Principle Component Analysis II

Principle Component Analysis (PCA)

Produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance and are mutually uncorrelated.

Used to:

- Visualize structure in data
- Learn about latent meta-variables
- Produce imputs for subsequent supervised learning

Handwritten Letters

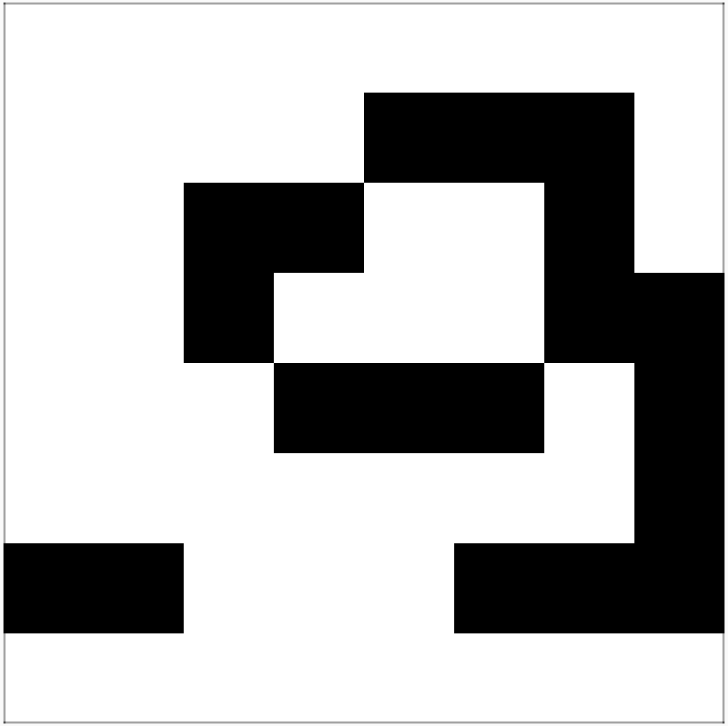
How much information is encoded in a 8 x 8 image of a handwritten letters?

Activity 4

Find the code to download the handwritten data set on the website under Week 11.

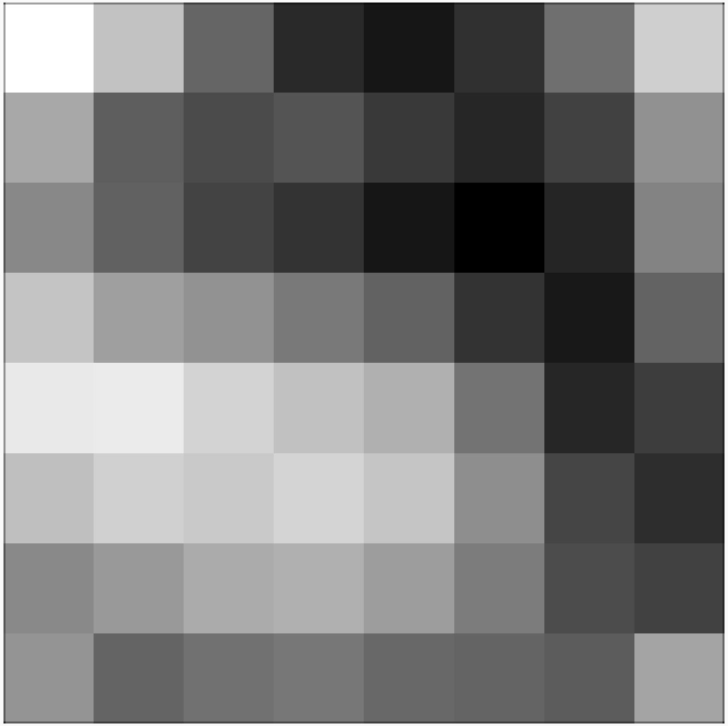
1. What do the columns and rows appear to represent in this dataset?
2. Select a letter of the alphabet and create a new dataset that includes only the images of that letter.
3. Visualize a few of those images using `plot_letter()` function.
4. Compute the *mean* image for that letter and visualize it.

