

Principal components regression

Matthew S. Shotwell, Ph.D.

Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, TN, USA

January 27, 2020

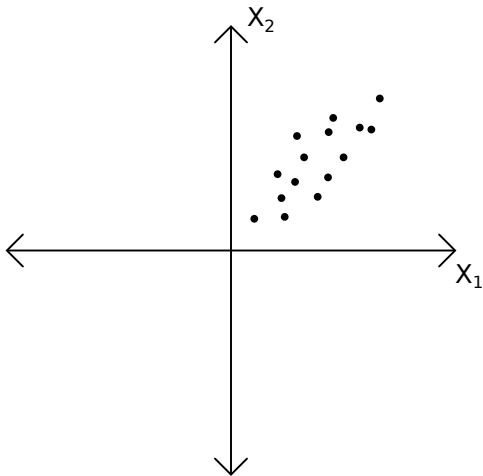
Overview

- ▶ Principle components regression involves creating new variables from existing ones (i.e. feature extraction)
- ▶ It is an unsupervised process – outcome data are not used
- ▶ The process consists of rotating the axes of X to better describe variability and minimize correlation between inputs
- ▶ If correlation was present, it may be possible to find a lower dimensional set of inputs that retain most of the information in the data
- ▶ Projecting points onto the eigenvectors of the estimated covariance matrix

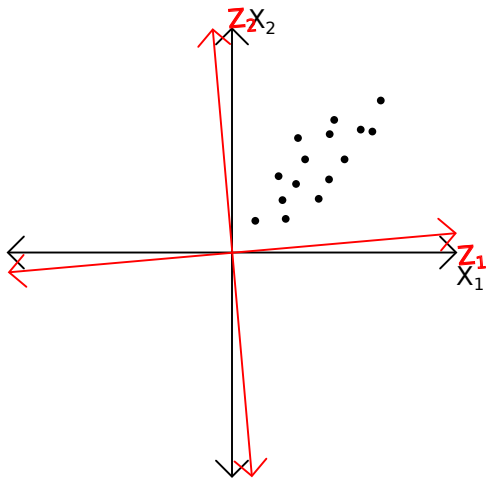
Principal Components Analysis

- ▶ to understand principal components regression, first need principal components analysis (PCA)
- ▶ suppose we have an $n \times p$ input matrix X
- ▶ all inputs must be numeric or dummy coded
- ▶ PCA transforms X into a new matrix Z with the same number of rows and columns
- ▶ columns of Z are called principal components (PCs)
- ▶ the new, transformed inputs (columns Z_1 , Z_2 , etc) are no longer correlated (they're "independent")
- ▶ variance of Z_1 is largest, then Z_2 , and so on
- ▶ variance of some Z may be very small or zero

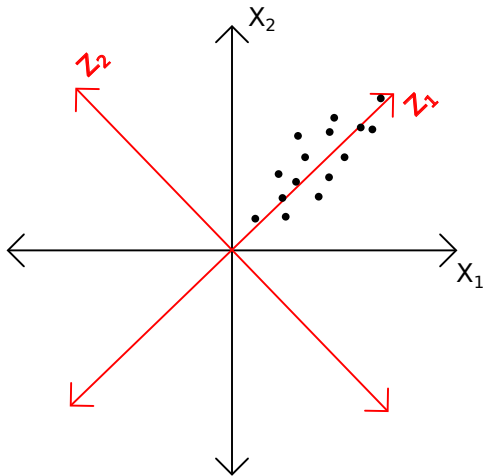
- ▶ two inputs substantially correlated



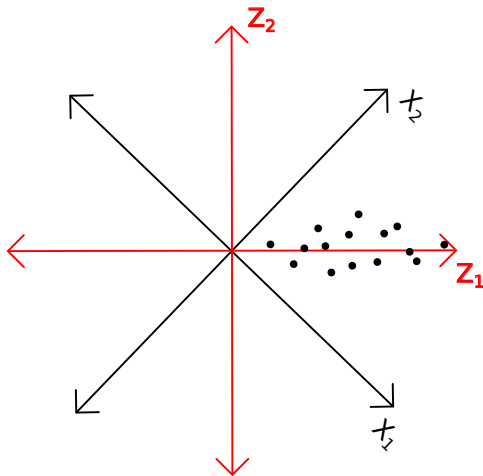
- PCA creates new inputs Z_1 and Z_2 by rotating the axes



