Statistical Learning

Week 3 - Dimension reduction techniques

Pedro Galeano
Department of Statistics
UC3M-BS Institute on Financial Big Data
Universidad Carlos III de Madrid
pedro.galeano@uc3m.es

Academic Year 2017/2018

Master in Big Data Analytics

uc3m | Universidad Carlos III de Madrid

Introduction

- Well-structured data set: Consider any well-structured data set.
- Data matrix: X of size $n \times p$.
- Sample size: n.
- Dimension: p.
- Curse of dimensionality: If the ratio n/p is not large enough, some problems might be intractable.
- Particularly: If p is large (even larger than n), data visualization becomes very difficult (if not impossible) and standard classification methods perform poorly.
- Thus: In these scenarios, it is very complicated to find interesting features in the data because of the accumulation of noise.

Introduction

- Noise features: Data sets with many variables use to contain many uninformative features.
- Dimension reduction: Transform the data matrix X into another data matrix Z with a smaller dimension (same sample size).
- Important: Z should contain the important features in X but should not contain the noise features in X.
- Thus:
  - Z should be more simple to analyze and to visualize.
  - Z should have larger discriminant power than X, if possible.
- Dimension reduction tools are: More of a means to an end rather than an end in themselves, because they frequently serve as an intermediate step in another analysis.

Introduction

- Principal component analysis (PCA): The most popular method for dimension reduction.

- Idea: Perform a linear transformation of the original data matrix, X, preserving its important features and reducing the noise.

- Properties of PCA:

  - The transformed variables are uncorrelated, thus they do not share linear information.

  - Powerful method to interpret the relationship between the variables in the data set.

  - Use to reveal unsuspected relationships and thereby allows interesting interpretations.

  - Clusters and outliers in the original data set are usually clearly shown in the transformed data set.

  - Sometimes increases the discriminatory power of the data set.

Introduction

- PCA: Depends solely on the sample covariance (or correlation) matrix of X.
- Sparse Principal Component Analysis (SPCA): Similar to PCA but attempt to simplify the interpretation of the PCs.
- Independent Component Analysis (ICA): Tries to obtain independent variables instead of uncorrelated variables.
- Nevertheless: The mathematical treatment of ICA and other alternatives becomes more difficult and computation becomes much more complex.

Introduction

- The rest of this chapter is devoted to:
  - Establish the main ideas of the principal component analysis.
  - Describe how to perform principal component analysis in practice.
  - Introduce sparse principal component analysis and independent component analysis.
  - Illustrate these techniques with real data examples.

Principal component analysis

- Data matrix: X of size $n \times p$.
- Quantitative variables: X should only contains quantitative variables.
- Binary variables: There is not a consensus on the inclusion of binary variables in a PCA.
- Sample covariance and sample correlation matrices: PCA are based on the information given by one of these two matrices.
- Interpretation: The meaning of the sample covariance and correlation coefficients between a quantitative variable and a binary variable differ from those between quantitative variables.

Principal component analysis

- Center the data: PCA starts by centering the variables in the data matrix.
- Why?: The linearly transformed data set will be centered as well, thus, we avoid sample mean vectors for the new variables.
- Centered data matrix: $\widetilde{X} = X - 1_n \overline{x}'$, where $\overline{x}$ be the sample mean vector of X and 1_n is the $n \times 1$ vector of ones.
- Goal of PCA: Obtain a linear transformation of $\widetilde{X}$, $Z = \widetilde{X} C$, where C is a matrix of size $p \times r$ such that:
  1. Z has smaller dimension than X, i.e., $r < p$.
  2. Z contains the important features in X.
  3. Z does not contain the irrelevant features X.

