# Chapter 5

# Estimation in Nonparametric Models

Estimation methods studied in this chapter are useful for nonparametric models as well as for parametric models in which the parametric model assumptions might be violated (so that robust estimators are required) or the number of unknown parameters is exceptionally large. Some such methods have been introduced in Chapter 3; for example, the methods that produce UMVUE's in nonparametric models, the U- and V-statistics, the LSE's and BLUE's, the Horvitz-Thompson estimators, and the sample (central) moments.

The theoretical justification for estimators in nonparametric models, however, relies more on asymptotics than that in parametric models. This means that applications of nonparametric methods usually require large sample sizes. Also, estimators derived using parametric methods are asymptotically more efficient than those based on nonparametric methods when the parametric models are correct. Thus, to choose between a parametric method and a nonparametric method, we need to balance the advantage of requiring weaker model assumptions (robustness) against the drawback of losing efficiency which results in requiring a larger sample size.

It is assumed in this chapter that a sample $X = (X_1, ..., X_n)$ is from a population in a nonparametric family, where $X_i$'s are random vectors.

## 5.1 Distribution Estimators

In many applications the c.d.f.'s of $X_i$'s are determined by a single c.d.f. $F$ on $\mathcal{R}^d$; for example, $X_i$'s are i.i.d. random $d$-vectors. In this section we

consider the estimation of $F$ or $F(t)$ for several $t$'s, under a nonparametric model in which very little is assumed about $F$.

### 5.1.1 Empirical c.d.f.'s in i.i.d. cases

For i.i.d. random variables $X_1, ..., X_n$, the empirical c.d.f. $F_n$ is defined in (2.31). The definition of the empirical c.d.f. in the case of $X_i \in \mathcal{R}^d$ is analogous. For $t \in \mathcal{R}^d$, let $A(t)$ be the set of all $s \in \mathcal{R}^d$ such that all components of $t - s$ are nonnegative. Then the empirical c.d.f. based on $X$ is defined by

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{A(t)}(X_i), \qquad t \in \mathcal{R}^d. \tag{5.1}$$

Similar to the case of $d = 1$ (Example 2.26), $F_n(t)$ as an estimator of $F(t)$ has the following properties. For any $t \in \mathcal{R}^d$, $nF_n(t)$ has the binomial distribution $Bi(F(t), n)$; $F_n(t)$ is unbiased with variance $F(t)[1 - F(t)]/n$; $F_n(t)$ is the UMVUE under some nonparametric models; and $F_n(t)$ is $\sqrt{n}$-consistent for $F(t)$. For any $m$ fixed distinct points $t_1, ..., t_m$ in $\mathcal{R}^d$, it follows from the multivariate CLT (Corollary 1.2) and (5.1) that as $n \to \infty$,

$$\sqrt{n}\big[\big(F_n(t_1), ..., F_n(t_m)\big) - \big(F(t_1), ..., F(t_m)\big)\big] \to_d N_m(0, \Sigma), \tag{5.2}$$

where $\Sigma$ is the $m \times m$ matrix whose $(i, j)$th element is

$$P\big(X_1 \in A(t_i) \cap A(t_j)\big) - F(t_i)F(t_j).$$

Note that these results hold without *any* assumption on $F$.

Considered as a function of $t$, $F_n$ is a random element taking values in $\mathcal{F}$, the collection of all c.d.f.'s on $\mathcal{R}^d$. As $n \to \infty$, $\sqrt{n}(F_n - F)$ converges in some sense to a random element defined on some probability space. A detailed discussion of such a result is beyond our scope, and can be found, for example, in Shorack and Wellner (1986). To discuss some global properties of $F_n$ as an estimator of $F \in \mathcal{F}$, we need to define a closeness measure between the elements (c.d.f.'s) in $\mathcal{F}$.

**Definition 5.1.** Let $\mathcal{F}_0$ be a subset of $\mathcal{F}$. A function $\varrho$ from $\mathcal{F}_0 \times \mathcal{F}_0$ to $[0, \infty)$ is called a *distance* or *metric* on $\mathcal{F}_0$ if and only if for any $G_j$ in $\mathcal{F}_0$,
(a) $\varrho(G_1, G_2) = 0$ if and only if $G_1 = G_2$;
(b) $\varrho(G_1, G_2) = \varrho(G_2, G_1)$;
(c) $\varrho(G_1, G_2) \le \varrho(G_1, G_3) + \varrho(G_3, G_2)$. ∎

The most commonly used distance is the sup-norm distance $\varrho_\infty$ defined on $\mathcal{F}$:

$$\varrho_\infty(G_1, G_2) = \sup_{t \in \mathcal{R}^d} |G_1(t) - G_2(t)|, \qquad G_j \in \mathcal{F}. \tag{5.3}$$

The following result concerning the sup-norm distance between $F_n$ and $F$ is due to Dvoretzky, Kiefer, and Wolfowitz (1956).

**Lemma 5.1.** (DKW's inequality). Let $F_n$ be the empirical c.d.f. based on i.i.d. $X_1, ..., X_n$ from a c.d.f. $F$ on $\mathcal{R}^d$.
(i) When $d = 1$, there exists a positive constant $C$ (not depending on $F$) such that

$$P\big(\varrho_\infty(F_n, F) > z\big) \le Ce^{-2nz^2}, \quad z > 0,\ n = 1, 2, ....$$

(ii) When $d \ge 2$, for any $\epsilon > 0$, there exists a positive constant $C_{\epsilon,d}$ (not depending on $F$) such that

$$P\big(\varrho_\infty(F_n, F) > z\big) \le C_{\epsilon,d}e^{-(2-\epsilon)nz^2}, \quad z > 0,\ n = 1, 2, .... \quad \blacksquare$$

The proof of this lemma is omitted. The following results useful in statistics are direct consequences of Lemma 5.1.

**Theorem 5.1.** Let $F_n$ be the empirical c.d.f. based on i.i.d. $X_1, ..., X_n$ from a c.d.f. $F$ on $\mathcal{R}^d$. Then
(i) $\varrho_\infty(F_n, F) \to_{a.s.} 0$ as $n \to \infty$;
(ii) $E[\sqrt{n}\varrho_\infty(F_n, F)]^s = O(1)$ for any $s > 0$.
**Proof.** (i) From DKW's inequality,

$$\sum_{n=1}^{\infty} P\big(\varrho_\infty(F_n, F) > z\big) < \infty.$$

Hence, the result follows from Theorem 1.8(v).
(ii) Using DKW's inequality with $z = y^{1/s}/\sqrt{n}$ and the result in Exercise 45 of §1.6, we obtain that

$$\begin{aligned} E[\sqrt{n}\varrho_\infty(F_n, F)]^s &= \int_0^\infty P\big(\sqrt{n}\varrho_\infty(F_n, F) > y^{1/s}\big)dy \\ &\le C_{\epsilon,d}\int_0^\infty e^{-(2-\epsilon)y^{2/s}}dy \\ &= O(1) \end{aligned}$$

as long as $2 - \epsilon > 0$. $\blacksquare$

Theorem 5.1(i) means that $F_n(t) \to_{a.s.} F(t)$ uniformly in $t \in \mathcal{R}^d$, a result stronger than the strong consistency of $F_n(t)$ for every $t$. Theorem 5.1(ii) implies that $\sqrt{n}\varrho_\infty(F_n, F) = O_p(1)$, a result stronger than the $\sqrt{n}$-consistency of $F_n(t)$. Again, these results hold without any condition on the unknown $F$.

Let $p \geq 1$ and $\mathcal{F}_p = \{G \in \mathcal{F} : \int \|t\|^p dG < \infty\}$, which is the subset of c.d.f.'s in $\mathcal{F}$ having finite $p$th moments. Mallows' distance between $G_1$ and $G_2$ in $\mathcal{F}_p$ is defined to be

$$\varrho_{M_p}(G_1, G_2) = \inf(E\|Y_1 - Y_2\|^p)^{1/p}, \tag{5.4}$$

where the infimum is taken over all pairs of $Y_1$ and $Y_2$ having c.d.f.'s $G_1$ and $G_2$, respectively. Let $\{G_j : j = 0, 1, 2, ...\} \subset \mathcal{F}_p$. Then $\varrho_{M_p}(G_j, G_0) \to 0$ as $j \to \infty$ if and only if $\int \|t\|^p dG_j \to \int \|t\|^p dG_0$ and $G_j(t) \to G_0(t)$ for every $t \in \mathcal{R}^d$ at which $G_0$ is continuous. It follows from Theorem 5.1 and the SLLN (Theorem 1.13) that $\varrho_{M_p}(F_n, F) \to_{a.s.} 0$ if $F \in \mathcal{F}_p$.

When $d = 1$, another useful distance for measuring the closeness between $F_n$ and $F$ is the $L_p$ distance ($p \geq 1$):

$$\varrho_{L_p}(G_1, G_2) = \left[\int |G_1(t) - G_2(t)|^p dt\right]^{1/p}, \qquad G_j \in \mathcal{F}_1. \tag{5.5}$$

A result similar to Theorem 5.1 is given as follows.

**Theorem 5.2.** Let $F_n$ be the empirical c.d.f. based on i.i.d. random variables $X_1, ..., X_n$ from a c.d.f. $F \in \mathcal{F}_1$. Then
(i) $\varrho_{L_p}(F_n, F) \to_{a.s.} 0$;
(ii) $E[\sqrt{n}\varrho_{L_p}(F_n, F)] = O(1)$ if $1 \leq p < 2$ and $\int \{F(t)[1 - F(t)]\}^{p/2} dt < \infty$, or $p \geq 2$.

**Proof.** (i) Since $[\varrho_{L_p}(F_n, F)]^p \leq [\varrho_\infty(F_n, F)]^{p-1}[\varrho_{L_1}(F_n, F)]$ and, by Theorem 5.1, $\varrho_\infty(F_n, F) \to_{a.s.} 0$, it suffices to show the result for $p = 1$. Let $Y_i = \int_{-\infty}^0 [I_{(-\infty,t]}(X_i) - F(t)]dt$. Then $Y_1, ..., Y_n$ are i.i.d. and

$$E|Y_i| \leq \int E|I_{(-\infty,t]}(X_i) - F(t)|dt = 2\int F(t)[1 - F(t)]dt,$$

which is finite under the condition that $F \in \mathcal{F}_1$. By the SLLN,

$$\int_{-\infty}^0 [F_n(t) - F(t)]dt = \frac{1}{n}\sum_{i=1}^n Y_i \to_{a.s.} E(Y_1) = 0. \tag{5.6}$$

Since $[F_n(t) - F(t)]_- \leq F(t)$ and $\int_{-\infty}^0 F(t)dt < \infty$ (Exercise 45 in §1.6), it follows from Theorem 5.1 and the dominated convergence theorem that

$$\int_{-\infty}^0 [F_n(t) - F(t)]_- dt \to_{a.s.} 0,$$

which with (5.6) implies

$$\int_{-\infty}^0 |F_n(t) - F(t)|dt \to_{a.s.} 0. \tag{5.7}$$

The result follows since we can similarly show that (5.7) holds with $\int_{-\infty}^{0}$ replaced by $\int_0^\infty$.
(ii) When $1 \le p < 2$, the result follows from

$$\begin{aligned} E[\varrho_{L_p}(F_n, F)] &\le \left\{ \int E|F_n(t) - F(t)|^p dt \right\}^{1/p} \\ &\le \left\{ \int [E|F_n(t) - F(t)|^2]^{p/2} dt \right\}^{1/p} \\ &= n^{-1/2} \left\{ \int \{F(t)[1 - F(t)]\}^{p/2} dt \right\}^{1/p} \\ &= O(n^{-1/2}), \end{aligned}$$

where the two inequalities follow from Jensen's inequality. When $p \ge 2$,

$$\begin{aligned} E[\varrho_{L_p}(F_n, F)] &\le E\left\{ [\varrho_\infty(F_n, F)]^{1-2/p} [\varrho_{L_2}(F_n, F)]^{2/p} \right\} \\ &\le \left\{ E[\varrho_\infty(F_n, F)]^{(1-2/p)q} \right\}^{1/q} \left\{ E[\varrho_{L_2}(F_n, F)]^2 \right\}^{1/p} \\ &= \left\{ O(n^{-(1-2/p)q/2}) \right\}^{1/q} \left\{ E \int |F_n(t) - F(t)|^2 dt \right\}^{1/p} \\ &= O(n^{-(1-2/p)/2}) \left\{ \frac{1}{n} \int F(t)[1 - F(t)] dt \right\}^{1/p} \\ &= O(n^{-1/2}), \end{aligned}$$

where $\frac{1}{q} + \frac{1}{p} = 1$; the second inequality follows from Hölder's inequality (see, e.g., Serfling (1980, p. 352)); and the first equality follows from Theorem 5.1(ii). ∎

### 5.1.2 Empirical likelihoods

In §4.4 and §4.5, we have shown that the method of using likelihoods provides some asymptotically efficient estimators. We now introduce some likelihoods in nonparametric models. This not only provides another justification for the use of the empirical c.d.f. in (5.1), but also leads to a useful method of deriving estimators in various (possibly non-i.i.d.) cases, some of which are discussed later in this chapter.

Let $X_1, ..., X_n$ be i.i.d. with $F \in \mathcal{F}$ and $P_G$ be the probability measure corresponding to $G \in \mathcal{F}$. Given $X_1 = x_1, ..., X_n = x_n$, the *nonparametric likelihood* function is defined to be the following functional from $\mathcal{F}$ to $[0, \infty)$:

$$\ell(G) = \prod_{i=1}^{n} P_G(\{x_i\}), \qquad G \in \mathcal{F}. \tag{5.8}$$

Apparently, $\ell(G) = 0$ if $P_G(\{x_i\}) = 0$ for at least one $i$. The following result, due to Kiefer and Wolfowitz (1956), shows that the empirical c.d.f. $F_n$ is a nonparametric maximum likelihood estimator of $F$.

**Theorem 5.3.** Let $X_1, ..., X_n$ be i.i.d. with $F \in \mathcal{F}$ and $\ell(G)$ be defined by (5.8). Then $F_n$ maximizes $\ell(G)$ over $G \in \mathcal{F}$.

**Proof.** We only need to consider $G \in \mathcal{F}$ such that $\ell(G) > 0$. Let $c \in (0, 1]$ and $\mathcal{F}(c)$ be the subset of $\mathcal{F}$ containing $G$'s satisfying $p_i = P_G(\{x_i\}) > 0$, $i = 1, ..., n$, and $\sum_{i=1}^n p_i = c$. We now apply the Lagrange multiplier method to solve the problem of maximizing $\ell(G)$ over $G \in \mathcal{F}(c)$. Define

$$H(p_1, ..., p_n, \lambda) = \prod_{i=1}^n p_i + \lambda \left( \sum_{i=1}^n p_i - c \right),$$

where $\lambda$ is the Lagrange multiplier. Set

$$\frac{\partial H}{\partial \lambda} = \sum_{i=1}^n p_i - c = 0, \qquad \frac{\partial H}{\partial p_j} = p_j^{-1} \prod_{i=1}^n p_i + \lambda = 0, \qquad j = 1, ..., n.$$

The solution is $p_i = c/n$, $i = 1, ..., n$, $\lambda = -(c/n)^{n-1}$. It can be shown (exercise) that this solution is a maximum of $H(p_1, ..., p_n, \lambda)$ over $p_i > 0$, $i = 1, ..., n$, $\sum_{i=1}^n p_i = c$. This shows that

$$\max_{G \in \mathcal{F}(c)} \ell(G) = (c/n)^n,$$

which is maximized at $c = 1$ for any fixed $n$. The result follows from $P_{F_n}(\{x_i\}) = n^{-1}$ for given $X_i = x_i$, $i = 1, ..., n$. ∎

From the proof of Theorem 5.3, $F_n$ maximizes the likelihood $\ell(G)$ in (5.8) over $p_i > 0$, $i = 1, ..., n$, and $\sum_{i=1}^n p_i = 1$, where $p_i = P_G(\{x_i\})$. This method of deriving an estimator of $F$ can be extended to various situations with some modifications of (5.8) and/or constraints on $p_i$'s. Modifications of the likelihood in (5.8) are called *empirical likelihoods* (Owen, 1988, 1990; Qin and Lawless, 1994). An estimator obtained by maximizing an empirical likelihood is then called a *maximum empirical likelihood estimator* (MELE). We now discuss several applications of the method of empirical likelihoods.

Consider first the estimation of $F$ with auxiliary information about $F$ (and i.i.d. $X_1, ..., X_n$). For instance, suppose that there is a known function $u$ from $\mathcal{R}^d$ to $\mathcal{R}^s$ such that

$$\int u(x) dF = 0 \tag{5.9}$$

(e.g., some components of the mean of $F$ are 0). It is then reasonable to expect that any estimate $\hat{F}$ of $F$ has property (5.9), i.e., $\int u(x) d\hat{F} = 0$,

which is not true for the empirical c.d.f. $F_n$ in (5.1), since

$$\int u(x)dF_n = \frac{1}{n}\sum_{i=1}^{n} u(X_i) \neq 0$$

even if $E[u(X_1)] = 0$. Using the method of empirical likelihood, a natural solution is to put another constraint in the process of maximizing the likelihood. That is, we maximize $\ell(G)$ in (5.8) subject to

$$p_i > 0, \quad i = 1, ..., n, \quad \sum_{i=1}^{n} p_i = 1, \quad \text{and} \quad \sum_{i=1}^{n} p_i u(x_i) = 0, \tag{5.10}$$

where $p_i = P_G(\{x_i\})$. Using the Lagrange multiplier method and a similar argument to the proof of Theorem 5.3, it can be shown (exercise) that an MELE of $F$ is

$$\hat{F}(t) = \sum_{i=1}^{n} \hat{p}_i I_{A(t)}(X_i), \tag{5.11}$$

where

$$\hat{p}_i = \{n[1 + u(X_i)\lambda_n^\tau]\}^{-1}, \qquad i = 1, ..., n, \tag{5.12}$$

and $\lambda_n \in \mathcal{R}^s$ is the Lagrange multiplier satisfying

$$\sum_{i=1}^{n} \hat{p}_i u(X_i) = \sum_{i=1}^{n} \frac{u(X_i)}{n[1 + u(X_i)\lambda_n^\tau]} = 0. \tag{5.13}$$

Note that $\hat{F}$ reduces to $F_n$ if $u \equiv 0$.

To see that (5.13) has a solution asymptotically, note that

$$\frac{\partial}{\partial\lambda}\left[\frac{1}{n}\sum_{i=1}^{n}\log\big(1 + u(X_i)\lambda^\tau\big)\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{u(X_i)}{1 + u(X_i)\lambda^\tau}$$

and

$$\frac{\partial^2}{\partial\lambda\partial\lambda^\tau}\left[\frac{1}{n}\sum_{i=1}^{n}\log\big(1 + u(X_i)\lambda^\tau\big)\right] = -\frac{1}{n}\sum_{i=1}^{n}\frac{[u(X_i)]^\tau u(X_i)}{[1 + u(X_i)\lambda^\tau]^2},$$

which is negative definite if $\mathrm{Var}(u(X_1))$ is positive definite. Also,

$$E\left\{\frac{\partial}{\partial\lambda}\left[\frac{1}{n}\sum_{i=1}^{n}\log\big(1 + u(X_i)\lambda^\tau\big)\right]\Big|_{\lambda=0}\right\} = E[u(X_1)] = 0.$$

Hence, using the same argument as in the proof of Theorem 4.18, we can show that there exists a unique sequence $\{\lambda_n(X)\}$ such that as $n \to \infty$,

$$P\left(\sum_{i=1}^{n}\frac{u(X_i)}{n[1 + u(X_i)\lambda_n^\tau]} = 0\right) \to 1 \qquad \text{and} \qquad \lambda_n \to_p 0. \tag{5.14}$$

**Theorem 5.4.** Let $X_1, ..., X_n$ be i.i.d. with $F \in \mathcal{F}$, $u$ be a function on $\mathcal{R}^d$ satisfying (5.9), and $\hat{F}$ be given by (5.11)-(5.13). Suppose that $U = \mathrm{Var}(u(X_1))$ is positive definite. Then, for any $m$ fixed distinct $t_1, ..., t_m$ in $\mathcal{R}^d$,

$$\sqrt{n}[\big(\hat{F}(t_1), ..., \hat{F}(t_m)\big) - \big(F(t_1), ..., F(t_m)\big)] \to_d N_m(0, \Sigma_u), \tag{5.15}$$

where

$$\Sigma_u = \Sigma - W^\tau U^{-1} W,$$

$W = \big([W(t_1)]^\tau, ..., [W(t_m)]^\tau\big)$, $W(t_j) = E[u(X_1) I_{A(t_j)}(X_1)]$ ($A(t)$ is given in (5.1)), and $\Sigma$ is given in (5.2).

**Proof.** We prove the case of $m = 1$. The case of $m \geq 2$ is left as an exercise. Let $\bar{u} = n^{-1} \sum_{i=1}^n u(X_i)$. It follows from (5.13), (5.14), and Taylor's expansion that

$$\bar{u}^\tau = \frac{1}{n} \sum_{i=1}^n [u(X_i)]^\tau u(X_i) \lambda_n^\tau + o_p(\|\lambda_n\|).$$

By the SLLN and CLT,

$$U^{-1} \bar{u}^\tau = \lambda_n^\tau + o_p(n^{-1/2}).$$

Using Taylor's expansion and the SLLN again, we have

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n I_{A(t)}(X_i)(\hat{p}_i - 1) &= \frac{1}{n} \sum_{i=1}^n I_{A(t)}(X_i) \left[ \frac{1}{1 + u(X_i)\lambda_n^\tau} - 1 \right] \\
&= -\frac{1}{n} \sum_{i=1}^n I_{A(t)}(X_i) u(X_i) \lambda_n^\tau + o_p(n^{-1/2}) \\
&= -W(t) \lambda_n^\tau + o_p(n^{-1/2}) \\
&= -W(t) U^{-1} \bar{u}^\tau + o_p(n^{-1/2}).
\end{aligned}$$

Thus,

$$\begin{aligned}
\hat{F}(t) - F(t) &= F_n(t) - F(t) + \frac{1}{n} \sum_{i=1}^n I_{A(t)}(X_i)(\hat{p}_i - 1) \\
&= F_n(t) - F(t) - W(t) U^{-1} \bar{u}^\tau + o_p(n^{-1/2}) \\
&= \frac{1}{n} \sum_{i=1}^n \left\{ I_{A(t)}(X_i) - F(t) - W(t) U^{-1} [u(X_i)]^\tau \right\} + o_p(n^{-1/2}).
\end{aligned}$$

The result follows from the CLT and the fact that

$$\begin{aligned}
\mathrm{Var}\big(W(t) U^{-1} [u(X_i)]^\tau\big) &= W(t) U^{-1} U U^{-1} [W(t)]^\tau \\
&= W(t) U^{-1} [W(t)]^\tau \\
&= E\{W(t) U^{-1} [u(X_i)]^\tau I_{A(t)}(X_i)\} \\
&= \mathrm{Cov}\big(I_{A(t)}(X_i), W(t) U^{-1} [u(X_i)]^\tau\big). \quad \blacksquare
\end{aligned}$$

Comparing (5.15) with (5.2), we conclude that $\hat{F}$ is asymptotically more efficient than $F_n$.

**Example 5.1** (Survey problems). An example of situations in which we have auxiliary information expressed as (5.9) is a survey problem (Example 2.3) where the population $\mathcal{P} = \{y_1, ..., y_N\}$ consists of two-dimensional $y_j$'s, $y_j = (y_{1j}, y_{2j})$, and the population mean $\bar{Y}_2 = N^{-1}\sum_{j=1}^N y_{2j}$ is known. For example, suppose that $y_{1j}$ is the current year's income of unit $j$ in the population and $y_{2j}$ is the last year's income. In many applications the population total or mean of $y_{2j}$'s is known, for example, from tax return records. Let $X_1, ..., X_n$ be a simple random sample (see Example 2.3) selected from $\mathcal{P}$ *with* replacement. Then $X_i$'s are i.i.d. bivariate random vectors whose c.d.f. is

$$F(t) = \frac{1}{N}\sum_{j=1}^N I_{A(t)}(y_j). \tag{5.16}$$

If $\bar{Y}_2$ is known, then it can be expressed as (5.9) with $u(x_1, x_2) = x_2 - \bar{Y}_2$. In survey problems $X_i$'s are usually sampled *without* replacement so that $X_1, ..., X_n$ are not i.i.d. However, for a simple random sample without replacement, (5.8) can still be treated as an empirical likelihood, given $X_i$'s. Note that $F$ in (5.16) is the c.d.f. of $X_i$, regardless of whether $X_i$'s are sampled with replacement.

If $X = (X_1, ..., X_n)$ is not a simple random sample, then the likelihood (5.8) has to be modified. Suppose that $\pi_i$ is the probability that the $i$th unit is selected (see Theorem 3.15). Given $X = \{y_i, i \in \boldsymbol{s}\}$, an empirical likelihood is

$$\ell(G) = \prod_{i\in\boldsymbol{s}}[P_G(\{y_i\})]^{1/\pi_i} = \prod_{i\in\boldsymbol{s}} p_i^{1/\pi_i}, \tag{5.17}$$

where $p_i = P_G(\{y_i\})$. With the auxiliary information (5.9), an MELE of $F$ in (5.16) can be obtained by maximizing $\ell(G)$ in (5.17) subject to (5.10). In this case $F$ may not be the c.d.f. of $X_i$, but the c.d.f.'s of $X_i$'s are determined by $F$ and $\pi_i$'s. It can be shown (exercise) that an MELE is given by (5.11) with

$$\hat{p}_i = \frac{1}{\pi_i[1 + u(y_i)\lambda_n^\tau]}\Big/\sum_{i\in\boldsymbol{s}}\frac{1}{\pi_i} \tag{5.18}$$

and

$$\sum_{i\in\boldsymbol{s}}\frac{u(y_i)}{\pi_i[1 + u(y_i)\lambda_n^\tau]} = 0. \tag{5.19}$$

If $\pi_i$ = a constant, then the MELE reduces to that in (5.11)-(5.13). If

$u(x) = 0$ (no auxiliary information), then the MELE is

$$\hat{F}(t) = \sum_{i \in \boldsymbol{s}} \frac{1}{\pi_i} I_{A(t)}(y_i) \Big/ \sum_{i \in \boldsymbol{s}} \frac{1}{\pi_i},$$

which is a ratio of two Horvitz-Thompson estimators (§3.4.2). Some asymptotic properties of the MELE $\hat{F}$ can be found in Chen and Qin (1993). ∎

The second part of Example 5.1 shows how to use empirical likelihoods in a non-i.i.d. problem. Applications of empirical likelihoods in non-i.i.d. problems are usually straightforward extensions of those in i.i.d. cases. The following is another example.

