Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Dealing with missing values – part 1

Applied Multivariate Statistics – Spring 2013

Overview

- Bad news: Data Processing Inequality
- Types of missing values: MCAR, MAR, MNAR
- Methods for dealing with missing values:
  - Case-wise deletion
  - Single Imputation
  (- Multiple Imputation in Part 2)

Information Theory 101

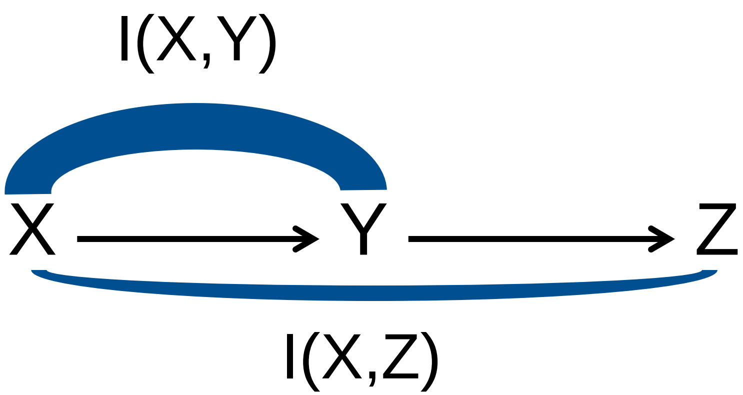
- Entropy: Amount of uncertainty

$$H(X) = -\sum_{x \in X} p(x) \log(p(x))$$

- Mutual Information btw. X and Y
  - What do you learn about X, if you know Y?
  - Decrease in entropy of X, if Y is known

$$I(X,Y) = H(X) - H(X|Y)$$

Information Theory 101: Data Processing Inequality



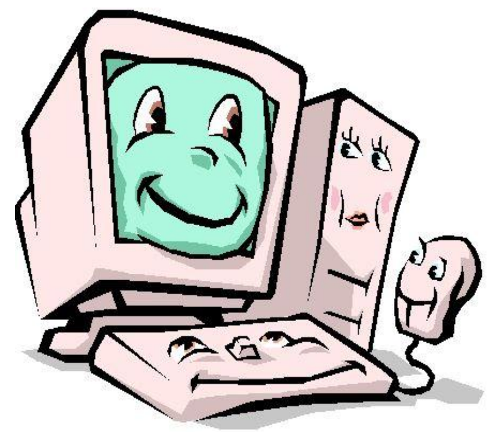
For a Markov Chain: $I(X,Z) \cdot \; I(X,Y)$

Postprocessing can never add information

Natur

.raw

.jpg



Postprocessing can never add information

Natur

Data with
missing values

A	B	C
1.3	5.4	7.2
3.2	?	?
?	8.3	?

After dealing with
missing values
somehow

A	B	C
1.3	5.4	7.2
3.2	7.2	5.6
8.1	8.3	8.2

Information Theory on dealing with missing values

- The information is lost!
  You cannot retrieve it just from the data!
- Try to avoid missing values where possible!
- When dealing with the data, don't waste even more information!
  Use clever methods!

Get an overview of missing values in data

- R: Function “md.pattern” in package “mice”

Types of missing values

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

OK

PROBLEM

Distribution of Missingness

Complete data Y_{com}

A	B	C
1.3	2.5	6.3
2.0	3.6	5.4
1.6	2.3	4.3

Some values are missing

Y_{obs}

A	B	C
1.3	2.5	
2.0		5.4
1.6		4.3

Y_{mis}

A	B	C
		6.3
	3.6	
	2.3	

R

A	B	C
1	1	0
1	0	1
1	0	1

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Example: Blood Pressure

- 30 participants in January (X) and February (Y)
- MCAR: Delete 23 Y values randomly
- MAR: Keep Y only where X > 140 (follow-up)
- MNAR: Record Y only where Y > 140 (test everybody again but only keep values of critical participants)

	Y			
X	Complete	MCAR	MAR	MNAR
Data for individual participants				
169	148	148	148	148
126	123	—	—	—
132	149	—	—	149
160	169	—	169	169
105	138	—	—	—
116	102	—	—	—
125	88	—	—	—
112	100	—	—	—
133	150	—	—	150
94	113	—	—	—
109	96	—	—	—
109	78	—	—	—
106	148	—	—	148
176	137	—	137	—
128	155	—	—	155
131	131	—	—	—
130	101	101	—	—
145	155	—	155	155
136	140	—	—	—
146	134	—	134	—
111	129	—	—	—
97	85	85	—	—
134	124	124	—	—
153	112	—	112	—
118	118	—	—	—
137	122	122	—	—
101	119	—	—	—
103	106	106	—	—
78	74	74	—	—
151	113	—	113	—

