independent with densities $f_i(x, \theta)$ the likelihood equation simplifies to

$$\sum_{i=1}^{n} \nabla_\theta \log f_i(X_i, \widehat{\theta}) = 0, \tag{2.2.27}$$

which again enables us to analyze the behavior of $\widehat{\theta}$ using known properties of sums of independent random variables. Evidently, there may be solutions of (2.2.27) that are not maxima or only local maxima, and as we have seen in Example 2.2.5, situations with θ well defined but (2.2.27) doesn't make sense. Nevertheless, the dual point of view of (2.2.22) and (2.2.27) is very important and we shall explore it extensively in the natural and favorable setting of multiparameter exponential families in the next section.

Here are two simple examples with θ real.

Example 2.2.6. Consider a population with three kinds of individuals labeled 1, 2, and 3 and occurring in the Hardy–Weinberg proportions

$$p(1,\theta) = \theta^2,\ p(2,\theta) = 2\theta(1-\theta),\ p(3,\theta) = (1-\theta)^2$$

where $0 < \theta < 1$ (see Example 2.1.4). If we observe a sample of three individuals and obtain $x_1 = 1, x_2 = 2, x_3 = 1$, then

$$L_{\mathbf{x}}(\theta) = p(1,\theta)p(2,\theta)p(1,\theta) = 2\theta^5(1-\theta).$$

The likelihood equation is

$$\frac{\partial}{\partial\theta} l_{\mathbf{x}}(\theta) = \frac{5}{\theta} - \frac{1}{1-\theta} = 0,$$

which has the unique solution $\widehat{\theta} = \frac{5}{6}$. Because

$$\frac{\partial^2}{\partial\theta^2} l_{\mathbf{x}}(\theta) = -\frac{5}{\theta^2} - \frac{1}{(1-\theta)^2} < 0$$

for all $\theta \in (0,1)$, $\frac{5}{6}$ maximizes $L_{\mathbf{x}}(\theta)$. In general, let n_1, n_2, and n_3 denote the number of $\{x_1,\ldots,x_n\}$ equal to 1, 2 and 3, respectively. Then the same calculation shows that if $2n_1 + n_2$ and $n_2 + 2n_3$ are both positive, the maximum likelihood estimate exists and is given by

$$\widehat{\theta}(\mathbf{x}) = \frac{2n_1 + n_2}{2n}. \tag{2.2.28}$$

If $2n_1 + n_2$ is zero, the likelihood is $(1-\theta)^{2n}$, which is maximized by $\theta = 0$, so the MLE does not exist because $\Theta = (0,1)$. Similarly, the MLE does not exist if $n_2 + 2n_3 = 0$. □

Example 2.2.7. Let X denote the number of customers arriving at a service counter during n hours. If we make the usual simplifying assumption that the arrivals form a Poisson process, then X has a Poisson distribution with parameter $n\lambda$, where λ represents the expected number of arrivals in an hour or, equivalently, the rate of arrival. In practice, λ is an unknown positive constant and we wish to estimate λ using X. Here X takes on values $\{0,1,2,\ldots\}$ with probabilities,

$$p(x,\lambda) = \frac{e^{-\lambda n}(\lambda n)^x}{x!},\ x = 0,1,\ldots \tag{2.2.29}$$

The likelihood equation is

$$\frac{\partial}{\partial\lambda}l_x(\lambda) = -n + \frac{x}{\lambda} = 0,$$

which has the unique solution $\widehat{\lambda} = x/n$. If x is positive, this estimate is the MLE of λ (see (2.2.30)). If $x = 0$ the MLE does not exist. However, the maximum is approached as $\lambda \downarrow 0$. □

To apply the likelihood equation successfully we need to know when a solution is an MLE. A sufficient condition, familiar from calculus, is that l be *concave* in θ. If l is twice differentiable, this is well known to be equivalent to

$$\frac{\partial^2}{\partial\theta^2}l_{\mathbf{x}}(\theta) \leq 0, \tag{2.2.30}$$

for *all* θ. This is the condition we applied in Example 2.2.6. A similar condition applies for vector parameters.

Example 2.2.8. *Multinomial Trials.* As in Example 1.6.7, consider an experiment with n i.i.d. trials in which each trial can produce a result in one of k categories. Let $X_i = j$ if the ith trial produces a result in the jth category, let $\theta_j = P(X_i = j)$ be the probability of the jth category, and let $N_j = \sum_{i=1}^n 1[X_i = j]$ be the number of observations in the jth category. We assume that $n \geq k - 1$. Then, for an experiment in which we observe $n_j = \sum_{i=1}^n 1[X_i = j]$, $p(\mathbf{x}, \boldsymbol{\theta}) = \prod_{j=1}^k \theta_j^{n_j}$, and

$$l_{\mathbf{x}}(\boldsymbol{\theta}) = \sum_{j=1}^k n_j \log\theta_j,\ \boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta} : \theta_j \geq 0, \sum_{j=1}^k \theta_j = 1\}. \tag{2.2.31}$$

To obtain the MLE $\widehat{\boldsymbol{\theta}}$ we consider l as a function of $\theta_1, \ldots, \theta_{k-1}$ with

$$\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j. \tag{2.2.32}$$

We first consider the case with all the n_j positive. Then $p(\mathbf{x}, \boldsymbol{\theta}) = 0$ if any of the θ_j are zero; thus, the MLE must have all $\widehat{\theta}_j > 0$, and must satisfy the likelihood equations

$$\frac{\partial}{\partial\theta_j}l_{\mathbf{x}}(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_j}\sum_{l=1}^k n_l \log\theta_l = \sum_{l=1}^k \frac{n_l}{\theta_l}\frac{\partial\theta_l}{\partial\theta_j} = 0,\ j = 1, \ldots, k-1.$$

By (2.2.32), $\partial\theta_k/\partial\theta_j = -1$, and the equation becomes $(\widehat{\theta}_k/\widehat{\theta}_j) = (n_k/n_j)$. Next use (2.2.32) to find

$$\widehat{\theta}_j = \frac{n_j}{n},\ j = 1, \ldots, k.$$

To show that this $\widehat{\boldsymbol{\theta}}$ maximizes $l_{\mathbf{x}}(\boldsymbol{\theta})$, we check the concavity of $l_{\mathbf{x}}(\boldsymbol{\theta})$: let $1 \le r \le k-1$, $1 \le j \le k-1$, then

$$\begin{aligned} \frac{\partial}{\partial\theta_r}\frac{\partial}{\partial\theta_j} l_{\mathbf{x}}(\boldsymbol{\theta}) &= \frac{\partial}{\partial\theta_r}\left(\frac{n_j}{\theta_j} - \frac{n_k}{\theta_k}\right) \\ &= -\left(\frac{n_r}{\theta_r^2} + \frac{n_k}{\theta_k^2}\right) < 0,\ r = j \\ &= -\frac{n_k}{\theta_k^2} < 0,\ r \neq j. \end{aligned} \tag{2.2.33}$$

It follows that in this $n_j > 0$, $\theta_j > 0$ case, $l_{\mathbf{x}}(\boldsymbol{\theta})$ is strictly concave and $\widehat{\boldsymbol{\theta}}$ is the unique maximizer of $l_{\mathbf{x}}(\boldsymbol{\theta})$. Next suppose that $n_j = 0$ for some j. Then $\widehat{\boldsymbol{\theta}}$ with $\widehat{\theta}_j = n_j/n$, $j = 1, \ldots, k$, is still the unique MLE of $\boldsymbol{\theta}$. See Problem 2.2.30. The $0 < \theta_j < 1$, $1 \le j \le k$, version of this example will be considered in the exponential family case in Section 2.3.

Example 2.2.9. Suppose that $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with μ and σ^2 both unknown. Using the concavity argument, we find that for $n \ge 2$ the unique MLEs of μ and σ^2 are $\widehat{\mu} = \bar{X}$ and $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$ (Problem 2.2.11(a)).

Maximum likelihood and least squares

We conclude with the link between least squares and maximum likelihood. Suppose the model $\mathcal{P}_0$ (see (2.2.1)) of Example 2.1.1 holds and $\mathbf{X} = (Y_1, \ldots, Y_n)^T$. Then

$$\begin{aligned} l_{\mathbf{X}}(\boldsymbol{\beta}) &= \log \prod_{i=1}^n \frac{1}{\sigma_0}\varphi\left(\frac{Y_i - g(\boldsymbol{\beta}, \mathbf{z}_i)}{\sigma_0}\right) \\ &= -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^n [Y_i - g(\boldsymbol{\beta}, \mathbf{z}_i)]^2. \end{aligned} \tag{2.2.34}$$

Evidently maximizing $l_{\mathbf{X}}(\boldsymbol{\beta})$ is equivalent to minimizing $\sum_{i=1}^n [Y_i - g(\boldsymbol{\beta}, \mathbf{z}_i)]^2$. Thus, least squares estimates are maximum likelihood for the particular model $\mathcal{P}_0$. As we have seen and shall see more in Section 6.6, these estimates viewed as an algorithm applied to the set of data $\mathbf{X}$ make sense much more generally. It is easy to see that weighted least squares estimates are themselves maximum likelihood estimates of $\boldsymbol{\beta}$ for the model Y_i independent $\mathcal{N}(g(\boldsymbol{\beta}, \mathbf{z}_i), w_i\sigma_0^2)$, $1 \le i \le n$. More generally, we can consider $\widehat{\boldsymbol{\beta}}$ minimizing $\sum_{i,j}[Y_i - g(\boldsymbol{\beta}, \mathbf{z}_i)][Y_j - g(\boldsymbol{\beta}, \mathbf{z}_j)]w_{ij}$ where $W = \|w_{ij}\|_{n\times n}$ is a symmetric positive definite matrix, as maximum likelihood estimates for $\boldsymbol{\beta}$ when $\mathbf{Y}$ is distributed as $\mathcal{N}_n((g(\boldsymbol{\beta}, \mathbf{z}_1), \ldots, g(\boldsymbol{\beta}, \mathbf{z}_n))^T, \sigma_0^2 W^{-1})$, see Problem 2.2.28.

Summary. In Section 2.2.1 we consider *least squares estimators* (LSEs) obtained by minimizing a contrast of the form $\sum_{i=1}^n [Y_i - g_i(\boldsymbol{\beta})]^2$, where $E(Y_i) = g_i(\boldsymbol{\beta})$, g_i, $i = 1, \ldots, n$, are known functions and $\boldsymbol{\beta}$ is a parameter to be estimated from the independent observations $Y_1, \ldots, Y_n$, where $\mathrm{Var}(Y_i)$ does not depend on i. This approach is applied to experiments in which for the ith case in a study the mean of the response Y_i depends on

a set of available covariate values $z_{i1}, \ldots, z_{id}$. In particular we consider the case with $g_i(\boldsymbol{\beta}) = \sum_{j=1}^{d} z_{ij}\beta_j$ and give the LSE of $\boldsymbol{\beta}$ in the case in which $\|z_{ij}\|_{n\times d}$ is of rank d. Extensions to weighted least squares, which are appropriate when $\mathrm{Var}(Y_i)$ depends on i or the Y's are correlated, are given. In Section 2.2.2 we consider *maximum likelihood estimators* (MLEs) $\widehat{\boldsymbol{\theta}}$ that are defined as maximizers of the likelihood $L_x(\theta) = p(x, \boldsymbol{\theta})$. These estimates are shown to be equivalent to minimum contrast estimates based on a contrast function related to Shannon entropy and Kullback–Leibler information divergence. In the case of independent response variables Y_i that are modeled to have a $\mathcal{N}(g_i(\boldsymbol{\beta}), \sigma^2)$ distribution, it is shown that the MLEs coincide with the LSEs.

2.3 MAXIMUM LIKELIHOOD IN MULTIPARAMETER EXPONENTIAL FAMILIES

Questions of existence and uniqueness of maximum likelihood estimates in canonical exponential families can be answered completely and elegantly. This is largely a consequence of the strict concavity of the log likelihood in the natural parameter η, though the results of Theorems 1.6.3 and 1.6.4 and Corollaries 1.6.1 and 1.6.2 and other exponential family properties also play a role. Concavity also plays a crucial role in the analysis of algorithms in the next section. Properties that derive solely from concavity are given in Propositon 2.3.1.

We start with a useful general framework and lemma. Suppose $\Theta \subset R^p$ is an open set. Let $\partial\Theta = \bar{\Theta} - \Theta$ be the boundary of Θ, where $\bar{\Theta}$ denotes the closure of Θ in $[-\infty, \infty]^p$. That is, $\partial\Theta$ is the set of points outside of Θ that can be obtained as limits of points in Θ, including all points with $\pm\infty$ as a coordinate. For instance, if $X \sim \mathcal{N}(\theta_1, \theta_2)$, $\Theta = R \times R^+$ and

$$\partial\Theta = \{(a, b) : a = \pm\infty, 0 \leq b \leq \infty\} \cup \{(a, b) : a \in R, b \in \{0, \infty\}\}.$$

In general, for a sequence $\{\boldsymbol{\theta}_m\}$ of points from Θ open, we define $\boldsymbol{\theta}_m \to \partial\Theta$ as $m \to \infty$ to mean that for any subsequence $\{\theta_{m_k}\}$ either $\boldsymbol{\theta}_{m_k} \to \mathbf{t}$ with $\mathbf{t} \notin \Theta$, or $\boldsymbol{\theta}_{m_k}$ diverges with $|\boldsymbol{\theta}_{m_k}| \to \infty$, as $k \to \infty$, where $|\ |$ denotes the Euclidean norm. For instance, in the $\mathcal{N}(\theta_1, \theta_2)$ case, $(a, m^{-1}), (m, b), (-m, b), (a, m), (m, m^{-1})$ all tend to $\partial\Theta$ as $m \to \infty$.

Lemma 2.3.1. *Suppose we are given a function $l : \Theta \to R$ where $\Theta \subset R^p$ is open and l is continuous. Suppose also that*

$$\lim\{l(\boldsymbol{\theta}) : \boldsymbol{\theta} \to \partial\Theta\} = -\infty. \tag{2.3.1}$$

Then there exists $\widehat{\boldsymbol{\theta}} \in \Theta$ such that

$$l(\widehat{\boldsymbol{\theta}}) = \max\{l(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}.$$

Proof. See Problem 2.3.5.

Existence and unicity of the MLE in exponential families depend on the strict concavity of the log likelihood and the condition of Lemma 2.3.1 only. Formally,

Proposition 2.3.1. *Suppose $X \sim \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, Θ open $\subset R^p$, with corresponding densities $p(x, \boldsymbol{\theta})$. If further $l_x(\boldsymbol{\theta}) \equiv \log p(x, \boldsymbol{\theta})$ is strictly concave and $l_x(\boldsymbol{\theta}) \to -\infty$ as $\boldsymbol{\theta} \to \partial\Theta$, then the MLE $\widehat{\boldsymbol{\theta}}(x)$ exists and is unique.*

Proof. From (B.9) we know that $\boldsymbol{\theta} \to l_x(\boldsymbol{\theta})$ is continuous on Θ. By Lemma 2.3.1, $\widehat{\boldsymbol{\theta}}(x)$ exists. If $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$ are distinct maximizers, then $l_x\left(\frac{1}{2}(\widehat{\boldsymbol{\theta}}_1 + \widehat{\boldsymbol{\theta}}_2)\right) > \frac{1}{2}(l_x(\widehat{\boldsymbol{\theta}}_1) + l_x(\widehat{\boldsymbol{\theta}}_2)) = l_x(\widehat{\boldsymbol{\theta}}_1)$, a contradiction.

Applications of this theorem are given in Problems 2.3.8 and 2.3.12. □

We can now prove the following.

Theorem 2.3.1. *Suppose $\mathcal{P}$ is the canonical exponential family generated by $(\mathbf{T}, h)$ and that*

(i) *The natural parameter space, $\mathcal{E}$, is open.*

(ii) *The family is of rank k.*

Let x be the observed data vector and set $\mathbf{t}_0 = \mathbf{T}(x)$.

(a) *If $\mathbf{t}_0 \in R^k$ satisfies*[1]

$$P[\mathbf{c}^T\mathbf{T}(X) > \mathbf{c}^T\mathbf{t}_0] > 0 \qquad \forall \mathbf{c} \neq \mathbf{0} \tag{2.3.2}$$

then the MLE $\widehat{\boldsymbol{\eta}}$ exists, is unique, and is a solution to the equation

$$\dot{A}(\boldsymbol{\eta}) = E_{\boldsymbol{\eta}}(\mathbf{T}(X)) = \mathbf{t}_0. \tag{2.3.3}$$

(b) *Conversely, if $\mathbf{t}_0$ doesn't satisfy (2.3.2), then the MLE doesn't exist and (2.3.3) has no solution.*

We, thus, have a necessary and sufficient condition for existence and uniqueness of the MLE given the data.

Define the *convex support* of a probability P to be the smallest convex set C such that $P(C) = 1$.

Corollary 2.3.1. *Suppose the conditions of Theorem* 2.3.1 *hold. If $C_{\mathbf{T}}$ is the convex support of the distribution of $\mathbf{T}(X)$, then $\widehat{\boldsymbol{\eta}}$ exists and is unique iff $\mathbf{t}_0 \in C^0_{\mathbf{T}}$ where $C^0_{\mathbf{T}}$ is the interior of $C_{\mathbf{T}}$.*

Proof of Theorem 2.3.1. We give the proof for the continuous case.

Existence and Uniqueness of the MLE $\widehat{\boldsymbol{\eta}}$. Without loss of generality we can suppose $h(x) = p(x, \boldsymbol{\eta}_0)$ for some reference $\boldsymbol{\eta}_0 \in \mathcal{E}$ (see Problem 1.6.27). Furthermore, we may also assume that $\mathbf{t}_0 = \mathbf{T}(x) = 0$ because $\mathcal{P}$ is the same as the exponential family generated by $\mathbf{T}(x) - \mathbf{t}_0$. Then, if $l_x(\boldsymbol{\eta}) \equiv \log p(x, \boldsymbol{\eta})$ with $\mathbf{T}(x) = 0$,

$$l_x(\boldsymbol{\eta}) = -A(\boldsymbol{\eta}) + \log h(x).$$

We show that if $\{\boldsymbol{\eta}_m\}$ has no subsequence converging to a point in $\mathcal{E}$, then $l_x(\boldsymbol{\eta}_m) \to -\infty$, which implies existence of $\widehat{\boldsymbol{\eta}}$ by Lemma 2.3.1. Write $\boldsymbol{\eta}_m = \lambda_m \mathbf{u}_m$, $\mathbf{u}_m = \frac{\eta_m}{\|\eta_m\|}$, $\lambda_m =$

$\|\boldsymbol{\eta}_m\|$. So, $\|\mathbf{u}_m\| = 1$. Then, if $\{\boldsymbol{\eta}_m\}$ has no subsequence converging in $\mathcal{E}$ it must have a subsequence $\{\boldsymbol{\eta}_{m_k}\}$ that obeys either case 1 or 2 as follows.

