Lecture 3

Bootstrap

Let X be a real-valued random variable, that is a $(\mathcal{F}, \mathcal{B})$-measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$, where $(\Omega, \mathcal{F}, P)$ is the probability space associated with some random experiment. Consider n independent repetitions of this random experiment and denote $(x_1, \ldots, x_n)$ the resulting set of observations, called a **data set**. The corresponding random vector $\mathbf{X} = (X_1, \ldots, X_n) : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}^n)$ is called a **sample of size n from a population with distribution P**. By construction, a sample $\mathbf{X}$ is a vector of n independent, identically distributed random variables, which will be denoted by $X_1, \ldots, X_n \overset{i.i.d}{\sim} P$. The corresponding distribution function will be denoted by F and p.d.f by f.

Any measurable function S of $\mathbf{X}$, $S(\mathbf{X})$ is called **statistic**, if $S(\mathbf{X})$ has a known value for known $\mathbf{X}$. A sample $\mathbf{X}$ is a trivial statistic. To make inference about $S(\mathbf{X})$ (confidence intervals, hypothesis tests) one has to know the distribution of $S(\mathbf{X})$. However, it is often impossible to derive the exact distribution of $S(\mathbf{X})$, either because S is complex or because P is unknown (even though in many cases an asymptotic distribution is available). In many cases **bootstrap** is an attractive way to estimate the distribution of $S(\mathbf{X})$.

A set of probability measures P_θ on $(\Omega, \mathcal{F})$ indexed by a parameter $\theta \in \Theta$ is said to be a **parametric family** if and only if $\Theta \subset \mathbb{R}^d$ for some fixed positive integer d and each P_θ is a known probability measure when θ is known. A **parametric statistical model** refers to the assumption that $\mathbf{X} = (X_1, \ldots, X_n)$ is a sample from the population with distribution $P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ for a given parametric family. In the following our statistic $S(\mathbf{X})$ will be an estimator of a population characteristics θ (which might be a parameter of the distribution or e.g., median or moment).

If we knew P, we could sample many times from P to get many realisations of $S(\mathbf{X})$ and herewith the empirical distribution of $S(\mathbf{X})$. Recall that for random variables $Y_1, \ldots, Y_n$ the empirical distribution function is defined via $F_n(y) = n^{-1} \sum_{i=1}^n \mathbb{I}(Y_i \leq y)$. However, in practice P is unknown. The idea of the bootstrap is to sample from an empirical distribution function. Of course, there many ways to estimate the distribution function and to sample from it. If the sample of size n is from a continuous distribution, then each observation has a probability $1/n$ and sampling from F_n would be equivalent to draw with replacement from the sample.

Herewith is the bootstrap algorithm

1. Draw n times with replacement from $\mathbf{X}$ to get a bootstrap sample $\mathbf{X}_1^*$ of size n. Repeat R times to get R bootstrap samples $\mathbf{X}_1^*, \dots, \mathbf{X}_R^*$, each of size n.
2. Compute bootstrap statistics $S(\mathbf{X}_1^*), \dots, S(\mathbf{X}_R^*)$.
3. Make inference about θ based on $S(\mathbf{X}_1^*), \dots, S(\mathbf{X}_R^*)$.

How good are estimators (point or interval) based on the bootstrap sample? We consider bootstrap confidence intervals in detail. First recall the definition of a confidence interval.

Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample from a population with distribution $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$. Let $C(\mathbf{X})$ depend only on the sample $\mathbf{X}$ and $\theta \in \Theta$ be an unknown parameter of interest. If

$$\inf_{P \in \mathcal{P}} P(\theta \in C(\mathbf{X})) \geq 1 - \alpha$$

for a fixed constant $\alpha \in (0, 1)$, then $C(\mathbf{X})$ is called a **confidence set** for θ with **level of significance** $1 - \alpha$.

If the parameter θ is real-valued, then $C(\mathbf{X}) = \left[\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})\right]$, for a pair of real-valued statistics $\underline{\theta}$ and $\bar{\theta}$ is called a **confidence interval** for θ.

Example
Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ with unknown $\mu \in \mathbb{R}$ and known $\sigma^2 > 0$. Since $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, we get

$$1 - \alpha = P\left(-z_{1-\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = P\left(\overline{X} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right),$$

where z_α is the α-quantile of the standard normal distribution.
Hence, the confidence interval is given by $C(\mathbf{X}) = \left[\overline{X} - z_{1-\alpha/2}\sigma/\sqrt{n}, \overline{X} + z_{1-\alpha/2}\sigma/\sqrt{n}\right]$, which has length $2z_{1-\alpha/2}\sigma/\sqrt{n}$.

