Afi Escuela de Finanzas

Análisis Multivariante

Máster en Data Science y Big Data en Finanzas

Javier Nogales – PhD Matemáticas
Catedrático, Estadística e IO UC3M
www.est.uc3m.es/nogales @fjnogales 2022

Organization

- Subject organized in 3 topics
- 12 hours in total: 3 sessions
- Practical course: 50% basic concepts + 50% computer labs (using R)
- Evaluation: 100% final exercise

Objectives

- Capacity to describe and interpret multivariate datasets
- Understand properties of multivariate distributions and make inference
- Find and analyze relationships between many variables
- Reduce the dimensionality of a dataset and identify real associations
- Handle the R language for multivariate data analysis

Outline

Some Complementary References

- Paul Newbold. Statistics for Business and Economics. 2012
- Michael Barrow. Statistics for Economics Accounting and Business Studies. 2010
- Härdle, W. K. and Simar, L. Applied Multivariate Statistical Analysis. 2007 Springer.
- Richard A. Johnson and Dean W. Wichern. Applied multivariate statistical analysis. 2007
- T. W. Anderson. An Introduction to Multivariate Statistical Analysis. 2009

1. Descriptive Analysis and Inference

Introduction

A typical dataset in Statistics and Machine Learning:

FTHG	FTAG	FTR	HTHG	HTAG	HTR	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR	B365H	B365D	B365A	BWH	BWD	BWA	IWH	IWD	IWA	LBH	LBD	LBA	PSH	PSD	PSA	WHH	WHD	WHA	SJH	SJD	SJA
2	0	H	1	0	H	16	15	6	2	13	6	6	5	1	1	0	0	1,73	3,6	4,75	1,72	3,6	4,75	1,7	3,6	4,7	1,66	3,6	5	1,79	3,74	5,12	1,75	3,75	4,5	1,67	3,6	5,5
1	0	H	0	0	D	9	11	1	2	15	23	9	6	3	5	0	0	1,53	4	6	1,57	4	5,5	1,55	3,9	5,6	1,57	3,75	6	1,56	4,33	6,78	1,55	4	6	1,57	4	6
1	2	A	1	1	D	8	13	2	3	10	8	5	5	1	0	0	0	2,5	3,3	2,8	2,6	3,3	2,65	2,4	3,3	2,75	2,6	3,3	2,6	2,69	3,35	2,86	2,5	3,3	2,8	2,63	3,25	2,75
7	0	H	6	0	H	22	4	13	1	15	16	9	3	1	3	0	0	1,08	10	26	1,08	10,5	23	1,1	8	20	1,08	10	23	1,09	14	30	1,08	11	23	1,09	9,5	29
1	2	A	0	2	A	14	13	5	4	15	17	7	6	1	4	0	0	2	3,3	3,75	2	3,3	3,8	2	3,3	3,6	2	3,3	3,5	2,13	3,37	3,95	2,15	3,25	3,5	2	3,4	3,8
2	1	H	1	1	D	20	11	9	4	11	20	5	7	1	2	0	0	1,17	7	17	1,16	7,75	13,5	1,17	6,5	14	1,15	6,5	15	1,16	9	19,3	1,17	7,5	15	1,17	7	17
1	3	A	1	1	D	14	16	5	6	12	13	1	9	4	4	0	0	2,8	3,3	2,5	2,7	3,4	2,5	2,6	3,2	2,6	2,75	3,3	2,5	2,89	3,39	2,64	2,8	3,3	2,5	2,63	3,4	2,63
2	3	A	1	0	H	15	14	2	4	17	14	6	10	2	0	0	0	2,6	3,2	2,75	2,55	3,25	2,75	2,5	3,3	2,65	2,37	3,2	2,87	2,7	3,32	2,87	2,5	3,2	2,9	2,5	3,25	2,88
2	2	D	1	0	H	15	6	10	5	23	12	7	2	1	4	0	0	2,2	3,2	3,4	2,15	3,4	3,3	2,2	3,3	3,1	2,2	3,3	3,25	2,23	3,33	3,69	2,15	3,25	3,5	2,2	3,3	3,4
3	0	H	2	0	H	11	8	5	1	16	11	3	4	2	2	0	0	2,25	3,25	3,25	2,2	3,3	3,25	2,1	3,3	3,3	2,2	3,3	3,3	2,31	3,42	3,39	2,25	3,3	3,2	2,25	3,25	3,3
2	0	H	1	0	H	8	18	2	5	15	20	5	10	2	2	0	0	1,62	3,75	5,5	1,62	3,7	5,75	1,7	3,6	4,7	1,66	3,6	5,5	1,65	4,14	5,82	1,67	3,6	5,5	1,62	4	5,5
2	2	D	1	2	A	14	16	5	7	15	17	2	7	2	1	0	0	1,85	3,5	4,2	1,8	3,4	4,6	1,85	3,45	4	1,85	3,5	4,2	1,91	3,63	4,54	1,83	3,5	4,4	1,83	3,6	4,33
1	1	D	1	0	H	12	11	2	4	9	12	4	2	4	3	0	0	3,2	3,4	2,2	3,4	3,3	2,15	3	3,3	2,25	3,2	3,3	2,25	3,32	3,47	2,27	3,2	3,3	2,25	3,13	3,5	2,25
3	1	H	1	1	D	13	8	3	2	17	12	6	3	3	3	0	0	3	3,25	2,38	3,1	3,3	2,25	2,9	3,3	2,3	3,1	3,3	2,3	3,18	3,43	2,41	3,2	3,3	2,25	3	3,4	2,38
2	1	H	1	1	D	17	6	6	2	11	25	8	3	4	2	0	0	1,85	3,6	4	1,85	3,4	4,33	1,85	3,45	4	1,9	3,4	4	1,88	3,7	4,6	1,85	3,5	4,33	1,83	3,5	4,5
5	0	H	3	0	H	14	7	5	1	11	9	8	3	0	1	0	0	1,29	5,25	10	1,34	4,75	9,25	1,35	4,8	7,6	1,33	5	9	1,32	5,6	11,58	1,33	5	9	1,3	5,25	11
1	2	A	0	0	D	20	13	6	4	9	18	12	8	1	3	0	1	1,8	3,5	4,5	1,72	3,7	4,75	1,7	3,7	4,5	1,7	3,6	5	1,86	3,64	4,78	1,75	3,6	4,75	1,73	3,6	5
0	0	D	0	0	D	8	18	3	3	17	18	5	7	3	5	0	0	3,5	3,3	2,1	3,5	3,25	2,1	2,85	3,3	2,35	3,2	3,25	2,25	3,54	3,45	2,21	3,4	3,3	2,15	3,4	3,3	2,2
0	1	A	0	1	A	9	15	3	11	14	12	2	11	3	2	0	0	13	6	1,22	14	6,25	1,2	10,3	5,5	1,25	12	6	1,22	13,7	6,78	1,25	12	7	1,2	15	6,5	1,2
0	1	A	0	1	A	8	21	3	8	14	10	5	8	4	2	0	0	7,5	5,5	1,33	9,25	5,25	1,3	10	5,2	1,27	9	5,5	1,3	9,11	5,52	1,37	8	5,5	1,33	10	5,5	1,3
2	2	D	2	1	H	9	15	5	9	18	18	4	5	3	5	1	0	2	3,3	3,8	2	3,3	3,8	2,1	3,3	3,3	2,05	3,4	3,5	2,07	3,68	3,79	2,05	3,4	3,6	2,05	3,4	3,8
1	2	A	0	1	A	23	6	8	2	14	17	8	1	4	3	1	0	1,73	3,6	4,75	1,8	3,7	4,2	2	3,3	3,6	1,8	3,5	4,5	1,83	3,74	4,8	1,83	3,75	4	1,83	3,6	4,5
1	1	D	1	0	H	17	3	4	1	9	6	6	5	1	2	0	0	2,2	3,3	3,3	2,2	3,2	3,4	2,1	3,3	3,3	2,2	3,3	3,3	2,23	3,47	3,53	2,25	3,2	3,3	2,15	3,3	3,6
0	3	A	0	2	A	9	9	3	4	15	14	7	8	3	2	0	0	2,5	3,2	2,88	2,5	3,25	2,8	2,5	3,3	2,65	2,45	3,2	2,87	2,65	3,31	2,93	2,62	3,1	2,8	2,6	3,2	2,88
1	0	H	0	0	D	21	16	6	0	19	8	10	5	2	2	0	0	2,1	3,3	3,5	2,1	3,3	3,5	2	3,3	3,6	2,15	3,3	3,4	2,27	3,37	3,53	2,25	3,2	3,3	2,1	3,4	3,6
0	0	D	0	0	D	10	8	3	3	19	14	7	7	2	4	0	0	2,05	3,4	3,6	2	3,5	3,6	1,9	3,45	3,8	2,05	3,4	3,5	2,11	3,58	3,76	2,05	3,4	3,6	2	3,5	3,8
3	1	H	2	0	H	20	12	7	2	15	13	9	6	1	2	0	0	1,17	7	16	1,16	7,75	13,5	1,2	6,5	10,3	1,18	7,5	12	1,18	8,4	17,8	1,17	7,5	15	1,2	7	15
2	2	D	1	1	D	15	15	7	6	17	13	10	1	3	3	0	1	1,62	3,8	5,5	1,57	3,9	5,75	1,65	3,8	4,7	1,66	3,75	5	1,62	4,14	6,27	1,6	4	5,5	1,67	3,75	5,5
1	2	A	0	1	A	10	16	4	9	19	21	5	5	2	4	0	0	2,9	3,3	2,38	2,8	3,25	2,5	2,75	3,3	2,4	2,87	3,4	2,37	2,88	3,51	2,58	2,8	3,3	2,5	2,75	3,4	2,6
2	3	A	2	3	A	16	20	10	9	20	11	5	7	2	3	0	0	5,5	4,33	1,53	6,25	4,4	1,48	5	3,9	1,6	6	4,33	1,5	6,13	4,54	1,57	5,5	4,33	1,55	5,75	4,33	1,57
4	2	H	2	1	H	20	7	8	2	15	17	8	1	2	4	0	0	1,25	6	11	1,26	5,5	11	1,27	5,2	10	1,28	5,5	9,5	1,27	6,32	13,73	1,25	5,5	12	1,25	6,25	12
3	2	H	1	0	H	19	7	10	2	12	17	6	9	2	5	0	0	1,18	7	15	1,18	7	13,5	1,15	7	15	1,16	7,5	13	1,19	8,48	15,71	1,17	7,5	13	1,17	7,5	17

