S3: Missing data

Hae-Won Uh

IMforFUTURE, Edinburgh

March 27, 2019

Who am I?

Short CV

- (Concert) Pianist, did math as a hobby
- PhD in mathematical statistics (UvA), postdoc in statistical genetics (LUMC)
- Ass. prof (LUMC, UMCU) enjoys working at (European) projects.

Current interest

- Integration of omics data
- Microbiome data analysis
- Wearable (sensor) data analysis

My group

- (Together with Jeanine): Zhujie Gu (IMforFUTURE), Said el Bouhaddani (MIMOmics, BD@H)
- Sven Kleine Bardenhorst (UU), Francesco Bussola (UBologna)

Outline

- MCAR, MAR, MNAR, etc & R exercise 1
- Multiple imputation & R exercise 2
- missForest
- Imputation of missing blocks
- Summary & guidelines
- References

Missing data is everywhere

- Survey data: non-reponse
- Clinical trials: dropout
- Data integration: different variables collected by different gourps or in different experiments
- ...

Mechanisms of missing data

MCAR: Missing completely at random

- Pattern of missingness independent of missing values
- Example: We run a taste study for 10 tomatoes. Each subject was asked to rate only 3 tomatoes at random.
- (Listwise) deletion: Unbiased results, but smaller sample size

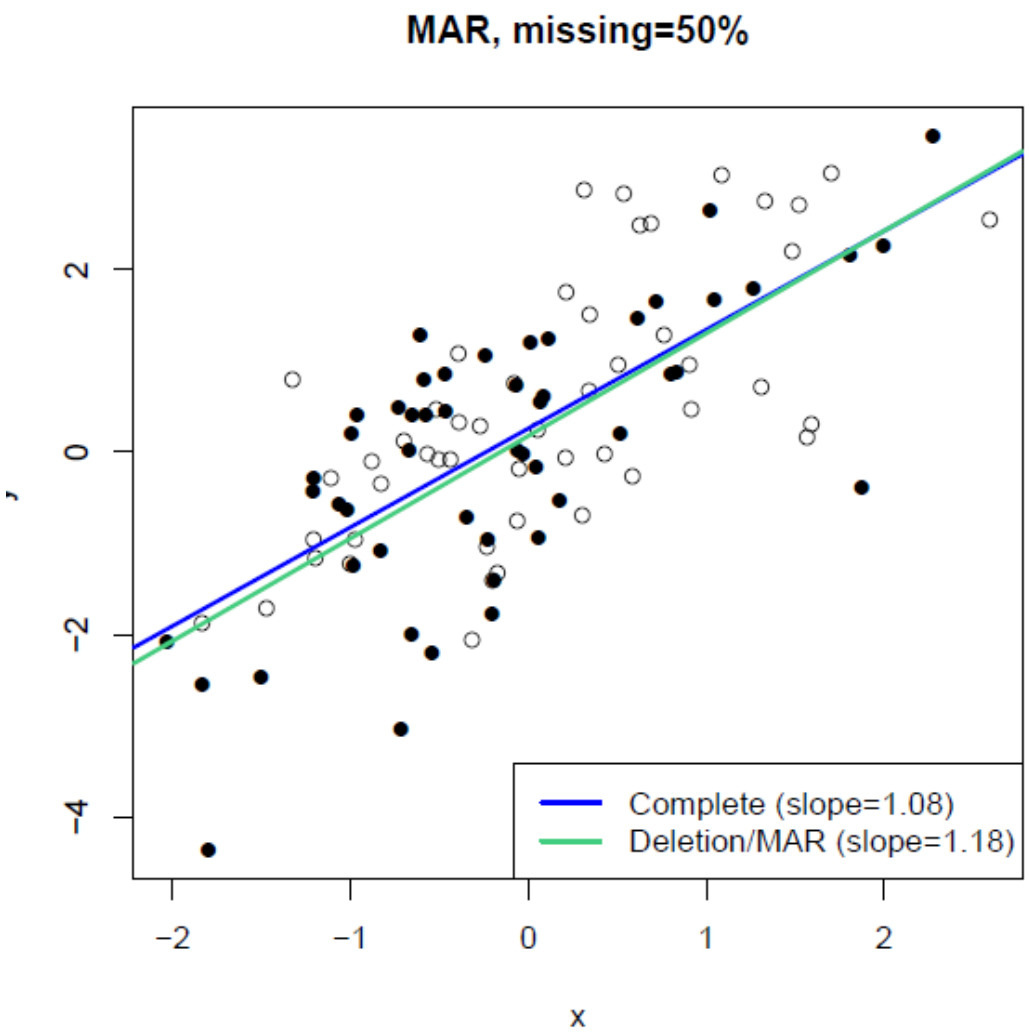
MAR: Missing at random

- Missingness depends only on the non-missing data (and thus in principle the missing values can be predicted from them).
- Example: In a survey, poor subjects were less likely to answer a question about drug use than wealthy subjects. $\rightarrow$ Related to observed predictors (income) but not drug use.

