Jacknife and Bootstrap

UBCO MDS — DATA 571

Intro

- We've now seen cross validation as a way of predicting the true MSE of a model.
- We previously largely ignored all the estimated models during the cross-validation process: we were only seeking an estimate of the long-run MSE of the model.
- The "jackknife" approach can actually use additional information gained from those CV fits to provide insights into the parameter estimates.

Jackknife

- Essentially LOOCV applied to estimating the bias and variance of parameter estimates. Here it is algorithmically:

  1. Let $i = 1$ be the index of the first observation
  2. Remove observation i from the data
  3. Estimate your assumed model, report any parameter estimates, say $\hat{\alpha}_i$
  4. Set $i = i + 1$, return to 2 if $i \leq n$.
  5. Now we have n estimates of parameter α. We can estimate the standard error and bias of an estimator $\hat{\alpha}$ by

$$\hat{SE}(\hat{\alpha}) = \sqrt{\frac{(n-1)}{n}\sum_{i=1}^{n}(\hat{\alpha}_i - \bar{\hat{\alpha}})^2} \qquad \hat{Bias}(\hat{\alpha}) = (n-1)(\bar{\hat{\alpha}} - \hat{\alpha})$$

where $\bar{\hat{\alpha}} = \sum_{i=1}^{n} \hat{\alpha}_i / n$

Traditional Example

- Before applying the jackknife in a modelling scenario, let's look at it from a simple, intro-stats-type example
- Estimating μ from a normal distribution!
- We generate a sample of size 25 from the normal distribution with the following (known) parameters $\mu = 0$ and $\sigma = 10$
- $\bar{X}$ is a traditional estimator for μ (among infinitely many estimators)

Traditional Example

- So, what is $SE(\bar{X})$? . . . 2
- We know this because X and $\bar{X}$ are especially neatly intertwined when X is normal.
- But suppose we lived in a world void of nice statistical theory (note that this is the case for several estimators)...the jackknife provides us a way of estimating $SE(\bar{X})$

Traditional Example

```
> set.seed(5141)
> x <- rnorm(25, 0, 10)
> xbarfull <- mean(x)
> xbarjack <- NA
> for(i in 1:25) xbarjack[i] <- mean(x[-i])
> sqrt((25-1)/(25)*sum((xbarjack-mean(xbarjack))^2))
[1] 1.873911

> set.seed(511)
> x <- rnorm(25, 0, 10)
> xbarfull <- mean(x)
> xbarjack <- NA
> for(i in 1:25) xbarjack[i] <- mean(x[-i])
> sqrt((25-1)/(25)*sum((xbarjack-mean(xbarjack))^2))
[1] 2.314983
```

Traditional Example

- What about the bias of $\bar{X}$? $\ldots 0$
- Part of the reason $\bar{X}$ is considered a good estimator for μ is that it's unbiased.
- Again, supposing we didn't know that $\bar{X}$ was unbiased, we could estimate its bias using the jackknife...

Traditional Example

```
> set.seed(5141)
> x <- rnorm(25, 0, 10)
> xbarfull <- mean(x)
> xbarjack <- NA
> for(i in 1:25) xbarjack[i] <- mean(x[-i])
> (25-1)*(mean(xbarjack) - mean(x))
[1] 0

> set.seed(511)
> x <- rnorm(25, 0, 10)
> xbarfull <- mean(x)
> xbarjack <- NA
> for(i in 1:25) xbarjack[i] <- mean(x[-i])
> (25-1)*(mean(xbarjack) - mean(x))
[1] -1.332268e-15
```

Traditional Example — Faster...

```
> set.seed(5141)
> x <- rnorm(25, 0, 10)
> xbarfull <- mean(x)
> xbarjack <- NA
> for(i in 1:25) xbarjack[i] <- mean(x[-i])
> (25-1)*(mean(xbarjack) - mean(x))
[1] 0
> sqrt((25-1)/(25)*sum((xbarjack-mean(xbarjack))^2))
[1] 1.873911

> library(bootstrap)
> jfit <- jackknife(x, mean)
> jfit$jack.bias
[1] 0
> jfit$jack.se
[1] 1.873911
```

