Chapter 4. Method of Maximum Likelihood

1 Introduction

Many statistical procedures are based on statistical models which specify under which conditions the data are generated. Usually the assumption is made that the set of observations $x_1, \ldots, x_n$ is a set of (i) independent random variables (ii) identically distributed with common pdf $f(x_i, \theta)$. Once this model is specified, the statistician tries to find optimal solutions to his problem (usually related to the inference on a set of parameters $\theta \in \Theta \subset R^k$, characterizing the uncertainty about the model).

The procedure just described is not always easy to carry out. In fact, when confronted with a set of data three attitudes are possible:

- The statistician may be a "pessimist" who does not believe in any particular model $f(x, \theta)$. In this case he must be satisfied with descriptive methods (like exploratory data analysis) without the possibility of inductive inference.
- The statistician may be an "optimist" who strongly believes in one model. In this case the analysis is straightforward and optimal solutions may often be easily obtained.
- The statistician may be "realist": he would like to specify a particular model $f(x, \theta)$ in order to get operational results but he may have either some doubt about the validity of this hypothesis or some difficulty in choosing a particular parametric family.

Let us illustrate this kind of preoccupation with an example. Suppose that the parameter of interest is the "center" of some population. In many situations, the statistician may argue that, due to a central limit effect, the data are generated by a normal pdf. In this case the problem is restricted to the problem of inference on μ, the mean of the population. But in some cases, he may have some doubt about these central limit effects and may suspect some skewness and/or some kurtosis or he may suspect that some observations are generated by other models (leading to the presence of outliers).

In this context three types of question may be raised to avoid gross errors in the prediction, or in the inference:

- Does the optimal solution, computed for assumed model $f(x, \theta)$, still have "good" properties if the true model is a little different?

- Are the optimal solutions computed for other models near to the original one really substantially different?
- Is it possible to compute (exactly or approximately) optimal solutions for a wider class of models based on very few assumptions?

The first question is concerned with the sensitivity of a given criterion to the hypotheses (criterion robustness). In the second question, it is the sensitivity of the inference which is analyzed (inference robustness). The last question may be viewed as a tentative first step towards the development of nonparametric methods (i.e. methods based on a very large parametric space).

2 Information Bound

Any statistical inference starts from a basic family of probability measures, expressing our prior knowledge about the nature of the probability measures from where the observations originate. Or a model P is a collection of probability measures P on $(\mathcal{X}, \mathcal{A})$ where X is the sample space with a σ-field of subsets A. If

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}, \;\; \Theta \subset R^s$$

for some k, then P is a *parametric model*. On the other hand, if

$$\mathcal{P} = \{\text{all } P \text{ on } (\mathcal{X}, \mathcal{A})\},$$

then P is often referred to as a *nonparametric model*.

Suppose that we have a fully specified parametric family of models. Denote the parameter of interest by θ. Suppose that we wish to calculate from the data a single value representing the "best estimate" that we can make of the unknown parameter. We call such a problem one of *point estimation*.

Define the *information matrix* as the $s \times s$ matrix

$$\mathbf{I}(\theta) = \|\mathbf{I}_{ij}(\theta)\|,$$

where

$$\mathbf{I}_{ij}(\theta) = E_\theta \left[\frac{\partial \log f(X;\theta)}{\partial \theta_i} \frac{\partial \log f(X;\theta)}{\partial \theta_j} \right].$$

When $k = 1$, $\mathbf{I}(\theta)$ is known as the *Fisher information*. Under regularity conditions, we have

$$E \left[\frac{\partial}{\partial \theta_i} \log f(X;\theta) \right] = 0 \tag{1}$$

and

$$\mathbf{I}_{ij}(\theta) = cov\left[\frac{\partial}{\partial\theta_i}\log f(X;\theta), \frac{\partial}{\partial\theta_j}\log f(X;\theta)\right].$$

Being a covariance matrix, $\mathbf{I}(\theta)$ is then positive semidefinite and positive definite unless the $(\partial/\partial\theta_i)\log f(X;\theta)$, $i = 1,\ldots,s$ are affinely dependent (and hence, by (1), linear dependent). When the density also has the second derivatives, we have the following alternative expression for $\mathbf{I}_{ij}(\theta)$ which is

$$\mathbf{I}_{ij}(\theta) = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X;\theta)\right].$$

To make above statements correct, we make the following assumptions when $s = 1$:

(i) Θ is an open interval (finite, infinite, or semi-infinite).

(ii) The distribution P_θ have common support, so that without loss of generality the set $A = \{x : p_\theta(x) > 0\}$ is independent of θ. (2)

(iii) For any x in A and θ in Θ, the derivative $p'_\theta(x) = \partial p_\theta(x)/\partial\theta$ exists and is finite.

Lemma 1 *(i) If (2) holds, and the derivative with respect to θ of the left side of*

$$\int f(x;\theta)d\mu(x) = 1 \tag{3}$$

can be obtained by differentiating under the integral sign, then

$$E_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] = 0$$

and

$$\mathbf{I}(\theta) = var_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]. \tag{4}$$

(ii) If, in addition, the second derivative with respect to θ of $\log f(X;\theta)$ exists for all x and θ and the second derivative with respect to θ of the left side of (3) can be obtained by differentiating twice under the integral sign, then

$$I(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right].$$

Let us now derive the *information inequality* for $s = 1$.

Theorem 1 *Suppose (2) and (4) hold, and that $I(\theta) > 0$. Let δ be any statistic with $E_\theta(\delta^2) < \infty$ for which the derivative with respect to θ of $E_\theta(\delta)$ exists and can be obtained by differentiating under the integral sign. Then*

$$var_\theta(\delta) \geq \frac{\left[\frac{\partial}{\partial\theta}E_\theta(\delta)\right]^2}{I(\theta)}.$$

Proof. For any estimator δ of $g(\theta)$ and any function $\psi(x,\theta)$ with finite second moment, the *covariance inequality* states that

$$var_\theta(\delta) \geq \frac{[cov(\delta,\psi)]^2}{var(\psi)}. \tag{5}$$

Denote $g(\theta) = E_\theta\delta$ and set

$$\psi(X,\theta) = \frac{\partial}{\partial\theta}\log f(X;\theta).$$

If differentiation under the integral sign is permitted in $E_\theta\delta$, it then follows that

$$cov(\delta,\psi) = \int \delta(x)\frac{f^{'}(x;\theta)}{f(x;\theta)}f(x;\theta)dx = g^{'}(\theta)$$

and hence

$$var_\theta(\delta) \geq \frac{\left[g^{'}(\theta)\right]^2}{var\left[\frac{\partial}{\partial\theta}\log f(X,\theta)\right]}.$$

This completes the proof of this theorem.

If δ is an unbiased estimator of θ, then

$$var_\theta(\delta) \geq \frac{1}{nI(\theta)}.$$

The above inequality provides a lower bound for the variance of any estimator. In fact, the quantity $nI(\theta)$ is known as the "Cramer-Rao lower bound." Likeliwise, we can also have the information inequality for general s. We begin by generalizing the correlation inequality to one involving many ψ_i $(i = 1,\ldots,r)$.

Theorem 2 *For any unbiased estimator δ of $g(\theta)$ and any functions $\psi_i(x,\theta)$ with finite second moments, we have*

$$var(\delta) \geq \gamma^{'}C^{-1}\gamma, \tag{6}$$

where $\gamma = (\gamma_1,\cdots,\gamma_r)$ and $C = \|C_{ij}\|$ are defined by

$$\gamma_i = cov(\delta,\psi_i), \quad C_{ij} = cov(\psi_i,\psi_j). \tag{7}$$

Proof. Replace Y by δ and X_i by $\psi_i(X,\theta)$ in the following lemma. Then the fact that $\rho^{*2} \leq 1$ yields this theorem.

Let $(X_1,\ldots,X_r)$ and Y be random variables with finite second moment, and consider the correlation coefficient $corr(\sum a_iX_i, Y)$. Its maximum value ρ^* over all $(a_1,\ldots,a_r)$ is the *multiple correlation coefficient* between Y and the vector $(X_1,\ldots,X_r)$.

Lemma 2 *Let $(X_1, \ldots, X_r)$ and Y have finite second moment, let $\gamma_i = cov(X_i, Y)$ and Σ be the covariance matrix of the X's. Without loss of generality, suppose Σ is positive definite. Then*

$$\rho^{*2} = \frac{\gamma' \Sigma^{-1} \gamma}{var(Y)}. \tag{8}$$

Proof. Since a correlation coefficient is invariant under scale changes, the a's maximizing (8) are not uniquely determined. Without loss of generality, we therefore impose the condition $var(\sum_i a_i X_i) = \mathbf{a}' \Sigma \mathbf{a} = 1$. In view of $\mathbf{a}' \Sigma \mathbf{a} = 1$,

$$corr(\sum_i a_i X_i, Y) = \mathbf{a}' \gamma / \sqrt{var(Y)}.$$

The problem then becomes that of maximizing $\mathbf{a}' \gamma$ subject to $\mathbf{a}' \Sigma \mathbf{a} = 1$. Using the method of undetermined multipliers, one maximizes instead

$$\mathbf{a}' \gamma - \frac{\lambda}{2} \mathbf{a}' \Sigma \mathbf{a} \tag{9}$$

with respect to $\mathbf{a}$ and then determines λ so as to satisfy $\mathbf{a}' \Sigma \mathbf{a} = 1$. Differentiation with respect to the a_i of (9) leads to a system of linear equations with the unique solution

$$\mathbf{a} = \frac{1}{\lambda} \Sigma^{-1} \gamma, \tag{10}$$

and the side condition $\mathbf{a}' \Sigma \mathbf{a} = 1$ gives

$$\lambda = \pm \sqrt{\gamma' \Sigma^{-1} \gamma}.$$

Substituting these values of λ into (10), one finds that

$$\mathbf{a} = \frac{\pm \Sigma^{-1} \gamma}{\sqrt{\gamma' \Sigma^{-1} \gamma}}$$

and the maximum value of $corr(\sum_i a_i X_i, Y)$, ρ^*, is therefore the positive root of (8).

Note that always $0 \leq \rho^* \leq 1$, and that ρ^* is 1 if and only if constants $a_1, \ldots, a_r$ and b exist such that $Y = \sum_i a_i X_i + b$.

Let us now state the information inequality for the multiparameter case in which $\theta = (\theta_1, \ldots, \theta_s)$.

Theorem 3 *Suppose that (1) holds and that $\mathbf{I}(\theta)$ is positive definite Let δ be any statistic with $E_\theta(\delta^2) < \infty$ for which the derivative with respect to θ_i exists for each i and can be obtained by differentiating under the integral sign. Then*

$$var_\theta(\delta) \geq \alpha' \mathbf{I}^{-1}(\theta) \alpha, \tag{11}$$

where α' is the row matrix with ith element

$$\alpha_i = \frac{\partial}{\partial \theta_i} E_\theta(\delta(\mathbf{X})).$$

Proof. If the functions ψ_i of Theorem 2 are taken to be $\psi_i = (\partial/\partial\theta_i) \log f(\mathbf{X}; \theta)$, this theorem follows immediately.

