Fall 2021 — STAT 2221: Advanced Applied Multivariate Analysis

Homework 3. Due by Tuesday, October 5, 2021 at 9:00 a.m. EST

Topic: Linear dimensionality reduction

Name: Giang Vu

Additional instructions. Add `\newpage` immediately before each problem so that each has its own page. Add `\begin{proof}[Solution.] ... \end{proof}` below each problem for providing your solution. You are welcome to add additional packages to the preamble, but do not modify the existing commands and formatting.

Problem 3.1. Consider the `SwissBankNotes` dataset, available on Canvas and described in more detail in Problem 7.7 on p.235 of Izenman.

- Carry out PCA separately on (i) only the 100 genuine bank notes, (ii) only the 100 counterfeit bank notes, and then (iii) all bank notes.
    - Each time, do PCA using (a) the covariance matrix and (b) the correlation matrix.
        - For choosing the number of PCs, use (α) Kaiser's rule of unity, (β) the RC rank trace method, and (γ) the profile likelihood procedure in Zhu and Ghodsi (2006).

Remark: you can streamline the process by defining and then evaluating a suitable "PCA master function" ...think about it. Also, note that method (β) is not entirely automatic and therefore requires further specification.

Display several plots and summarize your findings. Comment on the similarities and differences observed in your results.

Proof. □

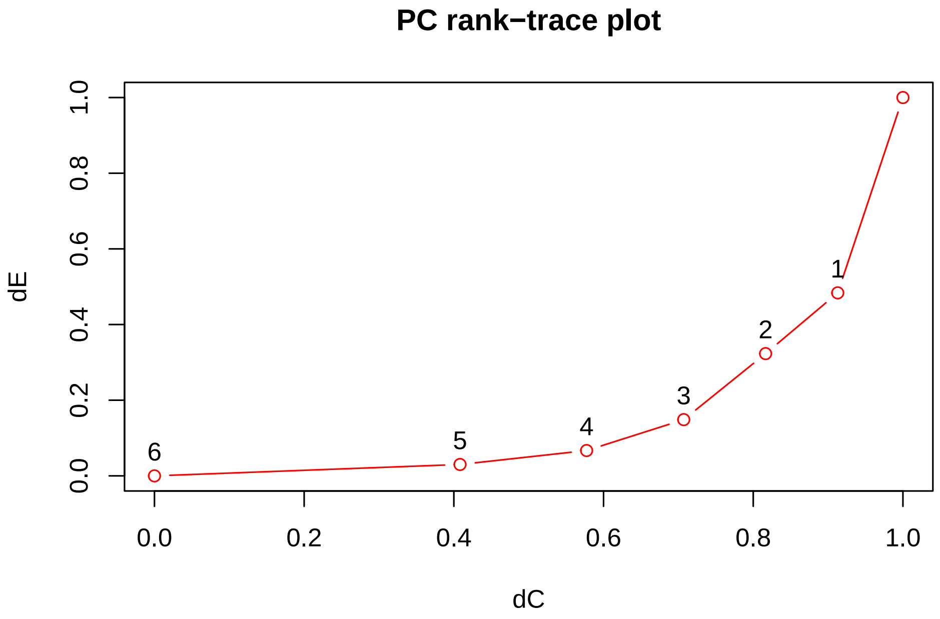
I have defined a master function to display PCA results, along with choices of dimensionality using the 3 following methods:
(α) Kaiser's rule of unity (drop all PC's with variance ≤ 1)
(β) RC rank trace method
(α) Zhu & Ghodsi's profile likelihood procedure
This master function's inputs include "data" which is a dataframe containing all our X variables, each containing in one column, as well as "cor" which is a logical value (TRUE if we want to use correlation matrix, FALSE if we want to use covariance matrix).
The function's outputs are as follows
pca_object: this is the full object containing PCA result from princomp() in base R. When we run the master function alone it will show the estimated standard deviation for each decomposition, but if we save the result of master function in a object, then we can index all the other details of a usual princomp object as well (loadings, scores, etc.).
kaiser: this is the result of choice for dimensionality (what number of PCs to retain) using method (α). TRUE means that that PC is retained, FALSE means that that PC is ommitted.
rank_trace_plot: this is the PC rank trace plot of method (β), plotting $\Delta\hat{C}^{(t)}$ and $\Delta\hat{\Sigma}^{(t)}_{\varepsilon\varepsilon}$ based on equations (7.38) and (7.39) in page 207, Izenman.
profile_likelihood_zhu_ghodsi: this is the result of the estimated q of method (γ), which is just the point up to which we should retain our PCs. I referred to the function dim_select() in packages "igraph" for this procedure.
Please refer to the attached R markdown file if you want more details about the master function.