Modelling Example

- Bringing this back around to machine learning, we can use the jackknife to investigate the bias and variance of our estimators within any particular model
- For example, $\hat{\beta}_0$ and $\hat{\beta}_1$ from SLR
- We simulate 30 observations from

$$Y = 2X + \epsilon$$

where $\epsilon \sim N(0, 0.25)$

SLR Example

```
> set.seed(311532)
> x <- runif(30, 0, 1)
> y <- 2*x + rnorm(30, sd=0.25)
> fullfit <- lm(y~x)
> jlist <- matrix(NA, nrow=30, ncol=2)
> for(i in 1:length(x)){
+   jlist[i,] <- lm(y[-i]~x[-i])$coef
+ }
> (30-1)*(mean(jlist[,1])-fullfit$coefficients[1])
  (Intercept)
-0.0005130173
> (30-1)*(mean(jlist[,2])-fullfit$coefficients[2])
            x
0.0008851131
```

SLR Example

```
> sqrt(((30-1)/30)*sum((jlist[,1]-mean(jlist[,1]))^2))
[1] 0.1042806
> #true se(hat beta0)
> .25*sqrt(1/30+(.5^2)/(sum((x-mean(x))^2)))
[1] 0.08246102

> sqrt(((30-1)/30)*sum((jlist[,2]-mean(jlist[,2]))^2))
[1] 0.1358563
> #true se(hat beta1)
> .25/sqrt(sum((x-mean(x))^2))
[1] 0.1373534
```

Modelling Example

- Again, traditional estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased...so it's good to see the jackknife bias estimates near 0.

- We also see that the jackknife estimates for the SEs of both estimators appear reasonably close to the known SEs

- With some additional work, we could use these SEs to provide jackknife versions of confidence intervals...but we'll leave that aside, as we're about to learn a 'simpler' option.

Notes

- We considered LOO jackknife...it is generalizable to 'leave j out', just as CV was.
- While the jackknife is a nonparametric method for estimating things like bias and variance of an estimator, that does not mean that it is void of assumptions, or rigorous statistical theory
- It has been shown consistent (essentially asymptotically 'good') for many common estimators, but not the median.
- The biggest assumption is that observations are iid — independent and identically distributed.

Bootstrap

- The (nonparametric) bootstrap is a tremendously versatile method for estimating the standard errors and bias of parameter estimates and has largely supplaced the jackknife.
- Instead of simulating new data from a known model (which will be impossible in real scenarios), we can randomly sample with replacement from our observed sampe of data!
- Oddly enough, with a bit of work, this provides good estimates of the bias and standard error of our estimators.

Bootstrap

- Here's the process:

  1. Set $j = 1$ as index for bootstrap sample number
  2. Take the j^{th} random sample of size n from your observations $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ with replacement
  3. Fit model, estimate parameter α with $\hat{\alpha}_j$ (for example)
  4. Set $j = j + 1$, and if $j \leq B$ return to step 2 ($B = 1000, 5000$ are standard amounts)
  5. Estimate the standard error and/or bias of the estimator

$$\hat{SE}(\hat{\alpha}) = \sqrt{\frac{\sum_{j=1}^{B}(\hat{\alpha}_j - \bar{\hat{\alpha}})^2}{B-1}} \qquad \hat{Bias}(\hat{\alpha}) = \bar{\hat{\alpha}} - \hat{\alpha}$$

where $\bar{\hat{\alpha}} = \sum_{j=1}^{B} \hat{\alpha}_j / B$

Bootstrap

- Visually, from your textbook

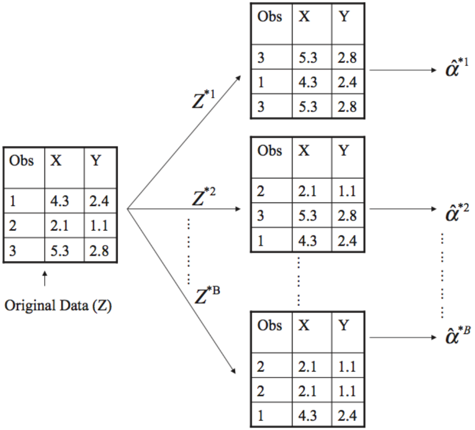


FIGURE 5.11. *A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α.*

Simulation Example

- We simulate 30 observations from

$$Y = 2X + \epsilon$$

where $\epsilon \sim N(0, 0.25)$

- In a simulation, we can simply continue generating data to investigate things like the distribution of $\hat{\beta}_1$

- So we do it...but to keep it simple, we'll hold X fixed...

