Jackknife notes

CRS

29 November 2023

Contents

1 Orientation **1**

2 Definitions **2**

3 Estimates of bias, variance, and standard errors **2**
3.1 Bias estimation 2
3.1.1 Bias correction 2
3.2 Variance estimation 2
3.3 Standard error estimation 3

4 When does the jackknife work? **3**

5 Why $n-1$? **3**
5.1 $n-1$ from bias correction 3
5.2 "Pseudovalues" and $n-1$ for variance 4
5.3 Pseudo-values and approximate Gaussian distributions 5

6 Code sketch **5**

7 Further reading **6**

References **7**

1 Orientation

We have a collection of data points $x_1, x_2, \ldots x_n$. Each one makes some contribution to an estimate $\hat{\theta}(x_1, x_2, \ldots x_n)$ of a quantity θ. We would like to be able to say something about the **sampling distribution** of $\hat{\theta}$, because that's our source of knowledge about things like bias, variance, and confidence intervals. Conceptually, what we want is to re-run the experiment, getting *random* variables $X_1, X_2, \ldots X_n$, and re-compute the estimate, $T = \hat{\theta}(X_1, X_2, \ldots X_n)$. If we could do that infinitely many times, we'd see the distribution of the random variable[1] T. Then the bias would be $\mathbb{E}[T] - \theta$, the variance would be $\mathrm{Var}[T]$, etc., etc. The difficulty though is that we have only *one* data set.

The **jackknife** is a technique for trying to get around this difficulty of having only one data set, by looking at how much the estimate would change if we slightly alter the data set. The resulting distribution of re-estimates is, under some assumptions, related to the sampling distribution of $\hat{\theta}$ that's what we'd really like to know.

[1] Notice that I am trying to distinguish between (1) the true value of what we're estimating, θ (a fixed but unknown number, vector, curve...); (2) the function $\hat{\theta}$ we use to do estimation, $\hat{\theta}$ (which is a function that reads in data sets and returns values in the same space as θ); and (3) the random variable T we get by applying $\hat{\theta}$ to our data source.

2 Definitions

The i^{th} **jackknife re-estimate** is the estimate we'd get if we omitted or held out x_i, i.e.,

$$T_{-i} \equiv \hat{\theta}(x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots x_n) \tag{1}$$

We therefore re-do our estimate n times, getting a distribution of jackknife re-estimates. The mean of this distribution is

$$\bar{\hat{\theta}} \equiv \frac{1}{n}\sum_{i=1}^{n} T_{-i} \tag{2}$$

3 Estimates of bias, variance, and standard errors

There are three major uses for the distribution of jackknife re-estimates.

3.1 Bias estimation

We *estimate* the bias of our estimator $\hat{\theta}$ by comparing the mean of the jackknife estimates to the estimate we got on the full data:

$$\widehat{\text{bias}} \equiv (n-1)(\bar{\hat{\theta}} - \hat{\theta}(x_1, \ldots x_n) \tag{3}$$

The factor of $n-1$ is to "scale up" the spread of re-estimates around the true estimate, to be more comparable to what we'd get in a truly new sample from the same population.

3.1.1 Bias correction

If we add our estimate of the bias to our original, full-data estimate, we get an (approximately) unbiased estimate. This process is called **bias correction** (naturally enough). It works for any (good) method of approximating the bias, the jackknife included.

3.2 Variance estimation

We estimate the variance of our estimator by (again) looking at the variance of the jackknife re-estimates:

$$\widehat{\text{Var}} \equiv (n-1)\frac{1}{n}\sum_{i=1}^{n}\left(T_{-i} - \bar{\hat{\theta}}\right)^2 \tag{4}$$

I write it in this way to make it clear that first we're taking the (sample) variance of the jackknife re-estimates around their mean, and then we're scaling it up by that same factor of $n-1$.

The jackknife variance estimate is known to usually *over*-estimate the true variance (i.e., to be biased upwards) (Efron and Stein 1981). This isn't *great*, but in situations where we're considering the jackknife at all, there usually isn't a much better alternative, and the amount of bias can be very small. (See [Efron and Stein (1981) for some details.)

3.3 Standard error estimation

The standard error of an estimator is always defined as the square root of its variance, so

$$\widehat{\mathrm{se}} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(T_{-i} - \overline{\hat{\theta}}\right)^2} \tag{5}$$

Since the jackknife tends to over-estimate variance, it also tends to over-estimate standard errors. In many contexts, over-estimating standard errors will lead us to being conservative in (e.g.) claiming that differences between groups are significant, so scientifically it can amount to "erring on the side of caution".