And below is the results of the function applied on (i) only 100 geniune bank notes, (ii) only 100 counterfeit bank notes, and (iii) all bank notes as well as my discussion.

(i) PCA on only 100 genuine bank notes

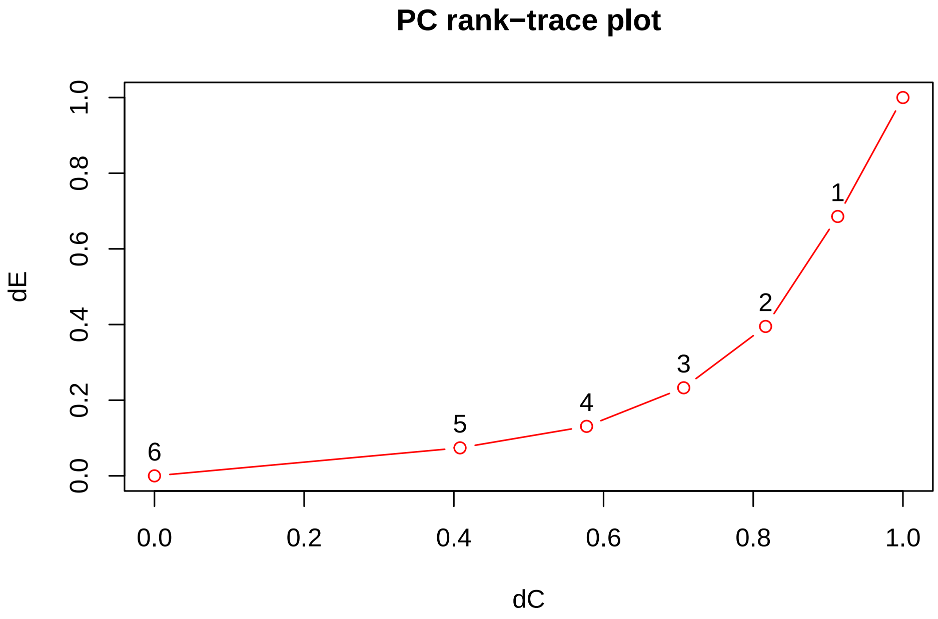
On only 100 genuine bank notes, using both covariance and correlation matrices, the methods of Kaiser and Zhu & Ghodsi's both gave similar numbers of PCs to retain. For covariance, they suggest we keep only 1 PC while for correlation matrix, they suggest we keep the first 2 PCs.
Whereas with the PC rank trace method, we end up with higher dimensionality. Looking at the rank trace plots for covariance matrix and correlation matrix, we are suggested to keep 3-4 and 4-5 PCs, respectively.
So if our goal is to reduce as much dimensionality as we can, then methods (α) and (γ) would be more suitable for that.