Under regularity conditions on the class of estimators $\hat{\theta}_n$ under consideration, it may be asserted that if $\hat{\theta}_n$ is $AN(\theta, n^{-1}\Sigma(\theta))$, then the condition

$$\Sigma(\theta) - \mathbf{I}(\theta)^{-1} \text{ is nonnegative definite}$$

must hold. (Read Ch2.6 and 2.7 of Lehmann (1983) for further details.) In this respect, an estimator $\hat{\theta}_n$ which is $AN(\theta, \Sigma_\theta)$ is "optimal." (Such an estimator need not exist.)

The following definition is thus motivated. An estimator $\hat{\theta}_n$ which is called *asymptotically efficient*, or *best asymptotically normal* (BAN). Under suitable regularity conditions, an asymptotically efficient estimate exists. One approach toward finding such estimates is the method of maximum likelihood. Neyman (1949) pointed out that these large-sample criteria were also satisfied by other estimates. He defined a class of best asymptotically normal estimates. So far, we have described three desirable properties $\hat{\theta}_n$. They are unbiasedness, consistency, and efficiency. We now describe a general procedure to produce an asymptotic unbiased, consistent, and asymptotic efficient estimator.

3 Maximum Likelihood Methodology

Many statistical techniques were invented in the nineteenth century by experimental scientists who personally applied their methods to authentic data sets. In these conditions the limits of what is computationally feasible are spontaneously observed. Until quite recently these limits were set by the capacity of the human calculator, equipped with pencil and paper and with such aids as the slide rule, tables of logarithms, and other convenient tables, which have been in constant use from the seventeenth century until well into the twentieth. Until the advent of the electronic computer, the powers of the human operator set the standard. This restriction has left its mark on statistical technique, and many new developments have taken place since it was lifted.

The first result of this modern computing revolution is that estimates defined by nonlinear equations can be established as a matter of routine by the appropriate iterative algorithms. This permits the use of nonlinear functional

forms. Although the progress of computing technology made nonlinear estimation possible, the statistical theory of Maximum Likelihood provided techniques and respectability. Its principle was first put forward as a novel and original method of deriving estimators by R.A. Fisher in the early 1920s. It very soon proved to be a fertile approach to statistical inference in general, and was widely adopted; but the exact properties of the ensuring estimators and test procedures were only gradually discovered.

Let observations $\mathbf{x} = (x_1, \ldots, x_n)$ be realized values of random variables $\mathbf{X} = (X_1, \ldots, X_n)$ and suppose that the random vector $\mathbf{X}$, having density $f_{\mathbf{X}}(\mathbf{x}; \theta)$ with respect to some σ-finite measure ν. Here θ is the scalar parameter to be determined. The likelihood function corresponding to an observed vector $\mathbf{x}$ from the density $f_{\mathbf{X}}(\mathbf{x}; \theta)$ is written

$$Lik_{\mathbf{X}}(\theta^{'}; \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}; \theta^{'}),$$

whose logarithm is denoted by $L(\theta^{'}; \mathbf{x})$. When the X_i are iid with probability density $f(\mathbf{x}; \theta)$ with respect to a σ-finite measure μ,

$$f(\mathbf{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

If the parameter space is Ω, then the maximum likelihood estimate (MLE) $\hat{\theta} = \hat{\theta}(\mathbf{x})$ is that value of $\theta^{'}$ maximizing $lik_{\mathbf{X}}(\theta^{'}; \mathbf{x})$, or equivalently its logarithm $L(\theta^{'}; \mathbf{x})$, over Ω. That is,

$$L(\hat{\theta}; \mathbf{x}) \geq L(\theta^{'}; \mathbf{x}) \ \ (\theta^{'} \in \Omega). \tag{12}$$

$L(\theta, \mathbf{x})$ is called the *log-likelihood*. Note that L is regarded as a function of θ with $\mathbf{x}$ fixed.

A MLE may not exist. It certainly exists if Ω is compact and $f(x; \theta)$ is upper semicontinuous in θ for all x. As an example, consider $U(\theta, \theta + 1)$. Later on, we shall use the shorthand notation $L(\theta)$ for $L(\theta, \mathbf{x})$ and $L^{'}(\theta), L(\theta)^{''}, \ldots$ for its derivatives with respect to θ. (Note that f is said to be upper semicontinuous if $\{x | f(x) < \alpha\}$ is an open set.)

Fisher was the first to study and establish optimum properties of estimates obtained by maximizing the likelihood function, using criteria such as consistency and efficiency (involving asymptotic variances) in large samples. At that time, however, the computation involved were hardly practicable, this prevented a widespread adoption of these methods. Fortunately, the new computer technology had become generally accessible. Therefore, Maximum Likelihood (ML) methodology is widely used now.

It is a constant theme of the history of the method that the use of ML techniques is not always accompanied by a clear appreciation of their limitations. Le Cam (1953) complains that

> $\cdots$ although all efforts at a proof of the general existence of [asymptotically] efficient estimates $\cdots$ as well as a proof of the efficiency of ML estimates were obviously inaccurate and although accurate proofs of similar statements always referred not to the general case but to particular classes of estimates $\cdots$ a general belief became established that the above statements are true in the most general sense.

As an illustration, consider the famous Neyman-Scott (1948) problem. In this example, the MLE is not even consistent.

Example 1. **Estimation of a Common Variance.** Let $X_{\alpha j}$ $(j = 1, \ldots, r)$ be independently distributed according to $N(\theta_\alpha, \sigma^2)$, $\alpha = 1, \ldots, n$. The MLEs are

$$\hat{\theta}_\alpha = X_{\alpha\cdot}, \quad \hat{\sigma}^2 = \frac{1}{rn} \sum \sum (X_{\alpha j} - X_{\alpha\cdot})^2.$$

Furthermore, these are the unique solutions of the likelihood equations.

However, in the present case, the MLE of σ^2 is not even consistent. To see this, note that the statistics

$$S_\alpha^2 = \sum (X_{\alpha j} - X_{\alpha\cdot})^2$$

are identically independently distributed with expectation

$$E(S_\alpha^2) = (r-1)\sigma^2$$

so that $\sum S_\alpha^2/n \to (r-1)\sigma^2$ and hence

$$\hat{\sigma}^2 \to \frac{r-1}{r}\sigma^2 \text{ in probability.}$$

Example 2. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a uniform distribution $U(0, \theta)$. The likelihood function is

$$L(\theta, \mathbf{x}) = \frac{1}{\theta^n}, \quad 0 < x_1, \ldots, x_n < \theta.$$

Clearly L cannot be maximized wrt θ by differentiation. However, it is not difficult to find $\hat{\theta}_n = X_{(n)}$ with density function nt^{n-1}/θ^n where $t \in (0, \theta)$. Then

$$E(\hat{\theta}_n) = \frac{n\theta}{n+1},$$

which is a biased estimator of θ. (But it is asymptotic unbiased.)

3.1 Efficient Likelihood Estimation

According to the example discussed by Neyman and Scott (1948), we will show that, under regularity conditions, the ML estimates are consistent, asymptotically normal, and asymptotically efficient. For simplicity, our treatment will be confined to the case of a 1-dimensional parameter.

We begin with the following regularity assumptions:

(A0) The distributions P_θ of the observations are distinct (otherwise, θ cannot be estimated consistently).

(A1) The distributions P_θ have common support.

(A2) The observations are $\mathbf{X} = (X_1, \ldots, X_n)$, where the X_i are iid with probability density $f(x_i, \theta)$ with respect to μ.

(A3) The parameter space Ω contains an open interval ω of which the true parameter value θ_0 is an interior point.

Theorem 4 *Under assumptions (A0)-(A2),*

$$P_{\theta_0}\{f(X_1, \theta_0) \cdots f(X_n, \theta_0) > f(X_1, \theta) \cdots f(X_n, \theta)\} \to 1$$

as $n \to \infty$ for any fixed $\theta \neq \theta_0$.

Proof. The inequality is equivalent to

$$\frac{1}{n} \sum_{i=1}^{n} \log\left[f(X_i, \theta)/f(X_i, \theta_0)\right] < 0.$$

By the strong law of large numbers, the left side tends with probability 1 toward

$$E_{\theta_0} \log[f(X, \theta)/f(X, \theta_0)].$$

Since $-\log$ is strictly convex, Jensen's inequality shows that

$$E_{\theta_0} \log[f(X, \theta)/f(X, \theta_0)] < \log E_{\theta_0}[f(X, \theta)/f(X, \theta_0)] = 0, \tag{13}$$

and the results follows. When θ_0 is the true value, the above proof gives a meaning to the numerical value of the Kullback-Leibler information number. Namely, the likelihood ratio converges to zero exponential fast, at rate $I(\theta, \eta)$.

Remark 1. Define the *Kullback-Leibler information number*

$$I(\theta, \eta) = E_\theta \left(\log \frac{f(X, \theta)}{f(X, \eta)} \right).$$

Note that $I(\theta, \eta) \geq 0$ with equality holding if and only if, $f(x, \theta) = f(x, \eta)$. $I(\theta, \eta)$ is a measure of the ability of the likelihood ratio to distinguish between $f(X, \theta)$ and $f(X, \theta_0)$ when the latter is true.

Remark 2. If $\hat{\theta}_n$ is an MLE of θ and if g is a function, then $g(\hat{\theta}_n)$ is an MLE of $g(\theta)$. When g is one-to-one, it holds obviously. If g is many-to-one, this result holds again when the derivative of g is nonzero.

By Theorem 4, the density of $\mathbf{X}$ at the true θ_0 exceeds that any other fixed θ with high probability when n is large. We do not know θ_0 but we can determine the value $\hat{\theta}$ of θ which maximizes the density of $\mathbf{X}$. However, Theorem 4 cannot guarantee that the MLE is consistent since we have to apply the law of large numbers to the right-hand side of (13) for all $\theta' \neq \theta$ simultaneously. However, if Ω is finite, the MLE $\hat{\theta}_n$ exists, it is unique with probability tending to 1, and it is consistent.

The following theorem is motivated by the simple fact by differentiating $\int f(x, \theta)\mu(dx) = 1$ with respect to θ. It leads to

$$E_{\theta_0} \frac{f'(X, \theta_0)}{f(X, \theta_0)} = 0.$$

Theorem 5 *Let $X_1, \ldots, X_n$ satisfy assumptions (A0)-(A3) and suppose that for almost all x, $f(x, \theta)$ is differentiable with respect to θ in w, with derivative $f'(x, \theta)$. Then with probability tending to 1 as $n \to \infty$, the likelihood equation*

$$\frac{\partial}{\partial \theta} [f(x_1, \theta) \cdots f(x_n, \theta)] = 0 \tag{14}$$

or, equivalently, the equation

$$L'(\theta, \mathbf{x}) = \sum_{i=1}^{n} \frac{f'(x_i, \theta)}{f(x_n, \theta)} = 0 \tag{15}$$

has a root $\hat{\theta}_n = \hat{\theta}_n(x_1, \ldots, x_n)$ such that $\hat{\theta}_n(X_1, \ldots, X_n)$ tends to the true values θ_0 in probability.

