Dimension reduction

Faculty of Information Technology, Monash University, Australia

FIT5149 week 11

Outline

High-Dimensional Data

MONASH University

- High-dimensional: data sets contains more features than observations, i.e., $p \gg n$
  - Genetic data: If we consider genes as variable, this means that our observations are in a space with thousands of dimensions.
    - Predict blood pressure with *single nucleotide polymorphisms* (SNPs)

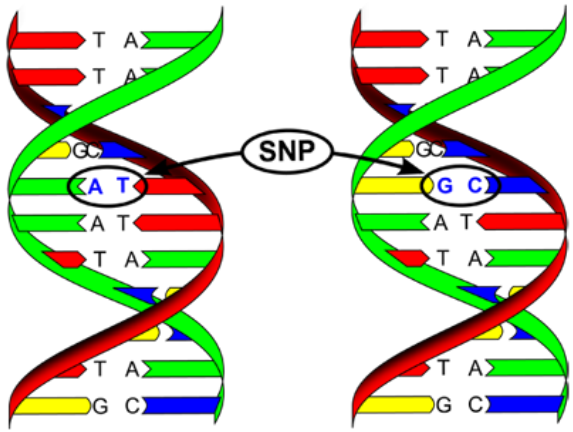


Figure: The picture is from http://www.viagenefertility.com/photos/snp.png

High-Dimensional Data

- High-dimensional: data sets contains more features than observations, i.e., $p \gg n$
  - The "bag of words" model:

Figure: The picture was generated by wordclouds.com

The Curse of Dimensionality: Overfitting

- When p is as large as, or larger than n, least squares cannot be perform.
  - Least squares will yield a set of coefficients that give a perfect fit to the data, such that the residuals are zero.

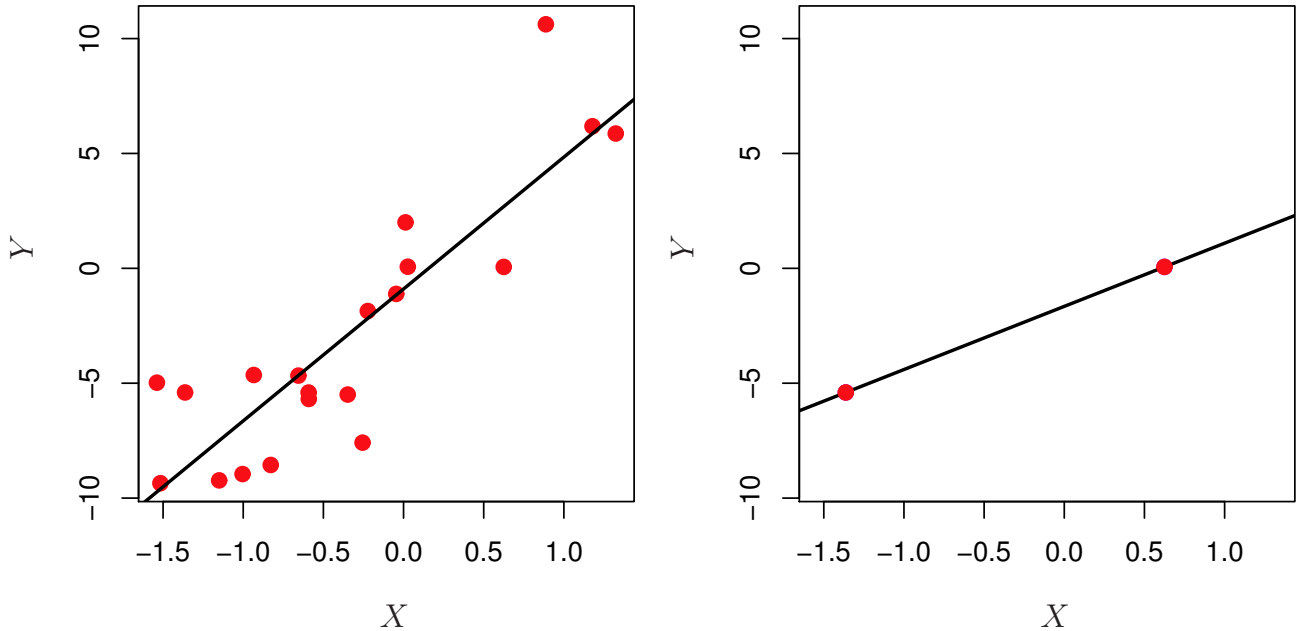


Figure: Left: Least squares regression in the low-dimensional setting. Right: Least squares regression with $n = 2$ observations and two parameters to estimate.

- Problem: When $p > n$ or $p \approx n$, least squares regression is too flexible and overfits the data.

The Curse of Dimensionality: Overfitting

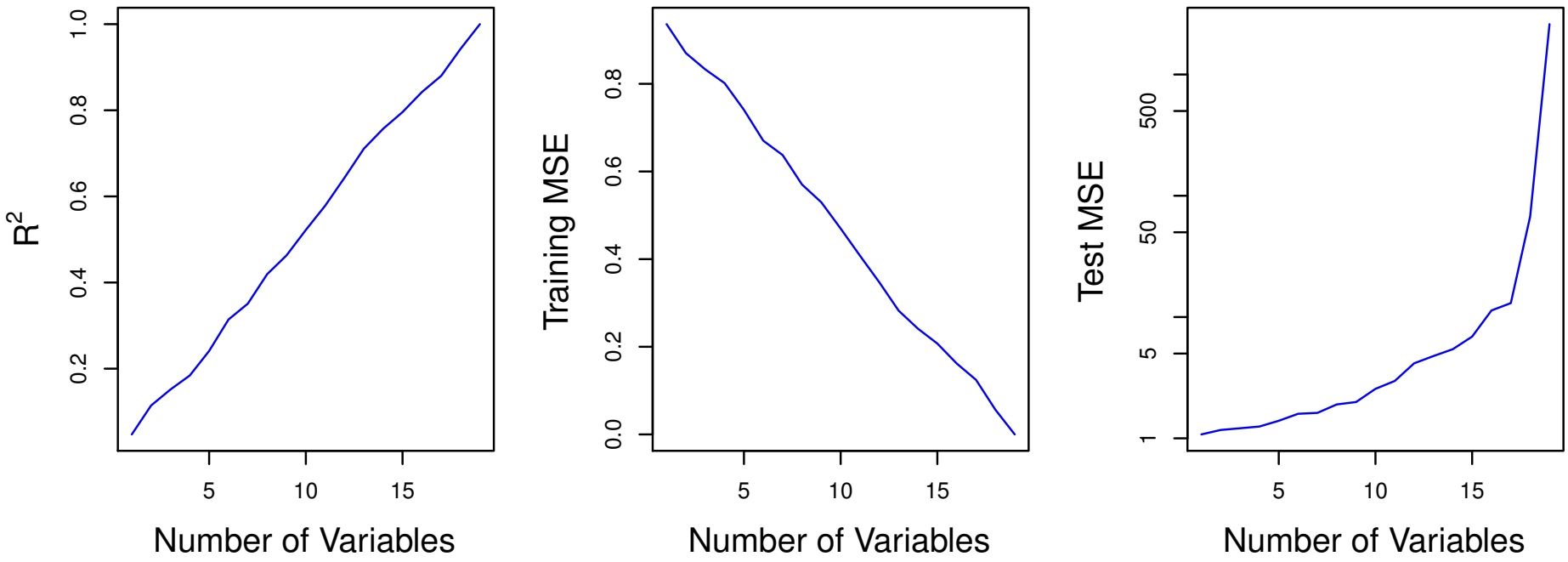
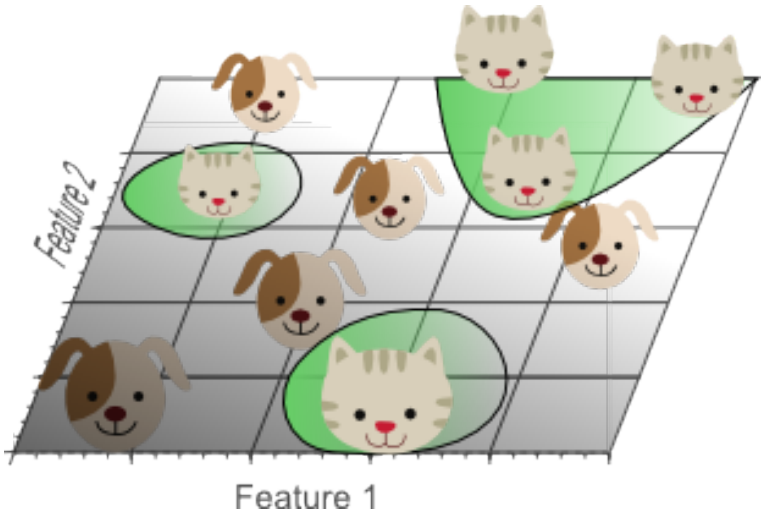
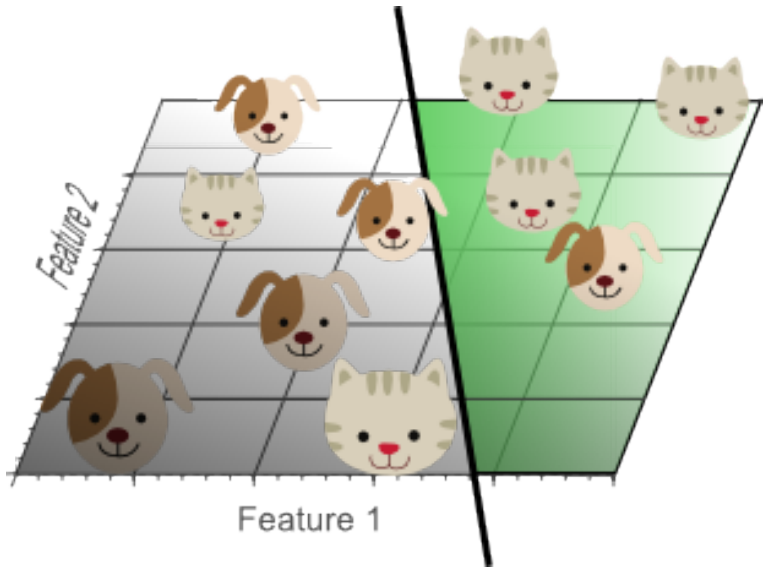


Figure: On a simulated example with $n = 20$ training observations, features that are completely unrelated to the outcome are added to the model.

The Curse of Dimensionality: Overfitting[1]



- Classifier trained using 3 features
- The decision hyperplane is projected onto a 2-dimensional space.
  - Corresponds to use a complicated non-linear classifier in the lower dimensional feature space



- Generalises much better to unseen data because it did not learn specific exceptions that were only in our training data by coincidence.

