Missing values

June 12, 2018

Poverty endangering of children in germany • Coding of missing values • Types of missing values • Consequences • Exploratory missingness analysis • Impact of missingness • Imputation methods • A simulated example • Multiple imputation • EM-Algorithm • Examples • Data augmentation • NORM • Chained equation • Combine estimates • Hints for imputation

Poverty endangering of children in germany

Example 2.16

- 2009 the DIW estimated the poverty risk of german children at 16.3% (using SOEP data for 2005)
- three weeks before the Bundestagswahl (27 Sep 2009) the figure was published by OECD (OECD average: 12.3%)
- heated debate about poverty risk of children in germany
- january 2010 increase of Kindergeld by 20 EUR ($\geq +10\%$)
- 2011 DIW published only a rate of 8.3% for 2005 without any further notice
- problem: in case of non-response the income of a family was set to **zero**
- DIW changed the methodology in 2010

Rademaker, Maike (2011). "Fehlerhafte Statistik: Kinderarmut nur halb so hoch wie gedacht". In: *Financial Times Deutschland*. url: www.ftd.de/politik/deutschland/:fehlerhafte-statistik-kinderarmut-nur-halb-so-hoch-wie-gedacht/60048191.html.

Coding of missing values I

- IEEE Standard for Floating-Point Arithmetic (IEEE 754)

$$(-1)^s \times c \times b^q$$

  - sign $s = 0$ or 1, base/radix $b = 2$ or 10,
  - a non-negative integer significand c and an integer exponent q
  - $-12.345 = (-1)^1 \times 12345 \times 10^{-3}$
- Format requieres special coding for
  - infinitiy: $+\infty$ and $-\infty$.
  - not a number (NaN): a quiet and a signaling one
- Floting points numbers
  - single precision: 1 bit sign, 8 bit exponent, 23 bit significand
  - double precision: 1 bit sign, 11 bit exponent, 52 bit significand
  - quiet NaN: $s = 0$, exponent bits are all set to 1
  - signaling NaN: $s = 1$, exponent bits are all set to 1
  - NaN: unused bits can carry playload, e.g. for source of NaN

Coding of missing values II

Listing 69.1: `example_na.R`

```r
# NA - datum not available
class(NA)
y <- c(1, 2, 3, NA)
is.na(y)
is.nan(y)
mean(y)
mean(y, na.rm=T)
# NaN - invalid operation
class(0/0)
y <- c(1, 2, 3, 0/0)
is.na(y)
is.nan(y)
mean(y)
mean(y, na.rm=T)
```

- `is.na(x)`

⚠ does not support a payload, e.g. for source of NA

- `is.nan(x)`

Types of missing values I

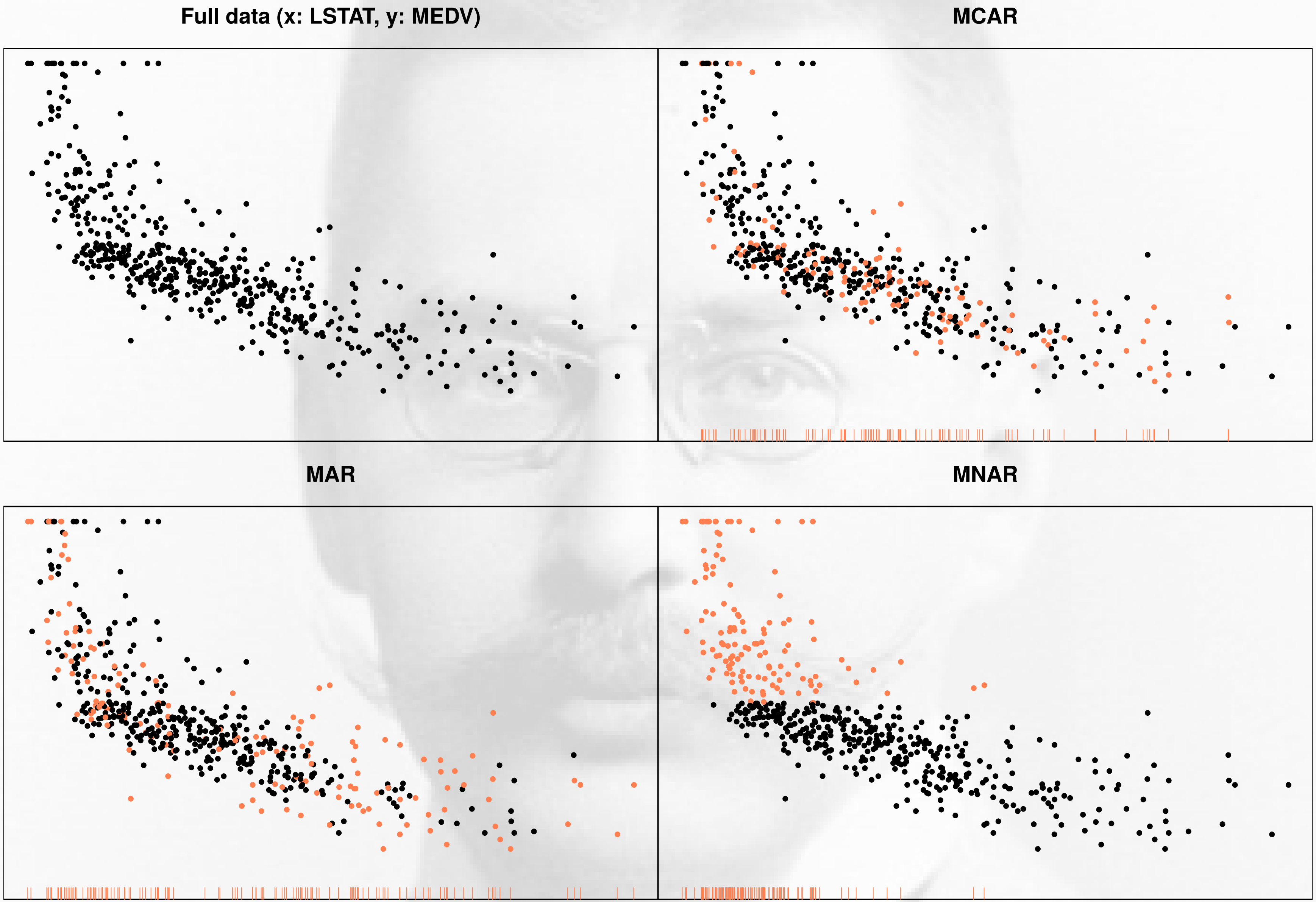
- Missing values can have various reasons
  - shame, (political) correctness
  - ignorance, non-applicable
  - exhaustion
  - misstyping
  - ...
- Types of missings values