Simulation example

```
newx <- list()
newy <- list()
modnew <- list()
coefs <- NA
for(i in 1:1000){
  newx[[i]] <- x
  newy[[i]] <- 2*newx[[i]] + rnorm(30, sd=0.25)
  modnew[[i]] <- lm(newy[[i]]~newx[[i]])
  coefs[i] <- modnew[[i]]$coefficients[2]
}
```

Simulation Example

- Now we have fit 1000 linear models to 1000 new simulations (of size 30)
- We stored all $\hat{\beta}_1$ in "coefs"
- sd(coefs) $= 0.1325$

Simulation Example

- BUT with real data, you will never have the option to just simulate a bunch more of it.
- Enter the bootstrap...

Simulation example

```
newboots <- list()
bootsmod <- list()
bootcoef <- NA
xy <- cbind(x,y)
for(i in 1:1000){
  newboots[[i]] <- xy[sample(1:30, 30, replace=TRUE),]
  bootsmod[[i]] <- lm(newboots[[i]][,2]~newboots[[i]][,1])
  bootcoef[i] <- bootsmod[[i]]$coefficient[2]
}
```

Simulation Example

- Now we have fit 1000 linear models to 1000 bootstrapped samples (of size 30)
- We stored all $\hat{\beta}_1$ in "bootcoef"
- sd(bootcoef) $= 0.1323$ (!!!!)

Simulation Example

- At this point, we can compare a few things

Method	SE($\hat{\beta}_1$)
Truth	0.1374
1000 Simulations	0.1325
Inferential summary(lm)	0.1294
Jackknife*	0.1359
Bootstrap*	0.1323

- * - Emphasis here is that for these methods, we made no assumptions about the underlying distribution of the error terms (though iid is assumed, and there are some conditions on $\hat{\beta}_1$).
- For summary(lm), we need to assume normal iid error
- For both the truth and 1000 simulations, we need the full form of $f(x) + \epsilon$...AKA, never feasible in practice

Simulation Example

- Change to uniform error?

Method	SE($\hat{\beta}_1$)
Truth	0.0624
1000 Simulations	0.0628
Inferential summary(lm)	0.0560
Jackknife*	0.0601
Bootstrap*	0.0598

- * - Emphasis here is that for these methods, we made no assumptions about the underlying distribution of the error terms (though iid is assumed, and there are some conditions on $\hat{\beta}_1$).
- For summary(lm), we need to assume normal iid error (AKA, wrong in this case)
- For both the truth and 1000 simulations, we need the full form of $f(x) + \epsilon$...AKA, never feasible in practice

Comments

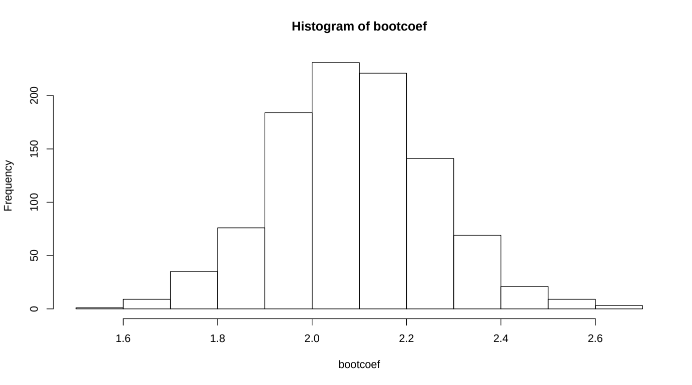
- This should give you some confidence in the power of bootstrapping!
- Bootstrap and jackknife are used for the same general purposes, but jackknife is older and was primarily used at a time when computing resources were limited.
- As such, bootstrapping is the predominant nonparametric method used for bias and standard error calculations.