Introduction

A typical dataset in Statistics and Machine Learning:

- Formally, we will have a data matrix with n rows (individuals) and p columns (variables)

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- Sometimes the main interest is in one variable: **supervised learning**
- Other times all the variables are equally important: **unsupervised learning**

Multivariate descriptive analysis

- Multivariate data analysis: tools for simultaneous analysis of data from several variables
- That is, try to extract information about the joint relationships between several variables
- In this way, we can improve predictions and also risk management
- Applications in: Marketing, Banking and Finance, Insurance, Economics, Energy, Sports, ...
- In general, first step to **gain knowledge** from data

Multivariate descriptive analysis

- Example: Socioeconomic data. 11 variables from 194 countries

	Country	Region	Population	Under15	Over60	FertilityRate	LifeExpectancy	ChildMortality	CellularSubscribers	LiteracyRate	GNI	PrimarySchoolEnrollmentMale	PrimarySchoolEnrollmentFemale
1	Afghanistan	Eastern Mediterranean	29825	47.42	3.82	5.40	60	98.5	54.26	*NA*	1140	*NA*	*NA*
2	Albania	Europe	3162	21.33	14.93	1.75	74	16.7	96.39	*NA*	8820	*NA*	*NA*
3	Algeria	Africa	38482	27.42	7.17	2.83	73	20.0	98.99	*NA*	8310	98.2	96.4
4	Andorra	Europe	78	15.20	22.86	*NA*	82	3.2	75.49	*NA*	*NA*	78.4	79.4
5	Angola	Africa	20821	47.58	3.84	6.10	51	163.5	48.38	70.1	5230	93.1	78.2
6	Antigua and Barbuda	Americas	89	25.96	12.35	2.12	75	9.9	196.41	99.0	17900	91.1	84.5
7	Argentina	Americas	41087	24.42	14.97	2.20	76	14.2	134.92	97.8	17130	*NA*	*NA*
8	Armenia	Europe	2969	20.34	14.06	1.74	71	16.4	103.57	99.6	6100	*NA*	*NA*
9	Australia	Western Pacific	23050	18.95	19.46	1.89	82	4.9	108.34	*NA*	38110	96.9	97.5
10	Austria	Europe	8464	14.51	23.52	1.44	81	4.0	154.78	*NA*	42050	*NA*	*NA*
11	Azerbaijan	Europe	9309	22.25	8.24	1.96	71	35.2	108.75	*NA*	8960	85.3	84.1
12	Bahamas	Americas	372	21.62	11.24	1.90	75	16.9	86.06	*NA*	*NA*	*NA*	*NA*
13	Bahrain	Eastern Mediterranean	1318	20.16	3.38	2.12	79	9.6	127.96	91.9	*NA*	*NA*	*NA*
14	Bangladesh	South-East Asia	155000	30.57	6.89	2.24	70	40.9	56.06	56.8	1940	*NA*	*NA*

- Difficult to gain knowledge from here

Multivariate descriptive analysis

- Socioeconomic data: Univariate description

```
> summary(WHO)
             Country                       Region      Population          Under15          Over60      FertilityRate   LifeExpectancy
 Afghanistan       :  1   Africa                :46   Min.   :      1   Min.   :13.12   Min.   : 0.81   Min.   :1.260   Min.   :47.00
 Albania           :  1   Americas              :35   1st Qu.:   1696   1st Qu.:18.72   1st Qu.: 5.20   1st Qu.:1.835   1st Qu.:64.00
 Algeria           :  1   Eastern Mediterranean:22   Median :   7790   Median :28.65   Median : 8.53   Median :2.400   Median :72.50
 Andorra           :  1   Europe                :53   Mean   :  36360   Mean   :28.73   Mean   :11.16   Mean   :2.941   Mean   :70.01
 Angola            :  1   South-East Asia       :11   3rd Qu.:  24535   3rd Qu.:37.75   3rd Qu.:16.69   3rd Qu.:3.905   3rd Qu.:76.00
 Antigua and Barbuda:  1   Western Pacific       :27   Max.   :1390000   Max.   :49.99   Max.   :31.92   Max.   :7.580   Max.   :83.00
 (Other)           :188                                                                                 NA's   :11
 ChildMortality    CellularSubscribers  LiteracyRate        GNI         PrimarySchoolEnrollmentMale PrimarySchoolEnrollmentFemale
 Min.   :  2.200   Min.   :  2.57      Min.   :31.10   Min.   :  340   Min.   : 37.20              Min.   : 32.50
 1st Qu.:  8.425   1st Qu.: 63.57      1st Qu.:71.60   1st Qu.: 2335   1st Qu.: 87.70              1st Qu.: 87.30
 Median : 18.600   Median : 97.75      Median :91.80   Median : 7870   Median : 94.70              Median : 95.10
 Mean   : 36.149   Mean   : 93.64      Mean   :83.71   Mean   :13321   Mean   : 90.85              Mean   : 89.63
 3rd Qu.: 55.975   3rd Qu.:120.81      3rd Qu.:97.85   3rd Qu.:17558   3rd Qu.: 98.10              3rd Qu.: 97.90
 Max.   :181.600   Max.   :196.41      Max.   :99.80   Max.   :86440   Max.   :100.00              Max.   :100.00
                   NA's   :10          NA's   :91      NA's   :32      NA's   :93                  NA's   :93
```

- Insights?

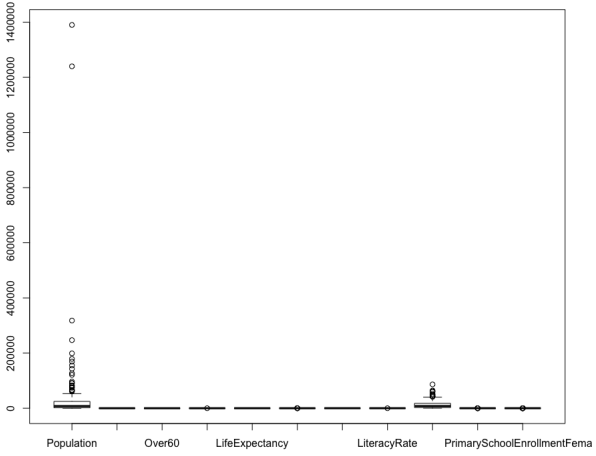
Multivariate descriptive analysis

Some insights, but not enough. Visualize data to get more knowledge:

- skewness
- multimodality
- outliers
- groups

Multivariate descriptive analysis

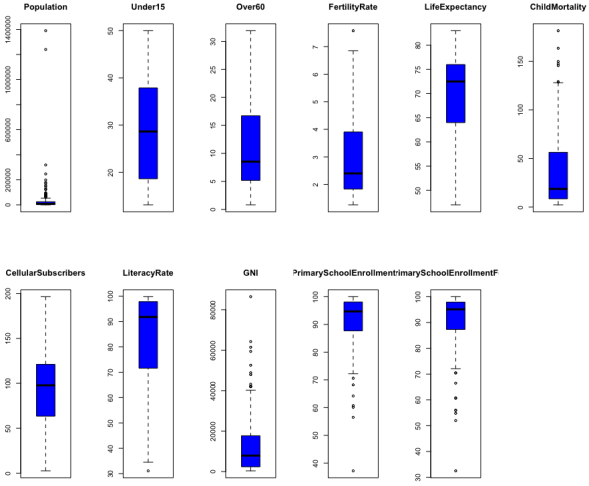
- Socioeconomic data: Multiple box-plots



- What happens?