$$
\begin{array}{ccc}
\text{can we ignore the reasons?} & \xrightarrow[\textit{No}]{\hspace{2cm}} & \text{MNAR} \\
 & & \text{Missing Not At Random} \\
\Big\downarrow \textit{Yes} & & \\
\text{are the missings completely random?} & \xrightarrow[\textit{No}]{\hspace{2cm}} & \text{MAR} \\
 & & \text{Missing At Random} \\
\Big\downarrow \textit{Yes} & & \\
\text{MCAR} & & \\
\text{Missing Completely At Random} & &
\end{array}
$$

Types of missing values II

- Formal definition:
  X variable with no missings values,
  Y variable with missing values
  - Missing Completely At Random: the missingness does not depend on X nor on Y
  - Missing At Random: the missingness depends on X, but not on Y
  - Missing Not At Random: the missingness depends on Y
- Examples:
  - MCAR: recording error
  - MAR: double sampling, sampling for non-response followup, planned missingness
  - MNAR: sample surveys with non-respondents (not at home, unwilling to answer), respondent out of range (in survival analysis)

Types of missing values III



Consequences

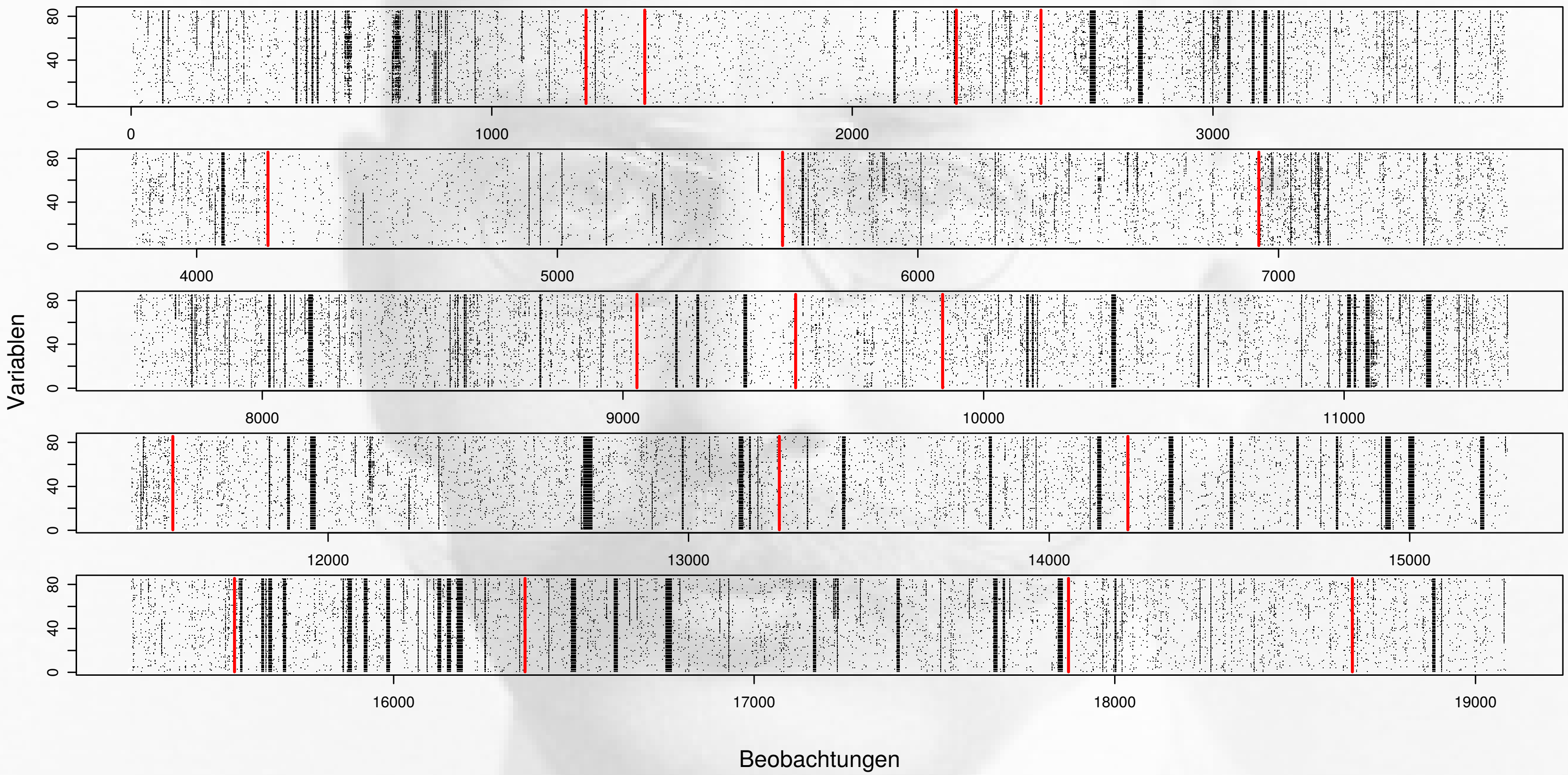
- MCAR
  - can be handeled with approriate methods
  - consistent estimators on the observed data are still consistent
  - may imply larger variances on (parameter) estimates
- MAR
  - there is no possibility to test wheter missings are MAR or MNAR
  - in most cases we can expect a departure from the MAR assumption
  - in many realistic cases a errorneous specification has only minor impact on estimates
- MNAR
  - needs to be modelled explicitly
  - obtain follow up answer from non-respondents

