ETC3250 - Assignment 2

Ha Lan Nhi (Chelsea) Le - 30100259 | Malo Hamon - 28754131 | Cecilia Li - 31882560

4/21/2020

1. (5pts)This question is about the normal distribution, and how it relates to the classification rule provided by quadratic discriminant analysis.

   a. Write down the density function for a univariate normal distribution ($p = 1$), with mean μ_k and variance σ_k.

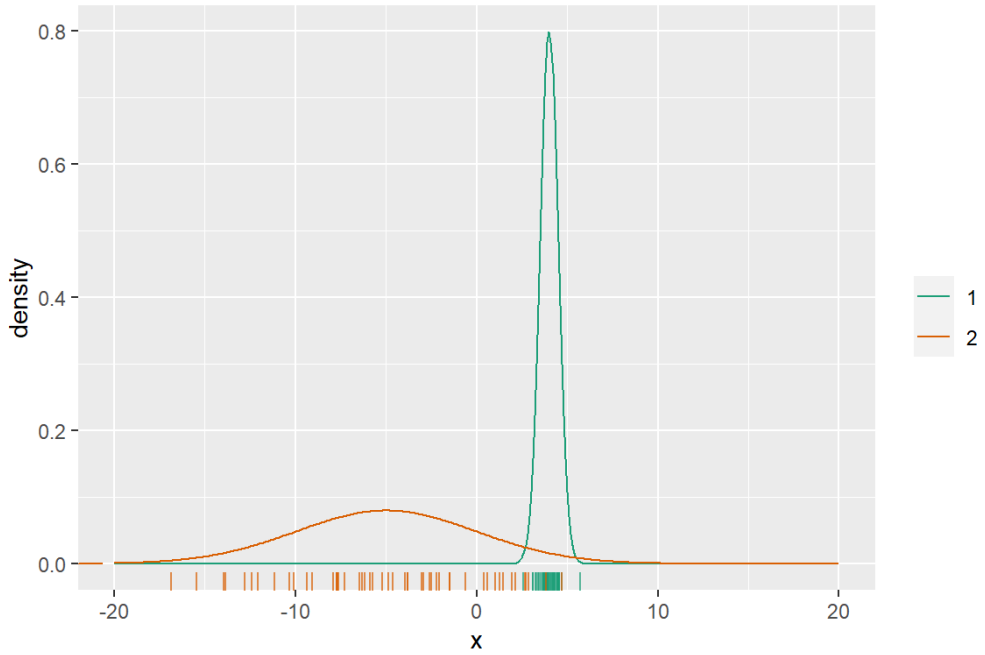
   $$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

   b. Show that the quadratic discriminant rule for two groups ($K = 2$), $\pi_1 = \pi_2$ is equal to: *Assign a new observation x_0 to group 1 if*

   $$-\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)x_0^2 + \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)x_0 - \frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) - \log\sigma_1 + \log\sigma_2 > 0$$

   Bayes classifier assign a new observation x_0 to group 1 if prior probability is highest i.e.
   $p_1(x_0) > p_2(x_0)$ $\frac{\pi_1\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2\sigma_1^2}(x_0-\mu_1)^2\right)}{\sum_{l=1}^{2}\pi_l\frac{1}{\sqrt{2\pi}\sigma_l}\exp\left(-\frac{1}{2\sigma_l}(x-\mu_l)^2\right)} > \frac{\pi_2\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{1}{2\sigma_2^2}(x_0-\mu_2)^2\right)}{\sum_{l=1}^{2}\pi_l\frac{1}{\sqrt{2\pi}\sigma_l}\exp\left(-\frac{1}{2\sigma_l}(x-\mu_l)^2\right)}$ Common denominator $\pi_1\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2\sigma_1^2}(x_0-\mu_1)^2\right) > \pi_2\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{1}{2\sigma_2^2}(x_0-\mu_2)^2\right)$ remove constant, log $\log\pi_1 - \log\sigma_1 - \frac{1}{2\sigma_1^2}(x_0-\mu_1)^2 > \log\pi_2 - \log\sigma_2 - \frac{1}{2\sigma_2^2}(x_0-\mu_2)^2$ shift sides, expand
   $-\frac{1}{2\sigma_1^2}\left(x_0^2 - 2\mu_1x_0 + \mu_1^2\right) + \frac{1}{2\sigma_2^2}\left(x_0^2 - 2\mu_2x_0 + \mu_2^2\right) + \log\pi_1 - \log\pi_2 - \log\sigma_1 + \log\sigma_2 > 0$
   simplify
   $-\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)x_0^2 + \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)x_0 - \frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) - \log\sigma_1 + \log\sigma_2 + \log\pi_1 - \log\pi_2 > 0$
   since $\pi_1 = \pi_2$ $-\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)x_0^2 + \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)x_0 - \frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) - \log\sigma_1 + \log\sigma_2 > 0$

   c. Suppose $\mu_1 = 4, \mu_2 = -5, \sigma_1 = 0.5, \sigma_2 = 5$ simulate a set of 50 observations from each population. Make a plot of the population model, and add these samples as a rug plot on the horizontal axis. (See the lecture notes for a similar plot and code for linear discriminant analysis.)



d. Write down the rule using these parameter values, and sketch the boundary corresponding to the rule on the previous plot.
Using the quadratic discriminate rule from Q1b and substituting $\mu_1 = 4, \mu_2 = -5, \sigma_1 = 0.5, \sigma_2 = 5$:

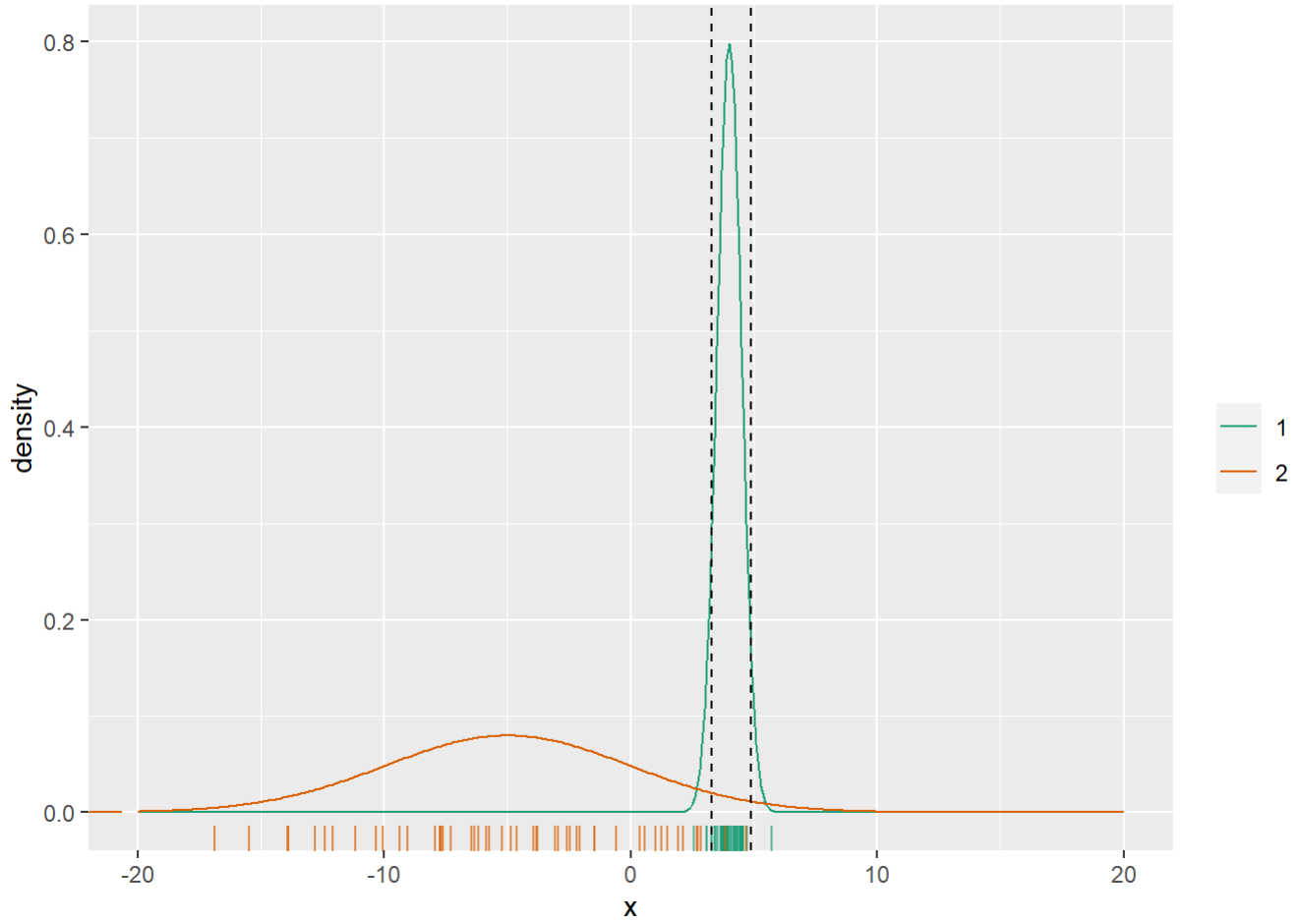
$$-\frac{1}{2}\left(\frac{1}{0.5^2} - \frac{1}{5^2}\right)x_0^2 + \left(\frac{4}{0.5^2} - \frac{-5}{5^2}\right)x_0 - \frac{1}{2}\left(\frac{4^2}{0.5^2} - \frac{(-5)^2}{5^2}\right) - \log 0.5 + \log 5 > 0$$

$$-1.98x_0^2 + 16.2x_0 - 31.5 - \log 0.5 + \log 5 > 0$$

Calculating the discriminate of the quadratic function:

$\Delta = 16.2^2 - 4 * -1.98 * (-31.5 - \log 2.5)$ $x_0 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

This mean when x_0 is between 3.3 and 4.88 assign to class 1.

