Latent Factor Approach to Missing not at Random

Naijia Liu
naijialiu.github.io; naijial@princeton.edu
Department of Politics, Princeton University

Overview

- Missing not at random
$$P(M|X,\phi) = P(M|X_{\text{mis}}, X_{\text{obs}}, \phi)$$
- Model on latent structure of missing pattern.
$$M_k \perp\!\!\!\perp X_k | X_{-k}, Z$$
- Sensitive question: self-reported ideology in China.
- Allow for a mixture of missing mechanisms in the dataset.
- Code available upon request.

Existing work

- Assume missing at random
  - Multiple Imputation by Chained Equation (White et al., 2011)
  - Amelia (Honaker and King, 2010)
  - Doubly robust estimator (Bang and Robins, 2005)

Assumptions

Latent factor captures confounding

Let Z be the latent factor behind missing matrix M: $Z \sim P(\cdot|M)$. For each variable k, we assume:

$$M_k \perp\!\!\!\perp X_k | X_{-k}, Z$$

Imputation to recover:

$$\mu = E(X) = E(E(X|\theta)) = \int x P(x|\theta, z) P(\theta, z) d\theta dz$$

By maximizing observed data likelihood as:

$$\left(\prod_i^n P(X_k|X_{-k}, Z, \theta_1)^{1\{M_k=0\}}\right)\left(\prod_i P(X_{-k}, Z, \theta_2)\right) f(\theta_3, Z)$$

where θ_1 determines $P(X_k|X_{-k})$, θ_2 determines $P(X_{-k}, Z)$ and $f(\theta_3, Z)$ is a function of latent factor itself.

Method

Step 1 Convert the dataset into a binary matrix M:
$M_{ik} = 1$ indicates observation i is missing kth variable and 0 otherwise.

Step 2 Conduct latent factor model on the binary matrix, to obtain the estimation of Z.

Step 3 Calculate pairwise distance, for observation i and j, d_{ij} using a kernel by both Z and observed data for each observation. Optimal bandwidth can be chosen via cross-validation (click to see more results).

$$d_{ij} = \mathbf{K}(\{Z_i, X_{ik}\}; \{Z_j, X_{jk'}\}),$$
$$\forall k, k' \text{ such that } M_{ik} = M_{jk'} = 0$$

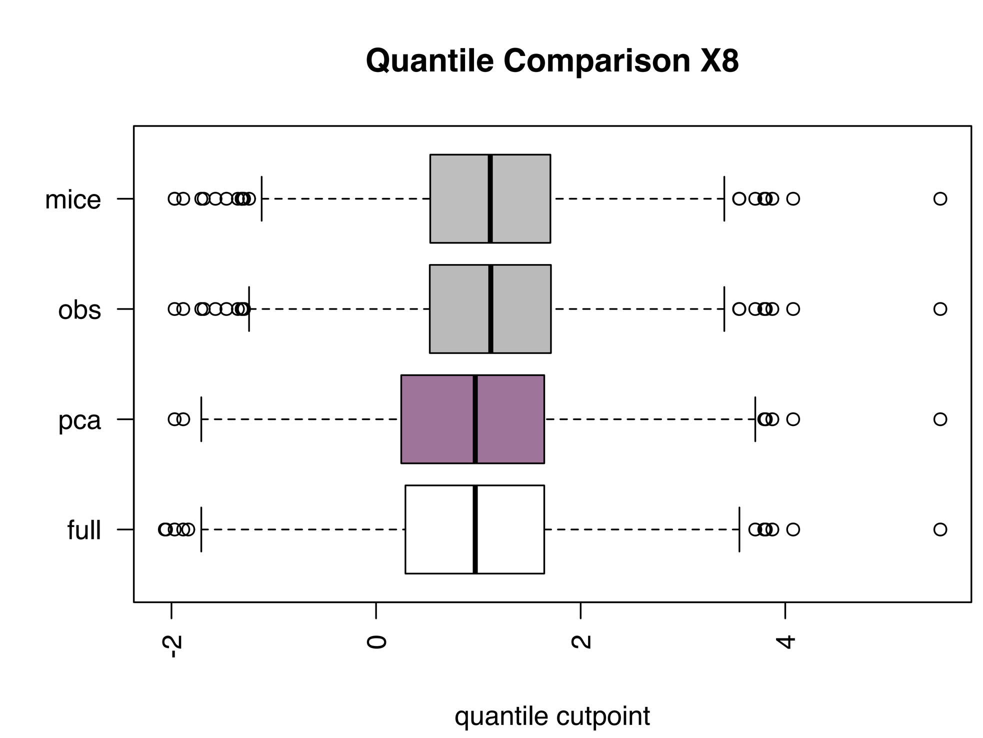
Step 4 Imputation using the kernel distance. If $M_{ik} = 1$, we impute the entry as follow:

$$x_{ik} = \sum_{j=1}^{N} w_{ij} x_{jk}, \quad \forall M_{jk} = 0 \text{ and } w_{ij} = \frac{d_{ij}}{\sum_{j=1}^{N} d_{ij}}$$