Case 1: $\lambda_{m_k} \to \infty$, $\mathbf{u}_{m_k} \to \mathbf{u}$. Write E_0 for $E_{\boldsymbol{\eta}_0}$ and P_0 for $P_{\boldsymbol{\eta}_0}$. Then

$$\begin{aligned} \underline{\lim}_k \int e^{\boldsymbol{\eta}_{m_k}^T \mathbf{T}(x)} h(x)dx &= \underline{\lim}_k E_0 e^{\lambda_{m_k} \mathbf{u}_{m_k}^T \mathbf{T}(X)} \\ &\geq \underline{\lim} e^{\lambda_{m_k}\delta} P_0[\mathbf{u}_{m_k}^T \mathbf{T}(X) > \delta] \\ &\geq \underline{\lim} e^{\lambda_{m_k}\delta} P_0[\mathbf{u}^T \mathbf{T}(X) > \delta] = \infty \end{aligned}$$

because for some $\delta > 0$, $P_0[\mathbf{u}^T\mathbf{T}(X) > \delta] > 0$. So we have

$$A(\boldsymbol{\eta}_{m_k}) = \log \int e^{\boldsymbol{\eta}_{m_k}^T \mathbf{T}(x)} h(x)dx \to \infty \text{ and } l_x(\boldsymbol{\eta}_{m_k}) \to -\infty.$$

Case 2: $\lambda_{m_k} \to \lambda$, $\mathbf{u}_{m_k} \to \mathbf{u}$. Then $\lambda\mathbf{u} \notin \mathcal{E}$ by assumption. So

$$lim_k E_0 e^{\lambda_{m_k} \mathbf{u}_{m_k}^T \mathbf{T}(X)} = E_0 e^{\lambda \mathbf{u}^T \mathbf{T}(X)} = \infty.$$

In either case $\lim_{m_k} l_x(\boldsymbol{\eta}_{m_k}) = -\infty$. Because any subsequence of $\{\boldsymbol{\eta}_m\}$ has no subsequence converging in $\mathcal{E}$ we conclude $l_x(\boldsymbol{\eta}_m) \to -\infty$ and $\widehat{\boldsymbol{\eta}}$ exists. It is unique and satisfies (2.3.3) by Theorem 1.6.4.

Nonexistence: If (2.3.2) fails, there exists $\mathbf{c} \neq 0$ such that $P_0[\mathbf{c}^T\mathbf{T} \leq 0] = 1 \Rightarrow E_{\boldsymbol{\eta}}(\mathbf{c}^T\mathbf{T}(X)) \leq 0$, for all $\boldsymbol{\eta}$. If $\widehat{\boldsymbol{\eta}}$ exists then $E_{\widehat{\boldsymbol{\eta}}}\mathbf{T} = 0 \Rightarrow E_{\widehat{\boldsymbol{\eta}}}(\mathbf{c}^T\mathbf{T}) = 0 \Rightarrow P_{\widehat{\boldsymbol{\eta}}}[\mathbf{c}^T\mathbf{T} = \mathbf{0}] = 1$, contradicting the assumption that the family is of rank k. □

Proof of Corollary 2.3.1. By (B.9.1) a point $\mathbf{t}_0$ belongs to the interior C of a convex set C iff there exist points in C^0 on either side of it, that is, iff, for every $\mathbf{d} \neq \mathbf{0}$, both $\{\mathbf{t} : \mathbf{d}^T\mathbf{t} > \mathbf{d}^T\mathbf{t}_0\} \cap C^0$ and $\{\mathbf{t} : \mathbf{d}^T\mathbf{t} < \mathbf{d}^T\mathbf{t}_0\} \cap C^0$ are nonempty open sets. The equivalence of (2.3.2) and Corollary 2.3.1 follow. □

Example 2.3.1. *The Gaussian Model.* Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, $\mu \in R$, $\sigma^2 > 0$. As we observed in Example 1.6.5, this is the exponential family generated by $T(\mathbf{X}) \equiv (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ and 1. Evidently, $C_T = R \times R^+$. For $n = 2$, $T(\mathbf{X})$ has a density by Problem B.2.7. The $n > 2$ case follows because it involves convolution with the $n = 2$ case. Thus, $C_T = C_T^0$ and the MLE always exists. For $n = 1$, $C_T^0 = \emptyset$ because $T(X_1)$ is always a point on the parabola $T_2 = T_1^2$ and the MLE does not exist. This is equivalent to the fact that if $n = 1$ the formal solution to the likelihood equations gives $\widehat{\sigma}^2 = 0$, which is impossible. □

Existence of MLEs when $\mathbf{T}$ has a continuous case density is a general phenomenon.

Theorem 2.3.2. *Suppose the conditions of Theorem* 2.3.1 *hold and* $\mathbf{T}_{k\times 1}$ *has a continuous case density on* R^k. *Then the MLE* $\widehat{\boldsymbol{\eta}}$ *exists with probability* 1 *and necessarily satisfies (2.3.3).*

Proof. The boundary of a convex set necessarily has volume 0 (Problem 2.3.9), thus, if $\mathbf{T}$ has a continuous case density $p_{\mathbf{T}}(\mathbf{t})$, then

$$P[\mathbf{T} \in \partial C_{\mathbf{T}}] = \int_{\partial C_{\mathbf{T}}} p_{\mathbf{T}}(\mathbf{t}) d\mathbf{t} = 0$$

and the result follows from Corollary 2.3.1. □

Remark 2.3.1. From Theorem 1.6.3 we know that $E_{\boldsymbol{\eta}} T(X) = \dot{A}(\boldsymbol{\eta})$. Thus, using (2.3.3), the MLE $\widehat{\boldsymbol{\eta}}$ in exponential families has an interpretation as a generalized method of moments estimate (see Problem 2.1.13 and the next example). When method of moments and frequency substitution estimates are not unique, the maximum likelihood principle in many cases selects the "best" estimate among them. For instance, in the Hardy–Weinberg examples 2.1.4 and 2.2.6, $\widehat{\theta}_1 = \sqrt{n_1/n}$, $\widehat{\theta}_2 = 1 - \sqrt{n_3/n}$ and $\widehat{\theta}_3 = (2n_1 + n_2)/2n$ are frequency substitution estimates (Problem 2.1.1), but only $\widehat{\theta}_3$ is a MLE. In Example 3.4.4 we will see that $\widehat{\theta}_3$ is, in a certain sense, the best estimate of θ.

A nontrivial application of Theorem 2.3.2 follows.

Example 2.3.2. *The Two-Parameter Gamma Family.* Suppose $X_1, \ldots, X_n$ are i.i.d. with density $g_{p,\lambda}(x) = \frac{\lambda^p}{\Gamma(p)} e^{-\lambda x} x^{p-1}$, $x > 0$, $p > 0$, $\lambda > 0$. This is a rank 2 canonical exponential family generated by $\mathbf{T} = (\sum \log X_i, \sum X_i)$, $h(x) = x^{-1}$, with

$$\eta_1 = p,\ \eta_2 = -\lambda,\ A(\eta_1, \eta_2) = n(\log \Gamma(\eta_1) - \eta_1 \log(-\eta_2))$$

by Problem 2.3.2(a). The likelihood equations are equivalent to (Problem 2.3.2(b))

$$\frac{\Gamma'}{\Gamma}(\widehat{p}) - \log \widehat{\lambda} = \overline{\log X} \tag{2.3.4}$$

$$\frac{\widehat{p}}{\widehat{\lambda}} = \bar{X} \tag{2.3.5}$$

where $\overline{\log X} \equiv \frac{1}{n}\sum_{i=1}^n \log X_i$. If $n = 2$, $\mathbf{T}$ has a density by Theorem B.2.2. The $n > 2$ case follows because it involves convolution with the $n = 2$ case. We conclude from Theorem 2.3.2 that (2.3.4) and (2.3.5) have a unique solution with probability 1. How to find such nonexplicit solutions is discussed in Section 2.4. □

If $\mathbf{T}$ is discrete MLEs need not exist. Here is an example.

Example 2.3.3. *Multinomial Trials.* We follow the notation of Example 1.6.7. The statistic of rank $k-1$ which generates the family is $\mathbf{T}_{(k-1)} = (T_1, \ldots, T_{k-1})^T$, where $T_j(\mathbf{X}) = \sum_{i=1}^n 1(X_i = j)$, $1 \le j \le k$. We assume $n \ge k-1$ and verify using Theorem 2.3.1 that in this case MLEs of $\eta_j = \log(\lambda_j/\lambda_k)$, $1 \le j \le k-1$, where $0 < \lambda_j \equiv P[X = j] < 1$, exist iff all $T_j > 0$. They are determined by $\widehat{\lambda}_j = T_j/n$, $1 \le j \le k$. To see this note that $T_j > 0$, $1 \le j \le k$ iff $0 < T_j < n$, $1 \le j \le k$. Thus, if we write $\mathbf{c}^T\mathbf{t}_0 = \sum\{c_j t_{j0} : c_j > 0\} + \sum\{c_j t_{j0} : c_j < 0\}$ we can increase $\mathbf{c}^T\mathbf{t}_0$ by replacing a t_{j0} by $t_{j0}+1$ in the first sum or a t_{j0} by $t_{j0}-1$ in the second. Because the resulting value of $\mathbf{t}$ is possible if $0 < t_{j0} < n$,

$1 \leq j \leq k$, and one of the two sums is nonempty because $\mathbf{c} \neq \mathbf{0}$, we see that (2.3.2) holds. On the other hand, if any $T_j = 0$ or n, $0 \leq j \leq k-1$ we can obtain a contradiction to (2.3.2) by taking $c_i = -1(i = j)$, $1 \leq i \leq k-1$. The remaining case $T_k = 0$ gives a contradiction if $\mathbf{c} = (1, 1, \ldots, 1)^T$. Alternatively we can appeal to Corollary 2.3.1 directly (Problem 2.3.10). □

Remark 2.3.1. In Example 2.2.8 we saw that in the multinomial case with the closed parameter set $\{\lambda_j : \lambda_j \geq 0, \sum_{j=1}^k \lambda_j = 1\}, n \geq k-1$, the MLEs of $\lambda_j, j = 1, \ldots, k$, exist and are unique. However, when we put the multinomial in canonical exponential family form, our parameter set is open. Similarly, note that in the Hardy–Weinberg Example 2.2.6, if $2n_1 + n_2 = 0$, the MLE does not exist if $\Theta = (0, 1)$, whereas if $\theta = [0, 1]$ it does exist and is unique. □

The argument of Example 2.3.3 can be applied to determine existence in cases for which (2.3.3) does not have a closed-form solution as in Example 1.6.8—see Problem 2.3.1 and Haberman (1974).

In some applications, for example, the bivariate normal case (Problem 2.3.13), the following corollary to Theorem 2.3.1 is useful.

Corollary 2.3.2. *Consider the exponential family*

$$p(x, \theta) = h(x) \exp\left\{\sum_{j=1}^{k} c_j(\theta) T_j(x) - B(\theta)\right\},\ x \in \mathcal{X},\ \theta \in \Theta.$$

Let C^0 denote the interior of the range of $(c_1(\theta), \ldots, c_k(\theta))^T$ and let x be the observed data. If the equations

$$E_\theta T_j(X) = T_j(x),\ j = 1, \ldots, k$$

have a solution $\widehat{\theta}(x) \in C^0$, then it is the unique MLE of θ.

When $\mathcal{P}$ is not an exponential family both existence and unicity of MLEs become more problematic. The following result can be useful. Let $Q = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, Θ open $\subset R^m$, $m \leq k-1$, be a curved exponential family

$$p(x, \boldsymbol{\theta}) = \exp\{c^T(\boldsymbol{\theta})\mathbf{T}(x) - A(c(\boldsymbol{\theta}))\}h(x). \tag{2.3.6}$$

Suppose $c : \Theta \to \mathcal{E} \subset R^k$ has a differential $\dot{c}(\boldsymbol{\theta}) \equiv \left\| \frac{\partial c_j}{\partial \theta_i}(\boldsymbol{\theta}) \right\|_{m \times k}$ on Θ. Here $\mathcal{E}$ is the natural parameter space of the exponential family $\mathcal{P}$ generated by $(\mathbf{T}, h)$. Then

Theorem 2.3.3. *If $\mathcal{P}$ above satisfies the condition of Theorem 2.3.1, $c(\Theta)$ is closed in $\mathcal{E}$ and $\mathbf{T}(\mathbf{x}) = \mathbf{t}_0$ satisfies (2.3.2) so that the MLE $\widehat{\boldsymbol{\eta}}$ in $\mathcal{P}$ exists, then so does the MLE $\widehat{\boldsymbol{\theta}}$ in Q and it satisfies the likelihood equation*

$$\dot{c}^T(\widehat{\boldsymbol{\theta}})(\mathbf{t}_0 - \dot{A}(c(\widehat{\boldsymbol{\theta}}))) = 0. \tag{2.3.7}$$

Note that $c(\widehat{\boldsymbol{\theta}}) \in c(\Theta)$ and is in general not $\widehat{\boldsymbol{\eta}}$. Unfortunately strict concavity of l_x is not inherited by curved exponential families, and unicity can be lost—take $\mathbf{c}$ not one-to-one for instance.

The proof is sketched in Problem 2.3.11.

Example 2.3.4. *Gaussian with Fixed Signal to Noise.* As in Example 1.6.9, suppose $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with $\mu/\sigma = \lambda_0 > 0$ known. This is a curved exponential family with $c_1(\mu) = \frac{\lambda_0^2}{\mu}$, $c_2(\mu) = -\frac{\lambda_0^2}{2\mu^2}$, $\mu > 0$, corresponding to $\eta_1 = \frac{\mu}{\sigma^2}$, $\eta_2 = -\frac{1}{2\sigma^2}$. Evidently $c(\Theta) = \{(\eta_1, \eta_2) : \eta_2 = -\frac{1}{2}\eta_1^2\lambda_0^{-2}, \eta_1 > 0, \eta_2 < 0\}$, which is closed in $\mathcal{E} = \{(\eta_1, \eta_2) : \eta_1 \in R, \eta_2 < 0\}$. As a consequence of Theorems 2.3.2 and 2.3.3, we can conclude that an MLE $\widehat{\mu}$ always exists and satisfies (2.3.7) if $n \geq 2$. We find

$$\dot{c}(\theta) = \lambda_0^2(-\mu^{-2}, \mu^{-3})^T,$$

and from Example 1.6.5

$$\dot{A}(\boldsymbol{\eta}) = \frac{1}{2}n(-\eta_1/\eta_2, \eta_1^2/2\eta_2^2 - 1/\eta_2)^T.$$

Thus, with $t_1 = \sum x_i$ and $t_2 = \sum x_i^2$, Equation (2.3.7) becomes

$$\lambda_0^2(-\mu^{-2}, \mu^{-3})(t_1 - n\mu, t_2 - n(\mu^2 + \lambda_0^2\mu^2))^T = 0,$$

which with $\widehat{\mu}_2 = n^{-1}\sum x_i^2$ simplifies to

$$\mu^2 + \lambda_0^2\bar{x}\mu - \lambda_0^2\widehat{\mu}_2 = 0$$

$$\widehat{\mu}_\pm = \frac{1}{2}[\lambda_0^2\bar{x} \pm \lambda_0\sqrt{\lambda_0^2\bar{x}^2 + 4\widehat{\mu}_2}].$$

Note that $\widehat{\mu}_+\widehat{\mu}_- = -\lambda_0^2\widehat{\mu}_2 < 0$, which implies $\widehat{\mu}_+ > 0$, $\widehat{\mu}_- < 0$. Because $\mu > 0$, the solution we seek is $\widehat{\mu}_+$. □

Example 2.3.5. *Location-Scale Regression.* Suppose that $Y_{j1}, \ldots, Y_{jm}$, $j = 1, \ldots, n$, are n independent random samples, where $Y_{jl} \sim \mathcal{N}(\mu_j, \sigma_j^2)$. Using Examples 1.6.5 and 1.6.10, we see that the distribution of $\{Y_{jl} : j = 1, \ldots, n, l = 1, \ldots, m\}$ is a $2n$-parameter canonical exponential family with $\eta_i = \mu_i/\sigma_i^2$, $\eta_{n+i} = -1/2\sigma_i^2$, $i = 1, \ldots, n$, generated by $h(\mathbf{Y}) = 1$ and

$$T(\mathbf{Y}) = \left(\sum_{l=1}^m Y_{1l}, \ldots, \sum_{l=1}^m Y_{nl}, \sum_{l=1}^m Y_{1l}^2, \ldots, \sum_{l=1}^m Y_{nl}^2\right)^T.$$

Next suppose, as in Example 1.6.10, that

$$\mu_i = \theta_1 + \theta_2 z_i,\ \sigma_i^2 = \theta_3(\theta_1 + \theta_2 z_i)^2,\ z_1 < \cdots < z_n$$

where $z_1, \ldots, z_n$ are given constants. Now $p(\mathbf{y}, \boldsymbol{\theta})$ is a curved exponential family of the form (2.3.6) with

$$c_i(\boldsymbol{\theta}) = \theta_3^{-1}(\theta_1 + \theta_2 z_i)^{-1},\ c_{n+i}(\boldsymbol{\theta}) = \frac{1}{2}\theta_3^{-1}(\theta_1 + \theta_2 z_i)^{-2},\ i = 1, \ldots, n.$$

If $m \geq 2$, then the full $2n$-parameter model satisfies the conditions of Theorem 2.3.1. Let $\mathcal{E}$ be the canonical parameter set for this full model and let

$$\Theta = \{\boldsymbol{\theta} : \theta_1 \in R, \theta_2 \in R, \theta_3 > 0\}.$$

Then $c(\Theta)$ is closed in $\mathcal{E}$ and we can conclude that for $m \geq 2$, an MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ exists and $\widehat{\theta}$ satisfies (2.3.7). □

Summary. In this section we derive necessary and sufficient conditions for existence of MLEs in canonical exponential families of full rank with $\mathcal{E}$ open (Theorem 2.3.1 and Corollary 2.3.1). These results lead to a necessary condition for existence of the MLE in curved exponential families but without a guarantee of unicity or sufficiency. Finally, the basic property making Theorem 2.3.1 work, strict concavity, is isolated and shown to apply to a broader class of models.

