Principle Component Analysis III

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.

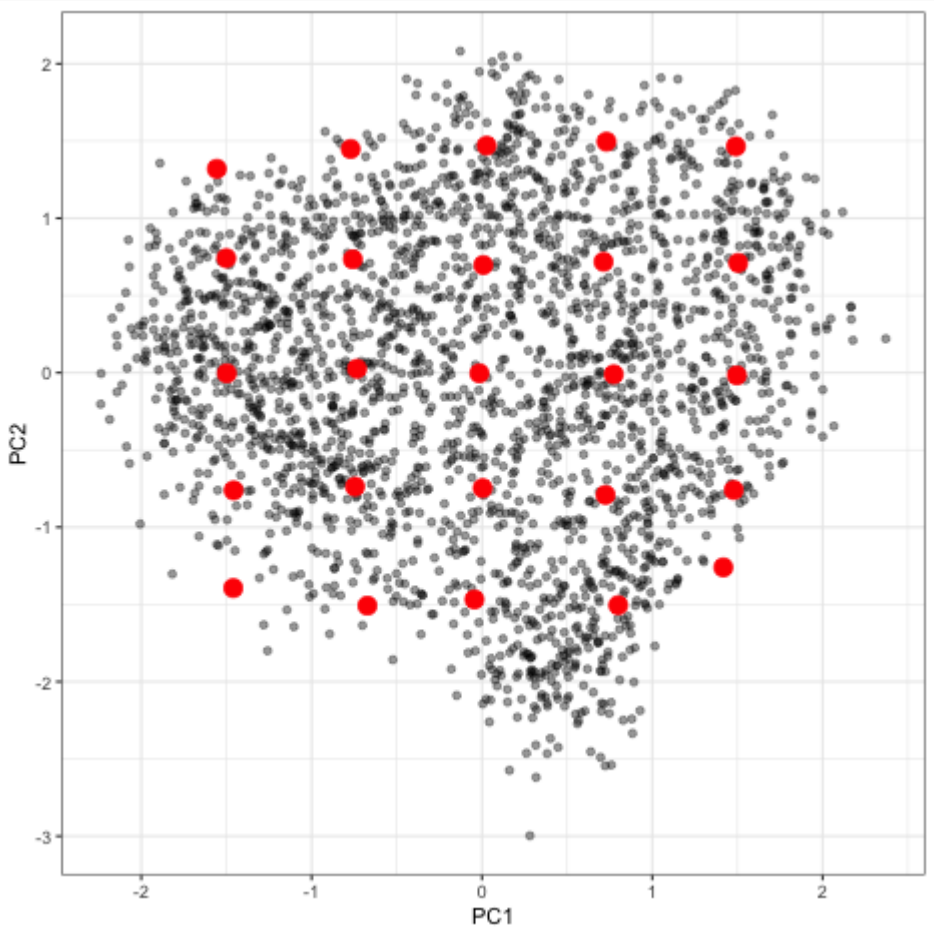
Used to:

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

Handwritten Letters

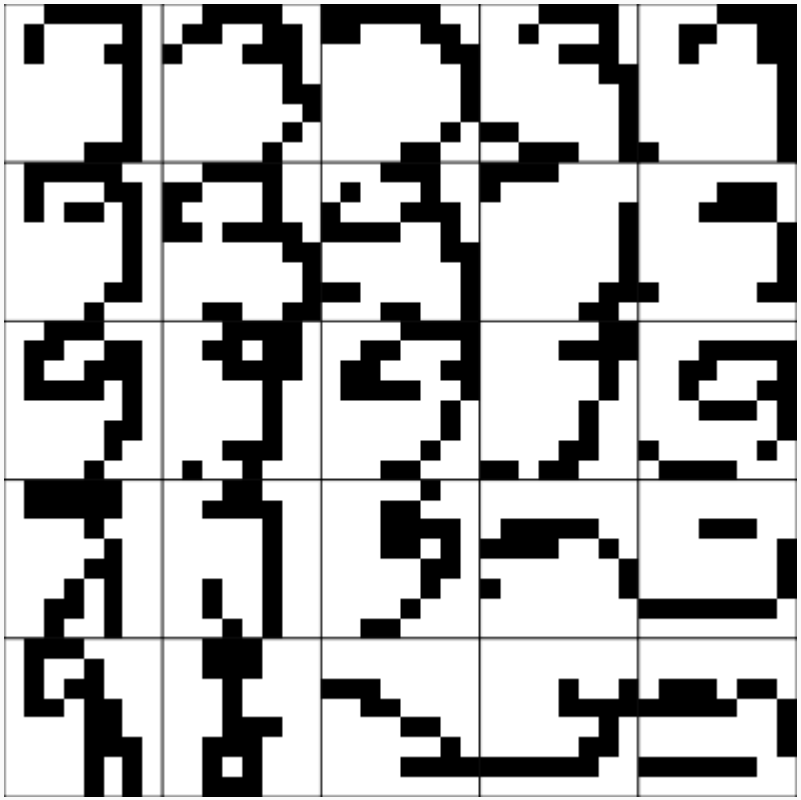
How much information is encoded in a 8 x 8 image of a handwritten letters?

