

Dimension Reduction

About PCA and FA

Danah Kim

Department of Applied Statistics
Yonsei University

DataScience Lab at Yonsei, April 5, 2019

Table of Contents

- 1 Introduction
- 2 Principal Component Analysis
- 3 Factor Analysis
- 4 More about PCA

Table of Contents

- 1 Introduction
- 2 Principal Component Analysis
- 3 Factor Analysis
- 4 More about PCA

What is Dimension in data?

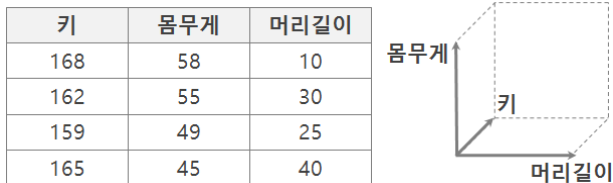
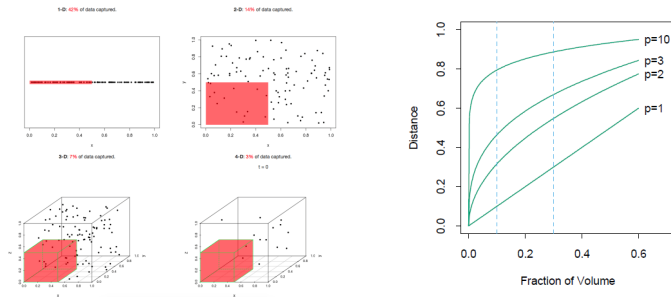


Figure: data in table

- Simply, Dimension = the number of variables.
- It's not always orthogonal in real data.

The Curse of Dimensionality



- Especially in machine learning and data mining, we call 'The Curse of Dimensionality', which means as the dimensionality increases, the size of space increases exponentially.
- $0.1^1 = 0.1$ vs $0.8^{10} = 0.1$

The Curse of Dimensionality

For instance, MNIST, one of the most well-known dataset for machine learning, has $28 \times 28 = 784$ dimensions.

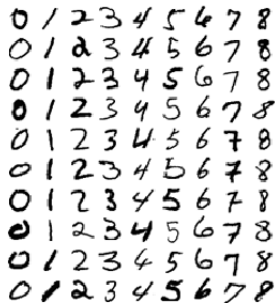
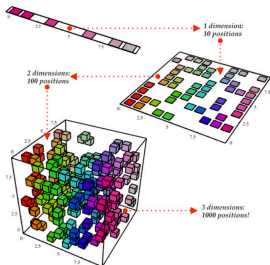


Figure: MNIST dataset

Dimension Reduction

- Dimension Reduction is the process of reducing the dimensions of data without losing much of information.
- We can reduce the number of dimensions(=features=variables) for those reasons :
 - to avoid the curse of dimensionality.
 - simplification of models to make them easy to interpret.
 - to shorter training times.
 - to enhance performance by reducing overfitting.



① Feature Selection

- Also known as variable selection. Select the subset of relevant features for use in model.
- Ex) Stepwise selection (forward/backward) , Lasso

② Feature Extraction

- Create a new feature with a combination of the original features.
- Ex) Principal Component Analysis, Linear Discriminant Analysis

Table of Contents

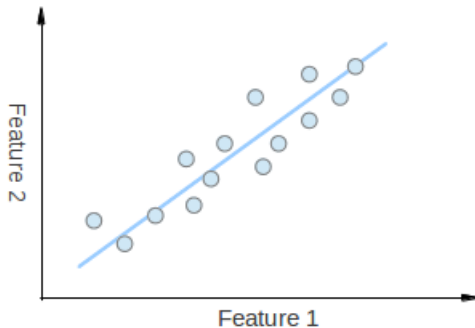
- 1 Introduction
- 2 Principal Component Analysis**
- 3 Factor Analysis
- 4 More about PCA

What is PCA?

- Principal Component Analysis(PCA) transforms a set of correlated response variables into a smaller set of **uncorrelated** variables called principal components.
- PCA seeks the linear combinations of the original variables such that the derived variables capture **maximal variance**.
- PCA can solve the multicollinearity.
- Idea
 - A few small principal components may contain almost all of the information that was available in the original data.
- Goal
 - Reduce the dimensionality of the data and discover the true dimensionality of the data.
 - identify new meaningful underlying variables

Basic Idea of PCA

- We can describe data with other axis which is a linear combination of original variables to reduce the dimension.
- What is the axis that contains the most information? Height or Gender?