(a) PCA using covariance matrix



```
## $pca_object
## Call:
## princomp(x = data, cor = cor)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
## 1.3830517 0.8870354 0.7919350 0.5385210 0.3615268 0.2565023
##
##  6  variables and  100 observations.
##
## $kaiser
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
##   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE
##
## $rank_trace_plot
##
## $profile_likelihood_zhu_ghodsi
## [1] 1
```

(b) PCA using correlation matrix

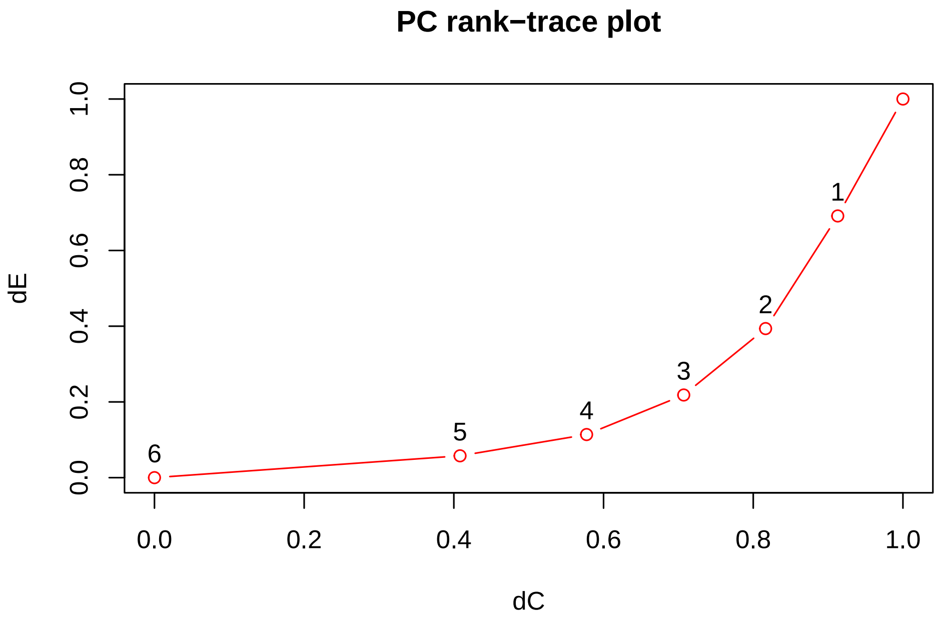


```
## $pca_object
## Call:
## princomp(x = data, cor = cor)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
## 1.4845355 1.3025778 0.9827302 0.7634784 0.5715609 0.4733979
##
##  6  variables and  100 observations.
##
## $kaiser
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
##   TRUE   TRUE  FALSE  FALSE  FALSE  FALSE
##
## $rank_trace_plot
##
## $profile_likelihood_zhu_ghodsi
## [1] 2
```

(ii) PCA on only 100 counterfeit bank notes

On only 100 counterfeit bank notes, using both covariance and correlation matrices, we get quite similar results, and the overall results don't seem too different from the data with 100 genuine bank notes. The methods of Kaiser and Zhu & Ghodsi's both gave fairly similar numbers of PCs to retain. For covariance matrix case, Kaiser's method suggests we keep only 1 PC while Zhu & Ghodsi's suggests we keep 2. For correlation matrix, they both suggest we keep the first 2 PCs.
Whereas with the PC rank trace method, we again have higher dimensionality. Looking at the rank trace plots for covariance matrix and correlation matrix, we are suggested to keep 5 and 4 PCs, respectively.
Again, we have method (γ) giving us the highest number of PCs to retain, thus lesser dimentionality reduction.

