Lecture Notes 12
Nonparametric Inference
See Chapters 7 and 20

Now we consider doing inference without assuming a parametric model. This is called *nonparametric inference*. Some examples we consider are:

1. Estimate the cdf F.

2. Estimate a density function $p(x)$.

3. Estimate a functional $T(P)$ of a distribution P for example $T(P) = \mathbb{E}(X) = \int x\, p(x)dx$.

1 The cdf

Given $X_1, \ldots, X_n \sim F$ where $X_i \in \mathbb{R}$ we use,

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

to estimate F. We saw earlier that

$$\mathbb{P}\Big(\sup_x |\widehat{F}(x) - F(x)| > \epsilon\Big) \leq 2e^{-2n\epsilon^2}.$$

Hence,

$$\sup_x |\widehat{F}(x) - F(x)| \xrightarrow{\mathrm{P}} 0$$

and

$$\sup_x |\widehat{F}(x) - F(x)| = O_P\left(\sqrt{\frac{1}{n}}\right).$$

It can be shown that this is the minimax rate of convergence. Also, we have a nonparametric confidence band:

$$\mathbb{P}(L_n(x) \leq F(x) \leq U_n(x) \text{ for all } x) \geq 1 - \alpha$$

where $L_n(x) = \widehat{F}_n(x) - \epsilon_n$, $U_n(x) = \widehat{F}_n(x) - \epsilon_n$ and

$$\epsilon_n = \sqrt{\frac{1}{2n}\log(2/\alpha)}.$$

2 Density Estimation

$X_1, \ldots, X_n$ are iid with density p where $X_i \in \mathbb{R}$. What happens if we try to do maximum likelihood? The likelihood is

$$L(p) = \prod_{i=1}^{n} p(X_i).$$

We can make this as large as we want by making p highly peaked at each X_i. So $\sup_p L(p) = \infty$ and the mle is the density that puts infinite spikes at each X_i. Thus likelihood is not very helpful here.

To proceed, we will need to put some restriction on p. For example

$$p \in \mathcal{P} = \left\{ p: \ p \geq 0, \ \int p = 1, \ \int |p''(x)|^2 dx \leq C \right\}.$$

The most commonly used nonparametric density estimator is probably the histogram. Another common estimator is the *kernel density estimator.* A *kernel* K is a symmetric density function with mean 0. The estimator is

$$\widehat{p}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h} \right)$$

where $h > 0$ is called the *bandwidth*.

The bandwidth controls the smoothness of the estimator. Larger h makes $\widehat{p}_n$ smoother. As a loss function we will use

$$L(p, \widehat{p}) = \int (p(x) - \widehat{p}(x))^2 dx.$$

The risk is

$$R = \mathbb{E}\left(L(p, \widehat{p})\right) = \int \mathbb{E}(p(x) - \widehat{p}(x))^2 dx = \int (b^2(x) + v(x)) dx$$

where

$$b(x) = \mathbb{E}(\widehat{p}(x)) - p(x)$$

is the bias and

$$v(x) = \mathrm{Var}(\widehat{p}(x)).$$

Theorem 1 *Suppose that $h \to 0$ as $n \to \infty$. The risk satisfies*

$$R_n = C_1 h^4 + \frac{C_2}{nh} + O\left(h^4 + \frac{1}{nh} \right)$$

for constants $C_1, C_2 > 0$. If $nh \to \infty$ as $n \to \infty$ then $R_n \to 0$. The risk is minimized by setting $h = Cn^{-1/5}$ for some $C > 0$. In this case $R_n = O(n^{-4/5})$.

Proof. Let

$$Y_i = \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

Then $\widehat{p}_n(x) = n^{-1}\sum_{i=1}^n Y_i$ and

$$\begin{aligned}
\mathbb{E}(\widehat{p}(x)) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right) = \mathbb{E}(Y_i) = \mathbb{E}\left(\frac{1}{h}K\left(\frac{X_i - x}{h}\right)\right) \\
&= \int \frac{1}{h}K\left(\frac{u-x}{h}\right)p(u)du \\
&= \int K(t)p(x+ht)dt \qquad\qquad \text{where } u = x + ht \\
&= \int K(t)\left(p(x) + htp'(x) + \frac{h^2t^2}{2}p''(x) + o(h^2)\right)dt \\
&= p(x)\int K(t)dt + hp'(x)\int tK(t)dt + \frac{h^2}{2}p''(x)\int t^2K(t)dt + o(h^2)dt \\
&= (p(x)\times 1) + (hp'(x)\times 0) + \frac{h^2}{2}p''(x)\kappa + o(h^2)
\end{aligned}$$