Taking it Further

- We alluded to the potential of computing confidence intervals with the jackknife, but with the bootstrap one option for doing so is quite straightforward.
- The B bootstrap estimates of parameter α provide an empirical estimate of the distribution of estimator $\hat{\alpha}$.
- So, for example, we can look at a histogram of bootstrapped $\hat{\beta}_1$ from our previous example...

Bootstrap CI

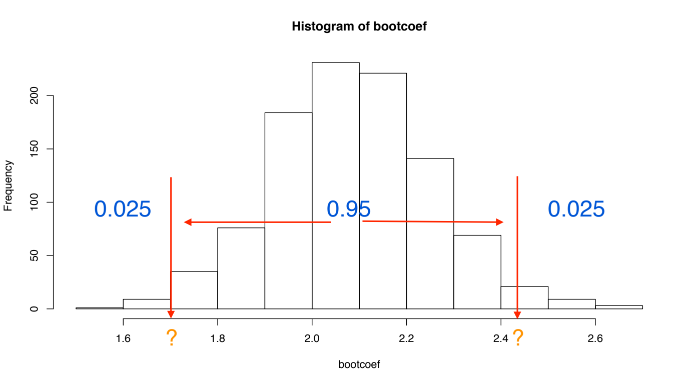
- hist(bootcoef)



- contains 1000 bootstrap $\hat{\beta}_1$ estimates...

Bootstrap CI

- The simplest method for providing a CI is to take percentiles. Suppose I want a 95% bootstrap CI for β_1...



Bootstrap CI

```
> quantile(bootcoef, .025)
    2.5%
1.76697

> quantile(bootcoef, .975)
   97.5%
2.296553
```

- So we have a 95% CI for true β_1 lying within the interval $(1.77, 2.30)$

Modern Bootstrapping

- While the bootstrap originated in the 70's, novel uses are still being developed:

ELSEVIER

Computational Statistics & Data Analysis

Volume 127, November 2018, Pages 160-171

Addressing overfitting and underfitting in Gaussian model-based clustering

Jeffrey L. Andrews

Show more

https://doi.org/10.1016/j.csda.2018.05.015

Get rights and content

Highlights

- Overfitting and underfitting are illustrated using the EM algorithm for clustering.
- A nonparametric bootstrap augmented EM-style algorithm is proposed.
- It is shown through applications to address both overfitting and underfitting.

Bootstrap Acclaim

Statistical 'rock star' wins coveted international prize
Nature.com - Nov. 12, 2018
Bradley **Efron**, at Stanford University, has won the US$80,000 ... 2018 International Prize in Statistics for pioneering the '**bootstrap**' method for ...

- Stats world's version of the Nobel Prize

Bootstrap methods: another look at the jackknife
B **Efron** - Breakthroughs in statistics, 1992 - Springer
We discuss the following problem given a random sample X=(X 1, X 2,..., X n) from an unknown probability distribution F, estimate the sampling distribution of some prespecified random variable R (X, F), on the basis of the observed data x.(Standard jackknife theory ...
Cited by 17085 Related articles All 8 versions Web of Science: 7439

[BOOK] An introduction to the **bootstrap**
B **Efron**, RJ Tibshirani - 1994 - books.google.com
Statistics is a subject of many uses and surprisingly few effective practitioners. The traditional road to statistical knowledge is blocked, for most, by a formidable wall of mathematics. The approach in An Introduction to the Bootstrap avoids that wall. It arms ...
Cited by 38703 Related articles All 10 versions

[BOOK] The jackknife, the **bootstrap**, and other resampling plans
B **Efron** - 1982 - books.google.com
The jackknife and the bootstrap are nonparametric methods for assessing the errors in a statistical estimation problem. They provide several advantages over the traditional parametric approach: the methods are easy to describe and they apply to arbitrarily ...
Cited by 9533 Related articles All 14 versions

- Actual cite counts should be much higher, bootstrapped SE/CIs are ubiquitous in applied papers across the natural sciences


UBC
THE UNIVERSITY OF BRITISH COLUMBIA