2.4 ALGORITHMIC ISSUES

As we have seen, even in the context of canonical multiparameter exponential families, such as the two-parameter gamma, MLEs may not be given explicitly by formulae but only implicitly as the solutions of systems of nonlinear equations. In fact, even in the classical regression model with design matrix $\mathbf{Z}_D$ of full rank d, the formula (2.1.10) for $\widehat{\boldsymbol{\beta}}$ is easy to write down symbolically but not easy to evaluate if d is at all large because inversion of $\mathbf{Z}_D^T\mathbf{Z}_D$ requires on the order of nd^2 operations to evaluate each of $d(d+1)/2$ terms with n operations to get $\mathbf{Z}_D^T\mathbf{Z}_D$ and then, if implemented as usual, order d^3 operations to invert. The packages that produce least squares estimates do not in fact use formula (2.1.10).

It is not our goal in this book to enter seriously into questions that are the subject of textbooks in numerical analysis. However, in this section, we will discuss three algorithms of a type used in different statistical contexts both for their own sakes and to illustrate what kinds of things can be established about the black boxes to which we all, at various times, entrust ourselves.

We begin with the bisection and coordinate ascent methods, which give a complete though slow solution to finding MLEs in the canonical exponential families covered by Theorem 2.3.1.

2.4.1 The Method of Bisection

The bisection method is the essential ingredient in the coordinate ascent algorithm that yields MLEs in k-parameter exponential families. Given f continuous on (a, b), $f \uparrow$ strictly, $f(a+) < 0 < f(b-)$, then, by the intermediate value theorem, there exists unique $x^* \epsilon (a, b)$ such that $f(x^*) = 0$. Here, in pseudocode, is the bisection algorithm to find x^*. Given tolerance $\epsilon > 0$ for $|x_{\text{final}} - x^*|$:

Find $x_0 < x_1$, $f(x_0) < 0 < f(x_1)$ by taking $|x_0|, |x_1|$ large enough. Initialize $x^+_{\text{old}} = x_1$, $x^-_{\text{old}} = x_0$.

(1) If $|x^+_{\text{old}} - x^-_{\text{old}}| < 2\epsilon$, $x_{\text{final}} = \frac{1}{2}(x^+_{\text{old}} + x^-_{\text{old}})$ and return x_{final}.

(2) Else, $x_{\text{new}} = \frac{1}{2}(x^+_{\text{old}} + x^-_{\text{old}})$.

(3) If $f(x_{\text{new}}) = 0$, $x_{\text{final}} = x_{\text{new}}$ and return x_{final}.

(4) If $f(x_{\text{new}}) < 0$, $x^-_{\text{old}} = x_{\text{new}}$.

(5) If $f(x_{\text{new}}) > 0$, $x^+_{\text{old}} = x_{\text{new}}$.

Go to (1).

End

Lemma 2.4.1. *The bisection algorithm stops at a solution* x_{final} *such that*

$$|x_{\text{final}} - x^*| \leq \epsilon.$$

Proof. If x_m is the mth iterate of x_{new}

$$|x_m - x_{m-1}| \leq \frac{1}{2}|x_{m-1} - x_{m-2}| \leq \cdots \leq \ldots \frac{1}{2^{m-1}}|x_1 - x_0|. \tag{1}$$

Moreover, by the intermediate value theorem,

$$x_m \leq x^* \leq x_{m+1} \text{ for all } m. \tag{2}$$

Therefore,

$$|x_{m+1} - x^*| \leq 2^{-m}|x_1 - x_0| \tag{3}$$

and $x_m \to x^*$ as $m \to \infty$. Moreover, for $m = \log_2(|x_1 - x_0|/\epsilon)$, $|x_{m+1} - x^*| \leq \epsilon$. □

If desired one could evidently also arrange it so that, in addition, $|f(x_{\text{final}})| \leq \epsilon$.
From this lemma we can deduce the following.

Theorem 2.4.1. *Let* $p(x, \eta)$ *be a one-parameter canonical exponential family generated by* (T, h)*, satisfying the conditions of Theorem* 2.3.1 *and* $T = t_0 \in C^0_T$*, the interior of the convex support of* p_T*. Then, the MLE* $\widehat{\eta}$*, which exists and is unique by Theorem* 2.3.1*, may be found (to tolerance* ϵ*) by the method of bisection applied to*

$$f(\eta) \equiv E_\eta T(X) - t_0.$$

Proof. By Theorem 1.6.4, $f'(\eta) = \text{Var}_\eta T(X) > 0$ for all η so that f is strictly increasing and continuous and necessarily because $\widehat{\eta}$ exists, $f(a+) < 0 < f(b-)$ if $\mathcal{E} = (a, b)$. □

Example 2.4.1. *The Shape Parameter Gamma Family.* Let $X_1, \ldots, X_n$ be i.i.d. $\Gamma(\theta, 1)$,

$$p(x, \theta) = \Gamma^{-1}(\theta)x^{\theta-1}e^{-x},\ x > 0,\ \theta > 0. \tag{2.4.1}$$

Because $T(\mathbf{X}) = \sum_{i=1}^{n} \log X_i$ has a density for all n the MLE always exists. It solves the equation

$$\frac{\Gamma'(\theta)}{\Gamma(\theta)} = \frac{T(\mathbf{X})}{n},$$

which by Theorem 2.4.1 can be evaluated by bisection. This example points to another hidden difficulty. The function $\Gamma(\theta) = \int_0^\infty x^{\theta-1}e^{-x}dx$ needed for the bisection method can itself only be evaluated by numerical integration or some other numerical method. However, it is in fact available to high precision in standard packages such as NAG or MATLAB. In fact, bisection itself is a defined function in some packages. □

2.4.2 Coordinate Ascent

The problem we consider is to solve numerically, for a canonical k-parameter exponential family,

$$E_{\boldsymbol{\eta}}(\mathbf{T}(\mathbf{X})) = \dot{A}(\boldsymbol{\eta}) = \mathbf{t}_0$$

when the MLE $\widehat{\boldsymbol{\eta}} \equiv \widehat{\boldsymbol{\eta}}(t_0)$ exists. Here is the algorithm, which is slow, but as we shall see, always converges to $\widehat{\eta}$.

The case $k = 1$: See Theorem 2.4.1.

The general case: Initialize

$$\widehat{\boldsymbol{\eta}}^0 = (\widehat{\eta}_1^0, \ldots, \widehat{\eta}_k^0).$$

Solve

$$\begin{aligned}
&\text{for } \widehat{\eta}_1^1 : \frac{\partial}{\partial \eta_1} A(\eta_1, \widehat{\eta}_2^0, \ldots, \widehat{\eta}_k^0) = t_1 \\
&\text{for } \widehat{\eta}_2^1 : \frac{\partial}{\partial \eta_2} A(\widehat{\eta}_1^1, \eta_2, \widehat{\eta}_3^0, \ldots, \widehat{\eta}_k^0) = t_2 \\
&\vdots \\
&\text{for } \widehat{\eta}_k^1 : \frac{\partial}{\partial \eta_k} A(\widehat{\eta}_1^1, \widehat{\eta}_2^1, \ldots, \eta_k) = t_k.
\end{aligned}$$

Set

$$\widehat{\boldsymbol{\eta}}^{01} \equiv (\widehat{\eta}_1^1, \widehat{\eta}_2^0, \ldots, \widehat{\eta}_k^0),\ \widehat{\boldsymbol{\eta}}^{02} \equiv (\widehat{\eta}_1^1, \widehat{\eta}_2^1, \widehat{\eta}_3^0, \ldots, \widehat{\eta}_k^0), \text{ and so on,}$$

and finally

$$\widehat{\boldsymbol{\eta}}^{0k} \equiv \widehat{\boldsymbol{\eta}}^{(1)} = (\widehat{\eta}_1^1, \ldots, \widehat{\eta}_k^1).$$

Repeat, getting $\widehat{\boldsymbol{\eta}}^{(r)}$, $r \geq 1$, eventually.

Notes:

(1) In practice, we would again set a tolerence to be, say ϵ, for each of the $\widehat{\boldsymbol{\eta}}^{jl}$, $1 \leq l \leq k$, in cycle j and stop possibly in midcycle as soon as

$$|\widehat{\boldsymbol{\eta}}^{jl} - \widehat{\boldsymbol{\eta}}^{j(l-1)}| \leq \epsilon.$$

(2) Notice that $\frac{\partial A}{\partial \eta_l}(\widehat{\eta}_1^j, \ldots, \widehat{\eta}_{l-2}^j, \eta_l, \widehat{\eta}_{l+1}^{j-1}, \ldots)$ is the expectation of $T_l(\mathbf{X})$ in the one-parameter exponential family model with all parameters save η_l assumed known. Thus, the algorithm may be viewed as successive fitting of one-parameter families. We pursue this discussion next.

Theorem 2.4.2. *If $\widehat{\boldsymbol{\eta}}^{(r)}$ are as above, (i), (ii) of Theorem* 2.3.1 *hold and $t_0 \in C_T^0$,*

$$\widehat{\boldsymbol{\eta}}^{(r)} \to \widehat{\boldsymbol{\eta}} \text{ as } r \to \infty.$$

Proof. We give a series of steps. Let $l(\boldsymbol{\eta}) = \mathbf{t}_0^T \boldsymbol{\eta} - A(\boldsymbol{\eta}) + \log h(\mathbf{x})$, the log likelihood.

(1) $l(\widehat{\boldsymbol{\eta}}^{ij}) \uparrow$ in j for i fixed and in i. If $1 \leq j \leq k$, $\widehat{\boldsymbol{\eta}}^{ij}$ and $\widehat{\boldsymbol{\eta}}^{i(j+1)}$ differ in only one coordinate for which $\widehat{\boldsymbol{\eta}}^{i(j+1)}$ maximizes l. Therefore, $\lim_{i,j} l(\widehat{\boldsymbol{\eta}}^{ij}) = \lambda$ (say) exists and is $> -\infty$.

(2) The sequence $(\widehat{\boldsymbol{\eta}}^{i1}, \ldots, \widehat{\boldsymbol{\eta}}^{ik})$ has a convergent subsequence in $\bar{\mathcal{E}} \times \cdots \times \bar{\mathcal{E}}$

$$(\widehat{\boldsymbol{\eta}}^{i_n 1}, \ldots, \widehat{\boldsymbol{\eta}}^{i_n k}) \to (\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^k).$$

But $\boldsymbol{\eta}^j \in \mathcal{E}$, $1 \leq j \leq k$. Else $\lim_i l(\widehat{\boldsymbol{\eta}}^{ij}) = -\infty$ for some j.

(3) $l(\boldsymbol{\eta}^j) = \lambda$ for all j because the sequence of likelihoods is monotone.

(4) $\frac{\partial l}{\partial \eta_j}(\boldsymbol{\eta}^j) = 0$ because $\frac{\partial l}{\partial \eta_j}(\widehat{\boldsymbol{\eta}}^{i_n j}) = 0, \forall n$.

(5) Because $\boldsymbol{\eta}^1, \boldsymbol{\eta}^2$ differ only in the second coordinate, (3) and (4) $\Rightarrow \boldsymbol{\eta}^1 = \boldsymbol{\eta}^2$. Continuing, $\boldsymbol{\eta}^1 = \cdots = \boldsymbol{\eta}^k$. Here we use the strict concavity of l.

(6) By (4) and (5), $\dot{A}(\boldsymbol{\eta}^1) = \mathbf{t}_0$. Hence, $\boldsymbol{\eta}^1$ is the unique MLE.

To complete the proof notice that if $\widehat{\boldsymbol{\eta}}^{(r_k)}$ is any subsequence of $\widehat{\boldsymbol{\eta}}^{(r)}$ that converges to $\widehat{\boldsymbol{\eta}}^*$ (say) then, by (1), $l(\widehat{\boldsymbol{\eta}}^*) = \lambda$. Because $l(\widehat{\boldsymbol{\eta}}^1) = \lambda$ and the MLE is unique, $\widehat{\boldsymbol{\eta}}^* = \widehat{\boldsymbol{\eta}}^1 = \widehat{\boldsymbol{\eta}}$. By a standard argument it follows that, $\widehat{\boldsymbol{\eta}}^{(r)} \to \widehat{\boldsymbol{\eta}}$. □

Example 2.4.2. *The Two-Parameter Gamma Family (continued).* We use the notation of Example 2.3.2. For $n \geq 2$ we know the MLE exists. We can initialize with the method of moments estimate from Example 2.1.2, $\widehat{\lambda}^{(0)} = \frac{\bar{X}}{\bar{\sigma}^2}$, $\widehat{p}^{(0)} = \frac{\bar{X}^2}{\bar{\sigma}^2}$. We now use bisection to get $\widehat{p}^{(1)}$ solving $\frac{\Gamma'}{\Gamma}(\widehat{p}^{(1)}) = \overline{\log X} + \log \widehat{\lambda}^{(0)}$ and then $\widehat{\lambda}^{(1)} = \frac{\widehat{p}^{(1)}}{\bar{X}}$, $\widehat{\boldsymbol{\eta}}^1 = (\widehat{p}^{(1)}, -\widehat{\lambda}^{(1)})$. Continuing in this way we can get arbitrarily close to $\widehat{\boldsymbol{\eta}}$. This two-dimensional problem is essentially no harder than the one-dimensional problem of Example 2.4.1 because the equation leading to $\widehat{\lambda}_{\text{new}}$ given $\widehat{p}_{\text{old}}$, (2.3.5), is computationally explicit and simple. Whenever we can obtain such steps in algorithms, they result in substantial savings of time. □

It is natural to ask what happens if, in fact, the MLE $\widehat{\boldsymbol{\eta}}$ doesn't exist; that is, $\mathbf{t}_0 \notin C_T^0$. Fortunately in these cases the algorithm, as it should, refuses to converge (in $\boldsymbol{\eta}$ space!)—see Problem 2.4.2.

We note some important generalizations. Consider a point we noted in Example 2.4.2: For some coordinates l, $\widehat{\eta}_l^j$ can be explicit. Suppose that this is true for each l. Then each step of the iteration both within cycles and from cycle to cycle is quick. Suppose that we can write $\boldsymbol{\eta}^T = (\boldsymbol{\eta}_1^T, \ldots, \boldsymbol{\eta}_r^T)$ where $\boldsymbol{\eta}_j$ has dimension d_j and $\sum_{j=1}^r d_j = k$ and the problem of obtaining $\widehat{\boldsymbol{\eta}}_l(\mathbf{t}_0, \boldsymbol{\eta}_j;\ j \neq l)$ can be solved in closed form. The case we have

just discussed has $d_1 = \cdots = d_r = 1$, $r = k$. Then it is easy to see that Theorem 2.4.2 has a generalization with cycles of length r, each of whose members can be evaluated easily. A special case of this is the famous Deming–Stephan proportional fitting of contingency tables algorithm—see Bishop, Feinberg, and Holland (1975), for instance, and Problems 2.4.9–2.4.10.

Next consider the setting of Proposition 2.3.1 in which $l_{\mathbf{x}}(\boldsymbol{\theta})$, the log likelihood for $\boldsymbol{\theta} \in \Theta$ open $\subset R^p$, is strictly concave. If $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ exists and $l_{\mathbf{x}}$ is differentiable, the method extends straightforwardly. Solve $\frac{\partial l_{\mathbf{x}}}{\partial \theta_j}(\theta_1^1, \ldots, \theta_{j-1}^1, \theta_j, \theta_{j+1}^0, \ldots, \theta_p^0) = 0$ by the method of bisection in θ_j to get θ_j^1 for $j = 1, \ldots, p$, iterate and proceed. Figure 2.4.1 illustrates the process. See also Problem 2.4.7.

The coordinate ascent algorithm can be slow if the contours in Figure 2.4.1 are not close to spherical. It can be speeded up at the cost of further computation by Newton's method, which we now sketch.

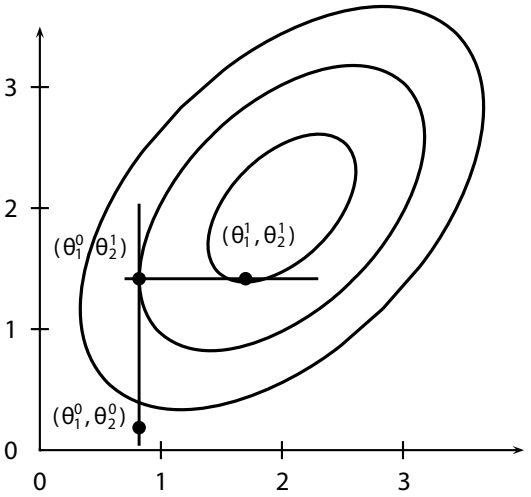

Figure 2.4.1. The coordinate ascent algorithm. The graph shows log likelihood contours, that is, values of $(\theta_1, \theta_2)^T$ where the log likelihood is constant. At each stage with one coordinate fixed, find that member of the family of contours to which the vertical (or horizontal) line is tangent. Change other coordinates accordingly.

2.4.3 The Newton–Raphson Algorithm

An algorithm that, in general, can be shown to be faster than coordinate ascent, when it converges, is the Newton–Raphson method. This method requires computation of the inverse of the Hessian, which may counterbalance its advantage in speed of convergence when it does converge. Here is the method: If $\widehat{\boldsymbol{\eta}}_{\text{old}}$ is the current value of the algorithm,

then

$$\widehat{\boldsymbol{\eta}}_{\text{new}} = \widehat{\boldsymbol{\eta}}_{\text{old}} - \ddot{A}^{-1}(\widehat{\boldsymbol{\eta}}_{\text{old}})(\dot{A}(\widehat{\boldsymbol{\eta}}_{\text{old}}) - \mathbf{t}_0). \tag{2.4.2}$$

The rationale here is simple. If $\widehat{\boldsymbol{\eta}}_{\text{old}}$ is close to the root $\widehat{\boldsymbol{\eta}}$ of $\dot{A}(\widehat{\boldsymbol{\eta}}) = \mathbf{t}_0$, then by expanding $\dot{A}(\widehat{\boldsymbol{\eta}})$ around $\widehat{\boldsymbol{\eta}}_{\text{old}}$, we obtain

$$\mathbf{t}_0 - \dot{A}(\widehat{\boldsymbol{\eta}}_{\text{old}}) = \dot{A}(\widehat{\boldsymbol{\eta}}) - \dot{A}(\widehat{\boldsymbol{\eta}}_{\text{old}}) \simeq \ddot{A}(\widehat{\boldsymbol{\eta}}_{\text{old}})(\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_{\text{old}}).$$

$\widehat{\boldsymbol{\eta}}_{\text{new}}$ is the solution for $\widehat{\boldsymbol{\eta}}$ to the approximation equation given by the right- and left-hand sides. If $\widehat{\boldsymbol{\eta}}_{\text{old}}$ is close enough to $\widehat{\boldsymbol{\eta}}$, this method is known to converge to $\widehat{\boldsymbol{\eta}}$ at a faster rate than coordinate ascent—see Dahlquist, Björk, and Anderson (1974). A hybrid of the two methods that always converges and shares the increased speed of the Newton–Raphson method is given in Problem 2.4.7.