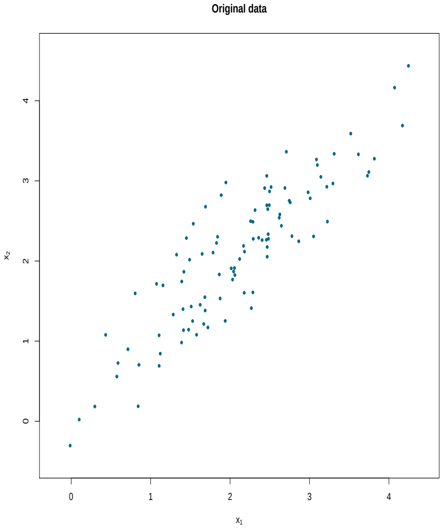
Principal component analysis

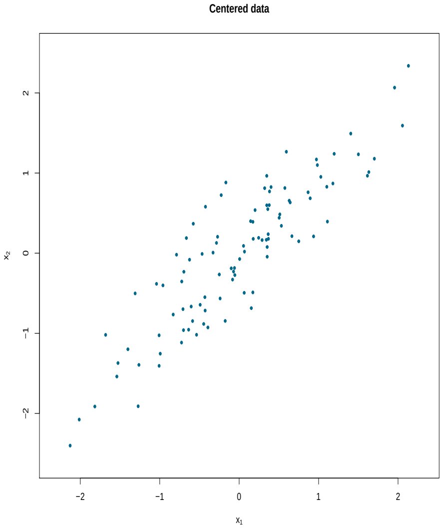
- Assume we want $r = 1$: Z has dimension $n \times 1$.
- The problem: Find the vector $C = (C_1, \ldots, C_p)'$ such that:
  - $Z = \widetilde{X} C$.
  - Z contains the most important features in X.
- The question is: How to do this?

Principal component analysis

- Toy example:
  - Sample size: $n = 100$.
  - Dimension: $p = 2$.
  - Data matrix: X of size 100×2.
  - First thing to do: Center the data, i.e., from X, obtain the centered data matrix $\widetilde{X}$.