**Example 5.2** (Biased sampling). Biased sampling is often used in applications. Suppose that $n = n_1 + \cdots + n_k$, $k \geq 2$; $X_i$'s are independent random variables; $X_1, ..., X_{n_1}$ are i.i.d. with $F$; and $X_{n_j+1}, ..., X_{n_{j+1}}$ are i.i.d. with the c.d.f.

$$\int_{-\infty}^{t} w_{j+1}(s) dF(s) \Big/ \int_{-\infty}^{\infty} w_{j+1}(s) dF(s),$$

$j = 1, ..., k-1$, where $w_j$'s are some nonnegative functions. A simple example is that $X_1, ..., X_{n_1}$ are sampled from $F$ and $X_{n_1+1}, ..., X_{n_2}$ are sampled from $F$ but conditional on the fact that each sampled value exceeds a given value $x_0$ (i.e., $w_2(s) = I_{(x_0,\infty)}(s)$). For instance, $X_i$'s are blood pressure measurements; $X_1, ..., X_{n_1}$ are sampled from ordinary people and $X_{n_1+1}, ..., X_{n_2}$ are sampled from patients whose blood pressures are higher than $x_0$. The name biased sampling comes from the fact that there is a bias in the selection of samples.

For simplicity we consider the case of $k = 2$, since the extension to $k \geq 3$ is straightforward. Denote $w_2$ by $w$. An empirical likelihood is

$$\begin{aligned} \ell(G) &= \prod_{i=1}^{n_1} P_G(\{x_i\}) \prod_{i=n_1+1}^{n} \frac{w(x_i) P_G(\{x_i\})}{\int w(s) dG(s)} \\ &= \left[ \sum_{i=1}^{n} p_i w(x_i) \right]^{-n_2} \prod_{i=1}^{n} p_i \prod_{i=n_1+1}^{n} w(x_i), \end{aligned} \tag{5.20}$$

where $p_i = P_G(\{x_i\})$. An MELE of $F$ can be obtained by maximizing the empirical likelihood (5.20) subject to $p_i > 0$, $i = 1, ..., n$, and $\sum_{i=1}^{n} p_i = 1$. Using the Lagrange multiplier method we can show (exercise) that an MELE $\hat{F}$ is given by (5.11) with

$$\hat{p}_i = [n_1 + n_2 w(X_i)/\hat{w}]^{-1}, \qquad i = 1, ..., n, \tag{5.21}$$

where $\hat{w}$ satisfies

$$\hat{w} = \sum_{i=1}^{n} \frac{w(X_i)}{n_1 + n_2 w(X_i)/\hat{w}}.$$

An asymptotic result similar to that in Theorem 5.4 can be established (Vardi, 1985; Qin, 1993). ∎

Our last example concerns an important application in survival analysis.

**Example 5.3** (Censored data). Let $Z_1, ..., Z_n$ be *survival times* that are i.i.d. nonnegative random variables from a c.d.f. $F$, and $Y_1, ..., Y_n$ be i.i.d. nonnegative random variables independent of $Z_i$'s. In a variety of applications in biostatistics and life-time testing, we are only able to observe the smaller of $Z_i$ and $Y_i$ and an indicator of which variables is smaller:

$$X_i = \min(Z_i, Y_i), \qquad \delta_i = I_{(0,Y_i)}(Z_i), \qquad i = 1, ..., n.$$

This is called a *random censorship model* and $Y_i$'s are called *censoring times*. We consider the estimation of the survival distribution $F$; see Kalbfleisch and Prentice (1980) for other problems involving censored data.

An MELE of $F$ can be derived as follows. Let $x_{(1)} \le \cdots \le x_{(n)}$ be ordered values of $X_i$'s and $\delta_{(i)}$ be the $\delta$-value associated with $x_{(i)}$. Consider a c.d.f. $G$ that assigns its mass to the points $x_{(1)}, ..., x_{(n)}$ and the interval $(x_{(n)}, \infty)$. Let $p_i = P_G(\{x_{(i)}\})$, $i = 1, ..., n$, and $p_{n+1} = 1 - G(x_{(n)})$. An MELE of $F$ is then obtained by maximizing

$$\ell(G) = \prod_{i=1}^{n} p_i^{\delta_{(i)}} \left( \sum_{j=i+1}^{n+1} p_j \right)^{1-\delta_{(i)}} \tag{5.22}$$

subject to

$$p_i > 0, \quad i = 1, ..., n+1, \qquad \sum_{i=1}^{n+1} p_i = 1. \tag{5.23}$$

It can be shown (exercise) that an MELE is

$$\hat{F}(t) = \sum_{i=1}^{n+1} \hat{p}_i I_{(X_{(i-1)}, X_{(i)})}(t), \tag{5.24}$$

where $X_{(0)} = 0$, $X_{(n+1)} = \infty$, $X_{(1)} \le \cdots \le X_{(n)}$ are order statistics, and

$$\hat{p}_1 = \tfrac{1}{n}, \quad \hat{p}_i = \tfrac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(1 - \tfrac{\delta_{(j)}}{n-j+1}\right), \quad i = 2, ..., n, \quad \hat{p}_{n+1} = 1 - \sum_{j=1}^{n} \hat{p}_j.$$

The $\hat{F}$ in (5.24) can also be written as (exercise)

$$\hat{F}(t) = 1 - \prod_{X_{(i)} \le t} \left(1 - \frac{\delta_{(i)}}{n-i+1}\right), \tag{5.25}$$

which is the well-known Kaplan-Meier (1958) *product-limit* estimator. Some asymptotic results for $\hat{F}$ in (5.25) can be found, for example, in Shorack and Wellner (1986). ∎

### 5.1.3 Density estimation

Suppose that $X_1, ..., X_n$ are i.i.d. random variables from $F$ and that $F$ is unknown but has a Lebesgue p.d.f. $f$. Estimation of $F$ can be done by estimating $f$, which is called *density estimation*. Note that estimators of $F$ derived in §5.1.1 and §5.1.2 do not have Lebesgue p.d.f.'s.

Since $f(t) = F'(t)$ a.e., a simple estimator of $f(t)$ is the difference quotient

$$f_n(t) = \frac{F_n(t+\lambda_n) - F_n(t-\lambda_n)}{2\lambda_n}, \qquad t \in \mathcal{R}, \tag{5.26}$$

where $F_n$ is the empirical c.d.f. given by (2.31) or (5.1) with $d = 1$, and $\{\lambda_n\}$ is a sequence of positive constants. Since $2n\lambda_n f_n(t)$ has the binomial distribution $Bi(F(t+\lambda_n) - F(t-\lambda_n), n)$,

$$E[f_n(t)] \to f(t) \qquad \text{if } \lambda_n \to 0 \text{ as } n \to \infty$$

and

$$\mathrm{Var}\big(f_n(t)\big) \to 0 \qquad \text{if } \lambda_n \to 0 \text{ and } n\lambda_n \to \infty.$$

Thus, we should choose $\lambda_n$ converging to 0 slower than $n^{-1}$. If we assume that $\lambda_n \to 0$, $n\lambda_n \to \infty$, and $f$ is continuously differentiable at $t$, then it can be shown (exercise) that

$$\mathrm{mse}_{f_n(t)}(F) = \frac{f(t)}{2n\lambda_n} + o\left(\frac{1}{n\lambda_n}\right) + O(\lambda_n^2) \tag{5.27}$$

and, under the additional condition that $n\lambda_n^3 \to 0$,

$$\sqrt{n\lambda_n}[f_n(t) - f(t)] \to_d N\big(0, \tfrac{1}{2}f(t)\big). \tag{5.28}$$

A useful class of estimators is the class of *kernel density estimators* of the form

$$\hat{f}(t) = \frac{1}{n\lambda_n} \sum_{i=1}^{n} w\left(\tfrac{t-X_i}{\lambda_n}\right), \tag{5.29}$$

where $w$ is a known Lebesgue p.d.f. on $\mathcal{R}$ and is called the kernel. If we choose $w(t) = \frac{1}{2}I_{[-1,1]}(t)$, then $\hat{f}(t)$ in (5.29) is essentially the same as the so-called histogram. The bias of $\hat{f}(t)$ in (5.29) is

$$\begin{aligned} E[\hat{f}(t)] - f(t) &= \frac{1}{\lambda_n}\int w\left(\tfrac{t-z}{\lambda_n}\right) f(z)dz - f(t) \\ &= \int w(y)[f(t-\lambda_n y) - f(t)]dy. \end{aligned}$$

If $f$ is bounded and is continuous at $t$, then, by the dominated convergence theorem (Theorem 1.1), the bias of $\hat{f}(t)$ converges to 0 as $\lambda_n \to 0$; if $f'$ is bounded and is continuous at $t$ and $\int |t|w(t)dt < \infty$, then the bias of $\hat{f}(t)$ is $O(\lambda_n)$. The variance of $f(t)$ is

$$\begin{aligned} \mathrm{Var}(\hat{f}(t)) &= \frac{1}{n\lambda_n^2}\mathrm{Var}\left(w\left(\tfrac{t-X_1}{\lambda_n}\right)\right) \\ &= \frac{1}{n\lambda_n^2}\int \left[w\left(\tfrac{t-z}{\lambda_n}\right)\right]^2 f(z)dz \\ &\quad - \frac{1}{n}\left[\frac{1}{\lambda_n}\int w\left(\tfrac{t-z}{\lambda_n}\right)f(z)dz\right]^2 \\ &= \frac{1}{n\lambda_n}\int [w(y)]^2 f(t-\lambda_n y)dy + O\left(\frac{1}{n}\right) \\ &= \frac{w_0 f(t)}{n\lambda_n} + o\left(\frac{1}{n\lambda_n}\right) \end{aligned}$$

if $f$ is bounded and is continuous at $t$ and $w_0 = \int [w(t)]^2 dt < \infty$. Hence, if $\lambda_n \to 0$ and $n\lambda_n \to \infty$ and if $f'$ is bounded and is continuous at $t$, then

$$\mathrm{mse}_{\hat{f}(t)}(F) = \frac{w_0 f(t)}{n\lambda_n} + O(\lambda_n^2).$$

Using the CLT (Theorem 1.15), one can show (exercise) that if $\lambda_n \to 0$, $n\lambda_n \to \infty$, $f$ is bounded and is continuous at $t$, and $\int [w(t)]^{2+\delta}dt < \infty$ for some $\delta > 0$, then

$$\sqrt{n\lambda_n}\{\hat{f}(t) - E[\hat{f}(t)]\} \to_d N\big(0, w_0 f(t)\big). \tag{5.30}$$

Furthermore, if $f'$ is bounded and is continuous at $t$ and $n\lambda_n^3 \to 0$, then

$$\sqrt{n\lambda_n}\{E[\hat{f}(t)] - f(t)\} = O\left(\sqrt{n\lambda_n}\lambda_n\right) \to 0$$

and, therefore, (5.30) holds with $E[\hat{f}(t)]$ replaced by $f(t)$.

Similar to the estimation of a c.d.f., we can also study global properties of $f_n$ or $\hat{f}$ as an estimator of the density curve $f$, using a suitably defined

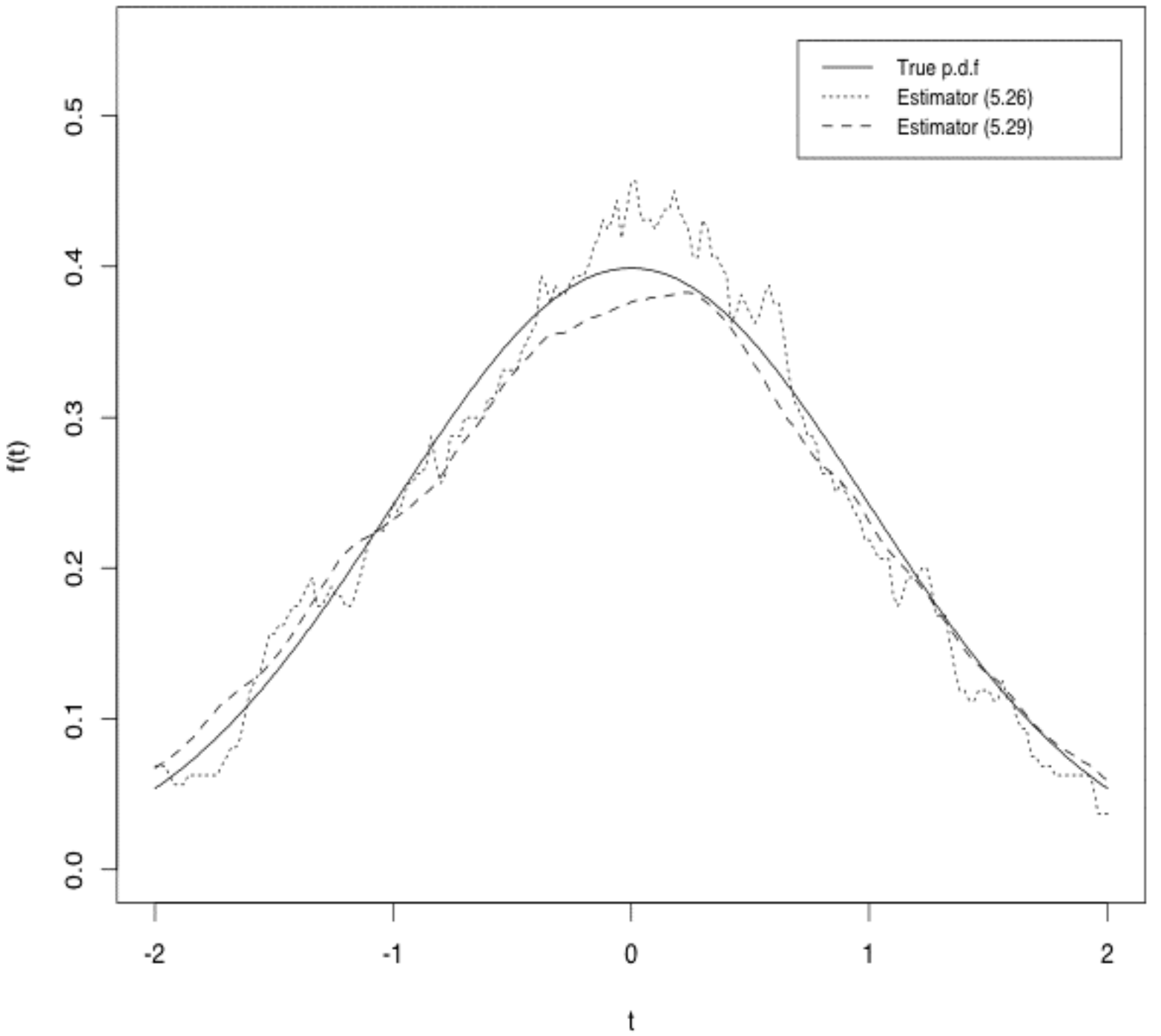


Figure 5.1: Density estimates in Example 5.4

distance between $f$ and its density estimator. For example, we may study the convergence of $\sup_{t\in\mathcal{R}} |\hat{f}(t) - f(t)|$ or $\int |\hat{f}(t) - f(t)|^2 dt$. More details can be found, for example, in Silverman (1986).

**Example 5.4.** An i.i.d. sample of size $n = 200$ was generated from $N(0, 1)$. Density curve estimates (5.26) and (5.29) are plotted in Figure 5.1 with the curve of the true p.d.f. For the kernel density estimator (5.29), $w(t) = \frac{1}{2}e^{-|t|}$ is used and $\lambda_n = 0.4$. From Figure 5.1, it seems that the kernel estimate (5.29) is much better than the estimate (5.26). ∎

There are many other density estimation methods, for example, the nearest neighbor method (Stone, 1977), the smoothing splines (Wahba, 1990), and the method of empirical likelihoods described in §5.1.2 (see, e.g., Jones (1991)), which produces estimators of the form

$$\hat{f}(t) = \frac{1}{\lambda_n} \sum_{i=1}^{n} \hat{p}_i w\left(\tfrac{t-X_i}{\lambda_n}\right).$$

## 5.2 Statistical Functionals

In many nonparametric problems we are interested in estimating some characteristics (parameters) of the unknown population, not the entire population. We assume in this section that $X_i$'s are i.i.d. from an unknown c.d.f. $F$ on $\mathcal{R}^d$. Most characteristics of $F$ can be written as $\mathtt{T}(F)$, where $\mathtt{T}$ is a functional from $\mathcal{F}$ to $\mathcal{R}^s$. If we estimate $F$ by the empirical c.d.f. $F_n$ in (5.1), then a natural estimator of $\mathtt{T}(F)$ is $\mathtt{T}(F_n)$, which is called a *statistical functional*.

Many commonly used statistics can be written as $\mathtt{T}(F_n)$ for some $\mathtt{T}$. Two simple examples are given as follows. Let $\mathtt{T}(F) = \int \psi(x)dF(x)$ with an integrable function $\psi$, and $\mathtt{T}(F_n) = \int \psi(x)dF_n(x) = n^{-1}\sum_{i=1}^n \psi(X_i)$. The sample moments discussed in §3.5.2 are particular examples of this kind of statistical functionals. For $d = 1$, let $\mathtt{T}(F) = F^{-1}(p) = \inf\{x : F(x) \geq p\}$, where $p \in (0, 1)$ is a fixed constant. $F^{-1}(p)$ is called the $p$th *quantile* of $F$. The statistical functional $\mathtt{T}(F_n) = F_n^{-1}(p)$ is called the $p$th *sample quantile*. More examples of statistical functionals are provided in §5.2.1 and §5.2.2.

In this section we study asymptotic distributions of $\mathtt{T}(F_n)$. We focus on the case of real-valued $\mathtt{T}$ ($s = 1$), since the extension to the case of $s \geq 2$ is straightforward.

### 5.2.1 Differentiability and asymptotic normality

Note that $\mathtt{T}(F_n)$ is a function of the "statistic" $F_n$. In Theorem 1.12 (and §3.5.1) we have studied how to use Taylor's expansion to establish asymptotic normality of differentiable functions of statistics that are asymptotically normal. This leads to the approach of establishing asymptotic normality of $\mathtt{T}(F_n)$ by using some generalized Taylor expansions for functionals and using asymptotic properties of $F_n$ given in §5.1.1.

First, we need a suitably defined differential of $\mathtt{T}$. Several versions of differentials are given in the following definition.

**Definition 5.2.** Let $\mathtt{T}$ be a functional on $\mathcal{F}_0$, a collection of c.d.f.'s on $\mathcal{R}^d$, and let $\mathcal{D} = \{c(G_1 - G_2) : c \in \mathcal{R},\ G_j \in \mathcal{F}_0,\ j = 1, 2\}$.
(i) A functional $\mathtt{T}$ on $\mathcal{F}_0$ is Gâteaux differentiable at $G \in \mathcal{F}_0$ if and only if there is a linear functional $\mathtt{L}_G$ on $\mathcal{D}$ (i.e., $\mathtt{L}_G(c_1\Delta_1 + c_2\Delta_2) = c_1\mathtt{L}_G(\Delta_1) + c_2\mathtt{L}_G(\Delta_2)$ for any $\Delta_j \in \mathcal{D}$ and $c_j \in \mathcal{R}$) such that $\Delta \in \mathcal{D}$ and $G + t\Delta \in \mathcal{F}_0$ imply

$$\lim_{t\to 0}\left[\frac{\mathtt{T}(G + t\Delta) - \mathtt{T}(G)}{t} - \mathtt{L}_G(\Delta)\right] = 0.$$

(ii) Let $\varrho$ be a distance on $\mathcal{F}_0$. Suppose that $\|c(G_1 - G_2)\| = |c|\varrho(G_1, G_2)$, $c \in \mathcal{R}$, $G_j \in \mathcal{F}_0$, defines a norm on $\mathcal{D}$ (i.e., $\|\Delta\| \geq 0$ and $= 0$ if and only

if $\Delta = 0$, $\|c\Delta\| = |c|\|\Delta\|$, and $\|\Delta + \tilde{\Delta}\| \le \|\Delta\| + \|\tilde{\Delta}\|$, $\Delta \in \mathcal{D}$, $\tilde{\Delta} \in \mathcal{D}$, $c \in \mathcal{R}$). A functional $\mathtt{T}$ on $\mathcal{F}_0$ is $\varrho$-Hadamard differentiable at $G \in \mathcal{F}_0$ if and only if there is a linear functional $\mathtt{L}_G$ on $\mathcal{D}$ such that for any sequence of numbers $t_j \to 0$ and $\{\Delta, \Delta_j, j = 1, 2, ...\} \subset \mathcal{D}$ satisfying $\|\Delta_j - \Delta\| \to 0$ and $G + t_j\Delta_j \in \mathcal{F}_0$,

$$\lim_{j\to\infty} \left[\frac{\mathtt{T}(G + t_j\Delta_j) - \mathtt{T}(G)}{t_j} - \mathtt{L}_G(\Delta_j)\right] = 0.$$

(iii) Let $\varrho$ be a distance on $\mathcal{F}_0$. A functional $\mathtt{T}$ on $\mathcal{F}_0$ is $\varrho$-Fréchet differentiable at $G \in \mathcal{F}_0$ if and only if there is a linear functional $\mathtt{L}_G$ on $\mathcal{D}$ such that for any sequence $\{G_j\}$ satisfying $G_j \in \mathcal{F}_0$ and $\varrho(G_j, G) \to 0$,

$$\lim_{j\to\infty} \frac{\mathtt{T}(G_j) - \mathtt{T}(G) - \mathtt{L}_G(G_j - G)}{\varrho(G_j, G)} = 0. \quad \blacksquare$$

The functional $\mathtt{L}_G$ is called the *differential* of $\mathtt{T}$ at $G$. If we define $h(t) = \mathtt{T}(G + t\Delta)$, then the Gâteaux differentiability is equivalent to the differentiability of the function $h(t)$ at $t = 0$, and $\mathtt{L}_G(\Delta)$ is simply $h'(0)$. Let $\delta_x$ denote the $d$-dimensional c.d.f. degenerated at the point $x$ and $\phi_G(x) = \mathtt{L}_G(\delta_x - G)$. Then $\phi_F(x)$ is called the *influence function* of $\mathtt{T}$ at $F$, which is an important tool in robust statistics (see Hampel (1974)).

If $\mathtt{T}$ is Gâteaux differentiable at $F$, then we have the following expansion (taking $t = n^{-1/2}$ and $\Delta = \sqrt{n}(F_n - F)$):

$$\sqrt{n}[\mathtt{T}(F_n) - \mathtt{T}(F)] = \mathtt{L}_F\big(\sqrt{n}(F_n - F)\big) + R_n. \tag{5.31}$$

Since $\mathtt{L}_F$ is linear,

$$\mathtt{L}_F\big(\sqrt{n}(F_n - F)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi_F(X_i) \to_d N(0, \sigma_F^2) \tag{5.32}$$

by the CLT, provided that

$$E[\phi_F(X_1)] = 0 \qquad \text{and} \qquad \sigma_F^2 = E[\phi_F(X_1)]^2 < \infty \tag{5.33}$$

(which is usually true when $\phi_F$ is bounded or when $F$ has some finite moments). By Slutsky's theorem and (5.32),

$$\sqrt{n}[\mathtt{T}(F_n) - \mathtt{T}(F)] \to_d N(0, \sigma_F^2) \tag{5.34}$$

if $R_n$ in (5.31) is $o_p(1)$.

Unfortunately, Gâteaux differentiability is too weak to be useful in establishing $R_n = o_p(1)$ (or (5.34)). This is why we need other types of

differentiability. Hadamard differentiability, which is also referred to as compact differentiability, is clearly stronger than Gâteaux differentiability but weaker than Fréchet differentiability (exercise). For a given functional $\mathtt{T}$, we can first find $\mathtt{L}_G$ by differentiating $h(t) = \mathtt{T}(G + t\Delta)$ at $t = 0$ and then check whether $\mathtt{T}$ is $\varrho$-Hadamard (or $\varrho$-Fréchet) differentiable with a given $\varrho$. The most commonly used distances on $\mathcal{F}_0$ are the sup-norm distance $\varrho_\infty$ in (5.3) and the $L_p$ distance $\varrho_{L_p}$ in (5.5). Their corresponding norms on $\mathcal{D}$ are $\|\Delta\|_\infty = \sup_x |\Delta(x)|$ and $\|\Delta\|_{L_p} = [\int |\Delta(x)|^p dx]^{1/p}$, $\Delta \in \mathcal{D}$.

**Theorem 5.5.** Let $X_1, ..., X_n$ be i.i.d. from a c.d.f. $F$ on $\mathcal{R}^d$.
(i) If $\mathtt{T}$ is $\varrho_\infty$-Hadamard differentiable at $F$, then $R_n$ in (5.31) is $o_p(1)$.
(ii) If $\mathtt{T}$ is $\varrho$-Fréchet differentiable at $F$ with a distance $\varrho$ satisfying

$$\sqrt{n}\varrho(F_n, F) = O_p(1), \tag{5.35}$$

then $R_n$ in (5.31) is $o_p(1)$.
(iii) In either (i) or (ii), if (5.33) is also satisfied, then (5.34) holds.
**Proof.** Part (iii) follows directly from (i) or (ii). The proof of (i) involves some high-level mathematics and is omitted; see, for example, Fernholz (1983). We now prove (ii). From Definition 5.2(iii), for any $\epsilon > 0$, there is a $\delta > 0$ such that $|R_n| < \epsilon\sqrt{n}\varrho(F_n, F)$ whenever $\varrho(F_n, F) < \delta$. Then

$$P\left(|R_n| > \eta\right) \le P\left(\sqrt{n}\varrho(F_n, F) > \eta/\epsilon\right) + P\left(\varrho(F_n, F) \ge \delta\right)$$

for any $\eta > 0$, which implies

$$\limsup_n P\left(|R_n| > \eta\right) \le \limsup_n P\left(\sqrt{n}\varrho(F_n, F) > \eta/\epsilon\right).$$

The result follows from (5.35) and the fact that $\epsilon$ can be arbitrarily small. ∎

Since $\varrho$-Fréchet differentiability implies $\varrho$-Hadamard differentiability, Theorem 5.5(ii) is useful when $\varrho$ is not the sup-norm distance. There are functionals that are not $\varrho_\infty$-Hadamard differentiable (and hence not $\varrho_\infty$-Fréchet differentiable). For example, if $d = 1$ and $\mathtt{T}(G) = g(\int x dG)$ with a differentiable function $g$, then $T$ is not necessarily $\varrho_\infty$-Hadamard differentiable, but is $\varrho_{L_1}$-Fréchet differentiable (exercise).

From Theorem 5.2, condition (5.35) holds for $\varrho_{L_p}$ under the moment conditions on $F$ given in Theorem 5.2.

Note that if $\varrho$ and $\tilde{\varrho}$ are two distances on $\mathcal{F}_0$ satisfying $\tilde{\varrho}(G_1, G_2) \le c\varrho(G_1, G_2)$ for a constant $c$ and all $G_j \in \mathcal{F}_0$, then $\tilde{\varrho}$-Hadamard (Fréchet) differentiability implies $\varrho$-Hadamard (Fréchet) differentiability. This suggests the use of the distance $\varrho_{\infty+p} = \varrho_\infty + \varrho_{L_p}$, which also satisfies (5.35) under the moment conditions in Theorem 5.2. The distance $\varrho_{\infty+p}$ is useful in some cases (Theorem 5.6).

