Principal Components Analysis 2

-Experimental Statistics II-

Lecturer: Darren Homrighausen, PhD

Two common uses of PCA

Exploratory Data Analysis (EDA): Looking at the estimated representation of the observations and explanatory variables

Principal Components Regression (PCR): Use the principal component scores as the inputs to a regression procedure to discover relationships between the explanatory variables and the response

We'll cover both of these, starting with EDA.

Using PCA for EDA

The goal here is two-fold:

We want to get an idea of the dimension of our data

We would like to know how the explanatory variables are related to each other

Using PCA for EDA: How Many Dimensions?

PCA finds the rotation that maximizes variance

We can order the PCs by how much variance each one explains.

Then, we retain the PCs that explain "enough" variance

Using PCA for EDA: How Many Dimensions?

We can visualize this via scree plots

We can use a variety of stopping rules

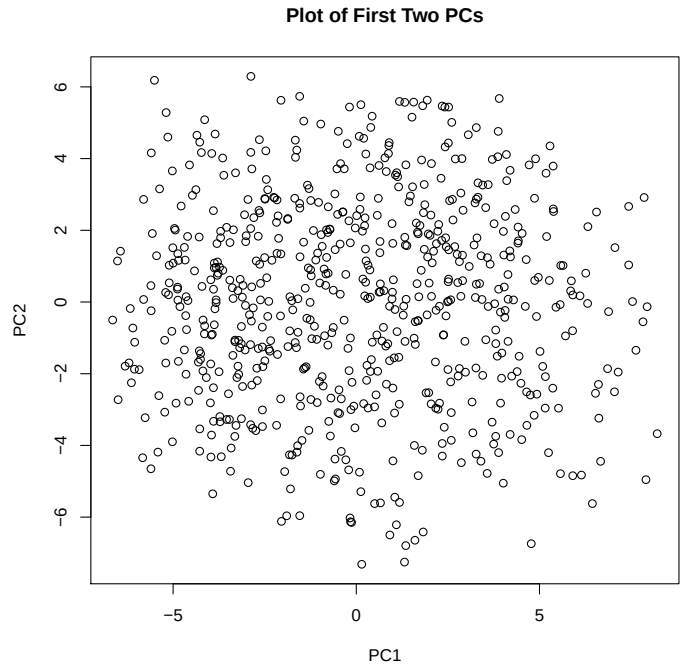
- stop at the elbow or kink in the scree plot
- at a proportion of variance condition

  (Reminder: the first pc is the direction that explains the most variance, the second explains the second most,...)
- a minimum amount of variance condition

  (Known as a "minimum eigenvalue condition")

Using PCA for EDA: Digits Example

We can plot the scores of the first two principal components versus each other:



Note: Each circle in this plot represents a hand written '3'.

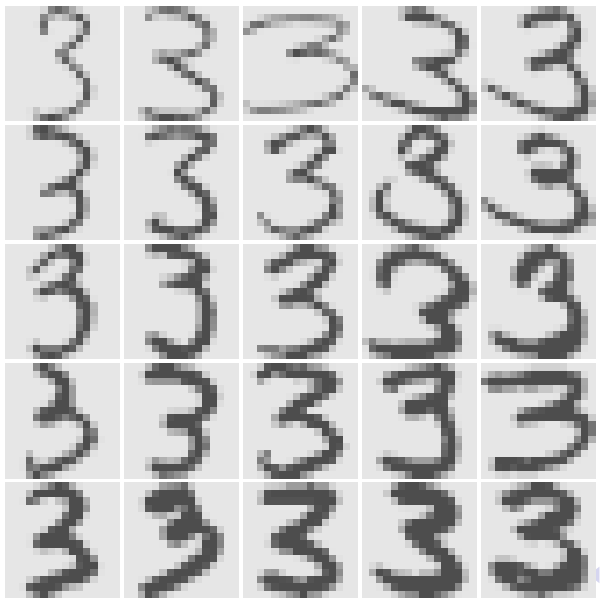
Using PCA for EDA: Digits Example, How Many Dimensions?

Each number represents a vector in $\mathbb{R}^{256}$

(as each square is 16x16 pixels)

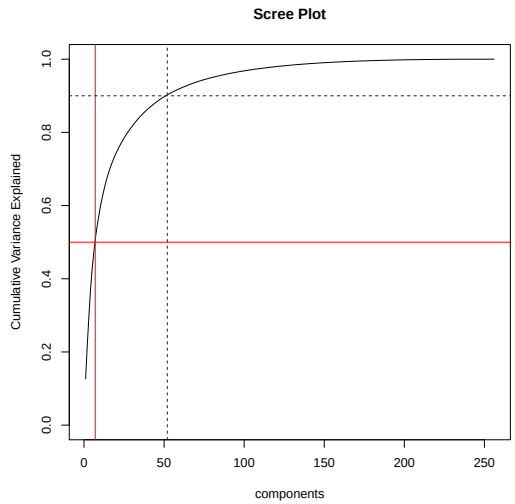
However, hopefully we can reduce this number by re-expressing the digits in PC-land

(For instance, the top-right pixel is always 0 and hence that explanatory variable is uninteresting)



Using PCA for EDA: Digits Example, How Many Dimensions?