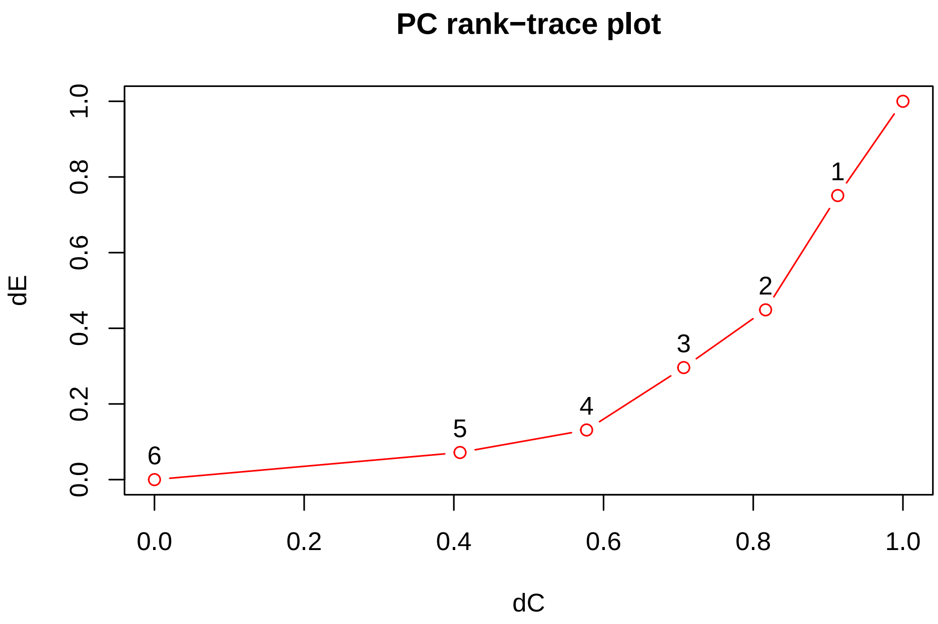
(a) PCA using covariance matrix



```
## $pca_object
## Call:
## princomp(x = data, cor = cor)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
## 1.1129305 0.9867567 0.7493592 0.5649973 0.4101951 0.3144159
##
##  6  variables and  100 observations.
##
## $kaiser
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
##   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE
##
## $rank_trace_plot
```

```
##
## $profile_likelihood_zhu_ghodsi
## [1] 2
```

(b) PCA using correlation matrix



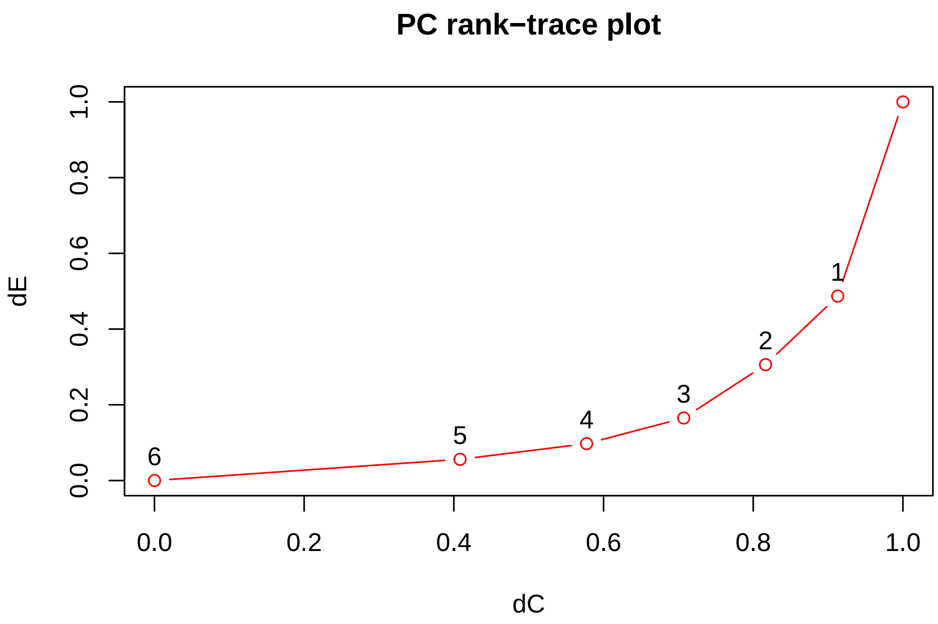
```
## $pca_object
## Call:
## princomp(x = data, cor = cor)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
## 1.3914793 1.3284814 0.9941399 0.8822512 0.5675470 0.4584009
##
##  6  variables and  100 observations.
##
## $kaiser
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
##   TRUE   TRUE  FALSE  FALSE  FALSE  FALSE
##
## $rank_trace_plot
##
## $profile_likelihood_zhu_ghodsi
## [1] 2
```