Exploratory missingness analysis I

- Which variables have a lot of missings?
- Which observations have a lot of missings?
  - possibly exclude variables or observations from further analysis
- Bivariate missingness
  - create for each pair of variables a binary variable (0=value not missing, 1=value missing)
  - is the missingness in pairs of variables are dependent?
- Missingness patterns
  - create frequency table of missing patterns
  - do we observe a specific patterns?
- Missingness map
  - use a scatterplot to plot the missing values
  - a point/bar is colored differently if missing or not
  - reorder columns or rows, e.g. after number of missings

Exploratory missingness analysis II

Missingness map for $n \approx 19000$ and $p = 89$



Exploratory missingness analysis III

Listing 69.2: `example_miss_missmap.R`

```r
library("MissingDataGUI") # for brfss data
library("Amelia")
missmap(brfss)
```

Listing 69.3: `example_miss_gui.R`

```r
library("MissingDataGUI") # for brfss data
MissingDataGUI(brfss)
```

- `Amelia::missmap(obj)`
- `MissingDataGUI::MissingDataGUI(obj)`

Impact of missingness

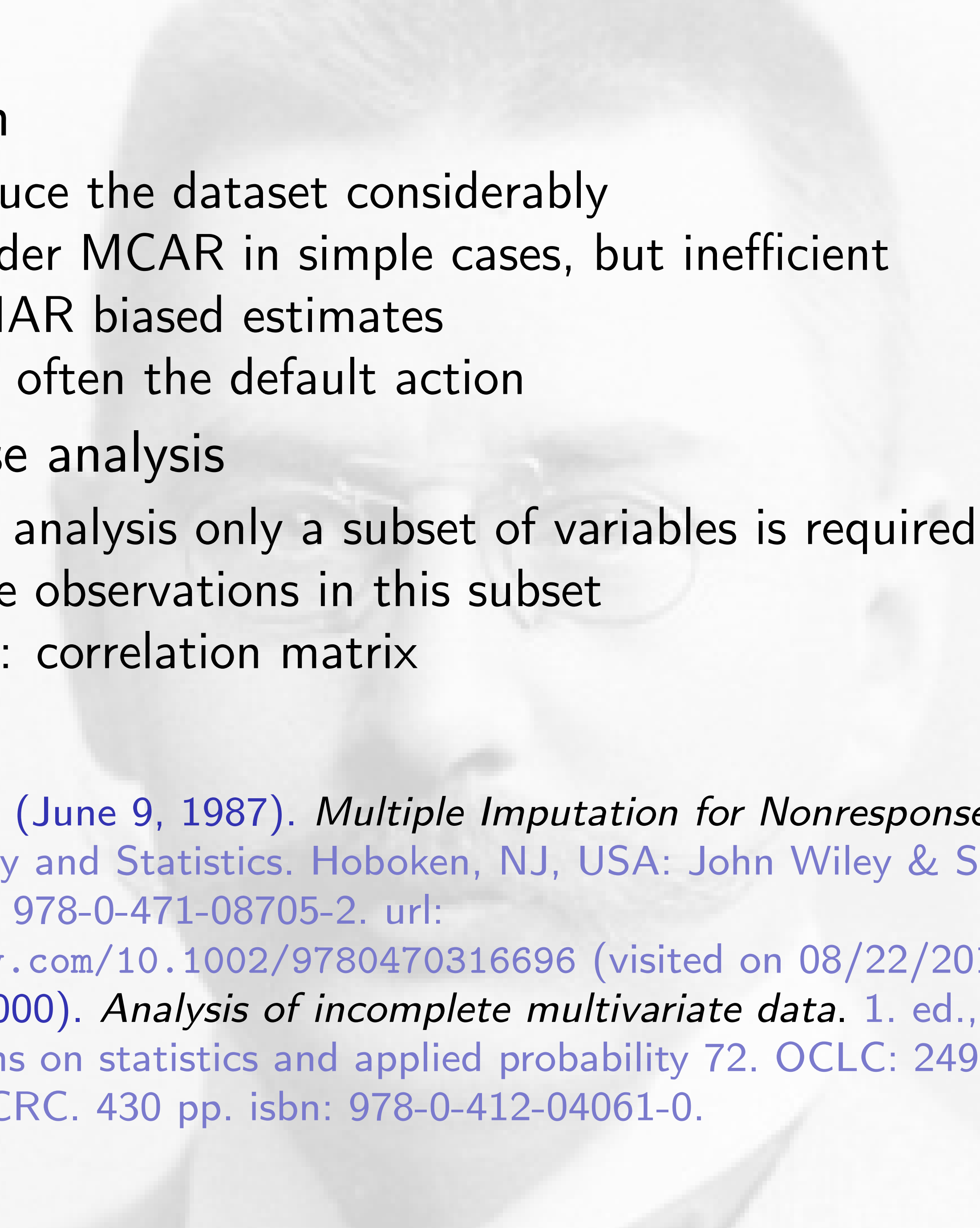
- create for each variable with missings a binary variable (0=value not missing, 1=value missing)
- on other variables apply subgroup analysis based on the binary variable
- if we have differences in the distribution or the parameter $\rightarrow$ MAR or MNAR
- if we have no differences in the distribution or the parameter $\rightarrow$ MCAR
- Little's test compares two models and tests for the MCAR case
- there is no test available to distinguish between MAR and MNAR case

Little, Roderick J. A. (Dec. 1988). "A Test of Missing Completely at Random for Multivariate Data with Missing Values". In: *Journal of the American Statistical Association* 83.404, p. 1198. issn: 01621459. doi: 10.2307/2290157. url: http://www.jstor.org/stable/2290157?origin=crossref (visited on 10/19/2017).

