A Latent Factor Approach to Missing Not at Random *

Naijia Liu
naijial@princeton.edu

June, 2021

Abstract

Missing data is prevalent among social science datasets. Existing multiple imputation methods assumes missing at random (MAR), which is a more restrictive assumption than MNAR. The problem becomes more challenging under missing not at random (MNAR) scenario, such as missingness in sensitive survey questions. This paper confronts MNAR by modeling the latent structure of the missingness to mitigate the influence of the unmeasured confounders that cause the missing values. This approach allows one to assume missing at random (MAR) conditional on the latent factor. The proposed method outperforms multiple imputation methods under MNAR. In addition to simulation comparison, I show an application using latent factor model to impute the missing values in a self-reported ideology question, which is considered to be a sensitive question in the 2017 Chinese Netizen Survey dataset. I conclude the paper with discussions of the scope of the method and potential extensions.

*R-package to implement the proposed method is available upon request.

1 Introduction

Missing data problem appears in many settings of social science studies, especially in observational datasets. One of the challenges in dealing with missing data is that missingness due to unobserved confounders. In other words, the missing values are systematically different from the observed ones and researchers do not observe enough information to impute. This phenomena is common in observational datasets, such as non-response bias in sensitive survey questions. Refusals under this setting may be due to some unobserved confounders and cannot be retrieved via observed covariates. One common example is social desirability and its impact on people's willingness to answer. In many cases, such missing values are nonignorable, preventing researchers to conduct analysis on the data.This paper is to propose a method to deal with the MNAR problem to enable scholars to proceed with data analysis, without imposing excessive substantive / theoretical structures on the story behind missing values.

Conventional multiple imputation methods assume missing at random (MAR) (King et al., 2001; White, Royston and Wood, 2011), where one assumes no systematic differences between observed and missing values. The MAR assumption may not fully capture the missing mechanism by ignoring the potential censoring or confounding powers behind missingness [1].

Little and Rubin (2014) defines missing not at random (MNAR) as

$$P(M \mid X, \phi) \neq P(M \mid X_{\text{obs}}, \phi), \quad \forall X_{\text{mis}}, X_{\text{obs}}, \phi \tag{1}$$

for missing indicator M, variable X and some parameter ϕ. Again, this definition shows that conditional on observed values X_{obs}, one cannot recover the full picture of the dataset.

To solve MNAR, one will need to obtain more information with regard to the distribution of unobserved confounders for missingness in the dataset. The missing pattern for all covariates among observations contains useful information to potentially capture the confounders behind it. To obtain such information, this paper proposes the factorization of the missingness distribution. Scholars try to uncover more information by matrix operations such as denoising and modeling the low-rank structure of the matrix (Kallus, Mao and Udell, 2018; Sportisse, Boyer and Josse, 2018). This project proposes a solution to the missing not at random (MNAR) problem by utilizing the latent structure behind missing data. More specifically, I factor the latent structure of binary missing patterns using latent factor model to improve imputation performance.

I start the paper by introducing the method in section 2. Then I provide simulation results and an application using sensitive survey questions. I conclude the paper by discussing the limitation and scope of the method. See appendix for more examples using different kinds of latent factor

[1] Notice that, listwise deletion assumes missing completely at random, which is an even stronger assumption to make and often unrealistic (King et al., 1998).

models.

1.1 Related Literature

Missing not at random problem has been studied by researchers across different fields. In social sciences, scholars worry about nonignorable missingness to sensitive survey questions (Miller, Saunders and Farhart, 2016; Barabas et al., 2014; Luskin and Bullock, 2011; Mondak and Davis, 2001; Pietryka and MacIntosh, 2013; Gibson and Caldeira, 2009). As a result, scholars propose list experiment to eliminate the social desirability issue respondents would otherwise face (Blair and Imai, 2012; Glynn, 2013). List experiment is a good example of using design to tackle the potential nonignorable missingness. However, a list experiment can handle only limited amount of questions and respondents sometimes have trouble fully understanding the rules behind it. It is complicated and expensive to implement for large surveys with many sensitive questions embedded.