4 When does the jackknife work?

The fundamental ideas on which the jackknife rests are (i) the estimator $\hat{\theta}$ should be "consistent", i.e., it should converge on the true value of θ as we get more and more data, and (ii) each data point should make a fairly small contribution to the estimate. Moreover, (iii) the data points should be statistically independent, and (iv) a sufficiently small change to the *true* distribution of the X_i should lead to only a small change in the *true* value of θ (continuity). This last point rules out using the jackknife for questions like "how many clusters should the mixture model have?",

5 Why $n-1$?

> This gets a little deep into the weeds, compared to the rest of this note. If you're willing to take the formulas for bias and variance on trust, you can skip this.

The appearance of the factor of $n-1$ should frankly bother you a little.

It should make sense that the different jackknife re-estimates are more similar to each other, and to $\hat{\theta}$, than we'd get if we really could re-run the experiment with independent data. After all, the original estimate and the jackknife re-estimates have $n-1$ data points in common, and any two re-estimates share $n-2$ data points. By contrast, if we repeat an experiment with continuous random variables, there will be *no* shared data points (with probability 1). So it should make sense that the distribution of jackknife re-estimates is more tightly concentrated around its mean, and around $\hat{\theta}(x_1, \ldots x_n)$, than is the sampling distribution of $\hat{\theta}(X_1, \ldots X_n)$. It should also make sense that this over-concentration or under-dispersion will get worse as n increases.

It would be *nice* if we could un-do this with a simple correction factor that just involves n (and not, e.g., some sort of complicated derivative of the estimator function). The usual lie told to children is that the factor of $n-1$ is because each jackknife re-estimate shares $n-1$ data points in common with the full data set, but this should be unsatisfying — at the very least it's not obvious why this is a better answer than n or $n-2$ or for that matter $2n+1$.

5.1 $n-1$ from bias correction

Suppose (what's often known to be true) that the bias of $\hat{\theta}$ can be expressed as a power series in n, that is,

$$\mathbb{E}\left[T_n\right] = \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + \ldots \tag{6}$$

(It'll be important to keep track of how much data we're using, hence the subscript in T_n.) If we based our estimate on any $n-1$ of the n data points, presumably the same formula would apply, only with powers of $n-1$ instead of n in the denominators.

Now let's *define*

$$T'_n \equiv T_n + (n-1)(T_n - T_{n-1}) \tag{7}$$

If we ask for the expectation value of T'_n, we get (after some character-building algebra) something slightly miraculous:

$$\mathbb{E}\left[T'_n\right] = \theta - \frac{a_2}{n^2} + \frac{a_2 + a_3}{n^3} + \ldots \tag{8}$$

In words: we've removed the bias of order $1/n$. Therefore

$$(n-1)\mathbb{E}\left[T_n - T_{n-1}\right] = -\frac{a_1}{n} + \frac{b_2}{n^2} + \ldots \tag{9}$$

That is, the expectation on the left *is* the bias of $\hat{\theta}$, or at least is the bias to leading order in n. If we used a different factor than $n-1$ on the left-hand side, we would *not* get the correct value a_1/n on the right-hand side.

Now, obviously we don't *know* $\mathbb{E}\left[T_n - T_{n-1}\right]$, but we can estimate it. Each one of our jackknife re-estimates T_{-i} is an example of T_{n-1}, so

$$\mathbb{E}\left[T - T_{-i}\right] = \mathbb{E}\left[T_n - T_{n-1}\right] \tag{10}$$

Since we have n values for $T - T_{-i}$, we might as well average them to get a better estimate of $\mathbb{E}\left[T_n - T_{n-1}\right]$. This is, fundamentally, why we use the factor of $n-1$ in the jackknife bias estimate.

5.2 "Pseudovalues" and $n-1$ for variance

The argument for why the factor of $n-1$ is right for variance is a bit more indirect. What we saw in the last sub-section convinced us that

$$V_i \equiv T + (n-1)(T - T_{-i}) \tag{11}$$

is (approximately) unbiased for θ. So is their average,

$$\overline{V} \equiv \frac{1}{n}\sum_{i=1}^{n} V_i = T + (n-1)(T - \overline{\hat{\theta}}) \tag{12}$$

These **pseudovalues** V_i (often) have two other important properties:

1. V_i and V_j are (approximately) independent of each other (when $i \neq j$).
   - This is because, to the extent that V_i differs from T, it's due to deleting X_i, while V_j's deviation from T is due to deleting X_j, and X_i and X_j, being
2. $\mathrm{Var}\left[V_i\right] \approx n\mathrm{Var}\left[T\right]$.
   - *overlineV*, the average of the V_i is the jackknife-bias-corrected estimator. Because we're *mostly* doing bias correction, this shouldn't really affect the variance compared to using the full data in our estimate, i.e., $\mathrm{Var}\left[\overline{V}\right] \approx \mathrm{Var}\left[T\right]$. But if the V_i are independent and identically distributed, then $\mathrm{Var}\left[\overline{V}\right] = n^{-2} n \mathrm{Var}\left[V_i\right] = n^{-1}\mathrm{Var}\left[V_i\right]$. So $\mathrm{Var}\left[V_i\right] \approx n\mathrm{Var}\left[T\right]$, as claimed.

