Stat 710: Mathematical Statistics Lecture 15

Jun Shao

Department of Statistics
University of Wisconsin
Madison, WI 53706, USA

Estimation in Nonparametric Models

Data $X = (X_1, ..., X_n)$, where X_i's are random d-vectors i.i.d. from an unknown c.d.f. F in a nonparametric family.
We study mainly two topics

- Estimation of the c.d.f. F.
- Estimation of $\theta = T(F)$, where T is a functional.

Empirical c.d.f.

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,t]}(X_i), \qquad t \in \mathscr{R}^d,$$

where $(-\infty, a]$ denotes the set $(-\infty, a_1] \times \cdots \times (-\infty, a_d]$ for any $a = (a_1, ..., a_d) \in \mathscr{R}^d$.
F_n is the distribution putting mass n^{-1} at each X_i, $i = 1, ..., n$.

Estimation in Nonparametric Models

Data $X = (X_1, ..., X_n)$, where X_i's are random d-vectors i.i.d. from an unknown c.d.f. F in a nonparametric family.

We study mainly two topics

- Estimation of the c.d.f. F.
- Estimation of $\theta = T(F)$, where T is a functional.

Empirical c.d.f.

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,t]}(X_i), \qquad t \in \mathscr{R}^d,$$

where $(-\infty, a]$ denotes the set $(-\infty, a_1] \times \cdots \times (-\infty, a_d]$ for any $a = (a_1, ..., a_d) \in \mathscr{R}^d$.

F_n is the distribution putting mass n^{-1} at each X_i, $i = 1, ..., n$.

Properties of empirical c.d.f.

For any $t \in \mathscr{R}^d$, $nF_n(t)$ 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;
$F_n(t)$ is $\sqrt{n}$-consistent for $F(t)$.
For any m fixed distinct points $t_1,...,t_m$ in $\mathscr{R}^d$, it follows from the multivariate CLT (Corollary 1.2) that as $n \to \infty$,

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

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

$$P\big(X_1 \in (-\infty,t_i] \cap (-\infty,t_j]\big) - F(t_i)F(t_j).$$

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 $\mathscr{F}$, the collection of all c.d.f.'s on $\mathscr{R}^d$.
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 in Shorack and Wellner (1986).

Properties of empirical c.d.f.

Sup-norm and sup-norm distance

$$\rho_\infty(G_1, G_2) = \|G_1 - G_2\|_\infty = \sup_{t \in \mathscr{R}^d} |G_1(t) - G_2(t)|, \quad G_j \in \mathscr{F}.$$

The following result is useful.
Its proof is omitted.

Lemma 5.1 (Dvoretzky, Kiefer, and Wolfowitz (DKW) inequality)

(i) When $d = 1$, there exists a positive constant C (not depending on F) such that

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

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

$$P\left(\rho_\infty(F_n, F) > z\right) \le C_{\varepsilon,d}e^{-(2-\varepsilon)nz^2}, \quad z > 0, n = 1, 2, \ldots.$$

Properties of empirical c.d.f.

Sup-norm and sup-norm distance

$$\rho_\infty(G_1,G_2) = \|G_1 - G_2\|_\infty = \sup_{t\in\mathscr{R}^d} |G_1(t) - G_2(t)|, \quad G_j \in \mathscr{F}.$$

The following result is useful.
Its proof is omitted.

Lemma 5.1 (Dvoretzky, Kiefer, and Wolfowitz (DKW) inequality)

(i) When $d = 1$, there exists a positive constant C (not depending on F) such that

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

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

$$P\left(\rho_\infty(F_n,F) > z\right) \le C_{\varepsilon,d}e^{-(2-\varepsilon)nz^2}, \quad z > 0, n = 1,2,....$$

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 $\mathscr{R}^d$.

(i) $\rho_\infty(F_n,F) \to_{a.s.} 0$ as $n \to \infty$;

(ii) $E[\sqrt{n}\rho_\infty(F_n,F)]^s = O(1)$ for any $s > 0$.

Proof

(i) From DKW's inequality,

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

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 55 of §1.6, we obtain that, as long as $2-\varepsilon > 0$,

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

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 $\mathscr{R}^d$.

(i) $\rho_\infty(F_n,F) \to_{a.s.} 0$ as $n \to \infty$;

(ii) $E[\sqrt{n}\rho_\infty(F_n,F)]^s = O(1)$ for any $s > 0$.

Proof

(i) From DKW's inequality,

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

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 55 of §1.6, we obtain that, as long as $2 - \varepsilon > 0$,

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

Remarks

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

L_p distance

When $d = 1$, another useful distance for measuring the closeness between F_n and F is the L_p distance ρ_{L_p} induced by the L_p-norm ($p \geq 1$)

$$\rho_{L_p}(G_1, G_2) = \|G_1 - G_2\|_{L_p} = \left[\int |G_1(t) - G_2(t)|^p dt\right]^{1/p}, \quad G_j \in \mathscr{F}_1,$$

where $\mathscr{F}_1 = \{G \in \mathscr{F} : \int |t|^p dG(t) < \infty\}$.

Remarks

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

L_p distance

When $d = 1$, another useful distance for measuring the closeness between F_n and F is the L_p distance ρ_{L_p} induced by the L_p-norm ($p \geq 1$)

$$\rho_{L_p}(G_1, G_2) = \|G_1 - G_2\|_{L_p} = \left[\int |G_1(t) - G_2(t)|^p dt\right]^{1/p}, \quad G_j \in \mathscr{F}_1,$$

where $\mathscr{F}_1 = \{G \in \mathscr{F} : \int |t|^p dG(t) < \infty\}$.