In medical and bio-statistic research, scholars also suffer from missing not at random problem, such as truncation by death. In fact, the doubly robust estimator for missing at random (Bang and Robins, 2005) was originated from such literature (assuming MAR). Acknowledging the limitation of MAR assumption, scholars propose new methods to deal with missing not at random in many specific settings (Yang, Wang and Ding, 2019). In most of the cases, a clinical trial is a well-designed randomized experiment. Thus, similar to list experiment, scholars have more structures to utilize from. Moreover, medical researchers often possess prior knowledge towards certain types of drugs / patients and thus can utilize them in the imputation process. One common practice is to assume prior distribution of a certain variable to aid the approximation and marginalization process.

Finally, statisticians have proposed new methods to deal with missing not at random in a data rich environment, where researchers have data with relatively high dimensionality (Kallus, Mao and Udell, 2018; Sportisse, Boyer and Josse, 2018). By taking advantage of the high dimensional matrix, one will be able to de-noise and recover information out of the missing structures. This project is more in line with this strand of literature. One of the advantages is that there is no additional design cost to the study. Unlike list experiment or a clinical trial, scholars only need a large survey dataset to begin with. And in most of the proposed methods, no follow-ups are required.

2 Proposed method

2.1 Latent factor

Let M denote the binary missing pattern matrix where $M_{ik} = 1$ if observation i is missing k-th variable and zero otherwise, Z as the latent factor estimated from missing matrix M. In other words, Z is the result of dimension reduction of the large matrix M.

$$Z \sim P(\cdot \mid M)$$

Note that M shares the same dimensionality with the original dataset.

Let X denote the complete set of covariates if we were to observe everything. More specifically, let X_{ik} denotes the value for observation i and variable k.

2.2 General Setup

The goal of the proposed method is to tackle the missing not at random problem (MNAR), where

$$P(M_{ik} = 1 \mid X_{-i,k}) \neq P(M_{ik} = 1 \mid X_{-i,k}, X_{i,k}), \quad \forall i, k \tag{2}$$

where $-i$ denotes other observations.

This is to say that with only the observed observations in the dataset, researchers do not have enough information to impute missing entries. And the missing probability will differ if not conditioning on the missing value itself.

Meanwhile, the proposed method allows missing at random (MAR) and missing completely at random (MCAR) to be present in the dataset. If MAR or MCAR are true, the latent factor Z derived from missing indicator M will not provide us with any additional information but pure noise.

Below, I introduce the assumptions that are necessary for the latent factor approach to missing not at random.

Assumption 2.1 (Ignorability conditional on the latent factor)**.**

$$M_{ik} \perp\!\!\!\perp X_{ik} \mid X_{-i,k}, Z_i, \quad \forall i, k \tag{3}$$

Where X_k denotes variable k and M_k denote the missing indicators for the same variable.

In other words, assumption 2.1 indicates that conditioning on the observed values and latent factor, one should be able to characterize the missing distribution. And by conditioning the observed values and latent factor, one will get the following equality:

$$P(M_{ik} = 1 \mid X_{-i,k}, Z_i) = P(M_{ik} = 1 \mid X_{ik}, X_{-i,k}, Z_i)$$

The researcher are allowed to choose any appropriate latent factor model to get Z for the dataset. In section 4, I will show an application using principal component analysis (Wold, Esbensen and Geladi, 1987). In appendix, I will repeat the same application but using latent utility model (Clinton, Jackman and Rivers, 2004).

2.3 Latent factor approach

Below I introduce latent factor approach to missing not at random. I first factorize the distribution of M matrix by latent factor model. Then I conduct imputation via a kernel method, by assigning differential weights to each observed values. The weights are calculated using the information obtained from latent factor model. The intuition behind the kernel distance is that we want to identify similar observations in the dataset, both by their observed values and missing patterns.