Principal component analysis

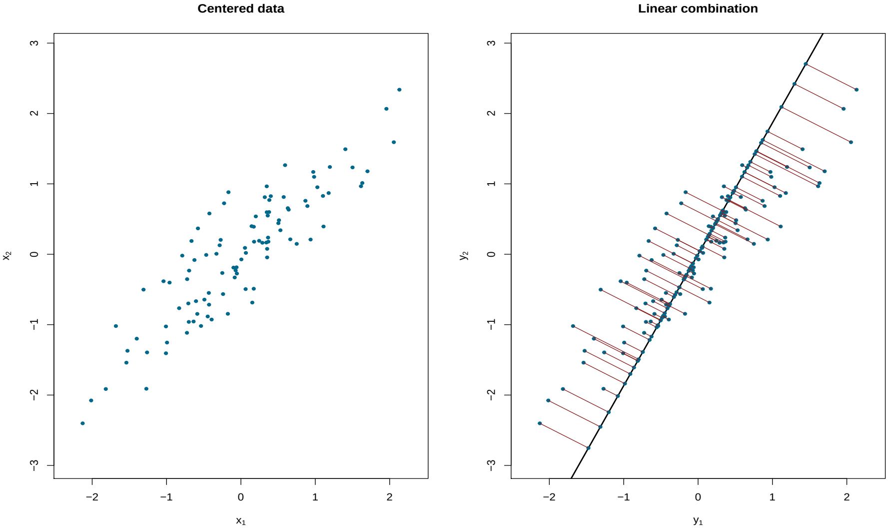




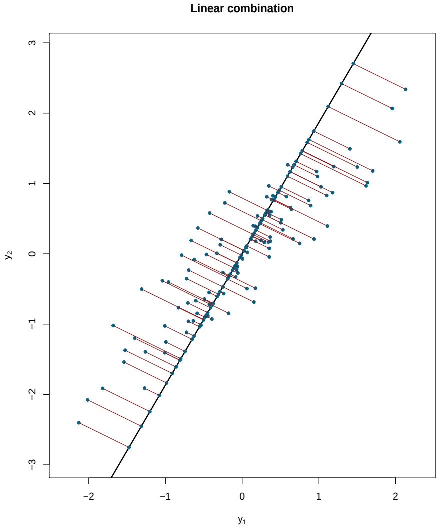
Principal component analysis

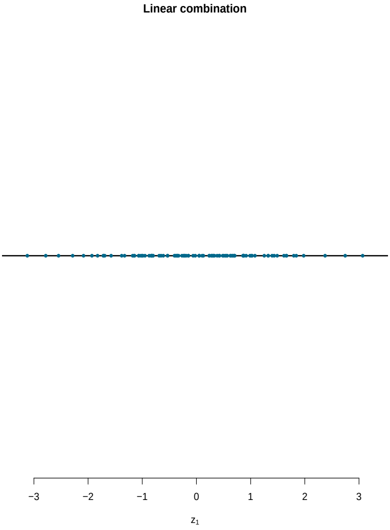
- Find: $Z = \widetilde{X} C$, where $C = (C_1, C_2)'$ of size 100×1.
- What is Z from a geometrical point of view?
- Idea: Project orthogonally the points in $\widetilde{X}$ into the straight line with slope given by $\frac{C_2}{C_1}$.
- Then: The points in Z are the points obtained after rotating this line (and thus the projected points) to the horizontal axe.

Principal component analysis



Principal component analysis

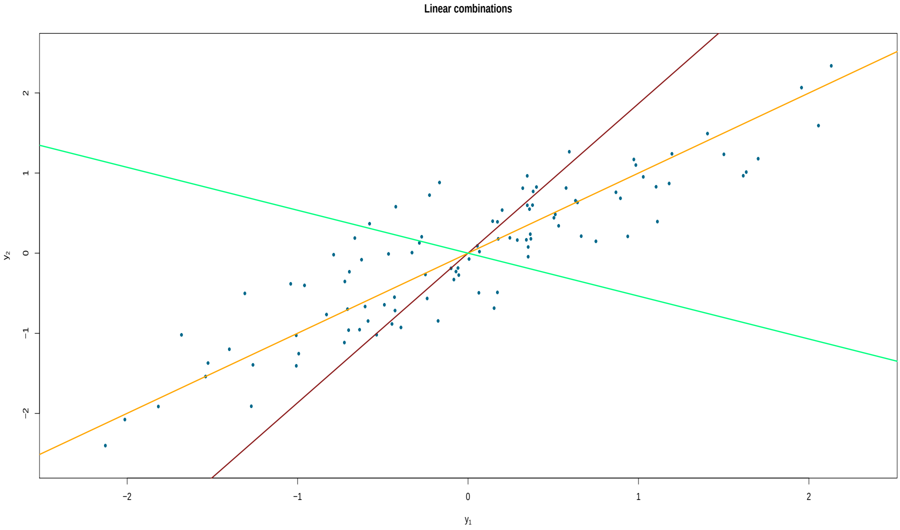




Principal component analysis

- The question is: Which vector $C = (C_1, C_2)'$ contains the most important features in X?
- See several possibilities in the next slide.
- Which one is the best option?

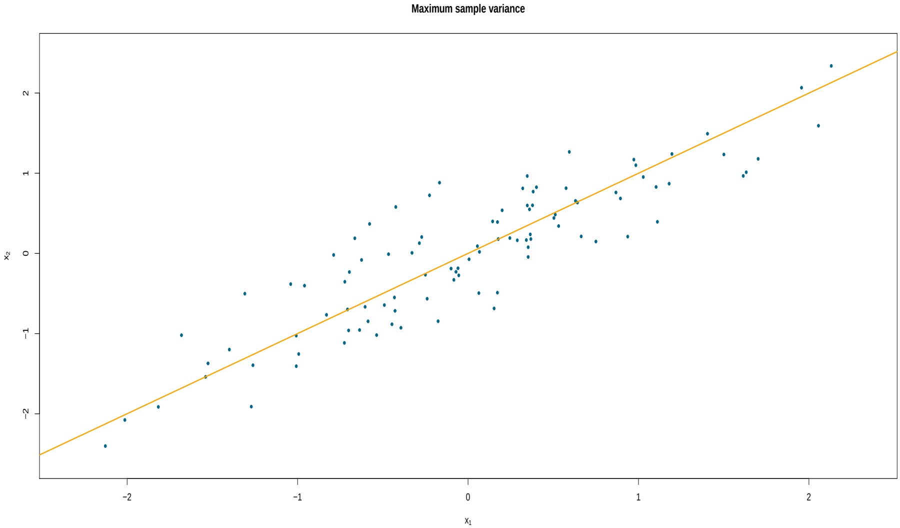
Principal component analysis



Principal component analysis

- PCA: C is the vector that maximizes the sample variance of the projected data.
- Problem: How to get such linear combination in practice?

