STAT40780 Data Programming with C (online)

Lab Sheet 11+12

Dr Marie Galligan

Summer 2015

This lab sheet poses questions relating to lecture material covered during this module, particularly that covered in weeks 11 and 12.

1 Find the median

With the help of the Rcpp and the `nth_element` algorithm from the Standard Template Library (STL), write a C++ function that's callable from R, to find the median of a vector passed from R.

2 Count values

With the help of the Rcpp and the `count` algorithm from the Standard Template Library (STL), write a C++ function that's callable from R, to count the number of missing values in an R vector.

3 Merge and sort

Write a C++ function that is callable from R, that takes as input two numeric vectors from R, and merges these into a single sorted vector, which is returned to R. The function should not sort the original vectors passed from R.

4 Simulate values from a multivariate normal density

Write a C++ function (callable from R) with the help of RcppArmadillo, to simulate n observations from a multivariate normal distribution, given a specified mean vector and covariance matrix (passed from R), and returns the simulated data in a matrix.

The following algorithm might help. It may be used to generate a single observation y (a vector) from a multivariate normal distribution with mean vector μ and covariance matrix Σ:

- First, use a Cholesky decomposition to factor the covariance matrix Σ into a left triangular matrix times its transpose

$$\mathbf{\Sigma = LL^T}$$

- Set p = number of columns (or equivalently number of rows) in Σ...the new observation will be of length p
- Fill a vector $\mathbf{x}$ with p values generated randomly from a standard normal distribution: N(0 , 1)
- A single observation $\mathbf{y}$ from a p-dimensional multivariate normal distribution with mean vector μ and covariance matrix Σ may obtained as:

$$\mathbf{y = Lx} + \mu$$

A proof of why this algorithm works may be found at:

To test your function, generate a matrix of 100 observations from a multivariate normal distribution with $\mu = (10, 5, -3)$ and

$$\Sigma = \begin{bmatrix} 1.0 & 0.9 & -0.3 \\ 0.9 & 1.0 & -0.4 \\ -0.3 & -0.4 & 1.0 \end{bmatrix}$$

5 Centering a matrix

Write a C++ function that takes as input a numeric matrix from R, and centers the data in the matrix, by calculating a vector of means (containing a mean for each column of the matrix), and subtracting the mean vector from each row of the matrix. The centered matrix should be returned to R. Compile using `cxxfunction()` and call the resulting R wrapper function `centerRcpp()`.

Run the following R code, which loads the `iris` dataset into R and extracts the first four columns (containing sepal length, sepal width, petal length and petal width), converts to a matrix and names the matrix Y. Pass the matrix Y to `centerRcpp()` to obtained a centered dataset.

Note that the 5th column of `iris` contains the categorical variable `Species` and hence is removed before passing to C++.

```
data(iris)  #load iris data into R
names(iris) #extract names of variables in the data.frame
?iris  #get information on the iris dataset

#remove last column of the iris data.frame,
#which contains the categorical variable Species
Y <-  as.matrix( iris[,-5] )
```

6 Outer product

Write a C++ functions (callable from R) that finds the outer product (i.e. cross-product) of two vectors $\mathbf{x}$ and $\mathbf{y}$.

Note that the outer product is equivalent to a matrix multiplication $\mathbf{x}\mathbf{y}^{\mathbf{T}}$ where both $\mathbf{x}$ and $\mathbf{y}$ are column vectors.