Multivariate descriptive analysis

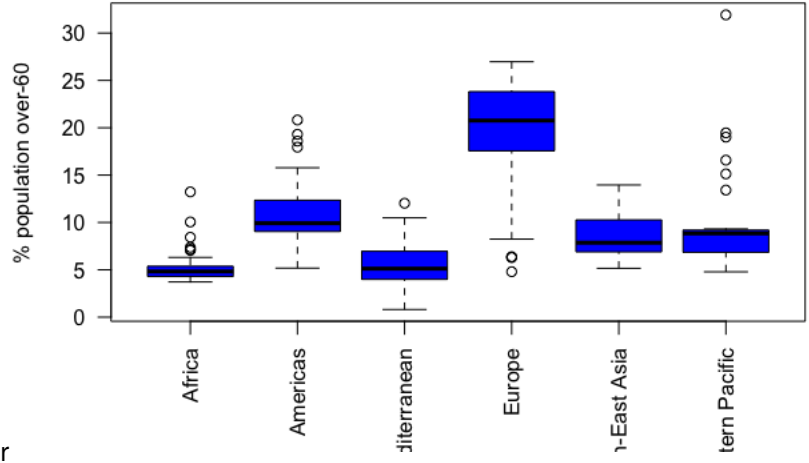
- Socioeconomic data: Multiple box-plots



- A bit better

Multivariate descriptive analysis

- Socioeconomic data: Multiple box-plots



- A bit better

Multivariate descriptive analysis

- Socioeconomic data:
- From previous graphs, we can extract insights for each variable
- But any insight regarding variable relationships?

Multivariate descriptive analysis

Some definitions:

- Remember the usual data matrix (n observations, p variables):

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- This is the main input for most of the analytical tools
- From this matrix, we can compute univariate statistics like the means, medians, standard deviations, proportions, etc.
- But much more...

Multivariate descriptive analysis

- The vector of means is defined as $\bar{x} = \begin{pmatrix} \bar{x}_1 & \cdots & \bar{x}_p \end{pmatrix}^T$

- The covariance matrix:

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

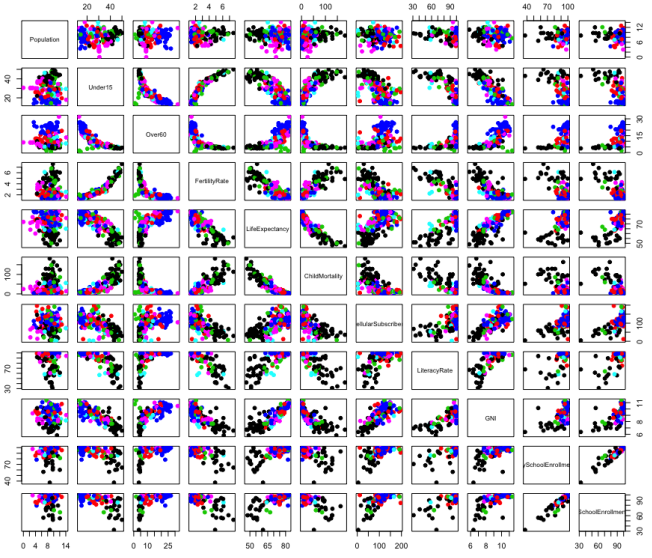
  where $s_{ij} = \frac{1}{n} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$ denotes the sample covariance between variables i and j, and denotes the sample variance of variable i when $i = j$

- Correlation matrix: $R = D^{-1/2} S D^{-1/2}$, where D is a diagonal matrix that contains the variances, s_{ii}, for each variable

- I.e., each component is $r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$ (unit-less)

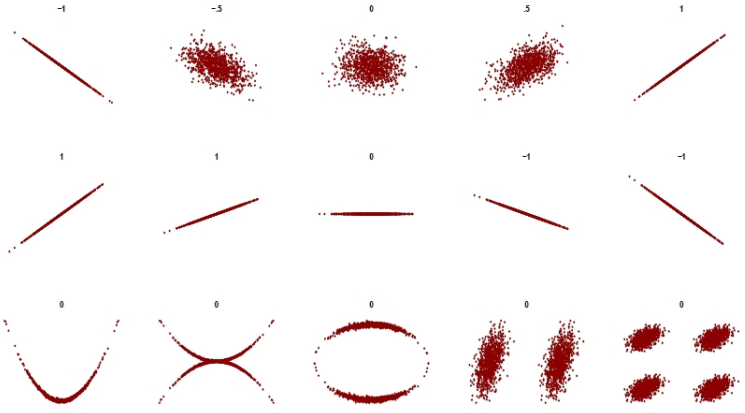
Multivariate descriptive analysis

- Socioeconomic data: Multiple scatter-plots (after log in Population and GNI)



Multivariate descriptive analysis

- Take care with correlations. They just give meaningful information if relations are linear



Multivariate descriptive analysis

- Generalized variance: $|S| = \det S$
  (representation of data volume, i.e. measure of linear dependence)

- Total variance: $\text{trace}(S) = s_{11} + \cdots + s_{pp}$
  (representation of total variability, but without relationships)

- Average variability: $|S|^{1/p}$
  (note $|S|^{1/p} \leq \text{trace}(S)/p$ and takes into account relationships)

- Largest and smallest eigenvalues of S: 6×10^7 and 1.52, respectively. Moreover, $\text{Tr}(S) = 6 \times 10^7$ and $|S| = 7 \times 10^{14}$: heterogeneous data units

- Largest and smallest eigenvalues of R: 4.7 and 0.01, respectively. Moreover, $\text{Tr}(R) = 6$ and $|R| = 0.0001$: highly correlated data

Multivariate descriptive analysis

Some insights:

- Variables are heterogeneous (groups), asymmetric, and contain outliers
- Relations between variables: linear or non-linear? Strong or weak?
- Take care: those relations are pairwise. More insights in high-dimensional relations
- Sample variances are quite different: variables are in different units. Hence, eigenvalues in S are quite extreme
- The correlation matrix, R, is unit independent. If variables are uncorrelated, then $|R| \simeq 1$
  If there exists a variable which is a linear combination of others, then $|R| \simeq 0$
- Eigenvalues and eigenvectors of S or R contain information about high-dimensional relations

Multivariate descriptive analysis

Some questions to answer:

- Are the variables related for countries in a given continent?
- Are the expected values comparable for countries in different continents?
- Can we rank the countries respect to their socio-economic development?
  For instance, the following index can be used to rank:
  index = -Under15 + Over60 - FertilityRate + LifeExpectancy -ChildMortality + GNI
- Can we improve that index?
- Can we reduce the dimension of the data set (to reduce noise)?
- How to classify many individuals into groups?

Multivariate Distributions and Inference

Multivariate Distributions and Inference

- Objective: jump from sample to population (inference)
- To do that, assume a data generating process, GDP: $X \equiv$ model $+$ noise
- Interest in how to estimate unknown population parameters and quantify the associated error
- In the multivariate case, with more information (mainly contained in the covariance matrix)
- Hence, better decisions can be made:
  - by understanding co-movements and make better predictions
  - by making better decisions based on risk (prediction error)
  - by developing indicators, make rankings, risk profiles, etc.

Random vectors

- A multivariate random variable (or vector) consists of p characteristics/features of an item/observation in a population
  - In previous example, different socio-economic variables
- They can be discrete or binary or categorical (but we focus on continuous ones)
- We can analyze each variable individually, or conditionally (respect to some others), or all together (multivariate)

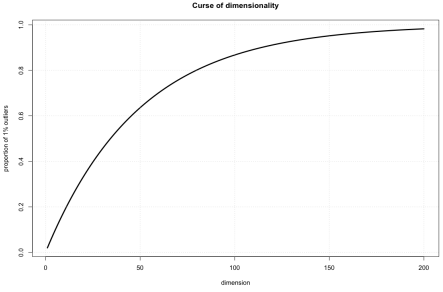
Curse of dimensionality

The complexity grows exponentially with the dimension, p

- As p increases, the volume increases exponentially, making the information scarcer
  (in high dimension, almost all the data are outliers)
- The number of parameters we need to obtain information typically grows quadratically with the dimension
  (ideally, we would need $n/p > 30$ to estimate efficiently the parameters, but this is not the case in practice)
- We can mitigate this problem if variables are highly dependent, or if they are completely independent, or using specific tools
- But we cannot eliminate this problem...