Imputation methods I

- Older methods
  - case deletion/listwise deletion
  - available case analysis
  - reweighting
- Single imputation
  - unconditional mean (item average)
  - unconditional distribution/hot deck
  - conditional mean
  - conditional distribution/predictive distribution
  - Maximum-Likelihood
- Multiple imputation

Imputation methods II



- case deletion
  - may reduce the dataset considerably
  - valid under MCAR in simple cases, but inefficient
  - under MAR biased estimates
  - ⚠ in SPSS often the default action
- available case analysis
  - if for an analysis only a subset of variables is required then use all complete observations in this subset
  - example: correlation matrix

Rubin, Donald B., ed. (June 9, 1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc. isbn: 978-0-470-31669-6 978-0-471-08705-2. url: http://doi.wiley.com/10.1002/9780470316696 (visited on 08/22/2016).

Schafer, Joseph L. (2000). *Analysis of incomplete multivariate data*. 1. ed., 1. CRC Press reprint. Monographs on statistics and applied probability 72. OCLC: 249266966. Boca Raton: Chapman & Hall/CRC. 430 pp. isbn: 978-0-412-04061-0.

Imputation methods III

Listing 71.1: `example_missing.R`

```
library("foreign")
x <- read.spss("ALLBUS2012.SAV", use.value.labels=FALSE, to.data.frame=
trust <- as.data.frame(x[,52:63])
# count number of missing values
r <- is.na(trust)
# no. per column
apply(r, 2, sum)
# no. per row
apply(r, 1, sum)
# number of complete cases
cc <- complete.cases(trust)
sum(cc)
# filter functions
head(na.omit(trust))
try(na.fail(trust))
```

- `complete.cases(x, ...)`
- `na.omit(x)`
- `na.fail(x)`

Imputation methods IV

Listing 71.2: `example_impute1.R`

```
library("foreign")
x <- read.spss("ALLBUS2012.SAV", to.data.frame=T)
body <- as.data.frame(x[,c(220,593,595)])
names(body) <- c("age", "height", "weight")
# number of NAs
nabody <- is.na(body)
apply(nabody, 2, sum)
# full data
mean(body$weight)
cor(body)
# case deletion
mean(body$weight, na.rm=T)
cor(body, use="complete.obs")
sum(complete.cases(body))
# available case analysis
cor(body, use="pairwise.complete.obs")
crossprod(!nabody)
```

A simulated example I

- X systolic blood pressure measured in january
- Y systolic blood pressure measured in february (follow-up)
- data simulated from

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 125 \\ 125 \end{pmatrix}, \begin{pmatrix} 625 & 375 \\ 375 & 625 \end{pmatrix}\right), \rho = 0.6$$

- MCAR on Y: select randomly from the Y observations
- MAR on Y: second reading if $X > 140$
- MNAR on Y: second recording if $Y > 140$
  - other possibility: second recording only when first and second reading differ
- Parameters to be analyzed: μ_Y, σ_Y, ρ, $\beta_{Y|X}$, $\beta_{X|Y}$

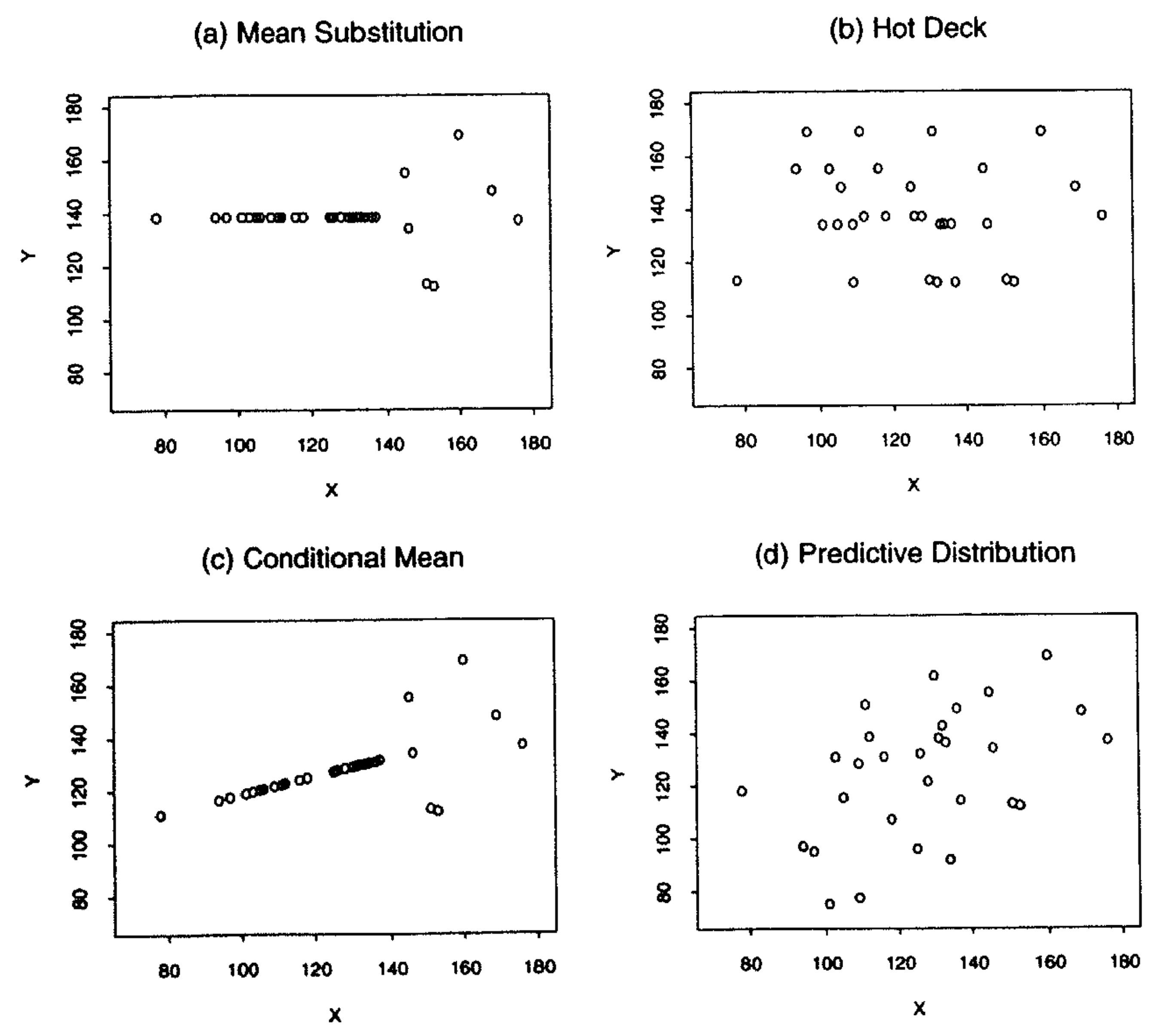
Schafer, Joseph L. and Graham, John W. (2002). "Missing data: Our view of the state of the art.". In: *Psychological Methods* 7.2, pp. 147–177. issn: 1939-1463, 1082-989X. doi: 10.1037/1082-989X.7.2.147. url: http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.7.2.147 (visited on 08/22/2016).