[1] Figures used in this slides are from http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

The Curse of Dimensionality: Dimensionality V.S. Classification Performance[2]

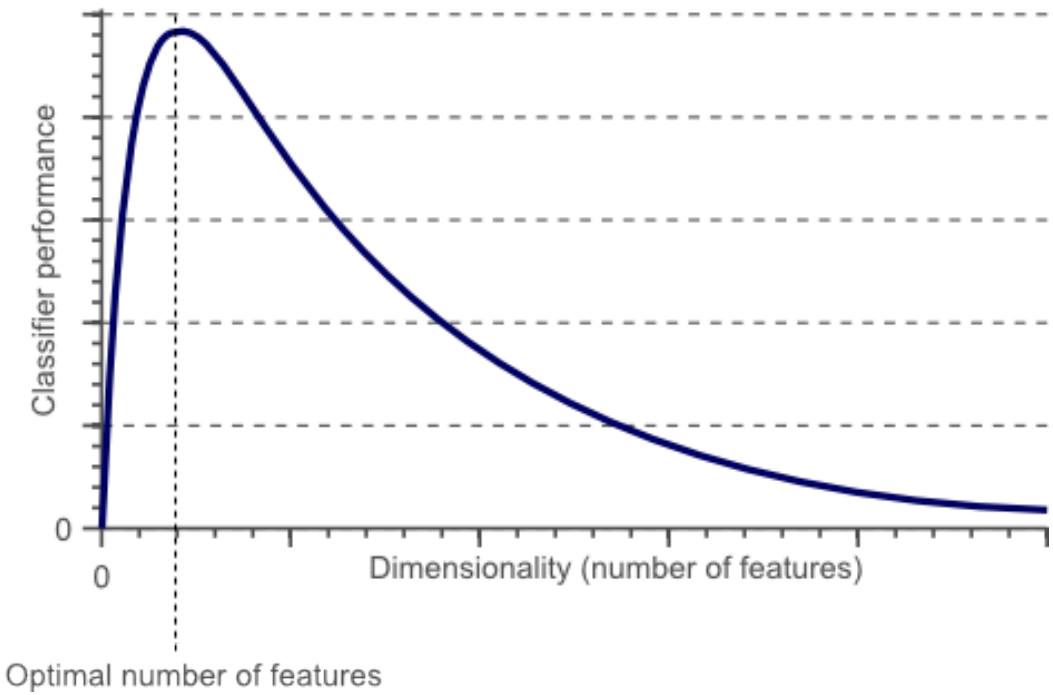


Figure: Fix the number of observations, increase the number of features used in training a classifier.

[2]Figure used in this slides is from http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

Regression in High Dimensions

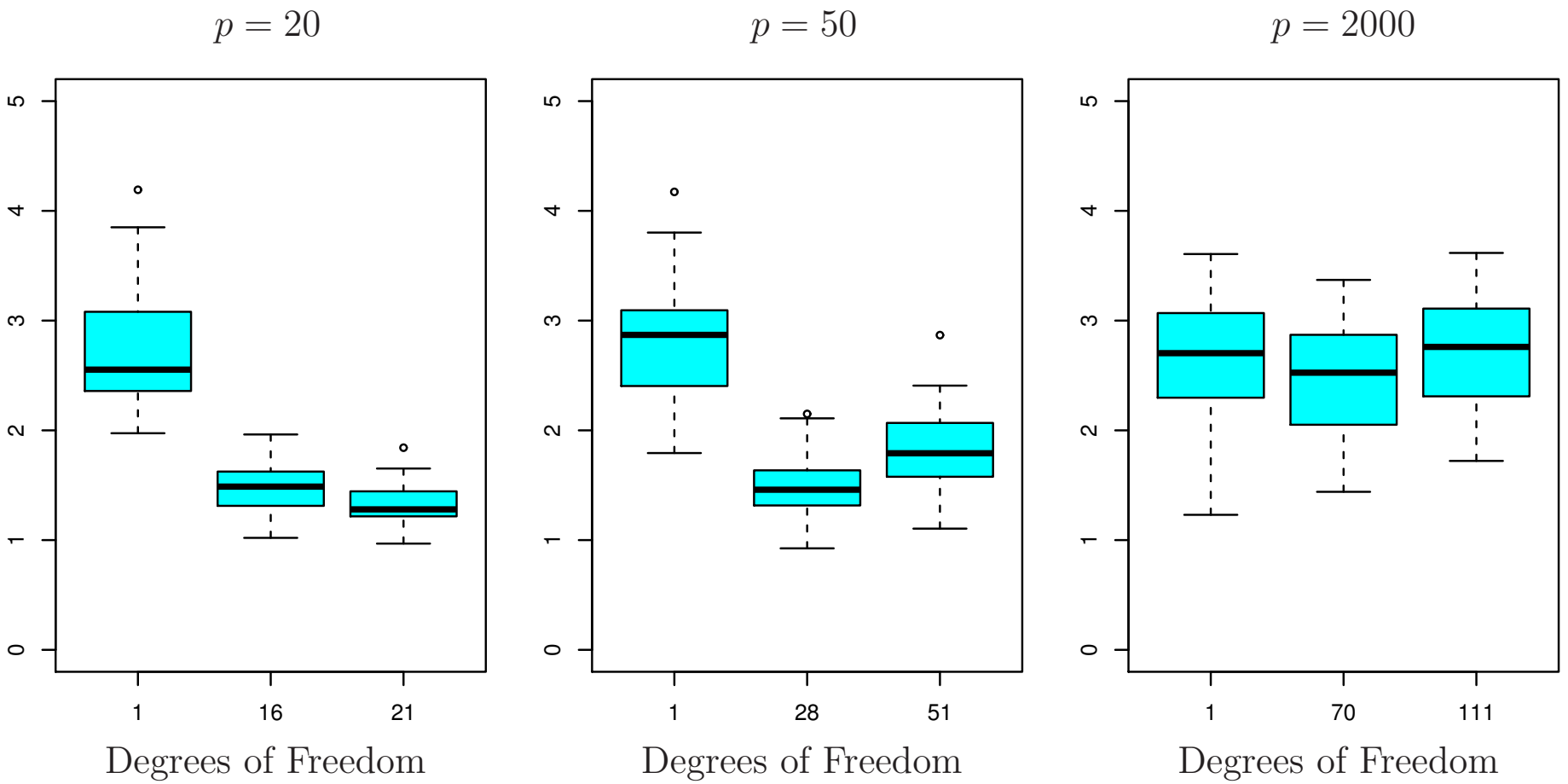
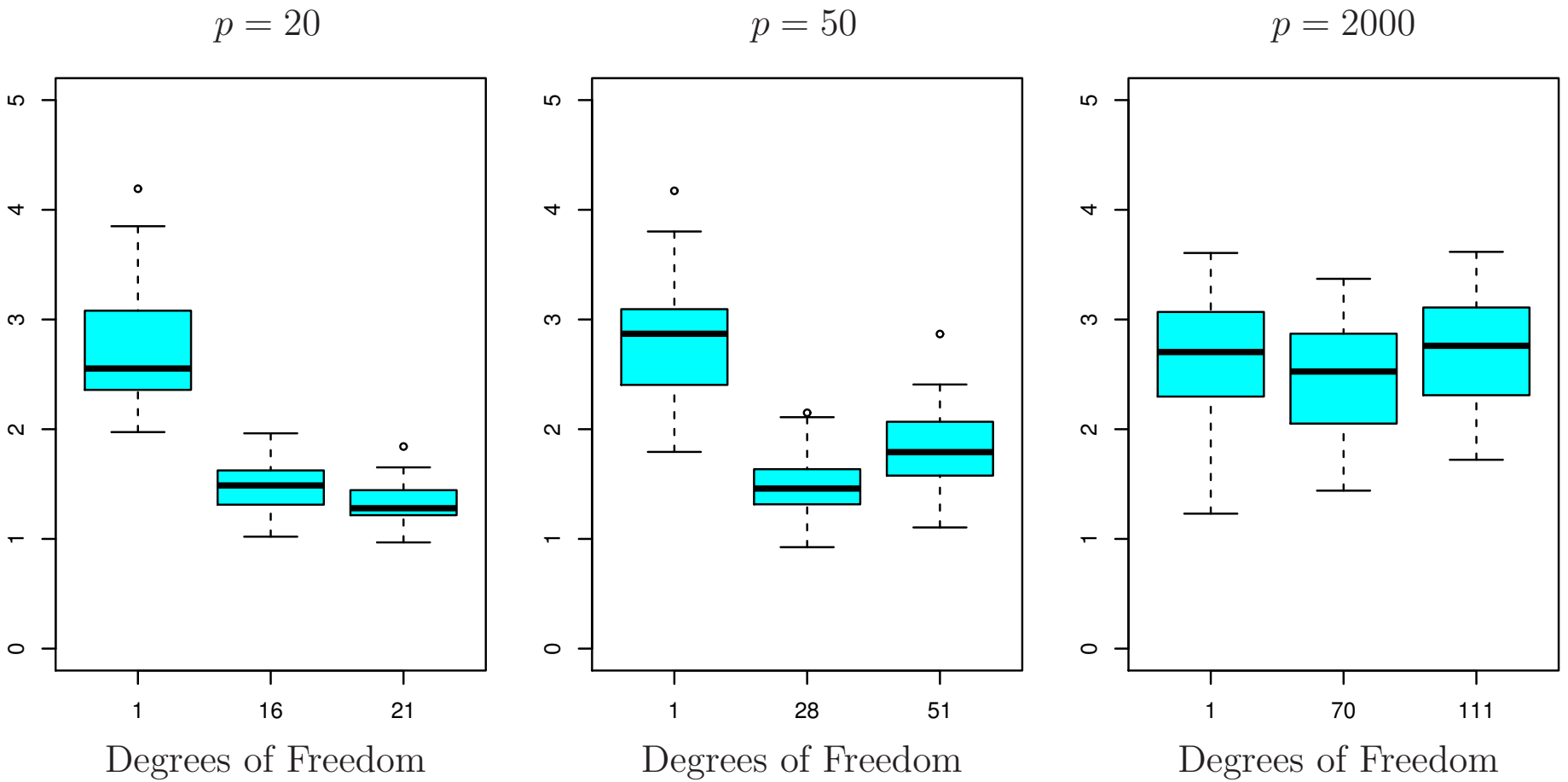


Figure: The lasso was performed with $n = 100$ observations and three values of p. Of the p features, 20 were associated with the response. The boxplots show the test MSEs that result using three different values of the tuning parameter λ.

Regression in High Dimensions



- Regularization or shrinkage plays a key role in high-dimensional problems,
- Appropriate tuning parameter selection is crucial for good predictive performance,
- and the test error tends to increase as the dimensionality of the problem increases, unless the additional features are truly associated with the response.