A $\varrho_\infty$-Hadamard differentiable $\mathtt{T}$ having a bounded and continuous influence function $\phi_F$ is *robust* in Hampel's sense (see, e.g., Huber (1981)). This is motivated by the fact that the asymptotic behavior of $\mathtt{T}(F_n)$ is determined by that of $\mathtt{L}_F(F_n - F)$, and a small change in the sample, i.e., small changes in all $x_i$'s (rounding, grouping) or large changes in a few of $x_i$'s (gross errors, blunders), will result in a small change of $\mathtt{T}(F_n)$ if and only if $\phi_F$ is bounded and continuous.

We now consider some examples. For the sample moments related to functionals of the form $\mathtt{T}(G) = \int \psi(x)dG(x)$, it is clear that $\mathtt{T}$ is a linear functional. Any linear functional is trivially $\varrho$-Fréchet differentiable for any $\varrho$. Next, if $F$ is one-dimensional and $F'(x) > 0$ for all $x$, then the quantile functional $\mathtt{T}(G) = G^{-1}(p)$ is $\varrho_\infty$-Hadamard differentiable at $F$ (Fernholz, 1983). Hence, Theorem 5.5 applies to these functionals. But the asymptotic normality of sample quantiles can be established under weaker conditions, which are studied in §5.3.1.

**Example 5.5** (Convolution functionals). Suppose that $F$ is on $\mathcal{R}$ and for a fixed $z \in \mathcal{R}$,

$$\mathtt{T}(G) = \int G(z-y)dG(y), \qquad G \in \mathcal{F}.$$

If $X_1$ and $X_2$ are i.i.d. with c.d.f. $G$, then $\mathtt{T}(G)$ is the c.d.f. of $X_1 + X_2$ (Exercise 42 in §1.6), and is also called the convolution of $G$ evaluated at $z$. For $t_j \to 0$ and $\|\Delta_j - \Delta\|_\infty \to 0$,

$$\mathtt{T}(G + t_j\Delta_j) - \mathtt{T}(G) = 2t_j \int \Delta_j(z-y)dG(y) + t_j^2 \int \Delta_j(z-y)d\Delta_j(y)$$

(for $\Delta = c_1G_1 + c_2G_2$, $G_j \in \mathcal{F}_0$, and $c_j \in \mathcal{R}$, $d\Delta$ denotes $c_1dG_1 + c_2dG_2$). Using Lemma 5.2, one can show (exercise) that

$$\int \Delta_j(z-y)d\Delta_j(y) = O(1). \tag{5.36}$$

Hence $\mathtt{T}$ is $\varrho_\infty$-Hadamard differentiable at any $G \in \mathcal{F}$ with $\mathtt{L}_G(\Delta) = 2\int \Delta(z-y)dG(y)$. The influence function, $\phi_F(x) = 2\int(\delta_x - F)(z-y)dF(y)$, is a bounded function and clearly satisfies (5.33). Thus, (5.34) holds. If $F$ is continuous, then $\mathtt{T}$ is robust in Hampel's sense (exercise). ∎

Three important classes of statistical functionals, i.e., L-estimators, M-estimators, and rank statistics and R-estimators, are considered in §5.2.2.

**Lemma 5.2.** Let $\Delta \in \mathcal{D}$ and $h$ be a function on $\mathcal{R}$ such that $\int h(x)d\Delta(x)$ is finite. Then

$$\left| \int h(x)d\Delta(x) \right| \le \|h\|_V \|\Delta\|_\infty,$$

where $\|h\|_V$ is the variation norm defined by

$$\|h\|_V = \lim_{a\to-\infty, b\to\infty} \left[\sup \sum_{j=1}^{m} |h(x_j) - h(x_{j-1})|\right]$$

with the supremum being taken over all partitions $a = x_0 < \cdots < x_m = b$ of the interval $[a, b]$. ∎

The proof of Lemma 5.2 can be found, for example, in Natanson (1961, p. 232).

The differentials in Definition 5.2 are first-order differentials. For some functionals, we can also consider their second-order differentials.

**Definition 5.3.** Let $\mathtt{T}$ be a functional on $\mathcal{F}_0$ and $\varrho$ be a distance on $\mathcal{F}_0$.
(i) $\mathtt{T}$ is second-order $\varrho$-Hadamard differentiable at $G \in \mathcal{F}_0$ if and only if there is a functional $\mathtt{Q}_G$ on $\mathcal{D}$ such that for any sequence of numbers $t_j \to 0$ and $\{\Delta, \Delta_j, j = 1, 2, ...\} \subset \mathcal{D}$ satisfying $\|\Delta_j - \Delta\| \to 0$ and $G + t_j\Delta_j \in \mathcal{F}_0$,

$$\lim_{j\to\infty} \frac{\mathtt{T}(G + t_j\Delta_j) - \mathtt{T}(G) - \mathtt{Q}_G(t_j\Delta_j)}{t_j^2} = 0,$$

where $\mathtt{Q}_G(\Delta) = \int\int \psi_G(x, y)d(G + t_j\Delta)(x)d(G + t_j\Delta)(y)$ for a function $\psi_G$ satisfying $\psi_G(x, y) = \psi_G(y, x)$, $\int\int \psi_G(x, y)dG(x)dG(y) = 0$, and $\|\cdot\|$ is the same as that in Definition 5.2(ii).
(ii) $\mathtt{T}$ is second-order $\varrho$-Fréchet differentiable at $G \in \mathcal{F}_0$ if and only if, for any sequence $\{G_j\}$ satisfying $G_j \in \mathcal{F}_0$ and $\varrho(G_j, G) \to 0$,

$$\lim_{j\to\infty} \frac{T(G_j) - T(G) - \mathtt{Q}_G(G_j - G)}{[\varrho(G_j, G)]^2} = 0,$$

where $\mathtt{Q}_G$ is the same as that in (i). ∎

For a second-order differentiable $\mathtt{T}$, we have the following expansion:

$$n[\mathtt{T}(F_n) - \mathtt{T}(F)] = nV_n + R_n, \tag{5.37}$$

where

$$V_n = \mathtt{Q}_G(F_n - F) = \int\int \psi_F(x, y)dF_n(x)dF_n(y) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} \psi_F(X_i, X_j)$$

is a "V-statistic" (§3.5.3) whose asymptotic properties are given by Theorem 3.16. If $R_n$ in (5.37) is $o_p(1)$, then the asymptotic behavior of $\mathtt{T}(F_n) - \mathtt{T}(F)$ is the same as that of $V_n$.

**Proposition 5.1.** Let $X_1, ..., X_n$ be i.i.d. from $F$.
(i) If $\mathtt{T}$ is second-order $\varrho_\infty$-Hadamard differentiable at $F$, then $R_n$ in (5.37) is $o_p(1)$.
(ii) If $\mathtt{T}$ is second-order $\varrho$-Fréchet differentiable at $F$ with a distance $\varrho$ satisfying (5.35), then $R_n$ in (5.37) is $o_p(1)$. ∎

Combining Proposition 5.1 with Theorem 3.16, we can summarize the asymptotic behavior of $\mathtt{T}(F_n) - \mathtt{T}(F)$ as follows. If

$$\zeta_1 = \mathrm{Var}\left(\int \psi_F(X_1, y)dF(y)\right)$$

is positive, then (5.34) holds with $\sigma_F^2 = 4\zeta_1$. If $\zeta_1 = 0$, then

$$n[\mathtt{T}(F_n) - \mathtt{T}(F)] \to_d \sum_{j=1}^{\infty} \lambda_j \chi_{1j}^2.$$

If $\mathtt{T}$ is also first-order differentiable, then it can be shown (exercise) that

$$\phi_F(x) = 2\int \psi_F(x, y)dF(y). \tag{5.38}$$

Then $\zeta_1 = 4^{-1}\mathrm{Var}(\phi_F(X_1))$ and $\zeta_1 = 0$ corresponds to the case of $\phi_F(x) \equiv 0$. However, second-order $\varrho$-Hadamard (Fréchet) differentiability does not in general imply first-order $\varrho$-Hadamard (Fréchet) differentiability (exercise).

The technique in this section can be applied to non-i.i.d. $X_i$'s when the c.d.f.'s of $X_i$'s are determined by an unknown c.d.f. $F$, provided that results similar to (5.32) and (5.35) (with $F_n$ replaced by some other estimator $\hat{F}$) can be established.

### 5.2.2 L-, M-, R-estimators and rank statistics

Three large classes of statistical functionals based on i.i.d. $X_i$'s are studied in this section.

#### L-estimators

Let $F$ be a c.d.f. on $\mathcal{R}$ and $J(t)$ be a function on $[0, 1]$. An *L-functional* is defined as

$$\mathtt{T}(G) = \int xJ(G(x))dG(x), \qquad G \in \mathcal{F}. \tag{5.39}$$

$\mathtt{T}(F_n)$ is called an L-estimator of $\mathtt{T}(F)$.

**Example 5.6.** The following are some examples of commonly used L-estimators.
(i) When $J \equiv 1$, $\mathtt{T}(F_n) = \bar{X}$, the sample mean.
(ii) When $J(t) = 4t - 2$, $\mathtt{T}(F_n)$ is called Gini's mean difference.
(iii) When $J(t) = (\beta - \alpha)^{-1} I_{(\alpha,\beta)}(t)$ for some constants $\alpha < \beta$, $\mathtt{T}(F_n)$ is called the trimmed sample mean. ∎

For an L-functional $\mathtt{T}$, it can be shown (exercise) that

$$\mathtt{T}(G) - \mathtt{T}(F) = \int \phi_F(x) d(G - F)(x) + R(G, F), \tag{5.40}$$

where

$$\phi_F(x) = -\int (\delta_x - F)(y) J(F(y)) dy, \tag{5.41}$$

$$R(G, F) = -\int W_G(x)[G(x) - F(x)]dx,$$

and

$$W_G(x) = \begin{cases} [G(x) - F(x)]^{-1} \int_{F(x)}^{G(x)} J(t)dt - J(F(x)) & G(x) \neq F(x) \\ 0 & G(x) = F(x). \end{cases}$$

A sufficient condition for (5.33) in this case is that $F$ has a finite variance (exercise). However, (5.33) is also satisfied if $\phi_F$ is bounded. The differentiability of $\mathtt{T}$ can be verified under some conditions on $J$.

**Theorem 5.6.** Let $\mathtt{T}$ be an L-functional defined by (5.39).
(i) Suppose that $J$ is bounded, $J(t) = 0$ when $t \in [0, \alpha] \cup [\beta, 1]$ for some constants $\alpha < \beta$, and that the set $D = \{x : J \text{ is discontinuous at } F(x)\}$ has Lebesgue measure 0. Then $\mathtt{T}$ is $\varrho_\infty$-Fréchet differentiable at $F$ with the influence function $\phi_F$ given by (5.41), and $\phi_F$ is bounded and continuous and satisfies (5.33).
(ii) Suppose that $J$ is bounded, the set $D$ in (i) has Lebesgue measure 0, and $J$ is continuous on $[0, \alpha] \cup [\beta, 1]$ for some constants $\alpha < \beta$. Then $\mathtt{T}$ is $\varrho_{\infty+1}$-Fréchet differentiable at $F$.
(iii) Suppose that $|J(t) - J(s)| \leq C|t - s|^{p-1}$, where $C > 0$ and $p > 1$ are some constants. Then $\mathtt{T}$ is $\varrho_{L_p}$-Fréchet differentiable at $F$.
(iv) If, in addition to the conditions in part (i), $J'$ is continuous on $[\alpha, \beta]$, then $\mathtt{T}$ is second-order $\varrho_\infty$-Fréchet differentiable at $F$ with

$$\psi_F(x, y) = \phi_F(x) + \phi_F(y) - \int (\delta_x - F)(z)(\delta_y - F)(z) J'(F(z)) dz.$$

(v) Suppose that $J'$ is continuous on $[0, 1]$. Then $\mathtt{T}$ is second-order $\varrho_{L_2}$-Fréchet differentiable at $F$ with the same $\psi_F$ given in (iv).

**Proof.** We prove (i)-(iii). The proofs for (iv) and (v) are similar and are left to the reader.
(i) Let $G_j \in \mathcal{F}$ and $\varrho_\infty(G_j, F) \to 0$. Let $c$ and $d$ be two constants such that $F(c) > \beta$ and $F(d) < \alpha$. Then, for sufficiently large $j$, $G_j(x) \in [0, \alpha] \cup [\beta, 1]$ if $x > c$ or $x < d$. Hence, for sufficiently large $j$,

$$\begin{aligned} |R(G_j, F)| &= \left| \int_d^c W_{G_j}(x)(G_j - F)(x)dx \right| \\ &\le \varrho_\infty(G_j, F) \int_d^c |W_{G_j}(x)|dx. \end{aligned}$$

Since $J$ is continuous at $F(x)$ when $x \notin D$ and $D$ has Lebesgue measure 0, $W_{G_j}(x) \to 0$ a.e. Lebesgue. By the dominated convergence theorem, $\int_d^c |W_{G_j}(x)|dx \to 0$. This proves that $\mathtt{T}$ is $\varrho_\infty$-Fréchet differentiable. The assertions on $\phi_F$ can be proved by noting that

$$\phi_F(x) = -\int_d^c (\delta_x - F)(y)J(F(y))dy.$$

(ii) From the proof of (i), we only need to show that

$$\left| \int_A W_{G_j}(x)(G_j - F)(x)dx \right| \Big/ \varrho_{\infty+1}(G_j, F) \to 0, \tag{5.42}$$

where $A = \{x : F(x) \le \alpha \text{ or } F(x) > \beta\}$. The quantity on the left-hand side of (5.42) is bounded by $\sup_{x \in A} |W_{G_j}(x)|$ which converges to 0 under the continuity assumption of $J$ on $[0, \alpha] \cup [\beta, 1]$. Hence (5.42) follows.
(iii) The result follows from

$$|R(G, F)| \le C \int |G(x) - F(x)|^p dx = O\left([\varrho_{L_p}(G, F)]^p\right)$$

and the fact that $p > 1$. ∎

An L-estimator with $J(t) = 0$ when $t \in [0, \alpha] \cup [\beta, 1]$ is called a trimmed L-estimator. Theorem 5.6(i) shows that trimmed L-estimators satisfy (5.34) and are robust in Hampel's sense. In case (ii) or (iii) of Theorem 5.6, (5.34) holds if $\mathrm{Var}(X_1) < \infty$, but $\mathtt{T}(F_n)$ may not be robust in Hampel's sense. It can be shown (exercise) that one or several of (i)-(v) of Theorem 5.6 can be applied to each of the L-estimators in Example 5.6.

### M-estimators

Let $F$ be a c.d.f. on $\mathcal{R}^d$ and $\rho(x, t)$ be a Borel function on $\mathcal{R}^d \times \mathcal{R}$. An *M-functional* is defined to be a solution of

$$\int \rho(x, \mathtt{T}(G))dG(x) = \min_{t \in \Theta} \int \rho(x, t)dG(x), \qquad G \in \mathcal{F}, \tag{5.43}$$

where $\Theta$ is an open subset of $\mathcal{R}$. $\mathtt{T}(F_n)$ is called an M-estimator of $\mathtt{T}(F)$. Assume that $\psi(x,t) = \partial\rho(x,t)/\partial t$ exists a.e. and

$$\lambda_G(t) = \int \psi(x,t)dG(x) = \frac{\partial}{\partial t}\int \rho(x,t)dG(x). \tag{5.44}$$

Then $\lambda_G(\mathtt{T}(G)) = 0$.

**Example 5.7.** The following are some examples of M-estimators.
(i) If $\rho(x,t) = (x-t)^2/2$, then $\psi(x,t) = t - x$; $\mathtt{T}(G) = \int xdG(x)$ is the mean functional; and $\mathtt{T}(F_n) = \bar{X}$ is the sample mean.
(ii) If $\rho(x,t) = |x-t|^p/p$, where $p \in [1,2)$, then

$$\psi(x,t) = \begin{cases} |x-t|^{p-1} & x < t \\ 0 & x = t \\ -|x-t|^{p-1} & x > t. \end{cases}$$

If $p = 1$, $\mathtt{T}(F_n)$ is the sample median. If $1 < p < 2$, $T(F_n)$ is called the $p$th least absolute deviations estimator or the minimum $L_p$ distance estimator.
(iii) Let $\mathcal{F}_0 = \{f_\theta : \theta \in \Theta\}$ be a parametric family of p.d.f.'s and $\rho(x,t) = -\log f_t(x)$. Then $\mathtt{T}(F_n)$ is an MLE. This indicates that M-estimators are extensions of MLE's in parametric models.
(iv) Huber (1964) considers

$$\rho(x,t) = \begin{cases} \frac{1}{2}(x-t)^2 & |x-t| \le C \\ C^2 & |x-t| > C \end{cases}$$

with

$$\psi(x,t) = \begin{cases} t-x & |x-t| \le C \\ 0 & |x-t| > C. \end{cases}$$

The corresponding $\mathtt{T}(F_n)$ is a type of trimmed sample mean.
(v) Huber (1964) considers

$$\rho(x,t) = \begin{cases} \frac{1}{2}(x-t)^2 & |x-t| \le C \\ C|x-t| - \frac{1}{2}C^2 & |x-t| > C \end{cases}$$

with

$$\psi(x,t) = \begin{cases} C & t-x > C \\ t-x & |x-t| \le C \\ -C & t-x < -C. \end{cases}$$

The corresponding $\mathtt{T}(F_n)$ is a type of Winsorized sample mean.
(vi) Hampel (1974) considers $\psi(x,t) = \psi_0(t-x)$ with $\psi_0(s) = -\psi_0(-s)$ and

$$\psi_0(s) = \begin{cases} s & 0 \le s \le a \\ a & a < s \le b \\ \frac{a(c-s)}{c-b} & b < s \le c \\ 0 & s > c, \end{cases}$$

where $0 < a < b < c$ are constants. A smoothed version of $\psi_0$ is

$$\psi_1(s) = \begin{cases} \sin(as) & 0 \le s < \pi/a \\ 0 & s > \pi/a. \end{cases} \quad \blacksquare$$

For bounded and continuous $\psi$, the following result shows that $\mathtt{T}$ is $\varrho_\infty$-Hadamard differentiable with a bounded and continuous influence function and, hence, $\mathtt{T}(F_n)$ satisfies (5.34) and is robust in Hampel's sense.

**Theorem 5.7.** Let $\mathtt{T}$ be an M-functional defined by (5.43). Assume that $\psi$ is a bounded and continuous function on $\mathcal{R}^d \times \mathcal{R}$ and that $\lambda_F(t)$ is continuously differentiable at $\mathtt{T}(F)$ and $\lambda_F'(\mathtt{T}(F)) \neq 0$. Then $\mathtt{T}$ is $\varrho_\infty$-Hadamard differentiable at $F$ with

$$\phi_F(x) = -\psi(x, \mathtt{T}(F))/\lambda_F'(\mathtt{T}(F)).$$

**Proof.** Let $t_j \to 0$, $\Delta_j \in \mathfrak{D}$, $\|\Delta_j - \Delta\|_\infty \to 0$, and $G_j = F + t_j\Delta_j \in \mathcal{F}$. Since $\lambda_G(\mathtt{T}(G)) = 0$,

$$|\lambda_F(\mathtt{T}(G_j)) - \lambda_F(\mathtt{T}(F))| = \left| t_j \int \psi(x, \mathtt{T}(G_j)) d\Delta_j(x) \right| \to 0$$

by $\|\Delta_j - \Delta\|_\infty \to 0$ and the boundedness of $\psi$. Note that $\lambda_F'(\mathtt{T}(F)) \neq 0$. Hence, the inverse of $\lambda_F(t)$ exists and is continuous in a neighborhood of $0 = \lambda_F(\mathtt{T}(F))$. Therefore,

$$\mathtt{T}(G_j) - \mathtt{T}(F) \to 0. \tag{5.45}$$

Let $h_F(\mathtt{T}(F)) = \lambda_F'(\mathtt{T}(F))$, $h_F(t) = [\lambda_F(t) - \lambda_F(\mathtt{T}(F))]/[t - \mathtt{T}(F)]$ if $t \neq \mathtt{T}(F)$,

$$R_{1j} = \int \psi(x, \mathtt{T}(F)) d\Delta_j(x) \left[ \frac{1}{\lambda_F'(\mathtt{T}(F))} - \frac{1}{h_F(\mathtt{T}(G_j))} \right],$$

$$R_{2j} = \frac{1}{h_F(\mathtt{T}(G_j))} \int [\psi(x, \mathtt{T}(G_j)) - \psi(x, \mathtt{T}(F))] d\Delta_j(x),$$

and

$$\mathtt{L}_F(\Delta) = -\frac{1}{\lambda_F'(\mathtt{T}(F))} \int \psi(x, \mathtt{T}(F)) d\Delta(x), \qquad \Delta \in \mathfrak{D}.$$

Then

$$\mathtt{T}(G_j) - \mathtt{T}(F) = -\mathtt{L}_F(t_j\Delta_j) + t_j(R_{1j} - R_{2j}).$$

By (5.45), $\|\Delta_j - \Delta\|_\infty \to 0$, and the boundedness of $\psi$, $R_{j1} \to 0$. The result then follows from $R_{2j} \to 0$, which follows from $\|\Delta_j - \Delta\|_\infty \to 0$ and the boundedness and continuity of $\psi$ (exercise). $\blacksquare$

Some $\psi$ functions in Example 5.7 satisfy the conditions in Theorem 5.7 (exercise). Under more conditions on $\psi$, it can be shown that an M-functional is $\varrho_\infty$-Fréchet differentiable at $F$ (Clarke, 1986; Shao, 1993). Some M-estimators that satisfy (5.34) but are not differentiable functionals are studied in §5.4.

**Rank statistics and R-estimators**

Assume that $X_1, ..., X_n$ are i.i.d. from a c.d.f. $F$ on $\mathcal{R}$. The *rank* of $X_i$ among $X_1, ..., X_n$, denoted by $R_i$, is defined to be the number of $X_j$'s satisfying $X_j \le X_i$, $i = 1, ..., n$. The rank of $|X_i|$ among $|X_1|, ..., |X_n|$ is similarly defined and denoted by $\tilde{R}_i$. A statistic that is a function of $R_i$'s or $\tilde{R}_i$'s is called a *rank statistic*. For $G \in \mathcal{F}$, let

$$\tilde{G}(x) = G(x) - G\big((-x)-\big), \qquad x > 0,$$

where $g(x-)$ denotes the left limit of the function $g$ at $x$. Define a functional $\mathtt{T}$ by

$$\mathtt{T}(G) = \int_0^\infty J(\tilde{G}(x))dG(x), \qquad G \in \mathcal{F}, \tag{5.46}$$

where $J$ is a function on $[0, 1]$ with a bounded $J'$. Then

$$\mathtt{T}(F_n) = \int_0^\infty J(\tilde{F}_n(x))dF_n(x) = \frac{1}{n}\sum_{i=1}^n J\left(\frac{\tilde{R}_i}{n}\right) I_{(0,\infty)}(X_i)$$

is a (one-sample) signed rank statistic. If $J(t) = t$, then $\mathtt{T}(F_n)$ is the well-known Wilcoxon signed rank test (§6.5.1).

Statistics based on ranks (or signed ranks) are robust against changes in values of $x_i$'s, but may not provide efficient inference procedures, since the values of $x_i$'s are discarded after ranks (or signed ranks) are determined.

It can be shown (exercise) that $\mathtt{T}$ in (5.46) is $\varrho_\infty$-Hadamard differentiable at $F$ with the differential

$$\mathtt{L}_F(\Delta) = \int_0^\infty J'(\tilde{F}(x))\tilde{\Delta}(x)dF(x) + \int_0^\infty J(\tilde{F}(x))d\Delta(x). \tag{5.47}$$

These results can be extended to the case where $X_1, ..., X_n$ are i.i.d. from a c.d.f. $F$ on $\mathcal{R}^2$. For any c.d.f. $G$ on $\mathcal{R}^2$, let $J$ be a function on $[0, 1]$ with $J(1-t) = -J(t)$ and a bounded $J'$,

$$\bar{G}(y) = [G(y, \infty) + G(\infty, y)]/2, \qquad y \in \mathcal{R},$$

and

$$\mathtt{T}(G) = \int J(\bar{G}(y))dG(y, \infty). \tag{5.48}$$

Let $X_i = (Y_i, Z_i)$, $R_i$ be the rank of $Y_i$, and $U_i$ be the number of $Z_j$'s satisfying $Z_j \le Y_i$, $i = 1, ..., n$. Then

$$\mathtt{T}(F_n) = \int J(\bar{F}_n(y))dF_n(y, \infty) = \frac{1}{n}\sum_{i=1}^{n} J\left(\tfrac{R_i+U_i}{2n}\right)$$

is called a two-sample linear rank statistic. It can be shown (exercise) that $\mathtt{T}$ in (5.48) is $\varrho_\infty$-Hadamard differentiable at $F$ with the differential

$$\mathtt{L}_F(\Delta) = \int J'(\bar{F}(y))\bar{\Delta}(y)dF(y, \infty) + \int J(\bar{F}(y))d\Delta(y, \infty). \tag{5.49}$$

Rank statistics (one-sample or two-sample) are asymptotically normal and robust in Hampel's sense (exercise). These results are useful in testing hypotheses (§6.5).

Let $F$ be a continuous c.d.f. on $\mathcal{R}$ symmetric about an unknown parameter $\theta \in \mathcal{R}$. An estimator of $\theta$ closely related to a rank statistic can be derived as follows. Let $X_i$ be i.i.d. from $F$ and $W_i = (X_i, 2t - X_i)$ with a fixed $t \in \mathcal{R}$. The functional $\mathtt{T}$ in (5.48) evaluated at the c.d.f. of $W_i$ is equal to

$$\lambda_F(t) = \int J\left(\tfrac{F(x)+1-F(2t-x)}{2}\right) dF(x). \tag{5.50}$$

If $J$ is strictly increasing and $F$ is strictly increasing in a neighborhood of $\theta$, then $\lambda_F(t) = 0$ if and only if $t = \theta$ (exercise). For $G \in \mathcal{F}$, define $\mathtt{T}(G)$ to be a solution of

$$\int J\left(\tfrac{G(x)+1-G(2\mathtt{T}(G)-x)}{2}\right) dG(x) = 0. \tag{5.51}$$

$\mathtt{T}(F_n)$ is called an *R-estimator* of $\mathtt{T}(F) = \theta$. When $J(t) = t - \frac{1}{2}$ (which is related to the Wilcoxon signed rank test), $\mathtt{T}(F_n)$ is the well-known Hodges-Lehmann estimator and is equal to any value between the two middle points of the values $(X_i + X_j)/2$, $i = 1, ..., n$, $j = 1, ..., n$.