MNAR: Missing NOT at random

- Example: IgG4 cannot be measured below a certain detection limit

MAR or MNAR?

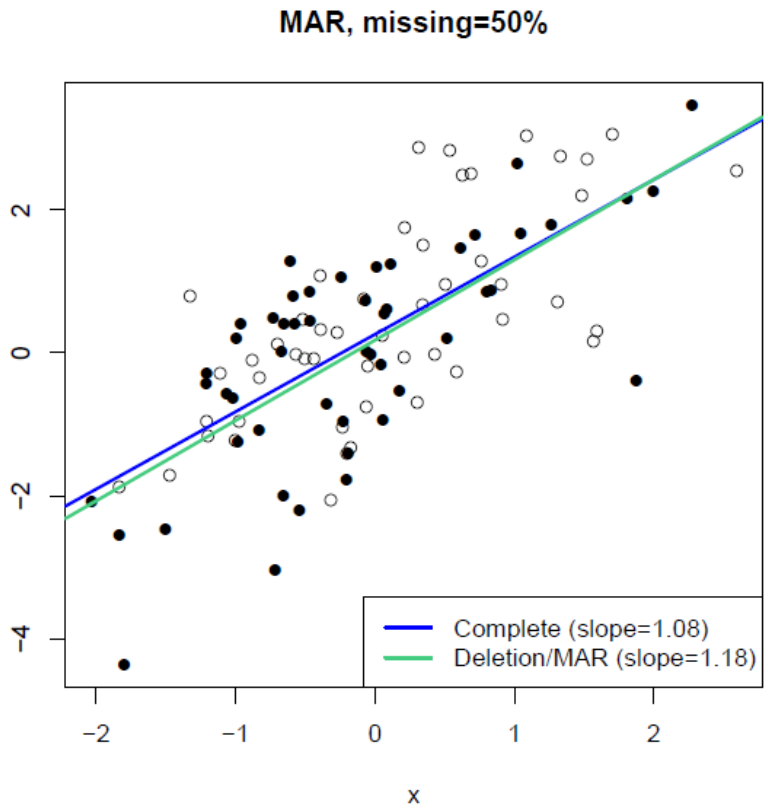


MAR, missing=50%
N=100

Complete: slope=1.08

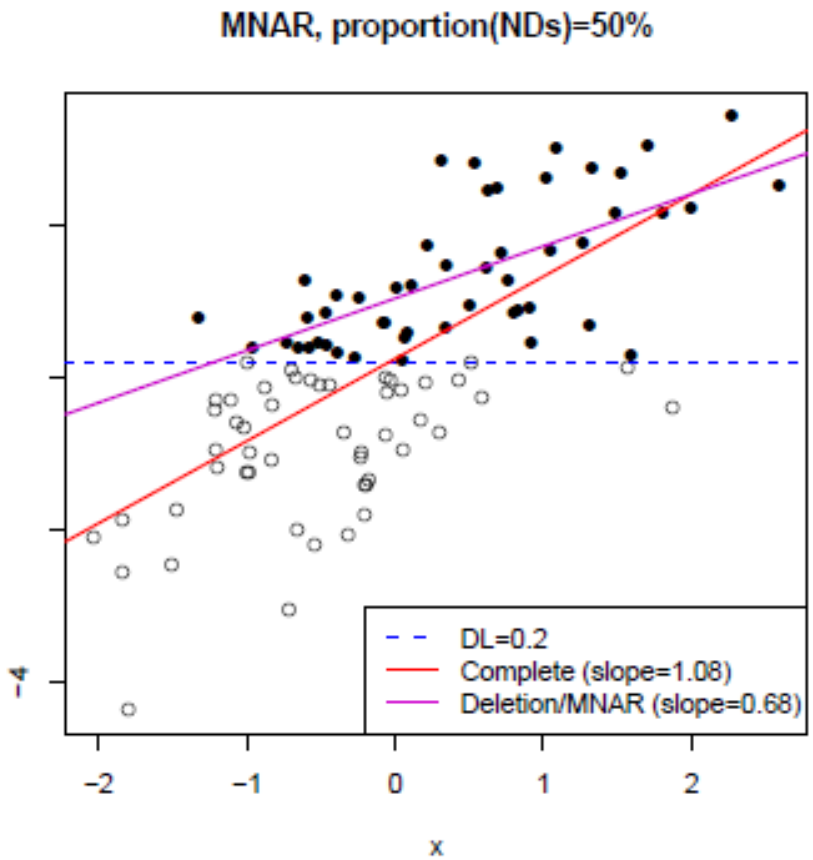
Deletion: slope=1.18

MAR or MNAR?



MAR, missing=50%
Complete data: slope=1.08
Deletion: slope=1.18

MNAR, Proportion(NDs)=50%
Complete: slope=1.08
Deletion: slope=0.68



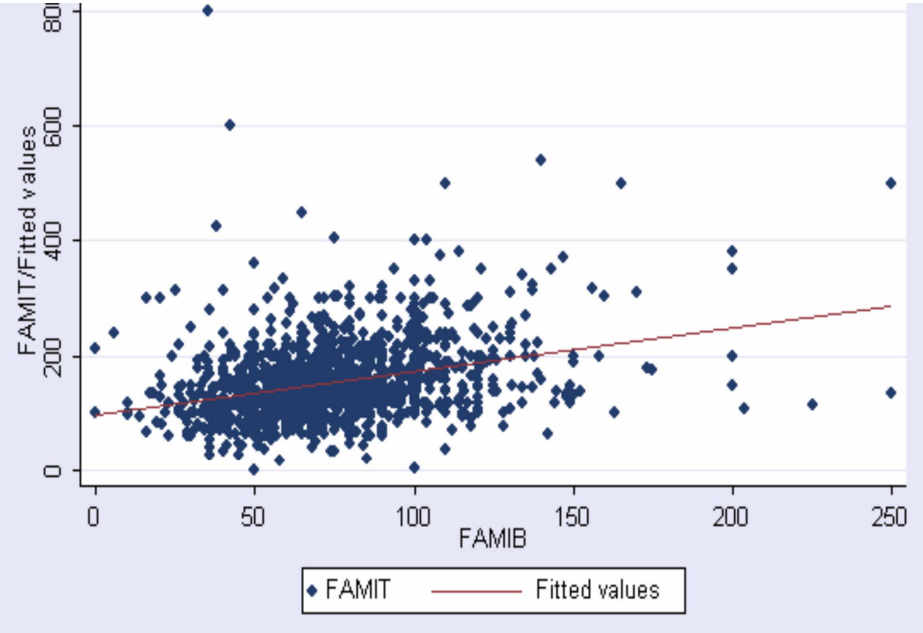
Recap: Lines of Best Fit and Linear Regression

There is some relationship between two variables, but it is not a perfect relationship. To describe a linear relationship we use a Line of Best Fit.

- This takes the form

$$\hat{y} = \alpha + \beta x.$$

- α: Intercept or constant
- β: Gradient or slope of the line (i.e. how steep it is).
- If $\beta = 0$, no relationship



Linear Regression estimates the equation of the Line of Best Fit, using a technique called Ordinary Least Squares (OLS).

R exercise 1

For all

1. Load the data some-data-1.Rda
2. Explore the data: how many, continuous/discrete, any missing etc
3. Linear regression: BMI outcome, 1 covariate Cholesterol
4. Make (randomly) 20 of Cholesterol values missing
5. Repeat the analysis of (3)
6. Compare the results of (3) and (5).
7. Repeat (4) - (6) with different values of missing & compare the results
8. Make summary (table)

For stats ESRs

- Reproduce the two plots, MAR & MNAR (DL)

R code

```
#== Some simulated data
load("some-data-1.Rda")

# Study the dataset
head(dat)
str(dat)

# Check the data for missing values.
sapply(dat, function(x) sum(is.na(x)))

# Fit a linear regression
fit <- lm(BMI ~ Cholesterol, data=dat)
summary(fit)
```