Outline

PCA: Principal Component Analysis

MONASH University

- PCA produces a low-dimensional representation of a dataset.
  - A sequence of linear combinations of the variables that have maximal variance and are mutually uncorrelated.
  - The first principal component is that (normalised) linear combination of the variables $(X_1, X_2, \ldots, X_p)$ with the largest variance.

$$Z_1 = \phi_{1,1} X_1 + \phi_{2,1} X_2 + \cdots + \phi_{p,1} X_p$$

    where $\sum_{j=1}^{p} \phi_{j,1}^2 = 1$.
  - We refer to the elements $\phi_{1,1}, \ldots, \phi_{p,1}$ as the loadings of the first principal component.
  - We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

PCA: Compute the First Principle Component

Suppose we have a $n \times p$ data set X. Since we are only interested in variance, we assume that each of the variables in X has been centred to have mean zero (that is, the column means of X are zero).

$$z_{i,1} = \phi_{1,1}x_{i,1} + \phi_{2,1}x_{i,2} + \cdots + \phi_{p,1}x_{i,p}$$

s.t. $\sum_{j=1}^{p} \phi_{j,1}^2 = 1$

- $z_{i,j}$ has mean zero since each of the $x_{i,j}$ has mean zero, due to the scaling.
- The sample variance of the $z_{i,1}$ can be written as

$$\frac{1}{n}\sum_{i=1}^{n} z_{i,1}^2$$

PCA: Compute the First Principle Component

Suppose we have a $n \times p$ data set X. Since we are only interested in variance, we assume that each of the variables in X has been centred to have mean zero (that is, the column means of X are zero).

$$z_{i,1} = \phi_{1,1}x_{i,1} + \phi_{2,1}x_{i,2} + \cdots + \phi_{p,1}x_{i,p}$$

s.t. $\sum_{j=1}^{p} \phi_{j,1}^2 = 1$

- The optimisation problem

$$\underset{\boldsymbol{\phi}_1}{maximize} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{\phi}_1^T \mathbf{x}_i \right)^2 \right\}$$

- Optimisation method: SVD (singular-value decomposition) of the matrix X, a standard techniques in linear algebra.

$$X_{n\times p} = U_{n\times n} S_{n\times p} V_{p\times p}^T$$

where $U^T U = I$, $V^T V = I$; the columns of U are orthonormal eigenvectors of XX^T , the columns of V are orthonormal eigenvectors of $X^T X$, and S is a diagonal matrix containing the square roots of eigenvalues from U or V in descending order. (`http://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm`)

PCA: Visualise First Principle Component

- The first principal component is that (normalised) linear combination of the variables with the largest variance.

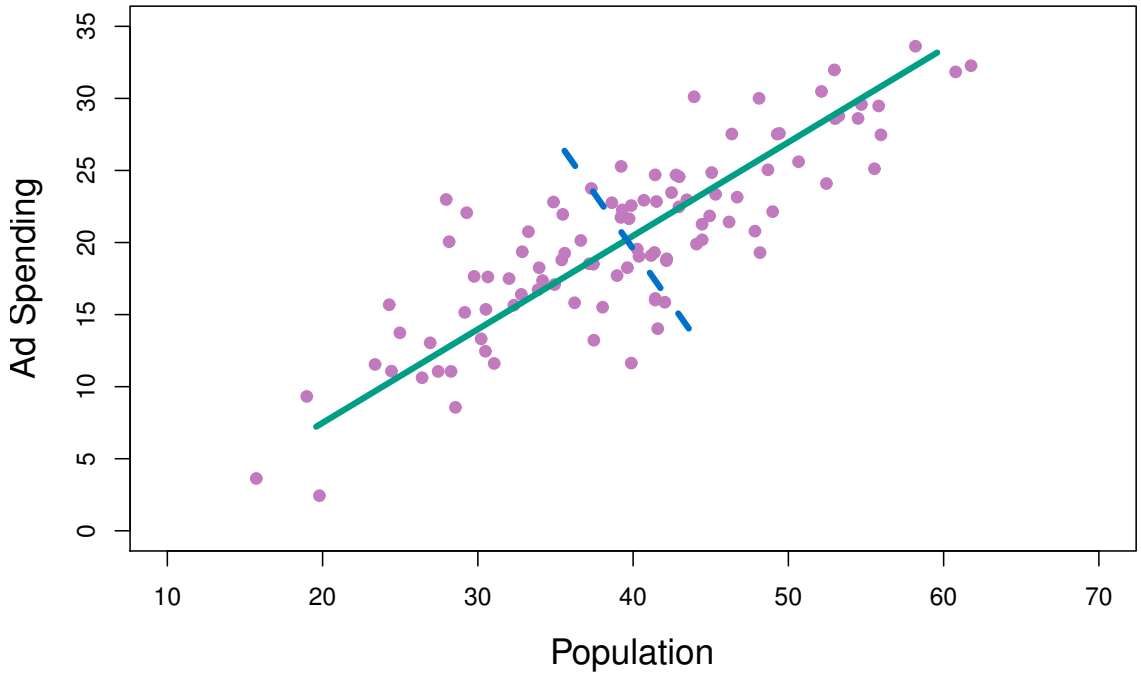


Figure: The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.

PCA: Visualise the First Principle Component

- The first principal component is that (normalised) linear combination of the variables with the largest variance.

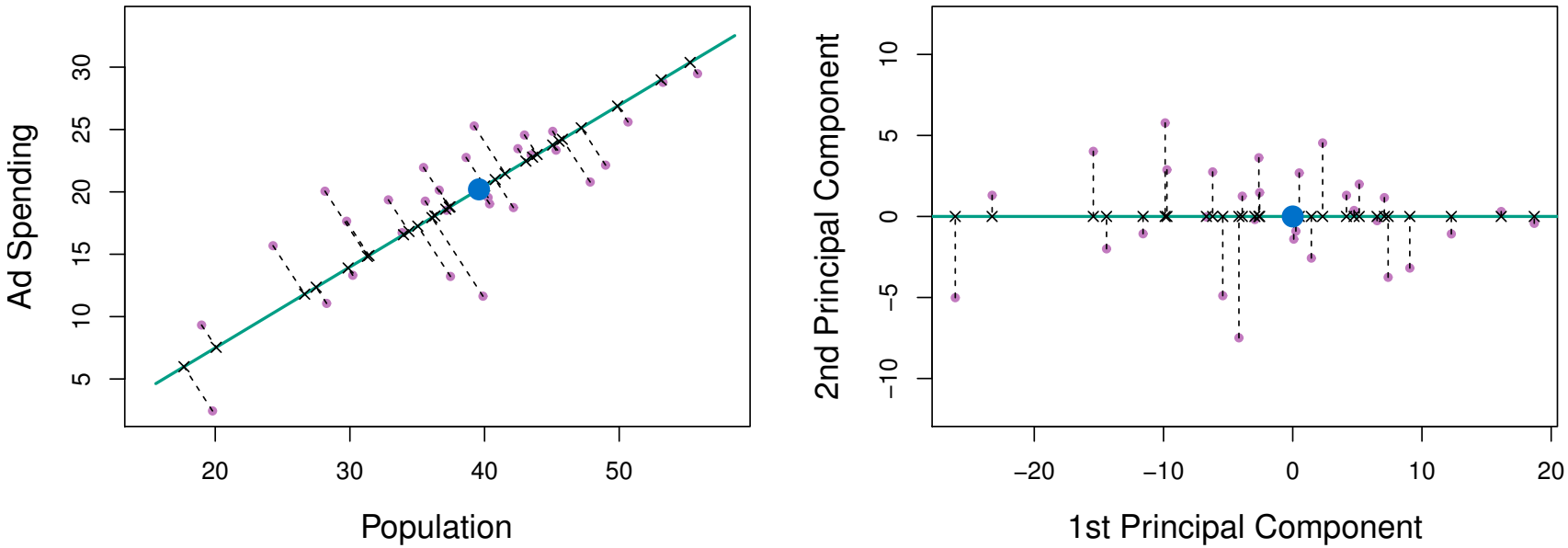
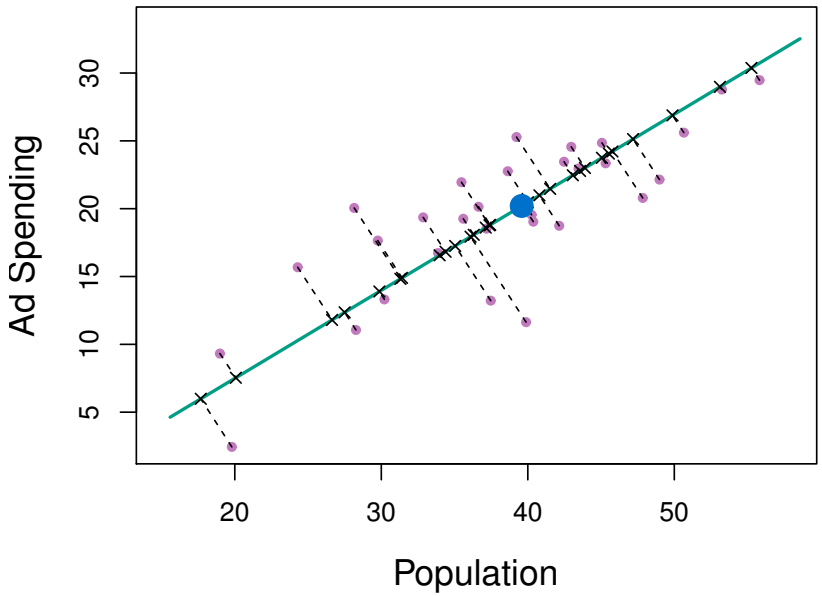
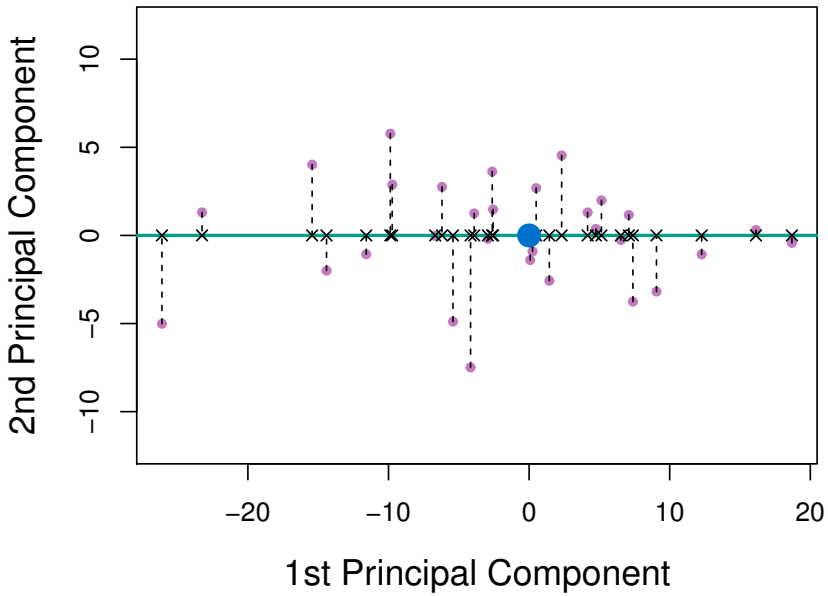


Figure: A subset of the advertising data. Left: The first principal component, chosen to minimise the sum of the squared perpendicular distances to each point, is shown in green. These distances are represented using the black dashed line segments. Right: The left-hand panel has been rotated so that the first principal component lies on the x-axis.