**Theorem 5.8.** Let $\mathtt{T}$ be the functional defined by (5.51). Suppose that $F$ is continuous and symmetric about $\theta$, the derivatives $F'$ and $J'$ exist, and $J'$ is bounded. Then $\mathtt{T}$ is $\varrho_\infty$-Hadamard differentiable at $F$ with the influence function

$$\phi_F(x) = \frac{J(F(x))}{\int J'(F(x))F'(x)dF(x)}.$$

**Proof.** Since $F$ is symmetric about $\theta$, $F(x) + F(2\theta - x) = 1$. Under the assumed conditions, $\lambda_F(t)$ is continuous and $\int J'(F(x))F'(x)dF(x) = -\lambda'_F(\theta) \neq 0$ (exercise). Hence the inverse of $\lambda_F$ exists and is continuous

at $0 = \lambda_F(\theta)$. Suppose that $t_j \to 0$, $\Delta_j \in \mathfrak{D}$, $\|\Delta_j - \Delta\|_\infty \to 0$, and $G_j = F + t_j\Delta_j \in \mathfrak{F}$. Then

$$\int [J(G_j(x,t)) - J(F(x,t))]dG_j(x) \to 0$$

uniformly in $t$, where $G(x,t) = [G(x) + 1 - G(2t - x)]/2$, and

$$\int J(F(x,t))d(G_j - F)(x) = \int (G_j - F)(x)J'(F(x,t))dF(x,t) \to 0$$

uniformly in $t$. Let $\lambda_G(t)$ be defined by (5.50) with $F$ replaced by $G$. Then

$$\lambda_{G_j}(t) - \lambda_F(t) \to 0$$

uniformly in $t$. Thus, $\lambda_F(\mathtt{T}(G_j)) \to 0$, which implies

$$\mathtt{T}(G_j) \to \mathtt{T}(F) = \theta. \tag{5.52}$$

Let $\xi_G(t) = \int J(F(x,t))dG(x)$, $h_F(t) = [\lambda_F(t) - \lambda_F(\theta)]/(t - \theta)$ if $t \neq \theta$, and $h_F(\theta) = \lambda'_F(\theta)$. Then $\mathtt{T}(G_j) - \mathtt{T}(F) - \int \phi_F(x)d(G_j - F)(x)$ is equal to

$$\xi_{G_j}(\theta)\left[\frac{1}{\lambda'_F(\theta)} - \frac{1}{h_F(\mathtt{T}(G_j))}\right] + \frac{\lambda_F(\mathtt{T}(G_j)) - \xi_{G_j}(\theta)}{h_F(\mathtt{T}(G_j))}. \tag{5.53}$$

Note that

$$\xi_{G_j}(\theta) = \int J(F(x))dG_j(x) = t_j \int J(F(x))d\Delta_j(x).$$

By (5.52), Lemma 5.2, and $\|\Delta_j - \Delta\|_\infty \to 0$, the first term in (5.53) is $o(t_j)$. The second term in (5.53) is the sum of

$$-\frac{t_j}{h_F(\mathtt{T}(G_j))}\int [J(F(x,\mathtt{T}(G_j))) - J(F(x))]d\Delta_j(x) \tag{5.54}$$

and

$$\frac{1}{h_F(\mathtt{T}(G_j))}\int [J(F(x,\mathtt{T}(G_j))) - J(G_j(x,\mathtt{T}(G_j)))]dG_j(x). \tag{5.55}$$

From the continuity of $J$ and $F$, the quantity in (5.54) is $o(t_j)$. Similarly, the quantity in (5.55) is equal to

$$\frac{1}{h_F(\mathtt{T}(G_j))}\int [J(F(x,\mathtt{T}(G_j))) - J(G_j(x,\mathtt{T}(G_j)))]dF(x) + o(t_j). \tag{5.56}$$

Using Taylor's expansion, (5.52), and $\|\Delta_j - \Delta\|_\infty \to 0$, the quantity in (5.56) is equal to

$$\frac{t_j}{h_F(\mathtt{T}(G_j))}\int J'(F(x))\Delta(x,\theta)dF(x) + o(t_j). \tag{5.57}$$

Since $J(1-t) = -J(t)$, the integral in (5.57) is 0. This proves that the second term in (5.53) is $o(t_j)$ and thus the result. ∎

It is clear that the influence function $\phi_F$ for an R-estimator is bounded and continuous if $J$ and $F$ are continuous. Thus, R-estimators satisfy (5.34) and are robust in Hampel's sense.

**Example 5.8.** Let $J(t) = t - \frac{1}{2}$. Then $\mathtt{T}(F_n)$ is the Hodges-Lehmann estimator. From Theorem 5.8, $\phi_F(x) = [F(x) - \frac{1}{2}]/\gamma$, where $\gamma = \int F'(x)dF(x)$. Since $F(X_1)$ has a uniform distribution on $[0,1]$, $\phi_F(X_1)$ has mean 0 and variance $(12\gamma^2)^{-1}$. Thus, $\sqrt{n}[\mathtt{T}(F_n) - \mathtt{T}(F)] \to_d N(0, (12\gamma^2)^{-1})$. ∎

## 5.3 Linear Functions of Order Statistics

In this section we study statistics that are linear functions of order statistics $X_{(1)} \le \cdots \le X_{(n)}$, based on independent random variables $X_1, ..., X_n$ (in §5.3.1 and §5.3.2, $X_1, ..., X_n$ are assumed i.i.d.). Order statistics, first introduced in Example 2.9, are usually sufficient and often complete (or minimal sufficient) for nonparametric families (Examples 2.12 and 2.14).

L-estimators defined in §5.2.2 are in fact linear functions of order statistics. If $\mathtt{T}$ is given by (5.39), then

$$\mathtt{T}(F_n) = \int xJ(F_n(x))dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} J\left(\tfrac{i}{n}\right) X_{(i)}, \tag{5.58}$$

since $F_n(X_{(i)}) = i/n$, $i = 1, ..., n$. If $J$ is a smooth function, such as those given in Example 5.6 or those satisfying the conditions in Theorem 5.6, the corresponding L-estimator is often called a smooth L-estimator. Asymptotic properties of smooth L-estimators can be obtained using Theorem 5.6 and the results in §5.2.1. Results on L-estimators that are slightly different from that in (5.58) can be found in Serfling (1980, Chapter 8).

In §5.3.1, we consider another useful class of linear functions of order statistics, the sample quantiles described in the beginning of §5.2. In §5.3.2, we study robust linear functions of order statistics (in Hampel's sense) and their relative efficiencies w.r.t. $\bar{X}$, an efficient but nonrobust estimator. In §5.3.3, extensions to linear models are discussed.

### 5.3.1 Sample quantiles

Recall that $G^{-1}(p)$ is defined to be $\inf\{x : G(x) \ge p\}$ for any c.d.f. $G$ on $\mathcal{R}$, where $p \in (0,1)$ is a fixed constant. For i.i.d. $X_1, ..., X_n$ from $F$, let $\theta_p = F^{-1}(p)$ and $\hat{\theta}_p = F_n^{-1}(p)$ denote the $p$th quantile of $F$ and the $p$th

sample quantile, respectively. Then

$$\hat{\theta}_p = c_{np} X_{(m_p)} + (1 - c_{np}) X_{(m_p+1)}, \tag{5.59}$$

where $m_p$ is the integer part of $np$, $c_{np} = 1$ if $np$ is an integer, and $c_{np} = 0$ if $np$ is not an integer. Thus, $\hat{\theta}_p$ is a linear function of order statistics.

Note that $F(\theta_p-) \le p \le F(\theta_p)$. If $F$ is not flat in a neighborhood of $\theta_p$, then $F(\theta_p - \epsilon) < p < F(\theta_p + \epsilon)$ for any $\epsilon > 0$.

**Theorem 5.9.** Let $X_1, ..., X_n$ be i.i.d. random variables from a c.d.f. $F$ satisfying $F(\theta_p - \epsilon) < p < F(\theta_p + \epsilon)$ for any $\epsilon > 0$. Then, for every $\epsilon > 0$ and $n = 1, 2,...,$

$$P\big(|\hat{\theta}_p - \theta_p| > \epsilon\big) \le 2Ce^{-2n\delta_\epsilon^2}, \tag{5.60}$$

where $\delta_\epsilon$ is the smaller of $F(\theta_p + \epsilon) - p$ and $p - F(\theta_p - \epsilon)$ and $C$ is the same constant in Lemma 5.1(i).

**Proof.** Let $\epsilon > 0$ be fixed. Note that $G(x) \ge t$ if and only if $x \ge G^{-1}(t)$ for any c.d.f. $G$ on $\mathcal{R}$ (exercise). Hence

$$\begin{aligned} P\big(\hat{\theta}_p > \theta_p + \epsilon\big) &= P\big(p > F_n(\theta_p + \epsilon)\big) \\ &= P\big(F(\theta_p + \epsilon) - F_n(\theta_p + \epsilon) > F(\theta_p + \epsilon) - p\big) \\ &\le P\big(\varrho_\infty(F_n, F) > \delta_\epsilon\big) \\ &\le Ce^{-2n\delta_\epsilon^2}, \end{aligned}$$

where the last inequality follows from DKW's inequality (Lemma 5.1(i)). Similarly,

$$P\big(\hat{\theta}_p < \theta_p - \epsilon\big) \le Ce^{-2n\delta_\epsilon^2}.$$

This proves (5.60). ∎

Result (5.60) implies that $\hat{\theta}_p$ is strongly consistent for $\theta_p$ (exercise) and that $\hat{\theta}_p$ is $\sqrt{n}$-consistent for $\theta_p$ if $F'(\theta_p-)$ and $F'(\theta_p+)$ (the left and right derivatives of $F$ at $\theta_p$) exist (exercise).

The exact distribution of $\hat{\theta}_p$ can be obtained as follows. Since $nF_n(t)$ has the binomial distribution $Bi(F(t), n)$ for any $t \in \mathcal{R}$,

$$\begin{aligned} P\big(\hat{\theta}_p \le t\big) &= P\big(F_n(t) \ge p\big) \\ &= \sum_{i=m_p}^{n} \binom{n}{i} [F(t)]^i [1 - F(t)]^{n-i}, \end{aligned} \tag{5.61}$$

where $m_p$ is given in (5.59). If $F$ has a Lebesgue p.d.f. $f$, then $\hat{\theta}_p$ has the Lebesgue p.d.f.

$$\varphi_n(t) = n\binom{n-1}{m_p - 1} [F(t)]^{m_p - 1} [1 - F(t)]^{n - m_p} f(t). \tag{5.62}$$

The following result provides an asymptotic distribution for $\sqrt{n}(\hat{\theta}_p - \theta_p)$.

**Theorem 5.10.** Let $X_1, ..., X_n$ be i.i.d. random variables from $F$.
(i) $P(\sqrt{n}(\hat{\theta}_p - \theta_p) \le 0) \to \Phi(0) = \frac{1}{2}$, where $\Phi$ is the c.d.f. of the standard normal.
(ii) If $F$ is continuous at $\theta_p$ and there exists $F'(\theta_p-) > 0$, then

$$P\big(\sqrt{n}(\hat{\theta}_p - \theta_p) \le t\big) \to \Phi(t/\sigma_F^-), \qquad t < 0,$$

where $\sigma_F^- = \sqrt{p(1-p)}/F'(\theta_p-)$.
(iii) If $F$ is continuous at $\theta_p$ and there exists $F'(\theta_p+) > 0$, then

$$P\big(\sqrt{n}(\hat{\theta}_p - \theta_p) \le t\big) \to \Phi(t/\sigma_F^+), \qquad t > 0,$$

where $\sigma_F^+ = \sqrt{p(1-p)}/F'(\theta_p+)$.
(iv) If $F'(\theta_p)$ exists and is positive, then

$$\sqrt{n}(\hat{\theta}_p - \theta_p) \to_d N(0, \sigma_F^2), \tag{5.63}$$

where $\sigma_F = \sqrt{p(1-p)}/F'(\theta_p)$.
**Proof.** The proof of (i) is left as an exercise. Part (iv) is a direct consequence of (i)-(iii) and the proofs of (ii) and (iii) are similar. Thus, we only give a proof for (iii).

Let $t > 0$, $p_{nt} = F(\theta_p + t\sigma_F^+ n^{-1/2})$, $c_{nt} = \sqrt{n}(p_{nt} - p)/\sqrt{p_{nt}(1-p_{nt})}$, and $Z_{nt} = [B_n(p_{nt}) - np_{nt}]/\sqrt{np_{nt}(1-p_{nt})}$, where $B_n(q)$ denotes a random variable having the binomial distribution $Bi(q, n)$. Then

$$\begin{aligned} P\big(\hat{\theta}_p \le \theta_p + t\sigma_F^+ n^{-1/2}\big) &= P\big(p \le F_n(\theta_p + t\sigma_F^+ n^{-1/2})\big) \\ &= P\big(Z_{nt} \ge -c_{nt}\big). \end{aligned}$$

Under the assumed conditions on $F$, $p_{nt} \to p$ and $c_{nt} \to t$. Hence, the result follows from

$$P\big(Z_{nt} < -c_{nt}\big) - \Phi(-c_{nt}) \to 0.$$

But this follows from the CLT (Example 1.26) and Pólya's theorem (Proposition 1.16). ∎

If both $F'(\theta_p-)$ and $F'(\theta_p+)$ exist and are positive, but $F'(\theta_p-) \neq F'(\theta_p+)$, then the asymptotic distribution of $\sqrt{n}(\hat{\theta}_p - \theta_p)$ has the c.d.f. $\Phi(t/\sigma_F^-)I_{(-\infty,0)}(t) + \Phi(t/\sigma_F^+)I_{[0,\infty)}(t)$, a mixture of two normal distributions. An example of such a case when $p = \frac{1}{2}$ is

$$F(x) = xI_{[0,\frac{1}{2})}(x) + (2x - \tfrac{1}{2})I_{[\frac{1}{2},\frac{3}{4})}(x) + I_{[\frac{3}{4},\infty)}(x).$$

When $F'(\theta_p-) = F'(\theta_p+) = F'(\theta_p) > 0$, (5.63) shows that the asymptotic distribution of $\sqrt{n}(\hat{\theta}_p - \theta_p)$ is the same as that of $\sqrt{n}[F_n(\theta_p) - F(\theta_p)]/F'(\theta_p)$ (see (5.2)). The following result reveals a stronger relationship between sample quantiles and the empirical c.d.f.

**Theorem 5.11** (Bahadur's representation). Let $X_1, ..., X_n$ be i.i.d. random variables from $F$. Suppose that $F'(\theta_p)$ exists and is positive. Then

$$\hat{\theta}_p = \theta_p + \frac{F(\theta_p) - F_n(\theta_p)}{F'(\theta_p)} + o_p\left(\tfrac{1}{\sqrt{n}}\right). \tag{5.64}$$

**Proof.** Let $t \in \mathcal{R}$, $\theta_{nt} = \theta_p + tn^{-1/2}$, $Z_n(t) = \sqrt{n}[F(\theta_{nt}) - F_n(\theta_{nt})]/F'(\theta_p)$, and $U_n(t) = \sqrt{n}[F(\theta_{nt}) - F_n(\theta_p)]/F'(\theta_p)$. It can be shown (exercise) that

$$Z_n(t) - Z_n(0) = o_p(1). \tag{5.65}$$

Note that $|p - F_n(\hat{\theta}_p)| \le n^{-1}$. Then

$$\begin{aligned} U_n(t) &= \sqrt{n}[F(\theta_{nt}) - p + p - F_n(\hat{\theta}_p)]/F'(\theta_p) \\ &= \sqrt{n}[F(\theta_{nt}) - p]/F'(\theta_p) + O(n^{-1/2}) \\ &\to t. \end{aligned} \tag{5.66}$$

Let $\xi_n = \sqrt{n}(\hat{\theta}_p - \theta_p)$. Then, for any $t \in \mathcal{R}$ and $\epsilon > 0$,

$$\begin{aligned} P\big(\xi_n \le t, Z_n(0) \ge t + \epsilon\big) &= P\big(Z_n(t) \le U_n(t), Z_n(0) \ge t + \epsilon\big) \\ &\le P\big(|Z_n(t) - Z_n(0)| \ge \epsilon/2\big) \\ &\quad + P\big(|U_n(t) - t| \ge \epsilon/2\big) \\ &\to 0 \end{aligned} \tag{5.67}$$

by (5.65) and (5.66). Similarly,

$$P\big(\xi_n \ge t + \epsilon, Z_n(0) \le t\big) \to 0. \tag{5.68}$$

It follows from the result in Exercise 97 of §1.6 that

$$\xi_n - Z_n(0) = o_p(1),$$

which is the same as (5.64). ∎

If $F$ has a positive Lebesgue p.d.f., then $\hat{\theta}_p$ viewed as a statistical functional (§5.2) is $\varrho_\infty$-Hadamard differentiable at $F$ (Fernholz, 1983) with the influence function

$$\phi_F(x) = [F(\theta_p) - I_{(-\infty,\theta_p]}(x)]/F'(\theta_p).$$

This implies result (5.64). Furthermore, $\hat{\theta}_p$ is robust in Hampel's sense; see also §5.3.2.

**Corollary 5.1.** Let $X_1, ..., X_n$ be i.i.d. random variables from $F$ having positive derivatives at $\theta_{p_j}$, where $0 < p_1 < \cdots < p_m < 1$ are fixed constants. Then

$$\sqrt{n}[(\hat{\theta}_{p_1}, ..., \hat{\theta}_{p_m}) - (\theta_{p_1}, ..., \theta_{p_m})] \to_d N_m(0, D),$$

where $D$ is the $m \times m$ symmetric matrix whose $(i, j)$th element is

$$p_i(1 - p_j)/[F'(\theta_{p_i})F'(\theta_{p_j})], \qquad i < j. \quad \blacksquare$$

The proof of this corollary is left to the reader.

**Example 5.9** (Interquartile range). One application of Corollary 5.1 is the derivation of the asymptotic distribution of the *interquartile range* $\hat{\theta}_{0.75} - \hat{\theta}_{0.25}$. The interquartile range is used as a measure of the variability among $X_i$'s. It can be shown (exercise) that

$$\sqrt{n}[(\hat{\theta}_{0.75} - \hat{\theta}_{0.25}) - (\theta_{0.75} - \theta_{0.25})] \to_d N(0, \sigma_F^2)$$

with

$$\sigma_F^2 = \frac{3}{16[F'(\theta_{0.75})]^2} + \frac{3}{16[F'(\theta_{0.25})]^2} - \frac{1}{8F'(\theta_{0.75})F'(\theta_{0.25})}. \quad \blacksquare$$

There are some applications of using extreme order statistics such as $X_{(1)}$ and $X_{(n)}$. One example is given in Example 2.34. Some other examples and references can be found in Serfling (1980, pp. 89-91).

### 5.3.2 Robustness and efficiency

Let $F$ be a c.d.f. on $\mathcal{R}$ symmetric about $\theta \in \mathcal{R}$ with $F'(\theta) > 0$. Then $\theta = \theta_{0.5}$ and is called the *median* of $F$. If $F$ has a finite mean, then $\theta$ is also equal to the mean. In this section we consider the estimation of $\theta$ based on i.i.d. $X_i$'s from $F$.

If $F$ is normal, it has been shown in previous chapters that $\bar{X}$ is the UMVUE, MRIE, and MLE of $\theta$, and is asymptotically efficient. On the other hand, if $F$ is the c.d.f. of the Cauchy distribution $C(\theta, 1)$, it follows from Exercise 50 in §1.6 that $\bar{X}$ has the same distribution as $X_1$, i.e., $\bar{X}$ is as variable as $X_1$, and is inconsistent as an estimator of $\theta$.

Why does $\bar{X}$ perform so differently? An important difference between the normal and Cauchy p.d.f.'s is that the former tends to 0 at the rate

$e^{-x^2/2}$ as $|x| \to \infty$, whereas the latter tends to 0 at the much slower rate $x^{-2}$, which results in $\int |x| dF(x) = \infty$. The poor performance of $\bar{X}$ in the Cauchy case is due to the high probability of getting extreme observations and the fact that $\bar{X}$ is sensitive to large changes in a few of the $X_i$'s. (Note that $\bar{X}$ is not robust in Hampel's sense, since the functional $\int x dG(x)$ has an unbounded influence function at $F$.) This suggests the use of a robust estimator that discards some extreme observations. The *sample median*, which is defined to be the 50%th sample quantile $\hat{\theta}_{0.5}$ described in §5.3.1, is insensitive to the behavior of $F$ as $|x| \to \infty$.

Since both the sample mean and the sample median can be used to estimate $\theta$, a natural question is when one is better than the other, using a criterion such as the amse. Unfortunately, a general answer does not exist, since the asymptotic relative efficiency between these two estimators depends on the unknown distribution $F$. If $F$ does not have a finite variance, then the sample median is certainly preferred since $\bar{X}$ is inconsistent but $\hat{\theta}_{0.5}$ is consistent and asymptotically normal as long as $F'(\theta) > 0$, and may still have a finite variance (Exercise 54). The following example, which compares the sample mean and median in some cases, shows that the sample median can be better even if $\text{Var}(X_1) < \infty$.

**Example 5.10.** Suppose that $\text{Var}(X_1) < \infty$. Then, by the CLT,

$$\sqrt{n}(\bar{X} - \theta) \to_d N(0, \text{Var}(X_1)).$$

By Theorem 5.10(iv),

$$\sqrt{n}(\hat{\theta}_{0.5} - \theta) \to_d N(0, [2F'(\theta)]^{-2}).$$

Hence, the asymptotic relative efficiency of $\hat{\theta}_{0.5}$ w.r.t. $\bar{X}$ is

$$e(F) = 4[F'(\theta)]^2 \text{Var}(X_1).$$

(i) If $F$ is the c.d.f. of $N(\theta, \sigma^2)$, then $\text{Var}(X_1) = \sigma^2$, $F'(\theta) = (\sqrt{2\pi}\sigma)^{-1}$, and $e(F) = 2/\pi = 0.637$.
(ii) If $F$ is the c.d.f. of the logistic distribution $LG(\theta, \sigma)$, then $\text{Var}(X_1) = \sigma^2\pi^2/3$, $F'(\theta) = (4\sigma)^{-1}$, and $e(F) = \pi^2/12 = 0.822$.
(iii) If $F(x) = F_0(x - \theta)$ and $F_0$ is the c.d.f. of the t-distribution $t_\nu$ with $\nu \geq 3$, then $\text{Var}(X_1) = \nu/(\nu - 2)$, $F'(\theta) = \Gamma(\frac{\nu+1}{2})/[\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})]$, $e(F) = 1.62$ when $\nu = 3$, $e(F) = 1.12$ when $\nu = 4$, and $e(F) = 0.96$ when $\nu = 5$.
(iv) If $F$ is the c.d.f. of the double exponential distribution $DE(\theta, \sigma)$, then $F'(\theta) = (2\sigma)^{-1}$ and $e(F) = 2$.
(v) Consider the Tukey model

$$F(x) = (1 - \epsilon)\Phi\left(\tfrac{x-\theta}{\sigma}\right) + \epsilon\Phi\left(\tfrac{x-\theta}{\tau\sigma}\right), \tag{5.69}$$

where $\sigma > 0$, $\tau > 0$ and $0 < \epsilon < 1$. Then $\text{Var}(X_1) = (1-\epsilon)\sigma^2 + \epsilon\tau^2\sigma^2$, $F'(\theta) = (1 - \epsilon + \epsilon/\tau)/(\sqrt{2\pi}\sigma)$, and $e(F) = 2(1 - \epsilon + \epsilon\tau^2)(1 - \epsilon + \epsilon/\tau)^2/\pi$. Note that $\lim_{\epsilon\to 0} e(F) = 2/\pi$ and $\lim_{\tau\to\infty} e(F) = \infty$. ∎

Since the sample median uses at most two actual values of $x_i$'s, it may go too far in discarding observations, which results in a possible loss of efficiency. The trimmed sample mean introduced in Example 5.6(iii) is a natural compromise between the sample mean and median. Since $F$ is symmetric, we consider $\beta = 1 - \alpha$ in the trimmed mean, which results in the following L-estimator

$$\bar{X}_\alpha = \frac{1}{n - 2m_\alpha} \sum_{j=m_\alpha+1}^{n-m_\alpha} X_{(j)}, \tag{5.70}$$

where $m_\alpha$ is the integer part of $n\alpha$ and $\alpha \in (0, \frac{1}{2})$. The estimator in (5.70) is called the $\alpha$-trimmed sample mean. It discards the $m_\alpha$ smallest and $m_\alpha$ largest observations. The sample mean and median can be viewed as two extreme cases of $\bar{X}_\alpha$ as $\alpha \to 0$ and $\frac{1}{2}$, respectively.

It follows from Theorem 5.6 that if $F(x) = F_0(x - \theta)$, where $F_0$ is symmetric about 0 and has a Lebesgue p.d.f. positive in the range of $X_1$, then

$$\sqrt{n}(\bar{X}_\alpha - \theta) \to_d N(0, \sigma_\alpha^2), \tag{5.71}$$

where

$$\sigma_\alpha^2 = \frac{2}{(1-2\alpha)^2}\left[\int_0^{\theta_{1-\alpha}} x^2 dF_0(x) + \alpha\theta_{1-\alpha}^2\right].$$

Lehmann (1983, §5.4) provides various values of the asymptotic relative efficiency $e_{\bar{X}_\alpha,\bar{X}}(F) = \text{Var}(X_1)/\sigma_\alpha^2$. For instance, when $F(x) = F_0(x - \theta)$ and $F_0$ is the c.d.f. of the t-distribution $t_3$, $e_{\bar{X}_\alpha,\bar{X}}(F) = 1.70$, 1.91, and 1.97 for $\alpha = 0.05$, 0.125, and 0.25, respectively; when $F$ is given by (5.69) with $\tau = 3$ and $\epsilon = 0.05$, $e_{\bar{X}_\alpha,\bar{X}}(F) = 1.20$, 1.19, and 1.09 for $\alpha = 0.05$, 0.125, and 0.25, respectively; when $F$ is given by (5.69) with $\tau = 3$ and $\epsilon = 0.01$, $e_{\bar{X}_\alpha,\bar{X}}(F) = 1.04$, 0.98, and 0.89 for $\alpha = 0.05$, 0.125, and 0.25, respectively.

Robustness and efficiency of other L-estimators can be discussed similarly. For an L-estimator $\mathtt{T}(F_n)$ with $\mathtt{T}$ given by (5.39), if the conditions in one of (i)-(iii) of Theorem 5.6 are satisfied, then (5.34) holds with

$$\sigma_F^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} J(F(x))J(F(y))[F(\min(x,y)) - F(x)F(y)]dxdy, \tag{5.72}$$

provided that $\sigma_F^2 < \infty$ (exercise). If $F$ is symmetric about $\theta$ and $J$ is symmetric about $\frac{1}{2}$, then $\mathtt{T}(F) = \theta$ (exercise) and, therefore, the asymptotic relative efficiency of $\mathtt{T}(F_n)$ w.r.t. $\bar{X}$ is $\text{Var}(X_1)/\sigma_F^2$.