A simulated example II

Performance of Listwise Deletion for Parameter Estimates and Confidence Intervals Over 1,000 Samples (N = 50 Participants)

Parameter	MCAR	MAR	MNAR
Average parameter estimate (with RMSE in parentheses)			
$\mu_Y = 125.0$	125.0 (6.95)	**143.3** (19.3)	**155.5** (30.7)
$\sigma_Y = 25.0$	24.6 (5.26)	**20.9** (5.84)	**12.2** (13.2)
$\rho = .60$	.59 (.19)	**.33** (.37)	**.34** (.36)
$\beta_{Y\|X} = .60$	.61 (.27)	.60 (.51)	**.21** (.43)
$\beta_{X\|Y} = .60$	.60 (.25)	**.20** (.44)	.60 (.52)
Coverage (with average interval width in parentheses)			
μ_Y	94.3 (30.0)	**18.8** (25.0)	**0.0** (14.7)
σ_Y	94.3 (23.3)	90.7 (19.4)	**17.4** (11.4)
ρ	95.4 (0.76)	**82.5** (0.93)	**82.7** (0.94)
$\beta_{Y\|X}$	94.6 (1.10)	95.9 (2.20)	**40.0** (0.73)
$\beta_{X\|Y}$	95.3 (1.08)	**37.7** (0.71)	96.6 (2.23)

A simulated example III



A simulated example IV

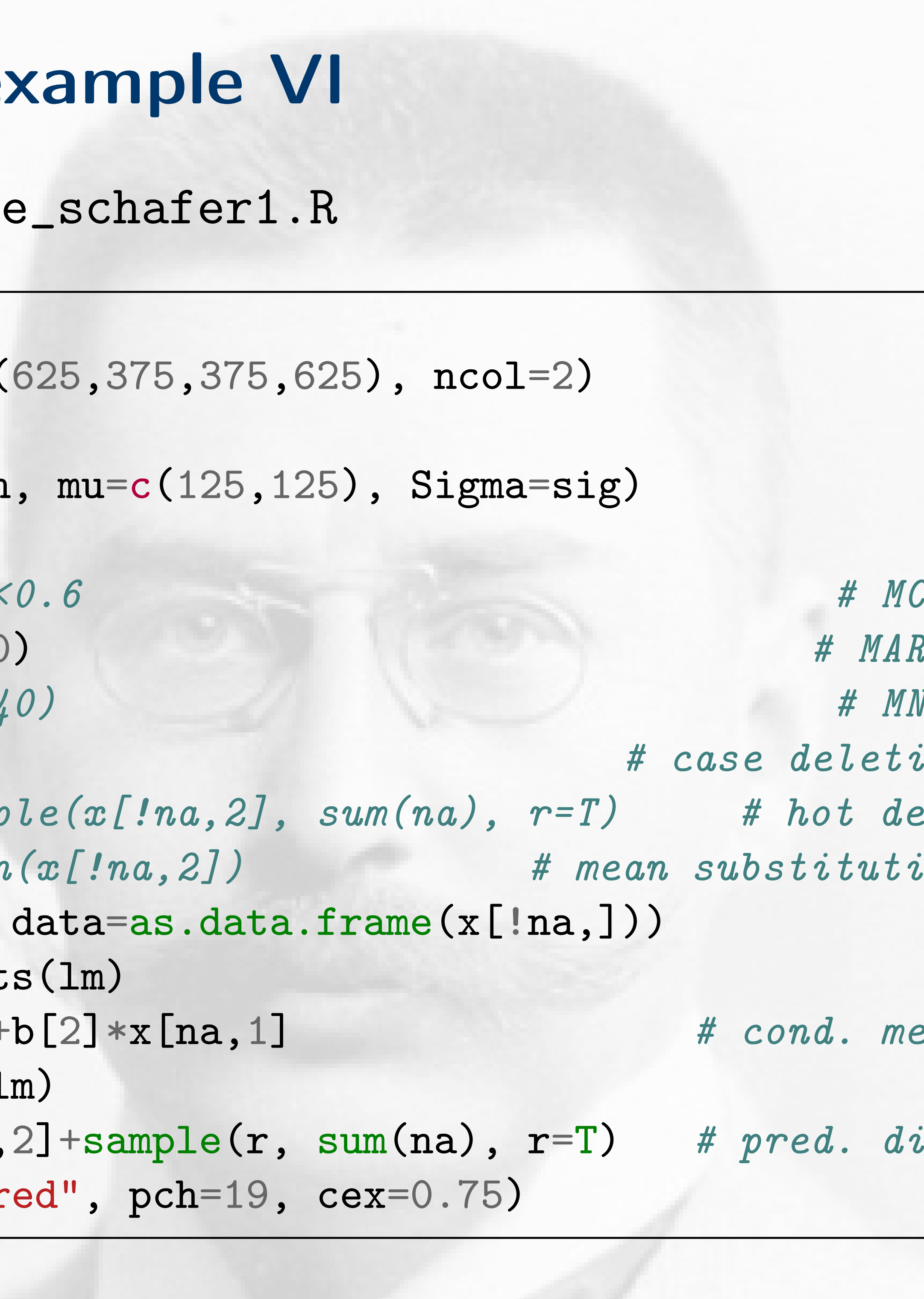
Table 3
Performance of Single-Imputation Methods for Parameter Estimates and Confidence Intervals Over 1,000 Samples (N = 50 Participants)