where $\kappa = \int t^2K(t)dt$. So $\mathbb{E}(\widehat{p}(x)) \approx p(x) + \frac{h^2}{2}p''(x)\kappa$ and

$$b(x) \approx \frac{h^2}{2}p''(x)\kappa.$$

Thus

$$\int b^2(x)dx = \frac{h^4}{4}\kappa^2\int (p''(x))^2dx = C_1h^4.$$

Now we compute the variance. We have

$$v(x) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right) = \frac{\mathrm{Var}Y_i}{n} = \frac{\mathbb{E}(Y_i^2) - (\mathbb{E}(Y_i))^2}{n}.$$

Now

$$\begin{aligned}
\mathbb{E}(Y_i^2) &= \mathbb{E}\left(\frac{1}{h^2}K^2\left(\frac{X_i - x}{h}\right)\right) \\
&= \int \frac{1}{h^2}K^2\left(\frac{u-x}{h}\right)p(u)du \\
&= \frac{1}{h}\int K^2(t)p(x+ht)dt \quad \mathrm{u = x + ht} \\
&\approx \frac{p(x)}{h}\int K^2(t)dt = \frac{p(x)\xi}{h}
\end{aligned}$$

where $\xi = \int K^2(t)dt$. Now

$$(\mathbb{E}(Y_i))^2 \approx \left(p(x) + \frac{h^2}{2}p''(x)\kappa\right)^2 = p^2(x) + O(h^2) \approx p^2(x).$$

So

$$v(x) = \frac{\mathbb{E}(Y_i^2)}{n} - \frac{(\mathbb{E}(Y_i))^2}{n} \approx \frac{p(x)}{nh} + p^2(x) = \frac{p(x)\xi}{nh} + o\left(\frac{1}{nh}\right) \approx \frac{p(x)\xi}{nh}$$

and

$$\int v(x)dx \approx \frac{C_2}{nh}.$$

Finally,

$$R \approx \frac{h^4}{4}\kappa^2 \int (p''(x))^2 dx + \frac{\xi}{nh} = C_1 h^4 + \frac{C_2}{nh}.$$

■

Note that

$$h \uparrow \quad \longrightarrow \quad \text{bias} \uparrow,\ \text{variance} \downarrow$$
$$h \downarrow \quad \longrightarrow \quad \text{bias} \downarrow,\ \text{variance} \uparrow.$$

If we choose $h = h_n$ to satisfy

$$h_n \to 0, \quad nh_n \to \infty$$

then we see that $\widehat{p}_n(x) \xrightarrow{\text{P}} p(x)$.

If we minimize over h we get

$$h = \left(\frac{\xi}{4nC}\right)^{1/5} = O\left(\frac{1}{n}\right)^{1/5}.$$

This gives

$$R = \frac{C_1}{n^{4/5}}$$

for some constant C_1.

Can we do better? The answer, based on minimax theory, is no.

Theorem 2 *Let*

$$\mathcal{P} = \left\{p : \int |p''(x)|^2 dx < M\right\}.$$

There is a constant a such that

$$\inf_{\widehat{p}} \sup_{p \in \mathcal{P}} R(p, \widehat{p}) \geq \frac{a}{n^{4/5}}.$$

We prove this in 10/36-702. So the kernel estimator achieves the minimax rate of convergence. The histogram converges at the sub-optimal rate of $n^{-2/3}$. There are many practical questions such as: how to choose h in practice, how to extend to higher dimensions etc. These are also discussed in 10/36-702.

3 Functionals

Let $X_1, \ldots, X_n \sim F$. Let $\mathcal{F}$ be all distributions. A map $T : \mathcal{F} \to \mathbb{R}$ is called a **statistical functional.** We write $\theta = T(F)$. We als write $\theta = T(P)$ where P is the distribution.