Plotting the PCs



A scatterplot of observations

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



What do the PCs mean?

Difficult to tell due to

- Low resolution
- Lots of information to synthesize

But there is another way to learn about these first two PCs...

boardwork

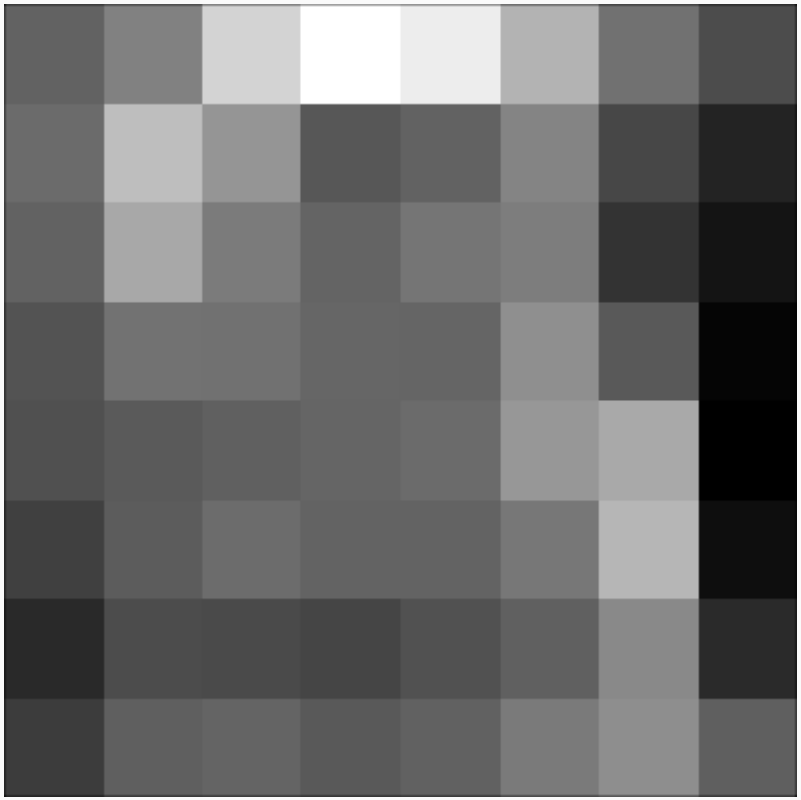
Activity 4, cont.

4.8) Extract the loadings (`rotation`) from your PCA object and use `plot_letter()` to plot your first several PCs. What variation do they seem to be capturing?

4.9) Do these loadings conform with variation seen across the 5x5 grid of letters across the first two PCs?

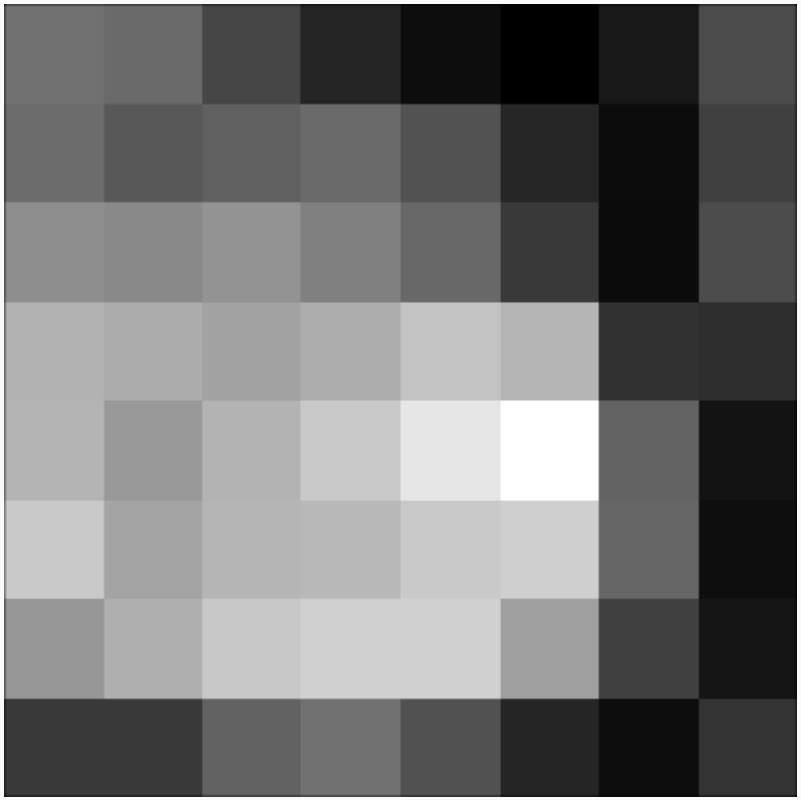
Loadings of PC1

```
plot_letter(pca1$rotation[, 1], hasletter = FALSE)
```

