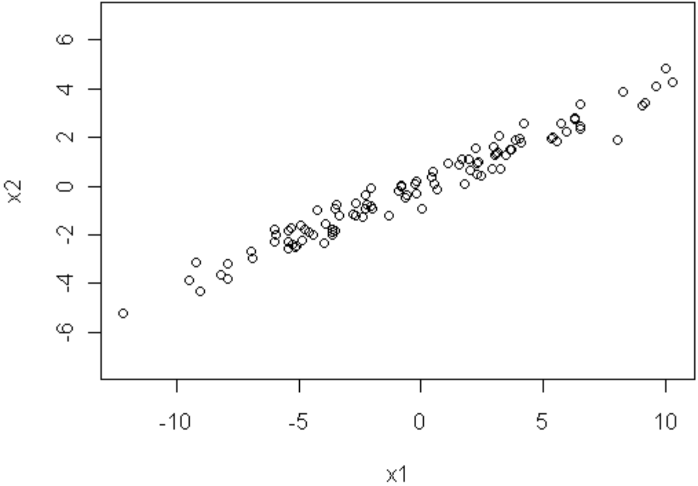
Math 5366 Notes

Principal Components

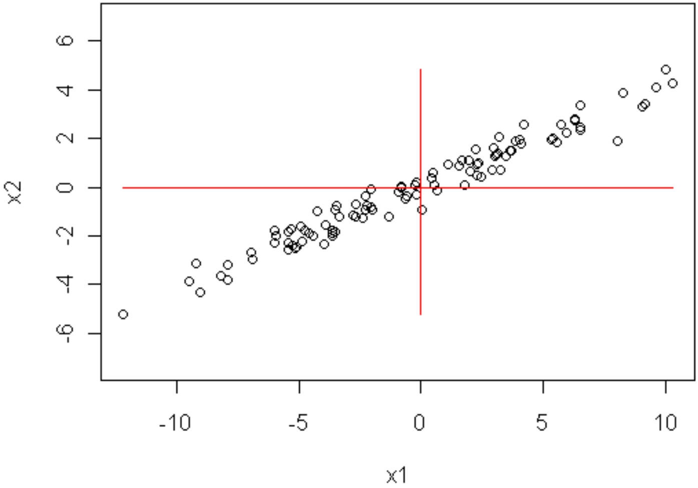
Jesse Crawford

Department of Mathematics
Tarleton State University

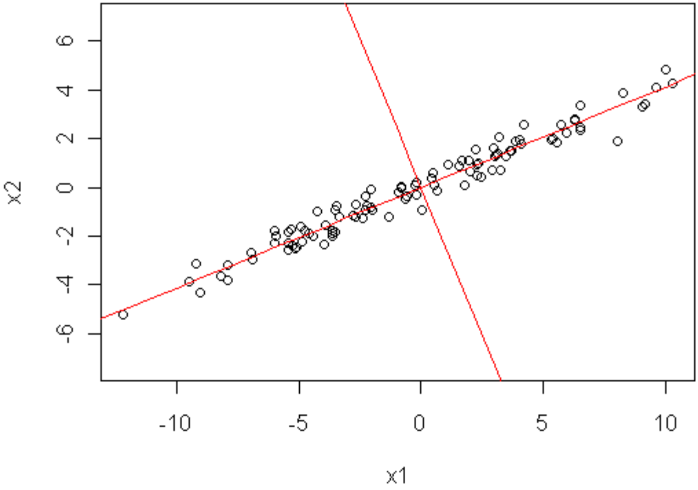
- Two variables X_1 and X_2
- Scatterplot:



Typical Coordinate System



Principal Components



Principal Components

Definition

- Consider a p-dimensional random vector X with covariance matrix Σ. Assume Σ is positive definite.
- Define

$$\lambda_1 = \max\{\mathrm{Var}(a'X) \mid a \in \mathbb{R}^p, a'a = 1\}.$$

- The vector a_1 where this maximum is attained is called the *first principal component*.
- Define

$$\lambda_2 = \max\{\mathrm{Var}(a'X) \mid a \in \mathbb{R}^p, a'a = 1, \mathrm{cov}(a'X, a_1'X) = 0\}.$$

- The vector a_2 where this maximum is attained is called the *second principal component*.