Notation. Let F be a distribution function. Let f denote the probability mass function if F is discrete and the probability density function if F is continuous. The integral $\int g(x)dF(x)$ is interpreted as follows:

$$\int g(x)dF(x) = \begin{cases} \sum_j g(x_j)p(x_j) & \text{if } F \text{ is discrete} \\ \int g(x)p(x)dx & \text{if } F \text{ is continuous.} \end{cases}$$

A **statistical functional** $T(F)$ is any function of of the cdf F. Examples include the mean $\mu = \int x\, dF(x)$, the variance $\sigma^2 = \int (x-\mu)^2 dF(x)$, the median $m = F^{-1}(1/2)$, and the largest eigenvalue of the covariance matrix Σ.

The **plug-in estimator** of $\theta = T(F)$ is defined by

$$\widehat{\theta}_n = T(\widehat{F}_n).$$

A functional of the form $\int a(x)dF(x)$ is called a **linear functional.** The empirical cdf $\widehat{F}_n(x)$ is discrete, putting mass $1/n$ at each X_i. Hence, if $T(F) = \int a(x)dF(x)$ is a linear functional then the plug-in estimator for linear functional $T(F) = \int a(x)dF(x)$ is:

$$T(\widehat{F}_n) = \int a(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} a(X_i).$$

Let $\widehat{\mathsf{se}}$ be an estimate of the standard error of $T(\widehat{F}_n)$.

Asymptotic Normality. If the functional F satisfies certain conditions, then

$$\frac{\widehat{\theta}_n - \theta}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1).$$

Thus, $\widehat{\theta}_n = T(\widehat{F}_n) \approx N(T(F), \widehat{\mathsf{se}}^2)$. In this case, an approximate $1-\alpha$ confidence interval for $T(F)$ is then

$$\widehat{\theta}_n \pm z_{\alpha/2}\, \widehat{\mathsf{se}}.$$

To test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

we can use the nonparametric version of the Wald statistic

$$W = \frac{\widehat{\theta}_n - \theta_0}{\text{se}}.$$

We reject H_0 if $|W| > z_{\alpha/2}$.

Example 3 (The mean) *Let* $\mu = T(F) = \int x\,dF(x)$. *The plug-in estimator is* $\widehat{\mu} = \int x\,d\widehat{F}_n(x) = \overline{X}_n$. *The standard error is* $\mathsf{se} = \sqrt{\mathsf{Var}(\overline{X}_n)} = \sigma/\sqrt{n}$. *If* $\widehat{\sigma}$ *denotes an estimate of* σ, *then the estimated standard error is* $\widehat{\mathsf{se}} = \widehat{\sigma}/\sqrt{n}$. *A Normal-based confidence interval for* μ *is* $\overline{X}_n \pm z_{\alpha/2}\,\widehat{\sigma}/\sqrt{n}$.

Example 4 (The variance) *Let* $\sigma^2 = \mathsf{Var}(X) = \int x^2\,dF(x) - \left(\int x\,dF(x)\right)^2$. *The plug-in estimator is*

$$\begin{aligned}
\widehat{\sigma}^2 &= \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2 && (1)\\
&= \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 && (2)\\
&= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2. && (3)
\end{aligned}$$

Example 5 (The skewness) *Let* μ *and* σ^2 *denote the mean and variance of a random variable* X. *The skewness — which measures the lack of symmetry of a distribution — is defined to be*

$$\kappa = \frac{\mathbb{E}(X-\mu)^3}{\sigma^3} = \frac{\int (x-\mu)^3 dF(x)}{\left\{\int (x-\mu)^2 dF(x)\right\}^{3/2}}.$$

To find the plug-in estimate, first recall that $\widehat{\mu} = n^{-1}\sum_{i=1}^n X_i$ *and* $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \widehat{\mu})^2$. *The plug-in estimate of* κ *is*

$$\widehat{\kappa} = \frac{\int (x-\mu)^3 d\widehat{F}_n(x)}{\left\{\int (x-\mu)^2 d\widehat{F}_n(x)\right\}^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu})^3}{\widehat{\sigma}^3}.$$