Newton's method also extends to the framework of Proposition 2.3.1. In this case, if $l(\theta)$ denotes the log likelihood, the argument that led to (2.4.2) gives

$$\widehat{\theta}_{\text{new}} = \widehat{\theta}_{\text{old}} - \ddot{l}^{-1}(\widehat{\theta}_{\text{old}})\dot{l}(\widehat{\theta}_{\text{old}}). \tag{2.4.3}$$

Example 2.4.3. Let $X_1, \ldots, X_n$ be a sample from the logistic distribution with d.f.

$$F(x, \theta) = [1 + \exp\{-(x - \theta)\}]^{-1}.$$

The density is

$$f(x, \theta) = \frac{\exp\{-(x - \theta)\}}{[1 + \exp\{-(x - \theta)\}]^2}.$$

We find

$$\begin{aligned} \dot{l}(\theta) &= n - 2\sum_{i=1}^{n} \exp\{-(X_i - \theta)\}F(X_i, \theta) \\ \ddot{l}(\theta) &= -2\sum_{i=1}^{n} f(X_i, \theta) < 0. \end{aligned}$$

The Newton–Raphson method can be implemented by taking $\widehat{\theta}_{\text{old}} = \widehat{\theta}_{\text{MOM}} = \bar{X}$. □

The Newton–Raphson algorithm has the property that for large n, $\widehat{\eta}_{\text{new}}$ after only one step behaves approximately like the MLE. We return to this property in Problem 6.6.10.

When likelihoods are non-concave, methods such as bisection, coordinate ascent, and Newton–Raphson's are still employed, though there is a distinct possibility of nonconvergence or convergence to a local rather than global maximum. A one-dimensional problem in which such difficulties arise is given in Problem 2.4.13. Many examples and important issues and methods are discussed, for instance, in Chapter 6 of Dahlquist, Björk, and Anderson (1974).

2.4.4 The EM (Expectation/Maximization) Algorithm

There are many models that have the following structure. There are ideal observations, $X \sim P_\theta$ with density $p(x,\theta)$, $\theta \in \Theta \subset R^d$. Their log likelihood $l_{p,x}(\theta)$ is "easy" to maximize. Say there is a closed-form MLE or at least $l_{p,x}(\theta)$ is concave in θ. Unfortunately, we observe $S \equiv S(X) \sim Q_\theta$ with density $q(s,\theta)$ where $l_{q,s}(\theta) = \log q(s,\theta)$ is difficult to maximize; the function is not concave, difficult to compute, and so on. A fruitful way of thinking of such problems is in terms of S as representing part of X, the rest of X is "missing" and its "reconstruction" is part of the process of estimating θ by maximum likelihood. The algorithm was formalized with many examples in Dempster, Laird, and Rubin (1977), though an earlier general form goes back to Baum, Petrie, Soules, and Weiss (1970). We give a few examples of situations of the foregoing type in which it is used, and its main properties. For detailed discussion we refer to Little and Rubin (1987) and MacLachlan and Krishnan (1997). A prototypical example follows.

Example 2.4.4. *Lumped Hardy–Weinberg Data.* As in Example 2.2.6, let X_i, $i = 1,\ldots,n$, be a sample from a population in Hardy–Weinberg equilibrium for a two-allele locus, $X_i = (\epsilon_{i1},\epsilon_{i2},\epsilon_{i3})$, where $P_\theta[X = (1,0,0)] = \theta^2$, $P_\theta[X = (0,1,0)] = 2\theta(1-\theta)$, $P_\theta[X = (0,0,1)] = (1-\theta)^2$, $0 < \theta < 1$. What is observed, however, is not $\mathbf{X}$ but $\mathbf{S}$ where

$$\begin{aligned} S_i &= X_i,\ 1 \le i \le m \\ S_i &= (\epsilon_{i1}+\epsilon_{i2},\epsilon_{i3}),\ m+1 \le i \le n. \end{aligned} \tag{2.4.4}$$

Evidently, $\mathbf{S} = \mathbf{S}(\mathbf{X})$ where $\mathbf{S}(\mathbf{X})$ is given by (2.4.4). This could happen if, for some individuals, the homozygotes of one type ($\epsilon_{i1} = 1$) could not be distinguished from the heterozygotes ($\epsilon_{i2} = 1$). The log likelihood of $\mathbf{S}$ now is

$$\begin{aligned} l_{q,s}(\theta) &= \sum_{i=1}^{m}[2\epsilon_{i1}\log\theta + \epsilon_{i2}\log 2\theta(1-\theta) + 2\epsilon_{i3}\log(1-\theta)] \\ &\quad + \sum_{i=m+1}^{n}[(\epsilon_{i1}+\epsilon_{i2})\log(1-(1-\theta)^2) + 2\epsilon_{i3}\log(1-\theta)] \end{aligned} \tag{2.4.5}$$

a function that is of curved exponential family form. It does turn out that in this simplest case an explicit maximum likelihood solution is still possible, but the computation is clearly not as simple as in the original Hardy–Weinberg canonical exponential family example. If we suppose (say) that observations $S_1,\ldots,S_m$ are not X_i but $(\epsilon_{i1},\epsilon_{i2}+\epsilon_{i3})$, then explicit solution is in general not possible. Yet the EM algorithm, with an appropriate starting point, leads us to an MLE if it exists in both cases. □

Here is another important example.

Example 2.4.5. *Mixture of Gaussians.* Suppose $S_1,\ldots,S_n$ is a sample from a population P whose density is modeled as a mixture of two Gaussian densities, $p(s,\theta) = (1-\lambda)\varphi_{\sigma_1}(s-\mu_1) + \lambda\varphi_{\sigma_2}(s-\mu_2)$ where $\boldsymbol{\theta} = (\lambda,(\mu_i,\sigma_i), i = 1,2)$ and $0 < \lambda < 1$, $\sigma_1,\sigma_2 > 0$, $\mu_1,\mu_2 \in R$ and $\varphi_\sigma(s) = \frac{1}{\sigma}\varphi\left(\frac{s}{\sigma}\right)$. It is not obvious that this falls under our

scheme but let

$$X_i = (\Delta_i, S_i),\ 1 \le i \le n \tag{2.4.6}$$

where Δ_i are independent identically distributed with $P_{\boldsymbol{\theta}}[\Delta_i = 1] = \lambda = 1 - P_{\boldsymbol{\theta}}[\Delta_i = 0]$. Suppose that given $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_n)$, the S_i are independent with

$$\mathcal{L}_{\boldsymbol{\theta}}(S_i \mid \boldsymbol{\Delta}) = \mathcal{L}_{\boldsymbol{\theta}}(S_i \mid \Delta_i) = \mathcal{N}(\Delta_i\mu_1 + (1-\Delta_i)\mu_2, \Delta_i\sigma_1^2 + (1-\Delta_i)\sigma_2^2).$$

That is, Δ_i tells us whether to sample from $\mathcal{N}(\mu_1, \sigma_1^2)$ or $\mathcal{N}(\mu_2, \sigma_2^2)$. It is easy to see (Problem 2.4.11), that under $\boldsymbol{\theta}$, $\mathbf{S}$ has the marginal distribution given previously. Thus, we can think of $\mathbf{S}$ as $S(\mathbf{X})$ where $\mathbf{X}$ is given by (2.4.6).

This five-parameter model is very rich permitting up to two modes and scales. The log likelihood similarly can have a number of local maxima and can tend to ∞ as $\boldsymbol{\theta}$ tends to the boundary of the parameter space (Problem 2.4.12). Although MLEs do not exist in these models, a local maximum close to the true $\boldsymbol{\theta}_0$ turns out to be a good "proxy" for the nonexistent MLE. The EM algorithm can lead to such a local maximum. □

The EM Algorithm. Here is the algorithm. Let

$$J(\theta \mid \theta_0) \equiv E_{\theta_0}\left(\log \frac{p(X,\theta)}{p(X,\theta_0)} \mid S(X) = s\right) \tag{2.4.7}$$

where we suppress dependence on s.

Initialize with $\theta_{\text{old}} = \theta_0$.

The first (E) step of the algorithm is to compute $J(\theta \mid \theta_{\text{old}})$ for as many values of θ as needed. If this is difficult, the EM algorithm is probably not suitable.

The second (M) step is to maximize $J(\theta \mid \theta_{\text{old}})$ as a function of θ. Again, if this step is difficult, EM is not particularly appropriate.

Then we set $\theta_{\text{new}} = \arg\max J(\theta \mid \theta_{\text{old}})$, reset $\theta_{\text{old}} = \theta_{\text{new}}$ and repeat the process.

As we shall see in important situations, including the examples, we have given, the M step is easy and the E step doable.

The rationale behind the algorithm lies in the following formulas, which we give for θ real and which can be justified easily in the case that $\mathcal{X}$ is finite (Problem 2.4.12)

$$\frac{q(s,\theta)}{q(s,\theta_0)} = E_{\theta_0}\left(\frac{p(X,\theta)}{p(X,\theta_0)} \mid S(X) = s\right) \tag{2.4.8}$$

and

$$\frac{\partial}{\partial\theta}\log q(s,\theta)\Big|_{\theta=\theta_0} = E_{\theta_0}\left(\frac{\partial}{\partial\theta}\log p(X,\theta) \mid S(X) = s\right)\Big|_{\theta=\theta_0} \tag{2.4.9}$$

for all θ_0 (under suitable regularity conditions). Note that (2.4.9) follows from (2.4.8) by taking logs in (2.4.8), differentiating and exchanging E_{θ_0} and differentiation with respect to θ at θ_0. Because, formally,

$$\frac{\partial J(\theta \mid \theta_0)}{\partial\theta} = E_{\theta_0}\left(\frac{\partial}{\partial\theta}\log p(X,\theta) \mid S(X) = s\right) \tag{2.4.10}$$

and, hence,

$$\left.\frac{\partial J(\theta \mid \theta_0)}{\partial \theta}\right|_{\theta_0} = \frac{\partial}{\partial \theta} \log q(s, \theta_0) \tag{2.4.11}$$

it follows that a fixed point $\widetilde{\theta}$ of the algorithm satisfies the likelihood equation,

$$\frac{\partial}{\partial \theta} \log q(s, \widetilde{\theta}) = 0. \tag{2.4.12}$$

The main reason the algorithm behaves well follows.

Lemma 2.4.1. *If $\theta_{\text{new}}, \theta_{\text{old}}$ are as defined earlier and $S(X) = s$,*

$$q(s, \theta_{\text{new}}) \geq q(s, \theta_{\text{old}}). \tag{2.4.13}$$

Equality holds in (2.4.13) *iff the conditional distribution of X given $S(X) = s$ is the same for θ_{new} as for θ_{old} and θ_{old} maximizes $J(\theta \mid \theta_{\text{old}})$.*

Proof. We give the proof in the discrete case. However, the result holds whenever the quantities in $J(\theta \mid \theta_0)$ can be defined in a reasonable fashion. In the discrete case we appeal to the product rule. For $x \in \mathcal{X}$, $S(x) = s$

$$p(x, \theta) = q(s, \theta) r(x \mid s, \theta) \tag{2.4.14}$$

where $r(\cdot \mid \cdot, \theta)$ is the conditional frequency function of X given $S(X) = s$. Then

$$J(\theta \mid \theta_0) = \log \frac{q(s, \theta)}{q(s, \theta_0)} + E_{\theta_0}\left\{ \log \frac{r(X \mid s, \theta)}{r(X \mid s, \theta_0)} \mid S(X) = s \right\}. \tag{2.4.15}$$

If $\theta_0 = \theta_{\text{old}}$, $\theta = \theta_{\text{new}}$,

$$\log \frac{q(s, \theta_{\text{new}})}{q(s, \theta_{\text{old}})} = J(\theta_{\text{new}} \mid \theta_{\text{old}}) - E_{\theta_{\text{old}}}\left\{ \log \frac{r(X \mid s, \theta_{\text{new}})}{r(X \mid s, \theta_{\text{old}})} \mid S(X) = s \right\}. \tag{2.4.16}$$

Now, $J(\theta_{\text{new}} \mid \theta_{\text{old}}) \geq J(\theta_{\text{old}} \mid \theta_{\text{old}}) = 0$ by definition of θ_{new}. On the other hand,

$$-E_{\theta_{\text{old}}}\left\{ \log \frac{r(X \mid s, \theta_{\text{new}})}{r(X \mid s, \theta_{\text{old}})} \mid S(X) = s \right\} \geq 0 \tag{2.4.17}$$

by Shannon's inequality, Lemma 2.2.1. □

The most important and revealing special case of this lemma follows.

Theorem 2.4.3. *Suppose $\{P_\theta : \theta \in \Theta\}$ is a canonical exponential family generated by (T, h) satisfying the conditions of Theorem* 2.3.1. *Let $S(X)$ be any statistic, then*

(a) *The EM algorithm consists of the alternation*

$$\dot{A}(\theta_{\text{new}}) = E_{\theta_{\text{old}}}(T(X) \mid S(X) = s) \tag{2.4.18}$$

$$\theta_{\text{old}} = \theta_{\text{new}}. \tag{2.4.19}$$

If a solution of (2.4.18) *exists it is necessarily unique.*

(b) *If the sequence of iterates* $\{\widehat{\theta}_m\}$ *so obtained is bounded and the equation*

$$\dot{A}(\theta) = E_\theta(T(X) \mid S(X) = s) \tag{2.4.20}$$

has a unique solution, then it converges to a limit $\widehat{\theta}^*$*, which is necessarily a local maximum of* $q(s, \theta)$.

Proof. In this case,

$$\begin{aligned} J(\theta \mid \theta_0) &= E_{\theta_0}\{(\theta - \theta_0)^T T(X) - (A(\theta) - A(\theta_0)) \mid S(X) = s\} \\ &= (\theta - \theta_0)^T E_{\theta_0}(T(X) \mid S(X) = y) - (A(\theta) - A(\theta_0)) \end{aligned} \tag{2.4.21}$$

Part (a) follows.

Part (b) is more difficult. A proof due to Wu (1983) is sketched in Problem 2.4.16. □

Example 2.4.4 (continued). $\mathbf{X}$ is distributed according to the exponential family

$$p(\mathbf{x}, \theta) = \exp\{\eta(2N_{1n}(\mathbf{x}) + N_{2n}(\mathbf{x})) - A(\boldsymbol{\eta})\}h(\mathbf{x}) \tag{2.4.22}$$

where

$$\eta = \log\left(\frac{\theta}{1-\theta}\right),\ h(\mathbf{x}) = 2^{N_{2n}(\mathbf{x})},\ A(\eta) = 2n\log(1 + e^\eta)$$

and $N_{jn} = \sum_{i=1}^n \epsilon_{ij}(x_i)$, $1 \le j \le 3$. Now,

$$A'(\eta) = 2n\theta \tag{2.4.23}$$

$$\begin{aligned} E_\theta(2N_{1n} + N_{2n} \mid \mathbf{S}) &= 2N_{1m} + N_{2m} \\ &\quad + E_\theta\left(\sum_{i=m+1}^n (2\epsilon_{i1} + \epsilon_{i2}) \mid \epsilon_{i1} + \epsilon_{i2},\ m+1 \le i \le n\right). \end{aligned} \tag{2.4.24}$$

Under the assumption that the process that causes lumping is independent of the values of the ϵ_{ij},

$$\begin{aligned} P_\theta[\epsilon_{ij} = 1 \mid \epsilon_{i1} + \epsilon_{i2} = 0] &= 0,\ 1 \le j \le 2 \\ P_\theta[\epsilon_{i1} = 1 \mid \epsilon_{i1} + \epsilon_{i2} = 1] &= \frac{\theta^2}{\theta^2 + 2\theta(1-\theta)} = \frac{\theta^2}{1-(1-\theta)^2} \\ &= 1 - P_\theta[\epsilon_{i2} = 1 \mid \epsilon_{i1} + \epsilon_{i2} = 1]. \end{aligned}$$

Thus, we see, after some simplification, that,

$$E_\theta(2N_{1n} + N_{2n} \mid \mathbf{S}) = 2N_{1m} + N_{2m} + \frac{2}{2 - \widehat{\theta}_{\text{old}}} M_n \tag{2.4.25}$$

where

$$M_n = \sum_{i=m+1}^n (\epsilon_{i1} + \epsilon_{i2}).$$

Thus, the EM iteration is

$$\widehat{\theta}_{\text{new}} = \frac{2N_{1m} + N_{2m}}{2n} + \frac{M_n}{n}(2 - \widehat{\theta}_{\text{old}})^{-1}. \tag{2.4.26}$$

It may be shown directly (Problem 2.4.15) that if $2N_{1m} + N_{2m} > 0$ and $M_n > 0$, then $\widehat{\theta}_m$ converges to the unique root of

$$\theta^2 - \left(\frac{2 + 2N_{1m} + N_{2m}}{2n}\right)\theta + \frac{2N_{1m} + N_{2m} + M_n}{n} = 0$$

in $(0, 1)$, which is indeed the MLE when $\mathbf{S}$ is observed. □

Example 2.4.6. Let $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ be i.i.d. as (Z, Y), where $(Z, Y) \sim \mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Suppose that some of the Z_i and some of the Y_i are missing as follows: For $1 \leq i \leq n_1$ we observe both Z_i and Y_i, for $n_1 + 1 \leq i \leq n_2$, we oberve only Z_i, and for $n_2 + 1 \leq i \leq n$, we observe only Y_i. In this case a set of sufficient statistics is

$$T_1 = \bar{Z},\ T_2 = \bar{Y},\ T_3 = n^{-1}\sum_{i=1}^{n} Z_i^2,\ T_4 = n^{-1}\sum_{i=1}^{n} Y_i^2,\ T_5 = n^{-1}\sum_{i=1}^{n} Z_i Y_i.$$

The observed data are

$$S = \{(Z_i, Y_i) : 1 \leq i \leq n_1\} \cup \{Z_i : n_1 + 1 \leq i \leq n_2\} \cup \{Y_i : n_2 + 1 \leq i \leq n\}.$$

