Writing better, faster code in R

Julia Wrobel

Overview

Today, we cover:

- Resampling methods
  - Permutation tests
- Improving code speed:
  - Benchmarking
  - Vectorization
  - Parallelization

Announcements:

- HW1 grades are up
- HW2 posted and due 2/11 at 10:00AM
- No class tomorrow (Thursday, January 29) but still have Office Hours

Permutation tests

Typically bootstrap is used for CI rather than hypothesis testing. For hypothesis testing and p-values, we can use a **permutation test**.

- Idea: use resampling to generate a **null distribution** for a test statistic, then compare it to the one you observe in the real data
- **Null distribution**: the distribution of a quantity of interest (i.e. $\hat{\beta}$) if the null hypothesis H_0 is true
- The null distribution is available theoretically in some cases. For example, assume $Y_i \sim N(\mu, \sigma^2), i = 1, \dots, n$.
  - Under $H_0 : \mu = 0$, we have $\bar{Y} \sim N(0, \sigma^2/n)$
    - Test H_0 by comparing $\bar{Y}$ with $N(0, \sigma^2/n)$
- Use **permutation test** when null distribution cannot be obtained theoretically

Permutation tests

The basic procedure of permutation test for H_0:

- Permute data under H_0 B times. Each time recompute the test statistics. The test statistics obtained from the permuted data form the null distribution.
- Compare the observed test statistics with the null distribution to obtain statistical significance.

Permutation test example: difference in means

Assume there are two sets of independent normal r.v.'s with the same known variance and different means:

- $X_i \sim N(\mu_1, \sigma^2)$
- $Y_i \sim N(\mu^2, \sigma^2)$

Our goal is to test $H_0 : \mu_1 = \mu_2$. Define test statistic: $t = \bar{X} - \bar{Y}$. Permutation test steps:

1. Randomly shuffle labels of X and Y
2. Compute $t^* = \bar{X}^* - \bar{Y}^*$
3. Repeat `nperm` times. Resulting t^* values form the **empirical null distribution** of t.
4. To compute p-values calculate $Pr(|t^*| > |t|)$

Permutation test example: difference in means

```r
set.seed(111)
x = rnorm(30, 2, 2)
y = rnorm(30, 0.5, 1.5)
mean_diff = mean(x) - mean(y)

nperm = 10000
perm_mean_diff = rep(NA, nperm)
for(perm in 1:nperm){
  combined_data = c(x, y)
  x_index = sample(1:length(combined_data), size = length(x), replace = FALSE)
  x_perm = combined_data[x_index]
  y_perm = combined_data[-x_index]

  perm_mean_diff[perm] = mean(x_perm) - mean(y_perm)
}

p = sum(abs(perm_mean_diff) >= abs(mean_diff))/nperm
p
```



```
[1] 0.0437
```

Permutation test example: regression coefficient

$$Y_i = \beta_0 + \beta_{treatment} X_{i1} + \mathbf{Z_i}^T \gamma + \epsilon_i$$

- Our goal is to test $H_0 : \beta_{treatment} = 0$ and we observe $\widehat{\beta}_{treatment} = 1.2$.

How will we perform the permutations?

Permutation test example: classification accuracy

We used a random forest model to build a binary classifier for disease detection and achieved accuracy of 78%.

- Goal is to test if this is significantly different than random guessing (50%)

How will we perform the permutations?

Choosing the number of permutations

- Monte Carlo error: If the true p-value is p, then

$$\hat{p} \sim Binomial(nperm, p)/nperm$$

and

$$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{nperm}}$$

Choosing the number of permutations

```
p_true = 0.05
nperm = c(10, 100, 10000)

# standard error
sqrt(p_true*(1-p_true)/nperm)
```

```
[1] 0.068920244 0.021794495 0.002179449
```

Lab: resampling methods

We will spend the next 20-30 minutes going through this lab.