Parameter	MCAR				MAR				MNAR			
	MS	HD	CM	PD	MS	HD	CM	PD	MS	HD	CM	PD
	Average parameter estimate (with RMSE in parentheses)											
$\mu_Y = 125.0$	125.1 (7.18)	125.2 (7.89)	125.2 (6.26)	125.1 (6.57)	143.5 (19.4)	143.5 (19.5)	124.9 (18.1)	124.8 (18.3)	155.5 (30.7)	155.5 (30.73)	151.6 (26.9)	151.6 (26.9)
$\sigma_Y = 25.0$	12.3 (13.0)	23.4 (5.40)	18.2 (8.57)	24.7 (5.37)	10.6 (14.6)	20.0 (6.68)	20.4 (10.7)	27.0 (8.77)	6.20 (18.9)	11.7 (13.7)	8.42 (16.9)	12.9 (12.7)
$\rho = .60$	.30 (.32)	.16 (.46)	.79 (.27)	.59 (.20)	.08 (.52)	.04 (.57)	.64 (.48)	.50 (.40)	.15 (.47)	.08 (.53)	.55 (.40)	.38 (.37)
$\beta_{Y\mid X} = .60$	.16 (.45)	.16 (.47)	.61 (.25)	.60 (.27)	.04 (.56)	.04 (.57)	.61 (.57)	.62 (.57)	.04 (.56)	.04 (.56)	.21 (.43)	.21 (.43)
$\beta_{X\mid Y} = .60$	.61 (.26)	.17 (.46)	1.12 (.64)	.60 (.24)	.20 (.44)	.06 (.56)	.78 (.75)	.45 (.40)	.61 (.55)	.19 (.53)	1.63 (1.72)	.76 (.68)
	Coverage (with average interval width in parentheses)											
μ_Y	39.2 (7.0)	60.0 (13.3)	58.5 (10.4)	71.0 (14.1)	0.2 (6.0)	2.4 (11.4)	25.7 (11.6)	32.3 (15.3)	0.0 (3.5)	0.0 (6.7)	0.0 (4.8)	0.0 (7.3)
σ_Y	0.7 (5.1)	63.7 (9.6)	31.3 (7.5)	65.4 (10.2)	0.1 (4.4)	45.3 (8.2)	30.0 (8.4)	49.4 (11.1)	0.0 (2.5)	1.7 (4.8)	0.7 (3.5)	4.4 (5.3)
ρ	25.5 (0.50)	5.5 (0.53)	21.7 (0.19)	65.0 (0.35)	0.0 (0.55)	0.0 (0.55)	19.6 (0.21)	40.7 (0.34)	2.2 (0.54)	0.5 (0.54)	37.6 (0.31)	50.0 (0.43)
$\beta_{Y\mid X}$	1.2 (0.27)	16.5 (0.54)	38.6 (0.22)	63.5 (0.44)	0.0 (0.25)	0.8 (0.47)	17.2 (0.22)	33.5 (0.45)	0.0 (0.14)	0.1 (0.27)	3.1 (0.13)	7.4 (0.26)
$\beta_{X\mid Y}$	98.1 (1.18)	23.9 (0.63)	8.9 (0.50)	71.1 (0.47)	91.2 (1.43)	14.5 (0.75)	18.6 (0.56)	60.0 (0.46)	97.4 (2.50)	71.3 (1.30)	19.1 (1.46)	56.2 (1.05)

A simulated example V

- table explanations
  - RMSE = Root Mean Squared Error
  - if parameter standard error estimate is one half of absolute size of parameter then serious bias in estimate, marked in bold
  - coverage = percentage of confidence intervals covering the parameters (should be 95%), less than 90% are in bold
- MNAR
  - none of the methods work
  - the unbiased estimate of $\beta_{X|Y}$ is an artefact (normal distribution)
- MAR/MCAR
  - Predictive distribution works
  - Mean subsitution works only for estimating μ_Y
- partially severe undercoverage
  - also due to less information

A simulated example VI



Listing 72.1: example_schafer1.R

```
library("MASS")
sig <- matrix(c(625,375,375,625), ncol=2)
n   <- 50
x   <- mvrnorm(n, mu=c(125,125), Sigma=sig)
plot(x)
#na <- runif(n)<0.6                                  # MCAR
na <- (x[,1]<140)                                   # MAR
#na <- (x[,2]<140)                                   # MNAR
#x <- x[!na,]                               # case deletion
#x[na,2] <- sample(x[!na,2], sum(na), r=T)       # hot deck
#x[na,2] <- mean(x[!na,2])               # mean substitution
lm <- lm(V2~V1, data=as.data.frame(x[!na,]))
b <- coefficients(lm)
x[na,2] <- b[1]+b[2]*x[na,1]                     # cond. mean
r <- residuals(lm)
x[na,2] <- x[na,2]+sample(r, sum(na), r=T)     # pred. dist
points(x, col="red", pch=19, cex=0.75)
```

Multiple imputation I

- Maximum-Likelihood
  - requires additional information
  - delivers better results (confidence intervals)
- EM-Algorithm
  - consider missing values as random variables
  - estimate unknown parameters via ML-Method
  - repeat the last two steps until estimate stabilizes
- Multiple imputation
  - fill missing values by data augmentation
  - does not help under MNAR
  - but provides at least information about the variability

Multiple imputation II

$$
\begin{tikzcd}
 & \text{imputed data set 1} \arrow[r] & \text{analysis result 1} \arrow[dr] & \\
\text{data set} \arrow[ur] \arrow[r] \arrow[dr] & \ldots\ldots \arrow[r] & \ldots\ldots \arrow[r] & \text{averaged result} \\
 & \text{imputed data set m} \arrow[r] & \text{analysis result m} \arrow[ur] &
\end{tikzcd}
$$