(iii) PCA on all bank notes

On the entire data of 200 bank notes, using both covariance and correlation matrices, we get quite similar results again, and the overall results seem similar from the 2 previous cases with partial data. The methods of Kaiser and Zhu & Ghodsi's give different numbers of PCs to retain now. For both covariance matrix and correlation matrix, Kaiser's method suggests we keep 2 PCs, while Zhu & Ghodsi's suggests we keep only 1. Using the PC rank trace method, we also have higher dimensionality. Looking at the rank trace plots for covariance matrix and correlation matrix, which are very similar to each other, we are suggested to keep 3-4 PCs.

Overall, we have fairly similar results regardless of what data we use, and out of the three methods to choose dimensions, methods (γ) always reduces the least, while the other two methods reduces dimensionality the most.

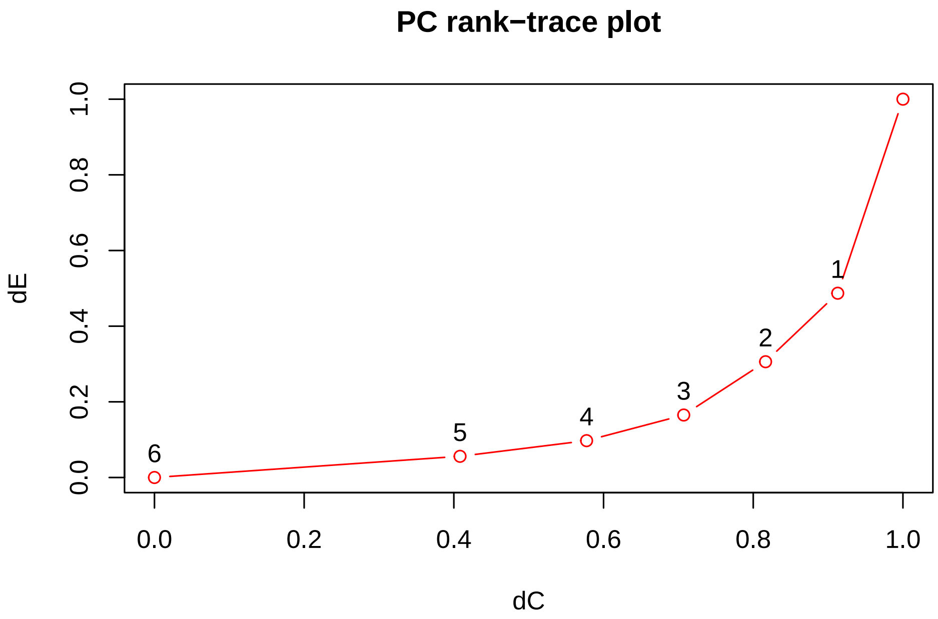
(a) PCA using covariance matrix



```
## $pca_object
## Call:
## princomp(x = data, cor = cor)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
## 1.7119668 1.1276938 0.9298857 0.6689692 0.5170431 0.4335153
##
##  6  variables and  200 observations.
##
## $kaiser
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
##   TRUE   TRUE  FALSE  FALSE  FALSE  FALSE
##
```

