Statistical Learning. Dimension reduction

Jose Ameijeiras Alonso
Departamento de Estatística e Investigación Operativa (USC)

Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data

Introduction

- In the session about regression we discussed some ways to deal with a large set of correlated variables $X_1, \ldots, X_p$
  - Subset Selection. Methods for selecting a subset of the p predictors. We then fit a model using least squares on the reduced set of variables
    - Best subset selection, stepwise selection, AIC, BIC, adjusted R^2, cross-validation methods,...
  - Shrinkage (regularization). This approach involves fitting a model involving all p predictors. But the estimated coefficients are shrunken towards zero relative to the least squares estimates.
    - Ridge regression, Lasso,...
  - Dimension Reduction. Methods for projecting the p predictors into a lower-dimensional subspace. Then, the projections are used as predictors to fit a linear regression model by least squares.

Dimension reduction

- Suppose that we have a random vector $X = (X_1, \ldots, X_p)^t$
- The components of the random vector are usually dependent and contain redundant information
- It can be useful, for example, to look for linear combinations of the original variables keeping as much of the original information as possible
- Principal component analysis (PCA) is possibly the dimension reduction technique most widely used in practice

Dimension reduction

- In practice we have n observations on the set of p features, $X_1, \ldots, X_p$.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- The objective is to reduce the dimensionality of a data set
  - interpretation
  - avoid overfitting
- Each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting
- We would like to find a low-dimensional representation of the data that captures as much of the information as possible
- Principal component analysis (PCA) finds a low-dimensional representation of a data set that contains as much as possible of the variation

Dimension reduction

- Suppose that we have a random vector $X = (X_1, \ldots, X_p)^t$
- The population mean vector is $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$$

- The population covariance matrix is Σ.

$$\Sigma = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \ldots & \mathrm{Var}(X_p) \end{pmatrix}$$

Dimension reduction

- Suppose that we have n observations on the set of p features, $X_1, \ldots, X_p$.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- We can estimate $\boldsymbol{\mu}$ with

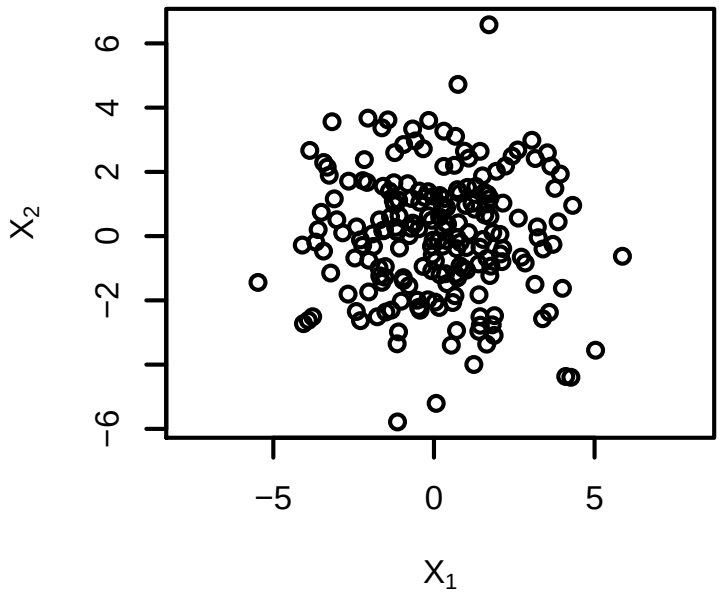
$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

- We can estimate Σ with the sample covariance matrix

$$S = \begin{pmatrix} s_1^2 & s_{12} & \ldots & s_{1d} \\ s_{21} & s_2^2 & \ldots & s_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ s_{d1} & s_{d2} & \ldots & s_d^2 \end{pmatrix}$$