Curse of dimensionality

- As p increases, the volume increases exponentially, making the information scarcer
- To avoid this curse, the sample size n should grow exponentially with p
- In high dimension, almost all the data are outliers ($1 - 0.98^d$, hypercube)



Random vectors

- Expected mean. The expectation or mean vector, μ, of a random vector X is

$$\mu = E(X) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{bmatrix}$$

- Some properties:
  - $E(c^T X) = c^T \mu$
  - $E(CX) = C\mu$
  - $E(X + Y) = E(X) + E(Y) = \mu_X + \mu_y$

Random vectors

- Covariance matrix. It is a symmetric and semi-definite positive matrix measuring pairwise dependence and defined as

$$\Sigma = \text{Cov}(X) = E\big((X - \mu)(X - \mu)^T\big)$$

  Note the diagonal elements contain the variances for each variable, whereas the off-diagonal elements contains the covariances

- The correlation matrix is defined as

$$R = \Delta^{-1/2}\, \Sigma\, \Delta^{-1/2}$$

  where $\Delta = \text{diag}(\Sigma)$

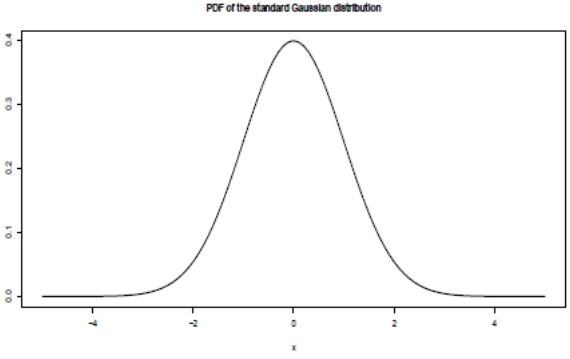
Random vectors

- Some properties
  - $\Sigma \succeq 0$
  - $\text{Var}(c^T X) = c^T \Sigma c$
  - If $Y = AX + b$, then $\mu_y = A\mu_x + b$, and $\Sigma_y = A\Sigma_x A'$
- Generalized variance: $|\Sigma| = \det \Sigma$
- Total variance: $\text{trace}(\Sigma) = \sigma_{11} + \cdots + \sigma_{pp}$
- Average variability: $|\Sigma|^{1/p}$

Multivariate Normal Distribution

- Remember the pdf of a univariate normal distribution, $X \sim \mathcal{N}(\mu, \sigma^2)$:

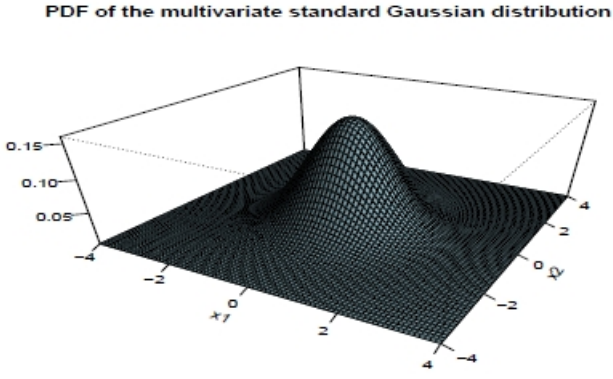
$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Multivariate Normal Distribution

- Now, the pdf of a multivariate normal distribution, $X \sim \mathcal{N}_p(\mu, \Sigma)$:

$$f(x) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)$$



Multivariate Normal: Properties

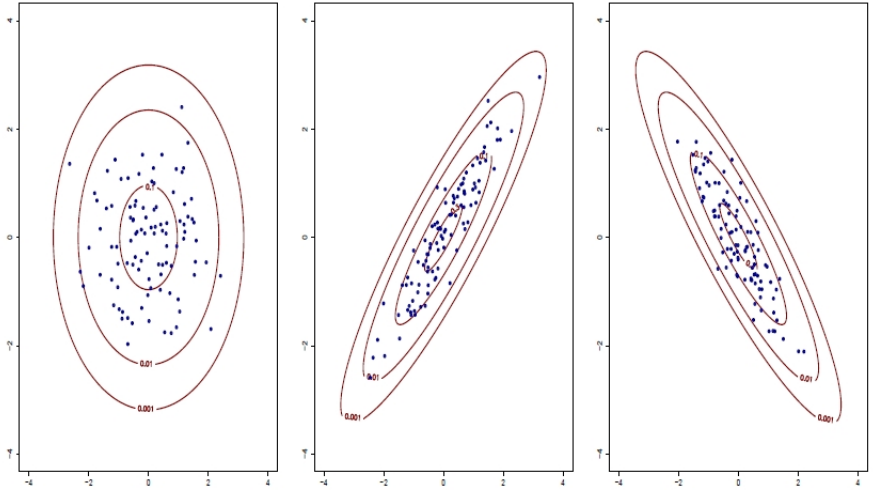
- The distribution is symmetric around μ and it is the unique maximum
- If $X \sim \mathcal{N}_p(\mu, \Sigma)$ and its components are uncorrelated (Σ is diagonal), then they are independent
- Standardization: if $X \sim \mathcal{N}_p(\mu, \Sigma)$, then

$$Z = \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}_p(0, I)$$

- Any marginal distribution is also normal
- Any linear transformation is also normal

Multivariate Normal: Properties

- Level curves (points with equal "probability"): $(x-\mu)^T\Sigma^{-1}(x-\mu) = c$



- They are ellipsoids

Multivariate Normal: Properties

- Level curves allow us to define distances between individuals
- Most common used distance when data follow a multivariate normal:
- Mahalanobis distance from point i to point j:

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)}$$

- If $X \sim \mathcal{N}_p(\mu, \Sigma)$, then

$$d_M^2(X, \mu) = (X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi_p^2$$

Hypothesis testing

- If $X \sim \mathcal{N}_p(\mu, \Sigma)$ and we want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, then (by the likelihood ratio test)

$$\lambda = n \log \frac{|S_0|}{|S|}$$

follows (when n is large) a χ^2_p, where

$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$ and

$S_0 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)(x_i - \mu_0)^T$

Hypothesis testing

- If $X \sim \mathcal{N}_p(\mu, \Sigma)$ and we want to test $H_0 : \Sigma = \Sigma_0$ against $H_1 : \Sigma \neq \Sigma_0$, then (by the likelihood ratio test)

$$\lambda = n \log \frac{|\Sigma_0|}{|S|} + n\text{Trace}(\Sigma_0^{-1} S) - np$$

  follows (when n is large) a $\chi^2_{p(p+1)/2}$

- If $X \sim \mathcal{N}_p(\mu, \Sigma)$ and we want to test $H_0 : \Sigma = \text{diag}$ against $H_1 : \Sigma \neq \text{diag}$, then (by the likelihood ratio test)

$$\lambda = -n \log |R|$$

  follows (when n is large) a $\chi^2_{p(p-1)/2}$

Some final conclusions

- Most of the time, we assume data follows a multivariate normal distribution. But that is almost never true...
- We can check the multivariate normal distribution through the Mahalanobis distance (that follows a χ^2 distribution)
- Anyway, assuming a normal distribution is quite practical!
- It allows us to develop formulas for tools, to compute prediction errors, to make infereces, etc.
- In a large data set, the presence of outliers is quite common. Unlike the univariate case, in the multivariate case it is very difficult to identify them

2. Dimension Reduction: PCA

Introduction

The data matrix X is $n \times p$. What happens if p is large?

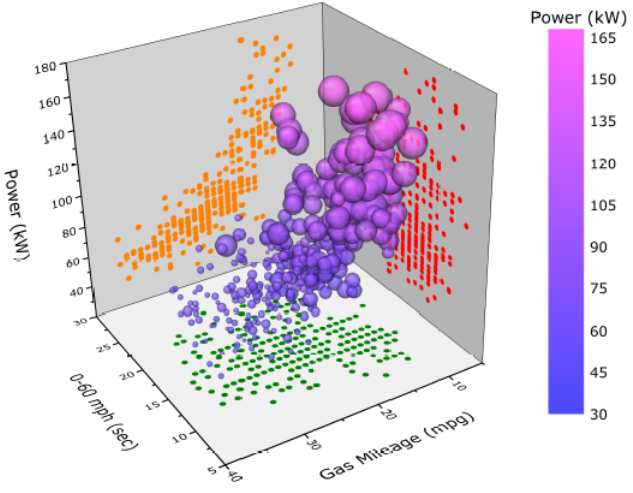
- S&P 500 based portfolios: $n = 1000$ trading days, $p = 500$ stocks. Hence we have $n \times p = 500000$ data points but we need to estimate $p + p(p+1)/2 = 125750$ parameters...
- Similar dimensions may be encountered nowadays in many fields like economics, environmental and medical studies

Curse of dimensionality: if p/n is large, then the analysis may become intractable

Principal Component Analysis

- Objective: represents the multivariate information with a smaller number of variables without loosing much information
- Possible if the variables are correlated
- These reduced variables are a few linear combinations of the original ones
- Possible to find hidden relationships between variables
- No distributional assumptions needed
- From now on, we must center X by subtracting off column means

PCA: from 3D to 2D



And in high dimension?