Principal component analysis



Principal component analysis

- First principal component: $Z = \widetilde{X} C_1$ such that Z has maximum sample variance.
- Sample variance of Z: $s_Z^2 = C_1' S_x C_1$, where S_x is the sample covariance matrix of X.
- However: $C_1' S_x C_1$ can be increased by multiplying C_1 with any constant larger than 1.
- Eliminate this indeterminacy: Restrict attention to coefficient vector of unit length, i.e., assume that $C_1' C_1 = 1$.
- Then: First PC corresponds to the linear combination, C_1, that solves:

$$\underset{s.t.\ C_1' C_1 = 1}{\arg\max}\ C_1' S_x C_1$$

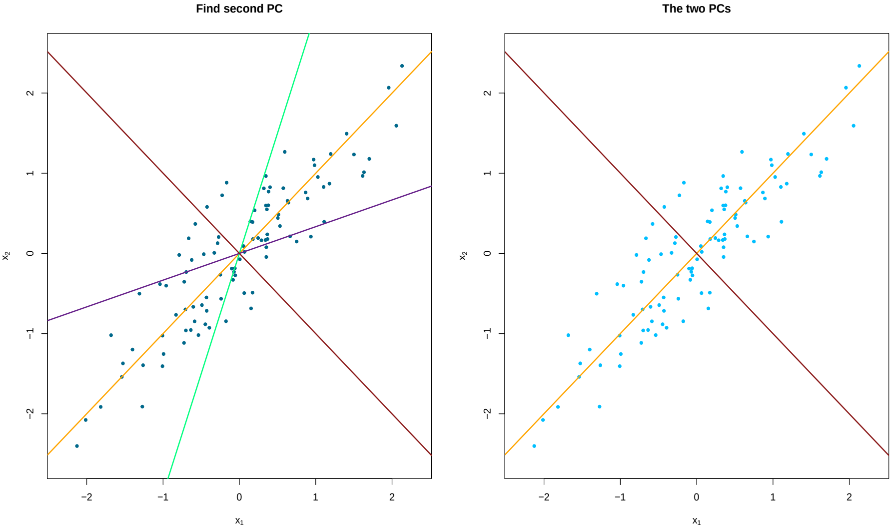
Principal component analysis

- Remember: S_x is a positive semi-definite matrix.
- Thus: S_x has p positive eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ with associated eigenvectors $v_1, \ldots, v_p$, such that, $S_x v_j = \lambda_j v_j$, for $j = 1, \ldots, p$.
- Solution to the optimization problem: C_1 is the eigenvector of S_x, v_1, associated with the largest eigenvalue, λ_1.
- First PC: $Z = \widetilde{X} v_1$.
- Sample variance of Z: $s_Z^2 = v_1' S_x v_1 = \lambda_1 v_1' v_1 = \lambda_1$
- In other words: The sample variance of the first PC is the largest eigenvalue of S_x, λ_1.