Linear Combination

Assume centered data

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots \\ x_{21} & x_{22} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} = (X_1 \ X_2 \ \dots \ X_p)$$

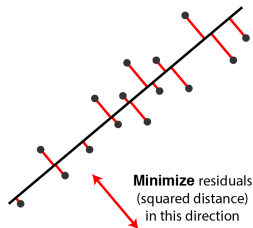
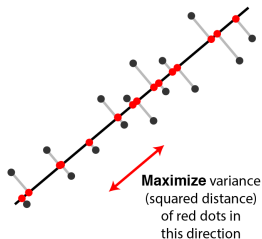
Consider the linear combination \mathbf{Y} instead of \mathbf{X} ,

$$\begin{cases} Y_1 = \mathbf{a}_1^T \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p \\ \vdots \\ Y_p = \mathbf{a}_p^T \mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p \end{cases}$$

Then,

$$\begin{aligned} \max(\text{Var}(Y_i)) &= \lambda_i : \text{eigen value} \\ \text{when } a_i &= v_i : \text{eigen vector} \end{aligned}$$

PCA on Covariance



$\mathbf{X}^T\mathbf{X} = \Sigma$: Covariance matrix

By Spectral Decomposition,

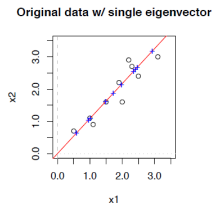
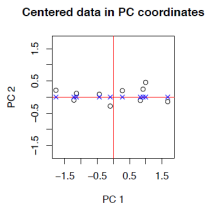
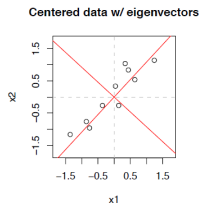
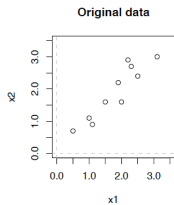
$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{V}^T$ where $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$ with $\mathbf{V} = \{v_1, v_2, \dots, v_p\}$

(eigenvectors) and $\mathbf{D} = \mathit{diag}\{\lambda_1, \lambda_2, \dots, \lambda_p\}$ with

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$. (eigenvalues)

$\mathit{Var}(a_i^T X) = a_i^T \mathit{Var}(X) a_i$ is maximized when $a_i = v_i$

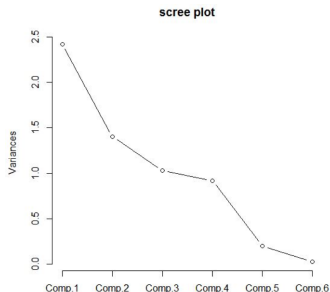
Example of PCA



$$PC1 = v_1^T X$$

$$PC2 = v_2^T X$$

Scree Plot



Determine the number of PC
Using eigenvalue and cumulative proportions of explained variance or an elbow in the plot.

The proportion of total variance due to the k -th principal component

$$= \frac{\lambda_k}{\sum_{i=1}^p \lambda_i}$$

The variables have much different variances, then standardize the data X or apply PCA on the correlation matrix.

- Result
 1. PC sequentially capture the maximum variability among the columns of \mathbf{X} , thus guaranteeing minimal information loss.
 2. PC are uncorrelated, so we can talk about one PC without referring to others.
- Drawback
 - Often difficult to interpret p variables and the derived PCs.

Table of Contents

- 1 Introduction
- 2 Principal Component Analysis
- 3 Factor Analysis**
- 4 More about PCA

What is Factor Analysis?

Factor Analysis reduces a p -dimensional random vector X into the fewer k latent variables, which is factor.

- Goal of FA
 1. Partition the p response variables into k subsets, each consisting of a group of variables tending to be more highly related to others.
 2. help understand the characteristics of data.
 3. Create a new set of uncorrelated variables, called 'underlying factors' or 'underlying characteristics'.
 4. Use the new variables in future analysis.