Of course, this leaves us needing to estimate $\mathrm{Var}\left[V_i\right]$, but we can do this with the sample variance of the pseudovalues. If we bias-correct this sample variance in the usual way, by using $n-1$ instead of n in the denominator, we get

$$\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(V_i - \overline{V}\right)^2 && (13)\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left((n-1)T_{-i} - (n-1)\overline{\hat{\theta}}\right)^2 && (14)\\
&= (n-1)\sum_{i=1}^{n}\sum_{i=1}^{n}\left(T_{-i} - \overline{\hat{\theta}}\right)^2 && (15)
\end{aligned}$$

The final estimate for the variance of T is then s^2/n, or

$$\frac{n-1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n}\left(T_{-i}-\overline{\hat{\theta}}\right)^2 \tag{16}$$

as promised.

(What if we *didn't* use the bias-correcting denominator $n-1$ in Eq. 13, but just n? Then we'd end up with $\frac{(n-1)^2}{n^2}\sum_{i=1}^{n}\left(T_{-i}-\overline{\hat{\theta}}\right)^2$. That is, our estimated variances would be uniformly smaller by a factor of $\frac{n-1}{n}$. This won't be a big deal when n is large, but it is, systematically, slightly too small.)

5.3 Pseudo-values and approximate Gaussian distributions

If we're willing to swallow the idea that the pseudovalues V_i are IID, and we use their average $\overline{V}$ as our bias-corrected estimator, then we're looking at a mean of n IID quantities, and the central limit theorem tells us that this is approaches a Gaussian as $n \to \infty$. Specifically, we should believe that $\overline{V} \approx \mathcal{N}(\theta, \mathrm{Var}\left[V_i\right]/n)$ for large n. We can use this to work out confidence intervals, hypothesis tests, etc., in the usual way. Notice, however, that we've piled lots of approximations on top of each other. The jackknifed estimator's being Gaussian is a *lot* more precarious than just the bias correction or variance formulas.

6 Code sketch

```
# Inputs: a data frame (data)
  # an estimator function (estimator)
# Outputs: a vector, giving the "raw" estimate, the estimated bias, and the estimated standard error
# Assumes: each row of data is an independent data frame
# Assumes; estimator() can work on data frames with any number of rows
# Assumes: estimator() returns a scalar
  # This approach works for vector-valued estimators, but the code gets a bit
  # clunkier
jackknife <- function(data, estimator) {
    n <- nrow(data)
    main.est <- estimator(data)
    # Define a utility function for holding back one row
    hold.one <- function(i) { estimator(data[-i,]) }
    # Applying that function to all the rows in this way is much faster, in
    # R, than writing the obvious for() loop
    jackknife.reests <- sapply(1:n, hold.one)
    bias.est <- (n-1)*(mean(jackknife.reests) - main.est)
    pseudovalues <- main.est + (n-1)*(main.est - jackknife.reests)
    var.est <- var(pseudovalues)/n # var() uses denominator of n-1
    return(c(raw=main.est, bias=bias.est, SE=sqrt(var.est)))
}
```

Now let's make up some data.

```
x <- runif(n=30, min=-2, max=37)
y <- x^2 + rnorm(n=length(x), sd=0.1)
some.parabolas.sorta <- data.frame(x=x, y=y)
```

It's well-known (e.g., Shao and Tu (1995), pp. 6–7) that if $\hat{\theta}$ is the sample mean, the jackknife approach (correctly) says that the ibas is 0, and returns the sample variance for the estimated variance. Let's try that out:

```
mean(some.parabolas.sorta$x) # Sample mean of X
```

```
## [1] 18.432
```

```
var(some.parabolas.sorta$x) # Sample variance of x, with denominator n-1
```

```
## [1] 142.359
```

```
sqrt(var(some.parabolas.sorta$x)/nrow(some.parabolas.sorta)) # Std. err. of mean
```

```
## [1] 2.178371
```

```
jackknife(data=some.parabolas.sorta, estimator=function(df) { mean(df$x) })
```

```
##       raw      bias        SE
## 18.432001  0.000000  2.178371
```

Since everything checks out in a case where we know the answer, let's apply it to the correlation between the two coordinates, where we don't:

```
jackknife(data=some.parabolas.sorta, estimator=function(df) { cor(df$x, df$y) })
```

```
##         raw        bias          SE
## 0.959637247 0.001995790 0.007212404
```

