MTH 511a - 2020: Lecture 22

Instructor: Dootika Vats

1 Resampling Methods

1.1 Bootstrapping

We have discussed cross-validation, which we use to choose model tuning parameters. However, once the final model is fit, we would like to make *inference*. That is, we want to account for the variability of the final estimators obtained.

If our estimates are MLEs then we know that under certain important conditions, MLEs have asymptotic normality, that is

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \overset{d}{\to} N(0, \sigma^2_{\text{MLE}})\,,$$

where σ^2_{MLE} is the inverse Fisher information. Then, if we can estimate σ^2_{MLE}, we can construct asymptotically normal confidence intervals:

$$\hat{\theta}_{\text{MLE}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{MLE}}{n}}\,.$$

We can also conduct hypothesis tests etc and go on to do regular statistical analysis. But sometimes we cannot use an asymptotic distribution:

1. when our estimates are not MLEs, like ridge and bridge regression
2. when the assumptions for asymptotic normality are not satisfied (I haven't shared these assumptions)
3. when n is not large enough for asymptotic normality to hold

We will try to approximate the distribution $\hat{\theta}$ using *Boostrap*, and from there we will obtain a confidence intervals.

Suppose $\hat{\theta}$ is some estimator of θ from sample $X_1, \dots, X_n \sim F$. Then since $\hat{\theta}$ is random it has a sampling distribution G_n that is unknown. If asymptotic normality holds, then $G_n \approx N(\cdot, \cdot)$ for large enough n, but in general we may not know much about G_n. If we could obtain many similar datasets, we could obtain an estimate from each dataset:

$$\hat{\theta}_1, \dots, \hat{\theta}_B \overset{iid}{\sim} G_n \,.$$

Once we have B realizations from G_n, we can easily estimate characteristics about G_n, like the overall mean, variance, quantiles, etc.

Thus, in order to learn things about the sampling distribution G_n, our goal is to draw more samples of such data. But this, of course is not easy in real-data scenarios. We could obtain more Monte Carlo datasets from F, but we typically do not know the true F. Instead of obtaining typical Monte Carlo datasets, we will "resample" from our current dataset. This would give us an approximate sample from our distribution G_n, and we could estimate characteristics of this distribution! This resampling using information from the current data is called *bootstrapping*. We will study two popular bootstrap methods: *nonparameteric bootstrap* and *parametric bootstrap*.

1.1.1 Nonparametric Bootstrap

In nonparametric bootstrap, we resample data of size n from within the sample of Xs (with replacement) and obtain estimates of θ using these samples. That is

$$\begin{aligned}
\text{Bootstrap sample 1: } & X^*_{11}, X^*_{21}, \dots, X^*_{n1} \Rightarrow \hat{\theta}^*_1 \\
\text{Bootstrap sample 2: } & X^*_{12}, X^*_{22}, \dots, X^*_{n2} \Rightarrow \hat{\theta}^*_2 \\
& \vdots \\
\text{Bootstrap sample B: } & X^*_{1B}, X^*_{2B}, \dots, X^*_{nB} \Rightarrow \hat{\theta}^*_B \,.
\end{aligned}$$

Each sample is called a bootstrap sample, and there are B bootstrap samples. Now, the idea is that $\hat{\theta}^*_1, \dots \hat{\theta}^*_B$ are B approximate samples from the distribution of $\hat{\theta}, G_n$.

We want to construct a $100(1-\alpha)\%$ confidence interval for $\hat{\theta}$. A confidence interval (L, U) is an interval such that

$$\Pr((L,U) \text{ contains } \theta) = 1 - \alpha \,.$$

Note that here L and U are random and θ is fixed. We can find the confidence interval by looking at the quantiles of the obtained bootstrap estimates. So if we order the bootstrap estimates

$$\hat{\theta}^*_{(1)} < \hat{\theta}^*_{(2)} < \cdots < \hat{\theta}^*_{(B)} \,,$$

and set

$$L = \hat{\theta}^*_{\lfloor \alpha/2*B \rfloor} \quad \text{ and } U = \quad \hat{\theta}^*_{\lfloor 1-\alpha/2*B \rfloor} \,.$$

Then $\left(\hat{\theta}^*_{\lfloor \alpha/2*B \rfloor}, \hat{\theta}^*_{\lfloor 1-\alpha/2*B \rfloor}\right)$, is a $100(1-\alpha)\%$ bootstrap confidence interval.