Dimension reduction

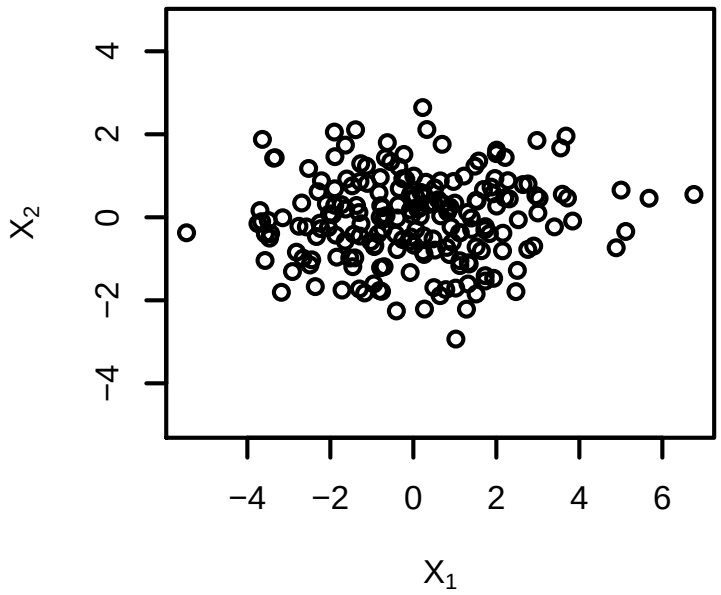
- $X = (X_1, X_2)^t$ normal with $\boldsymbol{\mu} = (0, 0)^t$ and $\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$




```
> colMeans(x)

## [1] 0.08422010 0.01200716

> cov(x)

##             [,1]       [,2]
## [1,]  4.2168611 -0.1404574
## [2,] -0.1404574  3.7607762
```

Dimension reduction

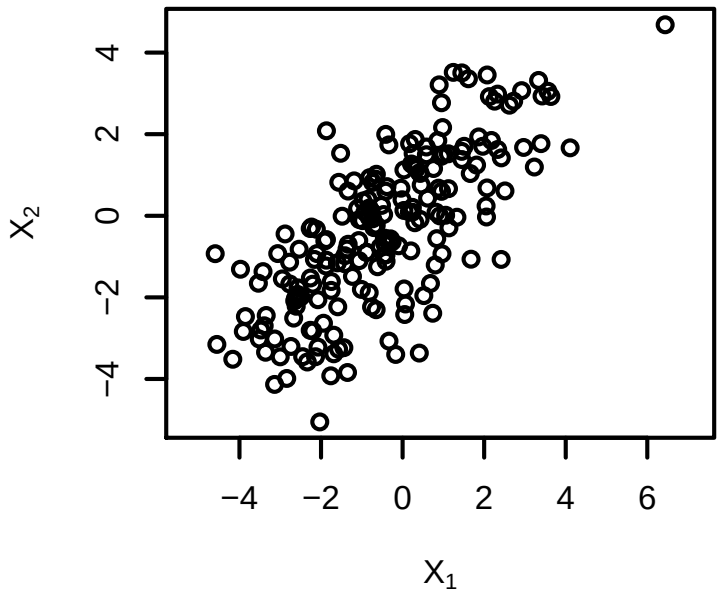
- $X = (X_1, X_2)^t$ normal with $\boldsymbol{\mu} = (0, 0)^t$ and $\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$




```
> colMeans(x)

## [1]  0.02743423 -0.04175770

> cov(x)

##             [,1]      [,2]
## [1,] 4.1284825 0.1236849
## [2,] 0.1236849 1.0401646
```

Dimension reduction

- $X = (X_1, X_2)^t$ normal with $\boldsymbol{\mu} = (0, 0)^t$ and $\Sigma = \begin{pmatrix} 4 & 3.2 \\ 3.2 & 4 \end{pmatrix}$




```
> colMeans(x)

## [1] -0.4521521 -0.3521426

> cov(x)

##          [,1]     [,2]
## [1,] 3.647091 2.699721
## [2,] 2.699721 3.672613
```

Principal component analysis

- PCA finds a low-dimensional representation of a data set that contains as much as possible of the variation
- PCA is concerned with explaining the variance-covariance structure of $X = (X_1, \ldots, X_p)^t$ through a smaller number of uncorrelated variables (the principal components)
- The principal components are the uncorrelated linear combinations of the features $X_1, \ldots, X_p$ whose variances are as large as possible.
- Recall that a linear combination of the features $X_1, \ldots, X_p$ can be written as

$$Z = a_1 X_1 + \ldots + a_p X_p = a^t X$$

  where $a = (a_1, \ldots, a_p)^t$.
- If $X = (X_1, \ldots, X_p)^t$ is a random vector with covariance matrix Σ, then for a given linear combination $Z = a^t X$, we have:

$$\text{Var}(Z) = a^t \Sigma a$$

- If $Y = b^t X$,

$$\text{Cov}(Z, Y) = a^t \Sigma b$$

Principal component analysis

- The principal components are the uncorrelated linear combinations of the features $X_1, \ldots, X_p$ whose variances are as large as possible.
- The first principal component of $X = (X_1, \ldots, X_p)^t$ is the linear combination