Proof. Let a be small enough so that $(\theta_0 - a, \theta_0 + a) \subset w$, and let

$$S_n = \{\mathbf{x} : L(\theta_0, \mathbf{x}) > L(\theta_0 - a, \mathbf{x}) \text{ and } L(\theta_0, \mathbf{x}) > L(\theta_0 + a, \mathbf{x})\}. \tag{16}$$

By Theorem 4, $P_{\theta_0}(S_n) \to 1$. For any $\mathbf{x} \in S_n$ there thus exists a value $\theta_0 - a < \hat{\theta}_n < \theta_0 + a$ at which $L(\theta)$ has a local maximum, so that $L'(\hat{\theta}_n) = 0$. Hence for any $a > 0$ sufficiently small, there exists a sequence $\hat{\theta}_n = \hat{\theta}_n(a)$ of roots such that

$$P_{\theta_0}(|\hat{\theta}_n - \theta_0| < a) \to 1.$$

It remains to show that we can determine such a sequence, which does not depend on a.

Let $\hat{\theta}_n^*$ be the root closest to θ_0, This exists because the limit of a sequence of roots is again a root by the continuity of $L(\theta)$.) Then clearly $P_{\theta_0}(|\hat{\theta}_n^* - \theta_0| < a) \to 1$ and this completes the proof.

Remarks. 1. This theorem does not establish the existence of a consistent estimator sequence since, with the true value θ_0 unknown, the data do not tell us which root to choose so as to obtain a consistent sequence. An exception, of course, is the case in which the root is unique.

2. It should also be emphasized that existence of a root $\hat{\theta}_n$ is not asserted for all $\mathbf{x}$ (or for a given n even for any $\mathbf{x}$). This does not affect consistency, which only requires $\hat{\theta}_n$ to be defined on a set $S_n^{'}$, the probability of which tends to 1 as $n \to \infty$.

Above theorem establishes the existence of a consistent root of the likelihood equation. The next theorem asserts that any such sequence is asymptotically normal and efficient.

Theorem 6 *Suppose that $X_1, \ldots, X_n$ are iid and satisfy the assumptions (A0)-(A3), the integral $\int f(x,\theta))d\mu(x)$ can be twice differentiated under the integral sign, and the existence of a third derivative satisfying*

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x,\theta)\right| \leq M(x) \tag{17}$$

for all $x \in A$, $\theta_0 - c < \theta < \theta_0 + c$ with $E_{\theta_0}[M(X)] < \infty$. Then any consistent sequence $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ of roots of the likelihood equation satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right).$$

Proof. For any fixed x, expand $L^{'}(\hat{\theta}_n)$ about θ_0

$$L^{'}(\hat{\theta}_n) = L^{'}(\theta_0) + (\hat{\theta}_n - \theta_0)L^{''}(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 L^{(3)}(\theta_n^*)$$

where θ_n^* lies between θ_0 and $\hat{\theta}_n$. By assumption, the left side is zero, so that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{(1/\sqrt{n})L^{'}(\theta_0)}{-(1/n)L^{''}(\theta_0) - (1/2n)(\hat{\theta}_n - \theta_0)L^{(3)}(\theta_n^*)}$$

where it should be remembered that $L(\theta)$, $L^{'}(\theta)$, and so on are functions of $(X_1, \ldots, X_n)$ as well as θ. The desired result follows if we can show that

$$\frac{1}{\sqrt{n}}L^{'}(\theta_0) \xrightarrow{d} N[0, I(\theta_0)], \tag{18}$$

that

$$-\frac{1}{n}L^{''}(\theta_0) \xrightarrow{P} I(\theta_0) \tag{19}$$

and that

$$\frac{1}{n}L^{(3)}(\theta_n^*) \text{ is bounded in probabiliy.} \tag{20}$$

Of the above statements, (18) follows from the fact that

$$\frac{1}{\sqrt{n}}L^{'}(\theta_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{f^{'}(X_i,\theta_0)}{f(X_i,\theta_0)} - E_{\theta_0}\frac{f^{'}(X_i,\theta_0)}{f(X_i,\theta_0)}\right]$$

since the expectation term is zero, and then from the CLT and the definition of $I(\theta)$.

Next,

$$-\frac{1}{n}L^{''}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{f^{'2}(X_i,\theta_0) - f(X_i,\theta_0)f^{''}(X_i,\theta_0)}{f^2(X_i,\theta_0)}.$$

By the law of large numbers, this tends in probability to

$$I(\theta_0) - E_{\theta_0}\frac{f^{''}(X_i,\theta_0)}{f(X_i,\theta_0)} = I(\theta_0).$$

Finally,

$$\frac{1}{n}L^{(3)}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta^3}\log f(X_i,\theta)$$

so that by (17)

$$\left|\frac{1}{n}L^{(3)}(\theta_n^*)\right| < \frac{1}{n}[M(X_1) + \cdots + M(X_n)]$$

with probability tending to 1. The right side tends in probability to $E_{\theta_0}[M(X)]$, and this completes the proof.

Remarks. 1. This is a strong result. It establishes several major properties of the MLE in addition to its consistency. The MLE is asymptotically normal, which is of great help for the derivation of (asymptotically valid) tests; it is asymptotically unbiased; and it is asympotically efficient, since the variance of its limiting distribution equals the Cramer-Rao lower bound.
2. As a rule we wish to supplement the parameter estimates by an estimate of their (asymptotic) variance. This will permit us to assess (asymptotic) t-ratios and (asymptotic) confidence interval. Although the variance may depend on the unknown parameter, we can just use MLE to get an estimate of variance.

The usual iterative methods for solving the likelihood equation $L^{'}(\theta) = 0$ are based on replacing $L^{'}(\theta)$ by the linear terms of its Taylor expansion about an approximate solution $\tilde{\theta}$. Suppose we can use estimation method such as the

method of moments to find a good estimate of θ. Denote it as $\tilde{\theta}$. Then it is quite natural to use $\tilde{\theta}$ as the initial solution of the iterative methods. Denote the MLE by $\hat{\theta}$. This leads to the approximation

$$0 = L^{'}(\hat{\theta}) \approx L^{'}(\tilde{\theta}) + (\hat{\theta} - \tilde{\theta})L^{''}(\tilde{\theta}),$$

and hence to

$$\hat{\theta} = \tilde{\theta} - \frac{L^{'}(\tilde{\theta})}{L^{''}(\tilde{\theta})}.$$

The procedure is then iterated according to the above scheme.

The following is a justification for the use of the above one-step approximation as an estimator of θ.

Theorem 7 *Suppose that the assumptions of Theorem 6 hold and that $\tilde{\theta}$ is not only a consistent but a $\sqrt{n}$-consistent estimator of θ, that is, that $\sqrt{n}(\tilde{\theta} - \theta_0)$ is bounded in probability so that $\tilde{\theta}$ tends to θ_0 at least at the rate of $1/\sqrt{n}$. Then the estimator sequence*

$$\delta_n = \tilde{\theta} - \frac{L^{'}(\tilde{\theta})}{L^{''}(\tilde{\theta})} \tag{21}$$

is asymptotically efficient.

Proof. As in the proof of Theorem 6, expand $L^{'}(\tilde{\theta})$ about θ_0 as

$$L^{'}(\tilde{\theta}_n) = L^{'}(\theta_0) + (\tilde{\theta}_n - \theta_0)L^{''}(\theta_0) + \frac{1}{2}(\tilde{\theta}_n - \theta_0)^2 L^{(3)}(\theta_n^*)$$

where θ_n^* lies between θ_0 and $\tilde{\theta}_n$. Substituting this expression into (21) and simplifying, we find

$$\begin{aligned} \sqrt{n}(\delta_n - \theta_0) &= \frac{(1/\sqrt{n})L^{'}(\theta_0)}{-(1/n)L^{''}(\tilde{\theta}_n)} + \sqrt{n}(\tilde{\theta}_n - \theta_0) \\ &\quad \times \left[1 - \frac{L^{''}(\theta_0)}{L^{''}(\tilde{\theta}_n)} - \frac{1}{2}(\tilde{\theta}_n - \theta_0)\frac{L^{(3)}(\theta_n^*)}{L^{''}(\tilde{\theta}_n)}\right]. \end{aligned} \tag{22}$$

Suppose we can show that the expression in square brackets on the right hand side of (22) tends to zero in probability and $L^{''}(\tilde{\theta}_n)/L^{''}(\theta_0) \to 1$ in probability. This theorem will follows accordingly. These follows from $\theta_n^* \to \theta_0$ in probability and use the expansion

$$\frac{1}{n}L^{''}(\tilde{\theta}_n) = \frac{1}{n}L^{''}(\theta_0) + \frac{1}{n}(\tilde{\theta}_n - \theta_0)L^{(3)}(\theta_n^{**})$$

where θ_n^{**} is between θ_0 and $\tilde{\theta}_n$.

3.2 The Multi-parameter Case

We just discuss the case that the distribution depends on a single parameter θ. When extending this theory to probability models involving several parameters $\theta_1, \ldots, \theta_s$, one may be interested either in simultaneous estimation of these parameters (or certain functions of them) or with the estimation of part of the parameters. The part of parameter is of intrinsic and the rest represents *nuisance or incidental parameters* that are necessary for a proper statistical model but of no interest in themselves. For instance, we are only interested in estimating σ^2 in Neyman-Scott problem. Then θ_α are called nuisance parameters.

Let $(X_1, \ldots, X_n)$ be iid with a distribution that depends on $\theta = (\theta_1, \ldots, \theta_s)$ and satisfies assumptions (A0)-(A3). The information matrix $I(\theta)$ is an $s \times s$ matrix with elements $I_{jk}(\theta)$, $j, k = 1, \ldots, s$, defined by

$$I_{jk}(\theta) = cov\left[\frac{\partial}{\partial\theta_j}\log f(X,\theta), \frac{\partial}{\partial\theta_k}\log f(X,\theta)\right].$$

We shall now show under regularity conditions that with probability tending to 1 there exists solutions $\hat{\theta}_n = (\hat{\theta}_{1n}, \ldots, \hat{\theta}_{sn})$ of the likelihood equations

$$\frac{\partial}{\partial\theta_j}[f(x_1,\theta)\cdots f(x_n,\theta)] = 0, \quad j = 1, \ldots, s,$$

or equivalently

$$\frac{\partial}{\partial\theta_j}[L(\theta)] = 0, \quad j = 1, \ldots, s$$

such that $\hat{\theta}_{jn}$ is consistent for estimating θ_j and asymptotically efficient in the sense of with asymptotic variance $[I(\theta)]^{-1}_{jj}$.

We first state some assumptions:

(A) There exists an open subset ω of Ω containing the true parameter point θ^0 such that for almost all x the density $f(x,\theta)$ admits all third derivatives $(\partial^3/\partial\theta_j\partial\theta_k\partial\theta_\ell)f(x,\theta)$ for all $\theta \in \omega$.