Example 1 (Mean of Gamma$(a, 1)$). Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Gamma$(a, 1)$. The mean of this distribution is $\theta = a$. Consider estimating θ with the sample mean

$$\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim G_n$$

Although a central limit theorem holds, so that $G_n \approx N$ for large n. However, we may not have many samples available in order to make confidence intervals. Thus, instead here we implement the nonparametric bootstrap:

$$\begin{gathered} X^*_{11}, X^*_{21}, \ldots, X^*_{n1} \Rightarrow \bar{X}^*_1 \\ X^*_{12}, X^*_{22}, \ldots, X^*_{n2} \Rightarrow \bar{X}^*_2 \\ \vdots \\ X^*_{1B}, X^*_{2B}, \ldots, X^*_{nB} \Rightarrow \bar{X}^*_B \\ . \end{gathered}$$

And find $\alpha/2$ and $1-\alpha/2$ sample quantiles from $\bar{X}^*_i, i = 1, \ldots, B$.

1.1.2 Parametric Bootstrap

Suppose $X_1, \ldots, X_n \sim F(\theta)$, where θ is a parameter we can estimate. Let $\hat{\theta}$ be a chosen estimator of θ. Instead of resampling within our data, in *parametric* bootstrap, we use our estimator of θ to obtain computer generated samples from $F(\hat{\theta})$:

$$\begin{gathered} X^*_{11}, X^*_{21}, \ldots, X^*_{n1} \sim F(\hat{\theta}) \Rightarrow \hat{\theta}^*_1 \\ X^*_{12}, X^*_{22}, \ldots, X^*_{n2} \sim F(\hat{\theta}) \Rightarrow \hat{\theta}^*_2 \\ \vdots \\ X^*_{1B}, X^*_{2B}, \ldots, X^*_{nB} \sim F(\hat{\theta}) \Rightarrow \hat{\theta}^*_B \end{gathered}$$

And again, we find the $\alpha/2$ and $1-\alpha/2$ quantiles of the $\hat{\theta}^*_i$s so that $\left(\hat{\theta}^*_{\lfloor \alpha/2*B \rfloor}, \hat{\theta}^*_{\lfloor 1-\alpha/2*B \rfloor}\right)$, is a $100(1-\alpha)\%$ bootstrap confidence interval.

Example 2 (Mean of Gamma$(a, 1)$). Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Gamma$(a, 1)$. And let $\hat{\theta} = \bar{X}$ be the chosen estimator of a. A parametric bootstrap estimator does:

$$\begin{gathered} X^*_{11}, X^*_{21}, \ldots, X^*_{n1} \sim \text{Gamma}(\bar{X}, 1) \Rightarrow \bar{X}^*_1 \\ X^*_{12}, X^*_{22}, \ldots, X^*_{n2} \sim \text{Gamma}(\bar{X}, 1) \Rightarrow \bar{X}^*_2 \\ \vdots \\ X^*_{1B}, X^*_{2B}, \ldots, X^*_{nB} \sim \text{Gamma}(\bar{X}, 1) \Rightarrow \bar{X}^*_B\,. \end{gathered}$$

And find $\alpha/2$ and $1-\alpha/2$ sample quantiles from $\bar{X}^*_i, i = 1, \ldots, B$.