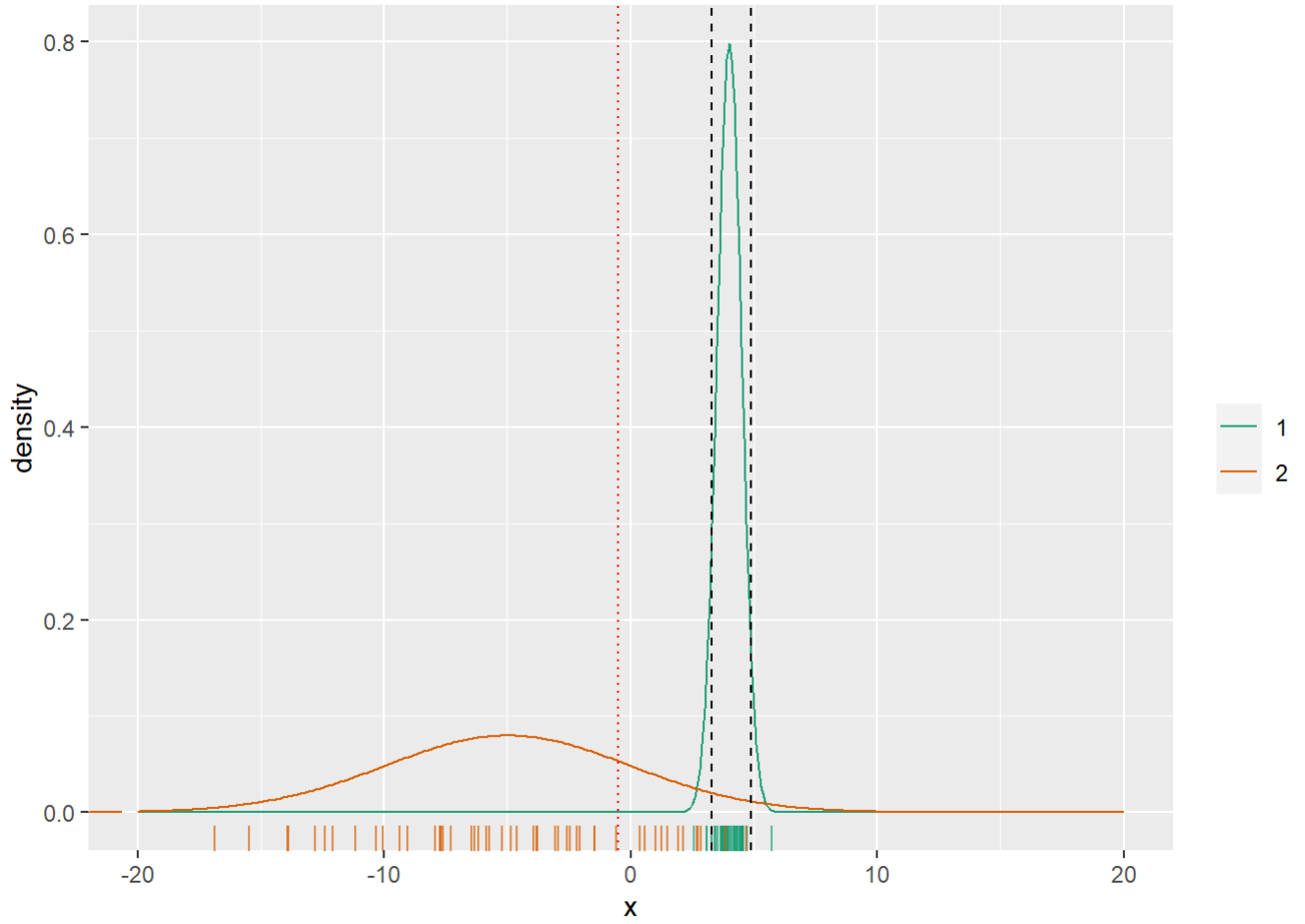


e. If instead you had made a mistake and assumed that the two variances were equal, this would have produced a linear discrimant rule. Mark this boundary on the previous plot. Explain why and how this differs from result of the QDA rule.

With the linear discrimant rule, if $K = 2$ and $\pi_1 = \pi_2$ then we assign x_0 to class if $x_0 > \frac{\mu_1+\mu_2}{2}$, therefore the boundary line is when $x_0 = \frac{4-5}{2}$ $x_0 = -0.5$

Difference between LDA and QDA is that:

- LDA assumes $\sigma_1 = \sigma_2$ whereas QDA does not assume they are the same. Therefore the decision boundary for LDA is when x_0 is the average of the 2 means.
- However in this example σ_1 is a lot lower than σ_2 which mean the assumption that variance is same is badly off. This will result in high bias such as observations from group 2 will be more likely to be misclassified as group 1 (i.e. observations belonging to the class with the higher variance will be more likely to be inaccurately assigned to the lower-variance class)
- QDA on the other hand results in a boundary that allows for the lower variance of group 1. In this case as $p = 1$ i.e. low dimension, the model is still reasonable given the small number of samples.
- QDA is more flexible as it fits a quadratic model whereas LDA fits a linear model.In this case, since the variances for the two classes are significantly different from each other, using QDA's boundaries is more appropriate.



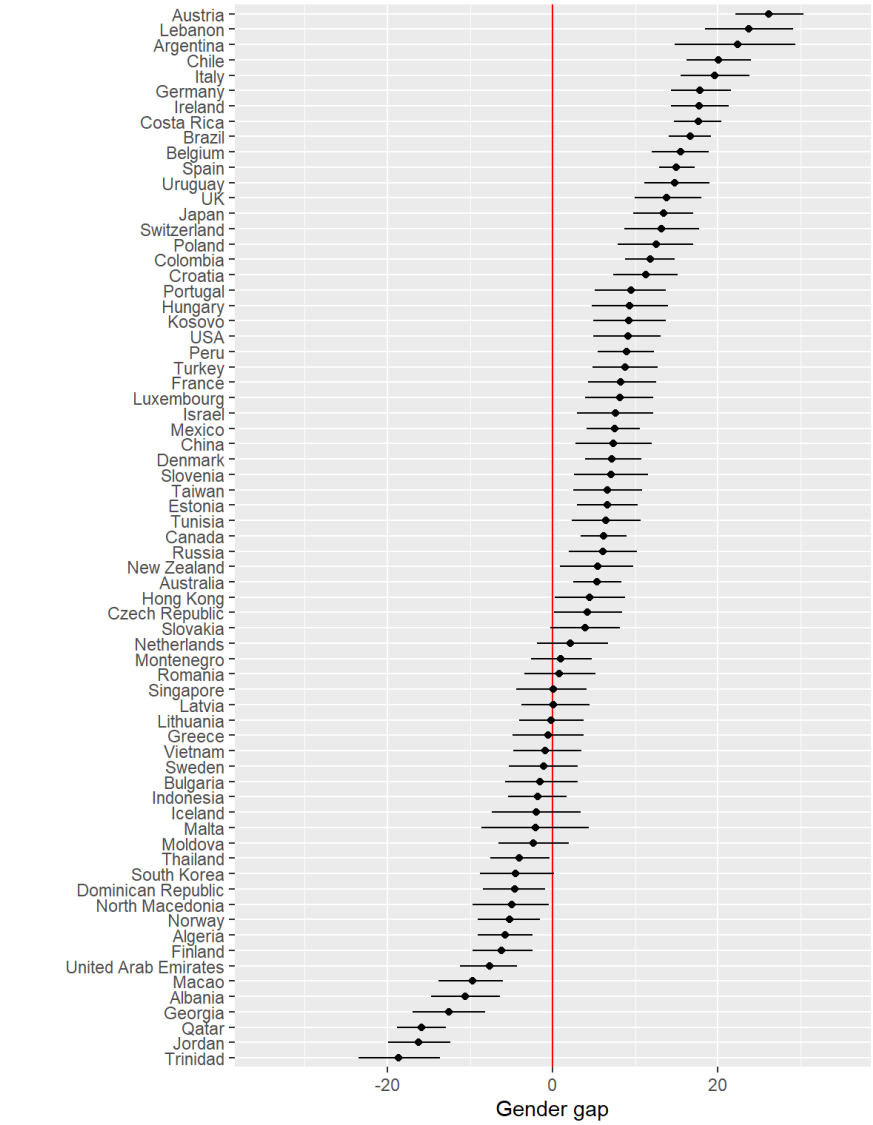
2. (4pts)In this question you are going to practice conducting bootstrap to obtain confidence intervals for a reasonably complicated yet simple analysis.

A significant gender gap in maths performance in favour of male students has returned, despite closing in 2015 Natassia Chrysanthos, Sydney Morning Herald (https://www.smh.com.au/national/nsw/urgent-need-to-address-maths-performance-as-nsw-slumps-in-international-test-20191203-p53ge2.html)

Last December, the 2018 OECD PISA results (http://www.oecd.org/pisa/data/) were released. These are standardised test scores in math, reading and science, of 15 year olds across the globe. It led to a flurry of articles in the news about slipping standards of Australian students. If you also browsed the news of other countries (including New Zealand, Indonesia, Finland), you would find that many had similarly woeful stories. The above headline focuses on the math gap. To explore this, we will compute bootstrap confidence intervals for the difference between weighted averages for boys and girls in each country. The data is from the 2015 results.

