Multivariate Analysis

Estimation with Missing Not at Random Data

Guillaume Blanc, Phd. Candidate,
University of Geneva

February 2022

Multivariate Data with Missing Values

	Y1	Y2	Y3
1	-0.34	-0.64	0.78
2	-1.16	-2.95	1.94
3	*NA*	-0.59	0.19
4	*NA*	-0.20	0.24
5	-1.75	-2.38	2.39
6	-0.99	-2.35	0.94
7	-1.30	-1.28	2.10
8	-1.32	-2.74	1.86
9	-0.55	0.06	0.59
10	-0.70	-1.54	2.41
11	*NA*	-0.42	1.08

	M1	M2	M3
1	0	0	0
2	0	0	0
3	1	0	0
4	1	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	0	0	0
11	1	0	0

Figure: Data with missing values and the mask.

Goal: estimate $\boldsymbol{\mu}$

Problem: missing values

Solution:

1. Impute by the mean (?!)
2. Correct the bias.

Requires:

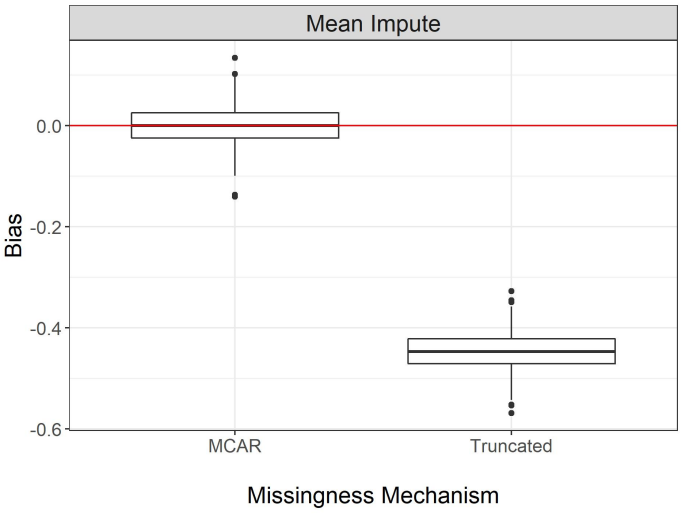
- Generative model for $\boldsymbol{Y}$
- Missingness mechanism

Experiment: Imputation by the Mean

Experiment: Simulate 1000 datasets with missing data. For each, get estimates of μ_1.

Two settings:

- MCAR : Missing Completely at Random.
- Truncated : Only smallest 50% of the data are observed.



Generative Model for $\boldsymbol{Y}$: Factor Analysis / PPCA

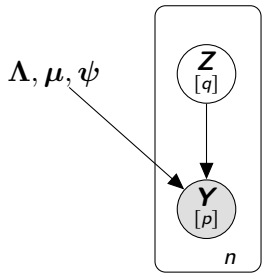


Figure: Generative model of the data.

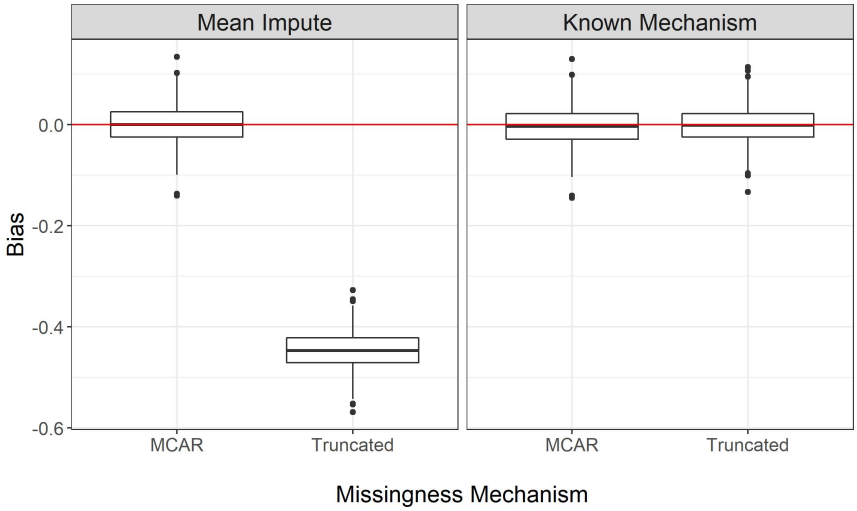
Assumptions: for $j, j' = 1, \ldots, p$:

- $Y_j^{obs} \perp Y_{j'}^{obs} | \boldsymbol{Z}$ whenever $j \neq j'$,
- $Y_j^{obs} | \boldsymbol{Z} = \boldsymbol{z} \sim N(\boldsymbol{z}^\top \boldsymbol{\Lambda}_{j\cdot} + \mu_j, \psi_j)$.

Consequence: the marginal distribution is

$$\boldsymbol{Y}^{obs} \sim MN(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\psi}).$$

Bias Correction with Known Missingness Mechanism



Modeling the Missingness Mechanism

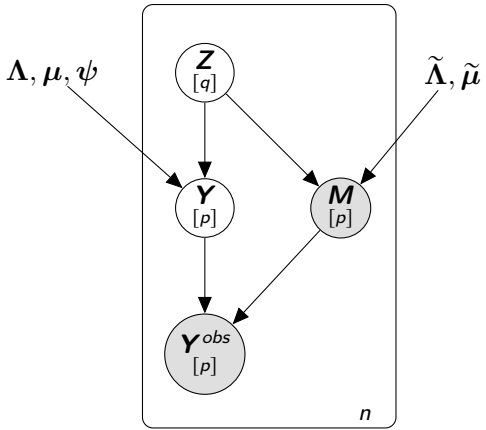


Figure: Generative model of the data and missingness mechanism.

Assumptions: for $j, j' = 1, \ldots, p$,

- $Y_j^{obs} \perp Y_{j'}^{obs} | \boldsymbol{Z}$ whenever $j \neq j'$,
- $Y_j^{obs} | \boldsymbol{Z} = \boldsymbol{z} \sim N(\boldsymbol{z}^\top \boldsymbol{\Lambda}_{j\cdot} + \mu_j, \psi_j)$,
- $M_j \perp M_{j'} | \boldsymbol{Z}$ whenever $j \neq j'$,
- $M_j | \boldsymbol{Z} = \boldsymbol{z} \sim \text{Bernoulli}(\text{sigmoid}(\boldsymbol{z}^\top \widetilde{\boldsymbol{\Lambda}}_{j\cdot} + \widetilde{\mu}_j))$.

Bias Correction with Estimated Missingness Mechanism