Let us implement this Gamma example in detail. Notice that a CLT holds here with asymptotic variance a, so that as $n \to \infty$.

$$\sqrt{n}(\bar{X} - a) \xrightarrow{d} N(0, a)\,.$$

So that an asymptotic 95% confidence interval is

$$\bar{X} \pm z_{.975}\sqrt{\frac{\bar{X}}{n}}\,.$$

We will make four comparisons:

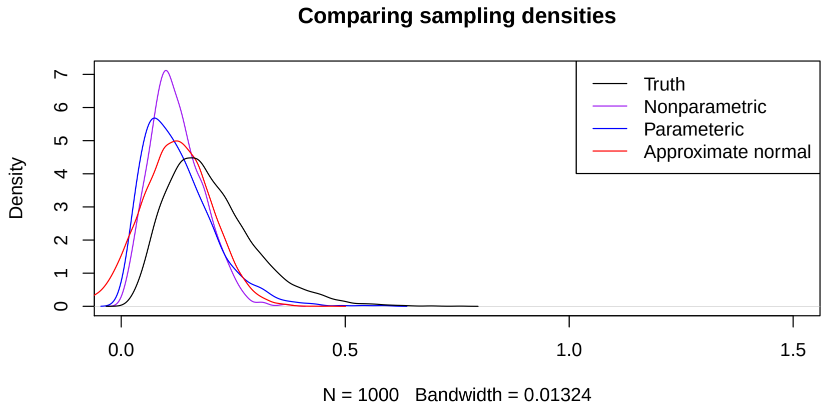
- The empirical distribution of nonparametric bootstrap estimates
- The empirical distribution of parametric bootstrap estimates
- The large sample normal distribution $N(a, a/n)$
- The true sampling distribution

```
###########################################
## Bootstrap for the Gamma distribution
###########################################
set.seed(10)
n <- 20
a <- .20
my.samp <- rgamma(n, shape = a, rate = 1)

barx <- mean(my.samp)

B <- 1e3 # number of bootstrap samples. 1e3 is standard.
boot.np <- numeric(length = B)
boot.p <- numeric(length = B)
for(b in 1:B)
{
  boot.samp.np <- sample(my.samp, replace = TRUE) # NP Bootstramp samples
  boot.np[b] <- mean(boot.samp.np) # NP Bootstrap estimator

  boot.samp.p <- rgamma(n, shape = barx, rate = 1) #parametric bootstrap
     samples
  boot.p[b] <- mean(boot.samp.p) # P bootstrap estimator
}

# 95% Bootstrap confidence interval
quantile(boot.np, probs = c(.025, .975)) # nonparameteric
#     2.5%     97.5%
#0.03058783 0.24917142

quantile(boot.p, probs = c(.025, .975)) #parametric
#     2.5%     97.5%
```

```
#0.02262155 0.31736283

# 95% asymptotic CI
c( barx - qnorm(.975)*sqrt(barx/n), barx + qnorm(.975)*sqrt(barx/n) )
#[1] -0.03029098 0.27848545

# Simulate repeated estimates to construct a 95% CI
true.samp <- numeric(length = 1e4)
for(i in 1:1e4)
{
  samp <- rgamma(n, shape = a, rate = 1)
  true.samp[i] <- mean(samp)
}
quantile(true.samp, probs = c(.025, .975))
#      2.5%      97.5%
#0.05446254 0.43878948
```

Since the sample size is low, all methods give different estimates. Certainly the CI based on the CLT is quite off since it proposed negative values in the interval, which is invalid. The parametric bootstrap is the closest to the truth. We can also compare their empirical densities against the truth.

```
plot(density(boot.np), col = "purple", xlim = c(0,1.5),
     main = "Comparing sampling densities")
lines(density(boot.p), col = "blue")
lines(density(rnorm(1e4, mean = barx, sd = sqrt(barx/n))), col = "red")
lines(density(true.samp))
legend("topright",lty = 1, legend = c("Truth", "Nonparametric",
    "Parameteric", "Approximate normal"), col = c("black", "purple", "blue",
    "red"))
```



Clearly the nonparametric estimator has lower than the real variance and this is the consequence of the really small sample size. The approximate normal interval assumes a symmetric sampling distribution, which is clearly not true.