Principal Components (cont.)

Definition

- Define

$$\lambda_j = \max\{\mathrm{Var}(a'X)\ |a \in \mathbb{R}^p, a'a = 1,$$
$$\mathrm{cov}(a'X, a_k'X) = 0, k = 1, \ldots, j-1\}.$$

- The vector a_j where this maximum is attained is called the *jth principal component*.
- There are p principle components $a_1, \ldots, a_p$, and $\lambda_j = \mathrm{Var}(a_j'X)$ for each j.

Converting to Linear Algebra

$$\mathrm{cov}(a'X, b'X) = a'\Sigma b$$
$$\mathrm{Var}(a'X) = a'\Sigma a$$

$$\lambda_j = \max\{\mathrm{Var}(a'X) \,|a \in \mathbb{R}^p, a'a = 1,$$
$$\mathrm{cov}(a'X, a_k'X) = 0, k = 1, \ldots, j-1\}.$$

$$\lambda_j = \max\{a'\Sigma a \,|a \in \mathbb{R}^p, a'a = 1,$$
$$a'\Sigma a_k = 0, k = 1, \ldots, j-1\}.$$

Relation to Eigenvectors and Eigenvalues

Theorem

- *Let* $\lambda_1 \geq \cdots \geq \lambda_p > 0$ *be the eigenvalues of* Σ*.*
- *Let* $a_1, \ldots, a_p$ *be the corresponding orthonormal eigenvectors.*
- *Then the principal components are* $a_1, \ldots, a_p$*, and* $\lambda_j = Var(a_j' X)$ *for each* j*.*

Spectral Theorem: Every real, symmetric matrix has an orthonormal eigenbasis.

$$(a_1 \cdots a_p)' \Sigma (a_1 \cdots a_p) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

Implementation in R

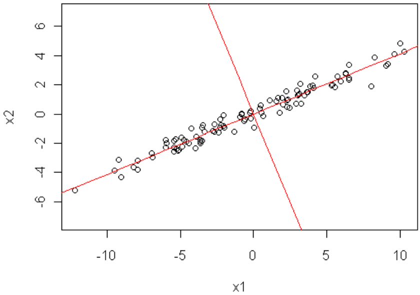
```
Console ~/
> X=cbind(x1,x2)
> S=cov(X)
> S
         x1        x2
x1 25.06959 10.262922
x2 10.26292  4.375917
> eigen(S)
$values
[1] 29.2961784  0.1493335

$vectors
           [,1]       [,2]
[1,] -0.9246567  0.3808018
[2,] -0.3808018 -0.9246567

> v1=eigen(S)$vectors[,1]
> v2=eigen(S)$vectors[,2]
> v1
[1] -0.9246567 -0.3808018
> v2
[1]  0.3808018 -0.9246567
>
```

R Provides Orthonormal Eigenvectors

```
Console ~/
> t(v1)%*%v1
     [,1]
[1,]    1
> t(v2)%*%v2
     [,1]
[1,]    1
> t(v1)%*%v2
              [,1]
[1,] -2.355429e-17
> 
```

Scatterplot with Principal Components

```
plot(x1,x2,asp=1)
lines(xrange,v1[2]/v1[1]*xrange,col='red')
lines(xrange,v2[2]/v2[1]*xrange,col='red')
```



Non-centered data will require intercept terms.

Dimension Reduction

- Last example: $\lambda_1 = 29.3$ and $\lambda_2 = 0.15$.

$$\text{Total Variance} = \text{trace}(S) = \lambda_1 + \lambda_2 = 29.5$$

Principal Component	% of Variance	Cumulative % of Var
a_1	99.5%	99.5%
a_2	0.5%	100%

- Rule of thumb: We can reduce the number of principal components to a set accounting for 90% or more of the total variance.

Applications

- It is often better to start with a correlation matrix instead of a covariance matrix so that each variable has comparable variability (R command: `cor(X)`).
- PCA can be used to reduce the dimension of a data set.
- It can be used to identify size and shape factors for biological organisms or other objects.
- Can be used to reduce variables in a regression model to avoid multicollinearity.
- Warning: principal components explaining over 90% of total variance may not be the best set of predictors, so one should remove the minimal number of principal components required to avoid multicollinearity.