This block of code will compute 90% bootstrap confidence intervals for the weighted mean difference.

This block of code will add country names, and make dotplots with confidence intervals for the math gap for each country.



Write a paragraph explaining what you learn about the math gap across the countries tested in 2015.

A large number of countries have math gaps significantly higher than 0, meaning that boys scored higher than girls in the PISA math test in 2015. This can be seen for countries from Austria to Australia (among which are big countries like Germany, UK, Japan, USA, France, China, Russia, New Zealand) in the above graph, where both the weighted mean difference and bootstrap 90% confidece do not intersect the 0 vertical benchmark.

For a couple other countries, there isn't a significantly different from zero gap between boys' and girls' performance in the test (i.e. countries from Hong Kong to Moldova). These are countries with 90% CI intersecting the 0 benchmark. Most notably are Singapore and Latvia where the weighted mean differences are exactly at 0, meaning almost no gap between boys' and girls' scores, a good sign that their math education was doing well at the time.

On the other side of the spectrum are countries where boys performed worse than girls in the 2015 math test. They go from Thailand to Trinidad in the above graph.

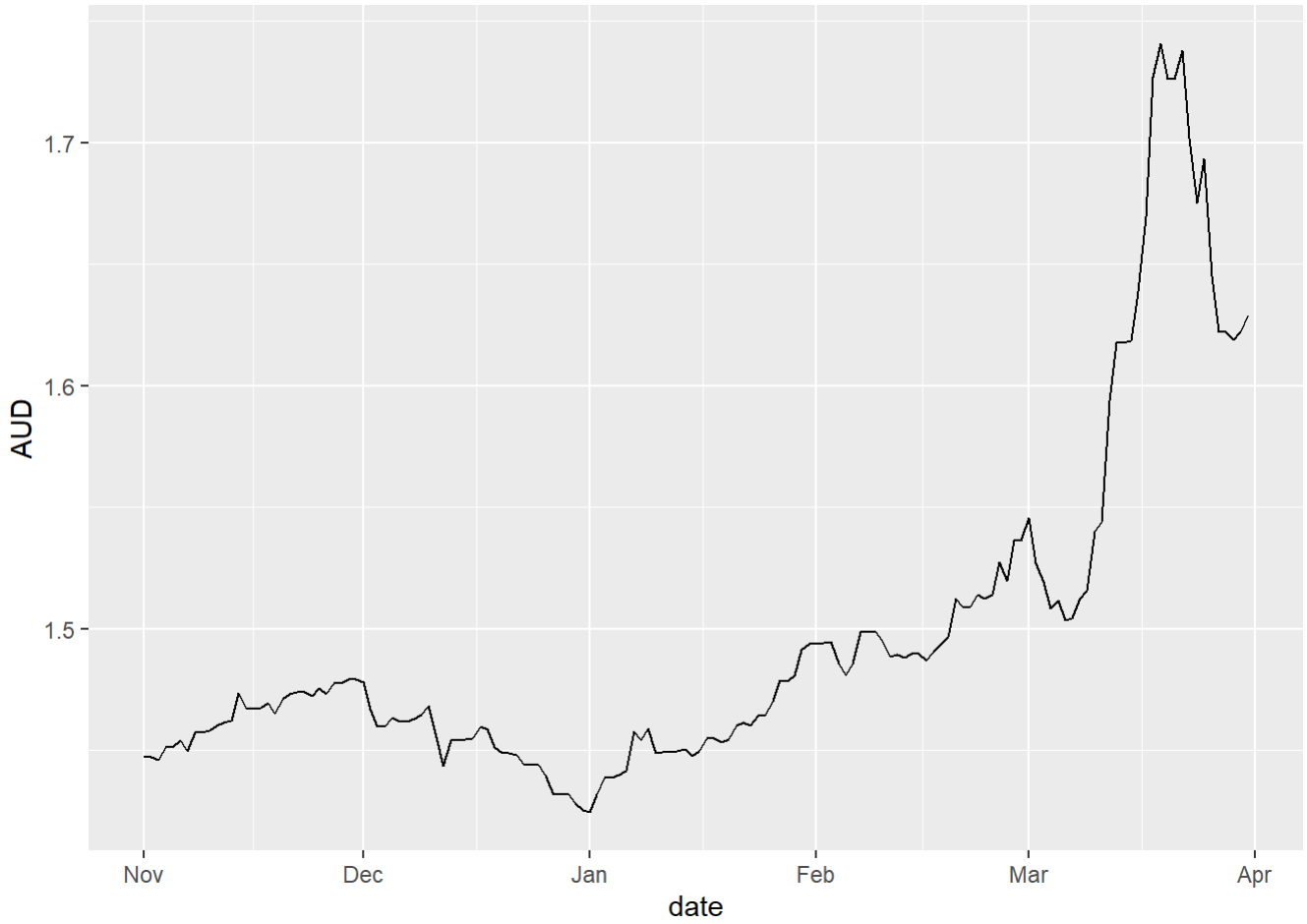
Overall, the results shows that a significant gender gap in maths performance in favour of male students exists for over half of the countries recorded. In other countries, performance is in favour of female students and in a small number of countries, there is not a significant difference.

3. (9pts)A cross-rate is *an exchange rate between two currencies computed by reference to a third currency, usually the US dollar.*

a. (1)What's the data? Make a plot of the Australian dollar against date. Explain how the Australian dollar has changed relative to the US dollar over the 5 month period.

The data is the daily cross-rate of many countries with US dollar as the reference currency (amount of local currency to exchange for 1 USD) from 1 November 2019 to 31 March 2020.

This is a plot of the AUD relative to one USD, meaning that increases are equivalent to depreciation of the AUD since it means one USD can now afford a larger value of the AUD, and vice versa. Over the 5 month period the Australian dollar has weakened against the US dollar, as seen from the upward trend from January onwards, with a big decline in mid-March (reaching a point where 1.75 AUD = 1 USD) as the coronavirus impact takes a toll on the Australian economy.