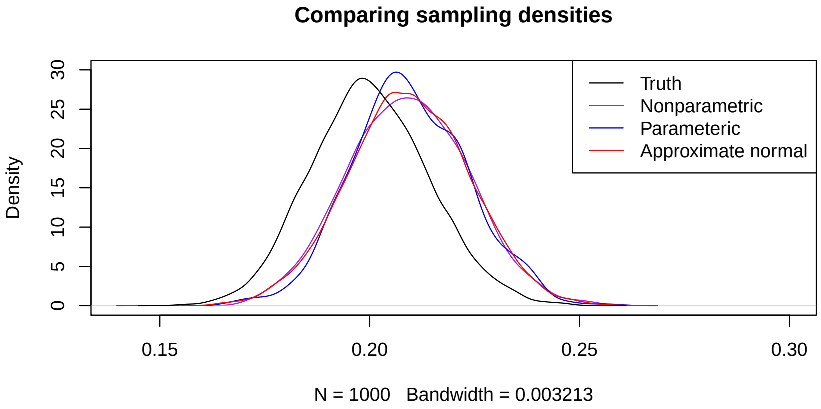
We can repeat the same thing for a larger n. Here, the asymptotic normal distribution coincides with the sampling distributions obtained via both bootstrap methods.

```
n <- 1000
a <- .20
my.samp <- rgamma(n, shape = a, rate = 1)

barx <- mean(my.samp)

B <- 1e3 # number of bootstrap samples
boot.np <- numeric(length = B)
boot.p <- numeric(length = B)

for(b in 1:B)
{
  boot.samp.np <- sample(my.samp, replace = TRUE) # NP Bootstramp samples
  boot.np[b] <- mean(boot.samp.np) # NP Bootstrap estimator

  boot.samp.p <- rgamma(n, shape = barx, rate = 1) #parametric bootstrap
      samples
  boot.p[b] <- mean(boot.samp.p) # P bootstrap estimator
}

# 95% Bootstrap confidence interval
quantile(boot.np, probs = c(.025, .975)) # nonparameteric
#     2.5%    97.5%
#0.1812819 0.2370416

quantile(boot.p, probs = c(.025, .975)) #parametric
#     2.5%    97.5%
#0.1832509 0.2374146

# 95% asymptotic CI
c( barx - qnorm(.975)*sqrt(barx/n), barx + qnorm(.975)*sqrt(barx/n) )
#[1] 0.1809254 0.2376330

# Simulate repeated estimates to construct a 95% CI
true.samp <- numeric(length = 1e4)
for(i in 1:1e4)
{
  samp <- rgamma(n, shape = a, rate = 1)
  true.samp[i] <- mean(samp)
}
quantile(true.samp, probs = c(.025, .975))
```

```
#     2.5%     97.5%
#0.1736390 0.2289925
```

All the estimated intervals are similar, however, they differ slightly from the true interval, which is natural, since the true interval is centered around the ground truth, and we are centered around $\bar{X}$.

```
plot(density(boot.np), col = "purple", xlim = c(0,1.5),
     main = "Comparing sampling densities")
lines(density(boot.p), col = "blue")
lines(density(rnorm(1e4, mean = barx, sd = sqrt(barx/n))), col = "red")
lines(density(true.samp))
legend("topright",lty = 1, legend = c("Truth", "Nonparametric",
    "Parameteric", "Approximate normal"), col = c("black", "purple", "blue",
    "red"))
```



2 Questions to think about

- We set $B = 1000$ in our experiments. What would happen if we increase of decrease B?
- How will you used bootstrapping to obtain confidence intervals for bridge regression coefficients?
- Do we need bootstrapping to obtain confidence intervals for ridge regression estimates?