PCA: Visualise the First Principle Component

- The first principal component is that (normalised) linear combination of the variables with the largest variance.





  - Given the loadings for the first principle component, we can compute the principle component scores.

$$z_{i,1} = 0.839 \times (pop_i - \overline{pop}) + 0.544 \times (ad_i - \overline{ad})$$

PCA: Visualise the First Principle Component

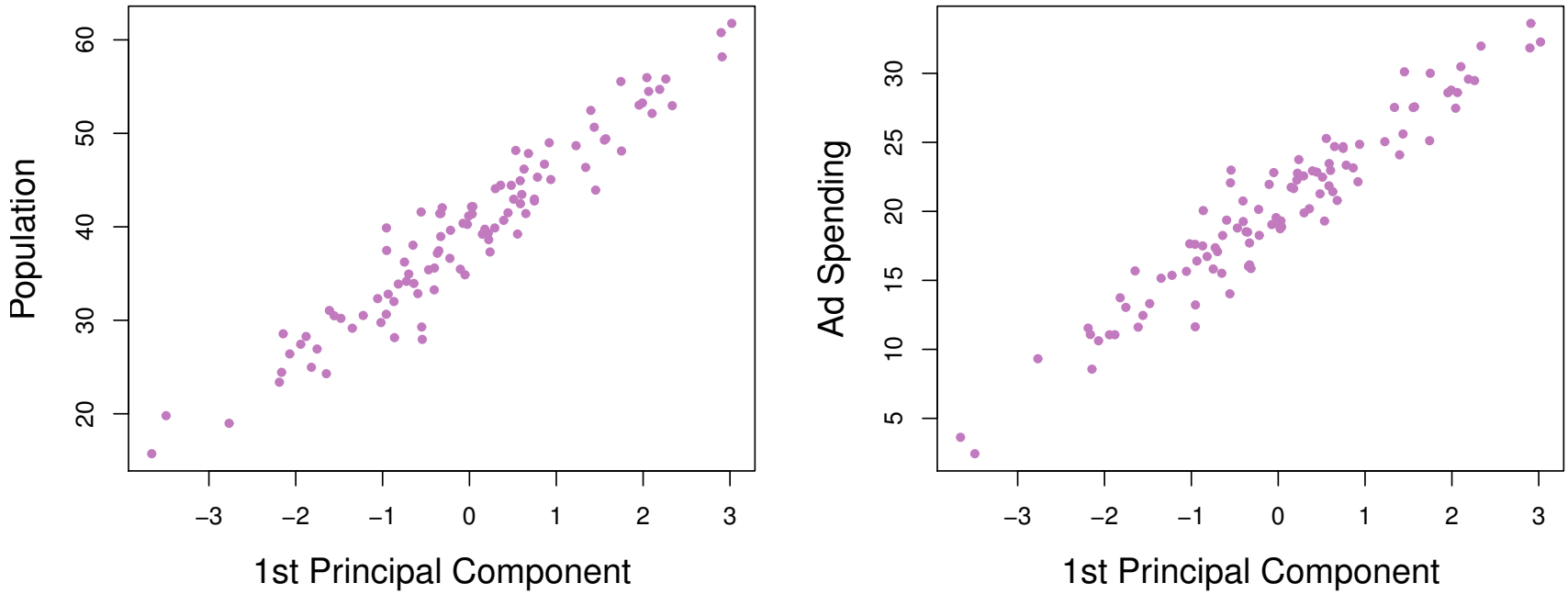


Figure: Plots of the first principal component scores $z_{i,1}$ versus pop and ad. The relationships are strong.

- pop and ad have approximately linear relationship (refer to slide-12)
- The above plots show a strong relationship between the first principle component and the two features.

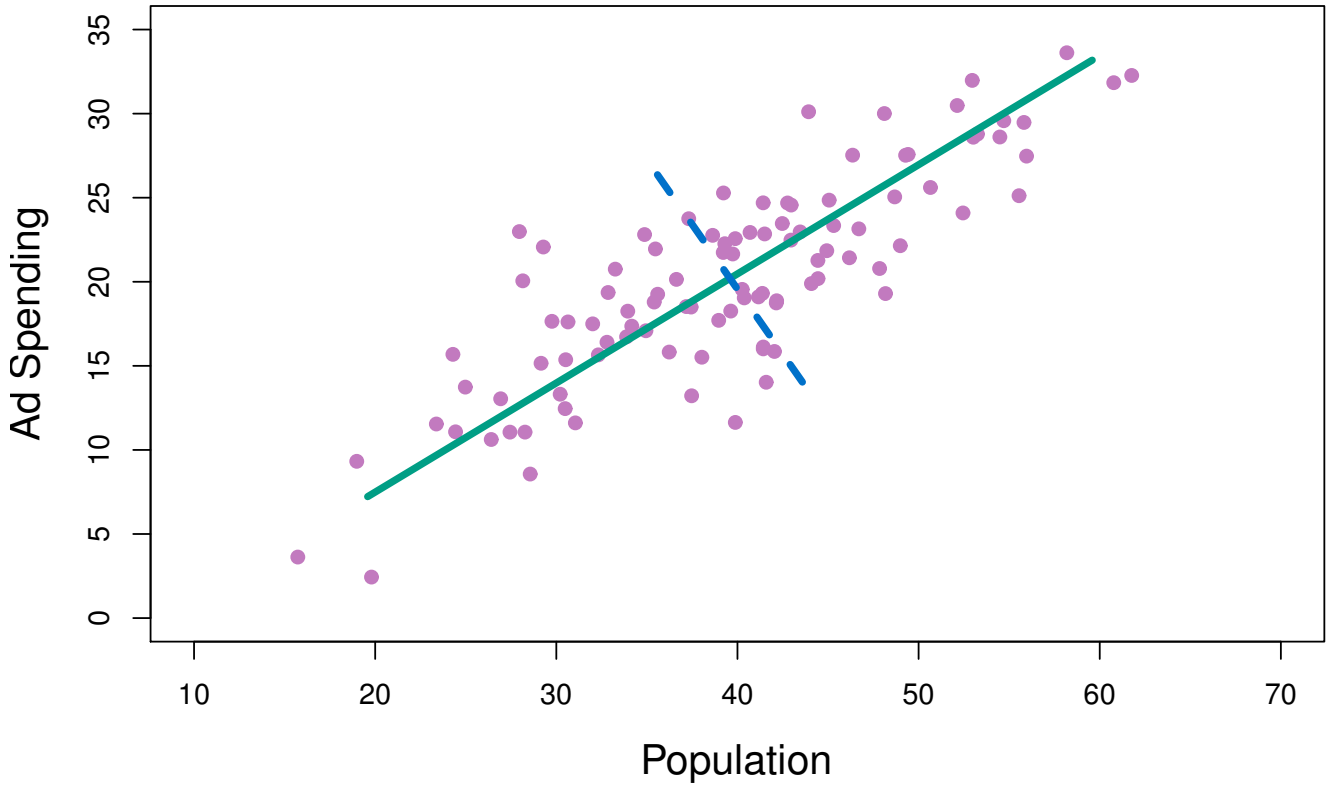
PCA: the Second Principle Component

- The second principal component is the linear combination of $X_1, X_2, \ldots, X_p$ that has maximal variance among all linear combinations that are uncorrelated with Z_1.

$$z_{i,2} = \phi_{1,2}x_{i,1} + \phi_{2,2}x_{i,2} + \cdots + \phi_{p,2}x_{i,p}$$

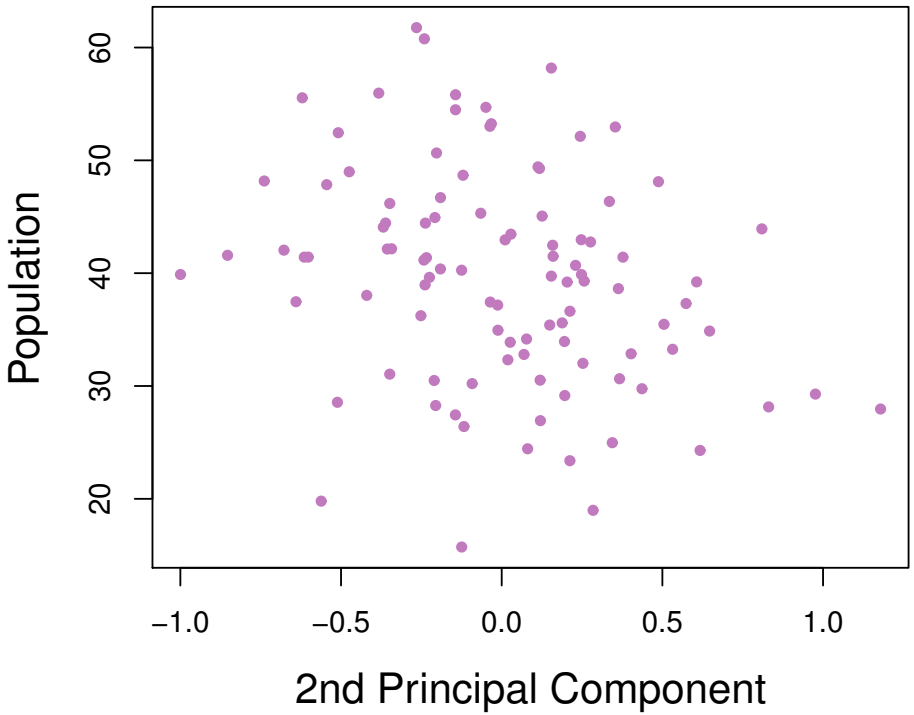
  where $\boldsymbol{\phi}_2$ is the second principle component loading vector.
- Constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction $\boldsymbol{\phi}_2$ to be orthogonal (perpendicular) to the direction $\boldsymbol{\phi}_1$.
- The principal component directions $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3, \ldots$ are the ordered sequence of right singular vectors of the matrix X, and the variances of the components are 1 times the squares of the singular values. There are at most $min(n-1, p)$ principal components.