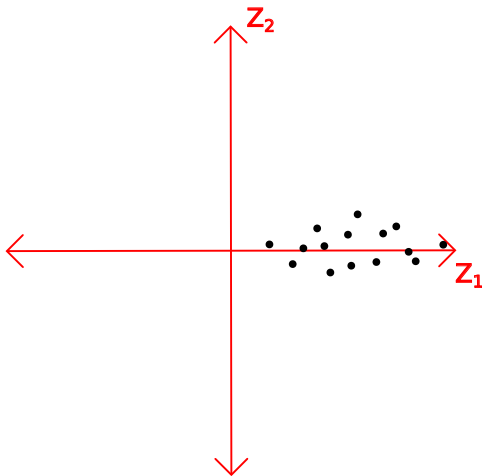
- such that new inputs Z_1 and Z_2 are not correlated



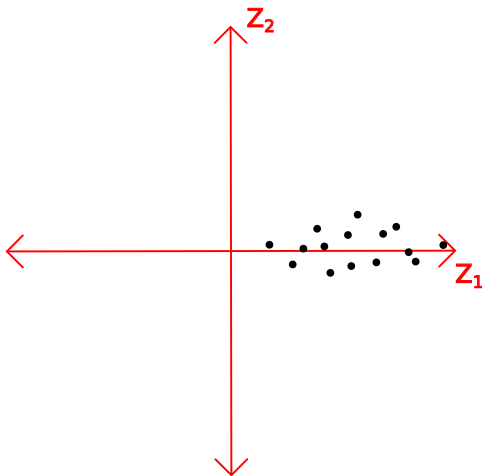
- rotate entire figure 45 degrees to view Z_1 and Z_2



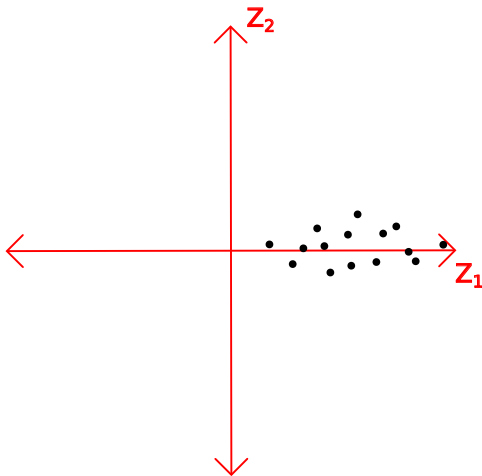
- drop the original axes



- no correlation between Z_1 and Z_2



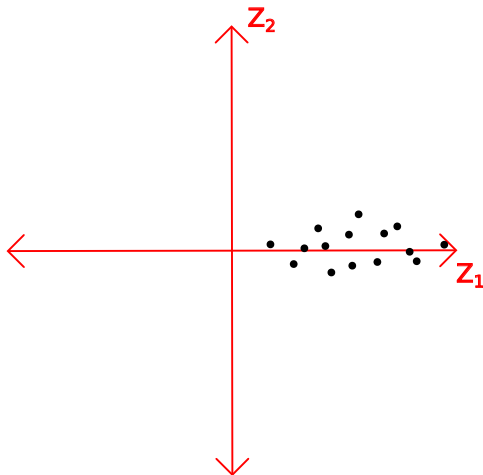
- variance of Z_1 greater than variance of Z_2