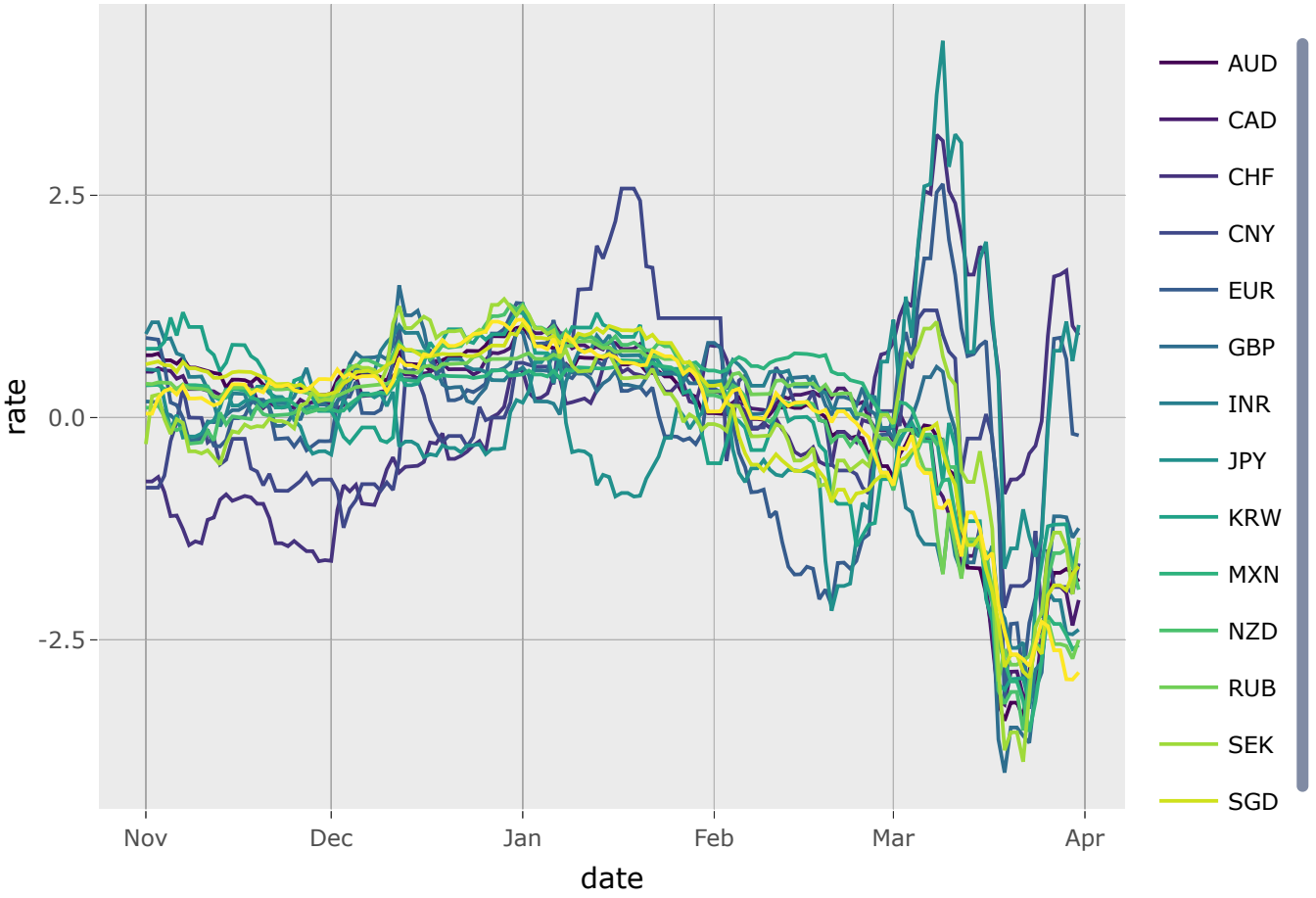
b. (1)You are going to work with these currencies: AUD, CAD, CHF, CNY, EUR, GBP, INR, JPY, KRW, MXN, NZD, RUB, SEK, SGD, ZAR. List the names of the countries and currency name that these codes refer to. Secondary question: why is the USD a constant 1 in this data.

- AUD - Australia, Australian dollar
- CAD - Canada, Canadian dollar
- CHF - Switzerland, Swiss Franc
- CNY - China, Chinese Yuan
- EUR - Most EU countries (e.g. France, Germany, Italy), Euro
- GBP - United Kingdom, British pound sterling
- INR - India, Indian Rupee
- JPY - Japan, Japanese Yen
- KRW - South Korea, South Korean Won
- MXN - Mexico, Mexican Peso
- NZD - New Zealand, New Zealand Dollar
- RUB - Russia, Russian Ruble
- SEK - Sweden, Swedish Krona
- SGD - Singapore, Singapore Dollar
- ZAR - South Africa, South Africa Rand

The US is the base rate, against which all other currencies are compared, that's why it is always 1. An increase in a currency's value against this USD base rate means that currency is depreciating relative to the USD and vice versa.

c. (2)The goal of the principal component analysis is to examine the relative movement of this subset of currencies, especially since coronavirus emerged until the end of March. PCA is used to summarise the volatility (variance) in the currencies, relative to each other. To do this you need to:

- Standardise all the currencies, individually. The resulting values will have a mean 0 and standard deviation equal to 1.
- Flip the sign so that high means the currency strengthened against the USD, and low means that it weakened. Its easier to explain trends, if you don't need to talk with double-negatives.
- Make a plot of all the currencies to check the result.



d. (5)Conduct a principal component analysis on the subset of currencies. You need to work from a wide format of the data, where dates are in the columns, and currencies are in the rows. Normally, PCA operate on standardised variables but for this data, you need to NOT standardise each date. Think about why this is best.

- Why is this data considered to be high-dimensional?

In this case, the data in wide format has dates on the columns as variables and currencies on the rows as observations. Dimension of the data, p, is defined as the number of variables. This data is considered high-dimensional as each day over the 5 months is one dimension. As there are 152 days, this represent the number of dimensions in the data. Moreover, there are also a lot more variables (dates) than there are observations (currencies) if we work with the wide format like this case.

```
- Make a scree plot to summarise the variance explained by cumulative principal components. How much of the total variation do two PCs explain?
```

Two PCs explain 81% (or 80.71%) of the total variation in the data.

```
- Plot the first two principal components. Write a summary of what you learn about the similarity and difference between the currencies.
```

The plot shows two clusters of currencies, one including AUD, CAD, GBP, INR, KRW, MXN, NZD, RUB, SEK, SGD, ZAR, and the other including EUR and JPY, which suggests that these currencies have similar currency movement profiles with regards to the USD over time. Chinese yuan (CNY) and Swiss Franc (CHF) seem to be far from the above two clusters of currencies, hence, they behave differently from each other and from the rest of the currencies in this subset. They might be outliers and should be investigated further.