Generate missing values (and save the data)

```
dat.mis <- dat

# set.seed(27032019)

# Make missing in each variable.
dat.mis[sample(1:nrow(dat.mis), 20), "Cholesterol"] <- NA
dat.mis[sample(1:nrow(dat.mis), 10), "Smoking"] <- NA
dat.mis[sample(1:nrow(dat.mis), 20), "Education"] <- NA
dat.mis[sample(1:nrow(dat.mis), 50), "Age"] <- NA

# Save the data
save(dat.mis, file="dat.mis.Rda")

# Study patterns of missing
sapply(dat.mis, function(x) sum(is.na(x)))
```

Dealing with missing data

Methods to avoid

- Observation (listwise) deletion
- Variable deletion
- Last observation carried forward, creating an extra category for the missing values, etc.
- Single imputation:
  1. Replace with the mean or median of the column (variable)
  2. Mean imputation using regression

We discuss today

- Multiple imputation
- missForest

Dealing with missing data

Methods to avoid

- Observation (listwise) deletion
- Variable deletion
- Last observation carried forward, creating an extra category for the missing values, etc.
- Single imputation:
  1. Replace with the mean or median of the column (variable)
  2. Mean imputation using regression

We discuss today

- Multiple imputation
- missForest

Multiple imputation

Basic idea

- For each variable in turn, impute a missing value by drawing from the conditional distribution of the variable, given the rest of the data. → meaning?

Practically,

1. Use many variables, including the dependent variable and variables not planned for inclusion in the finall model, to predict missings
2. Create several *Imputed/Complete* data sets.
3. Run Each analysis on Each Imputed Dataset
4. Combine the estimates by taking uncertainty into account.

Multiple imputation

Basic idea

- For each variable in turn, impute a missing value by drawing from the conditional distribution of the variable, given the rest of the data. → meaning?

Practically,

1. Use many variables, including the dependent variable and variables not planned for inclusion in the finall model, to predict missings
2. Create several *Imputed/Complete* data sets.
3. Run Each analysis on Each Imputed Dataset
4. Combine the estimates by taking uncertainty into account.

Rubin's Rules for slope & variance estimates

For $m\,(=5)$ imputed datasets

- Slope estimate

$$\bar{\beta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_i$$

- Variance estimate for $\bar{\beta}$ comprises of
  1. Within-imputation variance: $\bar{U} = 1/m\sum_{i=1}^{m}\mathrm{Var}(\hat{\beta}_i)$
  2. Between-imputation variance: $B = 1/(m-1)\sum_{i=1}^{m}(\hat{\beta}_i - \bar{\beta})^2$

- Total imputation variance of $\bar{\beta}$

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B.$$

Know more? Imputation: Stef van Buuren

Multiple imputation (MI) using mice package

```
library(mice)

# Study patterns of missing
md.pattern(dat.mis)

# library(VIM): optional
aggr(dat.mis, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(dat.mis), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))

# MI
dat.imp <- mice(dat.mis, m=5, maxit = 50, method = 'pmm',
                seed = 500)
summary(dat.imp)
```

Multiple imputation (MI) using mice package (2)

- Predictive mean matching (pmm): when filling in missings, find cases with similar predicted values to the case in question
  - Imputations for categorical variables always match the original scale (sex is always 0 or 1, not 0.75)
  - When a variable is badly skewed, the pmm always selects a realistic value.
- complete function can return any of the individual imputed data frames
- pool: outputs using Rubin's rule

R code for MI

```
library(mice)
load(``dat.mis.Rda'')

set.seed(28032019)
dat.imp <- mice(dat.mis, m=5, maxit = 50, method = 'pmm', s
summary(dat.imp)

# Build predictive model
fit.imp <- with(data = dat.imp, exp = lm(BMI~Cholesterol))

# Combine results of all 5 models
combine <- pool(fit.imp)
summary(combine)
```

R exercise 2: Multiple imputation using mice package

For all

1. Load the data dat.mis.Rda
2. Explore the missing pattern
3. Impute 10 sets using mice
4. Fit linear regression: BMI outcome, 1 covariate Cholesterol
5. Pool the results using mice package
6. Impute 1 set
7. Repeat the analysis of (4) and (5)
8. Compare the results of (5) and (7), and those of the original data.