Simulation

- 2000 observations with 18 variables.
- Multivariate normal with mild covariances.
- Missing not at random by unobserved confounders, with around 30% missing.
- Comparison among listwise deletion (gray-obs), multiple imputation using mice (gray-mice) and proposed method (purple).
- Proposed method recovers better the distribution of full data (white bar).

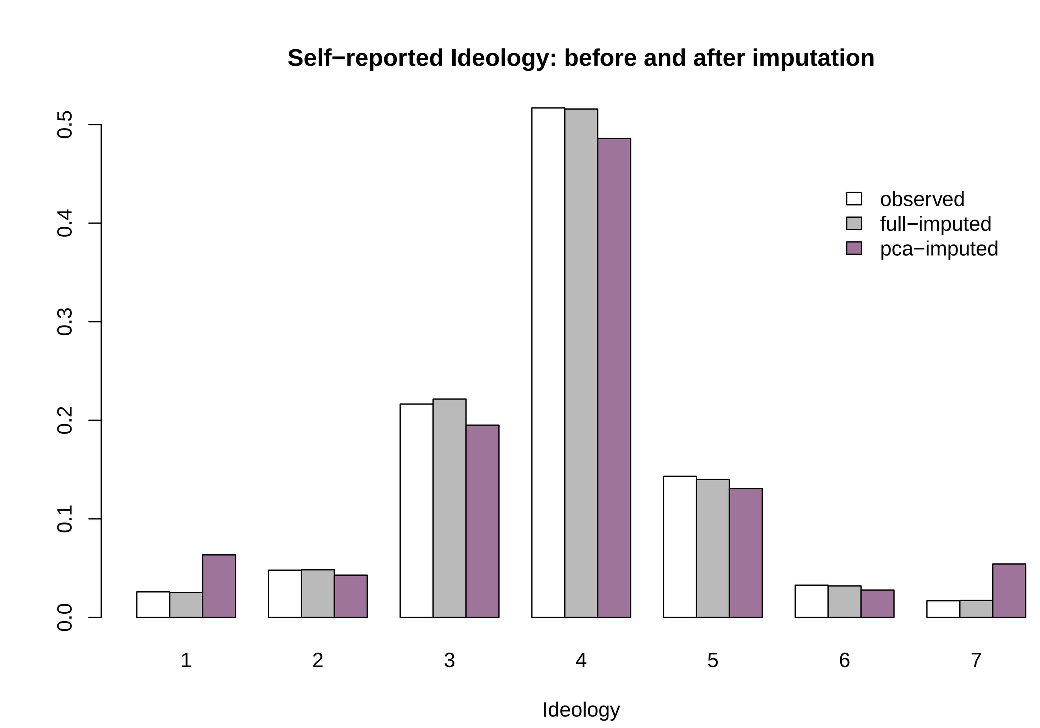
Figure 1:



Application

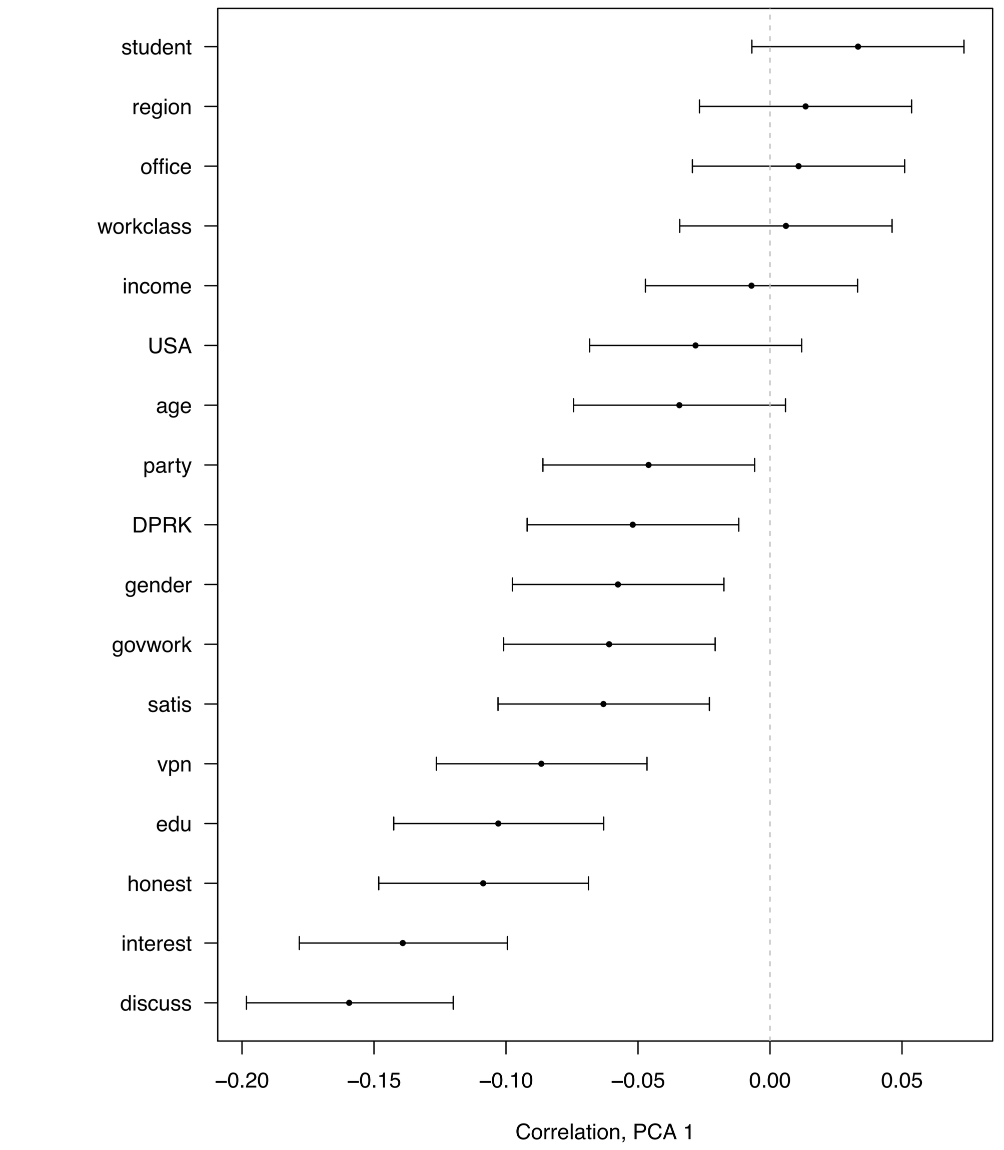
- 2017 Chinese Netizen Survey.
- Sensitive questions with a refusal option.
  - **Self-report ideology** (605 missing values).
- 2379 observations in total, 1314 complete observations.
- Comparison among listwise deletion (white), mice (gray) and proposed method (purple): Middle vs extreme.
- Pairwise correlation check between covariates and first components.
- Plot of std dev for each component.

Figure 2:Imputation result comparison



Application

Figure 3:Correlation with first component



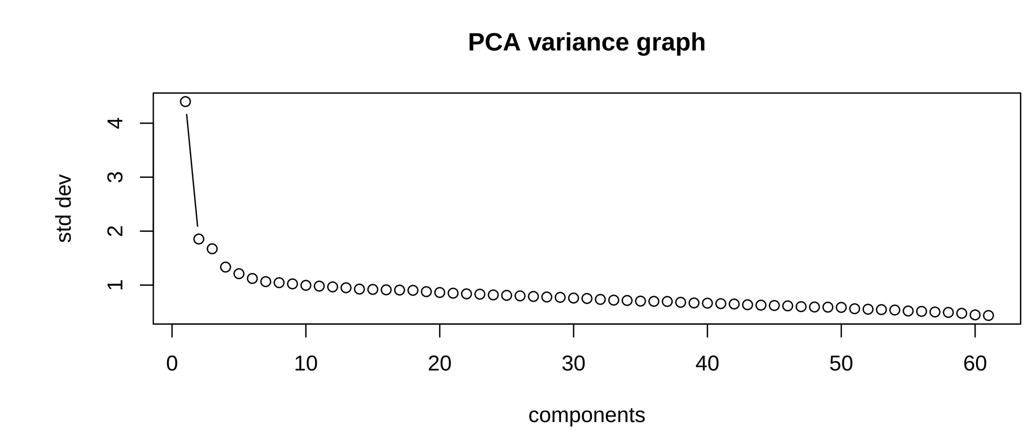


Figure 4:Variance of components

Conclusion and discussion

- Deals with MNAR with unobserved confounders, broad applications such as sensitive questions, censoring and etc.
- More simulation (click to see more) results to show superior performance relative to naive regression on binary missing indicators.
- Performs better in categorical variables than continuous variables for small samples.