Note that since $\overline{X}$ is unbiased for μ and the transformation is linear, we can write this confidence band for μ as $[c_{\alpha/2}, c_{1-\alpha/2}]$, where c_α is the α-quantile of $\mathcal{N}(\overline{X}, \sigma^2/n)$, that is, $c_{\alpha/2} = \overline{X} - z_{1-\alpha/2}\sigma/\sqrt{n}$ and $c_{1-\alpha/2} = \overline{X} + z_{1-\alpha/2}\sigma/\sqrt{n}$,.

Therefore, a natural way to construct the bootstrap confidence interval is to use empirical quantiles of the bootstrap distribution of $S(\mathbf{X})$: compute $\hat{\theta}_i^* = S(\mathbf{X}_i^*)$, $i = 1, \dots, R$ bootstrap statistics and set the confidence interval for θ by $[\hat{\theta}_L^*, \hat{\theta}_U^*]$, where $\hat{\theta}_L^*$ and $\hat{\theta}_U^*$ are

$\lfloor R\alpha/2 \rfloor$-th and $\lfloor R(1-\alpha)/2 \rfloor$ value in the ordered list of $\hat{\theta}_i^*$. Such confidence intervals are called **bootstrap percentile** confidence intervals.

Let us consider more formally under which conditions such intervals work properly. Let $\hat{\theta}^* = S(\mathbf{X}^*)$ be a bootstrap estimator for θ and P^* denotes the distribution of $\mathbf{X}^*$ conditional on $\mathbf{X}$. Define $F_B(x) = P^*(\hat{\theta}^* \le x)$. Let $\hat{\theta}_L^* = F_B^{-1}(\alpha/2)$ (above we used the same notation for the empirical version). Suppose there exists an increasing transformation ϕ_n such that

$$P\{\phi_n(\hat{\theta}) - \phi_n(\theta) \le x\} = \Psi(x) \tag{1}$$

holds for all possible F (including empirical c.d.f), where $\Psi(x)$ is continuous, strictly increasing and symmetric about 0. Note that $\hat{\theta} = \hat{\theta}_n$ is a statistic (an estimator for θ) that depends on the sample of size n. When $\Psi = \Phi$ (c.d.f on $\mathcal{N}(0,1)$), then ϕ is called the normalizing and variance stabilizing transformation (standardisation as in the example with normal distribution is one possible ϕ). If ϕ_n and Ψ are known, then the lower confidence bound for θ has the form $\phi_n^{-1}\{\phi_n(\hat{\theta}) + \Psi^{-1}(\alpha/2)\}$. We show that this bound is the same as $\hat{\theta}_L^*$:

$$\Psi\{\phi_n(\hat{\theta}_L^*) - \phi_n(\hat{\theta})\} = P^*\left\{\phi_n(\hat{\theta}^*) - \phi_n(\hat{\theta}) \le \phi_n(\hat{\theta}_L^*) - \phi_n(\hat{\theta})\right\} = P^*(\hat{\theta}^* \le \hat{\theta}_L^*) = \alpha/2.$$

Hence, $\phi_n(\hat{\theta}_L^*) - \phi_n(\hat{\theta}) = \Psi^{-1}(\alpha/2)$ and $\hat{\theta}_L^* = \phi_n^{-1}\{\phi_n(\hat{\theta}) + \Psi^{-1}(\alpha/2)\}$.
Thus, the bootstrap percentile confidence intervals will have coverage probability of $1-\alpha$ if assumption (1) holds exactly for all n. If (1) holds approximately for large n, then $\hat{\theta}_L^*$ is asymptotically correct and the confidence interval performance depends on how good the approximation is.

Efron (1981) considered a more general assumption

$$P\{\phi_n(\hat{\theta}) - \phi_n(\theta) + z_0 \le x\} = \Psi(x), \tag{2}$$

where z_0 a constant that may depend on F and n. Since $\Psi(0) = 1/2$, z_0 is a kind of "bias" of $\phi_n(\hat{\theta})$. If ϕ_n, z_0 and Ψ are known, then the lower confidence bound for θ is $\phi_n^{-1}\{\phi_n(\hat{\theta}) + z_0 + \Psi^{-1}(\alpha/2)\}$. Under assumption (2), we obtain

$$F_B(\hat{\theta}) = P^*(\phi_n(\hat{\theta}^*) - \phi_n(\hat{\theta}) + z_0 \le z_0) = \Psi(z_0),$$

so that $z_0 = \Psi^{-1}\{F_B(\hat{\theta})\}$. Also, from (2)