(B) the first and second logarithmic derivatives of f satisfy the equations

$$E_\theta\left[\frac{\partial}{\partial\theta_j}\log f(X,\theta)\right] = 0 \text{ for } j = 1, \ldots, s,$$

and

$$\begin{aligned} I_{jk}(\theta) &= E_\theta\left[\frac{\partial}{\partial\theta_j}\log f(X,\theta)\cdot\frac{\partial}{\partial\theta_k}\log f(X,\theta)\right] \\ &= E_\theta\left[-\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X,\theta)\right]. \end{aligned}$$

(C) Since the $s \times s$ matrix $I(\theta)$ is a covariance matrix, it is positive semidefinite. We shall assume that the $I_{jk}(\theta)$ are finite and that the matrix $I(\theta)$ is positive definite for all θ in ω, and hence that the s statistics

$$\frac{\partial}{\partial\theta_1}\log f(X,\theta),\ldots,\frac{\partial}{\partial\theta_s}\log f(X,\theta)$$

are affinely independent with probability 1.

(D) Finally, we shall suppose that there exists functions $M_{jk\ell}$ such that

$$\left|\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_\ell}\log f(x,\theta)\right| \le M_{jk\ell}(x) \text{ for all } \theta\in\omega$$

where $m_{jk\ell} = E_{\theta^0}\left[M_{jk\ell}(X)\right] < \infty$ for all j,k,ℓ.

Theorem 8 *Let $X_1,\ldots,X_n$ be iid each with a density $f(x,\theta)$ (with respect to μ) which satisfies (A0)-(A2) and assumptions (A)-(D) above. Then with probability tending to 1 as $n\to\infty$, there exist solutions $\hat\theta_n = \hat\theta_n(X_1,\ldots,X_n)$ of the likelihood equations such that*
(i) $\hat\theta_{jn}$ is consistent for estimating θ_j,
(ii) $\sqrt{n}(\hat\theta_n-\theta)$ is asymptotically normal with (vector) mean zero and covariance matrix $[I(\theta)]^{-1}$ and
(iii) $\hat\theta_{jn}$ is asymptotically efficient in the sense that

$$\sqrt{n}(\hat\theta_{jn}-\theta_j) \xrightarrow{L} N(0,[I(\theta)]^{-1}_{jj}).$$

Proof. (i) *Existence and Consistency.* To prove the consistence, with probability tending to 1, of a sequence of solutions of the likelihood equations which is consistent, we shall consider the behavior of the log likelihood $L(\theta)$ on the sphere Q_a with center at the true point θ^0 and radius a. We will show that for any sufficiently small a the probability tends to 1 that $L(\theta) < L(\theta^0)$ at all points θ on the surface of Q_a, and hence that $L(\theta)$ has a local maximum in the interior of Q_a. Since at a local maximum the likelihood equations must be satisfied it will follow that for any $a>0$, with probability tending to 1 as $n\to\infty$, the likelihood equations have a solution $\hat\theta_n(a)$ within Q_a and the proof can be completed as in the one-dimensional case.

To obtain the needed facts concerning the behavior of the likelihood on Q_a for small a, we expand the log likelihood about the true point θ^0 and divide by n to find

$$\begin{aligned}\frac{1}{n}L(\theta)-\frac{1}{n}L(\theta^0) &= \frac{1}{n}\sum A_j(x)(\theta_j-\theta_j^0)+\frac{1}{2n}\sum\sum B_{jk}(x)(\theta_j-\theta_j^0)(\theta_k-\theta_k^0)\\ &+ \frac{1}{6n}\sum_j\sum_k\sum_\ell(\theta_j-\theta_j^0)(\theta_k-\theta_k^0)(\theta_\ell-\theta_\ell^0)\sum_{i=1}^n\gamma_{jk\ell}(x_i)M_{jk\ell}(x_i)\\ &= S_1+S_2+S_3\end{aligned}$$

where

$$A_j(x) = \frac{\partial}{\partial\theta_j}L(\theta)\bigg|_{\theta=\theta^0}, \quad B_{jk}(x) = \frac{\partial^2}{\partial\theta_j\partial\theta_k}L(\theta)\bigg|_{\theta=\theta^0},$$

and where by assumption (D)

$$0 \le |\gamma_{jk\ell}(x)| \le 1.$$

To prove that the maximum of this difference for θ on Q_a is negative with probability tending to 1 if a is sufficiently small, we will show that with high probability the maximum of S_2 is negative while S_1 and S_3 are small compared to S_2. The basic tools for showing this are the facts that by (B) and the law of large numbers

$$\frac{1}{n}A_j(x) = \frac{1}{n}\frac{\partial}{\partial\theta_j}L(\theta)\bigg|_{\theta=\theta^0} \to 0 \text{ in probability.} \tag{23}$$

and

$$\frac{1}{n}B_{jk}(x) = \frac{1}{n}\frac{\partial^2}{\partial\theta_j\partial\theta_k}L(\theta)\bigg|_{\theta^0} \to -I_{jk}(\theta^0) \text{ in probability.} \tag{24}$$

Let us begin with S_1. On Q_a we have

$$|S_1| \le \frac{1}{n}a\sum_j |A_j(X)|.$$

For any given a, it follows from (23) that $|A_j(X)|/n < a^2$ and hence that $|S_1| < sa^3$ with probability tending to 1. Next consider

$$\begin{aligned} 2S_2 &= \sum\sum\left[-I_{jk}(\theta^0)(\theta_j-\theta_j^0)(\theta_k-\theta_k^0)\right] \\ &\quad + \sum\sum\left\{\frac{1}{n}B_{jk}(X) - [-I_{jk}(\theta^0)]\right\}(\theta_j-\theta_j^0)(\theta_k-\theta_k^0). \end{aligned}$$

For the second term it follows from an argument analogous to that for S_1 that its absolute value is less than s^2a^3 with probability tending to 1. The first term is a negative (nonrandom) quadratic form in the variables $(\theta_j - \theta_j^0)$. By an orthogonal transformation this can be reduced to diagonal form $\sum\lambda_i\xi_i^2$ with Q_a becoming $\sum\xi_i^2 = a^2$. Suppose that the λ's that are negative are numbered so that $\lambda_s \le \lambda_{s-1} \le \cdots \le \lambda_1 < 0$. Then $\sum\lambda_i\xi_i^2 \le \lambda_1\sum\xi_i^2 = \lambda_1 a^2$. Combining the first and second terms, we see that there exist $c > 0$, $a_0 > 0$ such that for $a < a_0$

$$S_2 < -ca^2$$

with probability tending to 1.

Finally, with probability tending to 1,

$$\left|\frac{1}{n}\sum M_{jk\ell}(X_i)\right| < 2m_{jk\ell}$$

and hence $|S_3| < ba^3$ on Q_a where

$$b = \frac{s^3}{3} \sum\sum\sum m_{jk\ell}.$$

Combining the three inequalities, we see that

$$\max(S_1 + S_2 + S_3) < -ca^2 + (b+s)a^3$$

which is less than zero if $a < c/(b+s)$, and this completes the proof of (i).

The proof of part (ii) of Theorem 8 is basically the same as that of Theorem 6. However, the single equation derived there from the expansion of $\hat{\theta}_n - \theta_0$ is now replaced by a system of s equations which must be solved for the differences $(\hat{\theta}_{jn} - \theta_j^0)$. In preparation, it will be convenient to consider quite generally a set of random linear equations in s unknowns

$$\sum_{k=1}^{s} A_{jkn} Y_{kn} = T_{jn} \quad (j = 1, \ldots, s). \tag{25}$$

Lemma 3 *Let $(T_{1n}, \ldots, T_{sn})$ be a sequence of random vectors tending weakly to $(T_1, \ldots, T_s)$ and suppose that for each fixed j and k, A_{jkn} is a sequence of random variables tending in probability to constants a_{jk} for which the matrix $A = \|a_{jk}\|$ is nonsingular. Let $B = \|b_{jk}\| = A^{-1}$. Then if the distribution of $(T_1, \ldots, T_s)$ has a density with respect to Lebesgue measure over E_s, the solution of (25) tend in probability to the solutions $(Y_1, \ldots, Y_s)$ of $\sum_{k=1}^{s} a_{jk} Y_k = T_j$, $1 \le j \le s$, given by $Y_j = \sum_{k=1}^{s} b_{jk} T_k$.*

In generalization of the proof of Theorem 6, expand $\partial L(\theta)/\partial\theta_j = L_j'(\theta)$ about θ^0 to obtain

$$L_j'(\theta) = L_j'(\theta^0) + \sum(\theta_k - \theta_k^0) L_{jk}''(\theta^0) + \frac{1}{2} \sum\sum(\theta_k - \theta_k^0)(\theta_\ell - \theta_\ell^0) L_{jk\ell}^{(3)}(\theta^*) \tag{26}$$

where L_{jk}'' and $L_{jk\ell}^{(3)}$ denote the indicated second and third derivatives of L and where θ^* is a point on the line segment connecting θ and θ^0. In this expansion, replace θ by a solution $\hat{\theta}_n$ of the likelihood equations, which by part (i) of this theorem can be assumed to exist with probability tending to 1 and to be consistent. The left side of (26) is zero and the resulting equations can be written as

$$\sqrt{n} \sum(\hat{\theta}_k - \theta_k^0) \left[\frac{1}{n} L_{jk}''(\theta^0) + \frac{1}{2n} L_{jk\ell}'''(\theta^*) \right] = -\frac{1}{\sqrt{n}} L_j'(\theta^0). \tag{27}$$

These have the form (26) with

$$Y_{kn} \;=\; \sqrt{n}(\hat{\theta}_k - \theta_k^0) \tag{28}$$

$$A_{jkn} = \frac{1}{n}L''_{jk}(\theta^0) + \frac{1}{2n}(\hat{\theta}_\ell - \theta^0_\ell)L^{(3)}_{jk\ell}(\theta^*) \quad (29)$$

$$T_{jn} = -\frac{1}{\sqrt{n}}L'_j(\theta^0) = -\frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta_j}\log f(X_i,\theta)\right]_{\theta=\theta^0}. \quad (30)$$

Since $E_{\theta^0}[(\partial/\partial\theta_j)\log f(X_i,\theta)] = 0$, the multivariate central limit theorem shows that $(T_{1n},\ldots,T_{sn})$ has a multivariate normal distribution with mean zero and covariance matrix $I(\theta^0)$.

On the other hand, it is easy to see-again in parallel to the proof of Theorem 6 that

$$A_{jkn} \xrightarrow{P} a_{jk} = E[L''_{jk}(\theta^0)] = -I_{jk}(\theta^0).$$

The limit distribution of the Y's is therefore that of the solution $(Y_1,\ldots,Y_s)$ of the equations

$$\sum_{k=1}^{s} I_{jk}(\theta^0)Y_k = T_j$$

where $T = (T_1,\ldots,T_s)$ is multivariate normal with mean zero and covariance matrix $I(\theta^0)$. It follows that the distribution of Y is that of $[I(\theta^0)]^{-1}T$, which is a multivariate distribution with zero mean and covariance matrix $[I(\theta^0)]^{-1}$. This completes the proof of asymptotic normality and efficiency.