Step 1 Convert the dataset into a binary matrix M, where $M_{ik} = 1$ 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.

Here we think of Z as a lower-dimensional representation of M. More discussions on the choice of latent factor model is provided later in the paper.

Step 3 Calculate pairwise distance, for observation i and j, d_{ij} using a kernel by both Z and observed data for each observation.

$$d_{ij} = \mathbf{K}(\{Z_i, X_i\}; \{Z_j, X_j\})$$

More discussions on the choice of kernel is provided later in the paper.

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

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

2.4 Likelihood of complete data

With assumption 2.1 to be true, the steps above involving observed data and latent factor should be able to produce consistent imputation results. I provide some intuition below, taking sample mean as an example.

Say one is interested in estimating parameter:

$$\begin{aligned}\mu = E(X) &= E(E(X|\theta)) \\ &= \int xP(x \mid \theta, z)P(\theta, z)d\theta dz\end{aligned}$$

where θ is a parameter for the data and z is the parameter for missingness. In other words, θ decides the distribution of data while z decides missing probability. To obtain an estimation of μ term, one would need to estimate the integral with the presence of missing data. As a result, let's further partition the probability into observed and missing ones.

Again, when $M = 1$ we do not observe the variable and $M = 0$ otherwise. We can write down the observed data density as the joint probability of missingness and observed values for each observation:

$$\begin{aligned}P(1\{M = 0\}X, \theta, z) &= P(X, \theta, z, M = 0)^{1\{M=0\}}P(M = 1, \theta, z)^{\mathbf{1}\{M=1\}} \\ &= \{P(X \mid \theta, M = 0, z)P(M = 0 \mid \theta, z)P(\theta, z)\}^{\mathbf{1}\{M=0\}}\{P(M = 1 \mid \theta, z)P(\theta, z)\}^{\mathbf{1}\{M=1\}}\end{aligned}$$

By assumption 2.1 we have conditional independence between M and X: $P(X|\theta, z) = P(X|\theta, M = 0, z)$

So the likelihood can be further written as:

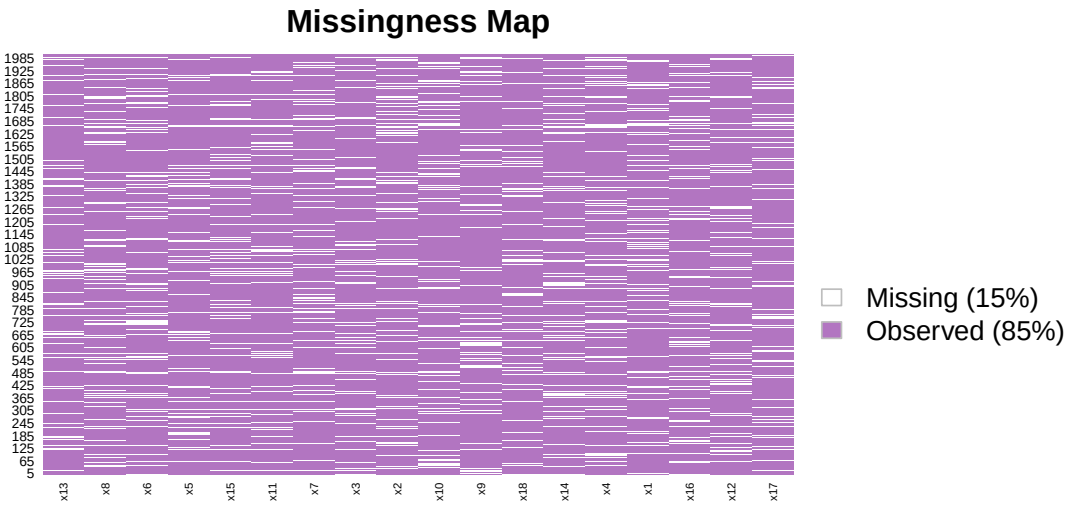
$$\begin{aligned}&\{P(X \mid \theta, z)P(M = 0 \mid \theta, z)P(\theta, z)\}^{\mathbf{1}\{M=0\}}\{P(M = 1 \mid \theta, z)P(\theta, z)\}^{\mathbf{1}\{M=1\}} \\ &= P(X \mid \theta, z)^{\mathbf{1}\{M=0\}}P(M \mid \theta, z)P(\theta, z)\end{aligned}$$