Eigenvalues and eigenvectors

- A squared matrix A has an eigenvalue λ associated with an eigenvector $v \neq 0$ if

$$Av = \lambda v$$

  Every squared and symmetric matrix with dimension n has n real eigenvalues with n eigenvectors

- If all the eigenvalues are positive, we say the matrix is positive definite

The spectral decomposition

- If A is symmetric, then $A = V\Lambda V^T$
  where V is a orthogonal matrix containing the eigenvectors (in columns) and Λ is a diagonal matrix containing the eigenvalues in the diagonal (in the same order)

- Spectral decomposition can also be written as

$$A = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \lambda_n v_n v_n^T$$

- As a consequence, $A^r = V\Lambda^r V^T$, $A^{-1} = V\Lambda^{-1} V^T$, and $A^{1/2} = V\Lambda^{1/2} V^T$ (which is called the square root)

Eigenvalues and eigenvectors: Example

Consider

$$A = \begin{pmatrix} 1.00 & 0.50 & 0.33 & 0.25 \\ 0.50 & 1.00 & 0.67 & 0.50 \\ 0.33 & 0.67 & 1.00 & 0.75 \\ 0.25 & 0.50 & 0.75 & 1.0 \end{pmatrix}$$

- Then, $A = V \Lambda V^T$ where

$$V = \begin{pmatrix} 0.0693 & -0.4422 & -0.8105 & 0.3778 \\ -0.3618 & 0.7420 & -0.1877 & 0.5322 \\ 0.7694 & 0.0486 & 0.3010 & 0.5614 \\ -0.5219 & -0.5014 & 0.4662 & 0.5088 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 0.2078 & 0 & 0 & 0 \\ 0 & 0.4078 & 0 & 0 \\ 0 & 0 & 0.8482 & 0 \\ 0 & 0 & 0 & 2.5362 \end{pmatrix}$$

- Then

$$A^{1/2} = \begin{pmatrix} 0.9594 & 0.2394 & 0.1237 & 0.0833 \\ 0.2394 & 0.8948 & 0.3199 & 0.1991 \\ 0.1237 & 0.3199 & 0.8566 & 0.3855 \\ 0.0833 & 0.1991 & 0.3855 & 0.8971 \end{pmatrix}$$

Principal Component Analysis

First principal component:

- For a given data matrix, X, the first component is a linear combination of the original variables: $Z_1 = Xa_1$
- To explain as much variability as possible, vector a_1 is chosen to maximize the variance $s^2_{z_1} = a_1^T S a_1$, where S is the sample covariance matrix
- Add the following constraint: $\|a_1\| = 1$ to normalize
- The solution is $a_1 =$ eigenvector of S with largest eigenvalue ($\lambda_1 = s^2_{z_1}$)

Principal Component Analysis

Second principal component:

- For a given data matrix, X, the second component is a linear combination of the original variables: $Z_2 = Xa_2$
- To explain as much variability as possible, vector a_2 is chosen to maximize the variance $s^2_{z_2} = a_2^T S a_2$, discounting the variance explained by Z_1
- Add the following constraint: $\|a_2\| = 1$ to normalize
- The solution is $a_2 =$ eigenvector of S with second largest eigenvalue ($\lambda_2 = s^2_{z_2}$)

Principal Component Analysis

r-th principal component:

- Analogously, the *r*-th principal component is $Z_r = Xa_r$, where $a_r =$ eigenvector of S with *r*-th largest eigenvalue ($\lambda_r = s^2_{z_r}$)

PCA: Example

For each of the 50 states in the US, we have the number of arrests per 100,000 residents for each of three crimes: Assault, Murder, and Rape. Moreover, we also have the UrbanPop (percentage living in in urban areas)

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

PC1 places equal weight on Assault, Murder, and Rape,but less weight on UrbanPop: hence it is an overall measure of serious crimes

PC2 places most of its weight on UrbanPop: hence basically the level of urbanization (maybe corrected by murder)

Principal Component Analysis

Main conclusions:

- The new matrix (principal component scores) is

$$Z = XA_r$$

  where A_r is the principal component loadings. That is, a $p \times r$ orthogonal matrix whose columns contain the "first"-r eigenvectors of S

- Because X is centered, the principal components matrix Z is also centered

- The variances of the r-principal components are the r-largest eigenvalues of S

- Principal components are unique up to sign and permutation of variables

Principal Component Analysis

Main conclusions:

- $\text{Cov}(Z) = \Lambda$: diagonal matrix with sorted eigenvalues (variances)
- Hence, principal components Z are uncorrelated: better interpretation
- However, principal components Z may not be independent (see example below)
- We can use PCA to identify possible latent variables which may have generated the data
- Or to discard noisy or highly-correlated variables
- We can also use PCA as a preprocessing tool for regression, clustering, etc.

Principal Component Analysis

Main conclusions:

- If $r = p$, then we attain the whole variability: $\text{tr}(S) = \lambda_1 + \cdots + \lambda_p$
- Proportion of variability explained by the r principal components:

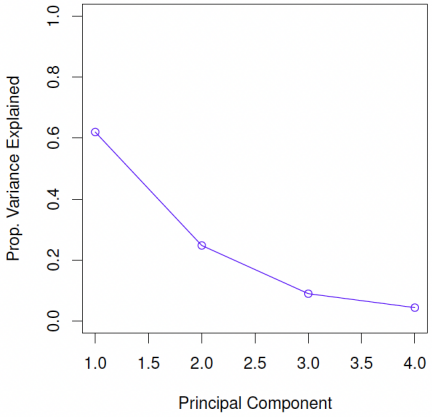
$$\text{PV}_r = \frac{\lambda_r}{\lambda_1 + \cdots + \lambda_p}$$

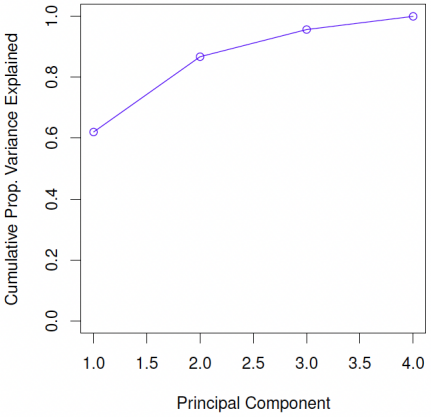
- Accumulated proportion of variability explained by the first r principal components:

$$\text{APV}_r = \frac{\lambda_1 + \cdots + \lambda_r}{\lambda_1 + \cdots + \lambda_p}$$

- Hence, how to choose r, the number of components?

PCA: Example





PC1 explains 62.0% of the variance in the data, while PC2 explains 24.7%

Together, the first two principal components explain almost 87% in the data

PCA: some insights:

- In general, the first principal component is usually interpreted as a global factor of size (in terms of variance), computed by a weighted average of all the variables
- The remaining components are usually interpreted as factors of shape and they typically contrast groups of variables respect to others
- But how many components do we need to explain most of data variability?
- And what happens if variables are measured in different units?

Principal Component Analysis

What happens if original variables are in different units?