### 5.3.3 L-estimators in linear models

In this section we extend L-estimators to the following linear model:

$$X_i = \beta Z_i^\tau + \varepsilon_i, \qquad i = 1, ..., n, \tag{5.73}$$

with i.i.d. $\varepsilon_i$'s having an unknown c.d.f. $F_0$ and a full rank $Z = (Z_1^\tau, ..., Z_n^\tau)^\tau$. Note that the c.d.f. of $X_i$ is $F_0(x - \beta Z_i^\tau)$. Instead of assuming $E(\varepsilon_i) = 0$ (as we did in Chapter 3), we assume that

$$\int xJ(F_0(x))dF_0(x) = 0, \tag{5.74}$$

where $J$ is a function on $[0, 1]$ (the same as that in (5.39)). Note that (5.74) may hold without any assumption on the existence of $E(\varepsilon_i)$. For instance, (5.74) holds if $F_0$ is symmetric about 0 and $J$ is symmetric about $\frac{1}{2}$ (Exercise 61).

Since $X_i$'s are not identically distributed, the use of the order statistics and the empirical c.d.f. based on $X_1, ..., X_n$ may not be appropriate. Instead, we consider the ordered values of residuals $r_i = X_i - \hat{\beta} Z_i^\tau$, $i = 1, ..., n$, and some empirical c.d.f.'s based on residuals, where $\hat{\beta} = XZ(Z^\tau Z)^{-1}$ is the LSE of $\beta$.

To illustrate the idea, let us start with the case where $\beta$ and $Z_i$ are univariate. First, assume that $Z_i \geq 0$ for all $i$. Let $\hat{F}_0$ be the c.d.f. putting mass $Z_i / \sum_{i=1}^n Z_i^2$ at $r_i$, $i = 1, ..., n$. An L-estimator of $\beta$ is defined to be

$$\hat{\beta}_L = \hat{\beta} + \int xJ(\hat{F}_0(x))d\hat{F}_0(x) \sum_{i=1}^n Z_i \Big/ \sum_{i=1}^n Z_i^2. \tag{5.75}$$

When $J(t) = (1-2\alpha)^{-1} I_{(\alpha, 1-\alpha)}(t)$ with an $\alpha \in (0, \frac{1}{2})$, which corresponds to the $\alpha$-trimmed sample mean in the i.i.d. case, $\hat{\beta}_L$ in (5.75) can be computed as follows. Order the residuals and trim off all observations corresponding to residuals $r_{(j)}$ with $w_j = \sum_{i=1}^j Z_{\delta_i} / \sum_{i=1}^n Z_i^2 \in [0, \alpha] \cup [1 - \alpha, 1]$, where $r_{(1)} \leq \cdots \leq r_{(n)}$ are ordered residuals and $\delta_i$ satisfies $r_{\delta_i} = r_{(i)}$. Then $\hat{\beta}_L$ is the LSE based on the remaining observations.

If some $Z_i$'s are negative, we can define L-estimators as follows. Let $Z_i^+ = \max(Z_i, 0)$ and $Z_i^- = Z_i^+ - Z_i$. Let $\hat{F}_0^\pm$ be the c.d.f. putting mass $Z_i^\pm / \sum_{i=1}^n Z_i^2$ at $r_i$, $i = 1, ..., n$. An L-estimator of $\beta$ is defined to be

$$\begin{aligned} \hat{\beta}_L = \hat{\beta} &+ \int xJ(\hat{F}_0^+(x))d\hat{F}_0^+(x) \sum_{i=1}^n Z_i^+ \Big/ \sum_{i=1}^n Z_i^2 \\ &- \int xJ(\hat{F}_0^-(x))d\hat{F}_0^-(x) \sum_{i=1}^n Z_i^- \Big/ \sum_{i=1}^n Z_i^2. \end{aligned}$$

For general $p$-vector $Z_i$, let $z_{ij}$ be the $j$th component of $Z_i$, $j = 1, ..., p$. Let $z_{ij}^+ = \max(z_{ij}, 0)$, $z_{ij}^- = z_{ij}^+ - z_{ij}$, and $\hat{F}_{0i}^{\pm}$ be the c.d.f. putting mass $z_{ij}^{\pm} / \sum_{i=1}^n z_{ij}^2$ at $r_i$, $i = 1, ..., n$. Then an L-estimator of $\beta$ is defined to be

$$\hat{\beta}_L = \hat{\beta} + (A^+ - A^-)(Z^\tau Z)^{-1}, \tag{5.76}$$

where

$$A^{\pm} = \left( \int x J(\hat{F}_{01}^{\pm}(x)) d\hat{F}_{01}^{\pm}(x) \sum_{i=1}^n z_{i1}^{\pm}, ..., \int x J(\hat{F}_{0p}^{\pm}(x)) d\hat{F}_{0p}^{\pm}(x) \sum_{i=1}^n z_{ip}^{\pm} \right).$$

Obviously, (5.76) reduces to (5.75) if $p = 1$ and $Z_i \geq 0$ for all $i$.

**Theorem 5.12.** Assume model (5.73) with i.i.d. $\varepsilon_i$'s from a c.d.f. $F_0$ satisfying (5.74) for a given $J$. Suppose that $F_0$ has a uniformly continuous, positive, and bounded derivative on the range of $\varepsilon_1$. Suppose further that the conditions on $Z_i$'s in Theorem 3.12 are satisfied.
(i) If the function $J$ is continuous on $(\alpha, \beta)$ and equals 0 on $[0, \alpha] \cup [\beta, 1]$, where $0 < \alpha < \beta < 1$ are constants, then

$$\sigma_{F_0}^{-1}(\hat{\beta}_L - \beta)(Z^\tau Z)^{1/2} \to_d N_p(0, I_p), \tag{5.77}$$

where $\sigma_{F_0}^2$ is given by (5.72) with $F = F_0$.
(ii) Result (5.77) also holds if $J'$ is bounded on $[0, 1]$, $E|\varepsilon_1| < \infty$, and $\sigma_{F_0}^2$ is finite. ∎

The proof of this theorem can be found in Bickel (1973). Robustness and efficiency comparisons between the LSE $\hat{\beta}$ and L-estimators $\hat{\beta}_L$ can be made in a similar way to those in §5.3.2.

## 5.4 Generalized Estimating Equations

The method of *generalized estimating equations* (GEE) is a powerful and general method of deriving point estimators, which includes many previously described methods as special cases. In §5.4.1, we begin with a description of this method and, to motivate the idea, we discuss its relationship with other methods that have been studied. Consistency and asymptotic normality of estimators derived from generalized estimating equations are studied in §5.4.2 and §5.4.3.

Throughout this section we assume that $X_1, ..., X_n$ are independent (not necessarily identically distributed) random vectors, where the dimension of $X_i$ is $d_i$, $i = 1, ..., n$ ($\sup_i d_i < \infty$), and that we are interested in estimating $\theta$, a $k$-vector of unknown parameters related to the unknown population.

### 5.4.1 The GEE method and its relationship with others

The sample mean and, more generally, the LSE in linear models are solutions of equations of the form

$$\sum_{i=1}^{n}(X_i - \gamma Z_i^\tau)Z_i = 0.$$

Also, MLE's (or RLE's) in §4.4 and, more generally, M-estimators in §5.2.2 are solutions of equations of the form

$$\sum_{i=1}^{n}\psi(X_i, \gamma) = 0.$$

This leads to the following general estimation method. Let $\Theta \subset \mathcal{R}^k$ be the range of $\theta$, $\psi_i$ be a Borel function from $\mathcal{R}^{d_i} \times \Theta$ to $\mathcal{R}^k$, $i = 1, ..., n$, and

$$s_n(\gamma) = \sum_{i=1}^{n}\psi_i(X_i, \gamma), \qquad \gamma \in \Theta. \tag{5.78}$$

If $\theta$ is estimated by $\hat{\theta} \in \Theta$ satisfying $s_n(\hat{\theta}) = 0$, then $\hat{\theta}$ is called a GEE estimator. The equation $s_n(\gamma) = 0$ is called a GEE. Apparently, the LSE's, RLE's, MQLE's, and M-estimators are special cases of GEE estimators.

Usually GEE's are chosen so that

$$E[s_n(\theta)] = \sum_{i=1}^{n}E[\psi_i(X_i, \theta)] = 0, \tag{5.79}$$

where the expectation $E$ may be replaced by an asymptotic expectation defined in §2.5.2 if the exact expectation does not exist. If this is true, then $\hat{\theta}$ is motivated by the fact that $s_n(\hat{\theta}) = 0$ is a sample analogue of $E[s_n(\theta)] = 0$.

To motivate the idea, let us study the relationship between the GEE method and other methods that have been introduced.

#### M-estimators

The M-estimators defined in §5.2.2 for unvariate $\theta = \mathtt{T}(F)$ in the i.i.d. case are special cases of GEE estimators. Huber (1981) also considers regression M-estimators in the linear model (5.73). A regression M-estimator of $\beta$ is defined as a solution to the GEE

$$\sum_{i=1}^{n}\psi(X_i - \gamma Z_i^\tau)Z_i = 0,$$

where $\psi$ is one of the functions given in Example 5.7.

**LSE's in linear and nonlinear regression models**

Suppose that

$$X_i = f(Z_i, \theta) + \varepsilon_i, \qquad i = 1, ..., n, \tag{5.80}$$

where $Z_i$'s are the same as those in (5.73), $\theta$ is an unknown $k$-vector of parameters, $f$ is a known function, and $\varepsilon_i$'s are independent random variables. Model (5.80) is the same as model (5.73) if $f$ is linear in $\theta$ and is called a *nonlinear regression model* otherwise. Note that model (4.64) is a special case of model (5.80). The LSE under model (5.80) is any point in $\Theta$ minimizing $\sum_{i=1}^n [X_i - f(Z_i, \gamma)]^2$ over $\gamma \in \Theta$. If $f$ is differentiable, then the LSE is a soluation to the GEE

$$\sum_{i=1}^n [X_i - f(Z_i, \gamma)] \frac{\partial f(Z_i, \gamma)}{\partial \gamma} = 0.$$

**Quasi-likelihoods**

This is a continuation of the discussion of the quasi-likelihoods introduced in §4.4.3. Assume first that $X_i$'s are univariate ($d_i \equiv 1$). If $X_i$'s follow a GLM, i.e., $X_i$ has the p.d.f. in (4.55) and (4.57) holds, and if (4.58) holds, then the likelihood equation (4.59) can be written as

$$\sum_{i=1}^n \frac{x_i - \mu_i(\gamma)}{v_i(\gamma)} G_i(\gamma) = 0, \tag{5.81}$$

where $\mu_i(\gamma) = \mu(\psi(\gamma Z_i^\tau))$, $G_i(\gamma) = \partial \mu_i(\gamma)/\partial \gamma$, $v_i(\gamma) = \mathrm{Var}(X_i)/\phi$, and we have used the following fact:

$$\psi'(t) = (\mu^{-1})'(g^{-1}(t))(g^{-1})'(t) = [\zeta''(\psi(t))]^{-1}(g^{-1})'(t).$$

Equation (5.81) is a quasi-likelihood equation if either $X_i$ does not have the p.d.f. in (4.55) or (4.58) does not hold. Note that this generalizes the discussion in §4.4.3: if $X_i$ does not have the p.d.f. in (4.55), then the problem is often nonparametric. Let $s_n(\gamma)$ be the left-hand side of (5.81). Then $s_n(\gamma) = 0$ is a GEE and $E[s_n(\beta)] = 0$ is satisfied as long as the first condition in (4.56), $E(X_i) = \mu_i(\beta)$, is satisfied.

For general $d_i$'s, let $X_i = (X_{i1}, ..., X_{id_i})$, $i = 1, ..., n$, where each $X_{it}$ satisfies (4.56) and (4.57), i.e.,

$$E(X_{it}) = \mu(\eta_{it}) = g^{-1}(\beta Z_{it}^\tau) \quad \text{and} \quad \mathrm{Var}(X_{it}) = \phi_i \mu'(\eta_{it}),$$

and $Z_{it}$'s are $k$-vector values of covariates. In biostatistics and life-time testing problems, components of $X_i$ are *repeated measurements* at different times from subject $i$ and are called *longitudinal data*. Although $X_i$'s are

assumed independent, $X_{it}$'s are likely to be dependent for each $i$. Let $R_i$ be the $d_i \times d_i$ correlation matrix whose $(t,l)$th element is the correlation coefficient between $X_{it}$ and $X_{il}$. Then

$$\mathrm{Var}(X_i) = \phi_i [D_i(\beta)]^{1/2} R_i [D_i(\beta)]^{1/2}, \tag{5.82}$$

where $D_i(\gamma)$ is the $d_i \times d_i$ diagonal matrix with the $t$th diagonal element $g^{-1}(\gamma Z_i^\tau)$. If $R_i$'s in (5.82) are known, then an extension of (5.81) to the multivariate $x_i$'s is

$$\sum_{i=1}^{n} [x_i - \mu_i(\gamma)] \{ [D_i(\gamma)]^{1/2} R_i [D_i(\gamma)]^{1/2} \}^{-1} G_i(\gamma) = 0, \tag{5.83}$$

where $\mu_i(\gamma) = (\mu(\psi(\gamma Z_{i1}^\tau)), ..., \mu(\psi(\gamma Z_{in_i}^\tau)))$ and $G_i(\gamma) = \partial \mu_i(\gamma)/\partial \gamma$. In most applications, $R_i$ is unknown and its form is hard to model. Let $\tilde{R}_i$ be a known correlation matrix (called a *working correlation matrix*). Replacing $R_i$ in (5.83) by $\tilde{R}_i$ leads to the quasi-likelihood equation

$$\sum_{i=1}^{n} [x_i - \mu_i(\gamma)] \{ [D_i(\gamma)]^{1/2} \tilde{R}_i [D_i(\gamma)]^{1/2} \}^{-1} G_i(\gamma) = 0. \tag{5.84}$$

For example, we may assume that the components of $X_i$ are independent and take $\tilde{R}_i = I_{d_i}$. Although the working correlation matrix $\tilde{R}_i$ may not be the same as the true unknown correlation matrix $R_i$, an MQLE obtained from (5.84) is still consistent and asymptotically normal (§5.4.2 and §5.4.3). Of course, MQLE's are asymptotically more efficient if $\tilde{R}_i$ is closer to $R_i$. Even if $\tilde{R}_i = R_i$ and $\phi_i \equiv \phi$, (5.84) is still a quasi-likelihood equation, since the covariance matrix of $X_i$ cannot determine the distribution of $X_i$ unless $X_i$ is normal.

Since an $\tilde{R}_i$ closer to $R_i$ results in a better MQLE, sometimes it is suggested to replace $\tilde{R}_i$ in (5.84) by $\hat{R}_i$, an estimator of $R_i$ (Liang and Zeger, 1986). The resulting equation is called a *pseudo-likelihood equation*. As long as $\max_{i \le n} \|\hat{R}_i - U_i\| \to_p 0$ as $n \to \infty$, where $U_i$'s are correlation matrices (not necessarily the same as the true correlation matrices), MQLE's are consistent and asymptotically normal.

### Profile empirical likelihoods

The previous discussion shows that the GEE method coincides with the method of deriving M-estimators, LSE's, MLE's, or MQLE's. The following discussion indicates that the GEE method is also closely related to the method of empirical likelihoods introduced in §5.1.2.

Assume that $X_i$'s are i.i.d. from a c.d.f. $F$ on $\mathcal{R}^d$ and $\psi_i = \psi$ for all $i$. Then condition (5.79) reduces to $E[\psi(X_1, \theta)] = 0$. This leads to the

consideration of the empirical likelihood

$$\ell(F) = \prod_{i=1}^{n} p_i \quad \text{subject to} \quad p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \psi(x_i, \theta) = 0,$$

where $p_i = P_F(\{x_i\})$, $i = 1, ..., n$. Maximizing this empirical likelihood is equivalent to maximizing

$$\ell(p_1, ..., p_n, \omega, \lambda, \theta) = \prod_{i=1}^{n} p_i + \omega \left(1 - \sum_{i=1}^{n} p_i\right) + \sum_{i=1}^{n} p_i \psi(x_i, \theta) \lambda^\tau,$$

where $\omega$ and $\lambda$ are Lagrange multipliers.

Suppose that $\ell(\theta, \xi)$ is a likelihood (or empirical likelihood) function, where $\theta$ and $\xi$ are not necessarily vector-valued. For example, $\xi = G$, a c.d.f. If maximizing $\ell(\theta, \xi)$ over $(\theta, \xi)$ is difficult, sometimes we can apply the method of *profile likelihoods*, which can be described as follows. For each fixed $\theta$, let $\xi(\theta)$ satisfy

$$\ell(\theta, \xi(\theta)) = \sup_{\xi} \ell(\theta, \xi).$$

The function

$$\ell_P(\theta) = \ell(\theta, \xi(\theta))$$

is called a *profile likelihood* function for $\theta$. Suppose that $\hat{\theta}_P$ maximizes $\ell_P(\theta)$. Then $\hat{\theta}_P$ is called a maximum profile likelihood estimator of $\theta$. Note that $\hat{\theta}_P$ may be different from an MLE of $\theta$.

In general, it is difficult to maximize the likelihood $\ell(p_1, ..., p_n, \omega, \lambda, \theta)$. We consider the following profile empirical likelihood. Let $\theta$ be fixed. It follows from (5.12) and (5.13) that

$$\omega = n, \qquad \tilde{p}_i = \{n[1 + \psi(x_i, \theta)\lambda_n^\tau]\}^{-1}$$

with a $\lambda_n = \lambda_n(\theta)$ satisfying

$$\sum_{i=1}^{n} \frac{\psi(x_i, \theta)}{n\{1 + \psi(x_i, \theta)[\lambda_n(\theta)]^\tau\}} = 0 \tag{5.85}$$

maximize $\ell(p_1, ...p_n, \omega, \lambda, \theta)$ for any fixed $\theta$. Substituting $\tilde{p}_i$ with $\sum_{i=1}^{n} \tilde{p}_i = 1$ into $\ell(p_1, ...p_n, \omega, \lambda, \theta)$ leads to the following profile empirical likelihood for $\theta$:

$$\ell_P(\theta) = \prod_{i=1}^{n} \frac{1}{n\{1 + \psi(x_i, \theta)[\lambda_n(\theta)]^\tau\}}. \tag{5.86}$$

Let $\hat{\theta}$ be a maximum of $\ell_P(\theta)$ in (5.86). It follows from the proof of Theorem 5.4 that

$$\lambda_n(\hat{\theta}) = O_p(n^{-1/2}) \qquad \text{and} \qquad \hat{\theta} - \theta = O_p(n^{-1/2})$$

(see also Qin and Lawless (1994)), under some conditions on $\psi$ and its derivatives. Using (5.85), we obtain that

$$0 = \sum_{i=1}^{n} \frac{\psi(x_i, \hat{\theta})}{1 + \psi(x_i, \hat{\theta})[\lambda_n(\hat{\theta})]^\tau} = \left[1 + O_p(n^{-1/2})\right] \sum_{i=1}^{n} \psi(x_i, \hat{\theta})$$

and, therefore,

$$P\left(\sum_{i=1}^{n} \psi(X_i, \hat{\theta}) = 0\right) \to 1.$$

That is, $\hat{\theta}$ is a GEE estimator.

### 5.4.2 Consistency of GEE estimators

We now study under what conditions (besides (5.79)) GEE estimators are consistent. For each $n$, let $\hat{\theta}_n$ be a GEE estimator, i.e., $s_n(\hat{\theta}_n) = 0$, where $s_n(\gamma)$ is defined by (5.78).

First, Theorem 5.7 and its proof can be extended to multivariate T in a straightforward manner. Hence, we have the following result.

**Proposition 5.2.** Suppose that $X_1, ..., X_n$ are i.i.d. from $F$ and $\psi_i \equiv \psi$, a bounded and continuous function from $\mathcal{R}^d \times \Theta$ to $\mathcal{R}^k$. Let $\Psi(t) = \int \psi(x,t)dF(x)$. Suppose that $\Psi(\theta) = 0$ and $\partial\Psi(t)/\partial t$ exists and is of full rank at $t = \theta$. Then $\hat{\theta}_n \to_p \theta$. ∎

For unbounded $\psi$ in the i.i.d. case, the following result and its proof can be found in Qin and Lawless (1994).

**Proposition 5.3.** Suppose that $X_1, ..., X_n$ are i.i.d. from $F$ and $\psi_i \equiv \psi$. Assume that $\varphi(x,\gamma) = \partial\psi(x,\gamma)/\partial\gamma$ exists in $N_\theta$, a neighborhood of $\theta$, and is continuous at $\theta$; there is a function $h(x)$ such that $\sup_{\gamma \in N_\theta} \|\varphi(x,\gamma)\| \le h(x)$, $\sup_{\gamma \in N_\theta} \|\psi(x,\gamma)\|^3 \le h(x)$, and $E[h(X_1)] < \infty$; $E[\varphi(X_1,\theta)]$ is of full rank; $E\{[\psi(X_1,\theta)]^\tau \psi(X_1,\theta)\}$ is positive definite; and (5.79) holds. Then, there exists a sequence of random vectors $\{\hat{\theta}_n\}$ such that

$$P\left(s_n(\hat{\theta}_n) = 0\right) \to 1 \qquad \text{and} \qquad \hat{\theta}_n \to_p \theta. \quad ∎ \tag{5.87}$$

Next, we consider non-i.i.d. $X_i$'s.

**Proposition 5.4.** Suppose that $X_1, ..., X_n$ are independent and $\theta$ is univariate. Assume that $\psi_i(x, \gamma)$ is real-valued and nonincreasing in $\gamma$ for all $i$; there is a $\delta > 0$ such that $\sup_i E|\psi_i(X_i, \gamma)|^{1+\delta} < \infty$ for any $\gamma$ in $N_\theta$, a neighborhood of $\theta$ (this condition can be replaced by $E|\psi(X_1, \gamma)| < \infty$ for any $\gamma$ in $N_\theta$ when $X_i$'s are i.i.d. and $\psi_i \equiv \psi$); $\psi_i(x, \gamma)$ are continuous in $N_\theta$; (5.79) holds; and

$$\limsup_n E[\Psi_n(\theta + \epsilon)] < 0 < \liminf_n E[\Psi_n(\theta - \epsilon)] \tag{5.88}$$

for any $\epsilon > 0$, where $\Psi_n(\gamma) = n^{-1}s_n(\gamma)$. Then, there exists a sequence of random variables $\{\hat{\theta}_n\}$ such that (5.87) holds. Furthermore, any sequence $\{\hat{\theta}_n\}$ satisfying $s_n(\hat{\theta}_n) = 0$ satisfies (5.87).

**Proof.** Since $\psi_i$'s are nonincreasing, the functions $\Psi_n(\gamma)$ and $E[\Psi_n(\gamma)]$ are nonincreasing. Let $\epsilon > 0$ be fixed so that $\theta \pm \epsilon \in N_\theta$. Under the assumed conditions,

$$\Psi_n(\theta \pm \epsilon) - E[\Psi_n(\theta \pm \epsilon)] \to_p 0$$

(Theorem 1.14(ii)). By condition (5.88),

$$P\big(\Psi_n(\theta + \epsilon) < 0 < \Psi_n(\theta - \epsilon)\big) \to 1.$$

The rest of the proof is left as an exercise. ∎

To establish the next result, we need the following lemma. First, we need the following concept. A sequence of functions $\{g_i\}$ from $\mathcal{R}^k$ to $\mathcal{R}^k$ is called *equicontinuous* on an open set $\mathcal{O} \subset \mathcal{R}^k$ if and only if for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that $\sup_i \|g_i(t) - g_i(s)\| < \epsilon$ whenever $t \in \mathcal{O}$, $s \in \mathcal{O}$, and $\|t - s\| < \delta_\epsilon$. Since a continuous function on a compact set is uniformly continuous, functions such as $g_i(\gamma) = g(t_i, \gamma)$ form an equicontinuous sequence on $\mathcal{O}$ if $t_i$'s vary in a compact set containing $\mathcal{O}$ and $g(t, \gamma)$ is a continuous function in $(t, \gamma)$.