$$Z_1 = \phi_{11} X_1 + \ldots + \phi_{1p} X_p = \phi_1^t X$$

  maximizing $\text{Var}(Z_1) = \phi_1^t \Sigma \phi_1$ subject to $\|\phi_1\| = 1$
- The second principal component of $X = (X_1, \ldots, X_p)^t$ is the linear combination

$$Z_2 = \phi_{21} X_1 + \ldots + \phi_{2p} X_p = \phi_2^t X$$

  maximizing $\text{Var}(Z_2) = \phi_2^t \Sigma \phi_2$ subject to $\|\phi_2\| = 1$ and $\text{Cov}(Z_1, Z_2) = 0$
- $\ldots$
- The k-th principal component of $X = (X_1, \ldots, X_p)^t$ is the linear combination

$$Z_k = \phi_{k1} X_1 + \ldots + \phi_{kp} X_p = \phi_k^t X$$

  maximizing $\text{Var}(Z_k) = \phi_k^t \Sigma \phi_k$ subject to $\|\phi_k\| = 1$ and $\text{Cov}(Z_k, Z_j) = 0$ for $j < k$.

Principal component analysis

- Let $X = (X_1, \ldots, X_p)^t$ be a random vector with covariance matrix Σ
- let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ denote the eigenvalues of the covariance matrix Σ
- Let the vectors $e_1, \ldots, e_p$ denote the corresponding eigenvectors with $\|e_i\| = 1$ for $i = 1, \ldots, p$.
- The variance for the k-th principal component is equal to the k-th eigenvalue and the elements of e_k will be the coefficients of the k-th principal component
- In particular,
  - the first principal component direction is the eigenvector associated with the largest eigenvalue λ_1
  - $Z_1 = e_1^t X$ has the largest variance amongst all normalized linear combinations of X

Principal component analysis

- Interpretation of the components in terms of the proportion of the full variation explained by each component
- The proportion of variation explained by the k-th principal component is

$$\frac{\lambda_k}{\lambda_1 + \ldots + \lambda_p}$$

- The proportion of variation explained by the first k principal components is

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p}$$

Principal component analysis

- Suppose that we have n observations on the set of p features, $X_1, \ldots, X_p$.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- The first sample principal component is the linear combination

$$Z_1 = \phi_{11} X_1 + \ldots + \phi_{1p} X_p = \phi_1^t X$$

  maximizing $\phi_1^t S \phi_1$ subject to $\|\phi_1\| = 1$

- ...
- The k-th sample principal component is the linear combination

$$Z_k = \phi_{k1} X_1 + \ldots + \phi_{kp} X_p = \phi_k^t X$$

  maximizing $\phi_k^t S \phi_k$ subject to $\|\phi_k\| = 1$ and $\phi_j^t S \phi_k = 0$ for $j < k$.

Principal component analysis

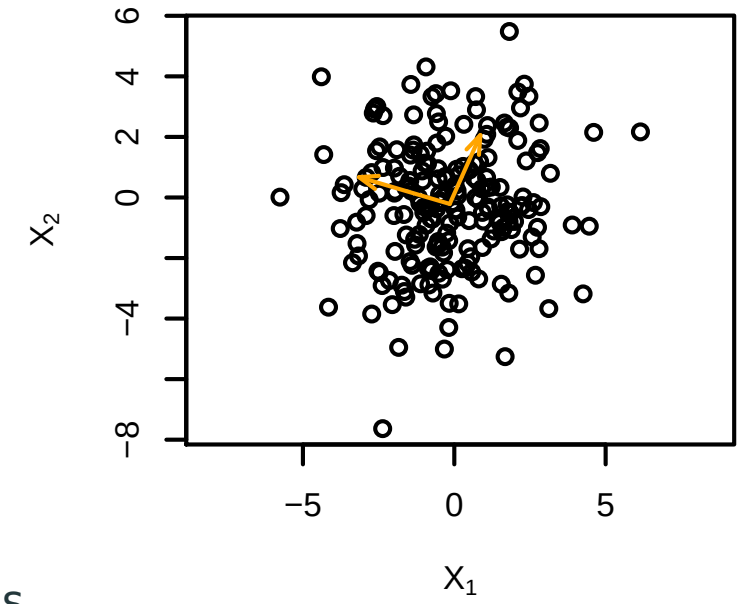
- Suppose that we have n observations on the set of p features, $X_1, \ldots, X_p$.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$ denote the eigenvalues of the sample covariance matrix S
- Let the vectors $\hat{e}_1, \ldots, \hat{e}_p$ denote the corresponding eigenvectors with $\|\hat{e}_i\| = 1$ for $i = 1, \ldots, p$.
- The variance for the k-th sample principal component is equal to the k-th eigenvalue and the elements of $\hat{e}_k$ will be the coefficients of the k-th sample principal component