- average results $\hat{\theta}_{MI} = \frac{1}{m}\sum_i \hat{\theta}_i$
- $\widehat{Var}(\hat{\theta}_i)$ can be computed easily
- m odd, not too large ($m = 3, 5, 7, 11$)

EM-Algorithm I

- Instead of full loglikelihood

$$l(\theta) = \sum_{\text{non-miss}} l_i(\theta) + \sum_{\text{miss}} l_i(\theta)$$

- Maximize $Q(\theta, \theta_t)$, the expectation over the missing observations

$$\begin{aligned}
Q(\theta, \theta_t) &= \int_{\text{miss}} l(\theta) \prod_{\text{miss}} f_i(\theta_t) d\text{miss} \\
&= \sum_{\text{non-miss}} l_i(\theta) + \sum_{\text{miss}} \int_{\text{miss}} l_i(\theta) \prod_{\text{miss}} f_i(\theta_t) d\text{miss} \\
&= \sum_{\text{non-miss}} l_i(\theta) + \underbrace{\sum_{\text{miss}} \int_{\text{miss}_i} l_i(\theta) f_i(\theta_t) d\text{miss}_i}_{H(\theta,\theta_t)}
\end{aligned}$$

EM-Algorithm II

Example 2.17

$$
\begin{aligned}
X_i \sim N(\mu;1) \quad &\rightarrow \quad f_i(\mu) = \exp(-0.5(x_i-\mu)^2) \\
L(\mu) &= \exp(-0.5(x_1-\mu)^2)\exp(-0.5(x_2-\mu)^2) \\
l(\mu) &= -0.5(x_1-\mu)^2 - 0.5(x_2-\mu)^2 \\
& \quad \text{2 obs., } x_2 \text{ missing} \\
Q(\mu,\mu_t) &= \int_{-\infty}^{\infty} (-0.5(x_1-\mu)^2) - 0.5(x_2-\mu)^2) * \\
& \exp(-0.5(x_2-\mu_t)^2)dx_2 \\
&= -0.5(x_1-\mu)^2 \\
& -0.5 \underbrace{\int_{-\infty}^{\infty} (x_2-\mu)^2 \exp(-0.5(x_2-\mu_t)^2)dx_2}_{H(\mu,\mu_t)} \\
\mu_{t+1} &= \max_{\mu} Q(\mu,\mu_t)
\end{aligned}
$$

EM-Algorithm III

- Initialize
  - find a prelimnary estimate θ_0, e.g. from non-missing observations
- E-Step: $E(l(X)) = \int l(x)f(x)dx$
  - compute $H(\theta, \theta_t)$
- M-Step:
  - compute θ_{t+1} as $max_\theta Q(\theta, \theta_t)$ (M-step)
- Repeat until estimate stabilizes (converges)
- Maximizing $l(\theta)$ is the "same" as maximizing $Q(\theta, \theta_t)$
  - it can be shown that the Kullback-Leibler-Divergence does not increase in the E-step
  - holds, e.g. if data are MAR and parameter of interest ψ are separate from the parameter(s) η which govern missingness ($\theta = (\psi, \eta)$)

Examples I

- $Y \sim N(\mu, \psi)$ with $\psi = \sigma^2$
- Full Maximum likehood estimator

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\psi} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \hat{\mu}^2$$

- EM-Iterations

$$\begin{aligned}
\hat{\mu}_{(t+1)} &= \frac{1}{n}\left(\sum_{\text{non-miss}} y_i + n_{\text{miss}}\hat{\mu}_{(t)}\right) \\
\hat{\psi}_{(t+1)} &= \frac{1}{n}\left(\sum_{\text{non-miss}} y_i^2 + n_{\text{miss}}(\hat{\psi}_{(t)} + \hat{\mu}^2_{(t)})\right) - \hat{\mu}^2_{(t+1)}
\end{aligned}$$

Examples II

- $Y \sim N(\mu, \Sigma)$ bivariate with second variable has missing values
- Full loglikelihood

$$-\frac{n}{2}\log \mid 2\pi\Sigma \mid -\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)$$

- Parameters to estimate: μ_i, σ_{ij} with $i, j = 1, 2$
- decompose $f_i(y_1, y_2) = f_i(y_1, \mu_1, \sigma_{11}) f_i(y_2, y_1, \beta_0, \beta_1, \sigma_r)$ for missing observations
- maximization and integrations leads to a linear regresson between y_1 and y_2

Examples III

- the following relationship is used

$$\begin{aligned}\beta_1 &= \sigma_{12}/\sigma_{11} \text{ (slope)}\\ \beta_0 &= \mu_2 - \beta_1\mu_1 \text{ (intercept)}\\ \sigma_r &= \sigma_{22} - \sigma_{12}^2/\sigma_{11} \text{ (std.dev. residuals)}\end{aligned}$$

- and can be estimated in closed form
  - $\hat{\mu}_1$ and $\hat{\sigma}_{11}$ with standard ML estimator
  - $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}_r$ only from non-missing observations
- e.g. μ_2 is estimated by

$$\hat{\mu}_2 = \frac{1}{n}\left(\sum_{\text{non-miss}} y_{i2} + \sum_{\text{miss}} \hat{y}_{i2}\right)$$

Examples IV

- Two binary variables with missings on both variables

	parameter		observed		
	$Y_2 = 0$	$Y_2 = 1$	$Y_2 = 0$	$Y_2 = 1$	
$Y_1 = 0$	θ_{11}	θ_{12}	n_{11}	n_{12}	n_{1+}
$Y_1 = 1$	θ_{21}	θ_{22}	n_{21}	n_{22}	n_{2+}
			n_{+1}	n_{+2}	n

- Full loglikelihood

$$l(\theta) = n_{11}\log(\theta_{11}) + n_{12}\log(\theta_{12}) + n_{21}\log(\theta_{22}) + n_{22}\log(\theta_{22})$$