Another way to look at this is to say there is only one cluster of currencies (AUD, CAD, GBP, INR, KRW, MXN, NZD, RUB, SEK, SGD, ZAR) and CNY, CHF, JPY and EUR might be the four outliers behaving differently from each other and from the cluster.

```
- Plot the loadings for PC1. Add a base line set at $1/\sqrt{152}$. Why use this as a guide? What time frame generated a big movement (or divergence) in the currencies? Which currencies strengthened relative to the USD in that period? What happened to the Australian dollar? Answer these questions in a paragraph, written in your own words.
```

We have chosen to add a base line at $\pm 1/\sqrt{152}$ instead as there are 152 loadings so therefore: $\sum_{j=1}^{152} \phi_{j1}^2 = 1$.
Since $\phi_{j1} = \phi_{k1}$ for all values of j and k between 1 and 152: $152\phi_{j1}^2 = 1$
$\phi_{j1} = \pm\frac{1}{\sqrt{152}}$

If the loadings are all similar i.e. close to $1/\sqrt{152}$, that means they are not significantly different from being equal. This is not the case for the cross-rate data set.

In particularly in March 2020, there were big movements and high volatility in the currencies (i.e. the absolute value for the loadings are high). We can verify this from the AUD plot in question 3a. PC1 loadings during this period are postive and significantly different from 0. This is likely due to the impact of COVID-19 on worldwide economies.

EUR, CHF and JPY strengthened during this period.

The AUD decreases in relative value to USD during this time.

```
- Do the same analysis for PC2. What time frame was there another movement of currencies? Which currencies primarily strengthened, and which weakened during this period?
```

There were two other period of movements detected by PC2: 1. during November and December, 2. from mid-January to end-February.

During November and December, CNY and CHF depreciated relative to the USD.

During January, only CNY strengthened and JPY weakened with regards to the USD, the rest of the currencies remained rather stable.

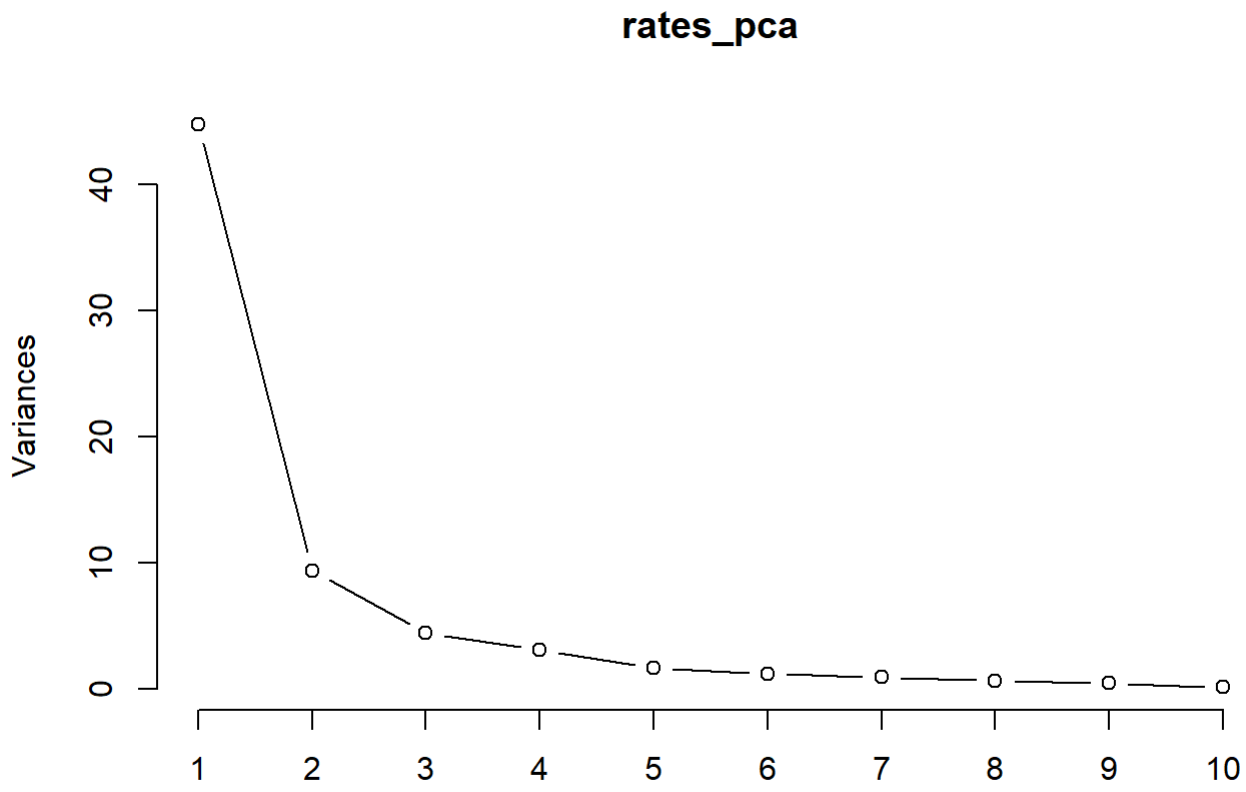
During February, CNY, JPY, CHF and especially EUR depreciated whereas the remaining currencies still remained flat.

During this time period, the AUD only saw a slight steady decrease, which can also be regarded as remaining stable.