To compute $E_\theta(\mathbf{T} \mid S = s)$, where $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \theta)$, we note that for the cases with Z_i and/or Y_i observed, the conditional expected values equal their observed values. For other cases we use the properties of the bivariate normal distribution (Appendix B.4 and Section 1.4), to conclude

$$\begin{aligned} E_\theta(Y_i \mid Z_i) &= \mu_2 + \rho\sigma_2(Z_i - \mu_1)/\sigma_1 \\ E_\theta(Y_i^2 \mid Z_i) &= [\mu_2 + \rho\sigma_2(Z_i - \mu_1)/\sigma_1]^2 + (1 - \rho^2)\sigma_2^2 \\ E_\theta(Z_i Y_i \mid Z_i) &= [\mu_2 + \rho\sigma_2(Z_i - \mu_1)/\sigma_1]Z_i \end{aligned}$$

with the corresponding Z on Y regression equations when conditioning on Y_i (Problem 2.4.1). This completes the E-step. For the M-step, compute (Problem 2.4.1)

$$\dot{A}(\theta) = E_\theta \mathbf{T} = (\mu_1, \mu_2, \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, \sigma_1\sigma_2\rho + \mu_1\mu_2).$$

We take $\widehat{\theta}_{\text{old}} = \widehat{\theta}_{\text{MOM}}$, where $\widehat{\theta}_{\text{MOM}}$ is the method of moment estimates $(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1^2, \widehat{\sigma}_2^2, r)$ (Problem 2.1.8) of θ based on the observed data. It may be shown directly (Problem 2.4.1) that the M-step produces

$$\begin{aligned} \widehat{\mu}_{1,\text{new}} &= T_1(\widehat{\theta}_{\text{old}}),\ \widehat{\mu}_{2,\text{new}} = T_2(\widehat{\theta}_{\text{old}}),\ \widehat{\sigma}_{1,\text{new}}^2 = T_3(\widehat{\theta}_{\text{old}}) - \widehat{T}_1^2 \\ \widehat{\sigma}_{2,\text{new}}^2 &= T_4(\widehat{\theta}_{\text{old}}) - \widehat{T}_2^2, \\ \widehat{\rho}_{\text{new}} &= [T_5(\widehat{\theta}_{\text{old}}) - \widehat{T}_1\widehat{T}_2]/\{[T_3(\widehat{\theta}_{\text{old}}) - \widehat{T}_1][\widehat{T}_4(\widehat{\theta}_{\text{old}}) - \widehat{T}_2]\}^{\frac{1}{2}} \end{aligned} \tag{2.4.27}$$

where $T_j(\theta)$ denotes T_j with missing values replaced by the values computed in the E-step and $\widehat{T}_j = T_j(\widehat{\theta}_{\text{old}})$, $j = 1, 2$. Now the process is repeated with $\widehat{\theta}_{\text{MOM}}$ replaced by $\widehat{\theta}_{\text{new}}$. □

Because the E-step, in the context of Example 2.4.6, involves imputing missing values, the EM algorithm is a form of *multiple imputation*.

Remark 2.4.1. Note that if $S(X) = X$, then $J(\theta \mid \theta_0)$ is $\log[p(X,\theta)/p(X,\theta_0)]$, which as a function of θ is maximized where the contrast $-\log p(X,\theta)$ is minimized. Also note that, in general, $-E_{\theta_0}[J(\theta \mid \theta_0)]$ is the Kullback–Leibler divergence (2.2.23).

Summary. The basic bisection algorithm for finding roots of monotone functions is developed and shown to yield a rapid way of computing the MLE in all one-parameter canonical exponential families with $\mathcal{E}$ open (when it exists). We then, in Section 2.4.2, use this algorithm as a building block for the general coordinate ascent algorithm, which yields with certainty the MLEs in k-parameter canonical exponential families with $\mathcal{E}$ open when it exists. Important variants of and alternatives to this algorithm, including the Newton–Raphson method, are discussed and introduced in Section 2.4.3 and the problems. Finally in Section 2.4.4 we derive and discuss the important EM algorithm and its basic properties.

2.5 PROBLEMS AND COMPLEMENTS

Problems for Section 2.1

1. Consider a population made up of three different types of individuals occurring in the Hardy–Weinberg proportions θ^2, $2\theta(1-\theta)$ and $(1-\theta)^2$, respectively, where $0 < \theta < 1$.

(a) Show that $T_3 = N_1/n + N_2/2n$ is a frequency substitution estimate of θ.

(b) Using the estimate of (a), what is a frequency substitution estimate of the odds ratio $\theta/(1-\theta)$?

(c) Suppose X takes the values $-1, 0, 1$ with respective probabilities p_1, p_2, p_3 given by the Hardy–Weinberg proportions. By considering the first moment of X, show that T_3 is a method of moment estimate of θ.

2. Consider n systems with failure times $X_1, \ldots, X_n$ assumed to be independent and identically distributed with exponential, $\mathcal{E}(\lambda)$, distributions.

(a) Find the method of moments estimate of λ based on the first moment.

(b) Find the method of moments estimate of λ based on the second moment.

(c) Combine your answers to (a) and (b) to get a method of moment estimate of λ based on the first two moments.

(d) Find the method of moments estimate of the probability $P(X_1 \geq 1)$ that one system will last at least a month.

3. Suppose that i.i.d. $X_1, \ldots, X_n$ have a beta, $\beta(\alpha_1, \alpha_2)$ distribution. Find the method of moments estimates of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ based on the first two moments.

Hint: See Problem B.2.5.

4. Let $X_1, \ldots, X_n$ be the indicators of n Bernoulli trials with probability of success θ.

(a) Show that $\bar{X}$ is a method of moments estimate of θ.

(b) Exhibit method of moments estimates for $\mathrm{Var}_\theta \bar{X} = \theta(1-\theta)/n$ first using only the first moment and then using only the second moment of the population. Show that these estimates coincide.

(c) Argue that in this case all frequency substitution estimates of $q(\theta)$ must agree with $q(\bar{X})$.

5. Let $X_1, \ldots, X_n$ and θ be as in Problem 2.1.4. Define $\psi : R^n \times (0,1) \to R$ by

$$\psi(X_1, \ldots, X_n, \theta) = (S/\theta) - (n-S)/(1-\theta)$$

where $S = \sum X_i$. Find V as defined by (2.1.3) and show that θ_0 is the unique solution of $V(\theta, \theta_0) = 0$. Find the estimating equation estimate of θ_0.

6. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the order statistics of a sample $X_1, \ldots, X_n$. (See Problem B.2.8.) There is a one-to-one correspondence between the empirical distribution function $\widehat{F}$ and the order statistics in the sense that, given the order statistics we may construct $\widehat{F}$ and given $\widehat{F}$, we know the order statistics. Give the details of this correspondence.

7. The jth cumulant $\widehat{c}_j$ of the empirical distribution function is called the jth *sample cumulant* and is a method of moments estimate of the cumulant c_j. Give the first three sample cumulants. See A.12.

8. Let $(Z_1, Y_1), (Z_2, Y_2), \ldots, (Z_n, Y_n)$ be a set of independent and identically distributed random vectors with common distribution function F. The natural estimate of $F(s,t)$ is the *bivariate empirical* distribution function $\widehat{F}(s,t)$, which we define by

$$\widehat{F}(s,t) = \frac{\text{Number of vectors } (Z_i, Y_i) \text{ such that } Z_i \leq s \text{ and } Y_i \leq t}{n}.$$

(a) Show that $\widehat{F}(\cdot,\cdot)$ is the distribution function of a probability $\widehat{P}$ on R^2 assigning mass $1/n$ to each point (Z_i, Y_i).

(b) Define the sample product moment of order (i,j), the sample covariance, the sample correlation, and so on, as the corresponding characteristics of the distribution $\widehat{F}$. Show that the sample product moment of order (i,j) is given by

$$\frac{1}{n}\sum_{k=1}^{n} Z_k^i Y_k^j.$$

The sample covariance is given by

$$\frac{1}{n}\sum_{k=1}^{n}(Z_k - \bar{Z})(Y_k - \bar{Y}) = \frac{1}{n}\sum_{k=1}^{n} Z_k Y_k - \bar{Z}\bar{Y},$$

where $\bar{Z}, \bar{Y}$ are the sample means of the $Z_1, \ldots, Z_n$ and $Y_1, \ldots, Y_n$, respectively. The sample correlation coefficient is given by

$$r = \frac{\sum_{k=1}^n (Z_k - \bar{Z})(Y_k - \bar{Y})}{\sqrt{\sum_{k=1}^n (Z_k - \bar{Z})^2 \sum_{k=1}^n (Y_k - \bar{Y})^2}}.$$

All of these quantities are natural estimates of the corresponding population characteristics and are also called method of moments estimates. (See Problem 2.1.17.) Note that it follows from (A.11.19) that $-1 \leq r \leq 1$.

9. Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ where the X_i are independent $\mathcal{N}(0, \sigma^2)$.

(a) Find an estimate of σ^2 based on the second moment.

(b) Construct an estimate of σ using the estimate of part (a) and the equation $\sigma = \sqrt{\sigma^2}$.

(c) Use the empirical substitution principle to construct an estimate of σ using the relation $E(|X_1|) = \sigma\sqrt{2\pi}$.

10. In Example 2.1.1, suppose that $g(\boldsymbol{\beta}, \mathbf{z})$ is continuous in $\boldsymbol{\beta}$ and that $|g(\boldsymbol{\beta}, \mathbf{z})|$ tends to ∞ as $|\boldsymbol{\beta}|$ tends to ∞. Show that the least squares estimate exists.

Hint: Set $c = \rho(X, \mathbf{0})$. There exists a compact set K such that for $\boldsymbol{\beta}$ in the complement of K, $\rho(X, \boldsymbol{\beta}) > c$. Since $\rho(X, \boldsymbol{\beta})$ is continuous on K, the result follows.

11. In Example 2.1.2 with $X \sim \Gamma(\alpha, \lambda)$, find the method of moments estimate based on $\widehat{\mu}_1$ and $\widehat{\mu}_3$.

Hint: See Problem B.2.4.

12. Let $X_1, \ldots, X_n$ be i.i.d. as $X \sim P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta \subset R^d$, with $\boldsymbol{\theta}$ identifiable. Suppose X has possible values $v_1, \ldots, v_k$ and that $q(\boldsymbol{\theta})$ can be written as

$$q(\boldsymbol{\theta}) = h(\mu_1(\boldsymbol{\theta}), \ldots, \mu_r(\boldsymbol{\theta}))$$

for some R^k-valued function h. Show that the method of moments estimate $\widehat{q} = h(\widehat{\mu}_1, \ldots, \widehat{\mu}_r)$ can be written as a frequency plug-in estimate.

13. *General method of moment estimates*[(1)]. Suppose $X_1, \ldots, X_n$ are i.i.d. as $X \sim P_{\boldsymbol{\theta}}$, with $\boldsymbol{\theta} \in \Theta \subset R^d$ and $\boldsymbol{\theta}$ identifiable. Let $g_1, \ldots, g_r$ be given linearly independent functions and write

$$\mu_j(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(g_j(X)),\ \widehat{\mu}_j = n^{-1} \sum_{i=1}^n g_j(X_i),\ j = 1, \ldots, r.$$

Suppose that X has possible values $v_1, \ldots, v_k$ and that

$$q(\boldsymbol{\theta}) = h(\mu_1(\boldsymbol{\theta}), \ldots, \mu_r(\boldsymbol{\theta}))$$

for some R^k-valued function h.

(a) Show that the method of moments estimate $\widehat{q} = h(\widehat{\mu}_1, \ldots, \widehat{\mu}_r)$ is a frequency plug-in estimate.

(b) Suppose $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is the k-parameter exponential family given by (1.6.10). Let $g_j(X) = T_j(X)$, $1 \leq j \leq k$. In the following cases, find the method of moments estimates

(i) Beta, $\beta(1, \theta)$

(ii) Beta, $\beta(\theta, 1)$

(iii) Rayleigh, $p(x, \theta) = (x/\theta^2)\exp(-x^2/2\theta^2)$, $x > 0$, $\theta > 0$

(iv) Gamma, $\Gamma(p, \theta)$, p fixed

(v) Inverse Gaussian, $IG(\mu, \lambda)$, $\boldsymbol{\theta} = (\mu, \lambda)$. See Problem 1.6.36.

Hint: Use Corollary 1.6.1.

14. When the data are not i.i.d., it may still be possible to express parameters as functions of moments and then use estimates based on replacing population moments with "sample" moments. Consider the Gaussian $AR(1)$ model of Example 1.1.5.

(a) Use $E(X_i)$ to give a method of moments estimate of μ.

(b) Suppose $\mu = \mu_0$ and $\beta = b$ are fixed. Use $E(U_i^2)$, where

$$U_i = (X_i - \mu_0) \Bigg/ \left(\sum_{j=0}^{i-1} b^{2j} \right)^{1/2},$$

to give a method of moments estimate of σ^2.

(c) If μ and σ^2 are fixed, can you give a method of moments estimate of β?

15. *Hardy–Weinberg with six genotypes.* In a large natural population of plants (Mimulus guttatus) there are three possible alleles S, I, and F at one locus resulting in six genotypes labeled SS, II, FF, SI, SF, and IF. Let θ_1, θ_2, and θ_3 denote the probabilities of S, I, and F, respectively, where $\sum_{j=1}^{3} \theta_j = 1$. The Hardy–Weinberg model specifies that the six genotypes have probabilities

Genotype	1	2	3	4	5	6
Genotype	SS	II	FF	SI	SF	IF
Probability	θ_1^2	θ_2^2	θ_3^2	$2\theta_1\theta_2$	$2\theta_1\theta_3$	$2\theta_2\theta_3$

Let N_j be the number of plants of genotype j in a sample of n independent plants, $1 \leq j \leq 6$ and let $\widehat{p}_j = N_j/n$. Show that

$$\begin{aligned} \widehat{\theta}_1 &= \widehat{p}_1 + \tfrac{1}{2}\widehat{p}_4 + \tfrac{1}{2}\widehat{p}_5 \\ \widehat{\theta}_2 &= \widehat{p}_2 + \tfrac{1}{2}\widehat{p}_4 + \tfrac{1}{2}\widehat{p}_6 \\ \widehat{\theta}_3 &= \widehat{p}_3 + \tfrac{1}{2}\widehat{p}_5 + \tfrac{1}{2}\widehat{p}_6 \end{aligned}$$

are frequency plug-in estimates of θ_1, θ_2, and θ_3.

16. Establish (2.1.6).
Hint: $[Y_i - g(\boldsymbol{\beta}, \boldsymbol{z}_i)] = [Y_i - g(\boldsymbol{\beta}_0, \boldsymbol{z}_i)] + [g(\boldsymbol{\beta}_0, \boldsymbol{z}_i) - g(\boldsymbol{\beta}, \boldsymbol{z}_i)]$.

17. *Multivariate method of moments.* For a vector $X = (X_1, \ldots, X_q)$, of observations, let the moments be

$$m_{jkrs} = E(X_r^j X_s^k),\ j \geq 0,\ k \geq 0;\ r, s = 1, \ldots, q.$$

For independent identically distributed $X_i = (X_{i1}, \ldots, X_{iq})$, $i = 1, \ldots, n$, we define the empirical or sample moment to be

$$\widehat{m}_{jkrs} = \frac{1}{n} \sum_{i=1}^{n} X_{ir}^j X_{is}^k,\ j \geq 0,\ k \geq 0;\ r, s = 1, \ldots, q.$$

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ can be expressed as a function of the moments, the method of moments estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained by replacing m_{jkrs} by $\widehat{m}_{jkrs}$. Let $X = (Z, Y)$ and $\boldsymbol{\theta} = (a_1, b_1)$, where (Z, Y) and (a_1, b_1) are as in Theorem 1.4.3. Show that method of moments estimators of the parameters b_1 and a_1 in the best linear predictor are

$$\widehat{b}_1 = \frac{n^{-1} \sum Z_i Y_i - \bar{Z}\bar{Y}}{n^{-1} \sum Z_i - (\bar{Z})^2},\ \widehat{a}_1 = \bar{Y} - \widehat{b}_1 \bar{Z}.$$

Problems for Section 2.2

1. An object of unit mass is placed in a force field of unknown constant intensity θ. Readings $Y_1, \ldots, Y_n$ are taken at times $t_1, \ldots, t_n$ on the position of the object. The reading Y_i differs from the true position $(\theta/2)t_i^2$ by a random error ϵ_i. We suppose the ϵ_i to have mean 0 and be uncorrelated with constant variance. Find the LSE of θ.

2. Show that the formulae of Example 2.2.2 may be derived from Theorem 1.4.3, if we consider the distribution assigning mass $1/n$ to each of the points $(z_1, y_1), \ldots, (z_n, y_n)$.

3. Suppose that observations $Y_1, \ldots, Y_n$ have been taken at times $z_1, \ldots, z_n$ and that the linear regression model holds. A new observation Y_{n+1} is to be taken at time z_{n+1}. What is the least squares estimate based on $Y_1, \ldots, Y_n$ of the best (MSPE) predictor of Y_{n+1}?

4. Show that the two sample regression lines coincide (when the axes are interchanged) if and only if the points (z_i, y_i), $i = 1, \ldots, n$, in fact, all lie on a line.
Hint: Write the lines in the form

$$\frac{(z - \bar{z})}{\widehat{\sigma}} = \widehat{\rho} \frac{(y - \bar{y})}{\widehat{\tau}}.$$

5. The regression line minimizes the sum of the squared vertical distances from the points $(z_1, y_1), \ldots, (z_n, y_n)$. Find the line that minimizes the sum of the squared *perpendicular* distance to the same points.

Hint: The quantity to be minimized is

$$\frac{\sum_{i=1}^{n}(y_i - \theta_1 - \theta_2 z_i)^2}{1+\theta_2^2}.$$

6. **(a)** Let $Y_1, \ldots, Y_n$ be independent random variables with equal variances such that $E(Y_i) = \alpha z_i$ where the z_i are known constants. Find the least squares estimate of α.

(b) Relate your answer to the formula for the best zero intercept linear predictor of Section 1.4.

7. Show that the least squares estimate is always defined and satisfies the equations (2.1.7) provided that g is differentiable with respect to β_i, $1 \leq i \leq d$, the range $\{g(\mathbf{z}_1, \boldsymbol{\beta}), \ldots, g(\mathbf{z}_n, \boldsymbol{\beta}), \boldsymbol{\beta} \in R^d\}$ is closed, and $\boldsymbol{\beta}$ ranges over R^d.

8. Find the least squares estimates for the model $Y_i = \theta_1 + \theta_2 z_i + \epsilon_i$ with ϵ_i as given by (2.2.4)–(2.2.6) under the restrictions $\theta_1 \geq 0$, $\theta_2 \leq 0$.