Improving speed

Two ways to find bottlenecks in code:

- Benchmarking
- Profiling

Comparing R code for speed

- R package `microbenchmark` is well-suited for comparing small chunks of code
- Your code can often be significantly improved

```
library(microbenchmark)
x <- 120
microbenchmark(
  sqrt(x),
  x^(0.5)
)
```

```
Unit: nanoseconds
    expr min lq  mean median uq  max neval
 sqrt(x)   0  0  9.84      0  0  779   100
 x^(0.5)  41 41 95.12     41 82 3403   100
```

Comparing R code for speed

```
p <- 1000
x <- runif(p, min = 100, max = 120)
microbenchmark(
  sqrt(x),
  x^(0.5)
)
```

```
Unit: nanoseconds
    expr  min   lq     mean median   uq   max neval
 sqrt(x)  861  984 1123.81   1025 1025  7585   100
 x^(0.5) 8774 8897 8971.21   8938 8979 11111   100
```

Comparing R code for speed

```
p <- 100000
x <- runif(p, min = 100, max = 120)
microbenchmark(
  sqrt(x),
  x^(0.5)
)
```

```
Unit: microseconds
    expr      min       lq     mean   median       uq      max neval
 sqrt(x)  87.576 116.645 127.5461 128.6785 135.2385  168.551   100
 x^(0.5) 855.752 870.348 949.2377 886.6455 898.6790 2964.710   100
```

Comparing R code for speed

- Take advantage of **crossprod** and **tcrossprod** functions. Suppose we want to calculate $x^T Ax$

```
p <- 30000
x <- rnorm(p)
A <- matrix(rnorm(p^2), p, p)
microbenchmark(
    t(x) %*% A %*% x,
    crossprod(x, A %*% x)
)
```

```
Unit: milliseconds
                  expr      min       lq     mean   median       uq      max
      t(x) %*% A %*% x 301.7449 314.4926 316.1829 316.7194 318.3924  336.7777
 crossprod(x, A %*% x) 972.4754 983.9149 997.3086 986.8418 991.9102 1960.5898
 neval
   100
```

Example: R Loops

R is very bad at resizing objects since it copies to resize

- Do not grow an object inside of a loop!
- Instead, make an empty object first and then fill elements.

```
bad = function (x){
  obj = c()
  for(i in 1:x){
    obj = c(obj, i)
    }
  return(obj)
}

better = function (x){
  obj = rep(NA, x)
  for (i in 1:x){
    obj[i] = i
    }
  return(obj)
```

Example: R Loops

```
microbenchmark(bad (100), better (100))
```

```
Unit: microseconds
         expr    min      lq     mean median     uq      max neval
    bad(100) 15.498 16.8305 31.88693 17.835 20.377 1264.932   100
 better(100)  2.501  2.6240 14.03061  2.665  2.788 1105.811   100
```

Measuring speed in simulations

If you are interested in measuring computation time in a simulation study, say, to compare how fast different methods are, I **would not** recommend `microbenchmark`. Instead, do the following:

```
library(tictoc)

tic()
## do some stuff
large_vector <- rnorm(1e7)  # Create a vector of 10 million random numbers
sum_large_vector <- sum(large_vector)
time_stamp = toc(quiet = TRUE) # stop the timer and print the time elapsed

time_stamp$toc - time_stamp$tic # human time in seconds
```

```
elapsed
  0.256
```

Vectorization

R is **vectorized**, meaning it efficiently performs operations on entire vectors or arrays in a single step, avoiding explicit loops and leveraging optimized low-level code for speed and simplicity.

- Often, there is more than one way to do something in R
- Take advantage of vectorization!
  - Often it is more concise and significantly faster

Vectorization

```r
# non vectorized squaring operation
x <- c(1, 2, 3, 4, 5)
result <- numeric(length(x))
for (i in seq_along(x)) {
  result[i] <- x[i]^2
}

# same operation, vectorized
x <- c(1, 2, 3, 4, 5)
result <- x^2
```

Why vectorization?

Another example- which is the vectorized version?

```
x <- matrix(rnorm(30), 10, 3)


colMeans(x)
```

```
[1]  0.05531146 -0.01858912 -0.04873593
```

```
apply(x, 2, mean)
```

```
[1]  0.05531146 -0.01858912 -0.04873593
```

Why vectorization?

- **colSums**, **colMeans** and corresponding row functions are vectorized

```
n <- 100
p <- 3000
A <- matrix(rnorm(n * p), n, p)
microbenchmark(
    colMeans(A),
    apply(A, 2, mean)
)
```

```
Unit: microseconds
              expr      min       lq      mean    median       uq      max
       colMeans(A)  101.639  102.951  107.3634  106.2105  110.536  124.476
 apply(A, 2, mean) 7047.121 7444.309 8431.4852 7690.7185 7880.118 13987.109
 neval
   100
```

Parallel computing

A modern CPU (Central Processing Unit) is at the heart of every computer. While traditional computers had a single CPU, modern computers can ship with multiple processors, which in turn can each contain multiple cores. These processors and cores are available to perform computations.