Distribution of Missingness

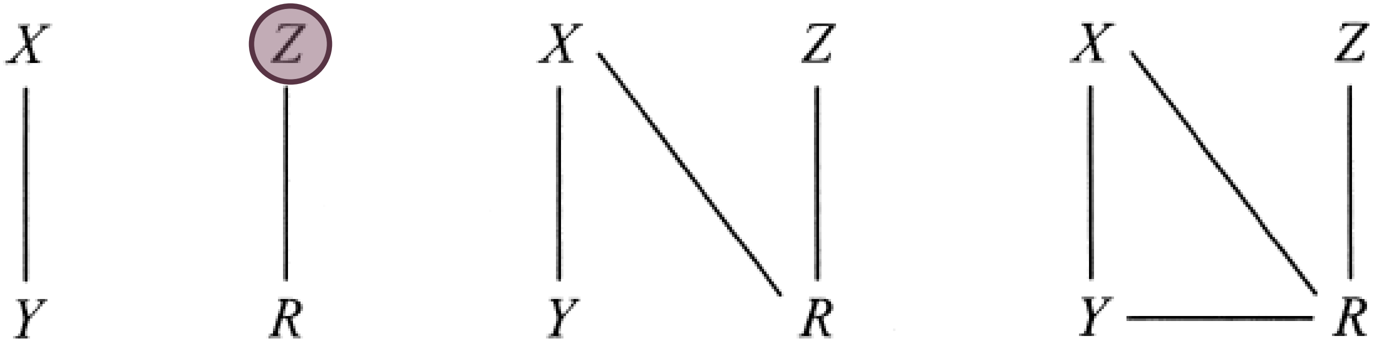
- MCAR $P(R|Y_{com}) = P(R)$
  Missingness does not depend on data

- MAR $P(R|Y_{com}) = P(R|Y_{obs})$
  Missingness depends only on observed data

- MNAR $P(R|Y_{com}) = P(R|Y_{mis})$
  Missingness depends on missing data

Distribution of Missingness: Intuition

Some unmeasured variables not related to X or Y



(a) MCAR (b) MAR (c) MNAR

Problems in practice

- Type is not testable.
- Pragmatic:
  - Use methods which hold in MAR
  - Don’t use methods which hold only in MCAR

Dealing with missing values

- Complete-case analysis - valid for MCAR
- Single Imputation - valid for MAR
- (Multiple Imputation – valid for MAR)

Complete-case analysis

- Delete all rows, that have a missing value
- Problem:
  - waste of information; inefficient
  - introduces bias if MAR
- OK, if 95% or more complete cases
- R: Function “complete.cases” in base distribution

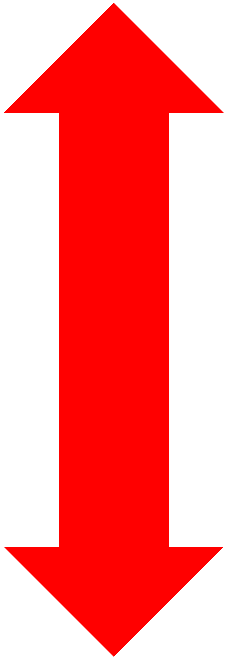
A	B	C	D
NA	3	4	6
3	2	3	NA
2	NA	5	4
5	7	NA	5
6	NA	9	2

- 25% missing values
- ZERO complete cases
  Complete-case analysis is useless

Single Imputation

- Unconditional Mean
- Unconditional Distribution
- Conditional Mean
- Conditional Distribution

Easy / Inaccurate



Hard / Accurate

Unconditional Mean: Idea

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

Mean = 4.75

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	4.75

Unconditional Distribution: Hot Deck Imputation

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

Randomly select observed value in column

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	6.3

Conditional Mean: E.g. Linear Regression

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

Estimate lm(C ~ A + B)
or something similar

Apply to predict C

Conditional Mean: E.g. Linear Regression

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

Prediction of

linear regression

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	8

Conditional Distribution: E.g. Linear Regression

- Start with Conditional Mean as before
- Add randomly sampled residual noise

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

Prediction of linear regression PLUS NOISE

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	8.3

Being pragmatic: Conditional Mean Imputation with missForest

- Use Random Forest (see later lecture) instead of linear regression
- Good trade-off between ease of use / accuracy
- Works with mixed data types (categorical, continuous and mixed)
- Estimates the quality of imputation
  OOBerror: Imputation error as percentage of total variation
  close to 0 - good
  close to 1 - bad

Idea of missForest

A	B	SEX
2.1	NA	M
3.4	3.7	F
4.1	4.5	NA

Idea of missForest

A	B	SEX
2.1	3.0	M
3.4	3.7	F
4.1	4.5	F

Fill in random values

Idea of missForest: Step 1

A	B	SEX
2.1	3.0	M
3.4	3.7	F
4.1	4.5	F

Apply B ~ A + SEX

Learn B ~ A + SEX
with Random Forest

Idea of missForest: Step 1

A	B	SEX
2.1	3.2	M
3.4	3.7	F
4.1	4.5	F

Apply B ~ A + SEX → update value

Learn B ~ A + SEX
with Random Forest

Idea of missForest: Step 2

A	B	SEX
2.1	3.2	M
3.4	3.7	F
4.1	4.5	F

Learn SEX ~ A + B
with Random Forest

Apply SEX ~ A + B → update

Repeat steps 1 & 2 until some stopping criterion is reached
(no real convergence;
stop if updates start getting bigger again)

Measuring quality of imputation

- Normalized Root Mean Squared Error (NRMSE):

$$NRMSE = \sqrt{\frac{mean(Y_{com} - Y_{imputed})^2}{var(Y_{com})}}$$

- Proportion of falsely classified entries (PFC) over all categorical values

$$PFC = \frac{nmb.\ missclassified}{nmb.\ categorical\ values}$$

Pros and Cons of missForest

- Effects are OK, if MAR holds
- Easily available: Function “missForest” in package “missForest”
- Estimation of imputation error
- Accuracy might be too optimistic, because
  - imputed values have no random scatter
  - model for prediction was taken to be the true model, but it is just an estimate
- Solution: Multiple Imputation

Concepts to know

- Data Processing Inequality and connection to missing values
- Distributions of missing values
- Case-wise deletion
- Methods for Single Imputation
- Idea of missForest; error measures for imputed values

R functions to know

- md.pattern
- complete.cases
- missForest