**Lemma 5.3.** Suppose that $\Theta$ is a compact subset of $\mathcal{R}^k$. Let $h_i(X_i) = \sup_{\gamma \in \Theta} \|\psi_i(X_i, \gamma)\|$, $i = 1, 2, ....$ Suppose that $\sup_i E|h_i(X_i)|^{1+\delta} < \infty$ and $\sup_i E\|X_i\|^\delta < \infty$ for some $\delta > 0$ (this condition can be replaced by $E|h(X_1)| < \infty$ when $X_i$'s are i.i.d. and $\psi_i \equiv \psi$). Suppose further that for any $c > 0$ and sequence $\{x_i\}$ satisfying $\|x_i\| \le c$, the sequence of functions $\{g_i(\gamma) = \psi_i(x_i, \gamma)\}$ is equicontinuous on any open subset of $\Theta$. Then

$$\sup_{\gamma \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n \{\psi_i(X_i, \gamma) - E[\psi_i(X_i, \gamma)]\} \right\| \to_p 0.$$

**Proof.** Since we only need to consider components of $\psi_i$'s, without loss of generality we can assume that $\psi_i$'s are functions from $\mathcal{R}^{d_i} \times \Theta$ to $\mathcal{R}$. For

any $c > 0$,

$$\sup_n E\left[\frac{1}{n}\sum_{i=1}^n h_i(X_i)I_{(c,\infty)}(\|X_i\|)\right] \le \sup_i E[h_i(X_i)I_{(c,\infty)}(\|X_i\|)].$$

Let $c_0 = \sup_i E|h_i(X_i)|^{1+\delta}$ and $c_1 = \sup_i E\|X_i\|^\delta$. By Hölder's inequality,

$$\begin{aligned} E[h_i(X_i)I_{(c,\infty)}(\|X_i\|)] &\le \left[E|h_i(X_i)|^{1+\delta}\right]^{1/(1+\delta)}\left[P(\|X_i\| > c)\right]^{\delta/(1+\delta)} \\ &\le c_0^{1/(1+\delta)} c_1^{\delta/(1+\delta)} c^{-\delta^2/(1+\delta)} \end{aligned}$$

for all $i$. For $\epsilon > 0$ and $\tilde{\epsilon} > 0$, choose a $c$ such that $c_0^{1/(1+\delta)} c_1^{\delta/(1+\delta)} c^{-\delta^2/(1+\delta)}$ $< \epsilon\tilde{\epsilon}/2$. Then, for any $\mathcal{O} \subset \Theta$, the probability

$$P\left(\frac{1}{n}\sum_{i=1}^n \left\{\sup_{\gamma\in\mathcal{O}} \psi_i(X_i,\gamma) - \inf_{\gamma\in\mathcal{O}} \psi_i(X_i,\gamma)\right\} I_{(c,\infty)}(\|X_i\|) > \frac{\epsilon}{2}\right) \tag{5.89}$$

is bounded by $\tilde{\epsilon}$ (exercise). From the equicontinuity of $\{\psi_i(x_i,\gamma)\}$, there is a $\delta_\epsilon > 0$ such that

$$\frac{1}{n}\sum_{i=1}^n \left\{\sup_{\gamma\in\mathcal{O}_\epsilon} \psi_i(X_i,\gamma) - \inf_{\gamma\in\mathcal{O}_\epsilon} \psi_i(X_i,\gamma)\right\} I_{[0,c]}(\|X_i\|) < \frac{\epsilon}{2}$$

for sufficiently large $n$, where $\mathcal{O}_\epsilon$ denotes any open ball in $\mathcal{R}^k$ with radius less than $\delta_\epsilon$. These results, together with Theorem 1.14(ii) and the fact that $\|\psi_i(X_i,\gamma)\| \le h_i(X_i)$, imply that

$$P\left(\frac{1}{n}\sum_{i=1}^n \left\{\sup_{\gamma\in\mathcal{O}_\epsilon} \psi_i(X_i,\gamma) - E\left[\inf_{\gamma\in\mathcal{O}_\epsilon} \psi_i(X_i,\gamma)\right]\right\} > \epsilon\right) \to 0. \tag{5.90}$$

Let $H_n(\gamma) = n^{-1}\sum_{i=1}^n \{\psi_i(X_i,\gamma) - E[\psi_i(X_i,\gamma)]\}$. Then

$$\sup_{\gamma\in\mathcal{O}_\epsilon} H_n(\gamma) \le \frac{1}{n}\sum_{i=1}^n \left\{\sup_{\gamma\in\mathcal{O}_\epsilon} \psi_i(X_i,\gamma) - E\left[\inf_{\gamma\in\mathcal{O}_\epsilon} \psi_i(X_i,\gamma)\right]\right\},$$

which with (5.90) implies that

$$P\big(H_n(\gamma) > \epsilon \text{ for all } \gamma \in \mathcal{O}_\epsilon\big) = P\left(\sup_{\gamma\in\mathcal{O}_\epsilon} H_n(\gamma) > \epsilon\right) \to 0.$$

Similarly we can show that

$$P\big(H_n(\gamma) < -\epsilon \text{ for all } \gamma \in \mathcal{O}_\epsilon\big) \to 0.$$

Since $\Theta$ is compact, there exists $m_\epsilon$ open balls $\mathcal{O}_{\epsilon,j}$ such that $\Theta \subset \cup \mathcal{O}_{\epsilon,j}$. Then, the result follows from

$$P\left(\sup_{\gamma\in\Theta} |H_n(\gamma)| > \epsilon\right) \le \sum_{j=1}^{m_\epsilon} P\left(\sup_{\gamma\in\mathcal{O}_{\epsilon,j}} |H_n(\gamma)| > \epsilon\right) \to 0. \quad \blacksquare$$

**Example 5.11.** Consider the quasi-likelihood equation (5.84). Let $\{\tilde{R}_i\}$ be a sequence of working correlation matrices and

$$\psi_i(x_i, \gamma) = [x_i - \mu_i(\gamma)]\{[D_i(\gamma)]^{1/2}\tilde{R}_i[D_i(\gamma)]^{1/2}\}^{-1}G_i(\gamma). \tag{5.91}$$

It can be shown (exercise) that $\psi_i$'s satisfy the conditions of Lemma 5.3 if $\Theta$ is compact and $\sup_i \|Z_i\| < \infty$. $\blacksquare$

**Proposition 5.5.** Assume (5.79) and the conditions in Lemma 5.3. Suppose that the functions $\Delta_n(\gamma) = E[n^{-1}s_n(\gamma)]$ have the property that $\lim_{n\to\infty} \Delta_n(\gamma) = 0$ if and only if $\gamma = \theta$. (If $\Delta_n$ converges to a function $\Delta$, then this condition and (5.79) means that $\Delta$ has a unique 0 at $\theta$.) Suppose that $\{\hat{\theta}_n\}$ is a sequence of GEE estimators and that $\hat{\theta}_n = O_p(1)$. Then $\hat{\theta}_n \to_p \theta$.

**Proof.** First, assume that $\Theta$ is a compact subset of $\mathcal{R}^k$. From Lemma 5.3 and $s_n(\hat{\theta}_n) = 0$, $\Delta_n(\hat{\theta}_n) \to_p 0$. By Theorem 1.8(vi), there is a subsequence $\{n_i\}$ such that

$$\Delta_{n_i}(\hat{\theta}_{n_i}) \to_{a.s.} 0. \tag{5.92}$$

Let $x_1, x_2, ...$ be a fixed sequence such that (5.92) holds and let $\theta_0$ be a limit point of $\{\hat{\theta}_n\}$. Since $\Theta$ is compact, $\theta_0 \in \Theta$ and there is a subsequence $\{m_j\} \subset \{n_i\}$ such that $\hat{\theta}_{m_j} \to \theta_0$. Using the argument in the proof of Lemma 5.3, it can be shown (exercise) that $\{\Delta_n(\gamma)\}$ is equicontinuous on any open subset of $\Theta$. Then

$$\Delta_{m_j}(\theta_{m_j}) - \Delta_{m_j}(\theta_0) \to 0,$$

which with (5.92) implies $\Delta_{m_j}(\theta_0) \to 0$. Under the assumed condition, $\theta_0 = \theta$. Since this is true for any limit point of $\{\hat{\theta}_n\}$, $\hat{\theta}_n \to_p \theta$.

Next, consider a general $\Theta$. For any $\epsilon > 0$, there is an $M_\epsilon > 0$ such that $P(\|\hat{\theta}_n\| \le M_\epsilon) > 1 - \epsilon$. The result follows from the previous proof by considering the closure of $\Theta \cap \{\gamma : \|\gamma\| \le M_\epsilon\}$ as the parameter space. $\blacksquare$

Condition $\hat{\theta}_n = O_p(1)$ in Proposition 5.5 is obviously necessary for the consistency of $\hat{\theta}_n$. It has to be checked in any particular problem.

If a GEE is a likelihood equation under some conditions, then we can often show, using a similar argument to the proof of Theorem 4.17 or 4.18, that there exists a consistent sequence of GEE estimators.

**Proposition 5.6.** Suppose that $s_n(\gamma) = \partial \log \ell_n(\gamma)/\partial\gamma$ for some function $\ell_n$; $I_n(\theta) = \mathrm{Var}(s_n(\theta)) \to 0$; $\varphi_i(x,\gamma) = \partial\psi_i(x,\gamma)/\partial\gamma$ exists and the sequence of functions $\{\varphi_{ij}, i = 1, 2, ...\}$ satisfies the conditions in Lemma 5.3 with $\Theta$ replaced by a compact neighborhood of $\theta$, where $\varphi_{ij}$ is the $j$th row of $\varphi_i$, $j = 1, ..., k$; and $-\liminf_n [I_n(\theta)]^{1/2} E[\nabla s_n(\theta)][I_n(\theta)]^{1/2}$ is positive definite, where $\nabla s_n(\gamma) = \partial s_n(\gamma)/\partial\gamma$. Then, there exists a sequence of estimators $\{\hat{\theta}_n\}$ satisfying (5.87). ∎

The proof of Proposition 5.6 is similar to that of Theorem 4.17 or Theorem 4.18 and is left as an exercise.

**Example 5.12.** Consider the quasi-likelihood equation (5.84) with $\tilde{R}_i = I_{d_i}$ for all $i$. Then the GEE is a likelihood equation under a GLM (§4.4.2) assumption. It can be shown (exercise) that the conditions of Proposition 5.6 are satisfied if $\sup_i \|Z_i\| < \infty$. ∎

### 5.4.3 Asymptotic normality of GEE estimators

Asymptotic normality of a consistent sequence of GEE estimators can be established under some conditions. We first consider the special case where $\theta$ is univariate and $X_1, ..., X_n$ are i.i.d.

**Theorem 5.13.** Let $X_1, ..., X_n$ be i.i.d. from $F$, $\psi_i \equiv \psi$, and $\theta \in \mathcal{R}$. Suppose that $\Psi(\gamma) = \int \psi(x,\gamma) dF(x) = 0$ if and only if $\gamma = \theta$, $\Psi'(\theta)$ exists and $\Psi'(\theta) \neq 0$.
(i) Assume that $\psi(x,\gamma)$ is nonincreasing in $\gamma$ and that $\int [\psi(x,\gamma)]^2 dF(x)$ is finite for $\gamma$ in a neighborhood of $\theta$ and is continuous at $\theta$. Then, any sequence of GEE estimators (M-estimators) $\{\hat{\theta}_n\}$ satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) \to_d N(0, \sigma_F^2), \tag{5.93}$$

where

$$\sigma_F^2 = \int [\psi(x,\theta)]^2 dF(x) / [\Psi'(\theta)]^2.$$

(ii) Assume that $\int [\psi(x,\theta)]^2 dF(x) < \infty$, $\psi(x,\gamma)$ is continuous in $x$, and $\lim_{\gamma\to\theta} \|\psi(\cdot,\gamma) - \psi(\cdot,\theta)\|_V = 0$, where $\|\cdot\|_V$ is the variation norm defined in Lemma 5.2. Then, any consistent sequence of GEE estimators $\{\hat{\theta}_n\}$ satisfies (5.93).
**Proof.** (i) Let $\Psi_n(\gamma) = n^{-1} s_n(\gamma)$. Since $\Psi_n$ is nonincreasing,

$$P(\Psi_n(t) < 0) \le P(\hat{\theta}_n \le t) \le P(\Psi_n(t) \le 0)$$

for any $t \in \mathcal{R}$. Then, (5.93) follows from

$$\lim_{n\to\infty} P\big(\Psi_n(t_n) < 0\big) = \lim_{n\to\infty} P\big(\Psi_n(t_n) \le 0\big) = \Phi(t)$$

for all $t \in \mathcal{R}$, where $t_n = \theta + t\sigma_F n^{-1/2}$. Let $s_{t,n}^2 = \text{Var}(\psi(X_1, t_n))$ and $Y_{ni} = [\psi(X_i, t_n) - \Psi(t_n)]/s_{t,n}$. Then, it suffices to show that

$$\lim_{n\to\infty} P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_{ni} \le -\frac{\sqrt{n}\Psi(t_n)}{s_{t,n}}\right) = \Phi(t)$$

for all $t$. Under the assumed conditions, $\sqrt{n}\Psi(t_n) \to \Psi'(\theta)t\sigma_F$ and $s_{t,n} \to -\Psi'(\theta)\sigma_F$. Hence, it suffices to show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_{ni} \to_d N(0,1).$$

Note that $Y_{n1}, ..., Y_{nn}$ are i.i.d. random variables. Hence we can apply Lindeberg's CLT (Theorem 1.15). In this case, Lindeberg's condition (1.53) is implied by

$$\lim_{n\to\infty} \int_{|\psi(x,t_n)|>\sqrt{n}\epsilon} [\psi(x,t_n)]^2 dF(x) = 0$$

for any $\epsilon > 0$. For any $\eta > 0$, $\psi(x, \theta+\eta) \le \psi(x, t_n) \le \psi(x, \theta-\eta)$ for all $x$ and sufficiently large $n$. Let $u(x) = \max\{|\psi(x, \theta-\eta)|, |\psi(x, \theta+\eta)|\}$. Then

$$\int_{|\psi(x,t_n)|>\sqrt{n}\epsilon} [\psi(x,t_n)]^2 dF(x) \le \int_{u(x)>\sqrt{n}\epsilon} [u(x)]^2 dF(x),$$

which converges to 0 since $\int [\psi(x,\gamma)]^2 dF(x)$ is finite for $\gamma$ in a neighborhood of $\theta$. This proves (i).

(ii) Let $\phi_F(x) = -\psi(x,\theta)/\Psi'(\theta)$. Following the proof of Theorem 5.7, we have

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi_F(X_i) + R_{1n} - R_{2n},$$

where

$$R_{1n} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi(X_i, \theta)\left[\frac{1}{\Psi'(\theta)} - \frac{1}{h_F(\hat{\theta}_n)}\right],$$

$$R_{2n} = \frac{\sqrt{n}}{h_F(\hat{\theta}_n)} \int [\psi(x, \hat{\theta}_n) - \psi(x,\theta)] d(F_n - F)(x),$$

and $h_F$ is defined in the proof of Theorem 5.7 with $\Psi = \lambda_F$. By the CLT and the consistency of $\hat{\theta}_n$, $R_{1n} = o_p(1)$. Hence the result follows if we can show that $R_{2n} = o_p(1)$. By Lemma 5.2,

$$|R_{2n}| \le \sqrt{n}|h_F(\hat{\theta}_n)|^{-1} \varrho_\infty(F_n, F) \|\psi(\cdot, \hat{\theta}_n) - \psi(\cdot,\theta)\|_V.$$

The result follows from the assumed condition on $\psi$ and the fact that $\sqrt{n}\varrho_\infty(F_n, F) = O_p(1)$ (Theorem 5.1). ∎

Note that the result in Theorem 5.13 coincides with the result in Theorem 5.7 and (5.34).

**Example 5.13.** Consider the M-estimators given in Example 5.7 based on i.i.d. random variables $X_1, ..., X_n$. If $\psi$ is bounded and continuous, then Theorem 5.7 applies and (5.93) holds. For case (ii), $\psi(x, \gamma)$ is not bounded but is nondecreasing in $\gamma$ ($-\psi(x, \gamma)$ is nonincreasing in $\gamma$). Hence Theorem 5.13 can be applied to this case.

Consider Huber's $\psi$ given in (v). Assume that $F$ is differentiable in neighborhoods of $\theta - C$ and $\theta + C$. Then

$$\Psi(\gamma) = \int_{\gamma-C}^{\gamma+C} (\gamma - x)dF(x) - CF(\gamma - C) + C[1 - F(\gamma + C)]$$

is differentiable at $\theta$ (exercise); $\Psi(\theta) = 0$ if $F$ is symmetric about $\theta$ (exercise); and

$$\int [\psi(x, \gamma)]^2 dF(x) = \int_{\gamma-C}^{\gamma+C} (\gamma - x)^2 dF(x) + C^2 F(\gamma - C) + C^2[1 - F(\gamma + C)]$$

is continuous at $\theta$ (exercise). Therefore, (5.93) holds with

$$\sigma_F^2 = \frac{\int_{\theta-C}^{\theta+C} (\theta - x)^2 dF(x) + C^2 F(\theta - C) + C^2[1 - F(\theta + C)]}{[F(\theta + C) - F(\theta - C)]^2}$$

(exercise). Note that Huber's M-estimator is robust in Hampel's sense. Asymptotic relative efficiency of $\hat{\theta}_n$ w.r.t. the sample mean $\bar{X}$ can be obtained (exercise). ∎

The next result is for general $\theta$ and independent $X_i$'s.

**Theorem 5.14.** Suppose that $\varphi_i(x, \gamma) = \partial\psi_i(x, \gamma)/\partial\gamma$ exists and the sequence of functions $\{\varphi_{ij}, i = 1, 2, ...\}$ satisfies the conditions in Lemma 5.3 with $\Theta$ replaced by a compact neighborhood of $\theta$, where $\varphi_{ij}$ is the $j$th row of $\varphi_i$; $\sup_i E\|\psi_i(X_i, \theta)\|^{2+\delta} < \infty$ for some $\delta > 0$ (this condition can be replaced by $E\|\psi(X_1, \theta)\|^2 < \infty$ if $X_i$'s are i.i.d. and $\psi_i \equiv \psi$); $E[\psi_i(X_i, \theta)] = 0$; $\liminf_n \lambda_-[n^{-1}\mathrm{Var}(s_n(\theta))] > 0$ and $\liminf_n \lambda_-[n^{-1}M_n(\theta)] > 0$, where $M_n(\theta) = -E[\nabla s_n(\theta)]$ and $\lambda_-[A]$ is the smallest eigenvalue of the matrix $A$. If $\{\hat{\theta}_n\}$ is a consistent sequence of GEE estimators, then

$$(\hat{\theta}_n - \theta)V_n^{-1/2} \to_d N_k(0, I_k), \tag{5.94}$$

where

$$V_n = [M_n(\theta)]^{-1}\mathrm{Var}(s_n(\theta))[M_n(\theta)]^{-1}. \tag{5.95}$$

**Proof.** The proof is similar to that of Theorem 4.17. By the consistency of $\hat{\theta}_n$, we can focus on the event $\{\hat{\theta}_n \in A_\epsilon\}$, where $A_\epsilon = \{\gamma : \|\gamma - \theta\| \le \epsilon\}$ with a given $\epsilon > 0$. For sufficiently small $\epsilon$, it can be shown (exercise) that

$$\max_{\gamma \in A_\epsilon} \frac{\|\nabla s_n(\gamma) - \nabla s_n(\theta)\|}{n} = o_p(1), \tag{5.96}$$

using a similar argument to the proof of Lemma 5.3. From the mean-value theorem and $s_n(\hat{\theta}_n) = 0$,

$$-s_n(\theta) = (\hat{\theta}_n - \theta) \int_0^1 \nabla s_n\big(\theta + t(\hat{\theta}_n - \theta)\big) dt.$$

It follows from (5.96) that

$$\frac{1}{n} \left\| \int_0^1 \nabla s_n\big(\theta + t(\hat{\theta}_n - \theta)\big) dt - \nabla s_n(\theta) \right\| = o_p(1).$$

Also, by Theorem 1.14(ii),

$$n^{-1}[\nabla s_n(\theta) + M_n(\theta)] = o_p(1).$$

This and $\liminf_n \lambda_-[n^{-1} M_n(\theta)] > 0$ imply

$$s_n(\theta)[M_n(\theta)]^{-1} = (\hat{\theta}_n - \theta)[1 + o_p(1)].$$

The result follows if we can show that

$$s_n(\theta)[M_n(\theta)]^{-1} V_n^{-1/2} \to_d N_k(0, I_k). \tag{5.97}$$

For any nonzero $l \in \mathcal{R}^k$,

$$\frac{1}{(l V_n l^\tau)^{1+\delta/2}} \sum_{i=1}^n E|\psi_i(X_i, \theta)[M_n(\theta)]^{-1} l^\tau|^{2+\delta} \to 0, \tag{5.98}$$

since $\liminf_n \lambda_-[n^{-1}\mathrm{Var}(s_n(\theta))] > 0$ and $\sup_i E\|\psi_i(X_i, \theta)\|^{2+\delta} < \infty$ (exercise). Applying the CLT (Theorem 1.15) with Liapunov's condition (5.98), we obtain that

$$s_n(\theta)[M_n(\theta)]^{-1} l^\tau / \sqrt{l V_n l^\tau} \to_d N(0, 1) \tag{5.99}$$

for any $l$, which implies (5.97) (exercise). ∎

Asymptotic normality of GEE estimators can be established under various other conditions; see, for example, Serfling (1980, Chapter 7) and He and Shao (1996).

If $X_i$'s are i.i.d. and $\psi_i \equiv \psi$, the asymptotic covariance matrix in (5.95) reduces to

$$V_n = n^{-1}\{E[\varphi(X_1,\theta)]\}^{-1}E\{[\psi(X_1,\theta)]^\tau[\psi(X_1,\theta)]\}\{E[\varphi(X_1,\theta)]\}^{-1},$$

where $\varphi(x,\gamma) = \partial\psi(x,\gamma)/\partial\gamma$. When $\theta$ is univariate, $V_n$ further reduces to

$$V_n = n^{-1}E[\psi(X_1,\theta)]^2/\{E[\varphi(X_1,\theta)]\}^2.$$

Under the conditions of Theorem 5.14,

$$E[\varphi(X_1,\theta)] = \int \frac{\partial\psi(x,\theta)}{\partial\theta}dF(x) = \frac{\partial}{\partial\theta}\int \psi(x,\theta)dF(x).$$

Hence, the result in Theorem 5.14 coincides with that in Theorem 5.13.

**Example 5.14.** Consider the quasi-likelihood equation in (5.84) and $\psi_i$ in (5.91). If $\sup_i \|Z_i\| < \infty$, then $\psi_i$ satisfies the conditions in Theorem 5.14 (exercise). Let $\tilde{V}_n(\gamma) = [D_i(\gamma)]^{1/2}\tilde{R}_i[D_i(\gamma)]^{1/2}$. Then

$$\mathrm{Var}(s_n(\theta)) = \sum_{i=1}^n [G_i(\theta)]^\tau[\tilde{V}_n(\theta)]^{-1}\mathrm{Var}(X_i)[\tilde{V}_n(\theta)]^{-1}G_i(\theta)$$

and

$$M_n(\theta) = \sum_{i=1}^n [G_i(\theta)]^\tau[\tilde{V}_n(\theta)]^{-1}G_i(\theta).$$

If $\tilde{R}_i = R_i$ (the true correlation matrix) for all $i$, then

$$\mathrm{Var}(s_n(\theta)) = \sum_{i=1}^n \phi_i[G_i(\theta)]^\tau[\tilde{V}_n(\theta)]^{-1}G_i(\theta).$$

If, in addition, $\phi_i \equiv \phi$, then

$$V_n = [M_n(\theta)]^{-1}\mathrm{Var}(s_n(\theta))[M_n(\theta)]^{-1} = \phi[M_n(\theta)]^{-1}. \quad \blacksquare$$

## 5.5 Variance Estimation

In statistical inference the accuracy of a point estimator is usually assessed by its mse or amse. If the bias or asymptotic bias of an estimator is (asymptotically) negligible w.r.t. its mse or amse, then assessing the mse or amse is equivalent to assessing variance or asymptotic variance. Since variances and asymptotic variances usually depend on the unknown population, we have to estimate them in order to report accuracies of point estimators. Variance estimation is an important part of statistical inference, not only for

assessing accuracy, but also for constructing inference procedures studied in Chapters 6 and 7. See also the discussion at the end of §2.5.1.

If the unknown population is in a parametric family indexed by a parameter $\theta$, then the covariance matrix or asymptotic covariance matrix of an estimator of $\theta$ is a function of $\theta$, say $V_n(\theta)$. If $\theta$ is estimated by $\hat{\theta}$, then it is natural to estimate $V_n(\theta)$ by $V_n(\hat{\theta})$. Thus, variance estimation in parametric problems is usually simple. This idea of substitution can be applied to nonparametric problems in which variance estimation is much more complex.

We introduce three commonly used variance estimation methods in this section, the substitution method, the jackknife, and the bootstrap.

### 5.5.1 The substitution method

Suppose that we can obtain a formula for the covariance or asymptotic covariance matrix of an estimator $\hat{\theta}_n$. Then a direct method of variance estimation is to substitute unknown quantities in the variance formula by some estimators. To illustrate, consider the simplest case where $X_1, ..., X_n$ are i.i.d. random $d$-vectors, $\hat{\theta}_n = g(\bar{X})$, and $g$ is a function from $\mathcal{R}^d$ to $\mathcal{R}^k$. Suppose that $E\|X_1\|^2 < \infty$ and $g$ is differentiable at $\mu = E(X_1)$. Then, by the CLT and Theorem 1.12(i),

$$(\hat{\theta}_n - \theta)V_n^{-1/2} \to_d N_k(0, I_k), \tag{5.100}$$

where $\theta = g(\mu)$ and

$$V_n = \nabla g(\mu)\mathrm{Var}(X_1)[\nabla g(\mu)]^\tau/n \tag{5.101}$$

is the asymptotic covariance matrix of $\hat{\theta}_n$ which depends on unknown quantities $\mu$ and $\mathrm{Var}(X_1)$. A substitution estimator of $V_n$ is

$$\hat{V}_n = \nabla g(\bar{X})S^2[\nabla g(\bar{X})]^\tau/n, \tag{5.102}$$

where $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^\tau(X_i - \bar{X})$ is the sample covariance matrix, an extension of the sample variance to the multivariate $X_i$'s.

An essential asymptotic requirement in variance estimation is the consistency of variance estimators according to the following definition.

**Definition 5.4.** Let $\{V_n\}$ be a sequence of $k \times k$ positive definite matrices and $\hat{V}_n$ be an estimator of $V_n$ for each $n$. Then $\{\hat{V}_n\}$ or $\hat{V}_n$ is said to be consistent for $V_n$ if and only if

$$l\hat{V}_n l^\tau / lV_n l \to_p 1 \qquad \text{for any } l \in \mathcal{R}^k. \tag{5.103}$$

$\hat{V}_n$ is strongly consistent if (5.103) holds with $\to_p$ replaced by $\to_{a.s.}$. ∎

If (5.103) holds, then $(\hat{\theta}_n - \theta)\hat{V}_n^{-1/2} \to_d N_k(0, I_k)$, a result useful for asymptotic inference as discussed in Chapters 6 and 7.

By the SLLN, $\hat{V}_n$ in (5.102) is strongly consistent for $V_n$ in (5.101), provided that $\nabla g(\mu) \neq 0$ and $\nabla g$ is continuous at $\mu$ so that $\nabla g(\bar{X}) \to_{a.s.} \nabla g(\mu)$.

**Example 5.15.** Let $Y_1, ..., Y_n$ be i.i.d. random variables with finite $\mu_y = EY_1$, $\sigma_y^2 = \text{Var}(Y_1)$, $\gamma_y = EY_1^3$, and $\kappa_y = EY_1^4$. Consider the estimation of $\theta = (\mu_y, \sigma_y^2)$. Let $\hat{\theta}_n = (\bar{X}, \hat{\sigma}_y^2)$, where $\hat{\sigma}_y^2 = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$. If $X_i = (Y_i, Y_i^2)$, then $\hat{\theta}_n = g(\bar{X})$ with $g(x) = (x_1, x_2 - x_1^2)$. Hence, (5.100) holds with

$$\text{Var}(X_1) = \begin{pmatrix} \sigma_y^2 & \gamma_y - \mu_y(\sigma_y^2 + \mu_y^2) \\ \gamma_y - \mu_y(\sigma_y^2 + \mu_y^2) & \kappa_y - (\sigma_y^2 + \mu_y^2)^2 \end{pmatrix}$$

and

$$\nabla g(x) = \begin{pmatrix} 1 & 0 \\ -2x_1 & 1 \end{pmatrix}.$$

The estimator $\hat{V}_n$ in (5.102) is strongly consistent, since $\nabla g(x)$ is obviously a continuous function. ∎

Similar results can be obtained for problems in Examples 3.21 and 3.23, and Exercises 92 and 93 in §3.6.

A key step in the previous discussion is the derivation of formula (5.101) for the asymptotic covariance matrix of $\hat{\theta}_n = g(\bar{X})$, via Taylor's expansion (Theorem 1.12) and the CLT. Thus, the idea can be applied to the case where $\hat{\theta}_n = \mathtt{T}(F_n)$, a differentiable statistical functional.

We still consider i.i.d. random $d$-vectors $X_1, ..., X_n$ from $F$. Suppose that $\mathtt{T}$ is a vector-valued functional whose components are $\varrho$-Hadamard differentiable at $F$, where $\varrho$ is either $\varrho_\infty$ or a distance satisfying (5.35). Let $\phi_F$ be the vector of influence functions of components of $\mathtt{T}$. If the components of $\phi_F$ satisfy (5.33), then (5.100) holds with $\theta = \mathtt{T}(F)$, $\hat{\theta}_n = \mathtt{T}(F_n)$, $F_n$ = the empirical c.d.f. in (5.1), and

$$V_n = \frac{\text{Var}(\phi_F(X_1))}{n} = \frac{1}{n}\int [\phi_F(x)]^\tau \phi_F(x) dF(x). \tag{5.104}$$

Formula (5.104) leads to a natural substitution variance estimator

$$\hat{V}_n = \frac{1}{n}\int [\phi_{F_n}(x)]^\tau \phi_{F_n}(x) dF_n(x) = \frac{1}{n^2}\sum_{i=1}^n [\phi_{F_n}(X_i)]^\tau \phi_{F_n}(X_i), \tag{5.105}$$

provided that $\phi_{F_n}(x)$ is well defined, i.e., the components of $\mathtt{T}$ are Gâteaux differentiable at $F_n$ for sufficiently large $n$. Under some more conditions on $\phi_{F_n}$ we can establish the consistency of $\hat{V}_n$ in (5.105).