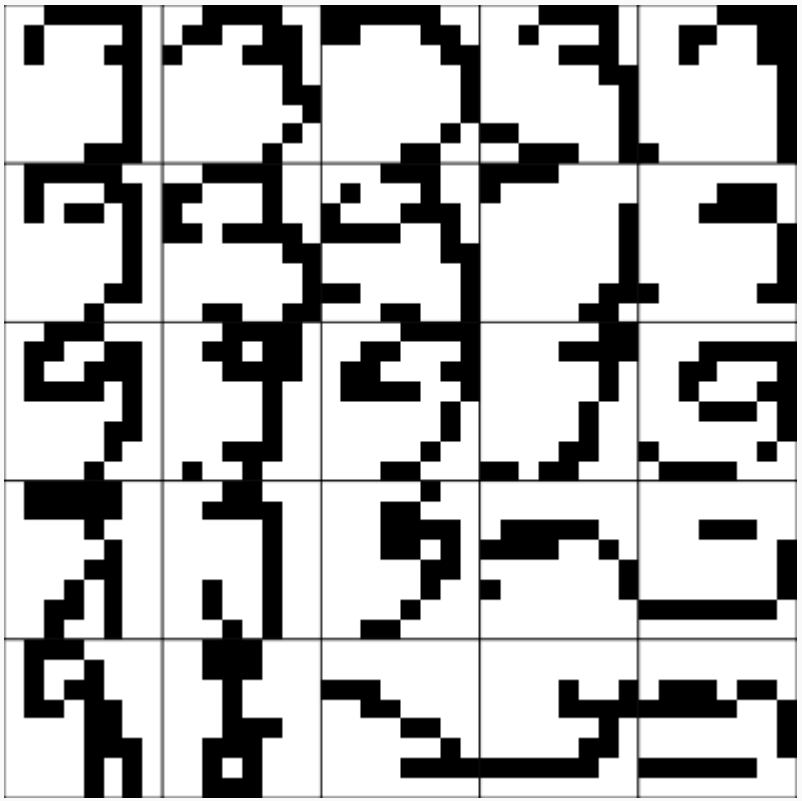


Loadings of PC2

```
plot_letter(pca1$rotation[, 2], hasletter = FALSE)
```



Did it work?



boardwork

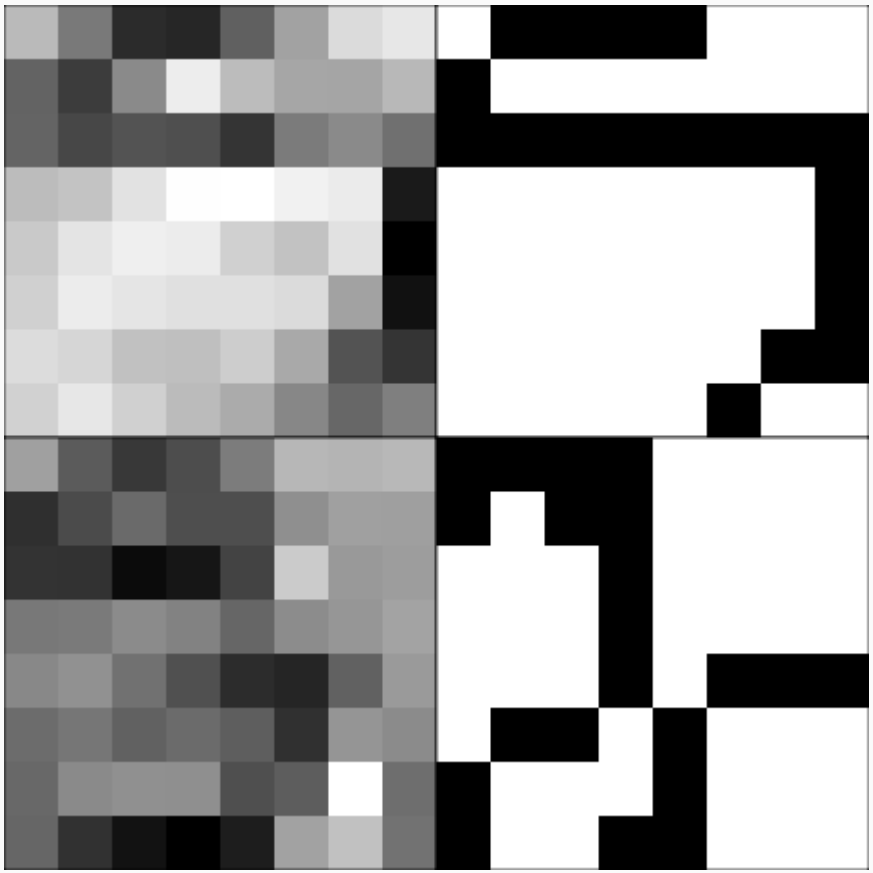
Activity 4, cont.