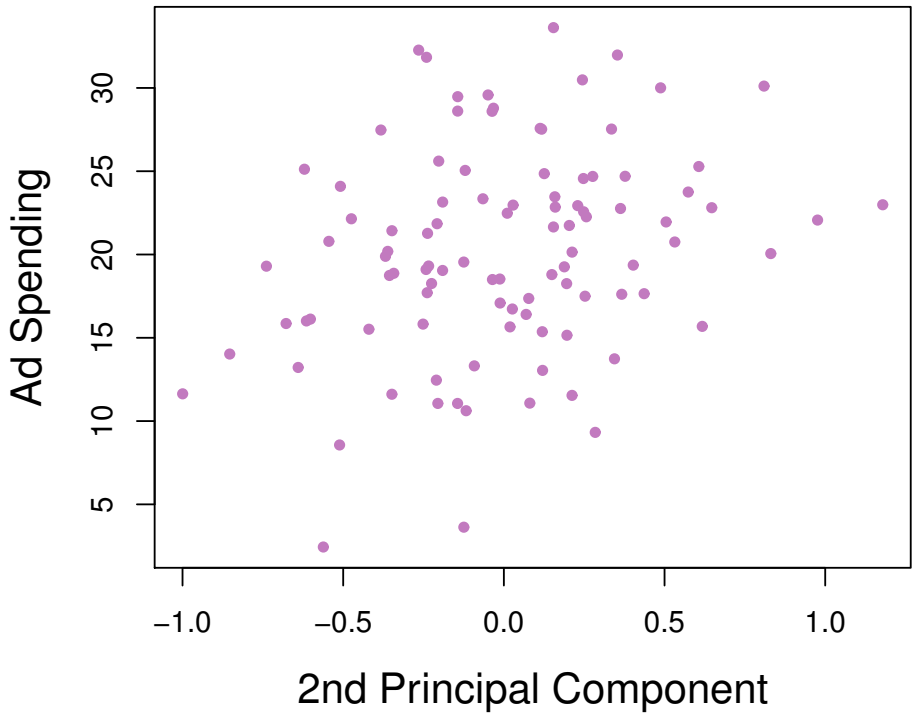
PCA: the Second Principle Component



- The first two components contain all the information that is in pop and ad.
- The first component contains the most information.
- The second component scores are much closer to zero → capture less information

PCA: the Second Principle Component





- Little relationship between the second principal component and these two predictors.

PCA: Geometry of PCA

- The loading vector ϕ_1 with elements $\phi_{1,1}, \phi_{2,1}, \ldots, \phi_{p,1}$ defines a direction in feature space along which the data vary the most.
- If we project the n data points $x_1, x_2, \ldots, x_n$ onto this direction, the projected values are the principal component scores $z_{1,1}, z_{2,1}, \ldots, z_{n,1}$ themselves.

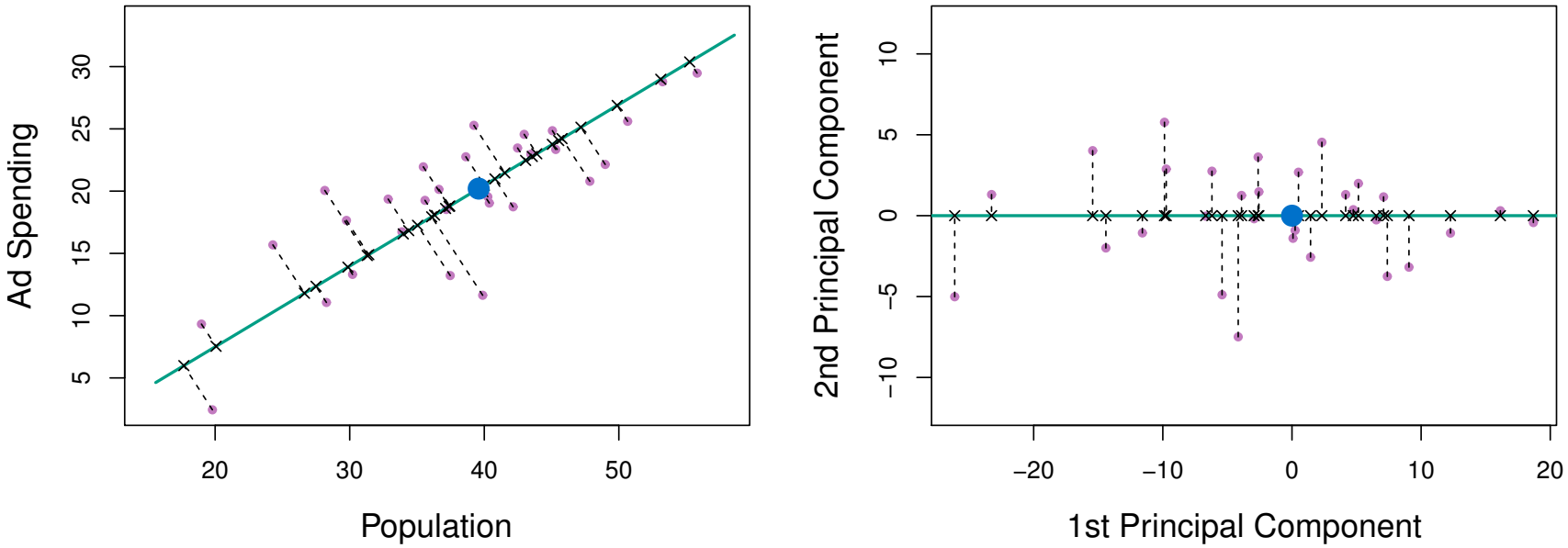


Figure: A subset of the advertising data. Left: The first principal component, chosen to minimise the sum of the squared perpendicular distances to each point, is shown in green. These distances are represented using the black dashed line segments. Right: The left-hand panel has been rotated so that the first principal component lies on the x-axis.

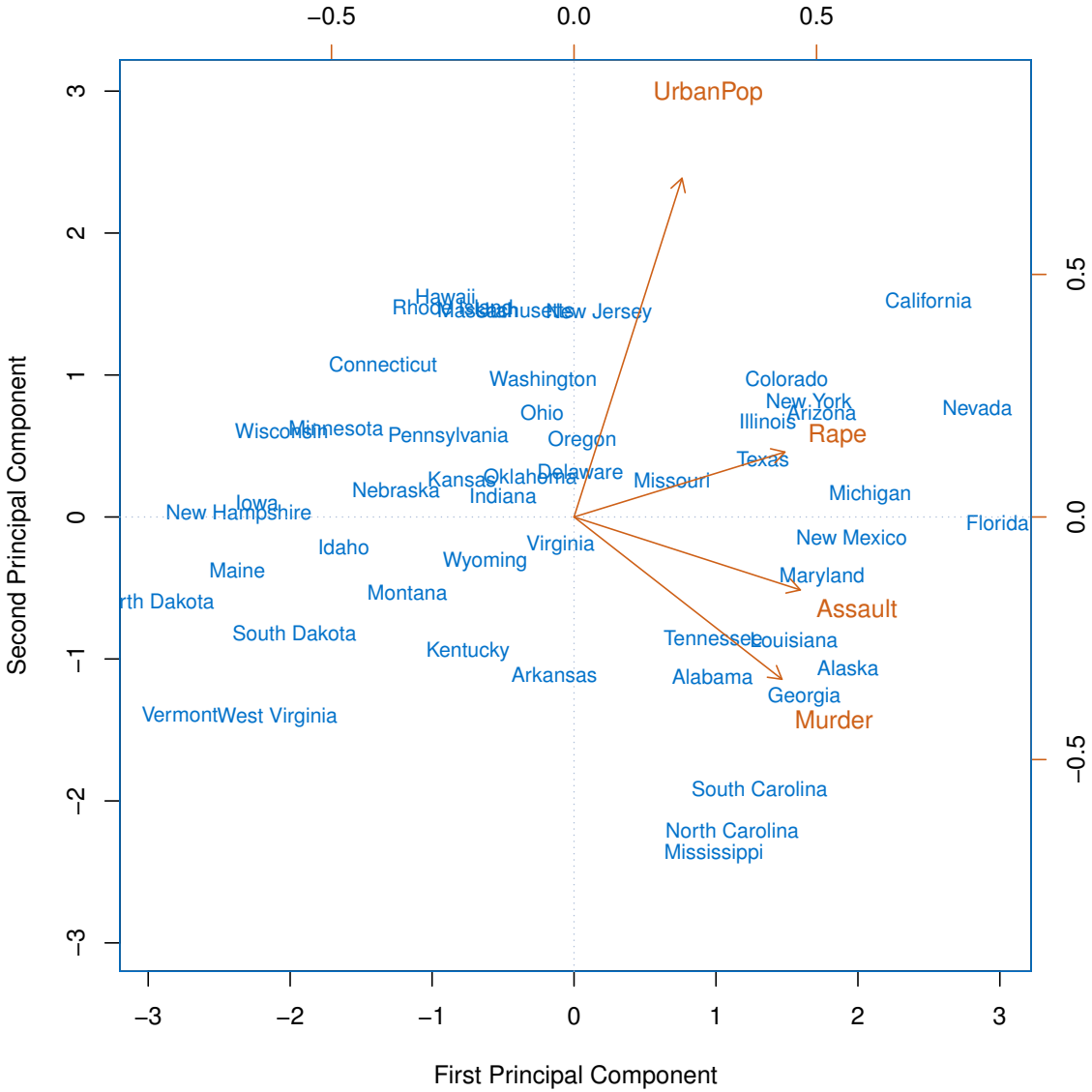
PCA: Illustration — Data

- USArrests data: For each of the fifty states in the United States, the data set contains the number of arrests per 100, 000 residents for each of three crimes: Assault, Murder, and Rape. We also record UrbanPop (the percent of the population in each state living in urban areas).
- The principal component score vectors have length $n = 50$, and the principal component loading vectors have length $p = 4$.

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

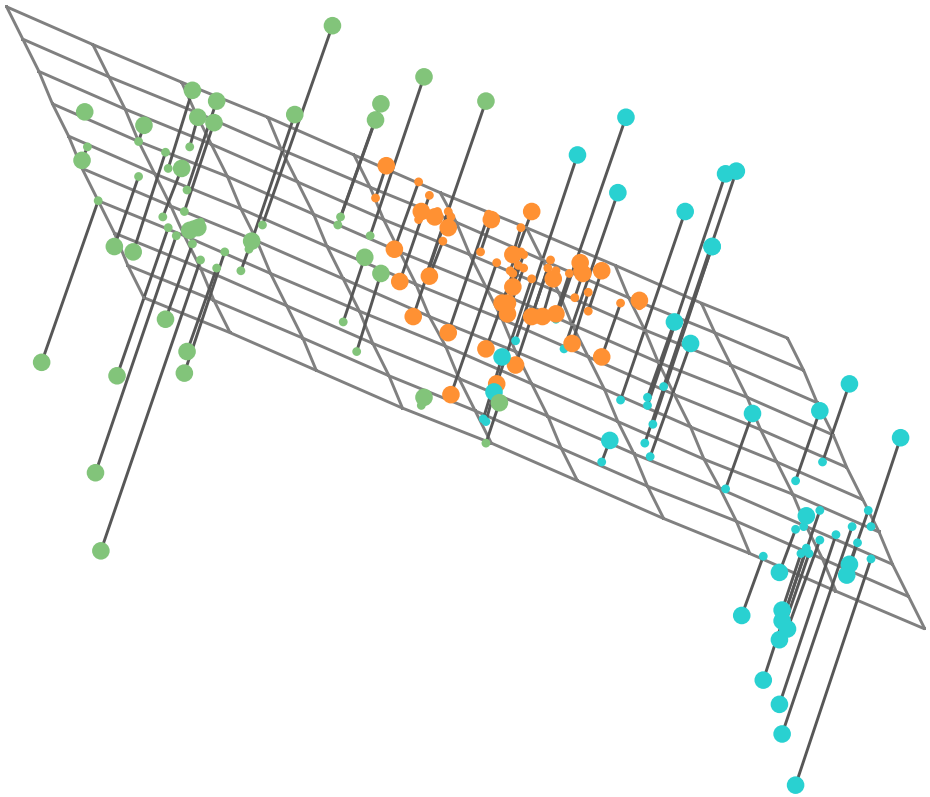
- PCA was performed after standardising each variable to have mean zero and standard deviation one.