**Theorem 5.15.** Let $X_1, ..., X_n$ be i.i.d. random $d$-vectors from $F$, $\mathtt{T}$ be a vector-valued functional whose components are Gâteaux differentiable at $F$ and $F_n$, and $\phi_F$ be the vector of influence functions of components of $\mathtt{T}$. Suppose that $\sup_{\|x\|\le c} \|\phi_{F_n}(x) - \phi_F(x)\| = o_p(1)$ for any $c > 0$ and that there exist a constant $c_0 > 0$ and a function $h(x) \ge 0$ such that $E[h(X_1)] < \infty$ and $P\big(\sup_{\|x\|\ge c_0} \|\phi_{F_n}(x)\|^2 \le h(x)\big) \to 1$. Then $\hat{V}_n$ in (5.105) is consistent for $V_n$ in (5.104).

**Proof.** Let $\zeta(x) = [\phi_F(x)]^\tau[\phi_F(x)]$ and $\zeta_n(x) = [\phi_{F_n}(x)]^\tau[\phi_{F_n}(x)]$. By the SLLN,

$$\frac{1}{n}\sum_{i=1}^n \zeta(X_i) \to_{a.s.} \int \zeta(x)dF(x).$$

Hence the result follows from

$$\frac{1}{n}\sum_{i=1}^n [\zeta_n(X_i) - \zeta(X_i)] = o_p(1).$$

Using the assumed conditions and the argument in the proof of Lemma 5.3, we can show that for any $\epsilon > 0$, there is a $c > 0$ such that

$$P\left(\frac{1}{n}\sum_{i=1}^n \|\zeta_n(X_i) - \zeta(X_i)\| I_{(c,\infty)}(\|X_i\|) > \frac{\epsilon}{2}\right) \le \epsilon$$

and

$$P\left(\frac{1}{n}\sum_{i=1}^n \|\zeta_n(X_i) - \zeta(X_i)\| I_{[0,c]}(\|X_i\|) > \frac{\epsilon}{2}\right) \le \epsilon$$

for sufficiently large $n$. This completes the proof. ∎

**Example 5.16.** Consider the L-functional defined in (5.39) and the L-estimator $\hat{\theta}_n = \mathtt{T}(F_n)$. Theorem 5.6 shows that $\mathtt{T}$ is Hadamard differentiable at $F$ under some conditions on $J$. It can be shown (exercise) that $\mathtt{T}$ is Gâteaux differentiable at $F_n$ with $\phi_{F_n}(x)$ given by (5.41) (with $F$ replaced by $F_n$). Then the difference $\phi_{F_n}(x) - \phi_F(x)$ is equal to

$$\int (F_n - F)(y)J(F_n(y))dy + \int (F - \delta_x)(y)[J(F_n(y)) - J(F(y))]dy.$$

One can show (exercise) that the conditions in Theorem 5.15 are satisfied if the conditions in Theorem 5.6(i) or (ii) (with $E|X_1| < \infty$) hold. ∎

Substitution variance estimators for M-estimators and U-statistics can also be derived (exercises).

The substitution method can clearly be applied to non-i.i.d. cases. For example, the LSE $\hat{\beta}$ in linear model (3.25) with a full rank $Z$ and i.i.d. $\varepsilon_i$'s

has $\text{Var}(\hat{\beta}) = \sigma^2(Z^\tau Z)^{-1}$, where $\sigma^2 = \text{Var}(\varepsilon_1)$. A consistent substitution estimator of $\text{Var}(\hat{\beta})$ can be obtained by replacing $\sigma^2$ in the formula of $\text{Var}(\hat{\beta})$ by a consistent estimator of $\sigma^2$ such as $SSR/(n-p)$ (see (3.36)).

We now consider variance estimation for the GEE estimators described in §5.4.1. By Theorem 5.14, the asymptotic covariance matrix of the GEE estimator $\hat{\theta}_n$ is given by (5.95), where

$$\text{Var}(s_n(\theta)) = \sum_{i=1}^{n} E\{[\psi_i(X_i,\theta)]^\tau \psi_i(X_i,\theta)\},$$

$$M_n(\theta) = \sum_{i=1}^{n} E[\varphi_i(X_i,\theta)],$$

and $\varphi_i(x,\gamma) = \partial\psi_i(x,\gamma)/\partial\gamma$. Substituting $\theta$ by $\hat{\theta}_n$ and the expectations by their empirical analogues, we obtain the substitution estimator $\hat{V}_n = \hat{M}_n^{-1}\widehat{\text{Var}}(s_n)\hat{M}_n^{-1}$, where

$$\widehat{\text{Var}}(s_n) = \sum_{i=1}^{n}[\psi_i(X_i,\hat{\theta}_n)]^\tau \psi_i(X_i,\hat{\theta}_n)$$

and

$$\hat{M}_n = \sum_{i=1}^{n} \varphi_i(X_i,\hat{\theta}_n).$$

The proof of the following result is left as an exercise.

**Theorem 5.16.** Let $X_1, ..., X_n$ be independent and $\{\hat{\theta}_n\}$ be a consistent sequence of GEE estimators. Assume the conditions in Theorem 5.14. Suppose further that the sequence of functions $\{h_{ij}, i = 1, 2, ...\}$ satisfies the conditions in Lemma 5.3 with $\Theta$ replaced by a compact neighborhood of $\theta$, where $h_{ij}(x,\gamma)$ is the $j$th row of $[\psi_i(x,\gamma)]^\tau\psi_i(x,\gamma)$. Let $V_n$ be given by (5.95). Then $\hat{V}_n = \hat{M}_n^{-1}\widehat{\text{Var}}(s_n)\hat{M}_n^{-1}$ is consistent for $V_n$. ∎

## 5.5.2 The jackknife

Applying the substitution method requires the derivation of a formula for the covariance matrix or asymptotic covariance matrix of a point estimator. There are variance estimation methods that can be used without actually deriving such a formula (only the existence of the covariance matrix or asymptotic covariance matrix is assumed), at the expense of requiring a large number of computations. These methods are called *resampling* methods, *replication* methods, or *data reuse* methods. The *jackknife* method introduced here and the *bootstrap* method in §5.5.3 are the most popular resampling methods.

The jackknife method was proposed by Quenouille (1949) and Tukey (1958). Let $\hat{\theta}_n$ be a vector-valued estimator based on independent $X_i$'s, where each $X_i$ is a random $d_i$-vector and $\sup_i d_i < \infty$. Let $\hat{\theta}_{-i}$ be the same estimator but based on $X_1, ..., X_{i-1}, X_{i+1}, ..., X_n$, $i = 1, ..., n$. Note that $\hat{\theta}_{-i}$ also depends on $n$ but the subscript $n$ is omitted for simplicity. Since $\hat{\theta}_n$ and $\hat{\theta}_{-1}, ..., \hat{\theta}_{-n}$ are estimators of the same quantity, the "sample covariance matrix"

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\theta}_{-i}-\bar{\theta}_n\right)^\tau\left(\hat{\theta}_{-i}-\bar{\theta}_n\right) \tag{5.106}$$

can be used as a measure of the variation of $\hat{\theta}_n$, where $\bar{\theta}_n$ is the average of $\hat{\theta}_{-i}$'s.

There are two major differences between the quantity in (5.106) and the sample covariance matrix $S^2$ previously discussed. First, $\hat{\theta}_{-i}$'s are not independent. Second, $\hat{\theta}_{-i} - \hat{\theta}_{-j}$ usually converges to 0 at a fast rate (such as $n^{-1}$). Hence, to estimate the asymptotic covariance matrix of $\hat{\theta}_n$, the quantity in (5.106) should be multiplied by a correction factor $c_n$. If $\hat{\theta}_n = \bar{X}$ ($d_i \equiv d$), then $\hat{\theta}_{-i} = (n-1)^{-1}(\bar{X} - X_i)$ and the quantity in (5.106) reduces to

$$\frac{1}{(n-1)^3}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^\tau\left(X_i-\bar{X}\right) = \frac{1}{(n-1)^2}S^2,$$

where $S^2$ is the sample covariance matrix. Thus, the correction factor $c_n$ is $(n-1)^2/n$ for the case of $\hat{\theta}_n = \bar{X}$, since, by the SLLN, $S^2/n$ is consistent for $\mathrm{Var}(\bar{X})$.

It turns out that the same correction factor works for many other estimators. This leads to the following *jackknife variance estimator* for $\hat{\theta}_n$:

$$\hat{V}_J = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{-i}-\bar{\theta}_n\right)^\tau\left(\hat{\theta}_{-i}-\bar{\theta}_n\right). \tag{5.107}$$

**Theorem 5.17.** Let $X_1, ..., X_n$ be i.i.d. random $d$-vectors from $F$ with finite $\mu = E(X_1)$ and $\mathrm{Var}(X_1)$, and let $\hat{\theta}_n = g(\bar{X})$. Suppose that $\nabla g$ is continuous at $\mu$ and $\nabla g(\mu) \neq 0$. Then the jackknife variance estimator $\hat{V}_J$ in (5.107) is strongly consistent for $V_n$ in (5.101).

**Proof.** From Definition 5.4, it suffices to show the case where $g$ is real-valued. Let $\bar{X}_{-i}$ be the sample mean based on $X_1, ..., X_{i-1}, X_{i+1}, ..., X_n$. From the mean-value theorem, we have

$$\begin{aligned}\hat{\theta}_{-i} - \hat{\theta}_n &= g(\bar{X}_{-i}) - g(\bar{X}) \\ &= \nabla g(\xi_{n,i})(\bar{X}_{-i} - \bar{X})^\tau \\ &= \nabla g(\bar{X})(\bar{X}_{-i} - \bar{X})^\tau + R_{n,i},\end{aligned}$$

where $R_{n,i} = \left[\nabla g(\xi_{n,i}) - \nabla g(\bar{X})\right](\bar{X}_{-i} - \bar{X})^\tau$ and $\xi_{n,i}$ is a point on the line segment between $\bar{X}_{-i}$ and $\bar{X}$. From $\bar{X}_{-i} - \bar{X} = (n-1)^{-1}(\bar{X} - X_i)$, it follows that $\sum_{i=1}^n (\bar{X}_{-i} - \bar{X}) = 0$ and

$$\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_{-i} - \hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n R_{n,i} = \bar{R}_n.$$

From the definition of the jackknife estimator in (5.107),

$$\hat{V}_J = A_n + B_n + 2C_n,$$

where

$$A_n = \frac{n-1}{n}\nabla g(\bar{X})\sum_{i=1}^n (\bar{X}_{-i} - \bar{X})^\tau (\bar{X}_{-i} - \bar{X})[\nabla g(\bar{X})]^\tau,$$

$$B_n = \frac{n-1}{n}\sum_{i=1}^n (R_{n,i} - \bar{R}_n)^2$$

and

$$C_n = \frac{n-1}{n}\sum_{i=1}^n (R_{n,i} - \bar{R}_n)\nabla g(\bar{X})(\bar{X}_{-i} - \bar{X})^\tau.$$

By $\bar{X}_{-i} - \bar{X} = (n-1)^{-1}(\bar{X} - X_i)$, the SLLN, and the continuity of $\nabla g$ at $\mu$,

$$A_n/V_n \to_{a.s.} 1.$$

Also,

$$(n-1)\sum_{i=1}^n \|\bar{X}_{-i} - \bar{X}\|^2 = \frac{1}{n-1}\sum_{i=1}^n \|X_i - \bar{X}\|^2 = O(1) \quad \text{a.s.} \tag{5.108}$$

Hence

$$\max_{i\le n} \|\bar{X}_{-i} - \bar{X}\|^2 \to_{a.s.} 0,$$

which, together with the continuity of $\nabla g$ at $\mu$ and $\|\xi_{n,i} - \bar{X}\| \le \|\bar{X}_{-i} - \bar{X}\|$, implies that

$$u_n = \max_{i\le n} \|\nabla g(\xi_{n,i}) - \nabla g(\bar{X})\| \to_{a.s.} 0.$$

From (5.101) and (5.108), $\sum_{i=1}^n \|\bar{X}_{-i} - \bar{X}\|^2/V_n = O(1)$ a.s. Hence

$$\frac{B_n}{V_n} \le \frac{n-1}{V_n n}\sum_{i=1}^n R_{n,i}^2 \le \frac{u_n}{V_n}\sum_{i=1}^n \|\bar{X}_{-i} - \bar{X}\|^2 \to_{a.s.} 0.$$

By the Cauchy-Schwarz inequality, $(C_n/V_n)^2 \le (A_n/V_n)(B_n/V_n) \to_{a.s.} 0$. This proves the result. ∎

A key step in the proof of Theorem 5.17 is that $\hat{\theta}_{-i} - \hat{\theta}_n$ can be approximated by $\nabla g(\bar{X})(\bar{X}_{-i} - \bar{X})^\tau$ and the contributions of the remainders, $R_{n,1}, ..., R_{n,n}$, are sufficiently small, i.e., $B_n/V_n \to_{a.s.} 0$. This indicates that the jackknife estimator (5.107) is consistent for $\hat{\theta}_n$ that can be well approximated by some linear statistic. In fact, the jackknife estimator (5.107) has been shown to be consistent when $\hat{\theta}_n$ is a U-statistic (Arvesen, 1969) or a statistical functional that is Hadamard differentiable and *continuously* Gâteaux differentiable at $F$ (which includes certain types of L-estimators and M-estimators). More details can be found in Shao and Tu (1995, Chapter 2).

The jackknife method can be applied to non-i.i.d. problems. A detailed discussion of the use of the jackknife method in survey problems can be found in Shao and Tu (1995, Chapter 6). We now consider the jackknife variance estimator for the LSE $\hat{\beta}$ in linear model (3.25). For simplicity, assume that $Z$ is of full rank. Assume also that $\varepsilon_i$'s are independent with $E(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma_i^2$. Then

$$\text{Var}(\hat{\beta}) = (Z^\tau Z)^{-1} \sum_{i=1}^n \sigma_i^2 Z_i^\tau Z_i (Z^\tau Z)^{-1}.$$

Let $\hat{\beta}_{-i}$ be the LSE of $\beta$ based on the data with the $i$th pair $(X_i, Z_i)$ deleted. Using the fact that $(A + c^\tau c)^{-1} = A^{-1} - A^{-1}c^\tau cA^{-1}/(1 + cA^{-1}c^\tau)$ for a matrix $A$ and a vector $c$, we can show that (exercise)

$$\hat{\beta}_{-i} = \hat{\beta} - r_i Z_i/(1 - h_{ii}), \tag{5.109}$$

where $r_i = X_i - \hat{\beta}Z_i^\tau$ is the $i$th residual and $h_{ii} = Z_i(Z^\tau Z)^{-1}Z_i^\tau$. Hence

$$\hat{V}_J = \frac{n-1}{n}(Z^\tau Z)^{-1}\left[\sum_{i=1}^n \frac{r_i^2 Z_i^\tau Z_i}{(1-h_{ii})^2} - \frac{1}{n}\sum_{i=1}^n \frac{r_i Z_i^\tau}{1-h_{ii}} \sum_{i=1}^n \frac{r_i Z_i}{1-h_{ii}}\right](Z^\tau Z)^{-1}.$$

Wu (1986) proposed the following weighted jackknife variance estimator that improves $\hat{V}_J$:

$$\hat{V}_{WJ} = \sum_{i=1}^n (1-h_{ii})\left(\hat{\beta}_{-i} - \hat{\beta}\right)^\tau \left(\hat{\beta}_{-i} - \hat{\beta}\right) = (Z^\tau Z)^{-1} \sum_{i=1}^n \frac{r_i^2 Z_i^\tau Z_i}{1-h_{ii}} (Z^\tau Z)^{-1}.$$

**Theorem 5.18.** Assume the conditions in Theorem 3.12 and that $\varepsilon_i$'s are independent. Then both $\hat{V}_J$ and $\hat{V}_{WJ}$ are consistent for $\text{Var}(\hat{\beta})$.

**Proof.** Let $l \in \mathcal{R}^p$ be a fixed nonzero vector and $l_i = l(Z^\tau Z)^{-1}Z_i^\tau$. Since $\max_{i \le n} h_{ii} \to 0$, the result for $\hat{V}_{WJ}$ follows from

$$\sum_{i=1}^n l_i^2 r_i^2 \Big/ \sum_{i=1}^n l_i^2 \sigma_i^2 \to_p 1. \tag{5.110}$$

By the WLLN (Theorem 1.14(ii)),

$$\sum_{i=1}^{n} l_i^2 \varepsilon_i^2 \Big/ \sum_{i=1}^{n} l_i^2 \sigma_i^2 \to_p 1.$$

Note that $r_i = \varepsilon_i + (\beta - \hat{\beta}) Z_i^\tau$ and

$$\max_{i \le n} [(\beta - \hat{\beta}) Z_i^\tau]^2 \le \|(\beta - \hat{\beta}) Z^\tau\|^2 \max_{i \le n} h_{ii} = o_p(1).$$

Hence (5.110) holds.

The consistency of $\hat{V}_J$ follows from (5.110) and

$$\frac{n-1}{n^2} \left( \sum_{i=1}^{n} \frac{l_i r_i}{1 - h_{ii}} \right)^2 \Big/ \sum_{i=1}^{n} l_i^2 \sigma_i^2 = o_p(1). \tag{5.111}$$

The proof of (5.111) is left as an exercise. ∎

Finally, let us consider the jackknife estimators for GEE estimators in §5.4.1. Under the conditions of Proposition 5.5 or 5.6, it can be shown that

$$\max_{i \le n} \|\hat{\theta}_{-i} - \hat{\theta}\| = o_p(1), \tag{5.112}$$

where $\hat{\theta}_{-i}$ is a root of $s_{ni}(\gamma) = 0$ and

$$s_{ni}(\gamma) = \sum_{j \ne i, j \le n} \psi_j(X_j, \gamma).$$

Using Taylor's expansion and the fact that $s_{ni}(\hat{\theta}_{-i}) = 0$ and $s_n(\hat{\theta}_n) = 0$, we obtain that

$$\psi_i(X_i, \hat{\theta}_{-i}) = (\hat{\theta}_{-i} - \hat{\theta}_n) \int_0^1 \nabla s_n\big(\hat{\theta}_n + t(\hat{\theta}_{-i} - \hat{\theta}_n)\big) dt.$$

Following the proof of Theorem 5.14, we obtain that

$$\hat{V}_J = [M_n(\theta)]^{-1} \sum_{i=1}^{n} [\psi_i(X_i, \hat{\theta}_{-i})]^\tau \psi_i(X_i, \hat{\theta}_{-i}) [M_n(\theta)]^{-1} + R_n,$$

where $R_n$ denotes a quantity satisfying $l R_n l^\tau / l V_n l^\tau = o_p(1)$ for $V_n$ in (5.95). Under the conditions of Theorem 5.16, it follows from (5.112) that $\hat{V}_J$ is consistent.

If $\hat{\theta}_n$ is computed using an iteration method, then the computation of $\hat{V}_J$ requires $n$ additional iteration processes. We may use the idea of a one-step MLE to reduce the amount of computation. For each $i$, let

$$\hat{\theta}_{-i} = \hat{\theta}_n - s_{ni}(\hat{\theta}_n) [\nabla s_{ni}(\hat{\theta}_n)]^{-1}, \tag{5.113}$$

which is the result from the first iteration when Newton-Raphson's method is applied in computing a root of $s_{ni}(\gamma) = 0$ and $\hat{\theta}_n$ is used as the initial point. Note that $\hat{\theta}_{-i}$'s in (5.113) satisfy (5.112) (exercise). If the jackknife variance estimator is based on $\hat{\theta}_{-i}$'s in (5.113), then

$$\hat{V}_J = [M_n(\theta)]^{-1} \sum_{i=1}^{n} [\psi_i(X_i, \hat{\theta}_n)]^\tau \psi_i(X_i, \hat{\theta}_n) [M_n(\theta)]^{-1} + R_n.$$

These results are summarized in the following theorem.

**Theorem 5.19.** Assume the conditions in Theorems 5.14 and 5.16. Assume further that $\hat{\theta}_{-i}$'s are given by (5.113) or GEE estimators satisfying (5.112). Then the jackknife variance estimator $V_J$ is consistent for $V_n$ given in (5.95). ∎

### 5.5.3 The bootstrap

The basic idea of the bootstrap method can be described as follows. Suppose that $P$ is a population or model that generates the sample $X$ and that we need to estimate $\mathrm{Var}(\hat{\theta})$, where $\hat{\theta} = \hat{\theta}(X)$ is an estimator, a statistic based on $X$. Suppose further that the unknown population $P$ is estimated by $\hat{P}$, based on the sample $X$. Let $X^*$ be a sample (called a *bootstrap sample*) taken from the estimated population $\hat{P}$ using the same or a similar sampling procedure used to obtain $X$, and let $\hat{\theta}^* = \hat{\theta}(X^*)$, which is the same as $\hat{\theta}$ but with $X$ replaced by $X^*$. If we believe that $P = \hat{P}$ (i.e., we have a perfect estimate of the population), then $\mathrm{Var}(\hat{\theta}) = \mathrm{Var}_*(\hat{\theta}^*)$, where $\mathrm{Var}_*$ is the conditional variance w.r.t. the randomness in generating $X^*$, given $X$. In general, $P \neq \hat{P}$ and, therefore, $\mathrm{Var}(\hat{\theta}) \neq \mathrm{Var}_*(\hat{\theta}^*)$. But $\hat{V}_B = \mathrm{Var}_*(\hat{\theta}^*)$ is an empirical analogue of $\mathrm{Var}(\hat{\theta})$ and can be used as an estimate of $\mathrm{Var}(\hat{\theta})$.

In a few cases, an explicit form of $\hat{V}_B = \mathrm{Var}_*(\hat{\theta}^*)$ can be obtained. First, consider i.i.d. $X_1, ..., X_n$ from a c.d.f. $F$ on $\mathcal{R}^d$. The population is determined by $F$. Suppose that we estimate $F$ by the empirical c.d.f. $F_n$ in (5.1) and that $X_1^*, ..., X_n^*$ are i.i.d. from $F_n$. For $\hat{\theta} = \bar{X}$, its bootstrap analogue is $\hat{\theta}^* = \bar{X}^*$, the average of $X_i^*$'s. Then

$$\hat{V}_B = \mathrm{Var}_*(\bar{X}^*) = \frac{1}{n^2} \sum_{i=1}^{n} (X_i - \bar{X})^\tau (X_i - \bar{X}) = \frac{n-1}{n^2} S^2,$$

where $S^2$ is the sample covariance matrix. In this case $\hat{V}_B = \mathrm{Var}_*(\bar{X}^*)$ is a strongly consistent estimator for $\mathrm{Var}(\bar{X})$. Next, consider i.i.d. random variables $X_1, ..., X_n$ from a c.d.f. $F$ on $\mathcal{R}$ and $\hat{\theta} = F_n^{-1}(\frac{1}{2})$, the sample

median. Suppose that $n = 2l - 1$ for an integer $l$. Let $X_1^*, ..., X_n^*$ be i.i.d. from $F_n$ and $\hat{\theta}^*$ be the sample median based on $X_1^*, ..., X_n^*$. Then

$$\hat{V}_B = \mathrm{Var}_*(\hat{\theta}^*) = \sum_{j=1}^{n} p_j \left( X_{(j)} - \sum_{i=1}^{n} p_i X_{(i)} \right)^2,$$

where $X_{(1)} \le \cdots \le X_{(n)}$ are order statistics and $p_j = P(\hat{\theta}^* = X_{(j)}|X)$. It can be shown (exercise) that

$$p_j = \sum_{t=0}^{l-1} \binom{n}{t} \frac{(j-1)^t (n-j+1)^{n-t} - j^t (n-j)^{n-t}}{n^n}. \tag{5.114}$$

However, in most cases $\hat{V}_B$ does not have a simple explicit form. When $P$ is known, the Monte Carlo method described in §4.1.4 can be used to approximate $\mathrm{Var}(\hat{\theta})$. That is, we draw repeatedly new data sets from $P$ and then use the sample covariance matrix based on the values of $\hat{\theta}$ computed from new data sets as a numerical approximation to $\mathrm{Var}(\hat{\theta})$. This idea can be used to approximate $\hat{V}_B$, since $\hat{P}$ is a known population. That is, we can draw $m$ bootstrap data sets $X^{*1}, ..., X^{*m}$ independently from $\hat{P}$ (conditioned on $X$), compute $\hat{\theta}^{*j} = \hat{\theta}(X^{*j})$, $j = 1, ..., m$, and approximate $\hat{V}_B$ by

$$\hat{V}_B^m = \frac{1}{m} \sum_{j=1}^{m} \left( \hat{\theta}^{*j} - \bar{\theta}^* \right)^\tau \left( \hat{\theta}^{*j} - \bar{\theta}^* \right),$$

where $\bar{\theta}^*$ is the average of $\hat{\theta}^{*j}$'s. Since each $X^{*j}$ is a data set generated from $\hat{P}$, $\hat{V}_B^m$ is a resampling estimator. From the SLLN, as $m \to \infty$, $\hat{V}_B^m \to_{a.s.} \hat{V}_B$, conditioned on $X$. Both $\hat{V}_B$ and its Monte Carlo approximation $\hat{V}_B^m$ are called *bootstrap variance estimators* for $\hat{\theta}$. $\hat{V}_B^m$ is more useful in practical applications, whereas in theoretical studies, we usually focus on $\hat{V}_B$.

The consistency of the bootstrap variance estimator $\hat{V}_B$ is a much more complicated problem than that of the jackknife variance estimator in §5.5.2. Some examples can be found in Shao and Tu (1995, §3.2.2).

The bootstrap method can also be applied to estimate quantities other than $\mathrm{Var}(\hat{\theta})$. For example, let $K(t) = P(\hat{\theta} \le t)$ be the c.d.f. of a real-valued estimator $\hat{\theta}$. From the previous discussion, a bootstrap estimator of $K(t)$ is the conditional probability $P(\hat{\theta}^* \le t|X)$, which can be approximated by the Monte Carlo approximation $m^{-1} \sum_{j=1}^{m} I_{(-\infty,t]}(\hat{\theta}^{*j})$. An important application of bootstrap distribution estimators in problems of constructing confidence sets is studied in §7.4. Here, we study the use of a bootstrap distribution estimator to form a consistent estimator of the asymptotic variance of a real-valued estimator $\hat{\theta}$.