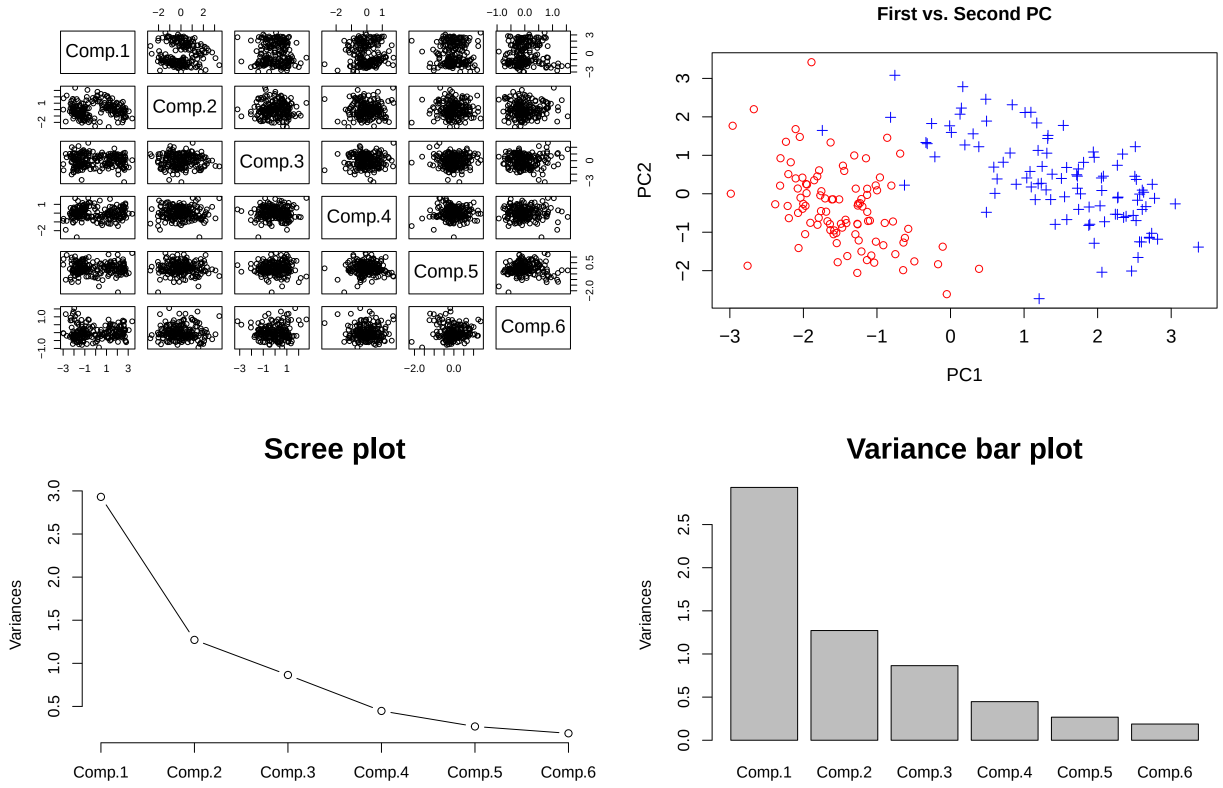
```
## $rank_trace_plot
##
## $profile_likelihood_zhu_ghodsi
## [1] 1
```

(b) PCA using correlation matrix



```
## $pca_object
## Call:
## princomp(x = data, cor = cor)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
## 1.7162629 1.1305237 0.9322192 0.6706480 0.5183405 0.4346031
##
##  6  variables and  200 observations.
##
## $kaiser
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
##   TRUE   TRUE  FALSE  FALSE  FALSE  FALSE
##
## $rank_trace_plot
##
## $profile_likelihood_zhu_ghodsi
## [1] 1
```

After looking at everything, I am inclined to choose the first 2 PCs. And for the full dataset with 200 bank notes using covariance matrix, I'm including the pair-wise plot for 6 PCs, as well as the standalone plot of the first and second PCs. And there's also a scree plot to see the variances explained by PCs too, which shows us that indeed the first 2 PCs explain most of the variances.


Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
First vs. Second PC
PC2
PC1
Scree plot
Variances
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
Variance bar plot
Variances
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6