Principal component analysis

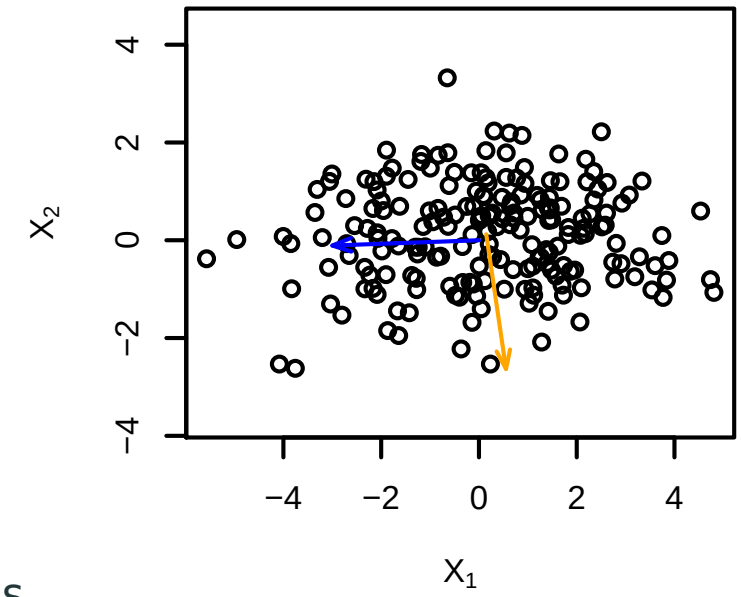
- $X = (X_1, X_2)^t$ normal with $\boldsymbol{\mu} = (0, 0)^t$ and $\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$




```
> eigen(S)$values

## [1] 4.359908 3.518505

> eigen(S)$vectors

##            [,1]       [,2]
## [1,] 0.4368357 -0.8995413
## [2,] 0.8995413  0.4368357
```

Principal component analysis

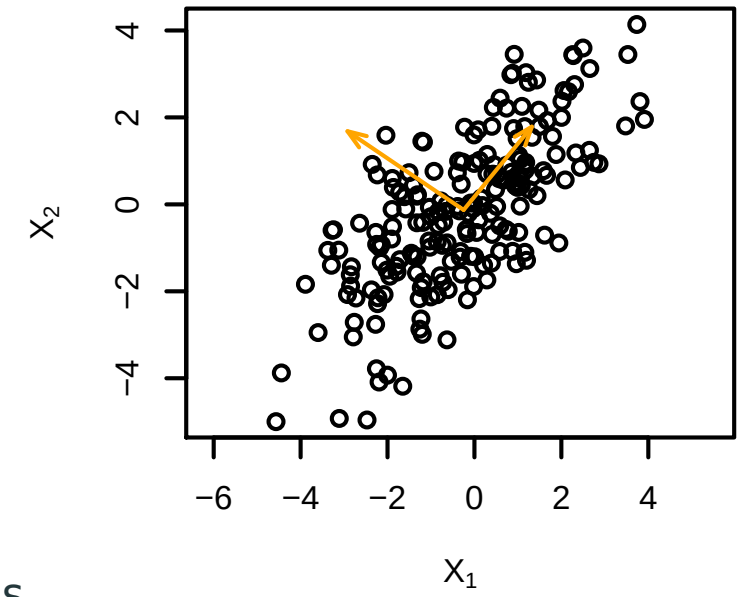
- $X = (X_1, X_2)^t$ normal with $\boldsymbol{\mu} = (0, 0)^t$ and $\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$




```
> eigen(S)$values

## [1] 3.909145 1.058256

> eigen(S)$vectors

##              [,1]        [,2]
## [1,] -0.99935801  0.03582688
## [2,] -0.03582688 -0.99935801
```

Principal component analysis

- $X = (X_1, X_2)^t$ normal with $\boldsymbol{\mu} = (0, 0)^t$ and $\Sigma = \begin{pmatrix} 4 & 3.2 \\ 3.2 & 4 \end{pmatrix}$




```
> eigen(S)$values

## [1] 5.1538187 0.8882358

> eigen(S)$vectors

##            [,1]       [,2]
## [1,] 0.6893498 -0.7244286
## [2,] 0.7244286  0.6893498
```

