Section 4*

Econ 240A - Second Half

Ingrid Haegele†
University of California, Berkeley

November 9, 2018

*These section notes rely on the notes prepared and revised by Markus Pelger, Raffaele Saggio and Seongjoo Min.

†E-mail: inha@berkeley.edu

1 Bayesian Bootstrap

In the section 2 we talked about the fact that prediction requires knowledge of the data distribution function. While a risk-minizing procedure in the absence of this knowledge is generally unavailable, we talked about ways on how to do prediction when only a random sample of size N from the target population is available. While our random sample does provide us with some information about the relative plausibility of different logically possible data distributions, it does not perfectly reveal the properties of the underlying distribution. Instead, we operate under some degree of uncertainty about the true population distribution.

This way of thinking about the data distribution as a random draw from the set of all logically possible data distributions is a great motivation for introducing the Bayesian perspective. We will first learn more about the concepts behind Bayesian inference and then see how this distinguishes the Bayesian Bootstrap from the frequentist Bootstrap. In order to provide you with the necessary tools to perform the Bayesian Bootstrap yourself, this Section focusses on developing important intution. For technical and mathematical details, please refer to the lecture notes.

1.1 Bayesian Inference

In the frequentist framework, we tend to think that the data are repeated samples from some distribution, and the parameter θ is some (unknown) fixed quantity (so it is non-random) that governs the distribution. That is, given a random process, we attempt to assign probability based on the "frequency" from the samples. For instance, the $1 - \alpha$ confidence interval is a frequestist notion: it uses the repeated samples to construct a random interval that contains the parameter value with probability $1 - \alpha$.

In constrast, Bayesian statistics view both the data and θ as random. In particular, we think that θ is a realization of a random variable $\Theta \sim \Pi$. $\Pi(\cdot)$ is called the **prior distribution**, which represents one's belief or hypothesis on the parameter before data are observed. Using information from the data, we update the prior to construct the **posterior distribution**, $\Pi(\cdot|D = d)$, where d denotes the observed data. For instance, we can use the posterior distribution to construct a credible interval, which is the Bayesian

counterpart of confidence interval, which is a fixed (non-random) interval that θ lies in with a certain probability.

We can obtain the posterior distribution using **Bayes' Theorem**, which states the following:

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{A \cap B\}}{\mathbb{P}\{B\}}$$

where A and B are subsets of the sample space Ω such that $\mathbb{P}\{B\} \neq 0$. Letting $\{C_i\}_i$ be a partition of Ω and using laws of probability, we have

$$= \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\sum_i \mathbb{P}\{B \cap C_i\}} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\sum_i \mathbb{P}\{B|C_i\}\mathbb{P}\{C_i\}}$$

Now, suppose that the distributions admit density functions. Then we can write the posterior density as

$$\pi(\theta|d) = \frac{p(d|\theta)\pi(\theta)}{p(d)} = \frac{p(d|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} p(d|\theta)\pi(\theta)d\theta} \propto_\theta p(d|\theta)\pi(\theta)$$

where $\pi(\cdot|d)$ is the conditional density of $\Theta|D = d$ (the posterior density), $\pi(\cdot)$ is the marginal density of Θ (the prior density), $p(\cdot)$ is the marginal density of D, and $p(\cdot|\theta)$ is the conditional density of $D|\Theta = \theta$. Note that $p(d|\theta)$ is the likelihood function evaluated at θ. The sign $\propto_\theta$ means "proportional" to the terms containing θ. This is useful because the term on the denominator is a normalizing constant, which ensures that the posterior density integrates to 1.

Further suppose that the data are iid, that is $d = (x_1, x_2, ...x_N)'$, $X_i|\Theta = \theta \overset{\text{iid}}{\sim} F(\cdot\,;\theta)$ with density function $f(\cdot\,;\theta)$. Then we have

$$\pi(\theta|d) \propto_\theta \Big(\prod_{i=1}^{N} f(x_i\,;\theta)\Big)\pi(\theta)$$

This is tractable given $f(\cdot;\theta)$ and our beliefs on $\pi(\cdot)$.

We typically choose the prior so that the prior and the posterior belong to the same family of distributions. In this case, we call it the **conjugate prior**. Because the posterior density is equal to the product of the likelihood and the prior density, the choice of prior depends on the form of the likelihood function. For example, if $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ where μ is a random variable, and σ^2 is known (fixed), the conjugate prior is a Gaussian distribution. If μ is known and σ^2 is a random variable, then the conjugate prior is an inverse gamma.