3.3 Efficiency and Adaptiveness

If the distribution of the X_i depends on $\theta = (\theta_1,\ldots,\theta_s)$, it is interesting to compare the estimation of θ_j when the other parameters are unknown with the situation in which they are known. Such a question arises naturally in the case that part of parameters are the *nuisance* parameter. For instance, consider estimating $\tilde{\mu}$ for a location family $f(x-\tilde{\mu})$ or the median. $\tilde{\mu}$, of a symmetric density, f. Then $\tilde{\mu}$ is the parameter of interest and f is the nuisance parameter. If f is known and continuously differentiable, the best asymptotic mean-squared error attainable for estimating $\tilde{\mu}$ is $(nI)^{-1}$ where

$$I = \int \frac{f'^2(x)}{f(x)}dx < \infty.$$

The question be asked is when can we estimate $\tilde{\mu}$ as well asymptotically not knowing f as knowing f. A necessary condition named the orthogonality condition is given in Stein (1956). If there exists an estimate achieving the bound $(nI)^{-1}$ when f is unknown, it is named as an adaptive estimate of $\tilde{\mu}$. According to Stein's condition, he indicated that such an estimator does exist for this problem. Completely definite results for this problem were obtained by Beran (1974) and Stone (1975).

Note that this problem is a so-called semiparametric estimation problem in which $\tilde{\mu}$ is the parametric component and f is the nonparametric component. Recently, the problem of estimating and testing hypotheses about the parametric component in the presence of an infinite dimensional nuisance parameter (nonparametric component) attracts a lot of attention. Main concerns are whether there exists either an adaptive or efficient estimate of the parametric component and the existence of a practical procedure to find them.

We now consider the finite-dimensional case and derive the orthogonality condition derived in Stein (1956). It was seen that under regularity conditions there exist estimator sequences $\hat{\theta}_{nj}$ of θ_j, when the other parameters are known, which are asymptotically efficient in the sense that

$$\sqrt{n}(\hat{\theta}_{jn} - \theta_j) \xrightarrow{d} N(0, \frac{1}{I_{jj}(\theta)}).$$

When the other parameters are unknown,

$$\sqrt{n}(\hat{\theta}_{jn} - \theta_j) \xrightarrow{d} N(0, [I(\theta)]^{-1}_{jj}).$$

These imply that

$$\frac{1}{I_{jj}(\theta)} \leq [I(\theta)]^{-1}_{jj}. \tag{31}$$

Stein (1956) raised the question whether we can estimate θ_j equally well no matter when the other parameters are known or not. This leads to the question of *efficiency* and *adaptiveness*.

The two sides of (31) are equal if

$$I_{ij}(\theta) = 0 \text{ for all } j \neq i, \tag{32}$$

as is seen from the definition of the inverse of a matrix, and in fact (32) is also necessary for equality in (31) by the following facts.

Fact. Let $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ be a partitioned matrix with A_{22} square and nonsingular, and let

$$B = \begin{pmatrix} I & -A_{12}A_{22}^{-1} \\ 0 & I \end{pmatrix}.$$

Note that

$$BA = \begin{pmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & 0 \\ A_{21} & A_{22} \end{pmatrix}.$$

It follows easily that $|A| = |A_{11} - A_{12}A_{22}^{-1}A_{21}| \cdot |A_{22}|$. Since A_{22} is nonsingular, $(A^{-1})_{11} = (A_{11})^{-1}$ if A_{12} is a zero matrix.

The equality in (31) implies that $I(\theta)$ is diagonal. Suppose the efficient estimator of θ_j depends on the remaining parameters and yet θ_j can be estimated without loss of efficiency when these parameters are unknown. The situation can then be viewed as a rather trivial example of the idea of adaptive estimation. On the other hand, it is known that $I_{jj}(\theta)$ is the smallest asymptotic mean-squared error attainable for estimating θ_j. If an estimator does achieve such a bound, it is called an efficient estimator. Then Stein (1956) states that the adaptation is not possible unless $I_{jk}(\theta) = 0$ for $k \neq j$.

We now study the bound of $[I(\theta)]_{11}^{-1}$. Write $I(\theta)$ as a partitioned matrix

$$\begin{pmatrix} I_{11}(\theta) & I_{1\cdot}(\theta) \\ I_{1\cdot}^T(\theta) & I_{\cdot\cdot}(\theta) \end{pmatrix},$$

where $I_{1\cdot}(\theta) = (I_{12}(\theta), \ldots, I_{1s}(\theta))$ and $I_{\cdot\cdot}(\theta)$ is the lower right submatrix of $I(\theta)$ with size $(s-1) \times (s-1)$. Then

$$[I(\theta)]_{11}^{-1} = \frac{1}{I_{11}(\theta) - I_{1\cdot}(\theta)[I_{\cdot\cdot}(\theta)]^{-1} I_{1\cdot}^T(\theta)}.$$

Recall that

$$I_{ij}(\theta) = E\left(\frac{\partial}{\partial\theta_i}\log f(X,\theta) \cdot \frac{\partial}{\partial\theta_j}\log f(X,\theta)\right).$$

Consider the minimization problem

$$\min_{a_j} E\left(\frac{\partial}{\partial\theta_1}\log f(X,\theta) - \sum_{j=2}^{s} a_j \frac{\partial}{\partial\theta_j}\log f(X,\theta)\right)^2.$$

and denote the minimizer as $\mathbf{a}^0 = (a_{20}, \ldots, a_{s0})^T$. By a simple algebra, $\mathbf{a}^0$ is the solution of norml equations $I_{\cdot\cdot}(\theta)\mathbf{a}^0 = I_{1\cdot}(\theta)$ or $\mathbf{a}^0 = I_{\cdot\cdot}^{-1}(\theta) I_{1\cdot}(\theta)$. It leads to

$$E\left(\frac{\partial}{\partial\theta_1}\log f(X,\theta) - \sum_{j=2}^{s} a_{j0} \frac{\partial}{\partial\theta_j}\log f(X,\theta)\right)^2 = I_{11}(\theta) - I_{1\cdot}(\theta)[I_{\cdot\cdot}(\theta)]^{-1} I_{1\cdot}^T(\theta).$$

Or

$$[I(\theta)]_{11}^{-1} = \min_{a_j} E\left(\frac{\partial}{\partial\theta_1}\log f(X,\theta) - \sum_{j=2}^{s} a_j \frac{\partial}{\partial\theta_j}\log f(X,\theta)\right)^2.$$

If $a_{j0} = 0$, $[I(\theta)]_{11}^{-1} = 1/I_{11}(\theta)$. Or, adaptivation is possible.

As illustrations, we will consider three examples. The first example is on the estimation of regression coefficients of a linear regression and the next two examples are on the estimation of parametric component in a semiparametric model. The two particular models we considered are the partial spline model (Wahba, 1984) and the two-sample proportional hazard model (Cox, 1972).

Example 1. (Linear Regression) Assume that $y = \beta_0 + \beta_1 x + \epsilon$ and $\epsilon \sim N(0, \sigma^2)$. Let $(\hat{\beta}_0, \hat{\beta}_1)$ denote the least squares estimate. It follows easily that

$$Var \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \sigma^2 \begin{pmatrix} n & \sum_i X_i \\ \sum_i X_i & \sum_i X_i^2 \end{pmatrix}^{-1} = \sigma^2 \begin{pmatrix} \frac{\sum_i X_i^2}{\sum_i (X_i - \bar{X})^2} & -\frac{\bar{X}}{\sum_i (X_i - \bar{X})^2} \\ -\frac{\bar{X}}{\sum_i (X_i - \bar{X})^2} & \frac{1}{\sum_i (X_i - \bar{X})^2} \end{pmatrix}^{-1}.$$

When β_0 is known, the variance of least squares estimate of β_1 is $\sigma^2 / \sum_i X_i^2$. Then the necessary and sufficient condition on guaranteeing adaptiveness is that $\bar{X} = 0$. When we write the model in matrix form, the condition $\bar{X} = 0$ can be explained as the two vectors $(1, \cdots, 1)^T$ and $(X_1, \cdots, X_n)^T$ are orthogonal. Note that $(1, \cdots, 1)^T$ and $(X_1, \cdots, X_n)^T$ are associated with β_0 and β_1, respectively. On the other hand, we can use the above derivation. Observe that

$$\frac{\partial \log f}{\partial \theta_0} = \frac{Y - \beta_0 - \beta_1 X}{\sigma^2} = \frac{\epsilon}{\sigma^2}, \quad \frac{\partial \log f}{\partial \theta_1} = \frac{(Y - \beta_0 - \beta_1 X) X}{\sigma^2} = \frac{X\epsilon}{\sigma^2}.$$

Then

$$\min_a E \left(\frac{\partial \log f}{\partial \beta_1} - a \frac{\partial \log f}{\partial \beta_0} \right)^2 = \min_a \frac{1}{\sigma^2} E(X - a)^2 = \frac{1}{\sigma^2} Var(X).$$

Example 2. (Partial Spline Model) Assume that $Y = \beta X + g(T) + \epsilon$ and $\epsilon \sim N(0, \sigma^2)$. Suppose that $g \in L_2[a, b]$, the set of all square integrable functions on the interval $[a, b]$. By proper taking care some mathematical subtlity, $g(T)$ can be written as $\sum_{j=1}^{\infty} b_j \phi_j(T)$ where $\{\phi_j\}$ is the complete bases of $L_2[a, b]$. Observe that

$$\frac{\partial \log f}{\partial \theta} = \frac{X\epsilon}{\sigma^2} \quad \frac{\partial \log f}{\partial b_j} = \frac{\epsilon \phi_j(T)}{\sigma^2}.$$

Then

$$\min_{a_j} E \left(\frac{\partial \log f}{\partial \beta} - \sum_j a_j \frac{\partial \log f}{\partial b_j} \right)^2 = \min_{a_j} E \frac{\epsilon^2}{\sigma^4} \left(X - - \sum_j a_j \phi_j(T) \right)^2$$
$$= \frac{1}{\sigma^2} \min_h E(X - h(T))^2.$$

Therefore, $h(T) = E(X|T)$. It means that when $E(X|T) = 0$ or X and T are uncorrelated, the adaption is possible. Otherwise, the efficient bound for estimating β is

$$\frac{\sigma^2}{E(X - E(X|T))^2} = \frac{\sigma^2}{EVar(X|T)}.$$

Refer to Chen (1988) and Speckman (1988) for further references and construction of efficient estimate of β.

Example 3. Let $t_1, \cdots, t_n$ be fixed constants. Suppose that $X_i \sim Bin(1, F(t_i))$ and $Y_i \sim Bin(1, F^{\theta}(t_i))$. We now derive a lower bound on the

asymptotic variance of estimate of θ. Again, we assume that $F(t) = \sum_i a_i\phi_i(t)$. According to the above discussion, it is equivalent to solving the following minimization problem:

$$\min_{b_j} E\left(\frac{\partial L}{\partial \theta} - \sum_{j=1}^{\infty} b_j \frac{\partial L}{\partial a_j}\right)^2,$$

where L is the log-likelihood function.