Plot letter



Mean letter

```
g_mean <- colSums(g_data[, -1])/nrow(g_data)
plot_letter(g_mean, hasletter = FALSE)
```



Dimension reduction

Can we capture most of the structure in a smaller number of dimensions?

$$m < p?$$

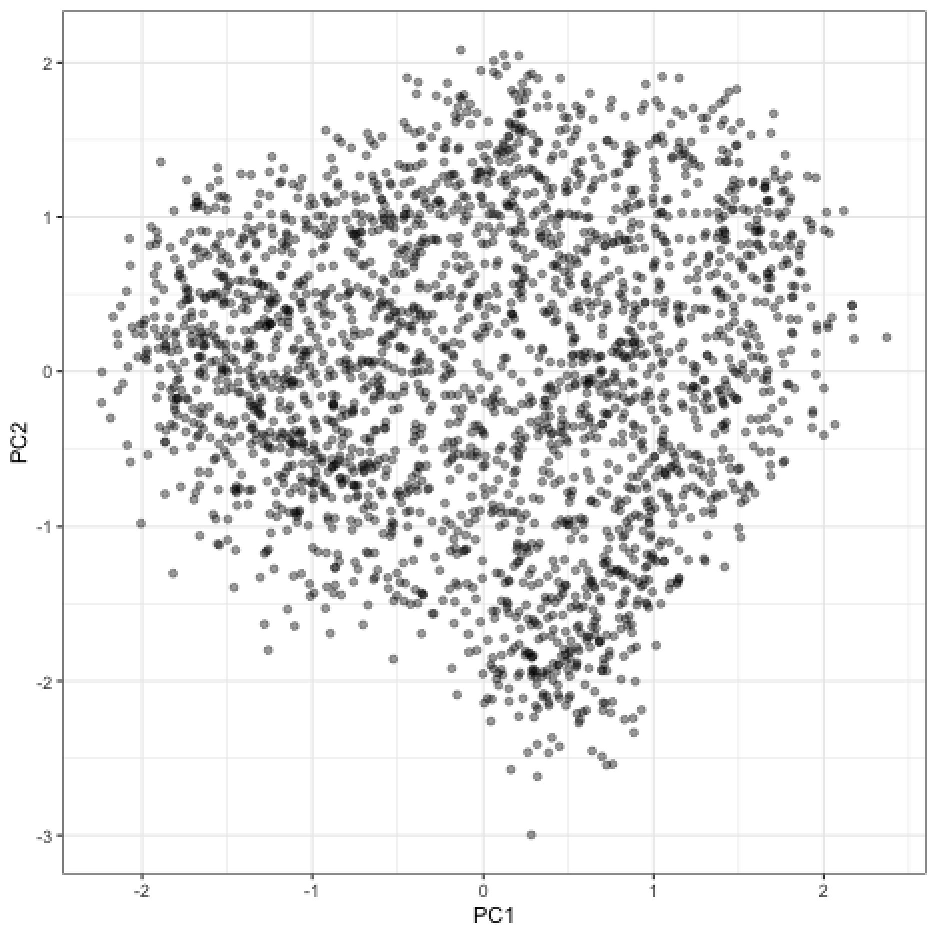
Activity 4, cont.

4.5) Perform PCA on your data set (for your particular letter) using the `prcomp()` function (detailed in the slides from last time). Create a scatterplot of the first two principle component scores of all observations of that letter. 4.6) Construct a scree plot showing the PVE for the first 20 PCs.

Plotting the PCs

```
pca1 <- prcomp(g_data[, -1])
d <- as.data.frame(pca1$x)
library(ggplot2)
p1 <- ggplot(d, aes(x = PC1, y = PC2)) +
  geom_point(alpha = .4) +
  theme_bw()
```