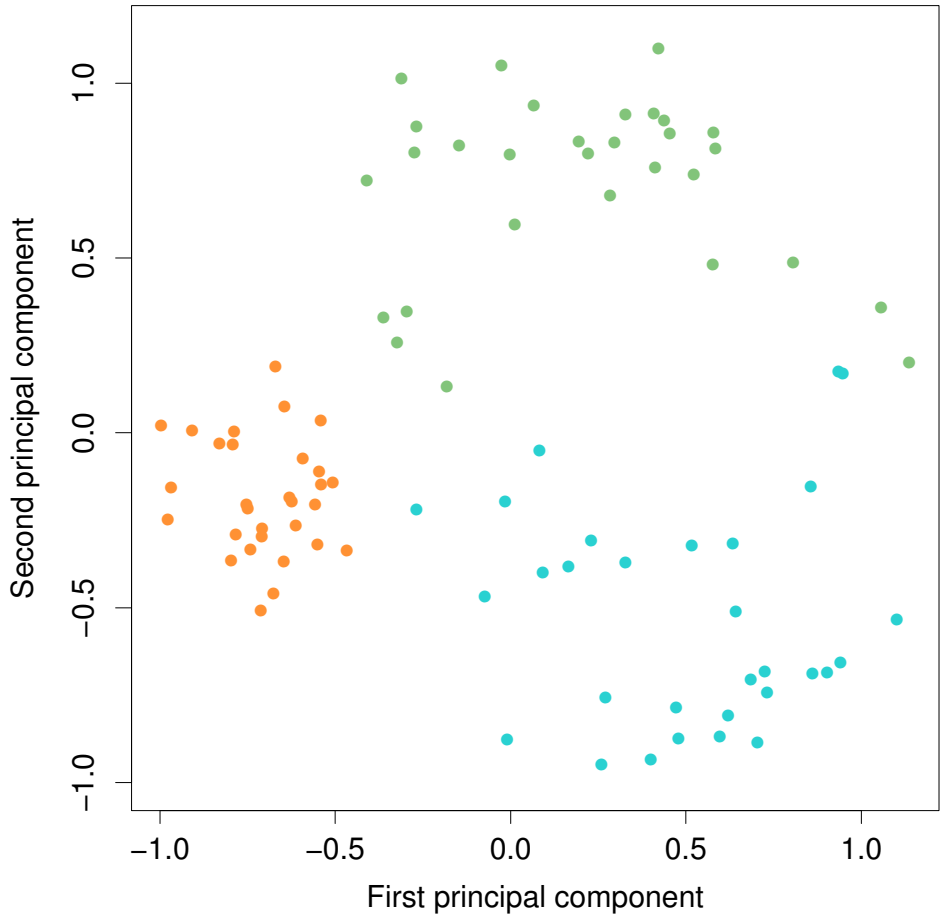
PCA: Illustration — biplot



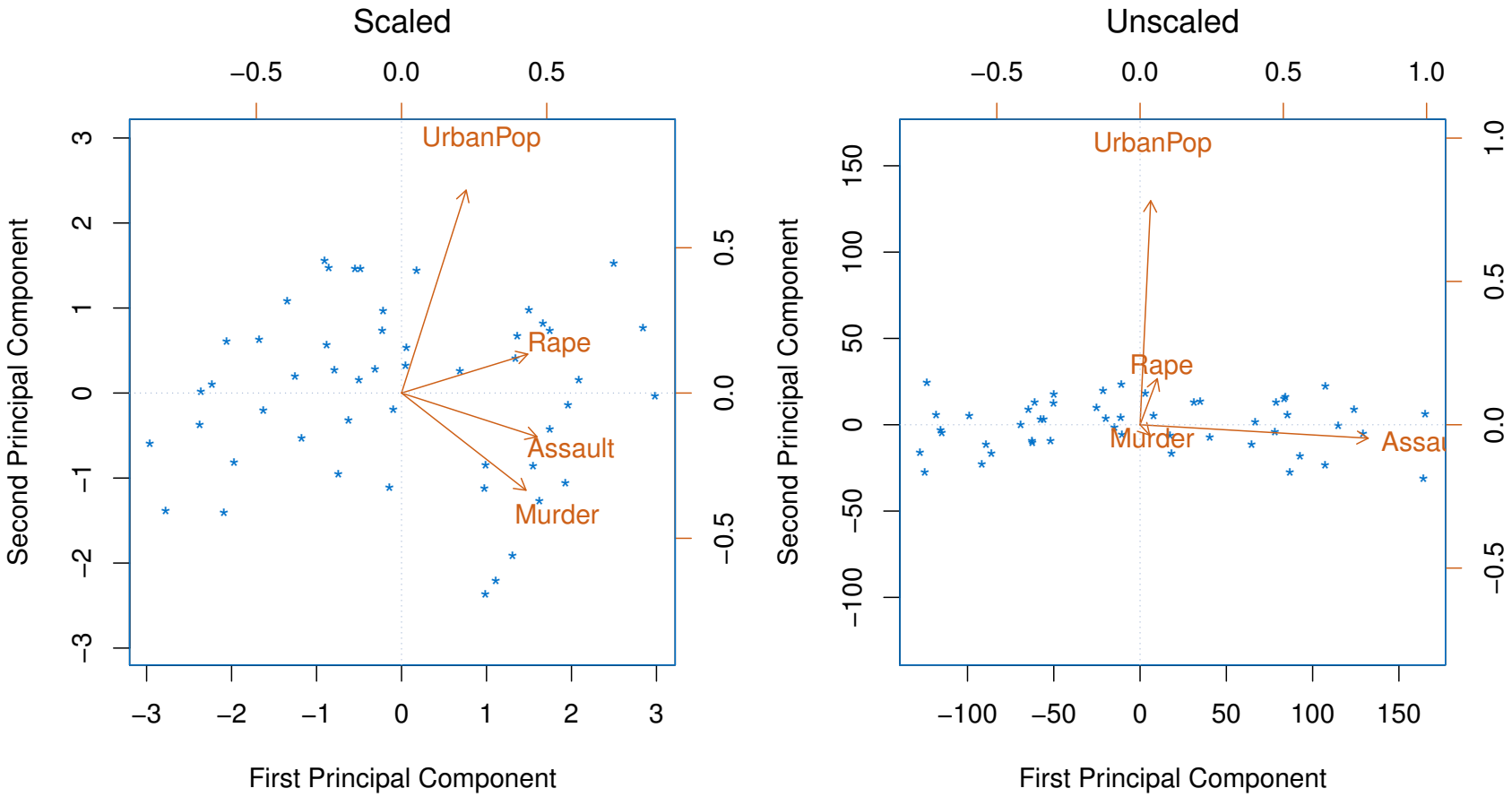
	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

PCA: Illustration — Projection





Scaling of the Variables



- The four variables (Murder, Rape, Assault and UrbanPop) have variance: 18.97, 87.73, 6945.16, and 209.5.
- If the variables are in different units, scaling each to have standard deviation equal to one is recommended

PVE—Proportion of Variance Explained

- PVE measures the strength of each component
- The total variance presented in a data set

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{i,j}^2$$

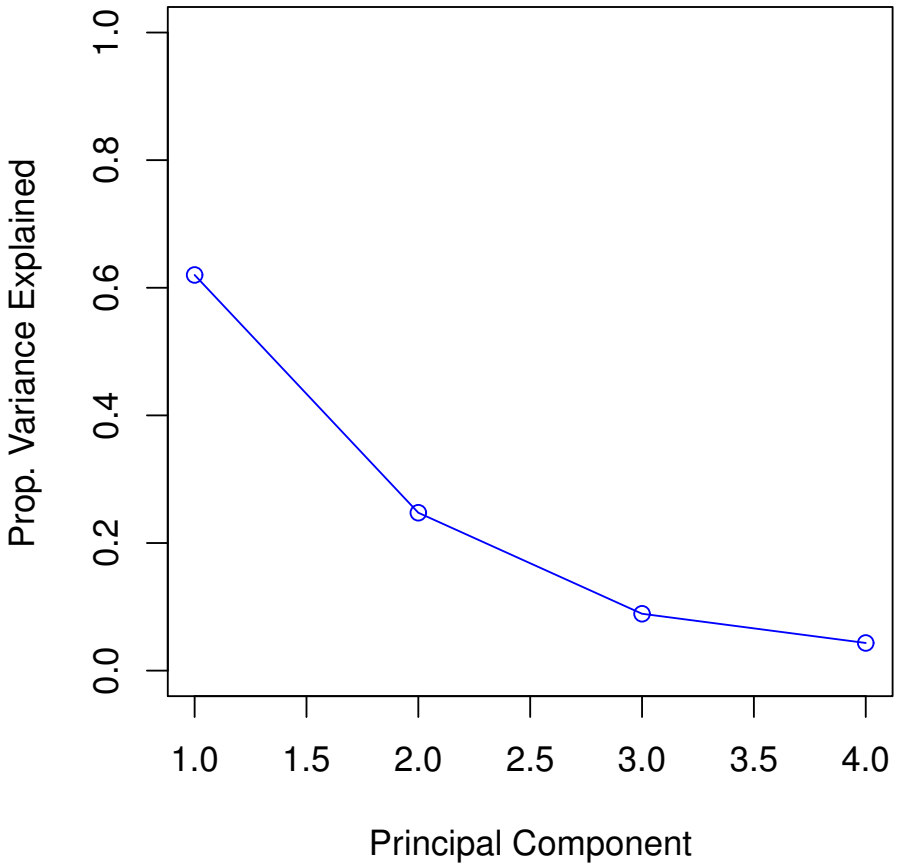
- The variance explained by the m-th PC is

$$\frac{1}{n} \sum_{i=1}^{n} z_{i,m}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j,m} x_{i,j} \right)^2$$

- The PVE of the m-th PC is given by

$$\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j,m} x_{i,j} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{i,j}^2}$$