- Remember the socioeconomic data set (first 14 countries):

	Country	Region	Population	Under15	Over60	FertilityRate	LifeExpectancy	ChildMortality	CellularSubscribers	LiteracyRate	GNI	PrimarySchoolEnrollmentMale	PrimarySchoolEnrollmentFemale
1	Afghanistan	Eastern Mediterranean	29825	47.42	3.82	5.40	60	98.5	54.26	NA	1140	NA	NA
2	Albania	Europe	3162	21.33	14.93	1.75	74	16.7	96.39	NA	8820	NA	NA
3	Algeria	Africa	38482	27.42	7.17	2.83	73	20.0	98.99	NA	8310	98.2	96.4
4	Andorra	Europe	78	15.20	22.86	NA	82	3.2	75.49	NA	NA	78.4	79.4
5	Angola	Africa	20821	47.58	3.84	6.10	51	163.5	48.38	70.1	5230	93.1	78.2
6	Antigua and Barbuda	Americas	89	25.96	12.35	2.12	75	9.9	196.41	99.0	17900	91.1	84.5
7	Argentina	Americas	41087	24.42	14.97	2.20	76	14.2	134.92	97.8	17130	NA	NA
8	Armenia	Europe	2969	20.34	14.06	1.74	71	16.4	103.57	99.6	6100	NA	NA
9	Australia	Western Pacific	23050	18.95	19.46	1.89	82	4.9	108.34	NA	38110	96.9	97.5
10	Austria	Europe	8464	14.51	23.52	1.44	81	4.0	154.78	NA	42050	NA	NA
11	Azerbaijan	Europe	9309	22.25	8.24	1.96	71	35.2	108.75	NA	8960	85.3	84.1
12	Bahamas	Americas	372	21.62	11.24	1.90	75	16.9	86.06	NA	NA	NA	NA
13	Bahrain	Eastern Mediterranean	1318	20.16	3.38	2.12	79	9.6	127.96	91.9	NA	NA	NA
14	Bangladesh	South-East Asia	155000	30.57	6.89	2.24	70	40.9	56.06	56.8	1940	NA	NA

Principal Component Analysis

- In this case, we cannot compare different variances
- Here, it is better to standardize the data: $Y = XD^{-1/2}$
- Equivalently, compute the principal components from the correlation matrix R instead of the covariance matrix S
- The new principal component scores are $Z = Y\ A_r^R$, where A_r^R represents the loadings (eigenvectors) of R

Socioeconomic data analysis

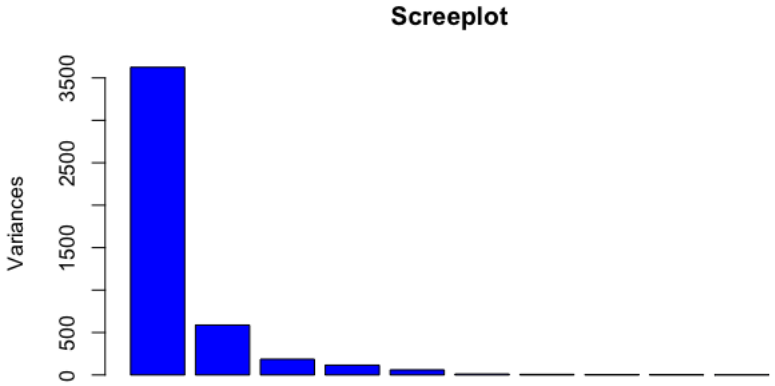
- After centering data, taking some logs, and considering correlations (instead of covariances)

	PC1	PC2	PC3	PC4	PC5
Population	-0.001997728	0.007636178	-0.024236914	0.02398467	-0.014921569
Under15	-0.162833319	-0.123541288	0.035955127	0.33082680	0.619477198
Over60	0.090982286	0.079321478	-0.041995379	-0.20016451	-0.631185560
FertilityRate	-0.024432631	-0.018929111	0.006983663	0.03513694	0.038120126
LifeExpectancy	0.157863386	0.092828759	0.001512155	0.03154670	-0.068010172
ChildMortality	-0.685890137	-0.559732987	-0.257124633	-0.33356834	-0.087438232
CellularSubscribers	0.595490618	-0.788140290	0.150545074	-0.02810403	-0.001239552
LiteracyRate	0.243478902	0.184117183	-0.156530313	-0.82107728	0.437616988
GNI	0.015289012	-0.001760446	-0.004272924	-0.03475459	-0.022932213
PrimarySchoolEnrollmentMale	0.142066989	-0.029660035	-0.671200329	0.18548579	-0.073780769
PrimarySchoolEnrollmentFemale	0.186902716	0.012797079	-0.657646072	0.16076163	0.080165524

- Eigenvalues:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	60.2081	24.2494	13.63980	10.76066	7.75678	3.27367	2.49848	1.74433 1

Socioeconomic data analysis



The first component explains 79% of variability, and the second one a 13%

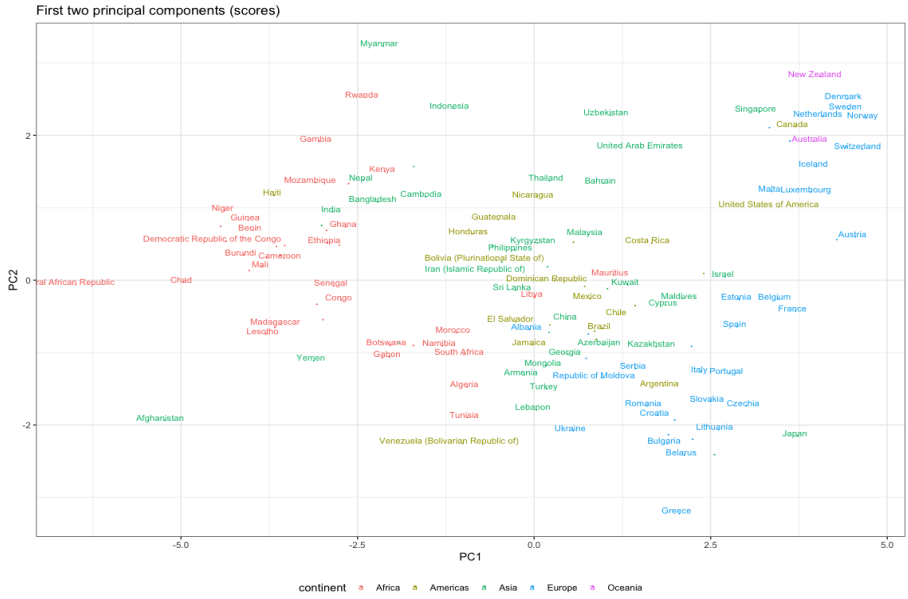
Socioeconomic data analysis

- First component interpretation:
  Look at positive signs
  Look at negative signs
- Hence, we interpret this component as a measure of development
- We can use Z_1 to rank countries

Socioeconomic data analysis

- Second component interpretation:
  Look at positive signs
  Look at negative signs
- More difficult interpretation
- We can also use Z_2 to rank countries
- The following graph helps to understand better these insights. Note PCs are uncorrelated but not independent

Socioeconomic data analysis



PCA: Extensions

- **Principal component regression**: handle better collinearity in regression and reduce the associated dimension

  Instead of fitting a high-dimensional regression model:

$$y = X\beta + \varepsilon, \quad \text{where } \beta \in \mathcal{R}^p$$

  apply PCA: $Z = XA_r$ with $r << p$ and fit

$$y = Z\beta + \varepsilon, \quad \text{where now } \beta \in \mathcal{R}^r$$

PCA: Extensions

- **Sparse PCA**: consider only just a few input variables (instead of all them as PCA)

  For instance, add a lasso-type penalty to each loading vector:

$$\begin{aligned} \max_{a_r} \quad & a_r^T \, S \, a_r - \lambda_r ||a_r||_1 \\ \text{s. t.} \quad & ||a_r||_2^2 = 1 \\ & a_h^T a_r = 0 \text{ for } h < r \end{aligned}$$

  Warning: previous problem is not convex

PCA: Extensions

- **Sparse PCA**: there are more alternatives
- For instance, consider this two-stage analysis:
  - Compute first classical PCA to obtain the scores $Z_1, \ldots, Z_p$
  - Then, use each score to find suitable sparse approximations:

$$\min_\beta \quad ||Z_r - X\beta||_2^2 + \lambda_2||\beta||_2^2 + \lambda_1||\beta||_1$$

    to finally obtain the sparse loadings: $a_r = \frac{\beta}{||\beta||_2}$

PCA: Extensions

- PCA is fast and accurate when $n >> p$, but it is less accurate when $n \simeq p$
- To improve numerical stability, use SVD decomposition
- In R:
  - prcomp(): singular-value decomposition of data matrix
  - princomp(): eigenvalue decomposition of covariance/correlation matrix
- SVD is preferred numerically to classical PCA when $X^T X$ is bad conditioned (high dimension in general) or when $p > n$, although it is computationally more expensive

PCA: an application

- *Genes mirror geography within Europe*
  Nature (2008)

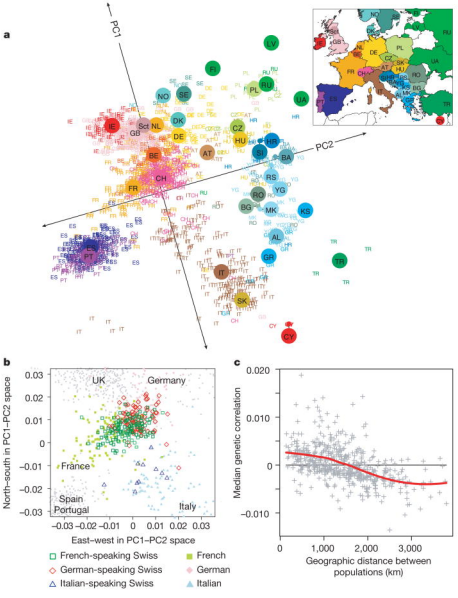
- $n = 3,000$ European individuals were genotyped

- Genetic variation in $p = 500,000$ variables in the European human-genome is reduced to just 2 dimensions

- Note $p >> n$

- A geographical map of Europe arose from the first 2 principal components
  I.e. close correspondence between genetic and geographic distances

PCA: an application



Summary

Final conclusions:

- PCA is a simple analytical tool to reduce the dimension of a data set or reveal hidden structure
- They are (a few) linear combinations of original variables that help us interpreting better the data
- Center always the original data set X
- If variables are in different units, first standardize X: $Y = XD^{-1/2}$,
  second compute the principal components from the cov matrix of Y or the correlation matrix of X (they are the same)
  Finally, $Z = Y\, A_r^R$
- Interpretation comes from A_r (loadings) or Z (scores)
- Use SVD and/or Sparse PCA in high dimension

3. Dimension Reduction: Factor Analysis

Factor Analysis: Introduction

The data matrix X is $n \times p$. What happens if p is large?