For stats ESRs

- After the step (3), use 5 complete datasets to fit the linear regression
- Pool the estimates using Rubin's rule

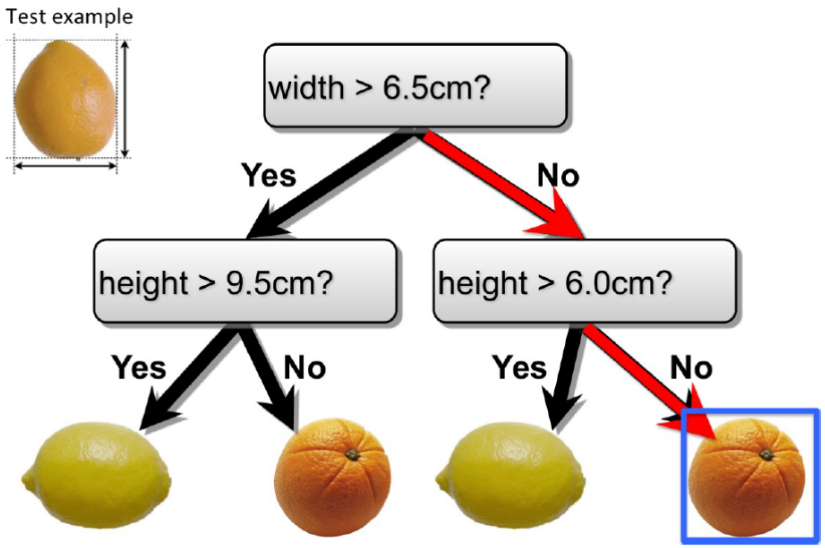
missForest

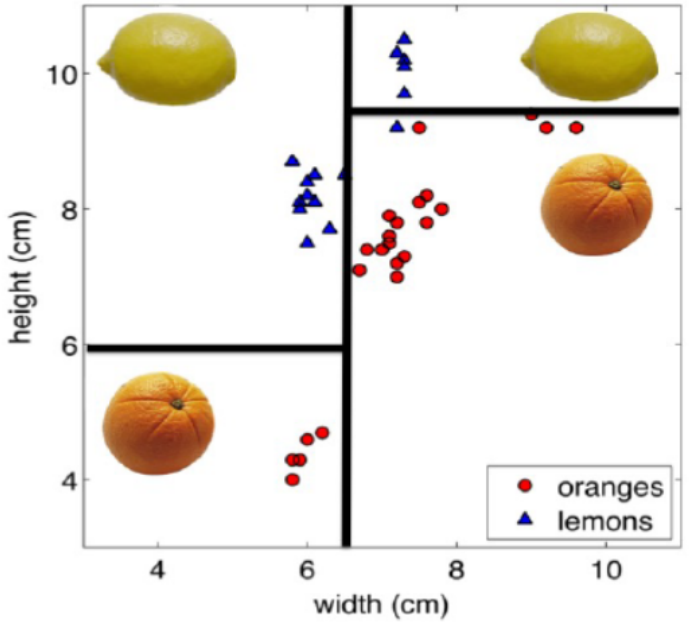
A Random Forest (RF) based imputation

- Start with a guess for the missing values, or keep the missing values (RF)
- For each variable in turn, predict the missing values using a RF with the other variables as targets. Fill in the missing values.
- Iterate this until no change.
- Use the imputed data to construct a predictor.

Decision tree

Example: Orange or lemon?