Principal component analysis

- The results of PCA depend on the scaling of the data
- Variables with the highest sample variances will tend to be emphasized in the first few principal components
- Principal component analysis using the covariance matrix is appropriate when the variables are measured in comparable units of measurement
- Principal component analysis using the correlation matrix is appropriate when the variables are measured in very different units of measurement

Principal component analysis

- Suppose that we have a random vector $X = (X_1, \ldots, X_p)^t$
- The correlation between variables X_i and X_j is

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

  where σ_{ij} denotes de covariance between X_i and X_j and σ_i and σ_j denote the standard deviation of X_i and X_j, respectively
- Then the correlation matrix is

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1p} \\ \rho_{21} & 1 & \ldots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \ldots & 1 \end{pmatrix}.$$

Principal component analysis

- Suppose that we have n observations on the set of p features, $X_1, \ldots, X_p$.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- We can estimate the correlation matrix with the sample correlation matrix

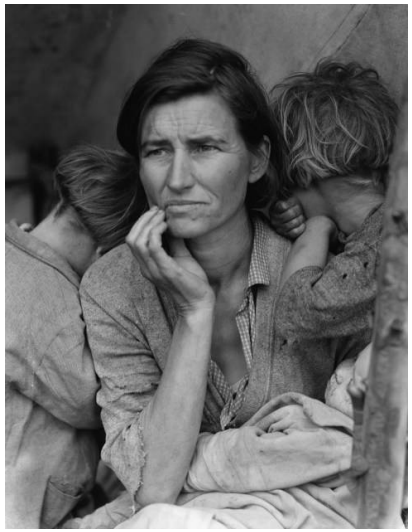
$$R = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix},$$

where

$$r_{jk} = \frac{s_{jk}}{s_j s_k}.$$

Principal component analysis

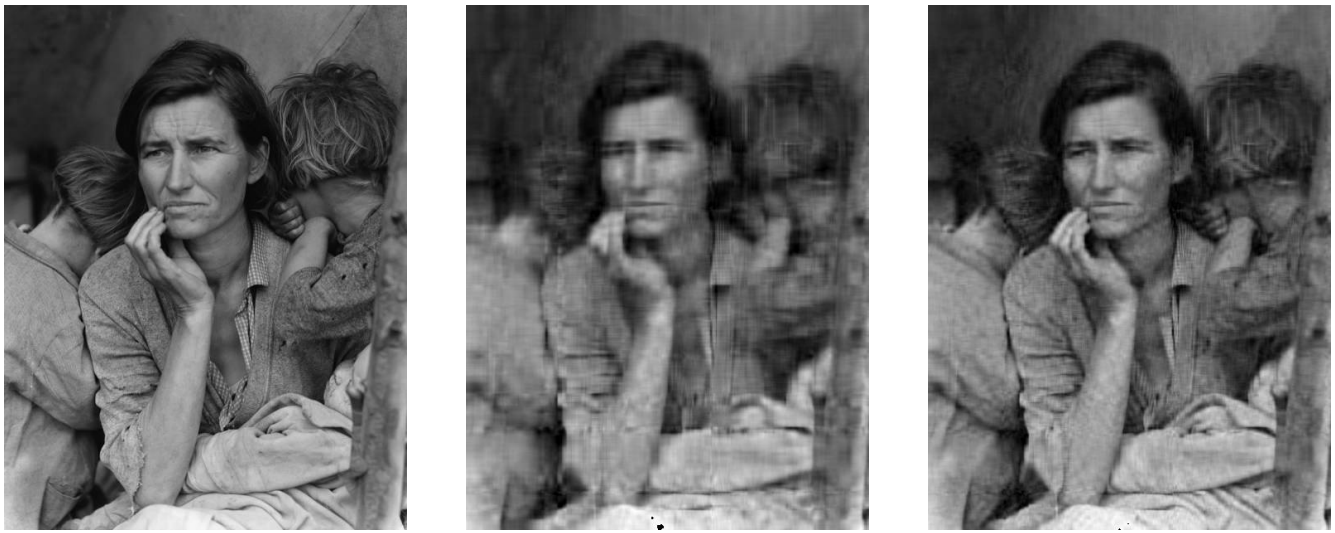
- Example: Principal Component Analysis applied to digital image compression (The Migrant Mother by Dorothea Lange)



```
> library(jpeg)
> imOrig <- readJPEG("lange.jpg")
> dim(imOrig)

## [1] 633 487
```

Principal component analysis

- Example: Principal Component Analysis applied to digital image compression (The Migrant Mother by Dorothea Lange)



Original image (633×487) and image reconstruction from $k = 20$ (middle) and $k = 40$ (left) principal components. The proportion of variation explained by the first 20 principal components is 93.77%. The proportion of variation explained by the first 40 principal components is 97.09%.