1.2 Frequentist vs Bayesian Bootstrap: Intuition

Bootstraping refers to a resampling technique where inferences are made based on simulated sets of samples, generated by re-drawing, with replacement, from the observed data. The (frequentist) bootstrap takes the data as a reasonable approximation to the unknown population distribution. Therefore, the sampling distribution of a statistic (a function of the data) can be approximated by repeatedly resampling the observations with replacement and computing the statistic for each sample.

To develop intuition, let $y = (y_1, ...y_n)$ denote the original data. We perform the bootstrap simulation by drawing B random samples of N with replacement from the data, where the b^{th} bootstrap sample is denoted as $y^b = ({y^b}_1, ...{y^b}_n)$. For each sample, we compute the test statistic of interest.

To fix ideas, consider the mean as an example of a test statistic. The mean of the bootstrap sample is given by

$$m_b = \frac{1}{n}\sum_{i=1}^{n} {y^b}_i$$

.

After performing B of these bootstrap replications, we use the distribution of m_b to approximate the sampling distribution from the unknown population.

Bayesian bootstrapping, as the name reveals, uses Bayesian learning to construct the posterior distribution, and the posterior is used for resampling. This particular way of performing Bootstrap simulation was introduced by Rubin (1981).

Under the Bayesian perspective, how would we compute our mean m_b? Again we let y denote the original data and y^b the bootstrap sample. In addition, we know that in each bootstrap sample y^b, each observation y_i occurs anywhere from 0 to n times. We can denote the number of times y_i occurs in y^b as ${h^b}_i$. Thus, ${h^b}_i \in \{0, 1, ..., n-1, n\}$ and we know that $\sum_{i=1}^{N} {h^b}_i = n$. Given h^b, we can construct a collection of nonnegative weights that sum to one: $w^b = \frac{h^b}{n}$, where ${w^b}_i = \frac{{h^b}_i}{n}$. Given this notation, we can express the mean of the bootstrap sample as

$$m_b = \sum_{i=1}^{n} {w^b}_i y_i$$

.

We can compute m_b by drawing w^b from its distribution and computing the dot product with y. Note, the distribution of w^b is determined by the way obervations are chosen for a bootstrap sample. In this setup, h^b has a multinominal distribution, and therefore the sampling distribution for the observations is multinomial. This implies that the prior distribution of the weights is a Dirichlet distribution.[1] Given Bayes rule, the multinominal likelihood and the Dirichlet prior, this implies that the posterior distribution of the weights is itself a member of the Dirichlet family. Remember, the Dirichlet is what is known as the conjugate prior of the multinomial distribution: when our prior beliefs about the distribution of probability mass across a finite set of support points takes the Dirichlet form, then our posterior beliefs, after observing a random sample, will take the same form.

Consequently, in order to perform the Bayesian Bootstrap, we first draw a random sample of N of weights and then compute the test statistic of interest. We repeat these steps B times.

1.3 Bayesian Bootstrapping: Formal Setup

After providing some basic intution on the procedure underlying the Bayesian procedure, a formal setup follows. Let $\mathbf{Z} = (\mathbf{Y}, \mathbf{X}')'$ be a discrete random vector with support $\{z_1, z_2, \ldots, z_j\}$, and $Z = (Z_1, Z_2, \ldots, Z_N)'$ be a random sample. Given $\theta = (\theta_1, \theta_2, \ldots, \theta_J)'$, let the conditional distribution be the following

$$\Pr(\mathbf{Z} = z_j|\theta) = \theta_j,\ j = 1, 2, \ldots, J$$

where

$$\theta \in \Theta = \mathbb{S}^{J-1} = \left\{\theta \in \mathbb{R}^J : \sum_{j=1}^{J} \theta_j = 1,\ \theta_j \geq 0,\ j = 1, 2, \ldots, J\right\}$$

$\mathbb{S}^{J-1}$ is called the $J-1$ simplex.

Define $N_j = \sum_{i=1}^{N} 1(Z_i = z_j),\ j = 1, 2, \ldots, J$, the number of observations that have value z_j. This tells us that $\mathbf{Z}|\theta$ has a multinomial distribution with pmf

$$f(Z|\theta) = \frac{N!}{N_1!N_2!\ldots N_J!} \prod_{j=1}^{J} \theta_j^{N_j}$$

[1]Note, the Dirichlet distribution is a distribution over proportions, in our case the weights, that sum to 1.