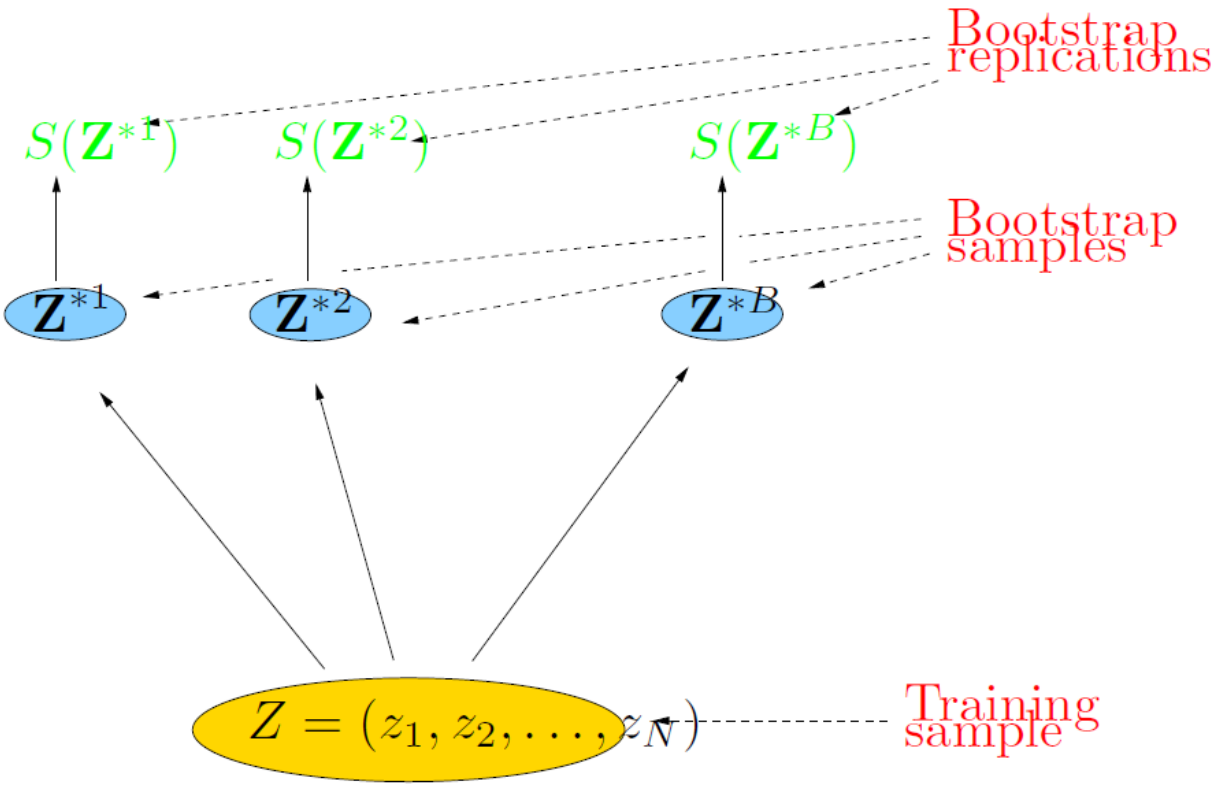




- Internal nodes test attributes
- Branching is determined by attribute values
- Leaf nodes are outputs (class assignments)

The Bootstrap (Efron*)

Resample the data *with replacement* from the actual data



Bootstrap vs CV

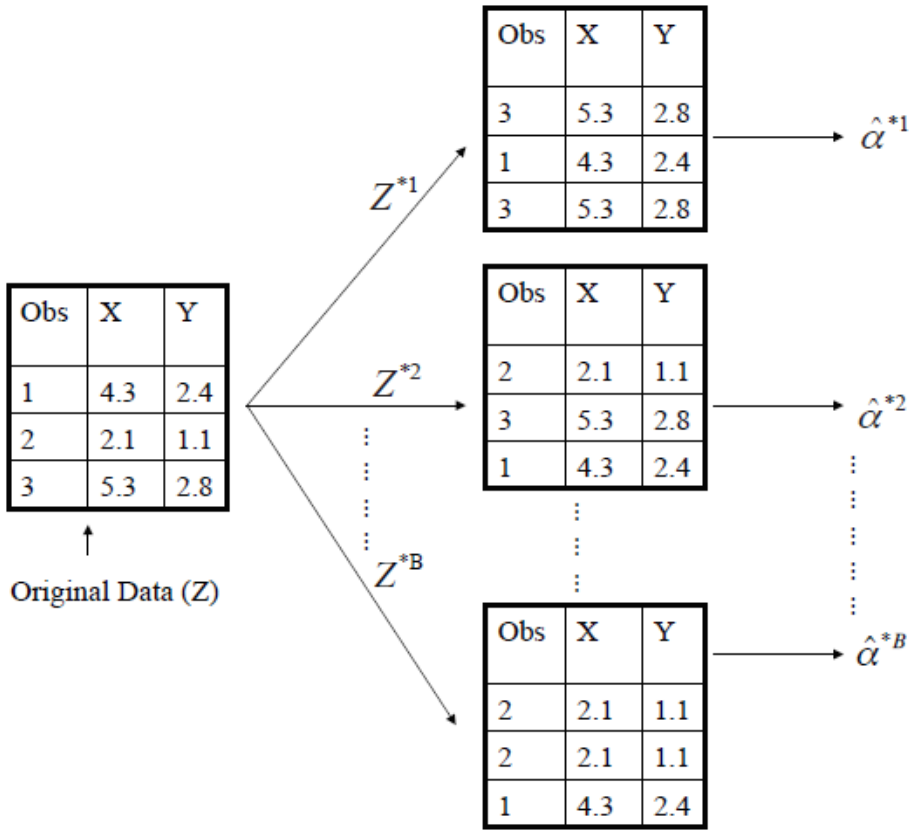
- Bootstrap: principally used to estimate error/uncertainty of parameter estimates, e.g. standard error, etc.
- Cross-validation: principally used to estimate prediction error.