Thus, the likelihood of the full model can be partitioned as

$$\prod_{i=1}^{n}\prod_{j=1}^{J} P(X_{ij} \mid \theta, z)^{\mathbf{1}\{M_{ij}=0\}}P(M_{ij} \mid \theta, z)P(\theta, z) \tag{4}$$

As a result, we can maximize full data likelihood by maximizing equation (4), using the observed values and the missing matrix. This proof is only to show the intuition of how we solve MNAR using the observed data and missing matrix. Scholars can decide the type of model to analyze the dataset on, and the proposed method is not constrained to simple parameters such as arithmetic mean.

Figure 1: Missingness heatmap



3 Simulation

Missing not at random with unobserved confounders

The data generating process follows a joint normal distribution with mild correlation among each other, with 2000 observations and 18 variables. The Gaussian kernel has a sigma of around 0.1, which is selected via cross validation (Ripley et al., 2013). Out of the 18 variables, most of them contain missingness at around 10% level or higher. I offer more discussion on the quality of data in the later part of the paper. Here, missingness is generated by a group of **unobserved** confounders that I generated separately. The confounders are correlated with multiple variables out of the 18. More importantly, they also determine missing probability for all the variables in the data. To provide a visualization of both the complete dataset and the missing ones, I present a heatmap below. In figure 1, we see a 2000 by 18 matrix, where purple indicates observed entries and white indicates missing entries.

The data generating process of the 18 variables are shown below. The mean of the 18 variables are randomly generated by a Poisson process. Variance covariance matrix is generated so that 18 variables share mild covariance, which will aid the multiple imputation method. A dataset with 18 variable is relatively large in size. With around 15% missing in the dataset, we will have

enough numbers of combination in missing patterns to work with.

$$X \sim \text{mvnorm}(\text{pois}(\lambda = 4), \sigma^2)$$

$$\sigma^2 \sim \begin{pmatrix} 2 & 1 & 1 & ... & 1 \\ 1 & 2 & 1 & ... & 1 \\ 1 & 1 & 2 & ... & 1 \\ & & ... & & \\ 1 & 1 & 1 & ... & 2 \end{pmatrix}$$

I then generated a group of unobserved confounders C as follow:

$$C \sim \begin{pmatrix} X & + & e_1 \\ X^2 & + & e_2 \\ X^2_{\text{odd}} & + & e_3 \\ X^2_{\text{even}} & + & e_4 \end{pmatrix}$$

$$e \sim \text{N}(0, 0.5)$$

As shown above, all confounders are generated as a linear or nonlinear function of the variables in the dataset with an uncertainty term. X_{odd} indicates all the odd number variables such as $X1, X3, X5$, whereas X_{even} indicates all the even number variables such as $X2, X4, X6$. Missingness in the dataset is then generated as a function of both the variable value and confounder value. For simplicity purpose, I make variables missing if the sum of confounders and the variable value is below 15% quantile, with an added uncertainty term.