9. Suppose $Y_i = \theta_1 + \epsilon_i$, $i = 1, \ldots, n_1$ and $Y_i = \theta_2 + \epsilon_i$, $i = n_1 + 1, \ldots, n_1 + n_2$, where $\epsilon_1, \ldots, \epsilon_{n_1+n_2}$ are independent $\mathcal{N}(0, \sigma^2)$ variables. Find the least squares estimates of θ_1 and θ_2.

10. Let $X_1, \ldots, X_n$ denote a sample from a population with one of the following densities or frequency functions. Find the MLE of θ.

(a) $f(x, \theta) = \theta e^{-\theta x}$, $x \geq 0$; $\theta > 0$. (exponential density)

(b) $f(x, \theta) = \theta c^{\theta} x^{-(\theta+1)}$, $x \geq c$; c constant > 0; $\theta > 0$. (Pareto density)

(c) $f(x, \theta) = c\theta^c x^{-(c+1)}$, $x \geq \theta$; c constant > 0; $\theta > 0$. (Pareto density)

(d) $f(x, \theta) = \sqrt{\theta} x^{\sqrt{\theta}-1}$, $0 \leq x \leq 1$, $\theta > 0$. (beta, $\beta(\sqrt{\theta}, 1)$, density)

(e) $f(x, \theta) = (x/\theta^2) \exp\{-x^2/2\theta^2\}$, $x > 0$; $\theta > 0$. (Rayleigh density)

(f) $f(x, \theta) = \theta c x^{c-1} \exp\{-\theta x^c\}$, $x \geq 0$; c constant > 0; $\theta > 0$. (Weibull density)

11. Suppose that $X_1, \ldots, X_n$, $n \geq 2$, is a sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

(a) Show that if μ and σ^2 are unknown, $\mu \in R$, $\sigma^2 > 0$, then the unique MLEs are $\widehat{\mu} = \bar{X}$ and $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

(b) Suppose μ and σ^2 are both known to be nonnegative but otherwise unspecified. Find maximum likelihood estimates of μ and σ^2.

12. Let $X_1, \ldots, X_n$, $n \geq 2$, be independently and identically distributed with density

$$f(x, \theta) = \frac{1}{\sigma} \exp\{-(x - \mu)/\sigma\}, \; x \geq \mu,$$

where $\theta = (\mu, \sigma^2)$, $-\infty < \mu < \infty$, $\sigma^2 > 0$.

(a) Find maximum likelihood estimates of μ and σ^2.

(b) Find the maximum likelihood estimate of $P_\theta[X_1 \geq t]$ for $t > \mu$.

Hint: You may use Problem 2.2.16(b).

13. Let $X_1, \ldots, X_n$ be a sample from a $\mathcal{U}[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$ distribution. Show that any T such that $X_{(n)} - \frac{1}{2} \leq T \leq X_{(1)} + \frac{1}{2}$ is a maximum likelihood estimate of θ. (We write $\mathcal{U}[a, b]$ to make $p(a) = p(b) = (b-a)^{-1}$ rather than 0.)

14. If $n = 1$ in Example 2.1.5 show that no maximum likelihood estimate of $\theta = (\mu, \sigma^2)$ exists.

15. Suppose that $T(\mathbf{X})$ is sufficient for θ and that $\widehat{\theta}(\mathbf{X})$ is an MLE of θ. Show that $\widehat{\theta}$ depends on $\mathbf{X}$ through $T(\mathbf{X})$ only provided that $\widehat{\theta}$ is unique.

Hint: Use the factorization theorem (Theorem 1.5.1).

16. (a) Let $\mathbf{X} \sim P_\theta$, $\theta \in \Theta$ and let $\widehat{\theta}$ denote the MLE of θ. Suppose that h is a one-to-one function from Θ onto $h(\Theta)$. Define $\eta = h(\theta)$ and let $f(\mathbf{x}, \eta)$ denote the density or frequency function of $\mathbf{X}$ in terms of η (i.e., reparametrize the model using η). Show that the MLE of η is $h(\widehat{\theta})$ (i.e., MLEs are unaffected by reparametrization, they are *equivariant* under one-to-one transformations).

(b) Let $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset R^p$, $p \geq 1$, be a family of models for $\mathbf{X} \in \mathcal{X} \subset R^d$. Let $\mathbf{q}$ be a map from Θ onto Ω, $\Omega \subset R^k$, $1 \leq k \leq p$. Show that if $\widehat{\boldsymbol{\theta}}$ is a MLE of $\boldsymbol{\theta}$, then $\mathbf{q}(\widehat{\boldsymbol{\theta}})$ is an MLE of $\boldsymbol{\omega} = \mathbf{q}(\boldsymbol{\theta})$.

Hint: Let $\Theta(\boldsymbol{\omega}) = \{\boldsymbol{\theta} \in \Theta : \mathbf{q}(\boldsymbol{\theta}) = \boldsymbol{\omega}\}$, then $\{\Theta(\boldsymbol{\omega}) : \boldsymbol{\omega} \in \Omega\}$ is a partition of Θ, and $\widehat{\boldsymbol{\theta}}$ belongs to only one member of this partition, say $\Theta(\widehat{\boldsymbol{\omega}})$. Because $\mathbf{q}$ is onto Ω, for each $\boldsymbol{\omega} \in \Omega$ there is $\boldsymbol{\theta} \in \Theta$ such that $\boldsymbol{\omega} = \mathbf{q}(\boldsymbol{\theta})$. Thus, the MLE of $\boldsymbol{\omega}$ is by definition

$$\widehat{\boldsymbol{\omega}}_{MLE} = \arg \sup_{\boldsymbol{\omega} \in \Omega} \sup\{L_{\mathbf{X}}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta(\boldsymbol{\omega})\}.$$

Now show that $\widehat{\boldsymbol{\omega}}_{MLE} = \widehat{\boldsymbol{\omega}} = \mathbf{q}(\widehat{\boldsymbol{\theta}})$.

17. *Censored Geometric Waiting Times.* If time is measured in discrete periods, a model that is often used for the time X to failure of an item is

$$P_\theta[X = k] = \theta^{k-1}(1-\theta),\ k = 1, 2, \ldots$$

where $0 < \theta < 1$. Suppose that we only record the time of failure, if failure occurs on or before time r and otherwise just note that the item has lived at least $(r+1)$ periods. Thus, we observe $Y_1, \ldots, Y_n$ which are independent, identically distributed, and have common frequency function,

$$f(k, \theta) = \theta^{k-1}(1-\theta),\ k = 1, \ldots, r$$

$$f(r+1, \theta) = 1 - P_\theta[X \leq r] = 1 - \sum_{k=1}^{r} \theta^{k-1}(1-\theta) = \theta^r.$$

(We denote by "$r+1$" survival for at least $(r+1)$ periods.) Let $M =$ number of indices i such that $Y_i = r+1$. Show that the maximum likelihood estimate of θ based on $Y_1, \ldots, Y_n$

is

$$\widehat{\theta}(\mathbf{Y}) = \frac{\sum_{i=1}^{n} Y_i - n}{\sum_{i=1}^{n} Y_i - M}.$$

18. Derive maximum likelihood estimates in the following models.

(a) The observations are indicators of Bernoulli trials with probability of success θ. We want to estimate θ and $\mathrm{Var}_\theta X_1 = \theta(1-\theta)$.

(b) The observations are $X_1 =$ the number of failures before the first success, $X_2 =$ the number of failures between the first and second successes, and so on, in a sequence of binomial trials with probability of success θ. We want to estimate θ.

19. Let $X_1, \ldots, X_n$ be independently distributed with X_i having a $\mathcal{N}(\theta_i, 1)$ distribution, $1 \leq i \leq n$.

(a) Find maximum likelihood estimates of the θ_i under the assumption that these quantities vary freely.

(b) Solve the problem of part (a) for $n = 2$ when it is known that $\theta_1 \leq \theta_2$. A general solution of this and related problems may be found in the book by Barlow, Bartholomew, Bremner, and Brunk (1972).

20. In the "life testing" problem 1.6.16(i), find the MLE of θ.

21. (Kiefer–Wolfowitz) Suppose $(X_1, \ldots, X_n)$ is a sample from a population with density

$$f(x, \theta) = \frac{9}{10\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) + \frac{1}{10}\varphi(x-\mu)$$

where φ is the standard normal density and $\theta = (\mu, \sigma^2) \in \Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$. Show that maximum likelihood estimates do not exist, but that $\sup_\sigma p(\mathbf{x}, \widehat{\mu}, \sigma^2) = \sup_{\mu,\sigma} p(\mathbf{x}, \mu, \sigma^2)$ if, and only if, $\widehat{\mu}$ equals one of the numbers $x_1, \ldots, x_n$. Assume that $x_i \neq x_j$ for $i \neq j$ and that $n \geq 2$.

22. Suppose X has a hypergeometric, $\mathcal{H}(b, N, n)$, distribution. Show that the maximum likelihood estimate of b for N and n fixed is given by

$$\widehat{b}(X) = \left[\frac{X}{n}(N+1)\right]$$

if $\frac{X}{n}(N+1)$ is *not* an integer, and

$$\widehat{b}(X) = \frac{X}{n}(N+1) \text{ or } \frac{X}{n}(N+1) - 1$$

otherwise, where $[t]$ is the largest integer that is $\leq t$.

Hint: Consider the ratio $L(b+1, x)/L(b, x)$ as a function of b.

23. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent samples from $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ populations, respectively. Show that the MLE of $\theta = (\mu_1, \mu_2, \sigma^2)$ is $\widehat{\theta} = (\bar{X}, \bar{Y}, \widetilde{\sigma}^2)$ where

$$\widetilde{\sigma}^2 = \left[\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2\right]/(m+n).$$

24. *Polynomial Regression.* Suppose $Y_i = \mu(\mathbf{z}_i) + \epsilon_i$, where ϵ_i satisfy (2.2.4)–(2.2.6). Set $\mathbf{z}^{\mathbf{j}} = z_1^{j_1} \cdots z_p^{j_p}$ where $\mathbf{j} \in \mathcal{J}$ and $\mathcal{J}$ is a subset of $\{(j_1, \ldots, j_p) : 0 \leq j_k \leq J,\ 1 \leq k \leq p\}$, and assume that

$$\mu(\mathbf{z}) = \sum\{\alpha_j \mathbf{z}^{\mathbf{j}} : \mathbf{j} \in \mathcal{J}\}.$$

In an experiment to study tool life (in minutes) of steel-cutting tools as a function of cutting speed (in feet per minute) and feed rate (in thousands of an inch per revolution), the following data were obtained (from S. Weisberg, 1985).

TABLE 2.6.1. Tool life data

Feed	Speed	Life	Feed	Speed	Life
-1	-1	54.5	$-\sqrt{2}$	0	20.1
-1	-1	66.0	$\sqrt{2}$	0	2.9
1	-1	11.8	0	0	3.8
1	-1	14.0	0	0	2.2
-1	1	5.2	0	0	3.2
-1	1	3.0	0	0	4.0
1	1	0.8	0	0	2.8
1	1	0.5	0	0	3.2
0	$-\sqrt{2}$	86.5	0	0	4.0
0	$\sqrt{2}$	0.4	0	0	3.5

The researchers analyzed these data using

$$Y = \log \text{ tool life},\ z_1 = (\text{feed rate} - 13)/6,\ z_2 = (\text{cutting speed} - 900)/300.$$

Two models are contemplated

(a) $Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \epsilon$

(b) $Y = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_1^2 + \alpha_4 z_2^2 + \alpha_5 z_1 z_2 + \epsilon$.

Use a least squares computer package to compute estimates of the coefficients (β's and α's) in the two models. Use these estimated coefficients to compute the values of the contrast function (2.1.5) for (a) and (b). Both of these models are approximations to the true mechanism generating the data. Being larger, the second model provides a better

approximation. However, this has to be balanced against greater variability in the estimated coefficients. This will be discussed in Volume II.

25. Consider the model (2.2.1), (2.2.4)–(2.2.6) with $g(\boldsymbol{\beta}, \mathbf{z}) = \mathbf{z}^T\boldsymbol{\beta}$. Show that the following are equivalent.

(a) The parameterization $\boldsymbol{\beta} \to \mathbf{Z}_D\boldsymbol{\beta}$ is identifiable.

(b) $\mathbf{Z}_D$ is of rank d.

(c) $\mathbf{Z}_D^T\mathbf{Z}_D$ is of rank d.

26. Let (Z, Y) have joint probability P with joint density $f(z, y)$, let $v(z, y) \geq 0$ be a weight funciton such that $E(v(Z, Y)Z^2)$ and $E(v(Z, Y)Y^2)$ are finite. The best linear weighted mean squared prediction error predictor $\beta_1(P) + \beta_2(P)Z$ of Y is defined as the minimizer of

$$E\{v(Z, Y)[Y - (b_1 + b_2Z)]^2\}.$$

(a) Let (Z^*, Y^*) have density $v(z, y)f(z, y)/c$ where $c = \int\int v(z, y)f(z, y)dzdy$. Show that $\beta_2(P) = \text{Cov}(Z^*, Y^*)/\text{Var } Z^*$ and $\beta_1(P) = E(Y^*) - \beta_2(P)E(Z^*)$.

(b) Let $\widehat{P}$ be the empirical probability defined in Problem 2.1.8 and let $v(z, y) = 1/\text{Var}(Y \mid Z = z)$. Show that $\beta_1(\widehat{P})$ and $\beta_2(\widehat{P})$ coincide with $\widehat{\beta}_1$ and $\widehat{\beta}_2$ of Example 2.2.3. That is, weighted least squares estimates are plug-in estimates.

27. Derive the weighted least squares normal equations (2.2.19).

28. Let $\boldsymbol{Z}_D = \|z_{ij}\|_{n\times d}$ be a design matrix and let $\boldsymbol{W}_{n\times n}$ be a known symmetric invertible matrix. Consider the model $\boldsymbol{Y} = \boldsymbol{Z}_D\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ has covariance matrix $\sigma^2\boldsymbol{W}$, σ^2 unknown. Let $\boldsymbol{W}^{-\frac{1}{2}}$ be a square root matrix of $\boldsymbol{W}^{-1}$ (see (B.6.6)). Set $\widetilde{\boldsymbol{Y}} = \boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{Y}$, $\widetilde{\boldsymbol{Z}}_D = \boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{Z}_D$ and $\widetilde{\boldsymbol{\epsilon}} = \boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{\epsilon}$.

(a) Show that $\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{Z}}_D\boldsymbol{\beta}+\widetilde{\boldsymbol{\epsilon}}$ satisfy the linear regression model (2.2.1), (2.2.4)–(2.2.6) with $g(\boldsymbol{\beta}, \boldsymbol{z}) = \widetilde{\boldsymbol{Z}}_D\boldsymbol{\beta}$.

(b) Show that if $\boldsymbol{Z}_D$ has rank d, then the $\widehat{\boldsymbol{\beta}}$ that minimizes

$$(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Z}}_D\boldsymbol{\beta})^T(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Z}}_D\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{Z}_D\boldsymbol{\beta})^T\boldsymbol{W}^{-1}(\boldsymbol{Y} - \boldsymbol{Z}_D\boldsymbol{\beta})$$

is given by (2.2.20).

29. Let $e_i = (\epsilon_i + \epsilon_{i+1})/2$, $i = 1, \ldots, n$, where $\epsilon_1, \ldots, \epsilon_{n+1}$ are i.i.d. with mean zero and variance σ^2. The e_i are called *moving average errors.*

Consider the model $Y_i = \mu + e_i$, $i = 1, \ldots, n$.

(a) Show that $E(Y_{i+1} \mid Y_1, \ldots, Y_i) = \frac{1}{2}(\mu + Y_i)$. That is, in this model the optimal MSPE predictor of the future Y_{i+1} given the past $Y_1, \ldots, Y_i$ is $\frac{1}{2}(\mu + Y_i)$.

(b) Show that $\bar{Y}$ is a multivariate method of moments estimate of μ. (See Problem 2.1.17.)

(c) Find a matrix $\boldsymbol{A}$ such that $\boldsymbol{e}_{n\times 1} = \boldsymbol{A}_{n\times(n+1)}\boldsymbol{\epsilon}_{(n+1)\times 1}$.

(d) Find the covariance matrix $\boldsymbol{W}$ of $\boldsymbol{e}$.

(e) Find the weighted least squares estimate of μ.

(f) The following data give the elapsed times $Y_1, \ldots, Y_n$ spent above a fixed high level for a series of $n = 66$ consecutive wave records at a point on the seashore. Use a weighted least squares computer routine to compute the weighted least squares estimate $\widehat{\mu}$ of μ. Is $\widehat{\mu}$ different from $\bar{Y}$?

TABLE 2.5.1. Elapsed times spent above a certain high level for a series of 66 wave records taken at San Francisco Bay. The data (courtesy S. J. Chou) should be read row by row.

2.968	2.097	1.611	3.038	7.921	5.476	9.858	1.397	0.155	1.301
9.054	1.958	4.058	3.918	2.019	3.689	3.081	4.229	4.669	2.274
1.971	10.379	3.391	2.093	6.053	4.196	2.788	4.511	7.300	5.856
0.860	2.093	0.703	1.182	4.114	2.075	2.834	3.968	6.480	2.360
5.249	5.100	4.131	0.020	1.071	4.455	3.676	2.666	5.457	1.046
1.908	3.064	5.392	8.393	0.916	9.665	5.564	3.599	2.723	2.870
1.582	5.453	4.091	3.716	6.156	2.039				

30. In the multinomial Example 2.2.8, suppose some of the n_j are zero. Show that the MLE of θ_j is $\widehat{\boldsymbol{\theta}}$ with $\widehat{\theta}_j = n_j/n$, $j = 1, \ldots, k$.

Hint: Suppose without loss of generality that $n_1 = n_2 = \cdots = n_q = 0, n_{q+1} > 0, \ldots, n_k > 0$. Then

$$p(\mathbf{x}, \boldsymbol{\theta}) = \prod_{j=q+1}^{k} \theta_j^{n_j},$$

which vanishes if $\theta_j = 0$ for any $j = q+1, \ldots, k$.

31. Suppose $Y_1, \ldots, Y_n$ are independent with Y_i uniformly distributed on $[\mu_i - \sigma, \mu_i + \sigma]$, $\sigma > 0$, where $\mu_i = \sum_{j=1}^{p} z_{ij}\beta_j$ for given covariate values $\{z_{ij}\}$. Show that the MLE of $(\beta_1, \ldots, \beta_p, \sigma)^T$ is obtained by finding $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ that minimizes the *maximum absolute value* contrast function $\max_i |y_i - \mu_i|$ and then setting $\widehat{\sigma} = \max_i |y_i - \widehat{\mu}_i|$, where $\widehat{\mu}_i = \sum_{j=1}^{p} z_{ij}\widehat{\beta}_j$.