How to do Factor Analysis

1. Choose the appropriate number of Factors. Use scree plot.
 2. Rotate factor matrix.
 3. Select variables that consist of the factor using the factor loadings.
- Method of Estimation
 1. Principal Component Method
 2. Principal Factor Method
 3. Maximal Likelihood Method

Orthogonal Factor Model

Let p -dimensional random vector $x^T = (x_1, x_2, \dots, x_p)$ with $E(X) = \mu$ and $Cov(X) = \Sigma$.

The model is,

$$\begin{cases} X_1 - \mu_1 = q_{11}f_1 + q_{12}f_2 + \dots + q_{1k}f_k + \epsilon_1 \\ \vdots \\ X_p - \mu_p = q_{p1}f_1 + q_{p2}f_2 + \dots + q_{pk}f_k + \epsilon_p \end{cases}$$

Then,

$$\mathbf{X} = \mathbf{QF} + \mu + \epsilon$$

where F = the k -dimensional vector of the k common factors ($k < p$), and $E(F) = \mathbf{0}$ and $Cov(F) = I_k$ and ϵ called specific factors with $E(\epsilon) = \mathbf{0}$ and $Cov(\epsilon) = \psi$.

Orthogonal Factor Model

Let's consider the last $p-k$ eigenvalues equal to zero ;

$$\lambda_1 \geq \dots \geq \lambda_k \geq \lambda_{k+1} = \dots = \lambda_p = 0.$$

$$\Sigma = VDV^T = V_1D_1V_1^T$$

We can obtain the approximation

$$\Sigma \approx QQ^T + \Psi$$

$$X^T X = V_1D_1V_1^T = V_1D_1^{1/2}D_1^{1/2}V_1^T$$

$$\therefore X = V_1D_1^{1/2}$$

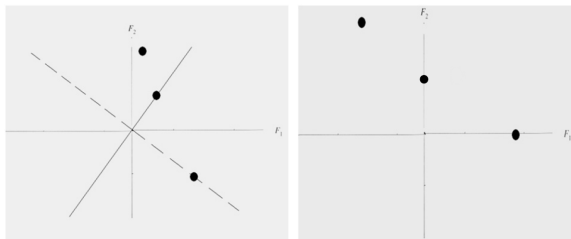
The communality ;

$$\text{Var}(X_j) = \sum_{l=1}^k q_{jl}^2 + \psi_{jj}$$

where $\sum_{l=1}^k q_{jl}^2$: communality. ψ_{jj} : specific variance

Rotated Factor Matrix

- The most commonly used method is Varimax Rotation.
- Maximizing V corresponding to “spreading out” the squares of loadings on each factor as much as possible.



Example of Factor Analysis

새로운 제품의 여러 가지 특성(변수)들이 소비자의 선호도와 관련하여 어떤 요인들을 형성하는지를 알려고 100명의 소비자를 대상으로 조사를 하였다. 선호도 정도에 따라 7점까지 점수를 준 설문 응답을 정리한 후 특성들에 대한 상관관계 행렬을 구하였더니 다음과 같았다.

- Correlation Matrix

특성(변수)		1	2	3	4	5
맛	1	1	.02	.96	.42	.01
가격	2	.02	1	.13	.71	.85
향기	3	.96	.13	1	.50	.11
적당한 요기거리	4	.42	.71	.50	1	.79
영양가	5	.01	.85	.11	.79	1

- group

(variable 1, 3) → taste

(variable 2, 4, 5) → cost-effectiveness

Example of Factor Analysis

Factor Loading

- Choose 2 Factors after looking the scree plot.
- Used a Principal Component Method
- Before rotating factor matrix

변수	Factor1	Factor2	커뮤니컬리티
1. 맛	-0.560	-0.816	0.979
2. 가격	-0.777	0.524	0.879
3. 향기	-0.645	-0.748	0.976
4. 요깃거리	-0.939	0.105	0.893
5. 영양가	-0.798	0.543	0.932
Variance	2.8531	1.8063	4.6594
% Var	0.571	0.361	0.932