The likelihood function is

$$\mathcal{L} = \left(\sum_i a_i\phi_i(t)\right)^x \left(1 - \sum_i a_i\phi_i(t)\right)^{1-x} \left[\left(\sum_i a_i\phi_i(t)\right)^\theta\right]^y \left\{1 - \left[\sum_i a_i\phi_i(t)\right]^\theta\right\}^{1-y}.$$

Thus the log likelihood function is

$$\begin{aligned} L = \log\mathcal{L} &= x\log\left(\sum_i a_i\phi_i(t)\right) + (1-x)\log\left(1 - \sum_i a_i\phi_i(t)\right) \\ &\quad +\theta y\log\left[\sum_i a_i\phi_i(t)\right] + (1-y)\log\left\{1 - \left[\sum_i a_i\phi_i(t)\right]^\theta\right\}. \end{aligned}$$

Observe that

$$\begin{aligned} \frac{\partial L}{\partial \theta} &= \log\left(\sum_i a_i\phi_i(t)\right) - (1-y)\log\left(1 - \sum_i a_i\phi_i(t)\right)\cdot\frac{1}{1-\left[\sum_i a_i\phi_i(t)\right]^\theta} \\ &= \log F(t) - (1-y)\log F(t)\cdot\frac{1}{1-F^\theta(t)} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial L}{\partial a_j} &= x\frac{\phi_j(t)}{\sum_i a_i\phi_i(t)} - (1-x)\frac{\phi_j(t)}{1-\sum_i a_i\phi_i(t)} + \theta y\frac{\phi_j(t)}{\sum_i a_i\phi_i(t)} \\ &\quad -(1-y)\frac{\theta\phi_j(t)[\sum_i a_i\phi_i(t)]^{\theta-1}}{1-[\sum_i a_i\phi_i(t)]^\theta} \\ &= \frac{x\phi_j(t)}{F(t)} - \frac{(1-x)\phi_j(t)}{1-F(t)} + \frac{\theta y\phi_j(t)}{F(t)} - \frac{(1-y)\theta F^{\theta-1}(t)\phi_j(t)}{1-F^\theta(t)}. \end{aligned}$$

According to the above discussion, a lower bound can be derived as

$$\min_{b_j} E\left(\frac{\partial L}{\partial \theta} - \sum_{j=1}^{\infty} b_j \frac{\partial L}{\partial a_j}\right)^2.$$

For notational simplicity, set

$$I = \frac{\partial L}{\partial \theta} - \sum_j b_j\frac{\partial L}{\partial a_j} \qquad \text{and} \qquad G(t) = \sum_j b_i\phi_j(t).$$

we then have

$$I = \log F(t) - (1-y)\log F(t)\cdot\frac{1}{1-F^\theta(t)}$$

$$
\begin{aligned}
& -\left[\frac{x}{F(t)}G(t) - \frac{(1-x)}{1-F(t)}G(t) + \frac{\theta y}{F(t)}G(t) - \frac{(1-y)\theta F^{\theta-1}(t)}{1-F^{\theta}(t)}G(t)\right] \\
= \;& \log F(t) - (1-y)\log F(t)\cdot\frac{1}{1-F^{\theta}(t)} \\
& -\left\{-\frac{1}{1-F(t)} + \frac{x}{F(t)[1-F(t)]} + \frac{\theta y}{F(t)[1-F^{\theta}(t)]} - \frac{\theta F^{\theta-1}(t)}{1-F^{\theta}(t)}\right\}G(t) \\
= \;& \frac{[\log F(t)][y-F^{\theta}(t)]}{1-F^{\theta}(t)} - \left\{\frac{x-F(t)}{F(t)[1-F(t)]} + \frac{\theta[y-F^{\theta}(t)]}{F(t)[1-F^{\theta}(t)]}\right\}G(t).
\end{aligned}
$$

Observe that

$$
\begin{aligned}
E(I^2|t) \;=\; & \left(\frac{\log F(t)}{1-F^{\theta}(t)}\right)^2 F^{\theta}(t)(1-F^{\theta}(t)) + G^2(t)\left\{\frac{1}{F(t)[1-F(t)]} + \frac{\theta^2 F^{\theta}(t)(1-F^{\theta}(t))}{F^2(t)[1-F^{\theta}(t)]^2}\right\} \\
& -2G(t)\frac{\theta(\log F(t))F^{\theta}(t)(1-F^{\theta}(t))}{F(t)[1-F^{\theta}(t)]^2} \\
=\; & \frac{F^{\theta}(t)(\log F(t))^2}{1-F^{\theta}(t)} + G^2(t)\left\{\frac{1}{F(t)[1-F(t)]} + \frac{\theta^2 F^{\theta}(t)}{F^2(t)[1-F^{\theta}(t)]}\right\} \\
& -2G(t)\frac{\theta(\log F(t))F^{\theta}(t)}{F(t)[1-F^{\theta}(t)]}.
\end{aligned}
$$

Then

$$
\min_G E(I^2) = \int (\log F(t))^2 \left[\frac{1}{\frac{1-F^{\theta}(t)}{F^{\theta}(t)}} + \theta^2\frac{1-F(t)}{F(t)}\right] dF^*(t),
$$

where F^* is the design measure of $t_1, \cdots, t_n$.

4 Other Methods of Estimation

Consider the following framework that $X \sim P \in \mathcal{P}$, usually $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for parametric models. More specifically, if $X_1, \ldots, X_n \overset{i.i.d.}{\to} P_\theta$, then

$$
\mathbf{P}_\theta = P_\theta \times \cdots \times P_\theta.
$$

Suppose that we are interested in an unknown parameters, $\nu(P)$ or $q(\theta) = \nu(P_\theta)$, which is a certain aspects of population. Recall that the empirical distribution is

$$
\hat{P}[X \in A] = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \quad \text{or} \quad \hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x).
$$

The substitution principle states that $\nu(P)$ can be estimate by $\nu(\hat{P})$. As to be seen later, most methods of estimation can be regarded as using *substitution principle*, since the functional form ν is not unique.

We now demonstrate this principle through a few examples. First, we consider $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Then $\mu = EX = \int x dF(x) = \mu(F)$ and

$\sigma^2 = \int x^2 dF(x) - \mu^2$. Hence,

$$\hat{\mu} = \mu(\hat{F}) = \int x d\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$\hat{\sigma}^2 = \int x^2 d\hat{F}(x) - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

This is also a non-parametric estimator, as the normality assumption has not been explicitly used.

Next, we consider a random sample $\{X_1, \ldots, X_n\}$, from a k-variate multinomial distribution in which there are k categories and the associated probabilities $p_1, \ldots, p_k$. Suppose that we are interested in parameters $p_1, \ldots, p_k$ and $q(p_1, \ldots, p_k)$. The empirical distribution gives

$$p_j = P(X = j) = F(j) - F(j-) = P_j(F).$$

Hence,

$$\hat{p}_j = P_j(\hat{F}) = \hat{F}(j) - \hat{F}(j-) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = j),$$

namely, the empirical frequency of getting j. Hence,

$$q(p_1, \ldots, p_k) = q(P_1(F), \ldots, P_k(F))$$

is estimated as

$$\hat{q} = q(\hat{p}_1, \ldots, \hat{p}_k)$$

which can be viewed as frequency substitution.

As an illustration, we consider the problem of sampling from a equilibrium population with respective to a gene with two alleles

$$\begin{cases} A & \text{with probability } \theta \\ a & \text{with probability } 1-\theta \end{cases},$$

A with prob. three genotypes can be observed with proportion

AA	Aa	aa
$p_1 = \theta^2$	$p_2 = 2\theta(1-\theta)$	$p_3 = (1-\theta)^2$

This is the so-called Hardy-Weinberg formula. In this example, one can estimate θ by $\sqrt{\hat{p}_1}$ or $1 - \sqrt{\hat{p}_3}$, etc. Thus, the representation

$$q(\theta) = h(p_1(\theta), \ldots, p_k(\theta))$$

is not necessarily unique, resulting in many different procedures.

4.1 Generalized method of moment (GMM)

Given that a random sample $X_1, X_2, \ldots, X_n$ are drawn from a population which is characterized by the parameter θ whose true value is θ_0. Let $g_1(X), \ldots, g_r(X)$ be given functions and write

$$\mu_j(\theta_0) = E_{\theta_0}[g_j(X)],$$

which are generalized moments of X. We first note, if the second moment of $g_j(X)$ exists, then law of large numbers implies

$$n^{-1}\sum_{i=1}^{n} g_j(X_i) \xrightarrow{P} E_{\theta_0}[g_j(X)].$$

It means that the population moment condition (11.1) can be approximated by the sample moment condition. Set $\hat{\mu}_j$ to be $n^{-1}\sum_{i=1}^{n} g_j(X_i)$. The GMM estimator $\hat{\theta}$ solves the equations

$$\hat{\mu}_j - \mu_j(\theta) = 0, \quad j = 1, \ldots, r.$$

From now on, we write them as

$$\frac{1}{n}\sum_{i=1}^{n} g_j(x_i; \theta) = 0, \quad j = 1, \ldots, r.$$

Suppose that the consistency of $\hat{\theta}$ is established. The proof of asymptotic normality is based on Taylor expansion and the central limit theorem. Given that $\hat{\theta}$ converges in probability to θ_0 and that g is differentiable with respect to θ, then for sufficiently large n the first-order Taylor expansion of

$$0 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(x_i; \hat{\theta}) \approx \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(x_i; \theta_0) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{g}(x_i; \theta_0)}{\partial \theta}(\hat{\theta} - \theta_0),$$

where $\mathbf{g} = (g_1, \ldots, g_r)$. Or,

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{g}(x_i; \theta_0)}{\partial \theta}\right]^{-1} \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(x_i; \theta_0).$$

around the true value θ_0 gives the following approximation

$$(0, \ldots, 0) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(X_i).$$

Provided that the second moment of $\partial \mathbf{g}(X_i; \theta_0)/\partial \theta$ exists, law of large numbers again implies

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{g}(X_i; \theta_0)}{\partial \theta} \xrightarrow{P} \mathbf{G}(\theta_0),$$

and the central limit theorem implies

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(X_i;\theta_0) - E[\mathbf{g}(X_i;\theta_0)]\right\} \xrightarrow{d} \mathbf{u} \sim N(0,\Omega(\theta_0)),$$

where $E\mathbf{g}(X;\theta_0) = 0$. Consequently,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} -\mathbf{G}(\theta_0)^{-1} \cdot \mathbf{u} \sim N(0, [\mathbf{G}(\theta_0)^{-1}]\Omega(\theta_0)[\mathbf{G}(\theta_0)^{-1}]^T).$$

In proving consistency and asymptotic normality of the GMM estimator, we have used law of large numbers and the central limit theorem. Obviously, certain assumptions are required before we can apply these theorems. We now state a set of regularity conditions that ensure the validity of the GMM estimation.

1. Conditions that ensure the differentiability of $\mathbf{g}(x;\theta)$ with respect to θ. For example, $\mathbf{g}(x;\theta)$ is usually assumed to be twice continuously differentiable with respect to θ.

2. Conditions that restrict the moments of $\mathbf{g}(X;\theta)$ and its derivatives with respect to θ. For example, the second moments of $g(x;\theta)$ and its first derivative are usually assumed to be finite.

3. Conditions that restrict the range of the possible values which the parameter θ can take. For example, θ is not allowed to have infinite value and the true value θ_0 may not be at the boundary of the permissible range of θ (if θ_0 is on the boundary of the permissible range of θ, then convergence to θ_0 cannot take place freely from all directions).