Plotting the PCs

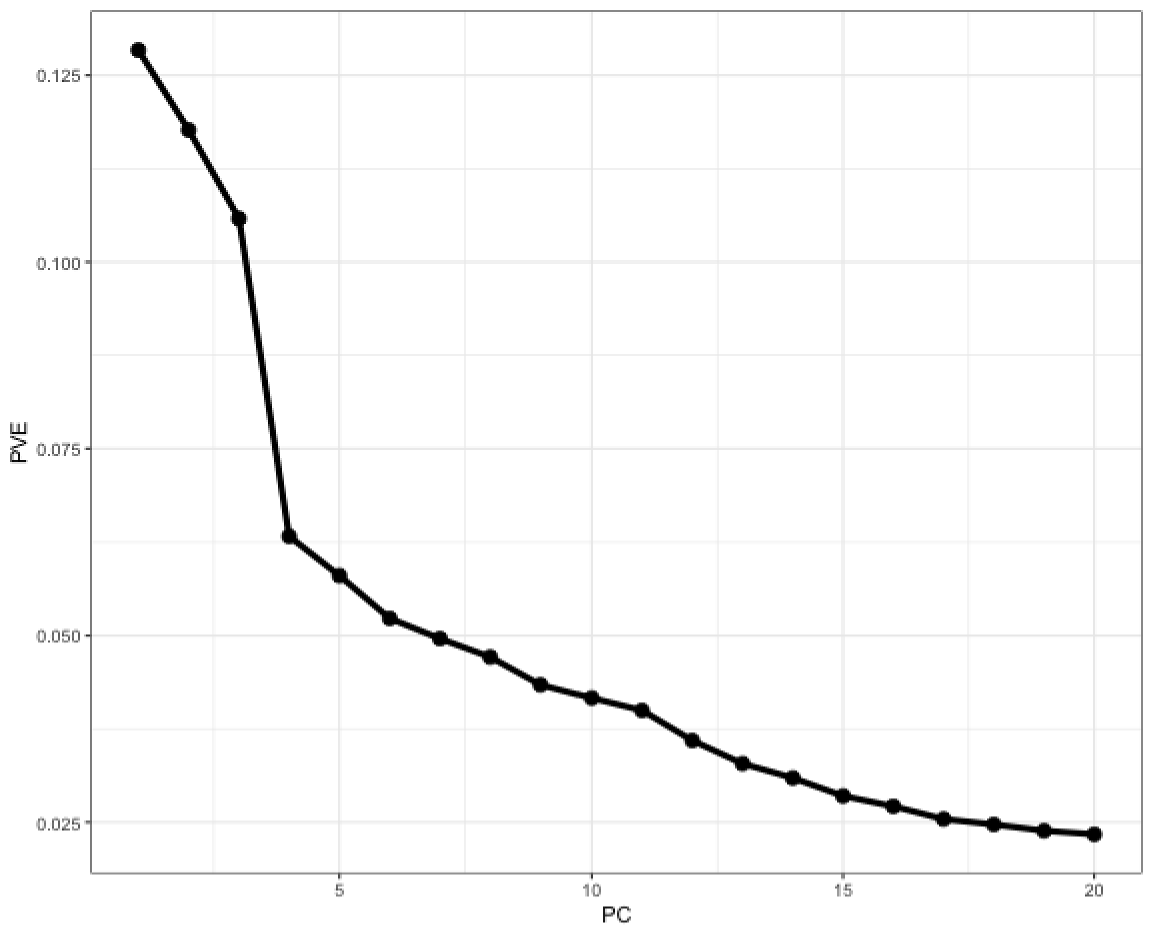


Scree plot

Used to visualize the proportion of variance explained (PVE) by each PC.

```r
d <- data.frame(PC = 1:20,
                PVE = pca1$sdev[1:20]^2 /
                  sum(pca1$sdev[1:20]^2))
p2 <- ggplot(d, aes(x = PC, y = PVE)) +
  geom_line(lwd = 1.5) +
  geom_point(size = 3) +
  theme_bw()
```