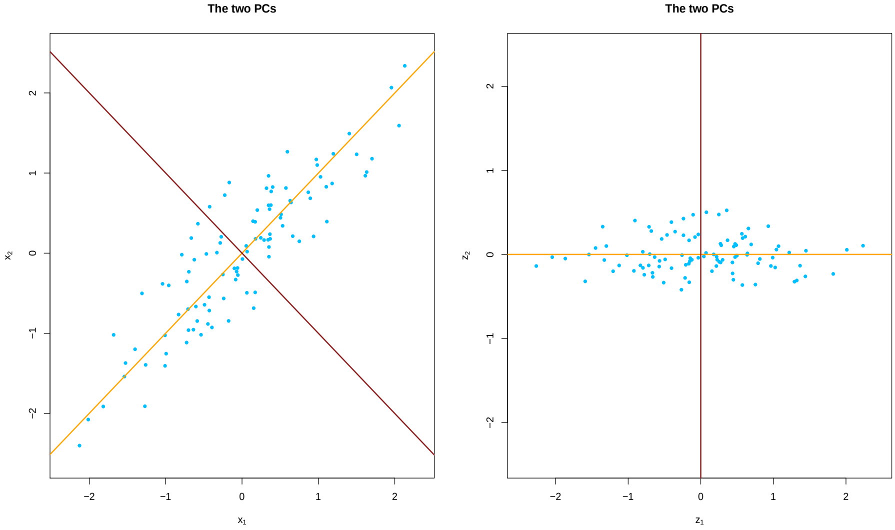
Principal component analysis

- Assume we want $r = 2$: $Z = \widetilde{X} C$, where C is a $p \times 2$ matrix.
- First PC: First column of Z is $Z_{\cdot 1} = \widetilde{X} v_1$.
- Second PC: Second column of Z is $Z_{\cdot 2} = \widetilde{X} C_2$.
- How to define C_2?
- See several possibilities in the next slide.
- Which one is the best option?

Principal component analysis



Principal component analysis



Principal component analysis

- Thus: The second PC is obtained with a similar argument adding the property that it is uncorrelated with the first PC.
- Why?: The new variables do not share common information.
- Second principal component: $Z_{\cdot 2} = \widetilde{X} C_2$ such that $Z_{\cdot 2}$ has maximum sample variance and it is uncorrelated to $Z_{\cdot 1}$.
- Sample variance of $Z_{\cdot 2}$: $s^2_{Z_{\cdot 2}} = C_2' S_x C_2$.
- Then: Second PC corresponds to the linear combination, C_2, that solves:

$$\underset{s.t.\ C_2' C_2 = 1,\ C_1' S_x C_2 = 0}{\arg\max} \quad C_2' S_x C_2$$

Principal component analysis

- Solution to the optimization problem: C_2 is the eigenvector of S_x, v_2, associated with the second largest eigenvalue, λ_2.
- Second PC: $Z_{\cdot 2} = \widetilde{X} v_2$.
- Sample variance of $Z_{\cdot 2}$: $s^2_{Z_{\cdot 2}} = v_2' S_x v_2 = \lambda_2 v_2' v_2 = \lambda_2$
- In other words: The sample variance of the second PC is the second largest eigenvalue of S_x, λ_2.

Principal component analysis

- More PCs: This argument can be extended for successive principal components.
- Assume we want r PCs: Define $V_r = [v_1|\dots|v_r]$ with columns the eigenvectors of S_x linked to the r largest eigenvalues $\lambda_1, \dots, \lambda_r$.
- Then: The r first PCs are given by the $n \times r$ matrix:

$$Z = \widetilde{X} V_r$$

- Characteristics of Z:
  1. Sample mean vector of Z: $\overline{z} = 0_r$.
  2. Sample covariance matrix of Z: S_z, is the diagonal matrix with elements $\lambda_1, \dots, \lambda_r$.
- PC scores: The observations in Z are usually called PC scores.

Principal component analysis

- Indeed: It is possible to take $r = p$, as in the two dimensional data set of the example.
- Total variability of X:

$$Tr\left(S_x\right) = \sum_{j=1}^{p} s_{x_j}^2$$

- Total variability of $Z = \widetilde{X} V_p$:

$$Tr\left(S_z\right) = \sum_{j=1}^{p} \lambda_j$$

- Total variability of X is preserved after a PCA transformation:

$$Tr\left(S_x\right) = Tr\left(S_z\right)$$

Principal component analysis

- Different units of measurement: X should be standardized first.
- Why?: Variables with large sample variances (due to the effect of the units of measurement) will tend to dominate the early components.
- Consequence: First, obtain $Y = \widetilde{X} D_x^{-1/2}$, where D_x is the diagonal matrix that contains the sample variances of the variables in X, and then, obtain PCs.
- Sample covariance of Y is the sample correlation of X:

$$S_y = D_x^{-1/2} S_x D_x^{-1/2} = R_x$$

- Therefore: The PCs should be constructed with the eigenvectors of R_x.

