APTS Statistical Computing 2024/25: Practical Lab 1 (Tuesday)

Here are some practical problems which aim to explore and reinforce some of the course material. Several problems here use simulated data: when developing statistical modelling code, it is often best to start out with data where you know what the truth is (and can generate further replicates). Do not consult the solutions[1] until you've made good attempts.

1. **PCA via SVD** The principal components analysis (PCA) of a multivariate data set was traditionally based on the eigen-decomposition of the sample covariance matrix of the data. The matrix of eigenvectors can be used to rotate the (centred) data observations to a set of uncorrelated random quantities ordered by decreasing variance. These rotated data are often known as the *scores*. We can write a small function to implement this as follows:

```
pcScoresEig = function(X) {
  Xc = sweep(as.matrix(X), 2, colMeans(X))
  eig = eigen(crossprod(Xc)/(nrow(Xc)-1), symmetric=TRUE)
  Xc %*% eig$vectors
}
```

   Look at the notes/slides to make sure you understand what this function is doing. We can test it on the Iris flower data (see `?iris`), as follows:

```
Xi = iris[,-5] # We exclude the 5th column, which is not numerical
scores = pcScoresEig(Xi)
plot(scores[,1], scores[,2], col=iris[,5], pch=19)
```

   This is essentially how the `princomp` function in R is implemented, and we can verify this:

```
head(scores, 3)
head(princomp(Xi)$scores, 3)
```

   (a) It turns out that the singular value decomposition of the (centred) data matrix can be used to construct the scores directly as $\mathbf{UD}$. Look at the notes/slides and think about why this is true, and write an R function, `pcScoresSvd` to implement this. Test it on the iris data, and don't worry about sign flipping (the sign of the score vectors is arbitrary).

   (b) This SVD-based method is more numerically stable than the eigendecomposition method, although if we are only interested in the first few components that is rarely a big deal. In the case of wide data ($p > n$) SVD can also be substantially more efficient. It also gives some additional insight into what the PCA "means". This is essentially how the `prcomp` function in R is implemented (which is almost always preferred to the `princomp` function). Compare your function with this (again, do not worry about the sign).

   (c) Simulate some random (eg.) $5,000 \times 1,000$ test data, and time your two implementations.

[1] Solutions will be made available from the course website before the end of the session.

(d) Think about how to use the SVD to compute the empirical variances or standard deviations of the scores, then create a function that takes as input the data matrix `X` and that uses its SVD to compute the standard deviations of the scores. Check it against `prcomp` for the iris data to make sure you've done it correctly.

2. **Ridge Regression**. For a linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

we know that the quadratic loss $L_0(\boldsymbol{\beta}) = \|\boldsymbol{\epsilon}\|^2 = \boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon}$ is minimised wrt $\boldsymbol{\beta}$ when $\boldsymbol{\beta}$ is a solution to the normal equations,

$$\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\mathsf{T}\mathbf{y}.$$

In *ridge regression*, the slightly modified quadratic loss function $L_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{\epsilon}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ is used, for some ridge penalty $\lambda > 0$, which encourages shrinkage of the regression coefficients towards zero.

(a) Show that for a given fixed $\lambda > 0$, the loss $L_\lambda(\boldsymbol{\beta})$ is minimised when $\boldsymbol{\beta}$ is a solution to

$$(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{X}^\mathsf{T}\mathbf{y}.$$

(b) Starting from the singular value decomposition, $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}$, show that the optimal $\widehat{\boldsymbol{\beta}}_\lambda$ can be written as

$$\widehat{\boldsymbol{\beta}}_\lambda = \mathbf{V}\mathbf{D}_\lambda\mathbf{U}^\mathsf{T}\mathbf{y},$$

where $\mathbf{D}_\lambda$ is a diagonal matrix with entries $d_i^\lambda = d_i/(d_i^2 + \lambda)$. Note that this means $\widehat{\boldsymbol{\beta}}_\lambda$ can be computed for as many different λ as desired, all for the cost of one single expensive SVD operation.

(c) In practice, both the data, $\mathbf{y}$, and the covariate matrix $\mathbf{X}$ are centred before ridge regression is applied, since then the model can be fit without an intercept, and typically you would not want to shrink the intercept. Write a function,

```
ridge(y, X, lambda)
```

which expects an n-vector `y`, an $n \times p$ matrix `X`, and a q-vector of λ values where the ridge solution is required. The function should return a $p \times q$ matrix of ridge regression parameters, with each column representing a solution for a given λ.

(d) For the `trees` dataset, regress volume on the other two variables for a range of shrinkage parameters.

```
ridge(trees[,3], trees[,1:2], c(0,exp(0:5)))
```

Ensure that your solution matches up with that of `lm` in the case $\lambda = 0$.

3. **Interactively exploring optimisation routines**. For this exercise, we will use the `FLtools` package from:

`https://bitbucket.org/finnlindgren/FLtools/`

developed by Finn Lindgren (a previous lecturer for this course). You can install it with:

```
library(devtools)
devtools::install_bitbucket("finnlindgren/FLtools")
```

If you don't have the `devtools` package, first install it with:

```
install.packages("devtools")
```

Once you have installed the `FLtools` package, you should be able to load it with

```
library(FLtools)
```

Make sure you have this package installed before proceeding to the next step.

(a) Start the `optimisation` shiny app:

```
FLtools::optimisation()
```

This should start a Shiny web application. It will also attempt to start up a tab in your browser connected to the session. If this doesn't work, just connect your browser to the URL of the Shiny app. Make sure the Shiny app is running in a browser window before proceeding to the next step.

(b) For the "Simple (1D)" and "Simple (2D)" functions, familiarise yourself with the "Step", "Converge", and "Reset" buttons.

(c) Choose different optimisation starting points by clicking in the figure.

(d) Explore the different optimisation methods and what they display in the figure for each optimisation step[2][3][4]. Also observe the diagnostic output box and how the number of function, gradient, and Hessian evaluations differ between the methods.

(e) For the "Rosenbrock (2D)" function, observe the differences in convergence behaviour for the four different optimisation methods.

(f) For the "Multimodal" functions, explore how the optimisation methods behave for different starting points.

(g) How far out can the optimisation start for the "Spiral" function? E.g., try the "Newton" method, starting in the top right corner of the figure.

4. Write your own code to optimise Rosenbrock's function

$$f(x, z) = 100(z - x^2)^2 + (1 - x)^2$$

by Newton's method. In particular, create the function:

```
newton_step <- function(x0, f, gf, hf, mh = 5, me=0.0001)
```

where:

- `x0` is the initial point;
- `f` is a function that evaluates $f(x, z)$;
- `gf` returns the gradient of $f(x, z)$;
- `hf` returns the Hessian of $f(x, z)$;
- `mh` is the maximum number of step-halving step use during back-tracking;

[2] LS stands for "line search".

[3] The simplex/triangle shapes are shown for each "Simplex" method step in blue. The "best" points for each simplex are connected (magenta).

[4] The Newton methods display the true quadratic Taylor approximations (red) as well as the approximations used to find the proposed steps (blue).

- `me` is the value at which any Hessian eigenvalue falling between 0 and `me` should be fixed.

Your function should return the estimated minimiser $[x^*, y^*]$ of the Rosenbrock function. Make sure that you function implements backtracking for step-length selection and that the Hessian is perturbed to positive definiteness. Ensure that you have implemented it correctly by comparing your output (and implementation) with that of the Shiny app from the first exercise.