Let's look at a scree plot

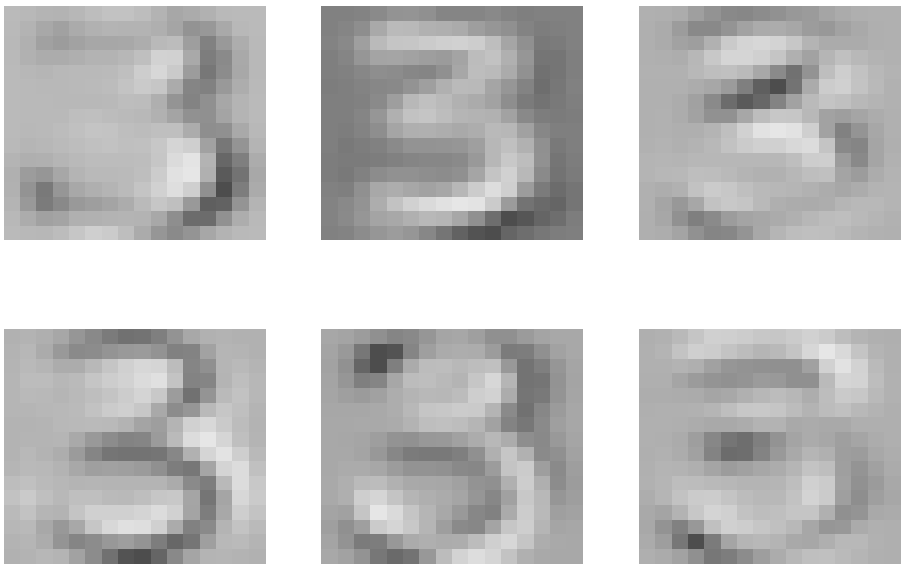


Using the "percent of variance explained" criterion, we can choose the dimension

There are vertical lines at the number of components that explain 50% and 90% of the variance has been explained (at 6 and 52 PCs, respectively)

Using PCA for EDA: Loadings

Looking at the loadings for the first 6 PCs (e.g. dimensions)

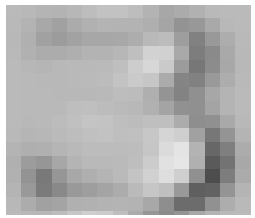
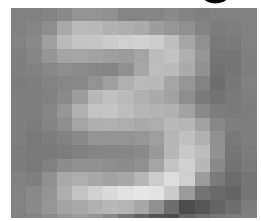
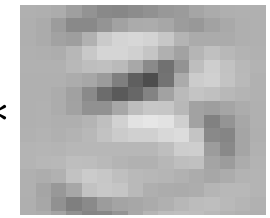


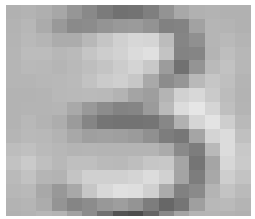
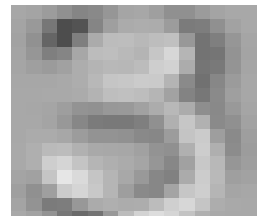
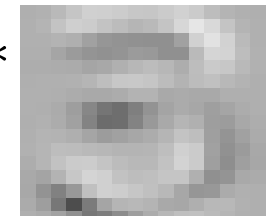
Using PCA for EDA: Looking deeper

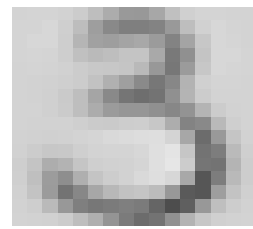
Reminder: Each digit can be represented by a linear combination of the PCs

The scores on the first 6 PCs are 2.52 0.64 2.02 0.17 -4.55 1.97

Hence, a representation of this 3 using these 6 dimensions is

$2.51*$  $+0.63*$  $+2.02*$ 

$0.16*$  $-4.55*$  $+1.96*$  $=$



Principal Components Regression (PCR)

PCR

KEY IDEA; Principal components combines explanatory variables into a linear combination

This linear combination is the "best" one possible, in the sense of inducing a minimum Euclidean distance distortion

Hence, it is natural to run a regression onto the PCA scores instead of the original explanatory variables

Due to being a linear combination, inference about the original explanatory variables in more heuristic

(Though the fact that the PCs are computed without the response means the references distributions are still valid)

We can use the loading to extend the inferences about the PCs back to the original explanatory variables.

Let's go to examples

Deforestation and debt

Reminder: It has been theorized that developing countries cut down their forests to pay off foreign debt:

```
DATA countryDebt;
  INPUT Country $ Debt Deforest Population;
  DATALINES;
    Brazil 86396 12150 128425
    Mexico 79613 2680 74195
    Ecuador 6990 1557 8751
    Colombia 10101 1500 27254
    Venezuela 24870 1430 16171
    Peru 10707 1250  18497
    Nicaragua 3985 550 3022
    Argentina 36664 400 29401
    Bollivia 3810 300 5971
    Paraguay 1479 250 3425
    CostaRica 3413 90 2440
;
```

Deforestation and debt

An issue we had with this data set is that there is substantial multicollininearity

Let's use PCA as a possible remediation

(Ridge regression would be another. Note that ridge regression and what we are proposing are actually very similar)

Do the following:

1. Compute the principal components

   ```
   PROC FACTOR DATA = countryDebt PLOTS=SCREE;
   RUN;
   ```

   (SAS calls them factors as there is a related technique known as Factor Analysis)
2. Decide how many principal components to use (call this number Q)

   (Using a scree plot, minimum eigenvalue condition, ...)
3. Use the first Q scores as an input to a GLM

Exploratory data analysis: Cars data set

Let's look at the cars data set as well

Let's look into this by

- Looking at and plotting the loadings
- Plotting the scores