Suppose that

$$\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, v), \tag{5.115}$$

where $v$ is unknown. Let $H_n(t)$ be the c.d.f. of $\sqrt{n}(\hat{\theta} - \theta)$ and

$$\hat{H}_B(t) = P(\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \le t | X) \tag{5.116}$$

be a bootstrap estimator of $H_n(t)$. If

$$\hat{H}_B(t) - H_n(t) \to_p 0$$

for any $t$, then, by (5.115),

$$\hat{H}_B(t) - \Phi\left(t/\sqrt{v}\right) \to_p 0,$$

which implies (exercise) that

$$\hat{H}_B^{-1}(\alpha) \to_p \Phi^{-1}\left(\alpha/\sqrt{v}\right) = \sqrt{v}\Phi^{-1}(\alpha)$$

for any $\alpha \in (0,1)$. Then, for $\alpha \neq \frac{1}{2}$,

$$\hat{H}_B^{-1}(1-\alpha) - \hat{H}_B^{-1}(\alpha) \to_p \sqrt{v}[\Phi^{-1}(1-\alpha) - \Phi^{-1}(\alpha)].$$

Therefore, a consistent estimator of $v/n$, the asymptotic variance of $\hat{\theta}$, is

$$\tilde{V}_B = \frac{1}{n}\left[\frac{\hat{H}_B^{-1}(1-\alpha) - \hat{H}_B^{-1}(\alpha)}{\Phi^{-1}(1-\alpha) - \Phi^{-1}(\alpha)}\right]^2.$$

The following result gives some conditions under which $\hat{H}_B(t) - H_n(t) \to_p 0$. The proof of part (i) is omitted. The proof of part (ii) is given in Exercises 97-99 in §5.6.

**Theorem 5.20.** Suppose that $X_1, ..., X_n$ are i.i.d. from a c.d.f. $F$ on $\mathcal{R}^d$. Let $\hat{\theta} = \mathtt{T}(F_n)$, where $\mathtt{T}$ is a real-valued functional, $\hat{\theta}^* = \mathtt{T}(F_n^*)$, where $F_n^*$ is the empirical c.d.f. based on a bootstrap sample $X_1^*, ..., X_n^*$ i.i.d. from $F_n$, and let $\hat{H}_B$ be given by (5.116).
(i) If $\mathtt{T}$ is $\varrho_\infty$-Hadamard differentiable at $F$ and (5.33) holds, then

$$\varrho_\infty(\hat{H}_B, H_n) \to_p 0. \tag{5.117}$$

(ii) If $d = 1$ and $\mathtt{T}$ is $\varrho_{L_p}$-Fréchet differentiable at $F$ ($\int\{F(t)[1-F(t)]\}^{p/2}dt < \infty$ if $1 \le p < 2$) and (5.33) holds, then (5.117) holds. ∎

Applications of the bootstrap method to non-i.i.d. cases can be found, for example, in Efron and Tibshirani (1993), Hall (1992), and Shao and Tu (1995).

## 5.6 Exercises

1. Let $\varrho_\infty$ be defined by (5.3).
   (a) Show $\varrho_\infty$ is a distance on $\mathcal{F}$.
   (b) Find an example of a sequence $\{G_j\} \subset \mathcal{F}$ for which $\lim_{j\to\infty} G_j(t) = G_0(t)$ for every $t$ at which $G_0$ is continuous, but $\varrho_\infty(G_j, G_0)$ does not converge to 0.

2. Let $X_1, ..., X_n$ be i.i.d. random $d$-vectors with c.d.f. $F$ and $F_n$ be the empirical c.d.f. defined by (5.1). Show that for any $t > 0$ and $\epsilon > 0$, there is a $C_{\epsilon,d}$ such that for all $n = 1, 2, ...,$

$$P\left(\sup_{m\geq n} \varrho_\infty(F_m, F) > t\right) \leq \frac{C_{\epsilon,d} e^{-(2-\epsilon)t^2 n}}{1 - e^{-(2-\epsilon)t^2}}.$$

3. Show that $\varrho_{M_p}$ defined by (5.4) is a distance on $\mathcal{F}_p$, $p \geq 1$.

4. Show that $\varrho_{L_p}$ defined by (5.5) is a distance on $\mathcal{F}_1$ for any $p \geq 1$.

5. Let $\mathcal{F}_1$ be the collection of c.d.f.'s on $\mathcal{R}$ with finite means.
   (a) Show that $\varrho_{M_1}(G_1, G_2) = \int_0^1 |G_1^{-1}(z) - G_2^{-1}(z)|dz$, where $G^{-1}(z) = \inf\{t : G(t) \geq z\}$ for any $G \in \mathcal{F}$.
   (b) Show that $\varrho_{M_1}(G_1, G_2) = \varrho_{L_1}(G_1, G_2)$.

6. Find an example of a sequence $\{G_j\} \subset \mathcal{F}$ for which
   (a) $\lim_{j\to\infty} \varrho_\infty(G_j, G_0) = 0$ but $\varrho_{M_2}(G_j, G_0)$ does not converge to 0;
   (b) $\lim_{j\to\infty} \varrho_{M_2}(G_j, G_0) = 0$ but $\varrho_\infty(G_j, G_0)$ does not converge to 0.

7. Repeat the previous exercise with $\varrho_{M_2}$ replaced by $\varrho_{L_2}$.

8. Let $X$ be a random variable having c.d.f. $F$. Show that
   (a) $E|X|^2 < \infty$ implies $\int\{F(t)[1 - F(t)]\}^{p/2}dt < \infty$ for $p \in (1, 2)$;
   (b) $E|X|^{2+\delta} < \infty$ with some $\delta > 0$ implies $\int\{F(t)[1 - F(t)]\}^{1/2}dt < \infty$.

9. For any one-dimensional $G_j \in \mathcal{F}_1$, $j = 1, 2$, show that $\varrho_{L_1}(G_1, G_2) \geq |\int x dG_1 - \int x dG_2|$.

10. In the proof of Theorem 5.3, show that $p_i = c/n$, $i = 1, ..., n$, $\lambda = -(c/n)^{n-1}$ is a maximum of the function $H(p_1, ..., p_n, \lambda)$ over $p_i > 0$, $i = 1, ..., n$, $\sum_{i=1}^n p_i = c$.

11. Show that (5.11)-(5.13) is a solution to the problem of maximizing $\ell(G)$ in (5.8) subject to (5.10).

12. In the proof of Theorem 5.4, prove the case of $m \geq 2$.

13. Show that a maximum of $\ell(G)$ in (5.17) subject to (5.10) is given by (5.11) with $\hat{p}_i$ defined by (5.18) and (5.19).

14. In Example 5.2, show that an MELE is given by (5.11) with $\hat{p}_i$'s given by (5.21).

15. In Example 5.3, show that
(a) maximizing (5.22) subject to (5.23) is equivalent to maximizing

$$\prod_{i=1}^{n} q_i^{\delta_{(i)}} (1-q_i)^{n-i+1-\delta_{(i)}},$$

where $q_i = p_i / \sum_{j=i}^{n+1} p_j$, $i = 1, ..., n$;
(b) $\hat{F}$ given by (5.24) maximizes (5.22) subject to (5.23); (Hint: use part (a) and the fact that $p_i = q_i \prod_{j=1}^{i-1}(1-q_j)$.)
(c) $\hat{F}$ given by (5.25) is the same as that in (5.24).
(d) if $\delta_i = 1$ for all $i$ (no censoring), then $\hat{F}$ in (5.25) is the same as the empirical c.d.f. in (5.1).

16. Let $f_n$ be given by (5.26).
(a) Show that $f_n$ is a Lebesgue p.d.f. on $\mathcal{R}$.
(b) Suppose that $f$ is continuously differentiable at $t$, $\lambda_n \to 0$, and $n\lambda_n \to \infty$. Show that (5.27) holds.
(c) Under $n\lambda_n^3 \to 0$ and the conditions of (b), show that (5.28) holds.
(d) Suppose that $f$ is continuous on $[a, b]$, $-\infty < a < b < \infty$, $\lambda_n \to 0$, and $n\lambda_n \to \infty$. Show that $\int_a^b f_n(t)dt \to_p \int_a^b f(t)dt$.

17. Let $\hat{f}$ be given by (5.29).
(a) Show that $\hat{f}$ is a Lebesgue p.d.f. on $\mathcal{R}$.
(b) Prove (5.30) under the condition that $\lambda_n \to 0$, $n\lambda_n \to \infty$, and $f$ is bounded and is continuous at $t$ and $\int [w(t)]^{2+\delta} dt < \infty$ for some $\delta > 0$. (Hint: check Liapunov's condition and apply Theorem 1.15.)
(c) Suppose that $\lambda_n \to 0$, $n\lambda_n \to \infty$, and $f$ is bounded and is continuous on $[a, b]$, $-\infty < a < b < \infty$. Show that $\int_a^b \hat{f}(t)dt \to_p \int_a^b f(t)dt$.

18. Show that $\varrho$-Fréchet differentiability implies $\varrho$-Hadamard differentiability.

19. Suppose that a functional $\mathtt{T}$ is Gâteaux differentiable at $F$ with a continuous differential $\mathrm{L}_F$ in the sense that $\varrho_\infty(\Delta_j, \Delta) \to 0$ implies $\mathrm{L}_F(\Delta_j) \to \mathrm{L}_F(\Delta)$. Show that $\phi_F$ is bounded.

20. Suppose that a functional $\mathtt{T}$ is Gâteaux differentiable at $F$ with a bounded and continuous influence function $\phi_F$. Show that the differential $\mathrm{L}_F$ is continuous in the sense described in the previous exercise.

21. Let $\mathtt{T}(G) = g(\int x dG)$ be a functional defined on $\mathcal{F}_1$, the collection of one-dimensional c.d.f.'s with finite means.
(a) Find a differentiable function $g$ for which the functional $\mathtt{T}$ is not $\varrho_\infty$-Hadamard differentiable at $F$.
(b) Show that if $g$ is a differentiable function, then $\mathtt{T}$ is $\varrho_{L_1}$-Fréchet differentiable at $F$. (Hint: use the result in Exercise 9.)

22. In Example 5.5, show that (5.36) holds. (Hint: for $\Delta = c(G_1 - G_2)$, show that $\|\Delta\|_V \le |c|(\|G_1\|_V + \|G_2\|_V) = 2|c|$.)

23. In Example 5.5, show that $\phi_F$ is continuous if $F$ is continuous.

24. In Example 5.5, show that $\mathtt{T}$ is not $\varrho_\infty$-Fréchet differentiable at $F$.

25. Prove Proposition 5.1(ii).

26. Suppose that $\mathtt{T}$ is first-order and second-order $\varrho$-Hadamard differentiable at $F$. Prove (5.38).

27. Find an example of a second-order $\varrho$-Fréchet differentiable functional $\mathtt{T}$ that is not first-order $\varrho$-Hadamard differentiable.

28. Prove (5.40) and that (5.33) is satisfied if $F$ has a finite variance.

29. Prove (iv) and (v) of Theorem 5.6.

30. Discuss which of (i)-(v) in Theorem 5.6 can be applied to each of the L-estimators in Example 5.6.

31. Obtain explicit forms of the influence functions for L-estimators in Example 5.6. Discuss which of them are bounded and continuous.

32. Provide an example in which the L-functional $\mathtt{T}$ given by (5.39) is not $\varrho_\infty$-Hadamard differentiable at $F$. (Hint: consider an untrimmed $J$.)

33. Discuss which M-functionals defined in (i)-(vi) of Example 5.7 satisfy the conditions of Theorem 5.7.

34. In the proof of Theorem 5.7, show that $R_{2j} \to 0$.

35. Show that the second equality in (5.44) holds when $\psi$ is Borel and bounded.

36. Show that the functional $\mathtt{T}$ in (5.46) is $\varrho_\infty$-Hadamard differentiable at $F$ with the differential given by (5.47). Obtain the influence function $\phi_F$ and show that it is bounded and continuous if $F$ is continuous.

37. Show that the functional T in (5.48) is $\varrho_\infty$-Hadamard differentiable at $F$ with the differential given by (5.49). Obtain the influence function $\phi_F$ and show that it is bounded and continuous if $F(y,\infty)$ and $F(\infty,z)$ are continuous.

38. Let $F$ be a continuous c.d.f. on $\mathcal{R}$. Suppose that $F$ is symmetric about $\theta$ and is strictly increasing in a neighborhood of $\theta$. Show that $\lambda_F(t)=0$ if and only if $t=\theta$, where $\lambda_F(t)$ is defined by (5.50) with a strictly increasing $J$ satisfying $J(1-t)=-J(t)$.

39. Show that $\lambda_F(t)$ in (5.50) is differentiable at $\theta$ and $\lambda_F'(\theta)$ is equal to $-\int J'(F(x))F'(x)dF(x)$.

40. Let $\mathrm{T}(F_n)$ be an R-estimator satisfying the conditions in Theorem 5.8. Show that (5.34) holds with

$$\sigma_F^2=\int_0^1[J(t)]^2dt\Big/\left[\int_{-\infty}^{\infty}J'(F(x))F'(x)dF(x)\right]^2.$$

41. Calculate the asymptotic relative efficiency of the Hodges-Lehmann estimator in Example 5.8 w.r.t. the sample mean based on an i.i.d. sample from $F$ when
(a) $F$ is the c.d.f. of $N(\mu,\sigma^2)$;
(b) $F$ is the c.d.f. of the logistic distribution $LG(\mu,\sigma)$;
(c) $F$ is the c.d.f. of the double exponential distribution $DE(\mu,\sigma)$;
(d) $F(x)=F_0(x-\theta)$, where $F_0(x)$ is the c.d.f. of the t-distribution $t_\nu$ with $\nu\geq 3$.

42. Let $G$ be a c.d.f. on $\mathcal{R}$. Show that $G(x)\geq t$ if and only if $x\geq G^{-1}(t)$.

43. Show that (5.60) implies that $\hat{\theta}_p$ is strongly consistent for $\theta_p$ and is $\sqrt{n}$-consistent for $\theta_p$ if $F'(\theta_p-)$ and $F'(\theta_p+)$ exist.

44. Under the condition of Theorem 5.9, show that for $\rho_\epsilon=e^{-2\delta_\epsilon^2}$,

$$P\left(\sup_{m\geq n}|\hat{\theta}_p-\theta_p|>\epsilon\right)\leq\frac{2C\rho_\epsilon^n}{1-\rho_\epsilon},\qquad n=1,2,....$$

45. Prove that $\varphi_n(t)$ in (5.62) is the Lebesgue p.d.f. of the $p$th sample quantile $\hat{\theta}_p$ when $F$ has the Lebesgue p.d.f. $f$, by
(a) differentiating the c.d.f. of $\hat{\theta}_p$ in (5.61);
(b) using result (5.59) and the result in Example 2.9.

46. Let $X_1,...,X_n$ be i.i.d. random variables from $F$ with a finite mean. Show that $\hat{\theta}_p$ has a finite $j$th moment for sufficiently large $n$, $j=1,2,....$

47. Prove Theorem 5.10(i).

48. Suppose that a c.d.f. $F$ has a Lebesgue p.d.f. $f$. Using the p.d.f. in (5.62) and Scheffé's theorem (Proposition 1.17), prove part (iv) of Theorem 5.10.

49. Let $\{k_n\}$ be a sequence of integers satisfying $k_n/n = p + o(n^{-1/2})$ with $p \in (0,1)$, and let $X_1, ..., X_n$ be i.i.d. random variables from a c.d.f. $F$ with $F'(\theta_p) > 0$. Show that
$$\sqrt{n}(X_{(k_n)} - \theta_p) \to_d N(0, p(1-p)/[F'(\theta_p)]^2).$$

50. In the proof of Theorem 5.11, prove (5.65), (5.68), and inequality (5.67).

51. Prove Corollary 5.1.

52. Prove the claim in Example 5.9.

53. Let $T(G) = G^{-1}(p)$ be the $p$th quantile functional. Suppose that $F$ has a positive derivative $F'$ in a neighborhood of $\theta = F^{-1}(p)$. Show that $T$ is Gâteaux differentiable at $F$ and obtain the influence function $\phi_F(x)$.

54. Let $X_1, ..., X_n$ be i.i.d. from the Cauchy distribution $C(0,1)$.
(a) Show that $E(X_{(j)})^2 < \infty$ if and only if $3 \le j \le n-2$.
(b) Show that $E(\hat{\theta}_{0.5})^2 < \infty$ for $n \ge 5$.

55. Suppose that $F$ is the c.d.f. of the uniform distribution $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathcal{R}$. Obtain the asymptotic relative efficiency of the sample median w.r.t. the sample mean, based on an i.i.d. sample of size $n$ from $F$.

56. Suppose that $F(x) = F_0(x - \theta)$ and $F_0$ is the c.d.f. of the Cauchy distribution $C(0,1)$ truncated at $c$ and $-c$, i.e., $F_0$ has the Lebesgue p.d.f. $(1+x^2)^{-1} I_{(-c,c)}(x) / \int_{-c}^{c} (1+x^2)^{-1} dt$. Obtain the asymptotic relative efficiency of the sample median w.r.t. the sample mean, based on an i.i.d. sample of size $n$ from $F$.

57. Show that $\bar{X}_\alpha$ in (5.70) is the L-estimator corresponding to the $J$ function given in Example 5.6(iii) with $\beta = 1 - \alpha$.

58. Let $X_1, ..., X_n$ be i.i.d. random variables from $F$, where $F$ is symmetric about $\theta$.
(a) Show that $X_{(j)} - \theta$ and $\theta - X_{(n-j+1)}$ have the same distribution.
(b) Show that $\sum_{j=1}^{n} w_j X_{(j)}$ has a c.d.f. symmetric about $\theta$, if $w_i$'s are constants satisfying $\sum_{i=1}^{n} w_i = 1$ and $w_j = w_{n-j+1}$ for all $j$.
(c) Show that the trimmed sample mean $\bar{X}_\alpha$ has a c.d.f. symmetric about $\theta$.

59. Under the conditions in one of (i)-(iii) of Theorem 5.6, show that (5.34) holds for $\mathtt{T}(F_n)$ with $\sigma_F^2$ given by (5.72), if $\sigma_F^2 < \infty$.

60. Prove (5.71) under the assumed conditions.

61. For the functional $\mathtt{T}$ given by (5.39), show that $\mathtt{T}(F) = \theta$ if $F$ is symmetric about $\theta$ and $J$ is symmetric about $\frac{1}{2}$.

62. Suppose that $F$ is the double exponential distribution $DE(\theta, 1)$, where $\theta \in \mathcal{R}$. Obtain the asymptotic relative efficiency of the trimmed sample mean $\bar{X}_\alpha$ w.r.t. the sample mean, based on an i.i.d. sample of size $n$ from $F$.

63. Show that the method of moments in §3.5.2 is a special case of the GEE method.

64. Let $\ell(\theta, \xi)$ be a likelihood. Show that a maximum profile likelihood estimator $\hat{\theta}$ of $\theta$ is an MLE if $\xi(\theta)$, the maximum of $\sup_\xi \ell(\theta, \xi)$ for a fixed $\theta$, does not depend on $\theta$.

65. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$. Derive the profile likelihood function for $\mu$ or $\sigma^2$. Discuss in each case whether the maximum profile likelihood estimator is the same as the MLE.

66. Complete the proof of Proposition 5.4.

67. In the proof of Lemma 5.3, show that the probability in (5.89) is bounded by $\epsilon$.

68. In Example 5.11, show that $\psi_i$'s satisfy the conditions of Lemma 5.3 if $\Theta$ is compact and $\sup_i \|Z_i\| < \infty$.

69. In the proof of Proposition 5.5, show that $\{\Delta_n(\gamma)\}$ is equicontinuous on any open subset of $\Theta$.

70. Prove Proposition 5.6.

71. Prove the claim in Example 5.12.

72. Prove the claims in Example 5.13.

73. For Huber's M-estimator discussed in Example 5.13, obtain a formula for $e(F)$, the asymptotic relative efficiency of $\hat{\theta}_n$ w.r.t. $\bar{X}$, when $F$ is given by (5.69). Show that $\lim_{\tau\to\infty} e(F) = \infty$. Find the value of $e(F)$ when $\epsilon = 0$, $\sigma = 1$, and $C = 1.5$.

74. Consider the $\psi$ function in Example 5.7(ii). Show that under some conditions on $F$, $\psi$ satisfies the conditions given in Theorem 5.13(i) or (ii). Obtain $\sigma_F^2$ in (5.93) in this case.

75. In the proof of Theorem 5.14, show that
(a) (5.96) holds;
(b) (5.98) holds;
(c) (5.99) implies (5.97). (Hint: use Theorem 1.9(iii).)

76. Prove the claim in Example 5.14, assuming some necessary moment conditions.

77. Derive the asymptotic distribution of the MQLE (the GEE estimator based on (5.84)), assuming that $X_i = (X_{i1}, ..., X_{id_i})$, $E(X_{it}) = me^{\eta_i}/(1+e^{\eta_i})$, $\text{Var}(X_{it}) = m\phi_i e^{\eta_i}/(1+e^{\eta_i})^2$, and (4.57) holds with $g(t) = \log \frac{t}{1-t}$.

78. Repeat the previous exercise under the assumption that $E(X_{it}) = e^{\eta_i}$, $\text{Var}(X_{it}) = \phi_i e^{\eta_i}$, and (4.57) holds with $g(t) = \log t$ or $g(t) = 2\sqrt{t}$.

79. In Theorem 5.14, show that result (5.94) still holds if $\tilde{R}_i$ is replaced by an estimator $\hat{R}_i$ satisfying $\max_{i \le n} \|\hat{R}_i - U_i\| = o_p(1)$, where $U_i$'s are correlation matrices.

80. Suppose that $X_1, ..., X_n$ are independent (not necessarily identically distributed) random $d$-vectors with $E(X_i) = \mu$ for all $i$. Suppose also that $\sup_i E\|X_i\|^{2+\delta} < \infty$ for some $\delta > 0$. Let $\mu = E(X_1)$, $\theta = g(\mu)$, and $\hat{\theta}_n = g(\bar{X})$. Show that
(a) (5.100) holds with $V_n = n^{-2}\nabla g(\mu) \sum_{i=1}^n \text{Var}(X_i)[\nabla g(\mu)]^\tau$;
(b) $\hat{V}_n$ in (5.102) is consistent for $V_n$ in part (a).

81. Consider the ratio estimator in Example 3.21. Derive the estimator $\hat{V}_n$ given by (5.102) and show that $\hat{V}_n$ is consistent for the asymptotic variance of the ratio estimator.

82. Derive a consistent variance estimator for $\hat{R}(t)$ in Example 3.23.

83. Prove the claims in Example 5.16.

84. Derive a consistent variance estimator for a U-statistic satisfying the conditions in Theorem 3.5(i).

85. Derive a consistent variance estimator for Huber's M-estimator discussed in Example 5.13.

86. Assume the conditions in Theorem 5.8. Let $r \in (0, \frac{1}{2})$.
(a) Show that $n^r \lambda_F(\mathtt{T}(F_n) + n^{-r}) \to_p \lambda_F(\mathtt{T}(F))$.
(b) Show that $n^r[\lambda_{F_n}(\mathtt{T}(F_n) + n^{-r}) - \lambda_F(\mathtt{T}(F_n) + n^{-r})] \to_p 0$.
(c) Derive a consistent estimator of the asymptotic variance of $\mathtt{T}(F_n)$, using the results in (a) and (b).

87. Prove Theorem 5.16.

88. Let $X_1, ..., X_n$ be random variables and $\hat{\theta} = \bar{X}^2$. Show that the jackknife estimator in (5.107) equals $\frac{4\bar{X}^2\hat{c}_2}{n-1} - \frac{4\bar{X}\hat{c}_3}{(n-1)^2} + \frac{\hat{c}_4-\hat{c}_2^2}{(n-1)^3}$, where $\hat{c}_j$'s are the sample central moments defined by (3.57).

89. Prove (5.109).

90. In the proof of Theorem 5.18, prove (5.111).

91. Show that $\hat{\theta}_{-i}$'s in (5.113) satisfy (5.112), under the conditions of Theorem 5.14.

92. Prove Theorem 5.19.

93. Prove (5.114).

94. Let $X_1, ..., X_n$ be random variables and $\hat{\theta} = \bar{X}^2$. Show that the bootstrap variance estimator based on i.i.d. $X_i^*$'s from $F_n$ is equal to $\hat{V}_B = \frac{4\bar{X}^2\hat{c}_2}{n} + \frac{4\bar{X}\hat{c}_3}{n^2} + \frac{\hat{c}_4}{n^3}$, where $\hat{c}_j$'s are the sample central moments defined by (3.57).

95. Let $X_1, ..., X_n$ be i.i.d. from a Lebesgue p.d.f. $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$ on $\mathcal{R}$, where $f$ is known. Let $H_n(t) = P(\sqrt{n}(\bar{X} - \mu)/S \le t)$ and $\hat{H}_B(t) = P(\sqrt{n}(\bar{X}^* - \bar{X})/S^* \le t|X)$ be the bootstrap estimator of $H_n$, where $S^2$ is the sample variance, $X_i^*$'s are i.i.d. from $\frac{1}{s}f\left(\frac{x-\bar{x}}{s}\right)$, given $\bar{X} = \bar{x}$ and $S = s$, and $S^*$ is the bootstrap analogue of $S$. Show that $\hat{H}_B \equiv H_n$.

96. Let $G$, $G_1$, $G_2$,..., be c.d.f.'s on $\mathcal{R}$. Suppose that $\varrho_\infty(G_j, G) \to 0$ as $j \to \infty$ and $G'(x)$ exists and is positive for all $x \in \mathcal{R}$. Show that $G_j^{-1}(p) \to G^{-1}(p)$ for any $p \in (0, 1)$.

97. Let $X_1, ..., X_n$ be i.i.d. from a c.d.f. $F$ on $\mathcal{R}^d$ with a finite $\mathrm{Var}(X_1)$. Let $X_1^*, ..., X_n^*$ be i.i.d. from the empirical c.d.f. $F_n$. Show that for almost all given sequences $X_1, X_2, ..., \sqrt{n}(\bar{X}^* - \bar{X}) \to_d N(0, \mathrm{Var}(X_1))$. (Hint: verify Lindeberg's condition.)

98. Let $X_1, ..., X_n$ be i.i.d. from a c.d.f. $F$ on $\mathcal{R}^d$, $X_1^*, ..., X_n^*$ be i.i.d. from the empirical c.d.f. $F_n$, and let $F_n^*$ be the empirical c.d.f. based on $X_i^*$'s. Using DKW's inequality (Lemma 5.1), show that
(a) $\varrho_\infty(F_n^*, F) \to_{a.s.} 0$;
(b) $\varrho_\infty(F_n^*, F) = O_p(n^{-1/2})$;
(c) $\varrho_{L_p}(F_n^*, F) = O_p(n^{-1/2})$, under the condition in Theorem 5.20(ii).

99. Using the results from the previous two exercises, prove Theorem 5.20(ii).

100. Under the conditions in Theorem 5.11, establish a Bahadur's representation for the bootstrap sample quantile $\hat{\theta}_p^*$.