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. $F \in \mathscr{F}_1$.
(i) $\rho_{L_p}(F_n,F) \to_{a.s.} 0$;
(ii) $E[\sqrt{n}\rho_{L_p}(F_n,F)] = O(1)$ if $1 \le p < 2$ and $\int \{F(t)[1-F(t)]\}^{p/2} dt < \infty$, or $p \ge 2$.

Proof

(i) Since $[\rho_{L_p}(F_n,F)]^p \le [\rho_\infty(F_n,F)]^{p-1}[\rho_{L_1}(F_n,F)]$ and, by Theorem 5.1, $\rho_\infty(F_n,F) \to_{a.s.} 0$, 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

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

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

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

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. $F \in \mathscr{F}_1$.
(i) $\rho_{L_p}(F_n,F) \to_{a.s.} 0$;
(ii) $E[\sqrt{n}\rho_{L_p}(F_n,F)] = O(1)$ if $1 \le p < 2$ and $\int\{F(t)[1-F(t)]\}^{p/2}dt < \infty$, or $p \ge 2$.

Proof

(i) Since $[\rho_{L_p}(F_n,F)]^p \le [\rho_\infty(F_n,F)]^{p-1}[\rho_{L_1}(F_n,F)]$ and, by Theorem 5.1, $\rho_\infty(F_n,F) \to_{a.s.} 0$, 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

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

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

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

Proof (continued)

Since $[F_n(t) - F(t)]_- \le F(t)$ and $\int_{-\infty}^0 F(t)dt < \infty$ (Exercise 55 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 $\int_{-\infty}^0 [F_n(t) - F(t)]dt \to_{a.s.} 0$ implies

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

The result follows since we can similarly show

$$\int_0^\infty |F_n(t) - F(t)|dt \to_{a.s.} 0.$$

(ii) Omitted.

Nonparametric MLE

In §4.4 and §4.5, we have shown that the method of using likelihoods provides some asymptotically efficient estimators.

Can we use the method of likelihoods in nonparametric models?

This not only provides another justification for the use of the empirical c.d.f., but also leads to a useful method of deriving estimators in various (possibly non-i.i.d.) cases.

Proof (continued)

Since $[F_n(t) - F(t)]_- \le F(t)$ and $\int_{-\infty}^0 F(t)dt < \infty$ (Exercise 55 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 $\int_{-\infty}^0 [F_n(t) - F(t)]dt \to_{a.s.} 0$ implies

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

The result follows since we can similarly show

$$\int_0^\infty |F_n(t) - F(t)|dt \to_{a.s.} 0.$$

(ii) Omitted.

Nonparametric MLE

In §4.4 and §4.5, we have shown that the method of using likelihoods provides some asymptotically efficient estimators.

Can we use the method of likelihoods in nonparametric models?

This not only provides another justification for the use of the empirical c.d.f., but also leads to a useful method of deriving estimators in various (possibly non-i.i.d.) cases.

Nonparametric MLE

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

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

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 MLE of F.

Theorem 5.3

Let $X_1, ..., X_n$ be i.i.d. with $F \in \mathscr{F}$.
The empirical c.d.f. F_n maximizes the nonparametric likelihood function $\ell(G)$ over $G \in \mathscr{F}$.

Nonparametric MLE

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

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

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 MLE of F.

Theorem 5.3

Let $X_1, ..., X_n$ be i.i.d. with $F \in \mathscr{F}$.
The empirical c.d.f. F_n maximizes the nonparametric likelihood function $\ell(G)$ over $G \in \mathscr{F}$.

Proof

We only need to consider $G \in \mathscr{F}$ such that $\ell(G) > 0$.
Let $c \in (0,1]$ and $\mathscr{F}(c)$ be the subset of $\mathscr{F}$ 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 $G \in \mathscr{F}(c)$.
Define

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

where λ 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 $p_i > 0$, $i = 1,...,n$, $\sum_{i=1}^n p_i = c$.

Proof (continued)

This shows that

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

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

An alternative proof

It suffices to show that

$$\prod_{i=1}^{n} p_i \le \left(\frac{c}{n}\right)^n$$

for any $p_i > 0$, $i = 1, ..., n$, $\sum_{i=1}^{n} p_i = c$.
Let Y be a random variable taking value p_i with probability n^{-1}, $i = 1, ..., n$.
From Jensen's inequality,

$$\frac{1}{n}\sum_{i=1}^{n} \log p_i = E(\log Y) \le \log E(Y) = \log\left(\frac{1}{n}\sum_{i=1}^{n} p_i\right) = \log\left(\frac{c}{n}\right),$$

which establishes the result.

Proof (continued)

This shows that

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

which is maximized at $c = 1$ for any fixed n.

The result follows from $P_{F_n}(\{x_i\}) = n^{-1}$ for given $X_i = x_i$, $i = 1,...,n$.

An alternative proof

It suffices to show that

$$\prod_{i=1}^{n} p_i \le \left(\frac{c}{n}\right)^n$$

for any $p_i > 0$, $i = 1,...,n$, $\sum_{i=1}^{n} p_i = c$.

Let Y be a random variable taking value p_i with probability n^{-1}, $i = 1,...,n$.

From Jensen's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\log p_i = E(\log Y) \le \log E(Y) = \log\left(\frac{1}{n}\sum_{i=1}^{n} p_i\right) = \log\left(\frac{c}{n}\right),$$

which establishes the result.