Bagging: Bootstrap aggregation

- Bootstrap, by taking repeated samples from the single training data set,
- Train the method on the bth bootstrapped training set to get $\hat{f}^{*b}(x)$,
- Average all the predictions,

$$\hat{f}_{\mathrm{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x).$$

Schematic of the bootstrap process



We wish to assess the (statistical) accuracy of $S(Z) = \alpha$.

- The standard error of the bootstrap estimates:

$$\sqrt{\frac{1}{B-1}\sum_{r}^{B}(\alpha^{*r} - \frac{1}{B}\sum_{r'}^{B}\hat{\alpha}^{*r'})^2}$$

Out-of-bag (OOB) error estimation

Without performing CV

- On average, each bagged tree makes use of around 2/3 of the observations.
- OOB obs: the remaining 1/3 of the observations not used to fit a given bagged tree
- Predict the response for the ith obs using each of the trees in which that obs was OOB: $B/3$ predictions
- To obtain a single prediction:
  - regression: average these predicted responses,
  - classification: take a majority vote.
- With B sufficiently large, OOB error $\approx$ LOOCV error.
- Very convenient when performing bagging on large data sets.

Random Forest (RF)

- Build a number forest of decision trees on bootstrapped training samples.
- At each tree split, a random sample of m features is drawn, and only those m features are considered for splitting.
- Typically $m = \sqrt{p}$, where p is the number of features
- For each tree grown on a bootstrap sample, the OOB error rate is monitored.
- Improves bagged trees by *decorrelating* trees. Each tree has the same expectation.

Q1: The algorithm is not even allowed to consider a majority of the available predictors ($m = \sqrt{p}$). A rational?
Q2: If $m = p$?

missForest

A Random Forest (RF) based imputation

- Start with a guess for the missing values, or keep the missing values (RF)
- For each variable in turn, predict the missing values using a RF with the other variables as targets. Fill in the missing values.
- Iterate this until no change.
- Use the imputed data to construct a predictor.

(Dis)Advantage

- Works for any type of (mixed) datasets
- Doesn't take account of the uncertainly in the imputation process.

missForest

A Random Forest (RF) based imputation

- Start with a guess for the missing values, or keep the missing values (RF)
- For each variable in turn, predict the missing values using a RF with the other variables as targets. Fill in the missing values.
- Iterate this until no change.
- Use the imputed data to construct a predictor.

(Dis)Advantage

- Works for any type of (mixed) datasets
- Doesn't take account of the uncertainly in the imputation process.

R code for missForest

```
library(missForest)
load(''dat.mis.Rda'')
dat.imp <- missForest(dat.mis)
#check imputed values
dat.imp$ximp
#check imputation error
dat.imp$OOBerror
#comparing actual data accuracy
dat.err <- mixError(dat.imp$ximp, dat.mis, dat)
dat.err

fit.rf <- lm(BMI~Cholesterol, data=dat.imp$ximp)
summary(fit.rf)
```

- NRMSE (normalized mean squared error) : from imputing continuous values.
- PFC (proportion of falsely classified): from imputing categorical values.