Principal component regression

- Principal component analysis can be used as a dimension reduction technique for regression
- Suppose that we observe a quantitative response Y and preditor variables $X = (X_1, \ldots, X_p)$
- The principal components regression (PCR) approach involves constructing the first k principal components, $Z_1, \ldots, Z_k$, and then using these components as the predictors in a linear regression model that is fit using least squares.

$$Y = \beta_0 + \beta_1 Z_1 + \ldots + \beta_k Z_k + \epsilon$$

- We assume that the directions in which $X_1, \ldots, X_p$ show the most variation are the directions that are associated with Y
- Estimating $k \ll p$ coefficients can mitigate overfitting

Principal component regression

- Example: data on samples of finely chopped pure meat
- 215 samples were measured
- For each sample, the fat content was measured along with a 100 channel spectrum of absorbances

```
> library(faraway)

> data(meatspec)
```

```
      V1      V2      V3      V4   ...     V100  fat
1 2.61776 2.61814 2.61859 2.61912   ...  2.81920 22.5
2 2.83454 2.83871 2.84283 2.84705   ...  3.17942 40.1
3 2.58284 2.58458 2.58629 2.58808   ...  2.54816  8.4
```

Principal component regression

- Example: data on samples of finely chopped pure meat
- We partition the data into a training sample (172 observations) and a test sample (43 observations)
- We fit a linear regression model with $p = 100$ predictors

```
> fit.lm <- lm(fat ~ ., data = train)
```

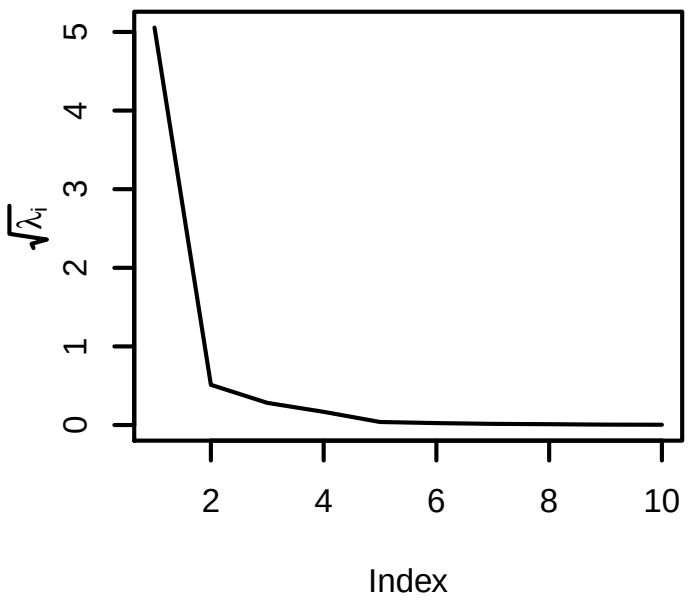
- We compute the MSE in the training sample and the MSE in the test sample

```
> MSEtrain

## [1] 0.4765372

> MSEtest

## [1] 14.54659
```

Principal component regression

- Example: data on samples of finely chopped pure meat
- We partition the data into a training sample (172 observations) and a test sample (43 observations)
- Now we compute the PCA on the training sample predictors:

```
> trainx <- train[, 1:p]
> pca <- princomp(trainx)
```



Principal component regression

- Example: data on samples of finely chopped pure meat
- We partition the data into a training sample (172 observations) and a test sample (43 observations)
- We use the first four PCs to predict the response:

```
> k <- 4
> fit.pcr <- lm(fat ~ ., data = pcax)
```

- We compute the MSE in the training sample and the MSE in the test sample

```
> MSEtrain

## [1] 16.52215

> MSEtest

## [1] 20.55699
```

Principal component regression

- Example: data on samples of finely chopped pure meat
- We partition the data into a training sample (172 observations) and a test sample (43 observations)
- We use the first 20 PCs to predict the response:

```
> k <- 20
> fit.pcr <- lm(fat ~ ., data = pcax)
```

- We compute the MSE in the training sample and the MSE in the test sample

```
> MSEtrain

## [1] 3.93473

> MSEtest

## [1] 5.205608
```