Principal component analysis

- How many PCs to select?
- Proportion of variability explained by r-th PC:

$$PV_r = \frac{\lambda_r}{\lambda_1 + \cdots + \lambda_p} \qquad r = 1, \ldots, p$$

where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of either S_x, or R_x.

- Accumulated proportion of variability explained by the first r PCs:

$$APV_r = \frac{\lambda_1 + \cdots + \lambda_r}{\lambda_1 + \cdots + \lambda_p} \qquad r = 1, \ldots, p$$

- Select r: APV_r larger than a certain quantity, such as 0.7, 0.8 or 0.9.
- Take into account: Trade off between APV_r and the number of PCs selected.

Principal component analysis

- Chapter 2.R script:
  - PCA: NCI60 data set.
  - PCA: College data set.
  - Detect outliers after a PCA: College data set.

Sparse principal component analysis

- Non-zero weights: As can be seen in the College data set, the PCAs are usually constructed with weights that are non-zero.
- All the variables contribute to all the PCs: This can be problematic when the number of variables is large.
- Two main reasons:
  1. Interpretation can be difficult.
  2. Estimation of eigenvectors can underweight important variables.

Sparse principal component analysis

- Sparse principal components: PCs with many weights forced to be 0.
- First sparse PC: Solve the following optimization problem:

$$\underset{s.t.\ C_1' C_1 = 1,\ \|C_1\|_1 \leq k}{\arg\max} \quad C_1' S_x C_1$$

  where $\|C_1\|_1 = \sum_{j=1}^{p} |C_{1j}| \leq k$, and k is an integer number.
- The number k: Controls the number of weights that are different than 0.
- First sparse PC: No closed form solution, say w_1.

Sparse principal component analysis

- Second sparse principal component: Solve the following optimization problem:

$$\underset{s.t.\ C_2' C_2=1,\ w_1' S_x C_2=0,\ \|C_2\|_1 \leq k}{\arg\max} C_2' S_x C_2$$

where $\|C_2\|_1 = \sum_{j=1}^{p} |C_{2j}| \leq k$, and k is an integer number (the same used before).

- Second sparse PC: No closed form solution, say w_2.

- Others: Follow the same arguments to get the p sparse principal components, say $w_1, w_2, \ldots, w_p$.

Sparse principal component analysis

- Complex optimization procedures: Resolution of the optimization problems is quite hard.
- Non-orthogonal scores: Usually, the solution obtained in general does not provide with orthogonal scores.
- Nevertheless: Sample correlations between sparse PCs are usually small.

Sparse principal component analysis

- Chapter 2.R script:
  - SPCA: College data set.

Independent component analysis

- PCA: Given X obtain $Z = \widetilde{X}C$, where Z of size $n \times r$ with $r < p$, contains uncorrelated variables.
- ICA: Given X obtain $Z = \widetilde{X}C$, where Z of size $n \times r$ with $r < p$, contains independent variables.
- Mathematical complexity: ICA is much more mathematically challenging than PCA, which is only based on eigenvectors and eigenvalues.
- Idea: Maximize the statistical independence of the independent component scores in Z by maximizing the non Gaussianity of the components of Z.
- Non-Gaussianity: Measured using a concept of the information theory called entropy.
- Entropy: A complex measure that depends on the joint density function of the variables in Z.

Independent component analysis

- Standardization: ICA always standardizes the variables in X.
- New variables: Z have sample mean vector 0_r and sample covariance matrix I_r (at least, it is expected).
- Consequence: The ICs have the same importance in Z.
- Fix r: It is necessary to fix r in advance.
- Role of r: Different values of r give different ICs.

Independent component analysis

- Chapter 2.R script:
  - ICA: College data set.