# Imputation for High Dimensional Data
## A comprehensive review

Edoardo Costantini

July 2020

# 1 Introduction

This is an amazing introduction. But keep reading for the methods

# 2 Methods

Here I describe the experiments

## 2.1 Experiment 1: Dimensionality and proportion of missingness

**Data Generation** The data was generated according to the standard normal multivariate model:

$$X = MVN(\mu_0 = \mathbf{0}, \mathbf{\Sigma_0}) \tag{1}$$

where $\mu_0$ is a $1 \times p$ vector of 0s, and $\Sigma_0$ is a correlation matrix. Variables were divided in three correlation blocks: high, mid and low correlation. Variables in block 1 are correlated among themselves with a correlation coefficient $\rho_1 = .6$. Variables in block 2 are correlated among themselves and with variables in block 1 with $\rho_2 = .3$. Variables in block 3 are correlated among themselves and with variables in block 1 and 2 with $\rho_3 = .01$.

| | $Block1$ | $Block2$ | $Block3$ |
|---|---|---|---|
| $Block1$ | $\rho_1$ | $\rho_2$ | $\rho_3$ |
| $Block2$ | $\rho_2$ | $\rho_2$ | $\rho_3$ |
| $Block3$ | $\rho_3$ | $\rho_3$ | $\rho_3$ |

Block 1 and 2 are made up of 10 variables each, and all the remaining variables belong to block 3.

After sampling the values of X from equation 1, all columns were rescaled to match mean, variances, and covariances of continuously treated items in EVS waves. Typical 10 points EVS items have means and variances around 5

(instead of 0). Hence, each variable in X was centred around 5 and scaled to have variance around 5 (instead of 1).

**Missing Data** Missing values were imposed on six variables, three in block 1 and three in block 2, using the following probit model

$$P(y_t = MISS|X) = \Phi(\tilde{X}\theta) \tag{2}$$

where $y_t$ is a variable target of missing values imposition ($t = 1, ..., 6$), $\Phi$ is the cumulative normal distribution, $\tilde{X}$ is a subset of $X$ including 4 determinants of missingness, and $\theta$ is a vector of regression coefficients. In Experiment 1, $\theta = (-1, -1, .667, -.333)$.

The probit model strategy facilitates manipulation of the proportion of missing cases. Defining ...

**Conditions** This procedure was repeated 500 times for each of 4 conditions. Two factors defined the experimental conditions: number of features and proportion of missing cases (per vairable). The number of features was either 50 or 500, representing a condition of low and high dimensionality respectively (the sample size $n$ was fixed at 200). The proportion of missing cases varied between .1 and .3. Table ... summarises the conditions of experiment 1:

Table 1: Experiment 1 conditions ($n = 200$)

| Cond | pm | p |
|---|---|---|
| 1 | .1 | 50 |
| 2 | .1 | 500 |
| 3 | .3 | 50 |
| 4 | .3 | 500 |

**Evaluation** After imputation, the MLE estimates of the means, variances and covariances of all variables with missing values where estimated and pooled across multiply imputed datasets.

The parameters of interest reference values were obtained by averaging the MLE obtained on the fully observed datasets. The bias of each estimate was computed as the difference between Monte Carlo averaged statistic and the reference values. For now, the size of the bias is shown per statistic as percentage of the target value.

A first run of experiment 1 took approximately 20h.

## 2.2 Experiment 2: Interactions and the like

Data for this experiment were generated in two steps. First, a matrix of predictors was generated as in experiment 1. Then, a dependent variable y was

generated using one predictor from each block. Depending on the condition, an interaction term was either included or not.

Missing data was imposed using a probit model to facilitate manipulation of the proportion of missing cases. In all conditions, seven variables where targeted by missingness: y, three in block 1, and three in block 2. Four variables where randomly selected from X to use as predictors in the probit model for each target variable. The same set of coefficients where used $(-1, -1, .666, -.333)$.

## 2.3 Experiment 3: Latent data

Bla bla this makes a lot of sense