4. The solution to the population moment condition $E_\theta[\mathbf{g}(X;\theta)] = 0$ must be unique and the unique solution must be the true value θ_0 of the parameter.

The first three categories of regularity conditions are somewhat technical and are routinely assumed. They can be replaced by other formulation. However, we do need to make special efforts to check the validity of the last one in each application of GMM. This last condition is referred to as the identification condition because it allows us to identify the true parameter value θ_0 for estimation. An obvious necessary condition for identification is that the number of individual population moment conditions r is cannot be smaller than the dimensionality of the parameter vector θ. Otherwise, the population moment condition will have multiple solutions of which all but one can be the true value so that the resulting GMM estimator does not necessarily converge to the true parameter value. This is the so-called under-identification problem.

The identification condition is implicitly assumed in the above analysis of the GMM estimation. In fact, we have made a stronger assumption that r is equal to the dimensionality of θ. We also assume that the derivative $\mathbf{G}(x;\theta)$ of $\mathbf{g}(x;\theta)$ with respect to θ is a square matrix and invertible. This is the so-called just-identification case. Refer to implicit function theorem to get ideas on the needed minimal assumptions. The GMM Interpretation of Both the least-squares estimation method for regression models and maximum likelihood estimation can be viewed as GMM.

If r is greater than the number of parameters, it is not possible to solve its sample counterpart. Instead, we can find θ that makes the sample moment condition as close to zero as possible based on the following quadratic form:

$$\sum_{j=1}^{r}[\hat{\mu}_j - \mu_j(\theta)]^2$$

(this has a scale problem) or more generally

$$[\hat{\mu} - \mu(\theta)]^T \Sigma^{-1} [\hat{\mu} - \mu(\theta)].$$

Here Σ is some positive definite weighting matrix of constants. Sometimes we can find Σ to optimize the performance of the estimator (GMM).

Consider a regression problem for any random sample $\{(\mathbf{X}_i, Y_i), i = 1, \ldots, , n\}$, define the coefficient of the best linear prediction under the loss function $d(\cdot)$ by

$$\beta(P) = arg\min_{\beta} E_P d(|Y - \beta^T \mathbf{X}|).$$

Thus, its substitution estimator is

$$\beta(\hat{P}) = arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n} d(|Y_i - \beta^T \mathbf{X}_i|).$$

Thus, $\beta(\hat{P})$ is always a consistent estimator of $\beta(P)$, whether the linear $Y = \beta^T \mathbf{X} + \epsilon$ holds or not. In this view, the least-squares estimator is a substitution estimator.

4.2 Minimum Contrast Estimator and Estimating Equations

Let $\rho(\mathbf{X}, \theta)$ be a contrast (discrepancy) function. Define

$$D(\theta_0, \theta) = E_{\theta_0}\rho(\mathbf{X}, \theta),$$

where θ_0 is the true parameter. Suppose that $D(\theta_0, \theta)$ has a unique minimum θ_0. Then, the minimum contrast estimator for a random sample is defined as

the minimizer of

$$\hat{D}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\rho(\mathbf{X}_i, \theta).$$

Under some regularity conditions, the estimator satisfies the estimating equations

$$\hat{D}'(\theta) = \frac{1}{n}\sum_{i=1}^{n}\rho'(\mathbf{X}_i, \theta).$$

In general, the method applies to general situation:

$$\hat{\theta} = arg\min_{\theta}\rho(\mathbf{X}, \theta),$$

as long as θ_0 minimizes

$$D(\theta, \theta_0) = E_{\theta_0}\rho(\mathbf{X}, \theta).$$

Usually, $\rho(\mathbf{X}, \theta) \to D(\theta, \theta_0)$ as $n \to \infty$.

Similarly, estimating equation method solves the equations

$$\phi_j(\mathbf{X}, \theta) = 0, \quad j = 1, \ldots, r$$

as long as

$$E_{\theta_0}\phi_j(\mathbf{X}, \theta_0) = 0, \quad j = 1, \ldots, r.$$

We now use a regression problem to demonstrate that these two approaches are closely related. Let $(\mathbf{X}_i, Y_i)$ be i.i.d. from

$$\begin{aligned} Y_i &= g(\mathbf{X}_i, \beta) + \epsilon_i \quad (\epsilon_i \sim N(0, \sigma^2)) \\ &= \mathbf{X}_i^T\beta + \epsilon_i \quad \text{linear model.} \end{aligned}$$

Then, by letting

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

be a contrast function, we have

$$\begin{aligned} D(\beta_0, \beta) &= E_{\beta_0}\rho(\mathbf{X}, \beta) = nE[\mathbf{Y} - g(\mathbf{X}, \beta)]^2 \\ &= nE\{g(\mathbf{X}, \beta_0) - g(\mathbf{X}, \beta)\}^2 + n\sigma^2, \end{aligned}$$

which is indeed minimized at $\beta = \beta_0$. Hence, the minimum contrast estimator is

$$\hat{\beta} = arg\min_{\beta}\sum_{i=1}^{n}[Y_i - g(\mathbf{X}_i, \beta)]^2.$$

It satisfies the system of equations

$$\sum_{i=1}^{n}[Y_i - g(\mathbf{X}_i, \hat{\beta})]\frac{\partial g(\mathbf{X}_i, \hat{\beta})}{\partial \beta_j} = 0, \quad j = 1, \ldots, d,$$

under some mild regularity conditions. One can easily check that

$$\psi_j(\beta) = [Y_i - g(\mathbf{X}_i, \beta)] \frac{\partial g(\mathbf{X}_i, \beta)}{\partial \beta_j}$$

satisfies

$$E_{\beta_0} \psi_j(\beta)|_{\beta=\beta_0} = E \left\{ [g(\mathbf{X}_i, \beta_0) - g(\mathbf{X}_i, \beta_0)] \frac{\partial g(\mathbf{X}_i, \beta_0)}{\partial \beta_j} \right\} = 0.$$

Thus, it is also an estimator based on the estimating equations.

Now we consider L_1-regression. Let $\mathbf{Y} = \mathbf{X}^T \beta_0 + \epsilon$. Consider

$$\rho(\mathbf{X}, \mathbf{Y}, \beta) = |\mathbf{Y} - \mathbf{X}^T \beta|.$$

Then,

$$D(\beta_0, \beta) = E_{\beta_0} |Y - \mathbf{X}^T \beta| = E|\mathbf{X}^T (\beta - \beta_0) + \epsilon|.$$

For any a, define

$$f(a) = E|\epsilon + a|.$$

Then,

$$\begin{aligned} f^{'}(a) &= E[sgn(\epsilon + a)] = P(\epsilon + a > 0) - P(\epsilon + a < 0) \\ &= 2P(\epsilon + a > 0) - 1. \end{aligned}$$

If $median(\epsilon) = 0$, then $f^{'}(0) = 0$. In other words, $f(a)$ is minimized at $a = 0$, or $D(\beta_0, \beta)$ is minimized at $\beta = \beta_0$! Thus, if $median(\epsilon) = 0$, then

$$\frac{1}{n} \sum_{i=1}^{n} |Y_i - \mathbf{X}_i^T \beta|.$$

is a minimum contrast estimator.

5 The EM algorithm

In order to deal with missing data, Dempster, Laird and Rubin(1977) and Baum, Petrie, Soules, and Weiss(1970) devised general algorithm for such a problem. Suppose that we have a situation in which the full likelihood $\mathbf{X} \sim p(\mathbf{x}, \theta)$ is easy to compute and to maximize. Unfortunately, we only observe the partial information $S = S(\mathbf{X}) \sim q(s, \theta)$. But $q(s, \theta)$ itself is hard to compute and to maximize. The algorithm is to maximize $q(s, \theta)$.

Consider *lumped Hardy-Weinberg data*. If we can observe $\{X_1, \ldots, X_n\}$, MLE leads to

$$\log p(\mathbf{x}, \theta) = n_1 \log \theta^2 + n_2 \log 2\theta(1 - \theta) + n_3 \log(1 - \theta)^2.$$

However, we only observe partial information:

$$\begin{aligned} S_i &= (X_{i1}, X_{i2}, X_{i3}), \quad i = 1, \ldots, m \\ S_i &= (X_{i1} + X_{i2}, X_{i3}), \quad i = m+1, \ldots, n. \end{aligned}$$

The likelihood of the available data is

$$\log q(s,\theta) = m_1 \log \theta^2 + m_2 \log 2\theta(1-\theta) + m_3 \log(1-\theta)^2 + n_{12}^* \log(1-(1-\theta)^2) + n_3^* \log(1-\theta)^2,$$

where $n_{12}^* = \sum_{i=m+1}^{n}(X_{i1} + X_{i2})$ and $n_3^* = \sum_{i=m}^{n} X_{i3}$. For many other problems, this log-likelihood can be hard to compute.

Intuitively, the E-M algorithm attempts to *guess* the full likelihood using the available and maximum the conjectured likelihood.

EM algorithm: Given an initial value θ_0,

E-step: Compute $\ell(\theta, \theta_0) = E_{\theta_0}\left[\ell(\mathbf{x}, \theta) | S(\mathbf{X}) = s\right]$,

M-step: $\hat{\theta} = arg\max \ell(\theta, \theta_0)$,

and iterate.

Rationale: $p(\mathbf{x}, \theta) = q(s, \theta) P_\theta(\mathbf{X} = \mathbf{x} | S = s) I(S(\mathbf{X}) = s)$. Let $r(x|s,\theta) = P_\theta(\mathbf{X} = \mathbf{x}|S = s) I(S(\mathbf{x}) = s)$. Then,

$$\ell(\theta, \theta_0) = \log q(s,\theta) + E_{\theta_0}\{\log r(\mathbf{X}|s,\theta) | S(\mathbf{X}) = s\}.$$

Hence

$$0 = \ell(\hat{\theta}, \theta_0)' = (\log q(s,\theta))'|_{\theta=\hat{\theta}} + E_{\theta_0}\left\{(\log r(\mathbf{X}|s,\theta))'|_{\theta=\hat{\theta}} | S = s\right\}.$$

If the algorithm converges to θ_1, then

$$(\log q(s,\theta))'|_{\theta=\theta_1} + E_{\theta_1}\left\{(\log r(\mathbf{X}|s,\theta))'|_{\theta=\theta_1} \Big| S = s\right\}.$$

Noticing that for any regular function f,

$$E_\theta(\log f(x,\theta))' = \int \frac{f'(x,\theta)}{f(x,\theta)} f(x,\theta) dx = 0.$$

Hence

$$\{\log q(s,\theta_1)\}' = 0,$$

which solves the likelihood equation based on the (partial) data. In other words, the EM algorithm converges to the true likelihood.

Theorem 9

$$\log q(s,\theta_{new}) \geq \log q(s,\theta_{old}),$$

namely, each iteration always increases the likelihood.