Because this is a simulation, we can re-run the simulation many times to see how good the jackknife estimates of bias and variance are. First we need a good value for the true, population correlation coefficient, which we get by doing an enormous simulation:

```
big.x <- runif(n=1e6, min=-2, max=37)
big.y <- big.x^2 + rnorm(n=length(big.x), sd=0.1)
(pop.cor <- cor(big.x, big.y))
```

```
## [1] 0.961022
```

Now we replicate our little simulation experiment many times:

```
sampling.dist.of.cor <- replicate(1e6, { x <- runif(n=30, min=-2, max=37); y <- x^2 + rnorm(n=length(x)
(bias <- mean(sampling.dist.of.cor) - pop.cor)
```

```
## [1] 0.001469903
```

```
(se <- sd(sampling.dist.of.cor))
```

```
## [1] 0.008004337
```

The jackknife isn't nailing everything exactly, but it's doing pretty well for something so simple (particular at getting the standard error).

EXERCISE: Modify my jackknife code to (also) return the vector of jackknife re-estimates. (It might help to make `jackknife()` return a list.) Plot their distribution here, and compare to the sampling distribution (as simulated in the last code block).

7 Further reading

For more on the theory and practice of the jackknife, see Efron and Tibshirani (1993), or Shao and Tu (1995).

The argument for why the factor of $n-1$ is right for the bias goes back to Quenouille (1956), in what is widely regarded as the first paper on the jackknife. The argument for why the factor of $n-1$ is right for the variance goes back to a one-paragraph abstract by the great statistician John W. Tukey (Tukey 1958). It

was, apparently, Tukey who introduced the name "jackknife", in drafts of a book on practical data analysis which were never formally published, but circulated widely during the 1960s[2]. That same unpublished book also seems to be the source of the term "pseudovalues"[3].

The pseudovalues approach tries to approximate a possibly-very-complicated estimator as a sum of IID terms, one from each data point. This turns out to be an important idea; for instance, the method of "influence functions" approximates messy statistics as sums of IID terms, one from each data point, which involves taking a complicated kind of derivative of the original statistic. This is a theoretically-advanced topic which however often leads to very practical estimators; see Vaart (1998) for some general theory, Laan and Rose (2011) for applications in causal inference, and Fisher and Kennedy (2021) for a tutorial introduction.

The bootstrap essentially originated with Bradley Efron thinking about how to modify the jackknife so as to get more information about the sampling distribution (Efron 1979, 1982).

For simplicity, I've covered only the case where the data are independent and identically distributed. But there are extensions of the jackknife to dependent data, such as stationary time series (Künsch 1989) and networks (Lin, Lunde, and Sarkar 2020).

References

Efron, Bradley. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7:1–26. https://doi.org/10.1214/aos/1176344552.

———. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: SIAM Press.

Efron, Bradley, and Charles Stein. 1981. "The Jackknife Estimate of Variance." *Annals of Statistics* 9:586–96. https://doi.org/10.1214/aos/1176345462.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman; Hall.

Fisher, Aaron, and Edward M. Kennedy. 2021. "Visually Communicating and Teaching Intuition for Influence Functions." *The American Statistician* 75:162–72. https://doi.org/10.1080/00031305.2020.1717620.

Künsch, Hans R. 1989. "The Jackknife and the Bootstrap for General Stationary Observations." *Annals of Statistics* 17:1217–41. https://doi.org/10.1214/aos/1176347265.

Laan, Mark J. van der, and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Berlin: Springer-Verlag. https://doi.org/10.1007/978-1-4419-9782-1.

Lin, Qiaohui, Robert Lunde, and Purnamrita Sarkar. 2020. "On the Theoretical Properties of the Network Jackknife." In *Proceedings of the 37th International Conference on Machine Learning [ICML 2020]*, edited by Hal Daumé and Aarti Singh, 6105–15. jmlr.org. https://arxiv.org/abs/2004.08935.

Quenouille, M. H. 1956. "Notes on Bias in Estimation." *Biometrika* 43:353–60. https://doi.org/10.2307/2332914.

Shao, Jun, and Dongsheng Tu. 1995. *The Jackknife and Bootstrap*. Berlin: Springer-Verlag. https://doi.org/10.1007/978-1-4612-0795-5.

Tukey, John W. 1958. "Bias and Confidence in Not-Quite Large Samples." *Annals of Mathematical Statistics* 29:614. https://www.jstor.org/stable/2237363.

Vaart, A. W. van der. 1998. *Asymptotic Statistics*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511802256.

[2] A jackknife is a small knife whose blade folds in to the handle. The metaphor seems to be that holding out one data point is like folding in a blade. (I don't think it's a great metaphor, but Tukey was a genius and I'm very much not, so...)

[3] Suppose that $\hat{\theta}$ is just the sample mean. If you go through the algebra for calculating V_i, you will find that you recover the original data value X_i. (Exercise: Go through the algebra.) With other statistics we don't get back the data values, hence "pseudo-" (=false) values.