$$\begin{aligned} 1-\alpha/2 &= \Psi\{-\Psi^{-1}(\alpha/2)\} = P^*\left\{\phi_n(\hat{\theta}^*) - \phi_n(\hat{\theta}) + z_0 \le -\Psi^{-1}(\alpha/2)\right\} \\ &= P^*\left(\hat{\theta}^* \le \phi_n^{-1}\{\phi_n(\hat{\theta}) - \Psi^{-1}(\alpha/2) - z_0\}\right), \end{aligned}$$

which implies $\phi_n^{-1}\{\phi_n(\hat\theta) - \Psi^{-1}(\alpha/2) - z_0\} = F_B^{-1}(1-\alpha/2)$. Since this equation holds for any α, we get for any $x \in (0,1)$ that

$$F_B^{-1}(x) = \phi_n^{-1}\{\phi_n(\hat\theta) + \Psi^{-1}(x) - z_0\}.$$

Since the lower confidence bound is given by $\phi_n^{-1}\{\phi_n(\hat\theta) + z_0 + \Psi^{-1}(\alpha/2)\}$, using the last equation, allows to rewrite the lower confidence bound as

$$F_B^{-1}\left[\Psi\{\Psi^{-1}(\alpha/2) + 2z_0)\}\right].$$

Assuming that Ψ is known (e.g., Φ) and using $z_0 = \Psi^{-1}\{F_B(\hat\theta)\}$, Efron (1981) suggests the bias-corrected lower confidence bound for θ

$$\hat\theta^*_{LC} = F_B^{-1}\left(\Psi\left[\Psi^{-1}(\alpha/2) + 2\Psi^{-1}\{F_B(\hat\theta)\}\right]\right).$$

Note that $\hat\theta^*_{LC}$ reduces to $\hat\theta^*_L$ if $F_B(\hat\theta) = 1/2$, i.e., $\hat\theta$ is the median of the bootstrap distribution F_B. Hence, $\hat\theta^*_{LC}$ is the bias corrected $\hat\theta^*_L$ and the bias correction is given by $2\Psi\{F_B(\hat\theta)\}$. Again, if (2) holds exactly, then the corresponding bias corrected bootstrap confidence interval will have coverage probability $1-\alpha$. If (2) holds approximately, then the performance of the bias corrected bootstrap confidence interval will depend on how good the approximation is.

Since in practice we rather use empirical version of F_B, to get the bias corrected bootstrap confidence intervals, set $\Psi = \Phi$ and calculate $\hat z_0 = \Phi^{-1}\left\{R^{-1}\sum_{i=1}^R \mathbb{I}(\hat\theta^*_i \le \hat\theta)\right\}$, then set $\alpha_1 = \Phi(z_{\alpha/2} + 2\hat z_0)$ and calculate $\hat\theta^*_{LC}$ as the $\lfloor R\alpha_1 \rfloor$ value in the ordered list of $\hat\theta^*_i$. Completely analogous, $\hat\theta^*_{UC}$ is the $\lfloor R\alpha_2 \rfloor$ value in the ordered list of $\hat\theta^*_i$, where $\alpha_2 = \Phi(z_{1-\alpha/2} + 2\hat z_0)$.

Bias corrected bootstrap confidence intervals improve bootstrap percentile confidence intervals by taking into account the bias. However, there are still many cases where assumption (2) is not fulfilled. Efron (1987) proposed a bootstrap accelerated bias-corrected confidence intervals, that further improves bias corrected bootstrap confidence intervals. One simple version of such confidence intervals is given by setting

$$\begin{aligned}\alpha_1 &= \Phi\left(\hat z_0 + \frac{\hat z_0 + z_{\alpha/2}}{1 - \hat a(\hat z_0 + z_{\alpha/2})}\right)\\ \alpha_2 &= \Phi\left(\hat z_0 + \frac{\hat z_0 + z_{1-\alpha/2}}{1 - \hat a(\hat z_0 + z_{1-\alpha/2})}\right),\end{aligned}$$

where

$$\hat a = \frac{\sum_{i=1}^n (\bar\theta_J - \hat\theta_{(i)})^3}{6\left\{\sum_{i=1}^n (\bar\theta_J - \hat\theta_{(i)})^2\right\}^{3/2}},$$

with $\bar{\theta}_J = n^{-1} \sum_{i=1}^{n} \hat{\theta}_{(i)}$, for $\hat{\theta}_{(i)}$ as the estimator of θ obtained without observation i, i.e., $\hat{\theta}_{(i)} = S(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.