32. Suppose $Y_1, \ldots, Y_n$ are independent with Y_i having the Laplace density

$$\frac{1}{2\sigma} \exp\{-|y_i - \mu_i|/\sigma\},\ \sigma > 0$$

where $\mu_i = \sum_{j=1}^{p} z_{ij}\beta_j$ for given covariate values $\{z_{ij}\}$.

(a) Show that the MLE of $(\beta_1, \ldots, \beta_p, \sigma)$ is obtained by finding $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ that minimizes the *least absolute deviation* contrast function $\sum_{i=1}^{n} |y_i - \mu_i|$ and then setting $\widehat{\sigma} =$

$n^{-1}\sum_{i=1}^n |y_i - \widehat{\mu}_i|$, where $\widehat{\mu}_i = \sum_{j=1}^p z_{ij}\widehat{\beta}_j$. These $\widehat{\beta}_1, \ldots, \widehat{\beta}_r$ and $\widehat{\mu}_1, \ldots, \widehat{\mu}_n$ are called *least absolute deviation estimates (LADEs)*.

(b) Suppose $\mu_i = \mu$ for each i. Show that the sample median $\widehat{y}$, as defined in Section 1.3, is the minimizer of $\sum_{i=1}^n |y_i - \mu|$.

Hint: Use Problem 1.4.7 with Y having the empirical distribution $\widehat{F}$.

33. The *Hodges–Lehmann (location)* estimate $\widehat{x}_{HL}$ is defined to be the median of the $\frac{1}{2}n(n+1)$ pairwise averages $\frac{1}{2}(x_i + x_j)$, $i \leq j$. An asymptotically equivalent procedure $\widetilde{x}_{HL}$ is to take the median of the distribution placing mass $\frac{2}{n^2}$ at each point $\frac{x_i+x_j}{2}$, $i < j$ and mass $\frac{1}{n^2}$ at each x_i.

(a) Show that the Hodges–Lehmann estimate is the minimizer of the contrast function

$$\rho(x, \theta) = \sum_{i \leq j} |x_i + x_j - 2\theta|.$$

Hint: See Problem 2.2.32(b).

(b) Define θ_{HL} to be the minimizer of

$$\int |x - 2\theta| d(F * F)(x)$$

where $F * F$ denotes convolution. Show that $\widetilde{x}_{HL}$ is a plug-in estimate of θ_{HL}.

34. Let X_i be i.i.d. as $(Z, Y)^T$ where $Y = Z + \sqrt{\lambda}W$, $\lambda > 0$, Z and W are independent $\mathcal{N}(0, 1)$. Find the MLE of λ and give its mean and variance.

Hint: See Example 1.6.3.

35. Let $g(x) = 1/\pi(1 + x^2)$, $x \in R$, be the Cauchy density, let X_1 and X_2 be i.i.d. with density $g(x - \theta)$, $\theta \in R$. Let x_1 and x_2 be the observations and set $\Delta = \frac{1}{2}(x_1 - x_2)$. Let $\widehat{\theta} = \arg\max L_{\mathbf{x}}(\theta)$ be "the" MLE.

(a) Show that if $|\Delta| \leq 1$, then the MLE exists and is unique. Give the MLE when $|\Delta| \leq 1$.

(b) Show that if $|\Delta| > 1$, then the MLE is not unique. Find the values of θ that maximize the likelihood $L_{\mathbf{x}}(\theta)$ when $|\Delta| > 1$.

Hint: Factor out $(\bar{x} - \theta)$ in the likelihood equation.

36. Problem 35 can be generalized as follows (Dharmadhikari and Joag–Dev, 1985). Let g be a probability density on R satisfying the following three conditions:

1. g is continuous, symmetric about 0, and positive everywhere.
2. g is twice continuously differentiable everywhere except perhaps at 0.
3. If we write $h = \log g$, then $h''(y) > 0$ for some nonzero y.

Let (X_1, X_2) be a random sample from the distribution with density $f(x, \theta) = g(x - \theta)$, where $x \in R$ and $\theta \in R$. Let x_1 and x_2 be the observed values of X_1 and X_2 and write $\bar{x} = (x_1 + x_2)/2$ and $\Delta = (x_1 - x_2)/2$. The likelihood function is given by

$$\begin{aligned} L_{\mathbf{x}}(\theta) &= g(x_1 - \theta)g(x_2 - \theta) \\ &= g(\bar{x} + \Delta - \theta)g(\bar{x} - \Delta - \theta). \end{aligned}$$

Let $\widehat{\theta} = \arg\max L_{\mathbf{x}}(\theta)$ be "the" MLE.

Show that

(a) The likelihood is symmetric about $\bar{x}$.

(b) Either $\widehat{\theta} = \bar{x}$ or $\widehat{\theta}$ is not unique.

(c) There is an interval (a, b), $a < b$, such that for every $y \in (a, b)$ there exists a $\delta > 0$ such that $h(y + \delta) - h(y) > h(y) - h(y - \delta)$.

(d) Use (c) to show that if $\Delta \in (a, b)$, then $\widehat{\theta}$ is not unique.

37. Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\theta, \sigma^2)$ and let $p(\mathbf{x}, \theta)$ denote their joint density. Show that the entropy of $p(\mathbf{x}, \theta)$ is $\frac{1}{2}n$ and that the Kullback–Liebler divergence between $p(\mathbf{x}, \theta)$ and $p(\mathbf{x}, \theta_0)$ is $\frac{1}{2}n(\theta - \theta_0)^2/\sigma^2$.

38. Let $\mathbf{X} \sim P_\theta$, $\theta \in \Theta$. Suppose h is a 1-1 function from Θ onto $\Omega = h(\Theta)$. Define $\eta = h(\theta)$ and let $p^*(\mathbf{x}, \eta) = p(\mathbf{x}, h^{-1}(\eta))$ denote the density or frequency function of $\mathbf{X}$ for the η parametrization. Let $K(\theta_0, \theta_1)$ $(K^*(\eta_0, \eta_1))$ denote the Kullback–Leibler divergence between $p(\mathbf{x}, \theta_0)$ and $p(x, \theta_1)$ $(p^*(\mathbf{x}, \eta_0)$ and $p^*(\mathbf{x}, \eta_1))$. Show that

$$K^*(\eta_0, \eta_1) = K(h^{-1}(\eta_0), h^{-1}(\eta_1)).$$

39. Let X_i denote the number of hits at a certain Web site on day i, $i = 1, \ldots, n$. Assume that $S = \sum_{i=1}^n X_i$ has a Poisson, $\mathcal{P}(n\lambda)$, distribution. On day $n + 1$ the Web Master decides to keep track of two types of hits (money making and not money making). Let V_j and W_j denote the number of hits of type 1 and 2 on day j, $j = n+1, \ldots, n+m$. Assume that $S_1 = \sum_{j=n+1}^{n+m} V_j$ and $S_2 = \sum_{j=n+1}^{n+m} W_j$ have $\mathcal{P}(m\lambda_1)$ and $\mathcal{P}(m\lambda_2)$ distributions, where $\lambda_1 + \lambda_2 = \lambda$. Also assume that S, S_1, and S_2 are independent. Find the MLEs of λ_1 and λ_2 based on S, S_1, and S_2.

40. Let $X_1, \ldots, X_n$ be a sample from the generalized Laplace distribution with density

$$\begin{aligned} f(x, \theta_1, \theta_2) &= \frac{1}{\theta_1 + \theta_2} \exp\{-x/\theta_1\},\ x > 0, \\ &= \frac{1}{\theta_1 + \theta_2} \exp\{x/\theta_2\},\ x < 0 \end{aligned}$$

where $\theta_j > 0$, $j = 1, 2$.

(a) Show that $T_1 = \sum X_i 1[X_i > 0]$ and $T_2 = \sum -X_i 1[X_i < 0]$ are sufficient statistics.

(b) Find the maximum likelihood estimates of θ_1 and θ_2 in terms of T_1 and T_2. Carefully check the "$T_1 = 0$ or $T_2 = 0$" case.

41. The mean relative growth of an organism of size y at time t is sometimes modeled by the equation (Richards, 1959; Seber and Wild, 1989)

$$\frac{1}{y}\frac{dy}{dt} = \beta\left[1 - \left(\frac{y}{\alpha}\right)^{\frac{1}{\delta}}\right],\ y > 0;\ \alpha > 0,\ \beta > 0,\ \delta > 0.$$

(a) Show that a solution to this equation is of the form $y = g(t;\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\alpha, \beta, \mu, \delta)$, $\mu \in R$, and

$$g(t,\boldsymbol{\theta}) = \frac{\alpha}{\{1+\exp[-\beta(t-\mu)/\delta]\}^{\delta}}.$$

(b) Suppose we have observations $(t_1, y_1), \ldots, (t_n, y_n)$, $n \geq 4$, on a population of a large number of organisms. Variation in the population is modeled on the log scale by using the model

$$\log Y_i = \log\alpha - \delta\log\{1+\exp[-\beta(t_i-\mu)/\delta]\} + \epsilon_i$$

where $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated with mean 0 and variance σ^2. Give the least squares estimating equations (2.1.7) for estimating α, β, δ, and μ.

(c) Let Y_i denote the response of the ith organism in a sample and let z_{ij} denote the level of the jth covariate (stimulus) for the ith organism, $i = 1, \ldots, n$; $j = 1, \ldots, p$. An example of a *neural net model* is

$$Y_i = \sum_{j=1}^{p} h(z_{ij};\boldsymbol{\lambda}_j) + \epsilon_i,\ i = 1, \ldots, n$$

where $\boldsymbol{\lambda} = (\alpha, \beta, \mu)$, $h(z;\boldsymbol{\lambda}) = g(z;\alpha,\beta,\mu,1)$, and $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated with mean zero and variance σ^2. For the case $p = 1$, give the least square estimating equations (2.1.7) for α, β, and μ.

42. Suppose $X_1, \ldots, X_n$ satisfy the autoregressive model of Example 1.1.5.

(a) If μ is known, show that the MLE of β is

$$\widehat{\beta} = \frac{-\sum_{i=2}^{n}(x_{i-1}-\mu)(x_i-\mu)}{\sum_{i=1}^{n-1}(x_i-\mu)^2}.$$

(b) If β is known, find the covariance matrix $\mathbf{W}$ of the vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ of autoregression errors. (One way to do this is to find a matrix $\mathbf{A}$ such that $\mathbf{e}_{n\times 1} = \mathbf{A}_{n\times n}\boldsymbol{\epsilon}_{n\times 1}$.) Then find the weighted least square estimate of μ. Is this also the MLE of μ?

Problems for Section 2.3

1. Suppose $Y_1, \ldots, Y_n$ are independent

$$P[Y_i = 1] = p(x_i,\alpha,\beta) = 1 - P[Y_i = 0],\ 1 \leq i \leq n,\ n \geq 2,$$
$$\log\frac{p}{1-p}(x,\alpha,\beta) = \alpha + \beta x,\ x_1 < \cdots < x_n.$$

Show that the MLE of α, β exists iff $(Y_1, \ldots, Y_n)$ is not a sequence of 1's followed by all 0's or the reverse.

Hint:

$$c_1\sum_{i=1}^n y_i + c_2\sum_{i=1}^n x_i y_i = \sum_{i=1}^n (c_1 + c_2 x_i) y_i \le \sum_{i=1}^n (c_1 + c_2 x_i) 1(c_2 x_i + c_1 \ge 0).$$

If $c_2 > 0$, the bound is sharp and is attained only if $y_i = 0$ for $x_i \le -\frac{c_1}{c_2}$, $y_i = 1$ for $x_i \ge -\frac{c_1}{c_2}$.

2. Let $X_1, \ldots, X_n$ be i.i.d. gamma, $\Gamma(\lambda, p)$.

(a) Show that the density of $\mathbf{X} = (X_1, \ldots, X_n)^T$ can be written as the rank 2 canonical exponential family generated by $\mathbf{T} = (\Sigma \log X_i, \Sigma X_i)$ and $h(x) = x^{-1}$ with $\eta_1 = p$, $\eta_2 = -\lambda$ and

$$A(\eta_1, \eta_2) = n[\log \Gamma(\eta_1) - \eta_1 \log(-\eta_2)],$$

where Γ denotes the gamma function.

(b) Show that the likelihood equations are equivalent to $(2.3.4)$ and $(2.3.5)$.

3. Consider the Hardy–Weinberg model with the six genotypes given in Problem 2.1.15. Let $\Theta = \{(\theta_1, \theta_2) : \theta_1 > 0, \theta_2 > 0, \theta_1 + \theta_2 < 1\}$ and let $\theta_3 = 1 - (\theta_1 + \theta_2)$. In a sample of n independent plants, write $x_i = j$ if the ith plant has genotype j, $1 \le j \le 6$. Under what conditions on $(x_1, \ldots, x_n)$ does the MLE exist? What is the MLE? Is it unique?

4. Give details of the proof of Corollary 2.3.1.

5. Prove Lemma 2.3.1.

Hint: Let $c = l(\mathbf{0})$. There exists a compact set $K \subset \Theta$ such that $l(\boldsymbol{\theta}) < c$ for all $\boldsymbol{\theta}$ not in K. This set K will have a point where the max is attained.

6. In the heteroscedastic regression Example 1.6.10 with $n \ge 3$, $0 < z_1 < \cdots < z_n$, show that the MLE exists and is unique.

7. Let $Y_1, \ldots, Y_n$ denote the duration times of n independent visits to a Web site. Suppose $\mathbf{Y}$ has an exponential, $\mathcal{E}(\lambda_i)$, distribution where

$$\mu_i = E(Y_i) = \lambda_i^{-1} = \exp\{\alpha + \beta z_i\},\ z_1 < \cdots < z_n$$

and z_i is the income of the person whose duration time is Y_i, $0 < z_1 < \cdots < z_n$, $n \ge 2$. Show that the MLE of $(\alpha, \beta)^T$ exists and is unique. See also Problem 1.6.40.

8. Let $X_1, \ldots, X_n \in R^p$ be i.i.d. with density,

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = c(\alpha) \exp\{-|\mathbf{x} - \boldsymbol{\theta}|^\alpha\},\ \boldsymbol{\theta} \in R^p,\ \alpha \ge 1$$

where $c^{-1}(\alpha) = \int_{R_p} \exp\{-|\mathbf{x}|^\alpha\} d\mathbf{x}$ and $|\cdot|$ is the Euclidean norm.

(a) Show that if $\alpha > 1$, the MLE $\widehat{\boldsymbol{\theta}}$ exists and is unique.

(b) Show that if $\alpha = 1$ and $p = 1$, the MLE $\widehat{\boldsymbol{\theta}}$ exists but is not unique if n is even.

9. Show that the boundary ∂C of a convex C set in R^k has volume 0.

Hint: If ∂C has positive volume, then it must contain a sphere and the center of the sphere is an interior point by (B.9.1).

10. Use Corollary 2.3.1 to show that in the multinomial Example 2.3.3, MLEs of η_j exist iff all $T_j > 0, 1 \leq j \leq k-1$.

Hint: The k points $(0,\ldots,0), (0,n,0,\ldots,0),\ldots,(0,0,\ldots,n)$ are the vertices of the convex set $\{(t_1,\ldots,t_{k-1}) : t_j \geq 0, 1 \leq j \leq k-1, \sum_{j=1}^{k-1} t_j \leq n\}$.

11. Prove Theorem 2.3.3.

Hint: If it didn't there would exist $\boldsymbol{\eta}_j = c(\theta_j)$ such that $\boldsymbol{\eta}_j^T t_0 - A(\boldsymbol{\eta}_j) \to \max\{\boldsymbol{\eta}^T \mathbf{t}_0 - A(\boldsymbol{\eta}) : \boldsymbol{\eta} \in c(\Theta)\} > -\infty$. Then $\{\boldsymbol{\eta}_j\}$ has a subsequence that converges to a point $\boldsymbol{\eta}^0 \in \mathcal{E}$. But $c(\Theta)$ is closed so that $\boldsymbol{\eta}^0 = c(\boldsymbol{\theta}^0)$ and $\boldsymbol{\theta}^0$ must satisfy the likelihood equations.

12. Let $X_1,\ldots,X_n$ be i.i.d. $\frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right)$, $\sigma > 0$, $\mu \in R$, and assume for $w \equiv -\log f_0$ that $w'' > 0$ so that w is strictly convex, $w(\pm\infty) = \infty$.

(a) Show that, if $n \geq 2$, the likelihood equations

$$\sum_{i=1}^{n} w'\left(\frac{X_i-\mu}{\sigma}\right) = 0$$

$$\sum_{i=1}^{n} \left\{\frac{(X_i-\mu)}{\sigma} w'\left(\frac{X_i-\mu}{\sigma}\right) - 1\right\} = 0$$

have a unique solution $(\widehat{\mu},\widehat{\sigma})$.

(b) Give an algorithm such that starting at $\widehat{\mu}^0 = 0$, $\widehat{\sigma}^0 = 1$, $\widehat{\mu}^{(i)} \to \widehat{\mu}$, $\widehat{\sigma}^{(i)} \to \widehat{\sigma}$.

(c) Show that for the logistic distribution $F_0(x) = [1+\exp\{-x\}]^{-1}$, w is strictly convex and give the likelihood equations for μ and σ. (See Example 2.4.3.)

Hint: **(a)** The function $D(a,b) = \sum_{i=1}^{n} w(aX_i - b) - n\log a$ is strictly convex in (a,b) and $\lim_{(a,b)\to(a_0,b_0)} D(a,b) = \infty$ if either $a_0 = 0$ or ∞ or $b_0 = \pm\infty$.

(b) Reparametrize by $a = \frac{1}{\sigma}$, $b = \frac{\mu}{\sigma}$ and consider varying a, b successively.

Note: You may use without proof (see Appendix B.9).

(i) If a strictly convex function has a minimum, it is unique.

(ii) If $\frac{\partial^2 D}{\partial a^2} > 0$, $\frac{\partial^2 D}{\partial b^2} > 0$ and $\frac{\partial^2 D}{\partial a^2}\frac{\partial^2 D}{\partial b^2} > \left(\frac{\partial^2 D}{\partial a \partial b}\right)^2$, then D is strictly convex.

13. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a sample from a $\mathcal{N}(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$ population.