Example of Factor Analysis

Factor Loading

- After rotating factor matrix

변수	Factor1	Factor2	커뮤니티
1. 맛	0.020	0.989	0.979
2. 가격	0.937	-0.011	0.879
3. 향기	0.129	0.979	0.976
4. 요깃거리	0.842	0.428	0.893
5. 영양가	0.965	-0.016	0.932
Variance	2.5374	2.122	4.6594
% Var	0.507	0.424	0.932

Table of Contents

- 1 Introduction
- 2 Principal Component Analysis
- 3 Factor Analysis
- 4 More about PCA

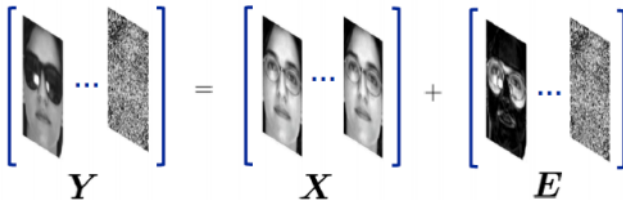
- It extends the classic method of PCA for the reduction of dimensionality of data by adding **sparsity constraint** on the input variables.
- When $\dim \rightarrow \infty$, PCA can be inconsistent and hard to interpret. Sparse PCA offer dimension reduction and variable selections simultaneously.
- Many methods, Lasso, Ridge, Elastic net, Change the L-norm.

$$\min_{U, V} \|X - UV\|_2^2 + \alpha \|V\|_1$$

$$\text{subject to } \|U_k\|_2 = 1 \quad \text{for all } 0 \leq k < n$$

Sparse Data												
Day	Sensor 1	Sensor 2	Sensor 3	Sensor 4	Sensor 5	Sensor 6	Sensor 7	Sensor 8	Sensor 9	Sensor 10	Sensor 11	Sensor 12
1-Jan	0	0.89	0	0	0	0	0	0	0	0	0	0.911
2-Jan	0	0	0	0	0	0	0	0	0	0	0	0.931
3-Jan	0	0.951	0	0	0	0	0	0	0	0	0	0.951
4-Jan	0.954	0.911	0	0	0	0	0	0	0	0	0	0.899
5-Jan	0	0	0	0	0	0	0	0	0	0	0	0.897
6-Jan	0	0.899	0	0	0	0	0	0	0	0	0	0.968
7-Jan	0.895	0.911	0	0	0	0	0	0	0	0	0	0.991
8-Jan	0.911	0.962	0	0	0	0	0	0	0	0	0	0.951
9-Jan	0	0.954	0	0	0	0	0	0	0	0	0	0.898
10-Jan	0.896	0.934	0	0	0	0	0	0	0	0	0	0.962

- Robust PCA is analogous to traditional PCA but instead of recovering a low rank approximation of the matrix under some Gaussian noise assumption, it decomposes it as the sum of of a **low rank matrix** and a **sparse** one.


$$Y = X + E$$

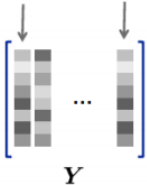
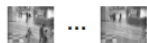
where unknown X is low-rank and E is sparse error.

Robust PCA

Static camera
surveillance video

200 frames,
144 x 172 pixels,

Significant foreground
motion



$$\text{Video } Y = \text{Low-rank approx. } X + \text{Sparse error } E$$



Figure: Background Modeling from video

Please visit the website below and try to do Robust PCA using image data.

<http://jeankossaifi.com/blog/rpca.html>

- Multiple correspondence analysis (MCA) is a data analysis technique for **nominal categorical data**, used to detect and represent underlying structures in a data set.
- It can also be seen as a generalization of principal component analysis when the variables to be analyzed are categorical instead of quantitative
- It does this by representing data as points in a low-dimensional Euclidean space. The procedure thus appears to be the counterpart of principal component analysis for categorical data.

- Function PCAmix for principal component analysis (PCA) of mixed data.
- Categorical + Continuous data

- [1] The Elements of Statistical Learning - Trevor Hastie, Robert Tibshirani, Jerome Friedman 2nd edition
- [2] Applied Multivariate Statistical Analysis – Hardle, Simar 3rd edition
- [3] Undergraduate Regression Analysis Class Note – prof. Jeon Yong Ho
- [4] Graduate Mutivariate Analysis Class Note – prof. Kim Hyun Jung