Proof. Note that

$$\begin{aligned}\ell(\theta_n, \theta_0) &= \log q(s, \theta_n) + E_{\theta_0}\{\log r(X|s, \theta_n)|S(\mathbf{X}) = s\} \\ &\geq \log q(s, \theta_0) + E_{\theta_0}\{\log r(X|s, \theta_0)|S(\mathbf{X}) = s\}.\end{aligned}$$

This implies

$$\begin{aligned}\log q(s, \theta_n) &\geq \log q(s, \theta_0) + E_{\theta_0}\left\{\log \frac{r(X|s, \theta_0)}{r(X|s, \theta_n)}\middle| S(\mathbf{X}) = s\right\} \\ &\geq \log q(s, \theta_0).\end{aligned}$$

Example. Let $X_1, \ldots, X_{n+4}$ be i.i.d. $N(\mu, 1/2)$. Suppose that we observe $S_1 = X_1, \ldots, S_n = X_n, S_{n+1} = X_{n+1}+2X_{n+2}$, and $S_{n+2} = X_{n+3}+X_{n+4}$. Use the EM algorithm to find the maximum likelihood estimator based on the observed data.

Solution. Note that the full likelihood is

$$\begin{aligned}\log p(\mathbf{X}, \mu) &= -\sum_{i=1}^{n+4}(X_i - \mu)^2 \\ &= -\sum_{i=1}^{n+4} X_i^2 + 2\mu\sum_{i=1}^{n+4} X_i - (n+4)\mu^2.\end{aligned}$$

At the E-step, we compute

$$\begin{aligned}E_{\mu_0}\{\log p(\mathbf{X}, \mu)|\mathbf{S}\} &= a(\mu_0) - 2\mu\left\{\sum_{i=1}^{n} X_i + E_{\mu_0}\{X_{n+1} + X_{n+2}|S_{n+1}\} + S_{n+2}\right\} \\ &\quad -(n+4)\mu^2.\end{aligned}$$

To compute $E_{\mu_0}\{X_{n+1}|S_{n+1}\}$, we note that $2X_{n+1} - X_{n+2}$ is uncorrelated with S_{n+1}. Hence, we have

$$\begin{aligned}E_{\mu_0}\{2X_{n+1} - X_{n+2}|S_{n+1}\} &= \mu_0 \\ E_{\mu_0}\{X_{n+1} + 2X_{n+2}|S_{n+1}\} &= S_{n+1}.\end{aligned}$$

Solving the above two equations gives

$$E_{\mu_0}(X_{n+1}|S_{n+1}) = (S_{n+1} + 2\mu_0)/5, \quad E_{\mu_0}\{X_{n+2}|S_{n+1}\} = (2S_{n+1} - \mu_0)/5.$$

and that

$$E_{\mu_0}\{X_{n+1} + X_{n+2}|S_{n+1}\} = (3S_{n+1} + \mu_0)/5.$$

Hence, the conditional likelihood is given by

$$\ell(\mu, \mu_0) = a(\mu_0) - 2\mu\left\{\sum_{i=1}^{n} X_i + 0.6S_{n+1} + 0.2\mu_0 + S_{n+2}\right\} - (n+4)\mu^2.$$

At the M-step, we maximize $\ell(\mu, \mu_0)$ with respect to μ, resulting in

$$\hat{\mu} = (n+4)^{-1} \left\{ \sum_{i=1}^{n} X_i + 0.6 S_{n+1} + 0.2\mu_0 + S_{n+2} \right\}.$$

The EM algorithm is to iterate the above step. When the algorithm converges, the estimate solves

$$\hat{\mu} = (n+4)^{-1} \left\{ \sum_{i=1}^{n} X_i + 0.6 S_{n+1} + 0.2\hat{\mu} + S_{n+2} \right\}.$$

or

$$\hat{\mu} = (n+3.8)^{-1} \left\{ \sum_{i=1}^{n} X_i + 0.6 S_{n+1} + S_{n+2} \right\}.$$

This is the maximum likelihood estimator for the *missing* data. the observed data.

Now we come back to the lumped Hardy-Weinberg data. The observed data likelihood is

$$\log p(x, \theta) = n_1 \log \theta^2 + n_2 \log 2\theta(1-\theta) + n_3 \log(1-\theta)^2.$$

E-step:

$$\begin{aligned} \ell(\theta, \theta_0) &= E_{\theta_0}(n_1|S) \log \theta^2 + E_{\theta_0}(n_2|S) \log 2\theta(1-\theta) + n_3 \log(1-\theta)^2. \\ E_{\theta_0}(n_1|S) &= m_1 + n_{12}^* \frac{\theta_0^2}{\theta_0^2 + 2\theta_0(1-\theta_0)}, \end{aligned}$$

and

$$E_{\theta_0}(n_2|S) = m_2 + n_{12}^* \frac{2\theta_0(1-\theta_0)}{\theta_0^2 + 2\theta_0(1-\theta_0)}.$$

M-step:

$$\hat{\theta} = \frac{2E_{\theta_0}(n_1|S) + E_{\theta_0}(n_2|S)}{2\left[E_{\theta_0}(n_1|S) + E_{\theta_0}(n_2|S) + n_3\right]} = \frac{n_{12} + E_{\theta_0}(n_1|S)}{2n},$$

where n_{12} is the number of data points for genotypes 1 and 2. When the algorithm converges, it solves the following equation:

$$2n\theta = n_{12} + m_1 + n_{12}^* \theta/(2-\theta).$$

Example. (Mixture normal distribution)
Assume that

$$S_1, \ldots, S_n \overset{i.i.d.}{\sim} \lambda N(\mu_1, \sigma_1^2) + (1-\lambda) N(\mu_2, \sigma_2^2)$$

. The likelihood of $S_1, \ldots, S_n$ is easy to write down, but hard to compute. We turn to EM algorithm to overcome the challenge in computing. Thinking of the

full information as $X_i = (\triangle_i, S_i)$, in which $\triangle_i$ tells the population under which it is drawn from, but missing.

$$\begin{aligned} P(\triangle_i = 1) &= \lambda, \\ P(S_i|\triangle_i) &\sim \begin{cases} N(\mu_1, \sigma_1^2), & \triangle_i = 1 \\ N(\mu_2, \sigma_2^2), & \triangle_i = 0 \end{cases} \end{aligned}$$

Then, the full likelihood is

$$\begin{aligned} p(\mathbf{x}, \theta) &= \lambda^{\sum \triangle_i}(1-\lambda)^{n-\sum \triangle_i} \prod_{\triangle_i=1} \left[\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(S_i-\mu_1)^2}{2\sigma_1^2} \right) \right] \\ &\quad \times \prod_{\triangle_i=0} \left[\frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(S_i-\mu_2)^2}{2\sigma_2^2} \right) \right]. \end{aligned}$$

It follows that

$$\begin{aligned} p(\mathbf{x}, \theta) &= \sum \triangle_i \log\lambda + \sum (1-\triangle_i)\log(1-\lambda) \\ &\quad + \sum_{\triangle_i=1} \left\{ \log\sigma_1 - \frac{(S_i-\mu_1)^2}{2\sigma_1^2} \right\} + \sum_{\triangle_i=0} \left\{ \log\sigma_2 - \frac{(S_i-\mu_2)^2}{2\sigma_2^2} \right\} \end{aligned}$$

To find the E-step, we need to find the conditional distribution of $\triangle_i|\mathbf{S}$. Note that

$$\begin{aligned} P_{\theta_0}\{\triangle_i = 1|\mathbf{S} = \mathbf{s}\} &= P\{\triangle_i = 1|\mathbf{S}_i \in \mathbf{s}_i \pm \epsilon\} \\ &= \frac{P\{\triangle_i = 1, \mathbf{S}_i \in \mathbf{s}_i \pm \epsilon\}}{P(\mathbf{S}_i \in \mathbf{s}_i \pm \epsilon)} \\ &= \frac{\lambda_0 \sigma_{10}^{-1} \phi\left(\frac{s_i-\mu_{10}}{\sigma_{10}}\right)}{\lambda_0 \sigma_{10}^{-1} \phi\left(\frac{s_i-\mu_{10}}{\sigma_{10}}\right) + (1-\lambda_0)\sigma_{20}^{-1}\phi\left(\frac{s_i-\mu_{20}}{\sigma_{20}}\right)} \\ &= p_i \end{aligned}$$

Then,

$$\begin{aligned} \ell(\theta, \theta_0) &= \sum_i p_i \log\lambda + \sum_i (1-p_i)\log(1-\lambda) \\ &\quad + \sum_i p_i \left\{ \log\sigma_1 - \frac{(S_i-\mu_1)^2}{\sigma_1^2} \right\} \\ &\quad + \sum_i (1-p_i) \cdots. \end{aligned}$$

The M-step is to maximize the above quantity with respective to, $\lambda, \sigma_1, \mu_1, \sigma_2, \mu_2$, which can be explicitly found. e.g.

$$\begin{aligned} \frac{\sum_{i=1}^n p_i}{\lambda} - \frac{\sum_{i=1}^n (1-p_i)}{1-\lambda} = 0 \quad &\rightarrow \quad \hat{\lambda} = \frac{\sum_{i=1}^n p_i}{n} \\ \sum_{i=1}^n p_i(s_i - \mu) = 0 \quad &\rightarrow \quad \hat{\mu} = \frac{\sum_{i=1}^n p_i s_i}{\sum_{i=1}^n p_i}. \end{aligned}$$

The EM algorithm is to iterate these two steps.

References

[1] Beran, R. (1974). Asymptotically efficient adaptive rank estimates in location models. *Ann. Statist.* **2** 63-74.

[2] Chen, H. (1988). Convergence rates for the parametric component in a partly linear model. *Ann. statist.* **16** 136-146.

[3] Cox, D.R. (1972). Regression models and life tables. *J. Roy. statist. Soc. Ser. B* **34** 187-202.

[4] Fergson, T.S. (1996). Chapters 17-20 of the book entitled *A Course in Large Sample Theory.*

[5] Le Cam, L. (1953). On some asymptotic properties of Maximum Likelihood estimates and related Bayes's estimates, *University of California Publications in Statistics* **1**: 277-330.

[6] Lehmann, E.L. (1983). *Theory of Point Estimation.* John Wiley & Sons, New York.

[7] Neyman, J. (1949). Contributions to the theory of χ^2 test. *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, edited by J. Neyman, pp. 230-273. Berkeley, Univ. of California Press.)

[8] Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1-32.

[9] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

[10] Simar, L. (1983). Protecting against gross errors: the aid of Bayesian methods. In *Specifying Statistical Models: From Parametric to Non-Parametric Using Bayesian or Non-Bayesian Approaches.* Edited by J.R. Florens, M. Mouchart, J.P. Raoult, L. Simar, and A.F.M. Smith. Lecture Notes in Statistics: Vol. 16. Springer-Verlag, New York.

[11] Speckman, P. (1988). Kernel smoothing in partial linear models. *J. roy. Statist. Soc. Ser. B* **50** 413-436.

[12] Stein, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 187-195.

[13] Stone, C.J. (1974). Asymptotic properties of estimators of a location parameter. *Ann. statist.* **2** 1127-1137.

[14] Wahba, G. (1984). Cross validated spline methods for the estimation of multivariate functions from data on functionals. In *Statistics: An Appraisal, Proceedings 50th Anniversary Conference Iowa State Statistical Laboratory* (H.A. David and H.T. David, eds) 205-235. Iowa State Univ. Press, Ames, IA.