PVE—Proportion of Variance Explained



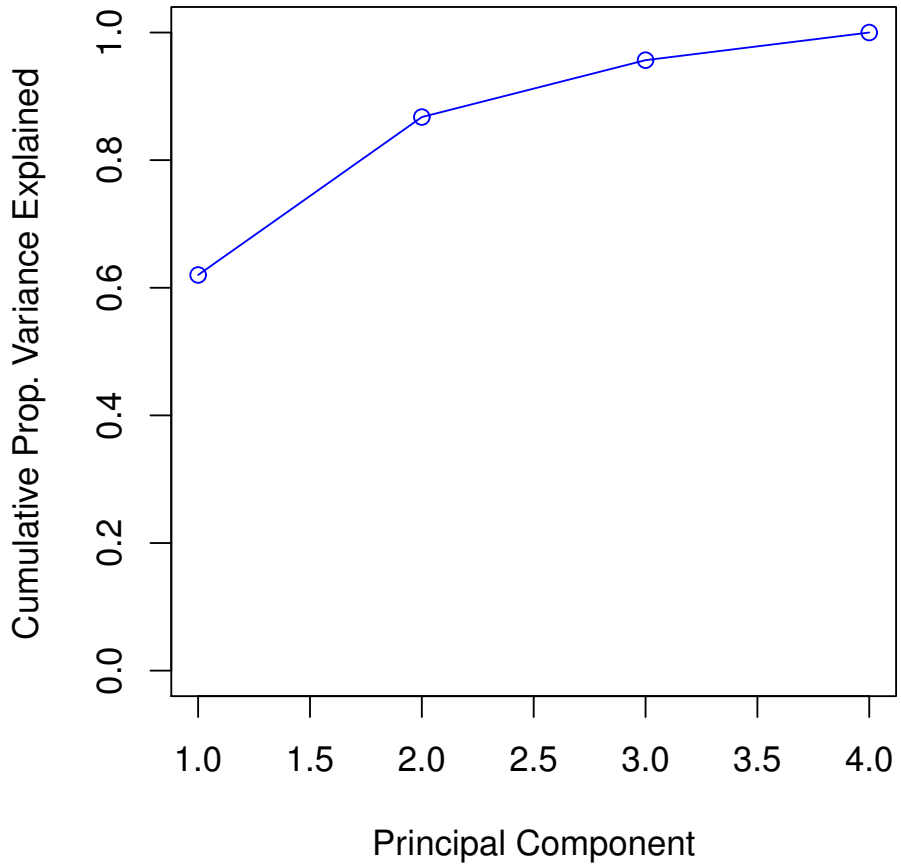


Figure: Left: a scree plot depicting the proportion of variance explained by each of the four PCs in the USArrests data. Right: the corresponding cumulative PVE.

Outline

Use PCs in linear regression

- Basic idea: transform the predictors and then fit a least squares model using the transformed variables.
  - Let $Z_1, Z_2, \ldots, Z_m$ represent $M < p$ linear combinations of our original p predictors.

$$Z_m = \sum_{j=1}^{p} \phi_{j,m} X_j = \boldsymbol{\phi}_m^T \mathbf{X}$$

    where $\boldsymbol{\phi}$ are some constants to be learned.
  - Then, the linear model to be fit is

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{i,m} + \epsilon_i, \quad i = 1, \ldots, n,$$

    where

$$z_{i,m} = \boldsymbol{\phi}_m^T \mathbf{x}_i$$

  - The regression coefficient: $\theta_1, \theta_2, \ldots, \theta_M$.

Use PCs in linear regression — continued

- Dimension reduction serves to constrain the estimated coefficients, β_j.

$$\begin{aligned}
\sum_{m=1}^{M} \theta_m z_{i,m} &= \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{j,m} x_{i,j} \\
&= \sum_{m=1}^{M} \sum_{j=1}^{p} \theta_m \phi_{j,m} x_{i,j} \\
&= \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{j,m} x_{i,j} \\
&= \sum_{j=1}^{p} \beta_j x_{i,j}
\end{aligned}$$

- Can win in the bias-variance tradeoff
  - If $p \gg n$, $m \ll p$ significantly reduce the variance of the estimated coefficients.
  - if $p = n$ and all Z_m are linearly independent, the model with dimension reduction is equivalent to the lease square model.

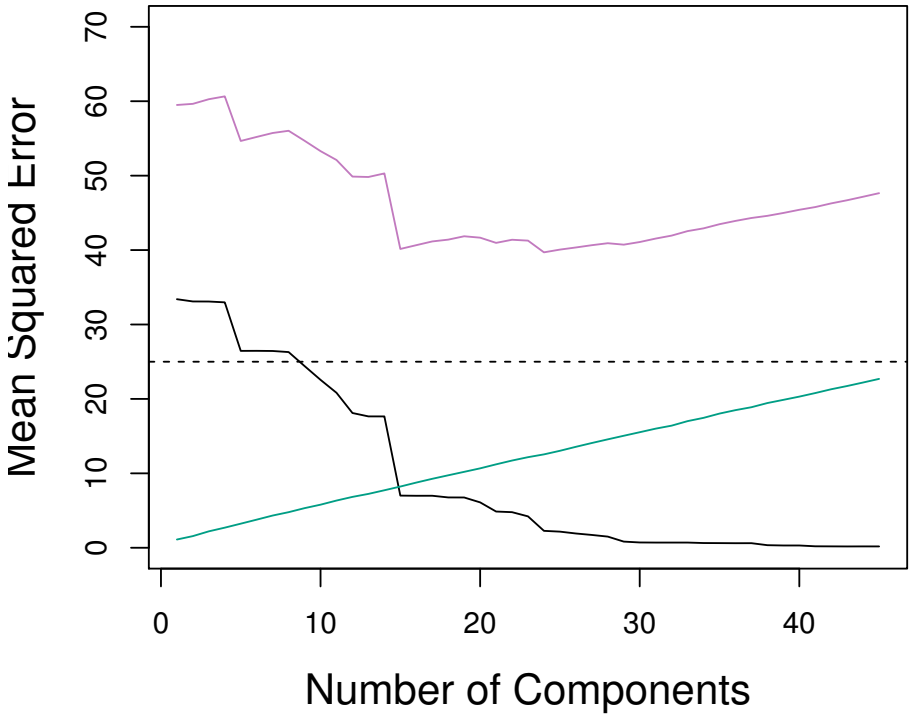
Principal Component Regression

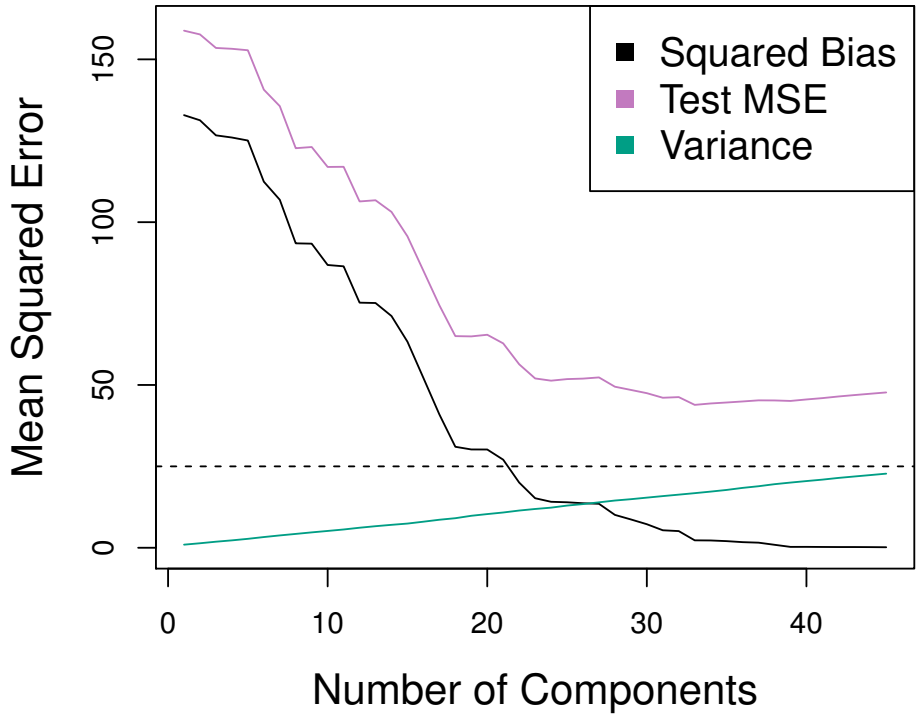
- Basic idea:
  1. Construct the first M principal components, $Z_1, Z_2, \ldots, Z_m$
  2. Fit a linear regression with the M components as predictors

$$Y = \beta_0 + \sum_{m=1}^{M} \beta_m Z_m + \epsilon$$

- Assumption: the directions in which $X_1, X_2, \ldots, X_p$ show the most variation are the directions that are associated with Y.
  - Not guaranteed to be true, but a reasonable enough approximation
  - If the assumption holds, fitting a least square model to the M principal components is better than to the P variables.
  - Mitigate overfitting: $M \ll P$

Principal Component Regression: Simulated data 1





- Data: $n = 50$ observations and $p = 45$ predictors
  - Left: the response variable as a function of all the 45 predictors
  - Right: the response variable as function of only two predictors
- Performance worse than the shrinkage methods: the data is generated in such a way that many principal components are required in order to adequately model the response variable.

