Chapter 5: Estimation in Non-Parametric Models
Lecture 9: Empirical c.d.f. and empirical likelihoods

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 $n \to \infty$,

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

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

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

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).

The following result is useful.
Its proof is omitted.

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

Define sup-norm distance

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

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

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

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

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

The following result holds without any condition on F.

Theorem 5.1

Let F_n be the empirical c.d.f. of 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 $n \to \infty$;
(This means that $F_n(t) \to_{a.s.} F(t)$ uniformly in $t \in \mathscr{R}^d$)
(ii) $E[\sqrt{n}\rho_\infty(F_n,F)]^s = O(1)$ for any $s > 0$.
(This implies that $\sqrt{n}\rho_\infty(F_n,F) = O_p(1)$)

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 $2 - \varepsilon > 0$,

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

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

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 $1 \leq p < 2$ and $\int \{F(t)[1 - F(t)]\}^{p/2} dt < \infty$, or $p \geq 2$.

Proof

(i) Since $[\rho_{L_p}(F_n, F)]^p \leq [\rho_\infty(F_n, F)]^{p-1}[\rho_{L_1}(F_n, F)]$ 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

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

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

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

Since $[F_n(t) - F(t)]_- \leq F(t)$ 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.

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

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

Proof

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 $G \in \mathscr{F}(c)$:

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

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 $p_i = c/n$, $i = 1,...,n$, $\lambda = -(c/n)^{n-1}$, which 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

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

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

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

Empirical likelihoods

The nonparametric MLE can be extended to various situations with some modifications of $\ell(G)$ and/or constraints on p_i's.

Modifications of the likelihood $\ell(G)$ are called *empirical likelihoods*.

An estimator obtained by maximizing an empirical likelihood is then called a *maximum empirical likelihood estimator* (MELE).

Estimation of F with auxiliary information about F

In some cases we have some information about F.
For instance, suppose that there is a known Borel function u from $\mathscr{R}^d$ to $\mathscr{R}^s$ such that

$$\int u(x)dF = 0$$

For example, let $X_i = (y_i, z_i)$, y_i is the income for the current year, and z_i is the income for the current year.
From tax return, we know $E(z_i) = c$.
Then $u(x) = z - c$.
It is reasonable to expect that any estimate $\widehat{F}$ of F has property $\int u(x)d\widehat{F} = 0$, which is not true for the empirical c.d.f. F_n, 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 likelihoods, a natural solution is to put another constraint in the process of maximizing the likelihood.

That is, we maximize $\ell(G)$ subject to

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

where $p_i = P_G(\{x_i\})$.

Using the Lagrange multiplier method and an argument similar to the proof of Theorem 5.3, it can be shown that an MELE of F is

$$\widehat{F}(t) = \sum_{i=1}^{n} \widehat{p}_i I_{(-\infty,t]}(X_i),$$

where

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

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

$$\sum_{i=1}^{n} \widehat{p}_i u(X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{u(X_i)}{1 + \lambda_n^{\tau} u(X_i)} = 0.$$

To see that the last equation has a solution asymptotically, note that

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

and

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

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

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

Hence, using the same argument as in the proof of Theorem 4.17, we can show that there exists a unique sequence $\{\lambda_n(X)\}$ such that

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

Theorem 5.4

Let u be a Borel function on $\mathscr{R}^d$ satisfying $\int u(x)dF = 0$ and $\widehat{F}$ be the MELE of F.
Suppose that $U = \mathrm{Var}(u(X_1))$ is positive definite.
Then, for any m fixed distinct $t_1, ..., t_m$ in $\mathscr{R}^d$,

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

where

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

Σ is the covariance matrix of $\sqrt{n}[(F_n(t_1), ..., F_n(t_m)) - (F(t_1), ..., F(t_m))]$, $W = (W(t_1), ..., W(t_m))$, and $W(t_j) = E[u(X_1)I_{(-\infty,t_j]}(X_1)]$.

Remark

$\widehat{F}$ is asymptotically more efficient than F_n, because of the use of the information $\int u(x)dF = 0$.
$\widehat{F}$ is better when U is less variable and the covariance W is larger.

Proof of Theorem 5.4

We prove the case of $m = 1$ only.
Let $\bar{u} = n^{-1}\sum_{i=1}^{n} u(X_i)$.
It follows from the estimation equations and Taylor's expansion that

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

By the SLLN and CLT,

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

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

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

Proof of Theorem 5.4 (continued)

Thus,

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

The result follows from the CLT and the fact that

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

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 $X_{n_1+\cdots+n_j+1}, ..., X_{n_1+\cdots+n_{j+1}}$ 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 Borel functions.
A simple example is that $X_1, ..., X_{n_1}$ are sampled from F and $X_{n_1+1}, ..., X_{n_1+n_2}$ 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 $X_{n_1+1}, ..., X_{n_1+n_2}$ 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 $k = 2$, $(w_2 = w)$.
An empirical likelihood with $p_i = P_G(\{x_i\})$ is

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

An MELE of F can be obtained by maximizing this empirical likelihood subject to $p_i > 0$, $i = 1, ..., n$, and $\sum_{i=1}^{n} p_i = 1$.
Using the Lagrange multiplier method we can show that an MELE $\widehat{F}$ is as previously given with

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

where $\widehat{w}$ satisfies

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

An asymptotic result similar to that in Theorem 5.4 can be established.