Scree plot

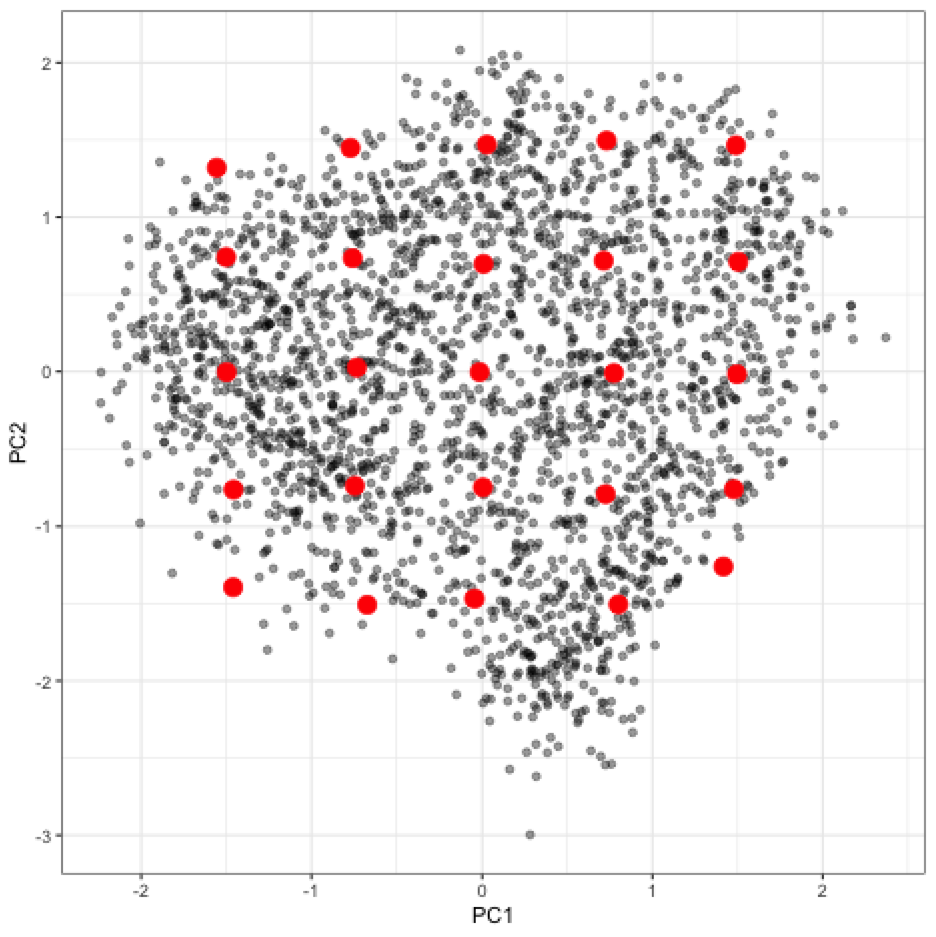


Scree plot

A good amount of the structure in the data resides in the first 4 PCs (PVE: 0.4151918)

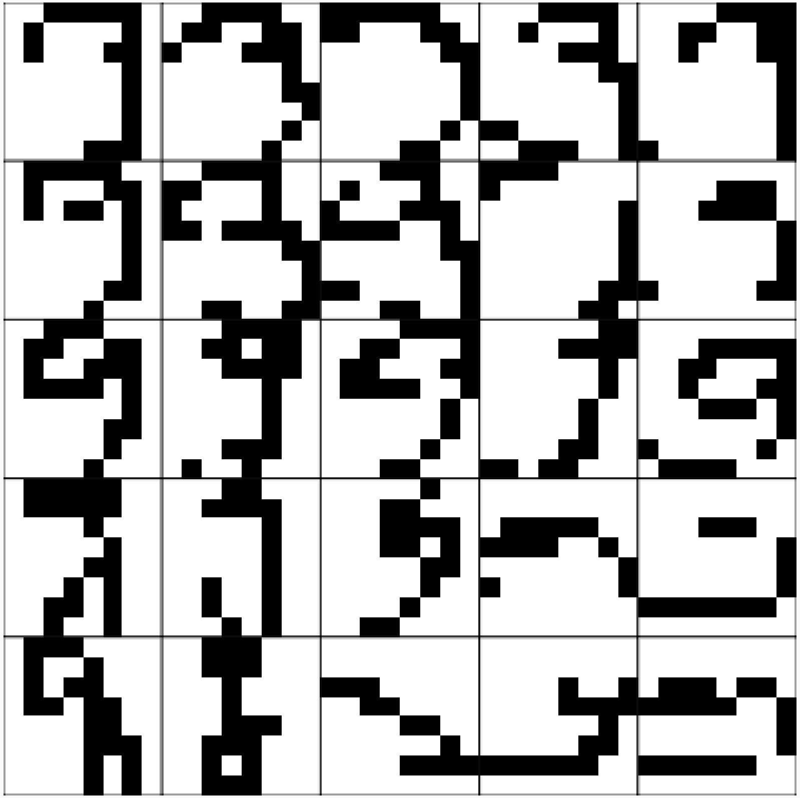
But what do these PCs actually mean?

Select a range of observations



A scatterplot of observations

```
pc_grid(pca1, g_data)
```



Activity 4, cont.

4.7) Use `pc_grid()` to plot a "scatterplot" of 26 observations across their first two principle components. What does each PC seem to be encoding?