Imputation of missing blocks

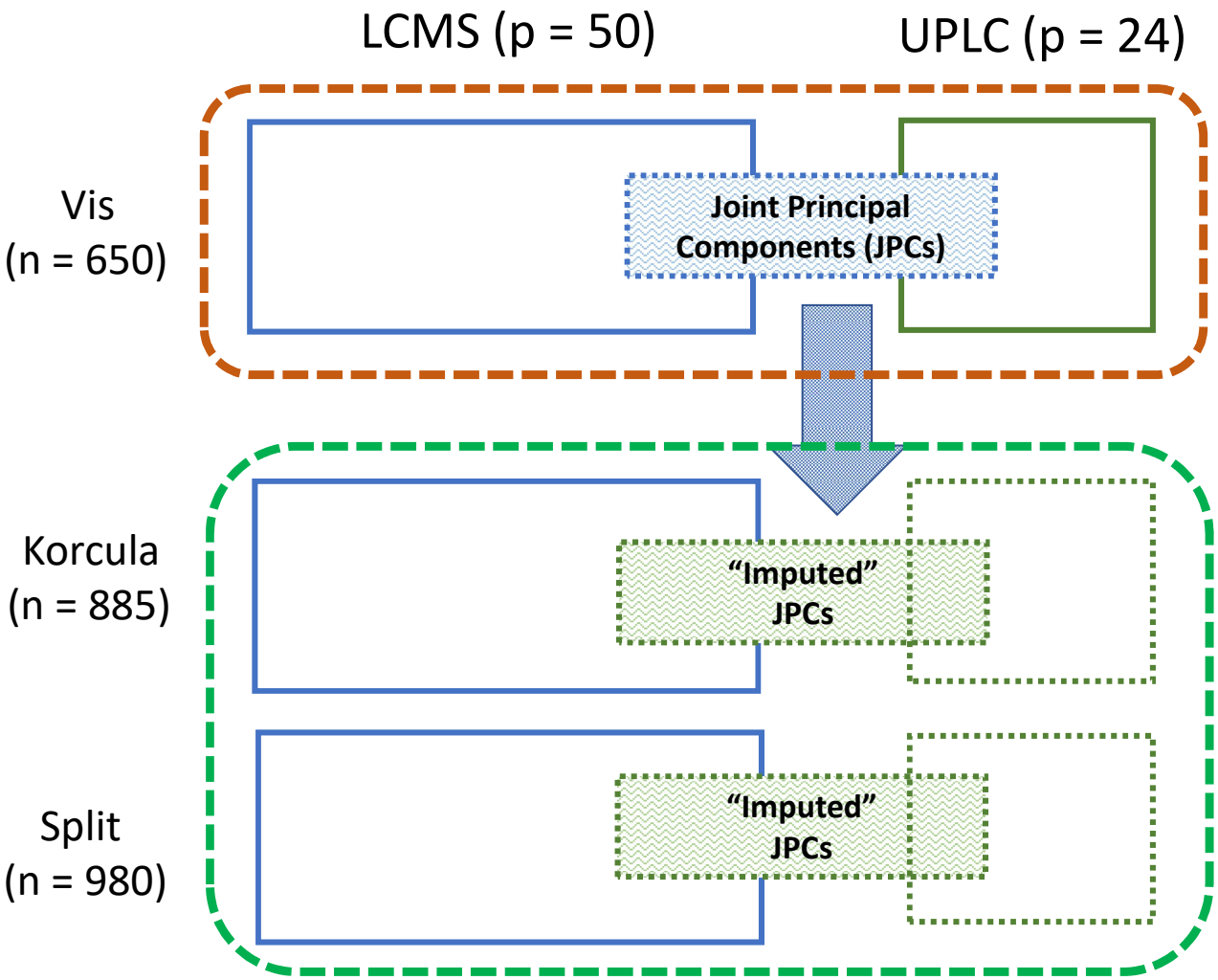
Missing data scenario

	Data 1	Data 2
Subset 1	X_1	Y_1
Subset 2	X_2	NA

What can we do?

- Imputation is easier in a latent space.
- Prediction/imputation based on Probabilistic Partial Least Squares (PPLS): Said el Bouhaddani

Schematic workflow



Horizontal Integration
Using PPLS
to obtain "training rule"

Imputation
Algorithm
using
"training rule"

$$\begin{array}{ccc} X & & Y \\ XW\Downarrow & & \Uparrow UC^{\top} \\ T & \overset{TB}{\Longrightarrow} & U \end{array}$$

Summary

- Visualize summaries or plots for the pattern of missingness.
- If a variable has too many missing values, you may want to exclude it from your analysis.
- When imputing, keep in mind that some variables are restricted to be positive or bounded.
- MI is most convenient under MAR.
- missForest is most convenient for mixed type of data imputation.

Some guidelines for reporting [Carpenter and Kenward]

- The proportion of missing data in key variables should be stated clearly, and possible reasons discussed.
- This information should motivate an analysis valid under the 'MAR' assumption, when using MI.

References

- D B Rubin. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.
- D Rubin. Multiple imputation after 18 years. Journal of the American Statistical Association, 91:473–490, 1996.
- J L Schafer. Analysis of incomplete multivariate data. London: Chapman and Hall, 1997.
- H W Uh et al. Evaluation of regression methods when immunological measurements are constrained by detection limits. BMC Immunol. 9:59, 2008.
- S van Buuren, K Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45 (3), 2011.
- D J Steckhoven, P B´uhlman. MissForest - non-parametric missing value imputation for mixed-type data. Bioinformatics, 28, 112-118, 2012.
- P Schmitt, J Mandel, M Guedj. A Comparison of Six Methods for Missing Data Imputation. J Biom Biostat 6:224, 2015.