Principal Component Regression: Simulated data 2

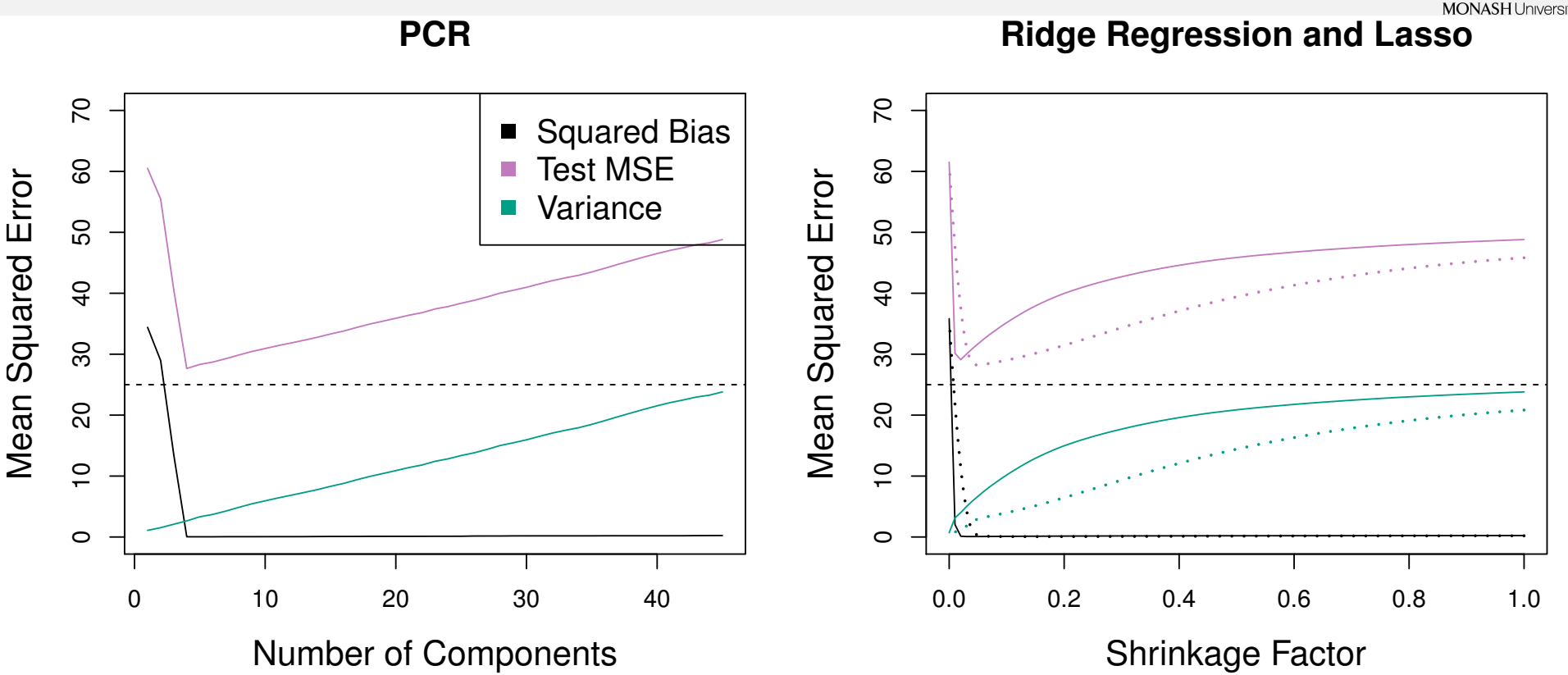


Figure: Left: results for PCR. Right: Results for lasso (solid) and ridge regression (dotted)

- A simulated dataset that is favourable to PCR.
  - The response variable is a function of the first five principal component.
- The plots show
  - The bias drops to zero rapidly as M increases.
  - $M = 5$ gives the minimum MSE

Principal Component Regression: M=?

- PCR is not a feature selection method.
  - A linear combination of all p of the original features (or predictors)
  - Closely related to ridge regression than to the lasso
- How many principal components should we use?

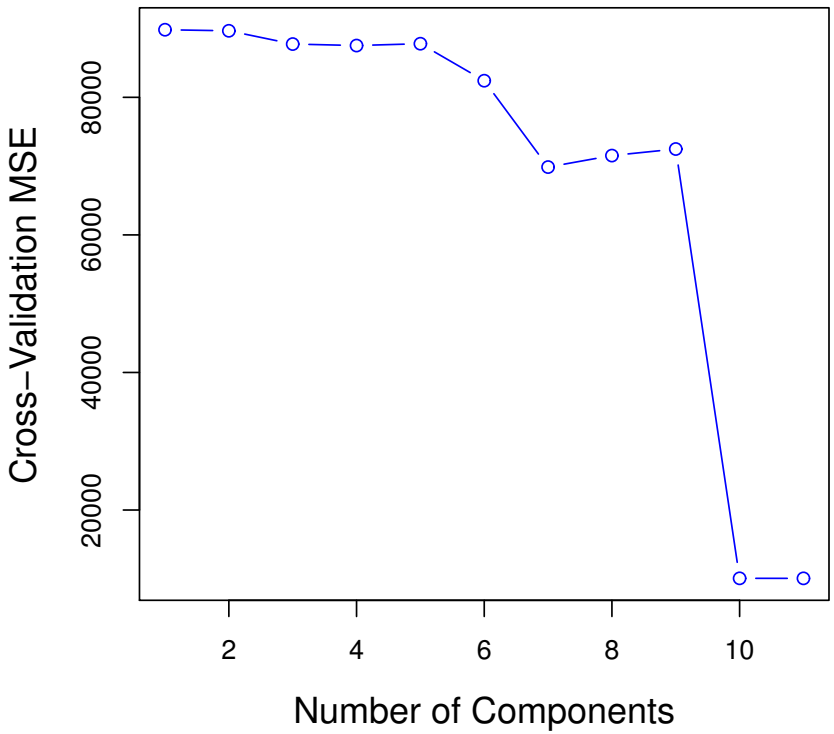


Figure: The 10-fold cross validation MSE obtained using PCR, as a function of M.

Principal Component Regression: Standardising Features

- The scale on which the variables are measured ultimately has an effect on the final PCR model
  - measures of tree size
    - the trunk diameter in cm, biomass of leaves in kg, number of branches, overall height in meters
- Variables with the highest sample variances will tend to be emphasised in the first few principal components.
- Principal component analysis using the covariance function should only be considered if all of the variables have the same units of measurement.
- Each component's eigenvalue represents how much variance it explains.
- Standardized variables before PCA:

$$\hat{X}_{i,j} = \frac{X_{i,j} - \overline{x_j}}{s_j}$$

where s_j is the standard deviation of the j-th variable.

Outline

Partial Least Squares: A supervised alternative to PCR

- PCR identifies linear combinations, or directions, that best represent the predictors $X_1, X_2, \ldots, X_p$.
- The principal componenets are identified in an unsupervised way:
  - the response does not supervise the identification of the principal components.
- A potentially serious drawback
  - No guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

Partial Least Squares: A supervised alternative to PCR

MONASH University

- What is similar to PCR:
  - First identifies a new set of features $Z_1, Z_2, \ldots, Z_M$ that are linear combinations of the original features.
  - Then, fits a linear model via least squares using these M new features.
- What is different to PCR:
  - Use the response variable Y to supervise the learning of $Z_1, Z_2, \ldots, Z_M$ so that the direction found by PLS can explain both the response and the predictors.

Partial Least Squares: How to compute the coefficients?

- Details of PLS:
  1. Standardise the p features (or predictors)
  2. Compute Z_1 by setting each $\phi_{j,1}$ $(1 \leq j \leq q)$ equal to the coefficient from the simple linear regression of Y onto X_j.
     - The coefficient is proportional to the correlation between Y and X_j
  3. Compute
  $$Z_1 = \sum_{j=1}^{p} \phi_{j,1} X_j$$
  where PLS places the highest weight on the variables that are most strongly related to the response.
  4. Subsequent directions are found by taking residuals and then repeating the above prescription.

PLS V.S. PCA

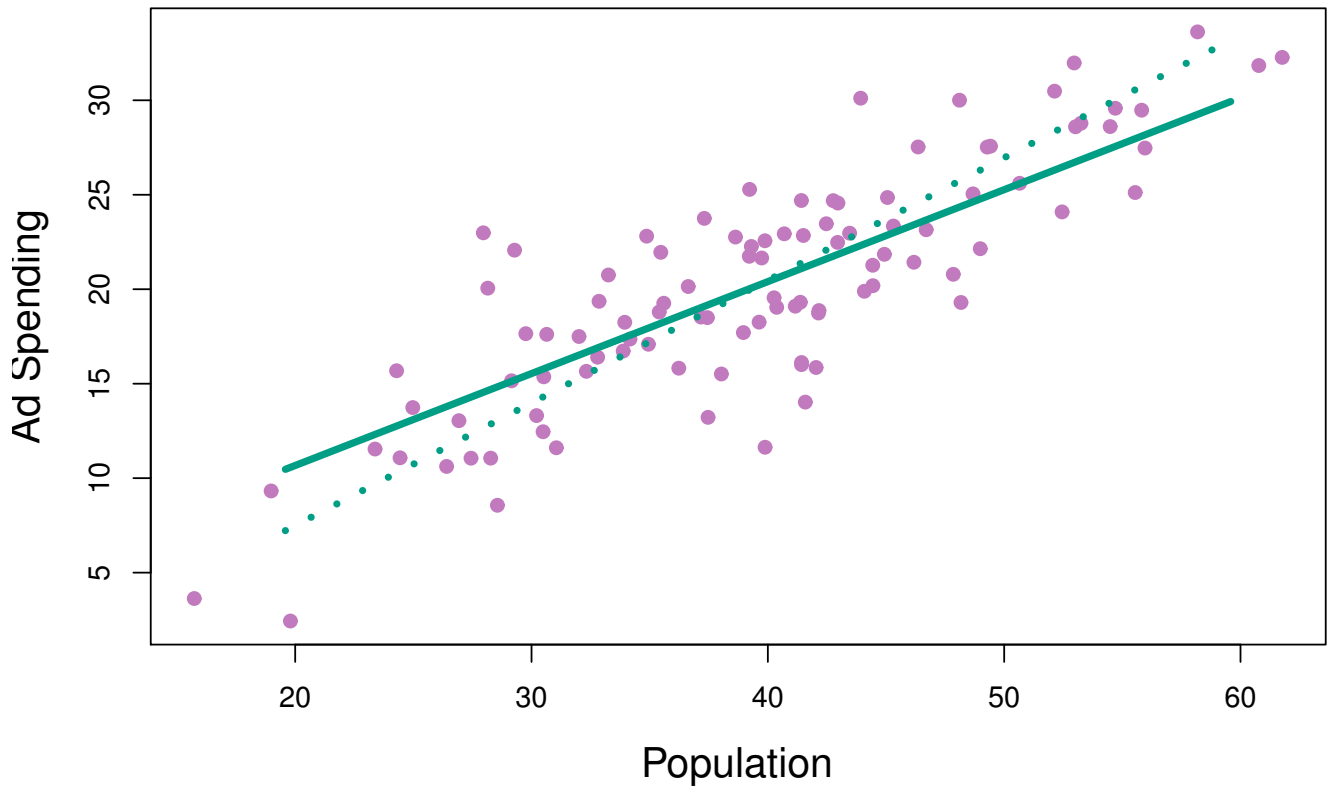


Figure: For the advertising data, the first PLS direction (solid line) and first PCR direction (dotted line) as shown.

- The comparison suggests
  - pop is more highly correlated with the response variable than is ad.

Summary

- What goes wrong in high dimensions?
- Dimension reduction methods
  - PCA
  - PCR
  - PLS
- Reading materials:
  - "Linear Model Selection and Regularisation", Chapter 6 of "Introduction to Statistical Learning"
    - Section 6.3 "Dimension Reduction Methods"
    - Section 6.4 "Considerations in High Dimensions"
  - "Unsupervised learning", Chapter 10 of "Introduction to Statistical Learning"
    - Section 10.2 "Principal Components Analysis"
- Acknowledgement:
  - Figures in this presentation were taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
  - Some of the slides are reproduced based on the slides from T. Hastie and R. Tibshirani