- Full Maximum likehood estimator

$$\hat{\theta}_{ij} = \frac{n_{ij}}{n}$$

Examples V

- n_{ij}^{A} non-missing observation
- n_{i+}^{B} missing observations in Y_2
- n_{+j}^{C} missing observations in Y_1
- EM-Iterations

$$\theta_{ij}^{(t+1)} = \frac{1}{n}\left(n_{ij}^{A} + n_{i+}^{B}\frac{\theta_{ij}^{(t)}}{\theta_{i+}^{(t)}} + n_{+j}^{C}\frac{\theta_{ij}^{(t)}}{\theta_{+j}^{(t)}} \right)$$

Data augmentation

- Idea of data augmentation
  - EM algorithm delivers an estimate $\hat{\theta}$
  - $\hat{\Theta}$ is random $\Longrightarrow$ $\hat{\Theta}$ has a distribution
  - draw m "plausible" $\hat{\theta}$'s from the distribution
- Markov-Chain-Monte-Carlo (MCMC) method
  - get a start estimate $\hat{\theta}_i^{(0)}$
  - impute from $\hat{\theta}_i^{(k)}$ and get a complete dataset
  - recompute an estimate $\hat{\theta}_i^{(k+1)}$ and repeat
  - $\hat{\theta}_i^{(k)}$ iterates through the distribution of $\hat{\theta}$
  - choose m "plausible" $\hat{\theta}$ (imputed datasets)

A. P. Dempster N. M. Laird, D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, pp. 1–38. issn: 00359246. url: http://www.jstor.org/stable/2984875.

Tanner, Martin A. and Wong, Wing Hung (June 1987). "The Calculation of Posterior Distributions by Data Augmentation". In: *Journal of the American Statistical Association* 82.398, p. 528. issn: 01621459. doi: 10.2307/2289457. url: http://www.jstor.org/stable/2289457?origin=crossref (visited on 08/22/2016).

NORM I

- Assumes data are multivariate normal distributed

$$X \sim N(\mu, \Sigma)$$

- Estimate μ and Σ, e.g. with EM algorithm
- Impute values by random draws from a multivariate normal distribution with estimated μ and Σ
- Reestimate μ and Σ until estimates converge
- Performs well in practice, even for ordinal data
- package `Amelia` (after Amelia Earhart)
  - Schafer developed a windows program `NORM`

NORM II

Listing 73.1: example_amelia.R

```
# run example_mar.R before
library("Amelia")
library("mitools")
# run NORM
aobj <- amelia(xmar, noms=c("chas","rad"))
# compute linear regressions
models <- lapply(aobj$imputations, function(x) {
  lm(medv~lstat, data=x)
})
# look at one model
summary(models[[1]])
# extract
beta <- MIextract(models, fun="coef")
vcov <- MIextract(models, fun="vcov")
summary(MIcombine(beta, vcov))
summary(MIcombine(models))
```

```
Amelia::amelia(x, m=5, parallel=c("no", "multicore", "snow"), noms,
    ords)
```

Chained equation I

- multiple imputations by chained equation
  - applies iteratively regression methods for imputing each variable

`numeric`	various linear regression models, mean/class-mean imputation, quadratic term regression model
`factor`, 2 levels	logistic regression
`factor`, >= 2 levels	polytomous logistic regression linear discriminant analysis
`ordered`, any	proportional odds model predictive mean matching, sample

- theoretical properties are not fully explored, but performs well in practice

Chained equation II

Listing 73.2: example_mice.R

```
# run example_mar.R before
source(list.files(pattern='example_mar.R', recursive=TRUE)[1])
#
library("mice")
# run NORM
xmar$chas <- factor(xmar$chas)
xmar$rad <- factor(xmar$rad)
mobj <- mice(xmar)
# compute linear regressions
models = list()
for (i in 1:5) models[[i]] <- lm(medv~lstat,
                                 data=complete(mobj, i))
# look at one model
summary(models[[1]])
```

```
mice::mice(data, m=5, method=vector("character",length=ncol(data)),
           predictorMatrix=(1-diag(1,ncol(data))))
```

```
mice::complete(x, action=1)
```

Combine estimates I

- multiple imputation estimate can easily combined by averaging
- allows also for variance estimation

$$\hat{\theta}_{MI} = \frac{1}{m}\sum_i \hat{\theta}_i$$

$$\widehat{Var}(\hat{\theta}_{MI}) = \underbrace{\frac{1}{m}\sum_i \widehat{Var}(\hat{\theta}_i)}_{=\text{within}} + \left(1-\frac{1}{m}\right) \underbrace{\frac{1}{m-1}\sum_i \left(\hat{\theta}_i - \hat{\theta}_{MI}\right)^2}_{=\text{between}}$$

Combine estimates II

Listing 73.3: `example_mi.R`

```r
# run example_mar.R/example_mice.R before
source(list.files(pattern='example_mar.R', recursive=TRUE)[1])
source(list.files(pattern='example_mice.R', recursive=TRUE)[1])
#
library("mitools")
# extract
beta <- MIextract(models, fun="coef")
vcov <- MIextract(models, fun="vcov")
summary(MIcombine(beta, vcov))
summary(MIcombine(models))
```

- `mitools::imputationList(datasets, ...)`
- `mitools::MIextract(results, expr, fun)`
- `mitools::MIcombine(results, ...)`

Hints for imputation

- Don't round off imputations for dummy variables. May introduce bias in estimation process.
- Don't transform skewed variables. Changes relationship between variables, might impute outliers.
- Use more imputations. Rule-of-thumb: number of imputations = percent of missing values.
- Create multiplicative terms before imputing. Creating the multiplicative terms after imputation may bias the regression parameters of the multiplicative term.
- Alternatives to multiple imputation aren't usually better. Other techniques (e.g. listwise deletion) impose more stringent assumptions than MAR.

Source: /www.theanalysisfactor.com/multiple-imputation-5-recent-findings-that-change-how-to-use-it