(a) Show that the MLEs of σ_1^2,σ_2^2, and ρ when μ_1 and μ_2 are assumed to be known are $\widetilde{\sigma}_1^2 = (1/n)\sum_{i=1}^{n}(X_i-\mu_1)^2$, $\widetilde{\sigma}_2^2 = (1/n)\sum_{i=1}^{n}(Y_i-\mu_2)^2$, and

$$\widetilde{\rho} = \left[\sum_{i=1}^{n}(X_i-\mu_1)(Y_i-\mu_2)/n\widetilde{\sigma}_1\widetilde{\sigma}_2\right]$$

respectively, provided that $n \geq 3$.

(b) If $n \geq 5$ and μ_1 and μ_2 are unknown, show that the estimates of μ_1, μ_2, σ_1^2, σ_2^2, ρ coincide with the method of moments estimates of Problem 2.1.8.

Hint: (b) Because (X_1, Y_1) has a density you may assume that $\widetilde{\sigma}_1^2 > 0, \widetilde{\sigma}_2^2 > 0, |\widetilde{\rho}| < 1$. Apply Corollary 2.3.2.

Problems for Section 2.4

1. EM for bivariate data.

(a) In the bivariate normal Example 2.4.6, complete the E-step by finding $E(Z_i \mid Y_i)$, $E(Z_i^2 \mid Y_i)$ and $E(Z_i Y_i \mid Y_i)$.

(b) In Example 2.4.6, verify the M-step by showing that

$$E_\theta \mathbf{T} = (\mu_1, \mu_2, \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, \rho\sigma_1\sigma_2 + \mu_1\mu_2).$$

2. Show that if T is minimal and $\mathcal{E}$ is open and the MLE doesn't exist, then the coordinate ascent algorithm doesn't converge to a member of $\mathcal{E}$.

3. Describe in detail what the coordinate ascent algorithm does in estimation of the regression coefficients in the Gaussian linear model

$$\mathbf{Y} = \mathbf{Z}_D\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \text{rank}(\mathbf{Z}_D) = k,\ \epsilon_1, \ldots, \epsilon_n \text{ i.i.d. } \mathcal{N}(0, \sigma^2).$$

(Check that you are describing the Gauss–Seidel iterative method for solving a system of linear equations. See, for example, Golub and Van Loan, 1985, Chapter 10.)

4. Let (I_i, Y_i), $1 \leq i \leq n$, be independent and identically distributed according to P_θ, $\theta = (\lambda, \mu) \in (0,1) \times R$ where

$$P_\theta[I_1 = 1] = \lambda = 1 - P_\theta[I_1 = 0],$$

and given $I_1 = j$, $Y_1 \sim \mathcal{N}(\mu, \sigma_j^2)$, $j = 0, 1$ and $\sigma_0^2 \neq \sigma_1^2$ known.

(a) Show that $\mathbf{X} \equiv \{(I_i, Y_i) : 1 \leq i \leq n\}$ is distributed according to an exponential family with $\mathbf{T} = \left(\frac{1}{\sigma_1^2}\sum_i Y_i I_i + \frac{1}{\sigma_0^2}\sum_i Y_i(1 - I_i),\ \sum_i I_i\right)$, $\eta_1 = \mu$, $\eta_2 = \log\left(\frac{\lambda}{1-\lambda}\right) + \frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)$.

(b) Deduce that $\mathbf{T}$ is minimal sufficient.

(c) Give explicitly the maximum likelihood estimates of μ and λ, when they exist.

5. Suppose the I_i in Problem 4 are not observed.

(a) Justify the following crude estimates of μ and λ,

$$\begin{aligned} \widetilde{\mu} &= \bar{Y} \\ \widetilde{\lambda} &= \left(\tfrac{1}{n}\textstyle\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sigma_0^2\right) \big/ (\sigma_1^2 - \sigma_0^2). \end{aligned}$$

Do you see any problems with $\widetilde{\lambda}$?

(b) Give as explicitly as possible the E- and M-steps of the EM algorithm for this problem.

Hint: Use Bayes rule.

6. Consider a genetic trait that is directly unobservable but will cause a disease among a certain proportion of the individuals that have it. For families in which one member has the disease, it is desired to estimate the proportion θ that has the genetic trait. Suppose that in a family of n members in which one has the disease (and, thus, also the trait), X is the number of members who have the trait. Because it is known that $X \geq 1$, the model often used for X is that it has the conditional distribution of a $\mathcal{B}(n, \theta)$ variable, $\theta \in [0, 1]$, given $X \geq 1$.

(a) Show that $P(X = x \mid X \geq 1) = \dfrac{\binom{n}{x}\theta^x(1-\theta)^{n-x}}{1-(1-\theta)^n}$, $x = 1, \ldots, n$, and that the MLE exists and is unique.

(b) Use (2.4.3) to show that the Newton–Raphson algorithm gives

$$\widehat{\theta}_1 = \widetilde{\theta} - \frac{\widetilde{\theta}(1-\widetilde{\theta})[1-(1-\widetilde{\theta})^n]\{x - n\widetilde{\theta} - x(1-\widetilde{\theta})^n\}}{n\widetilde{\theta}^2(1-\widetilde{\theta})^n[n-1+(1-\widetilde{\theta})^n] - [1-(1-\widetilde{\theta})^n]^2[(1-2\widetilde{\theta})x + n\widetilde{\theta}^2]},$$

where $\widetilde{\theta} = \widehat{\theta}_{\text{old}}$ and $\widehat{\theta}_1 = \widehat{\theta}_{\text{new}}$, as the first approximation to the maximum likelihood estimate of θ.

(c) If $n = 5$, $x = 2$, find $\widehat{\theta}_1$ of (b) above using $\widetilde{\theta} = x/n$ as a preliminary estimate.

7. Consider the following algorithm under the conditions of Theorem 2.4.2. Define $\widehat{\boldsymbol{\eta}}^0$ as before. Let

$$\widehat{\boldsymbol{\eta}}(\lambda) \equiv \widehat{\boldsymbol{\eta}}_{\text{old}} + \lambda \ddot{A}^{-1}(\widehat{\boldsymbol{\eta}}_{\text{old}})(\dot{A}(\widehat{\boldsymbol{\eta}}_{\text{old}}) - \mathbf{t}_0)$$

and

$$\widehat{\boldsymbol{\eta}}_{\text{new}} = \widehat{\boldsymbol{\eta}}(\lambda^*)$$

where λ^* maximizes

$$\mathbf{t}_0^T \widehat{\boldsymbol{\eta}}(\lambda) - A(\widehat{\boldsymbol{\eta}}(\lambda)).$$

Show that the sequence defined by this algorithm converges to the MLE if it exists.

Hint: Apply the argument of the proof of Theorem 2.4.2 noting that the sequence of iterates $\{\widehat{\boldsymbol{\eta}}_m\}$ is bounded and, hence, the sequence $(\widehat{\boldsymbol{\eta}}_m, \widehat{\boldsymbol{\eta}}_{m+1})$ has a convergent subsequence.

8. Let X_1, X_2, X_3 be independent observations from the Cauchy distribution about θ, $f(x, \theta) = \pi^{-1}(1 + (x-\theta)^2)^{-1}$. Suppose $X_1 = 0$, $X_2 = 1$, $X_3 = a$. Show that for a

sufficiently large the likelihood function has local maxima between 0 and 1 and between p and a.

(a) Deduce that depending on where bisection is started the sequence of iterates may converge to one or the other of the local maxima.

(b) Make a similar study of the Newton–Raphson method in this case.

9. Let $X_1, \ldots, X_n$ be i.i.d. where $X = (U, V, W)$, $P[U = a, V = b, W = c] \equiv p_{abc}$, $1 \le a \le A$, $1 \le b \le B$, $1 \le c \le C$ and $\sum_{a,b,c} p_{abc} = 1$.

(a) Suppose for all a, b, c,

(1) $\log p_{abc} = \mu_{ac} + \nu_{bc}$ where $-\infty < \mu, \nu < \infty$.

Show that this holds iff

$$P[U = a, V = b \mid W = c] = P[U = a \mid W = c]P[V = b \mid W = c],$$

i.e. iff U and V are independent given W.

(b) Show that the family of distributions obtained by letting μ, ν vary freely is an exponential family of rank $(C-1) + C(A+B-2) = C(A+B-1) - 1$ generated by $N_{++c}, N_{a+c}, N_{+bc}$ where $N_{abc} = \#\{i : X_i = (a, b, c)\}$ and "+" indicates summation over the index.

(c) Show that the MLEs exist iff $0 < N_{a+c}$, $N_{+bc} < N_{++c}$ for all a, b, c and then are given by

$$\widehat{p}_{abc} = \frac{N_{++c}}{n} \frac{N_{a+c}}{N_{++c}} \frac{N_{+bc}}{N_{++c}}.$$

Hint:

(b) Consider $N_{a+c} - N_{++c}/A$, $N_{+bc} - N_{++c}/B$, N_{++c}.

(c) The model implies $\widehat{p}_{abc} = \widehat{p}_{+bc}\widehat{p}_{a+c}/\widehat{p}_{++c}$ and use the likelihood equations.

10. Suppose X is as in Problem 9, but now

(2) $\log p_{abc} = \mu_{ac} + \nu_{bc} + \gamma_{ab}$ where μ, ν, γ vary freely.

(a) Show that this is an exponential family of rank

$$A + B + C - 3 + (A-1)(C-1) + (B-1)(C-1) + (A-1)(B-1)$$
$$= AB + AC + BC - (A + B + C).$$

(b) Consider the following "proportional fitting" algorithm for finding the maximum likelihood estimate in this model.

Initialize: $\widehat{p}_{abc}^{(0)} = \frac{N_{a++}}{n}\frac{N_{+b+}}{n}\frac{N_{++c}}{n}$

$$\begin{aligned}\widehat{p}_{abc}^{(1)} &= \frac{N_{ab+}}{n}\frac{\widehat{p}_{abc}^{(0)}}{\widehat{p}_{ab+}^{(0)}}\\ \widehat{p}_{abc}^{(2)} &= \frac{N_{a+c}}{n}\frac{\widehat{p}_{abc}^{(1)}}{\widehat{p}_{a+c}^{(1)}}\\ \widehat{p}_{abc}^{(3)} &= \frac{N_{+bc}}{n}\frac{\widehat{p}_{abc}^{(2)}}{\widehat{p}_{+bc}^{(2)}}.\end{aligned}$$

Reinitialize with $\widehat{p}_{abc}^{(3)}$. Show that the algorithm converges to the MLE if it exists and diverges otherwise.

Hint: Note that because $\{p_{abc}^{(0)}\}$ belongs to the model so do all subsequent iterates and that $\widehat{p}_{abc}^{(1)}$ is the MLE for the exponential family

$$p_{abc} = \frac{e^{\mu_{ab}}p_{abc}^{(0)}}{\sum\limits_{a',b',c'} e^{\mu_{a'b'}}p_{a'b'c'}^{(0)}}$$

obtained by fixing the "b, c" and "a, c" parameters.

11. **(a)** Show that **S** in Example 2.4.5 has the specified mixture of Gaussian distribution.

(b) Give explicitly the E- and M-steps of the EM algorithm in this case.

12. Justify formula (2.4.8).
Hint: $P_{\theta_0}[X = x \mid S(X) = s] = \frac{p(x,\theta_0)}{q(s,\theta_0)}1(S(x) = s)$.

13. Let $f_\theta(x) = f_0(x - \theta)$ where

$$f_0(x) = \frac{1}{3}\varphi(x) + \frac{2}{3}\varphi(x - a)$$

and φ is the $\mathcal{N}(0, 1)$ density. Show for $n = 1$ that bisection may lead to a local maximum of the likelihood, if a is sufficiently large.

14. Establish the last claim in part (2) of the proof of Theorem 2.4.2.
Hint: Use the canonical nature of the family and openness of $\mathcal{E}$.

15. Verify the formula given below (2.4.26) in Example 2.4.4 for the actual MLE in that example.
Hint: Show that $\{(\theta_m, \theta_{m+1})\}$ has a subsequence converging to (θ^*, θ^*) and necessarily $\theta^* = \widehat{\theta}_0$.

16. Establish part (b) of Theorem 2.4.3.
Hint: Show that $\{(\theta_m, \theta_{m+1})\}$ has a subsequence converging to (θ^*, θ^*) and, thus, necessarily θ^* is the global maximizer.

17. *Limitations of the missing value model of Example 2.4.6.* The assumption underlying Example 2.4.6 is that the conditional probability that a component X_j of the data vector X is missing given the rest of the data vector is not a function of X_j. That is, given $X - \{X_j\}$, the process determining whether X_j is missing is independent of X_j. This condition is called *missing at random.* For example, in Example 2.4.6, the probability that Y_i is missing may depend on Z_i, but not on Y_i. That is, given Z_i, the "missingness" of Y_i is independent of Y_i. If Y_i represents the seriousness of a disease, this assumption may not be satisfied. For instance, suppose all subjects with $Y_i \geq 2$ drop out of the study. Then using the E-step to impute values for the missing Y's would greatly underpredict the actual Y's because all the Y's in the imputation would have $Y \leq 2$. In Example 2.4.6, suppose Y_i is missing iff $Y_i \geq 2$. If $\mu_2 = 1.5$, $\sigma_1 = \sigma_2 = 1$ and $\rho = 0.5$, find the probability that $E(Y_i \mid Z_i)$ underpredicts Y_i.

18. *EM and Regression.* For $X = \{(Z_i, Y_i) : i = 1, \ldots, n\}$, consider the model

$$Y_i = \beta_1 + \beta_2 Z_i + \epsilon_i$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, $Z_1, \ldots, Z_n$ are i.i.d. $\mathcal{N}(\mu_1, \sigma_1^2)$ and independent of $\epsilon_1, \ldots, \epsilon_n$. Suppose that for $1 \leq i \leq m$ we observe both Z_i and Y_i and for $m+1 \leq i \leq n$, we observe only Y_i. Complete the E- and M-steps of the EM algorithm for estimating $(\mu_1, \beta_1, \sigma_1^2, \sigma^2, \beta_2)$.

2.6 NOTES

Notes for Section 2.1

(1) "Natural" now was not so natural in the eighteenth century when the least squares principle was introduced by Legendre and Gauss. For a fascinating account of the beginnings of estimation in the context of astronomy see Stigler (1986).

(2) The frequency plug-in estimates are sometimes called *Fisher consistent.* R. A. Fisher (1922) argued that only estimates possessing the substitution property should be considered and the best of these selected. These considerations lead essentially to maximum likelihood estimates.

Notes for Section 2.2

(1) An excellent historical account of the development of least squares methods may be found in Eisenhart (1964).

(2) For further properties of Kullback–Leibler divergence, see Cover and Thomas (1991).

Note for Section 2.3

(1) Recall that in an exponential family, for any A, $P[\mathbf{T}(X) \in A] = 0$ for all or for no $P \in \mathcal{P}$.

Note for Section 2.5

(1) In the econometrics literature (e.g. Appendix A.2; Campbell, Lo, and MacKinlay, 1997), a multivariate version of minimum contrasts estimates are often called generalized method of moment estimates.

2.7 REFERENCES

BARLOW, R. E., D. J. BARTHOLOMEW, J. M. BREMNER, AND H. D. BRUNK, *Statistical Inference Under Order Restrictions* New York: Wiley, 1972.

BAUM, L. E., T. PETRIE, G. SOULES, AND N. WEISS, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Ann. Math. Statist., 41*, 164–171 (1970).

BISHOP, Y. M. M., S. E. FEINBERG, AND P. W. HOLLAND, *Discrete Multivariate Analysis: Theory and Practice* Cambridge, MA: MIT Press, 1975.

CAMPBELL, J. Y., A. W. LO, AND A. C. MACKINLAY, *The Econometrics of Financial Markets* Princeton, NJ: Princeton University Press, 1997.

COVER, T. M., AND J. A. THOMAS, *Elements of Information Theory* New York: Wiley, 1991.

DAHLQUIST, G., A. BJÖRK, AND N. ANDERSON, *Numerical Analysis* New York: Prentice Hall, 1974.

DEMPSTER, A., M. M. LAIRD, AND D. B. RUBIN, "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm," *J. Roy. Statist. Soc. B*, 1–38 (1977).

DHARMADHIKARI, S., AND K. JOAG-DEV, "Examples of Nonunique Maximum Likelihood Estimators," *The American Statistician, 39*, 199–200 (1985).

EISENHART, C., "The Meaning of Least in Least Squares," *Journal Wash. Acad. Sciences, 54*, 24–33 (1964).

FAN, J., AND I. GIJBELS, *Local Polynomial Modelling and Its Applications* London: Chapman and Hall, 1996.

FISHER, R. A., "On the Mathematical Foundations of Theoretical Statistics," reprinted in *Contributions to Mathematical Statistics* (by R. A. Fisher 1950) New York: J. Wiley and Sons, 1922.

GOLUB, G. H., AND C. F. VAN LOAN, *Matrix Computations* Baltimore: John Hopkins University Press, 1985.

HABERMAN, S. J., *The Analysis of Frequency Data* Chicago: University of Chicago Press,1974.

KOLMOGOROV, A. N., "On the Shannon Theory of Information Transmission in the Case of Continuous Signals," *IRE Transf. Inform. Theory, IT2*, 102–108 (1956).

LITTLE, R. J. A., AND D. B. RUBIN, *Statistical Analysis with Missing Data* New York: J. Wiley, 1987.

MACLACHLAN, G. J., AND T. KRISHNAN, *The EM Algorithm and Extensions* New York: Wiley, 1997.

MOSTELLER, F., "Association and Estimation in Contingency Tables," *J. Amer. Statist. Assoc., 63*, 1–28 (1968).

RICHARDS, F. J., "A Flexible Growth Function for Empirical Use," *J. Exp. Botany, 10*, 290–300 (1959).

RUPPERT, D., AND M. P. WAND, "Multivariate Locally Weighted Least Squares Regression," *Ann. Statist., 22*, 1346–1370 (1994).

SEBER, G. A. F., AND C.J. WILD, *Nonlinear Regression* New York: Wiley, 1989.

SHANNON, C. E., "A Mathematical Theory of Communication," *Bell System Tech. Journal, 27*, 379–243, 623–656 (1948).

SNEDECOR, G. W., AND W. COCHRAN, *Statistical Methods*, 6th ed. Ames, IA: Iowa State University Press, 1967.

STIGLER, S., *The History of Statistics* Cambridge, MA: Harvard University Press, 1986.

WEISBERG, S., *Applied Linear Regression*, 2nd ed. New York: Wiley, 1985.

WU, C. F. J., "On the Convergence Properties of the EM Algorithm," *Ann. Statist., 11*, 95–103 (1983).