- In PCA, we showed a descriptive tool to reduce the dimension
- Now, we will study an analytical tool: model that will reduce the dimension
- Very useful in Social Sciences where some factors (like intelligence, quality of life, happiness) cannot be measured directly: latent variables
- We will approximate these variables through (many) observable variables like wealth, employment, physical and mental health, education level, leisure time, etc.: indicators

Factor Analysis: Introduction

- Factor analysis provides a tool to find the relationships between latent variables (non-observable) and indicators (observable)
- The focus here is to explain correlations between indicators due to common factors (latent variables)
- Remember the focus of PCA is to explain total variance
- FA is a model (provides a way to generate data) that explains common co-movements (between indicators)
- Original idea (C. Spearman): g factor (psychometrics)

Factor Analysis: Example

- Analyze the quality of life in several countries
- We can observe, for each country, three variables: the wealth (like GDP), the employment and the health/safety
- The factor model with one factor for these three variables is:

$$
\begin{aligned}
x_1 &= \mu_1 + L_{11}\, f + \varepsilon_1 \\
x_2 &= \mu_2 + L_{21}\, f + \varepsilon_2 \\
x_3 &= \mu_3 + L_{31}\, f + \varepsilon_3
\end{aligned}
$$

- f is an unobservable factor (possibly the quality of life), μ are the means of the three variables, and ε contains the error terms (with mean 0)
- Note this is a model: we are imposing the three variables are linear respect to the factor plus an error

Factor Analysis: General Formulation

- Assume $x = (x_1, \ldots, x_p)^T$ is a random vector, with mean $\mu_x = E(x) = (\mu_1, \ldots, \mu_p)^T$ and covariance matrix Σ_x

- A factor model has the form

$$x = \mu_x + L\, f + \varepsilon$$

where

  - L is the loading matrix with dimension $p \times r$ and $r < p$ is the number of factors. Note L contains constants to be estimated
  - $f = (f_1, \ldots, f_r)^T$ is a random vector with the latent variables (the factors), such that $E(f) = 0$ and $\text{Cov}(f) = I$
  - $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)^T$ is the noise term, such that $E(\varepsilon) = 0$ and $\text{Cov}(\varepsilon) = \Psi = \text{diag}(\psi_1, \ldots, \psi_p)$. Moreover, factors and errors are uncorrelated: $\text{Cov}(f, \varepsilon) = 0$

Factor Analysis: General Formulation

- Data formulation: if X is a data matrix, the factor model has the form

$$X = 1\mu_x^T + F\,L^T + E$$

where 1 is a vector of ones, F is the factor matrix, and E is the noise matrix

Factor Analysis: Properties

- For each component, the corresponding model is independent except for the presence of the common factors
- $\text{Cov}(x, f) = E((x - \mu_x)f^T) = L$
- If x is standardized, $\text{corr}(x, f) = L$
- That is, the loading matrix contains the covariances (correlations) between the variables and the factors
- Because $\text{Var}(f) = I$, L_j^T contains the β-coefficients of the j-regression:

$$x_j = \mu_j + f^T\beta + \varepsilon_j$$

Factor Analysis: Properties

- Factors and errors are uncorrelated, then decomposition:

$$\Sigma_x = LL^T + \Psi$$

- Hence, the original covariance matrix (with $p(p+1)/2$ elements) can be computed with just $p(r+1)$ elements

- Moreover, the following decomposition for the variance of each variable is attained

$$\sigma_j^2 = h_j + \psi_j, \quad j = 1, \ldots, p$$

where $h_j = L_{j1}^2 + \cdots + L_{jr}^2$ is the *j*-communality: proportion of common variance explained by factors

- Hence, ψ_j represents the specific variability of x_j not due to factors, and is the j-th uniqueness

- Take care: previous properties are only valid if model is true...

Factor Analysis: Estimation

- Mainly based on the decomposition: $\Sigma_x = LL^T + \Psi$
- Note the main interest is in L, not in f (which does not appear in the decomposition)
- Note the estimation depends on Σ_x, not on original variables x
- When factors are not observable, the factor model cannot be determined uniquely (even if they are uncorrelated)
- Then, the loading matrix and the factors will be determined up to an orthogonal transformation

Factor Analysis: Estimation

- That is, note the model will be the same if we rotate the factors:

$$L^* = L\,H, \quad \text{and} \quad f^* = H^T f, \quad \text{with } H \text{ a } r \times r \text{ orthogonal matrix}$$

- So, which rotation should we use?
- Several criteria: that rotation with smallest trace of Ψ, varimax, etc.
- When we rotate factors, the communalities and specific variances do not change
- With rotation we try to convert factors into uncorrelated factors to improve interpretation

Factor Analysis: Estimation

- Based on principal components factor analysis:
  - First, compute the sample covariance matrix S_x
  - Then, $\hat{L}_r = V_r D_r^{1/2}$, where V_r contains the first r principal components of S_x, and D is the corresponding diagonal matrix with the eigenvalues (variances)
  - Finally, $\hat{\Psi}_r = \text{diag}(S_x - \hat{L}_r \hat{L}_r^T)$

- Based on principal factor analysis: Same as before but replacing S_x with $S_x - \hat{\Psi}$, for some estimation of Ψ (we can use previous $\hat{\Psi}_r$). We can iterate this process

- Based on Maximum Likelihood Estimation: assume normality for f and ε, and then estimate L and Ψ using MLE (optimization algorithm)

Factor Analysis: Different Units

- Many times variables have different units
- Then, standardize: $y = \Delta_x^{-1/2}(x - \mu_x)$, where Δ_x contains the variances
- The factor model for y, given the general factor model for x, is:

$$y = Mf + \epsilon$$

  where $M = \Delta_x^{-1/2} L$ and $\text{Cov}(\epsilon) = \Delta_x^{-1/2} \Psi \Delta_x^{-1/2}$

- That is, work with the correlation matrix instead of the covariance matrix
- Then, the decomposition is $\rho_x = MM^T + \Phi$
- And in terms of communalities, $1 = g_j + \phi_j$, $j = 1, \ldots, p$

Factor Analysis: Estimation of factors, f

- In FA, the main interest is in L
- But, what happens if we want to estimate f (scores)?
  This is useful to rank observations
- Three methods:
  - Distribution-free estimation (non-parametrics): based on LS
    Consider L is fixed (already estimated) and F are the parameters in the regression $x - \mu_x = L\, f + \varepsilon$.
    Then, the LS estimator for the factor scores is

$$\hat{f}_i = (\hat{L}^T \hat{L})^{-1}\, \hat{L}^T (x_i - \bar{x}), \text{ for } i = 1, \ldots, n$$

  - WLS or Bartlett (also no distribution): each indicator has different variability (more weight to indicators with less specific variability)

$$\hat{f}_i = (\hat{L}^T \hat{\Psi}^{-1} \hat{L})^{-1}\, \hat{L}^T \hat{\Psi}^{-1} (x_i - \bar{x}), \text{ for } i = 1, \ldots, n$$

  - Based on MLE (regression approach): assume f is a random normal vector, then

$$\hat{f}_i = (\hat{L}^T \hat{\Psi}^{-1} \hat{L} + I)^{-1}\, \hat{L}^T \hat{\Psi}^{-1} (x_i - \bar{x}), \text{ for } i = 1, \ldots, n$$

Factor Analysis: Example

- Consider the socioeconomic data set (first 14 countries):