- A computer with one processor may still have 4 cores (quad-core), allowing 4 computations to be executed at the same time.
- **Parallel computing** is the simultaneous use of multiple processors or computers to solve a problem by dividing it into smaller, independent tasks, i.e. operating **in parallel**.

Parallel computing

You can check how many **cores** your computer has to see how many tasks can be run at once.

```
# Load the parallel package
library(parallel)

# Get the number of cores
detectCores()
# 12
```

When to parallelize

- Using 2 cores does not mean your code will be $2\times$ faster
- Not all tasks can be parallelized
- Loops and repetitive tasks are great candidates
  - What are some computations we have looked at already that might be good candidates for parallelization?

Example : foreach and doParallel

```
library(doParallel)
library(foreach)

# Set up parallel backend with 10 cores
num_cores = detectCores() - 2
cl = makeCluster(num_cores)
registerDoParallel(cl)

# define  Monte Carlo function to estimate Pi
monte_carlo_pi <- function(n) {
  inside_circle <- sum(runif(n, -1, 1)^2 + runif(n, -1, 1)^2 <= 1)
  return((inside_circle / n) * 4)
}

nsim = 100
```

Example : foreach and doParallel

```
pi_estimates = foreach(i = 1:nsim, .combine = c) %dopar% {
  n = 1000
  monte_carlo_pi(n)
}

# overall estimate of pi
mean(pi_estimates)
```

Example : foreach and doParallel

Useful arguments include multiple iterators, error catching, combine

```
foreach(
  ...,
  .combine,
  .init,
  .final = NULL,
  .inorder = TRUE,
  .multicombine = FALSE,
  .maxcombine = if (.multicombine) 100 else 2,
  .errorhandling = c("stop", "remove", "pass"),
  .packages = NULL,
  .export = NULL,
  .noexport = NULL,
  .verbose = FALSE
)
```

Resources

- Advanced R: chapters 22-24
- foreach vignette
- furrr package for tidyverse parallelization

Extra

Profiling your code

Some good references on profiling to learn more (a lot of overlap):

- Advanced R
- profvis R package
- Rstudio guide
- proftools R package

There has been some changes on how profiling works from R v.3 to v.4 which (sometimes) makes profiler output confusing

Profiling in R

- `profvis` is built on R's built-in profiler tool, `Rprof`
- `Rprof` is a statistical profiler that uses sampling
- When you run `profvis`, it stops the R interpreter every 10ms (default interval) and records which function is currently executing, as well as the entire call stack (i.e., which function called that function)
  - The results are not deterministic

Profiling in R

From R programming for Data Science

> *"Rprof() keeps track of the function call stack at regularly sampled intervals and tabulates how much time is spent inside each function. By default, the profiler samples the function call stack every 0.02 seconds. This means that if your code runs very quickly (say, under 0.02 seconds), the profiler is not useful. But if your code runs that fast, you probably don't need the profiler."*

Profiling: return to Ex. 2 (powers of a matrix)

Rprof() function gives a report of (approximately) how much time each function/operation within your code takes. To see the effect of memory allocation, enable tracking of **garbage collection** (GC)

```
Rprof(gc.profiling = TRUE) # start monitoring
invisible(powers1(x, 8)) # suppress function output
Rprof(NULL) # stop monitoring
summaryRprof() # see the report
```

Rprof()

- **by.total** divides the time spend in each function by the total run time
- **by.self** first subtracts out time spent in functions above the current function in the call stack (**more useful typically**)

Garbage collection (GC)

Garbage collection - freeing the memory from objects that are no longer in use (more in Section 2.6 of Advanced R)

If significant time of your program is spend in GC

- you may be doing a lot of dynamic memory allocation (the case of powers1)
- you may be storing a lot of temporary objects
- you may be consistently changing objects of large sizes

Profiling: return to Ex. 3 (powers of a matrix)

```
powers3 <- function(x, dg){
  # allocate memory in advance
  pw <- matrix(x, length(x), dg)
  prod <- x # current product
  for (i in 2:dg){
    prod <- prod * x
    pw[ , i] <- prod # no cbind
  }
  return(pw)
}
```