Example 6 (Correlation) *Let* $Z = (X, Y)$ *and let* $\rho = T(F) = \mathbb{E}(X-\mu_X)(Y-\mu_Y)/(\sigma_x\sigma_y)$ *denote the correlation between* X *and* Y, *where* $F(x,y)$ *is bivariate. We can write* $T(F) = a(T_1(F), T_2(F), T_3(F), T_4(F), T_5(F))$ *where*

$$\begin{array}{lll}
T_1(F) = \int x\,dF(z) & T_2(F) = \int y\,dF(z) & T_3(F) = \int xy\,dF(z)\\
T_4(F) = \int x^2\,dF(z) & T_5(F) = \int y^2\,dF(z) &
\end{array}$$

and

$$a(t_1, \ldots, t_5) = \frac{t_3 - t_1 t_2}{\sqrt{(t_4 - t_1^2)(t_5 - t_2^2)}}.$$

Replace F *with* $\widehat{F}_n$ *in* $T_1(F)$, ..., $T_5(F)$, *and take*

$$\widehat{\rho} = a(T_1(\widehat{F}_n), T_2(\widehat{F}_n), T_3(\widehat{F}_n), T_4(\widehat{F}_n), T_5(\widehat{F}_n)).$$

We get

$$\widehat{\rho} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \overline{X}_n)^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2}}$$

which is called the **sample correlation**.

Example 7 (Quantiles) *Let F be strictly increasing with density f. Let $T(F) = F^{-1}(p)$ be the p^{th} quantile. The estimate of $T(F)$ is $\widehat{F}_n^{-1}(p)$. We have to be a bit careful since $\widehat{F}_n$ is not invertible. To avoid ambiguity we define $\widehat{F}_n^{-1}(p) = \inf\{x : \ \widehat{F}_n(x) \geq p\}$. We call $\widehat{F}_n^{-1}(p)$ the p^{th}* **sample quantile**.

What if we do not know how to estimate the standard error. Then we use the *bootstrap* (stay tuned).

4 Nonparametric Confidence Interval For The Median

Suppose we want to find a confidence interval for the median θ of a distribution F. Let $Y_1, \ldots, Y_n \sim F$. Define

$$Z_i = \frac{\text{sign}(Y_i - \theta) + 1}{2}.$$

Note that

$$Z_i = \begin{cases} 1 & \text{if } Y_i > \theta \\ 0 & \text{if } Y_i < \theta. \end{cases}$$

Note that $\mathbb{P}(Z_i = 1) = 1/2$. Let $T = \sum_{i=1}^n Z_i$. Hence $T \sim \text{Binomial}(n, 1/2)$. Also, note that

$$T = \text{the number of } Y_i's \ > \theta.$$

Let k_1 and k_2 be chosen so that

$$\mathbb{P}(k_1 \leq \text{Binomial}(n, 1/2) \leq k_2) \geq 1 - \alpha.$$

Hence,

$$1 - \alpha \leq P(k_1 \leq T \leq k_2) = P(k_1 \leq (\text{the number of } Y_i's \ > \theta) \leq k_2).$$

Now

$$(\text{the number of } Y_i's \ > \theta) \geq k_1 \qquad \text{iff} \qquad \theta < Y_{(n-k_1+1)}$$

and

$$(\text{the number of } Y_i's \ > \theta) \leq k_2 \qquad \text{iff} \qquad Y_{(n-k_2)} \leq \theta.$$

So

$$1 - \alpha \leq P(Y_{(n-k_2)} \leq \theta \leq Y_{(n-k_1+1)}).$$

Therefore, $C_n = [Y_{(n-k_2)}, Y_{(n-k_1+1)}]$ is a nonparametric $1 - \alpha$ confidence interval for θ.

We can use Hoeffding's inequality to get expressions for k_1 and k_2. Let $S \sim \text{Binomial}(n, 1/2)$. Then

$$
\begin{aligned}
\mathbb{P}(S \geq k_2) &= \mathbb{P}\left(\frac{S}{n} - \frac{1}{2} \geq \frac{k_2}{n} - \frac{1}{2}\right) \\
&\leq \exp\left(-n(k_2/n - 1/2)^2\right).
\end{aligned}
$$

Set this to be less than $\alpha/2$ to get

$$
k_2 = \frac{n}{2} + \sqrt{n \log\left(\frac{2}{\alpha}\right)}.
$$

By a similar calculation,

$$
k_1 = \frac{n}{2} - \sqrt{n \log\left(\frac{2}{\alpha}\right)}.
$$