From here on, the confounders C is then excluded from all imputation process. This is to simulate a missing not at random situation, where researchers do not observe the confounders to use for imputation. It is problematic to use the dataset without dealing with missing values, since there maybe systematic bias induced by missing values. The goal of imputation is to recover the complete data distribution to the best extent. Figure 2 also shows comparison between MICE and proposed method. The Y-axis is the average difference in L2 norm between the imputed CDF and true CDF generated.

$$\sum_x \sqrt{(F_x - F'_x)^2} \tag{5}$$

Total variation measures at the worst point, how far is the imputed CDF to true CDF, while L2 norm measures on average how far the two are. A lower total variation and L2 norm indicates better performances. As it's shown in both figures, the proposed method outperforms multiple imputation method. The imputed variables have CDFs that are closer to the truth, by the proposed method.

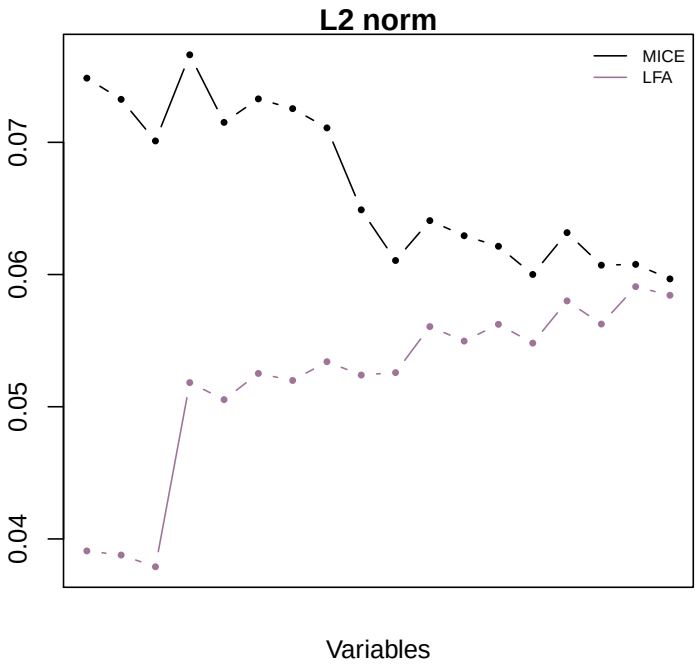


Figure 2: Comparison among methods on L2 norm over all variables

Naive missing indicator approach

The latent factor approach may remind readers of a more straightforward way to model missing patterns: adding missing indicator to each variable in the regression. This section presents simulation results of linear regression, by interacting independent variables with missing indicators. Here I used the same data generating process as in previous section. Again, the data generating process follow a joint normal distribution with mild correlation among each other, with 2000 observations and 18 variables. However, to better illustrate the indicator approach, I will deploy a different missing mechanism. Also, for the purpose of illustration, only part of the variables contain missing.

Here, I made only two variables to contain missingness. X_2 is missing as a function of both X_1 and X_2. X_{12} is missing as a function of both X_4 and X_{12}. Missing indicator for both X_2 and X_{12} are included in the regression. Missing indicator approach will add binary missing indicators for variables with missingness as additional regressors. In some scenario, one also adds interaction terms between these indicators and existing variables in the dataset. We will compare results between this approach and the oracle values simulated.

Table 1 shows the comparison results. For visualization purposes, I only show coefficients for related variables, while the regressions were run with full list of variables. The first column shows the oracle results using whole dataset. The second column shows the regression with binary

missing indicators as additional variables.

$$Y = \beta \cdot X + \beta_{m2} \cdot \mathbf{1}\{X_2 \text{ missing}\} + \beta_{m12} \cdot \mathbf{1}\{X_{12} \text{ missing}\}$$

The third column shows the regression with binary missing indicators interacted with **correct** choice of variables. That is:

$$\begin{aligned} Y = \beta \cdot X + \beta_{m2} \cdot \mathbf{1}\{X_2 \text{ missing}\} + \beta_{m12} \cdot \mathbf{1}\{X_{12} \text{ missing}\} \\ + \beta_{m2'} X_1 \cdot \mathbf{1}\{X_2 \text{ missing}\} + \beta_{m12'} X_4 \cdot \mathbf{1}\{X_{12} \text{ missing}\} \end{aligned}$$

The forth column shows a randomly chosen mis-specified model. That is, in addition to the correct setting, wrongfully specified interactions between $X_3 \cdot \mathbf{1}\{X_2 \text{ missing}\}$ and $X_{14} \cdot \mathbf{1}\{X_{12} \text{ missing}\}$ were added.