It turns out that the Dirichelet distribution[2] is a conjugate prior. So the marginal density of θ at p (the prior) is

$$\pi(p;\alpha) = \frac{\Gamma\Big(\sum\limits_{j=1}^{J}\alpha_j\Big)}{\prod\limits_{j=1}^{J}\Gamma(\alpha_j)}\prod_{j=1}^{J}p_j^{\alpha_j-1}$$

where $\alpha_j > 0,\ j = 1, 2, \ldots, J$ are parameters of the Dirichelet distribution.

Computation reveals that (as expected) the posterior distribution is also Dirichelet. The conditional density of θ given $\mathbf{Z} = Z$ is

$$\pi(p|Z;\alpha) = \frac{\Gamma\Big(\sum\limits_{j=1}^{J}N_j+\alpha_j\Big)}{\prod\limits_{j=1}^{J}\Gamma(N_j+\alpha_j)}\prod_{j=1}^{J}p_j^{N_j+\alpha_j-1}$$

1.4 Posterior Simulation

Let us return to the linear regression model, $\mathbb{E}^*[Y|X] = X'\beta$ for a concrete application example. We are interested in the linear regression coefficient $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$. To accommodate the Bayesian view, we can write β as

$$\beta = \beta(\theta) = \Big[\sum_{j=1}^{J}\theta_j x_j x_j'\Big]^{-1}\Big[\sum_{j=1}^{J}\theta_j x_j y_j\Big]$$

If we knew θ with certainty we could proceed as in the previous chapters. Unfortunately we do not know which value of θ indexes the sample population and hence

$\beta(\theta)$. In fact, θ is a random variable. That means we have to compute the expectation of β using the posterior distribution.

Our posterior for θ summarizes our beliefs, after observing Z = z, about the relatively plausibility of all possible joint distributions of X and Y . The posterior mean of the vector of coefficients indexing the best linear predictor of Y given X is

$$\bar{\beta} = \mathbb{E}[\beta(\theta)|Z;\alpha] = \int_{\mathbb{S}^{J-1}}\Big[\sum_{j=1}^{J}p_j x_j x_j'\Big]^{-1}\Big[\sum_{j=1}^{J}p_j x_j y_j\Big]\pi(p|Z;\alpha)dp$$

[2]The Dirichelet distribution is the multivariate version of the Beta distribution.

Computing this integral is very cumbersome. So we turn to simulation, drawing p from the posterior distribution.

How can we simulate from a posterior distribution? Note, that we assumed that the prior distribution and consequently also the posterior distribution are Dirichlet distributions. In order to randomly draw from the Dirichlet posterior distribution, we can use a common trick in Econometrics. For a formal derivation, please see the lecture notes. A short intuitive overview can be given as follows:

First, let $\{W_j\}_{j=1}^{J}$ be J independent random variables with $W_j \sim \text{Gamma}(\alpha_j, 1)$. Define

$$V_j = \frac{W_j}{\sum_{j=1}^{J} W_j}$$

It turns out that $(V_1, \ldots, V_J)$ coincides with a random draw from a Dirichlet distribution with parameter $\alpha = (\alpha_1, \ldots, \alpha_1)^\top$.

For computatinal ease, we can also consider $W^*{}_i \sim \text{Gamma}(1,1)$ and $V^*{}_i = \frac{W^*{}_i}{\sum_{i=1}^{N} W^*{}_i}$, which leads to an alternative representation for the posterior distribution. This implies that in order to sample from a Dirichlet posterior distribution, we can take random draws of the weights from a Gamma$(1,1)$ distribution.

We motivated this part by pointing out the necessity to simulate the posterior distribution of the coefficient vector indexing the best linear predictor of Y given X due to the randomness of θ. Using the second representation using the Gamma$(1,1)$ distribution, the expression of a random draw corresponds to the weighted least squares fit of Y onto X with the vector $V^* = (V^*{}_1, \ldots, V^*{}_N)$ containing the weights.

$$\beta = \Big[\sum_{i=1}^{J} V_i^* X_i X_i^\top\Big]^{-1} \times \Big[\sum_{i=1}^{J} V_i^* X_i Y_i\Big]$$

1.5 Practical Implementation

Steps to estimate β using Bayesian bootstrapping

1. Draw an N-vector of independent Gamma$(1,1)$ random variables. Form the sum normalized vector $V^*_{(b)}$

2. Compute

$$\hat{\beta}_{(b)} = \Big[\sum_{i=1}^{N} [V^*_{(b)}]_i X_i X_i^\top\Big]^{-1} \Big[\sum_{i=1}^{N} [V^*_{(b)}]_i X_i Y_i\Big]$$

3. Repeat the process for $b = 1, 2, \ldots, B$.