4.10) Use the first ten principle components to reconstruct (and visualize) the first two observations of the letter that you studied. For each observation, plot the true image and the reconstructed image on the same plot. How much information was lost?

G reconstructed

```
g_mean <- colMeans(g_data[ , -1])
phi <- pca1$rotation
z <- pca1$x
ncomp <- 10
x_star1 <- g_mean + z[1, 1:ncomp] %*% t(phi[, 1:ncomp])
x_star2 <- g_mean + z[2, 1:ncomp] %*% t(phi[, 1:ncomp])
```

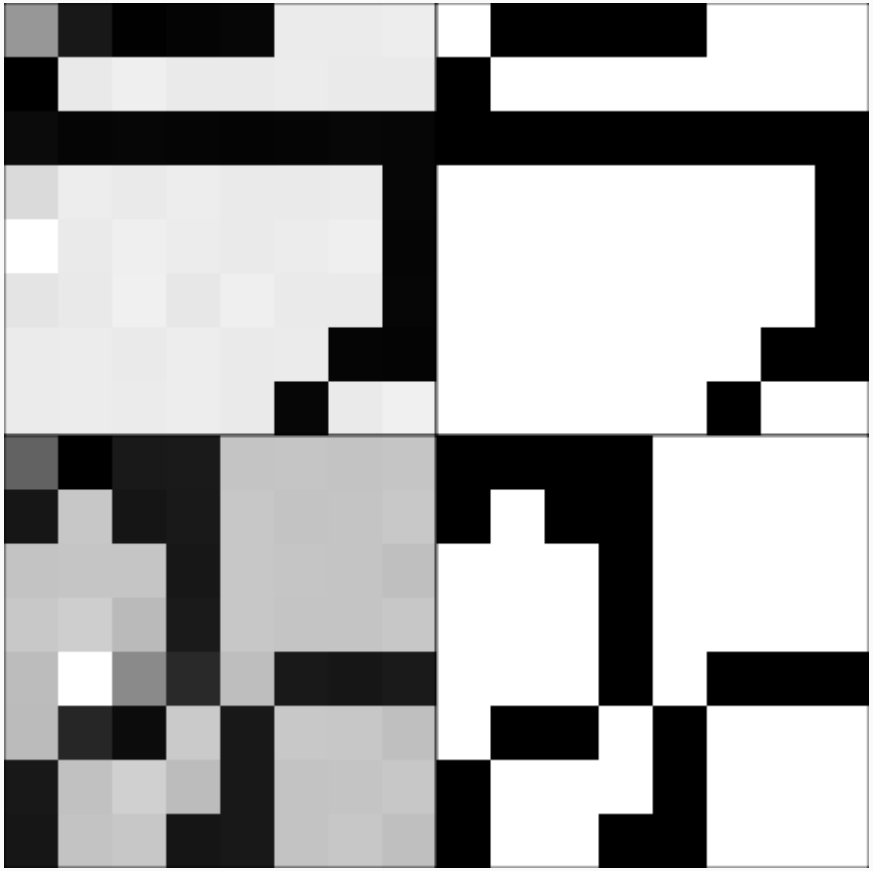
Reconstructed vs original



G almost fully reconstructed

```
g_mean <- colMeans(g_data[ , -1])
phi <- pca1$rotation
z <- pca1$x
ncomp <- 60
x_star1 <- g_mean + z[1, 1:ncomp] %*% t(phi[, 1:ncomp])
x_star2 <- g_mean + z[2, 1:ncomp] %*% t(phi[, 1:ncomp])
```

Almost fully reconstructed vs original



Activity 4, cont. (the last one!)

4.11) Choose a second letter (not g!) that you expect to have a very different scree plot than your first.

- Create a *scree plot*. How does the shape compare with your original?
- Create a *letter grid* and plots of the first two *loadings*. What do they appear to be encoding?
- *Reconstruct* the first observed letter based on 10 principle components. Is it more or less intact that your reconstruction of the other letter? What structure in the scree plot explains this difference (or similarity)?