Principal components analysis

- ▶ transforming X to get Z **is** PCA
- ▶ if X has p columns, then Z will have p columns
- ▶ what we can do with Z makes PCA useful
- ▶ **dimension reduction**

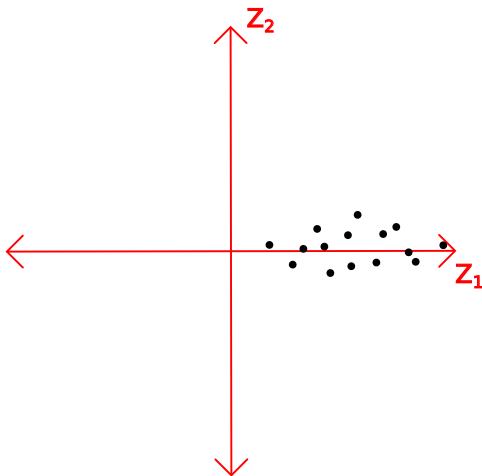
- ▶ most information in Z is captured by Z_1
- ▶ maybe we can simply ignore Z_2
- ▶ if so, the dimension of (transformed) input is reduced by 1



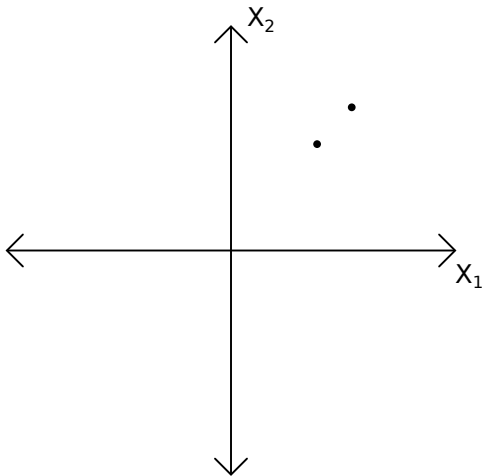
Principal components analysis

- ▶ ignoring some PCs generally causes loss of information
- ▶ exceptions:
 - ▶ if $n \leq p$, can drop $p - n + 1$ PCs without loss of info
 - ▶ if some inputs perfectly correlated, can drop some PCs without loss of info

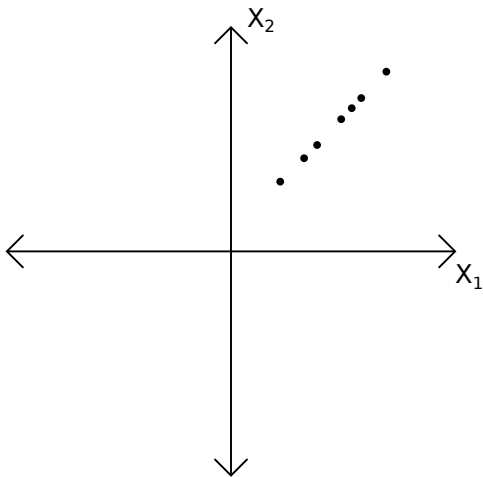
- ignoring Z_2 would cause loss of (a little) information



- ▶ when $n = p$, only $p - 1$ PCs needed; no info loss



- ▶ when X_1 and X_2 perfectly correlated, only 1 PC needed; no info loss



Principal components regression

- ▶ say X is a matrix of training inputs
- ▶ dimension reduction reduces the information in X
- ▶ less information means less flexible predictor based on X
- ▶ degree of dimension reduction (i.e., how many PCs ignored) affects bias-variance tradeoff
- ▶ principal components regression is simply linear regression using PCs as inputs, and after applying some dimension reduction
- ▶ number of PCs used is the tuning parameter

Principal components regression

- ▶ for some $0 \leq M \leq p$, use only first M PCs in regression
- ▶ $y = z_M \beta_M$
- ▶ where z_M is matrix of first M PCs
- ▶ fit β_M by minimizing training error
- ▶ tune M using testing error

Code example

pca-regression-example.R