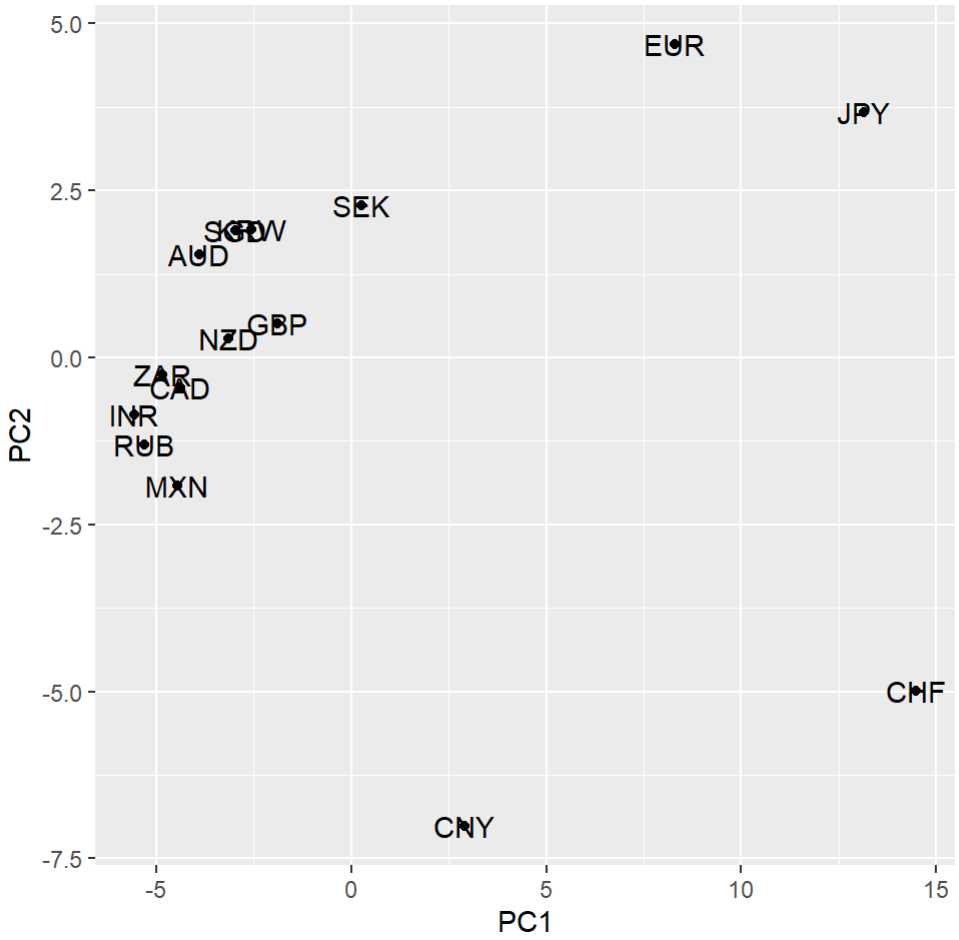
```
- Finish with a paragraph summarising what variability the principal components analysis is summarising. What dimension reduction is being done?
```

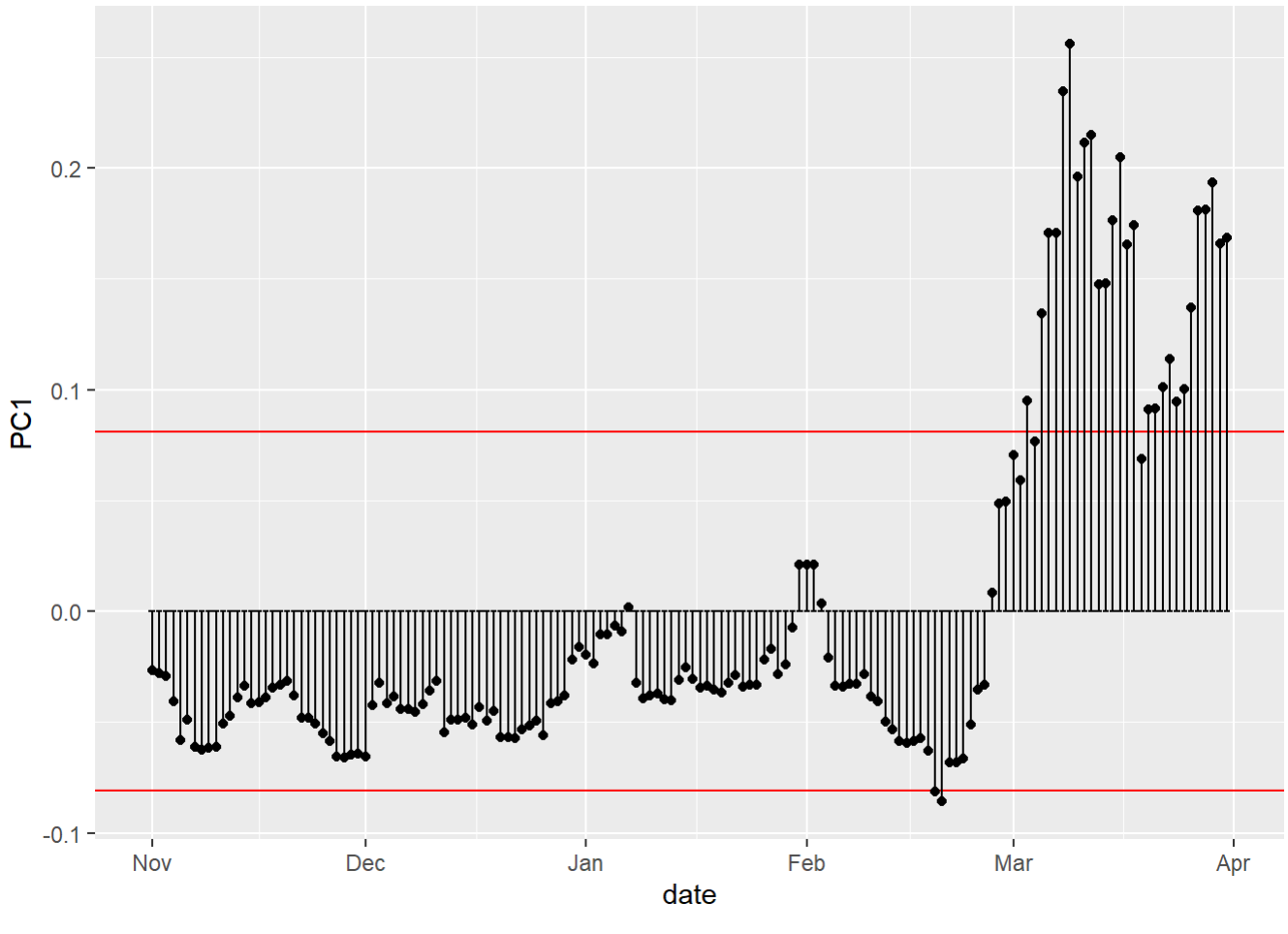
The principle components analysis summarises the variability in the currencies' movements relative to the USD and relative to each other, especially during the first few months of 2020 to indicate whether COVID-19 has any impact on the behaviour of different currencies. The dates are used as variables in this case because PCA seeks to understand the changes happening to currencies' movement profiles across the whole time period.