	Country	Region	Population	Under15	Over60	FertilityRate	LifeExpectancy	ChildMortality	CellularSubscribers	LiteracyRate	GNI	PrimarySchoolEnrollmentMale	PrimarySchoolEnrollmentFemale
1	Afghanistan	Eastern Mediterranean	29825	47.42	3.82	5.40	60	98.5	54.26	*NA*	1140	*NA*	*NA*
2	Albania	Europe	3162	21.33	14.93	1.75	74	16.7	96.39	*NA*	8820	*NA*	*NA*
3	Algeria	Africa	38482	27.42	7.17	2.83	73	20.0	98.99	*NA*	8310	98.2	96.4
4	Andorra	Europe	78	15.20	22.86	*NA*	82	3.2	75.49	*NA*	*NA*	78.4	79.4
5	Angola	Africa	20821	47.58	3.84	6.10	51	163.5	48.38	70.1	5230	93.1	78.2
6	Antigua and Barbuda	Americas	89	25.96	12.35	2.12	75	9.9	196.41	99.0	17900	91.1	84.5
7	Argentina	Americas	41087	24.42	14.97	2.20	76	14.2	134.92	97.8	17130	*NA*	*NA*
8	Armenia	Europe	2969	20.34	14.06	1.74	71	16.4	103.57	99.6	6100	*NA*	*NA*
9	Australia	Western Pacific	23050	18.95	19.46	1.89	82	4.9	108.34	*NA*	38110	96.9	97.5
10	Austria	Europe	8464	14.51	23.52	1.44	81	4.0	154.78	*NA*	42050	*NA*	*NA*
11	Azerbaijan	Europe	9309	22.25	8.24	1.96	71	35.2	108.75	*NA*	8960	85.3	84.1
12	Bahamas	Americas	372	21.62	11.24	1.90	75	16.9	86.06	*NA*	*NA*	*NA*	*NA*
13	Bahrain	Eastern Mediterranean	1318	20.16	3.38	2.12	79	9.6	127.96	91.9	*NA*	*NA*	*NA*
14	Bangladesh	South-East Asia	155000	30.57	6.89	2.24	70	40.9	56.06	56.8	1940	*NA*	*NA*

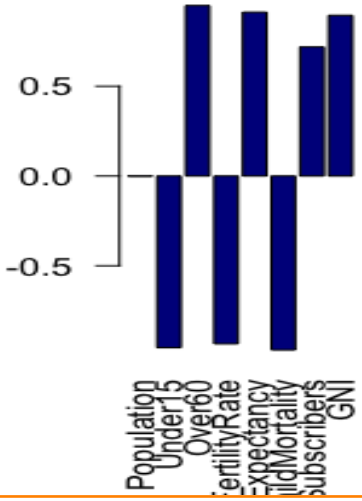
Factor Analysis: Socioeconomic example

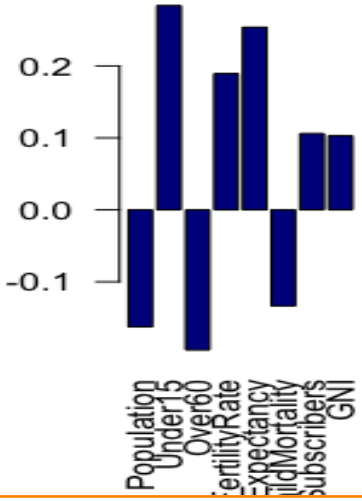
- The variables are in different units: work with the correlation matrix
- Moreover, after taking some logs, removing outliers and educational variables

	Factor1	Factor2	Uniquenesses
Population	-0.001894033	-0.1626864	0.97346131
Under15	-0.956331161	0.2838735	0.00500000
Over60	0.949326225	-0.1947343	0.06085962
FertilityRate	-0.934738914	0.1893143	0.09042268
LifeExpectancy	0.912684651	0.2537824	0.10260149
ChildMortality	-0.968801843	-0.1335775	0.04358040
CellularSubscribers	0.718876258	0.1060469	0.47194259
GNI	0.895185739	0.1031421	0.18800922

Factor Analysis: Socioeconomic example

- Factors seem to explain well the correlation between all variables, except Population

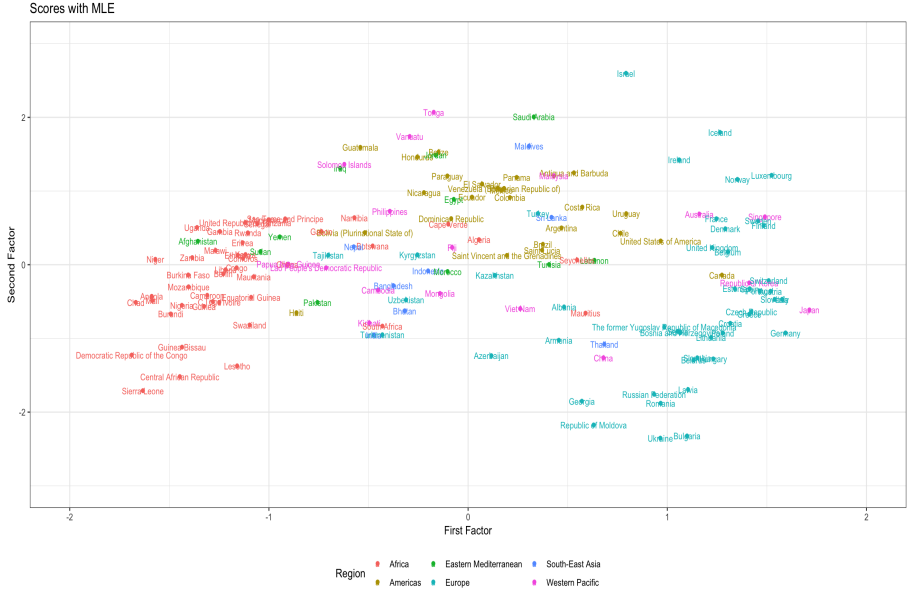




Factor Analysis: Socioeconomic example

- First factor: measure of development
- Second factor?

Factor Analysis: Socioeconomic example



Factor Analysis: final comments

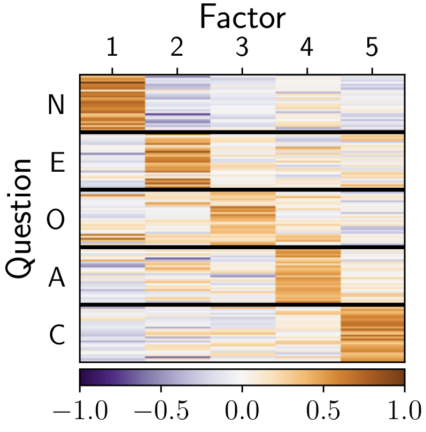
- The results and interpretation are only valid if the model is true
- Hence, check the model assumptions
- Diagnosis: check if the residuals $r = x - \hat{L}\hat{f}$ are $N_p(0, \hat{\Psi})$
- FA provides many models for the same data depending on:
  - The estimation of L and Ψ (based on principal components, based on MLE, ...)
  - How we rotate the factors (varimax, ...)
  - How we compute the factor scores (weighted least-squares, ridge regression, ...)

Human Behaviour Application

There are about five major personality domains that describe the personality profile of an individual

▸ Nature Human Behaviour 2, 735–742 (2018)

- From $p = 300$ items/featutes in a questionary
- From $n = 145388$ individuals
- $r = 5$ personality traits can be obtained: Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness
- Varimax rotation was used to interpret better the personality factors

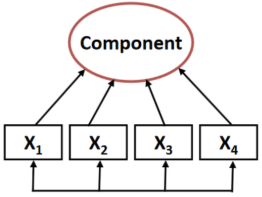


Factor Analysis: final considerations

- Factor analysis and PCA are quite related, but not the same
- PCA: find a linear combination that explains most of the **total variance**
- Factor analysis: find unobservable (common) factors that explain **indicator variances and covariances**
- In PCA, factors (components) are uncorrelated whereas in FA are not
- In PCA the factors (components) are linear combinations of indicators whereas in FA the indicators are linear combinations of factors. That implies FA is more interpretable but definition of factors is not explicit (more black-box)
- But if $\Psi \simeq 0$, both tools are roughly similar

Final considerations

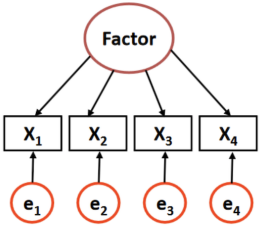
Principal Component Analysis



Linear combination that explains most of the total variance

The observable variables cause the component

Factor Analysis



Unobservable (common) factors that explain variable variances and covariances

The factors cause the observable variables

If error term, e, is small, then both tools are roughly similar

The end... Attained objectives (I hope...)

- Capacity to describe and interpret multivariate datasets
- Understand properties of multivariate distributions and make inference
- Find and analyze relationships between many variables
- Reduce the dimensionality of a dataset and identify real associations
- Handle the R language for multivariate data analysis

“Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.”

John W. Tukey

Afi Escuela de Finanzas