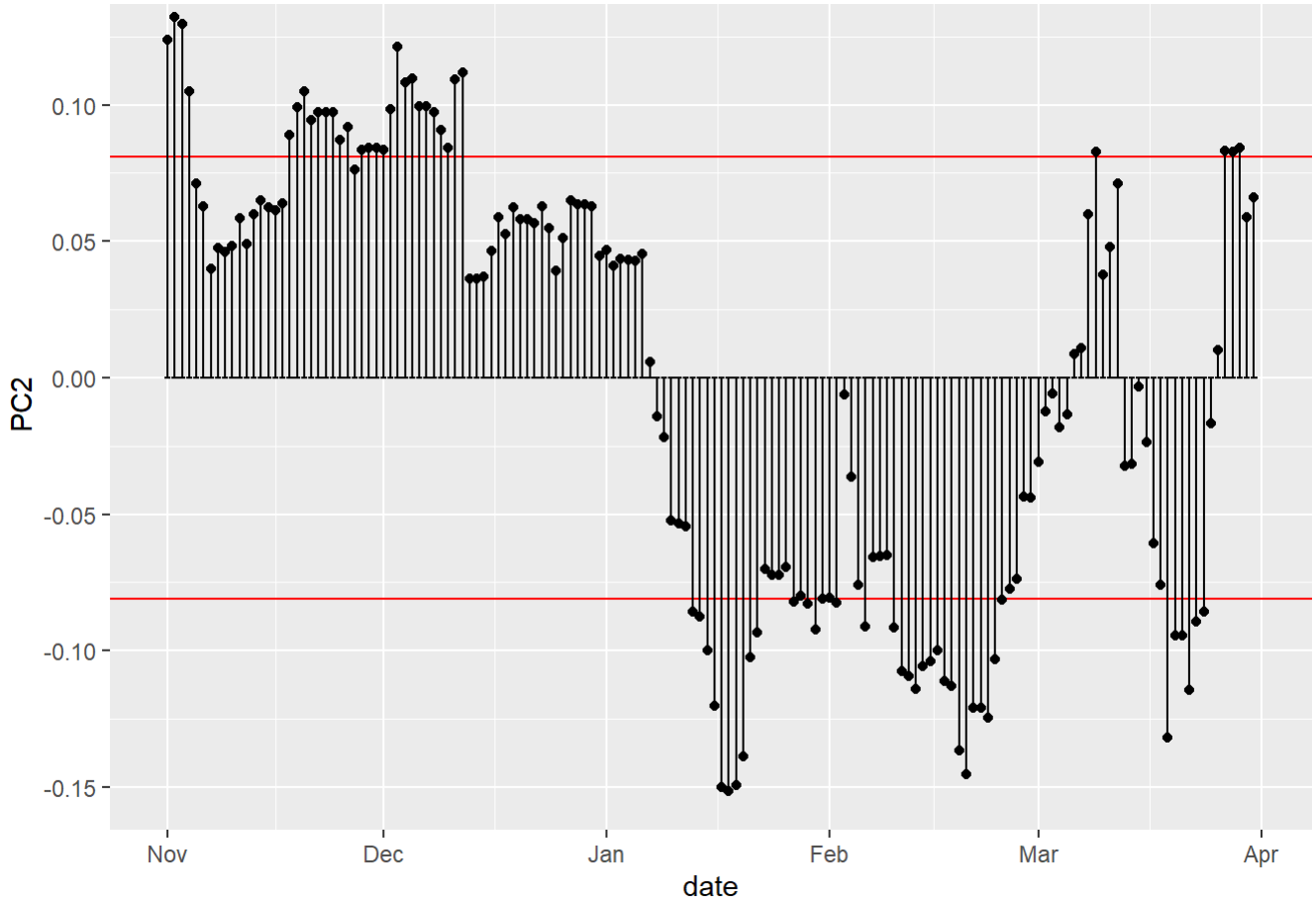
The dimension reduction technique being used here is principle component analysis, which simplifies the problem from estimating 153 coefficients (1 for the intercept and 152 for the 152 dates) to estimating only 2 - 3 coefficients (since we find that using 2 - 3 principle components suffice in explaining most of the variability in the data). Here, we see that only with the first two principle components, we were able to see two distinct periods where currencies varied the most: 1. mid-January to end-February, 2. the whole of March.



```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     6.6889 3.0604 2.11501 1.75429 1.28077 1.10197 0.96250
## Proportion of Variance 0.6674 0.1397 0.06673 0.04591 0.02447 0.01811 0.01382
## Cumulative Proportion  0.6674 0.8071 0.87387 0.91978 0.94425 0.96236 0.97618
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.79487 0.66991 0.37970 0.36507 0.30562 0.28223 0.25608
## Proportion of Variance 0.00942 0.00669 0.00215 0.00199 0.00139 0.00119 0.00098
## Cumulative Proportion  0.98561 0.99230 0.99445 0.99644 0.99783 0.99902 1.00000
##                           PC15
## Standard deviation     9.126e-16
## Proportion of Variance 0.000e+00
## Cumulative Proportion  1.000e+00
```


EUR
JPY
SEK
AUD
NZD
GBP
ZAR
CAD
INR
RUB
MXN
CHF
CNY
5.0
2.5
0.0
-2.5
-5.0
-7.5
PC2
-5
0
5
10
15
PC1


0.2
0.1
0.0
-0.1
PC1
Nov
Dec
Jan
Feb
Mar
Apr
date



4. (2pts)What's wrong with the following statement?

Principal component analysis is a dimension reduction technique.

- Although PCA seeks to summarise a large set of variables with a smaller set of representative variable, it is not a feature selection method as the principal components is a linear combination of the p original features and does not reduce the number of features. Hence, even when choosing M (the number of principle components to do principle component regression) smaller than p (the number of variables in the orginal data set), this will not result in the development of a model that relies entirely upon a small subset of the original features. Instead, dimension reduction simply means this approach helps "reduce the problem of estimating (p + 1) coefficients $\beta_0, \ldots, \beta_p$ to the simpler, more manageable problem of estimating (M + 1) coefficients $\theta_0, \ldots \theta_M$, where M < p. In other words, the dimension of the problem has been reduced from (p + 1) to (M + 1)." [1]. The dimension of the data is still the same, but the dimension of the model we are fitting has been reduced. Of course, this only holds if we find that the first few principle components suffice in explaining most of the variability in the data. If otherwise we find that the same number of principle components as the number of original variables are needed in regression, then no dimension reduction will take place.
- In additional to being a dimension reduction technique, PCA is also used for exploring linear correlation between variables.

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani 2017, An Introduction to Statistical Learning with Applications in R, Chapter 6, 6.3, page 229.