Table 1: Regressions with missing indicator

	Full	Indicator naive	Correct	Mispecified
x2	−1.012	−1.045	−1.041	−1.045
	(0.011)	(0.024)	(0.024)	(0.024)
x4	0.540	−0.565	0.392	−0.566
	(0.055)	(0.115)	(0.163)	(0.115)
x12	0.985	0.887	0.960	0.886
	(0.016)	(0.035)	(0.035)	(0.035)
ind_x2		1.257	1.261	1.084
		(0.032)	(0.032)	(0.645)
ind_x12		−0.913	4.660	−0.914
		(0.046)	(0.683)	(0.046)
x1:ind_x2			−0.001	0.0003
			(0.006)	(0.006)
x4:ind_x12			−1.854	
			(0.227)	
x3:ind_x2				−0.043
				(0.161)
x14:ind_x12				0.048
				(0.046)
Constant	0.946	4.570	1.281	4.670
	(0.437)	(0.697)	(0.795)	(0.775)
Observations	2,000	2,000	2,000	2,000
Adjusted R^2	1.000	0.999	0.999	0.999

The immediate next step is to further check the performance, with all possible interactions.

That is to interact missing indicators with all the variables in the regression. However, one might be concerned about overfitting due to the number of parameters. Table 2 shows the result after penalized linear regression (Tibshirani, 1996). Only the variables with a non-zero coefficient are reported.

Table 2: Glmnet after cross validation

	Lambda=0.09900804	OLS
Constant	−0.070	0.946
x1	5.181	5.199
x2	−0.839	-1.012
x3	0.459	1.023
x5	0.875	1.017
x6	0.702	0.97
x7	0.173	0.606
x8	−1.891	-1.998
x9	−0.881	-1.013
x10	−0.694	-1.188
x11	0.831	1.009
x12	0.630	0.985
x13	−0.893	-1.023
x14	0.368	0.507
x15	−0.874	-1.006
x16	0.826	0.992
x17	−0.870	-0.992
x18	0.845	0.986
x5*ind_x2	0.052	
x6*ind_x2	0.113	
x7*ind_x2	0.116	
x12*ind_x2	0.223	
x16*ind_x2	0.114	
x2*ind_x12	-0.136	
x4*ind_x12	-0.129	
x8*ind_x12	−0.002	

Furthermore, table 3 shows the OLS regression with only selected variables from previous step.

In summary, naively adding missing indicators to a regression will not solve MNAR problem. Even after lasso selection of interaction terms, the same problem remains. Furthermore, lasso may select out important variables that researchers would like to study, simply because the interaction term is not significant.

Table 3

	Estimate	Std. Error	truth	zscore
(Intercept)	4.044	0.979	0	4.131
x1	5.193	0.006	5	29.827
x2	−1.060	0.034	−1	−1.762
x3	0.669	0.115	1	−2.889
x4	−0.366	0.163	0.500	−5.323
x5			1	
x6	0.963	0.084	1	−0.443
x7	0.394	0.119	0.600	−1.729
x8	−1.996	0.034	−2	0.127
x9	−1.026	0.036	−1	−0.715
x10	−1.103	0.114	−1.200	0.850
x11	0.989	0.033	1	−0.323
x12	0.854	0.049	1	−2.962
x13	−1.038	0.046	−1	−0.839
x14	0.512	0.032	0.500	0.368
x15	−0.993	0.033	−1	0.226
x16	0.995	0.032	1	−0.153
x17	−1.000	0.046	−1	−0.003
x18	0.954	0.033	1	−1.415

4 Applications

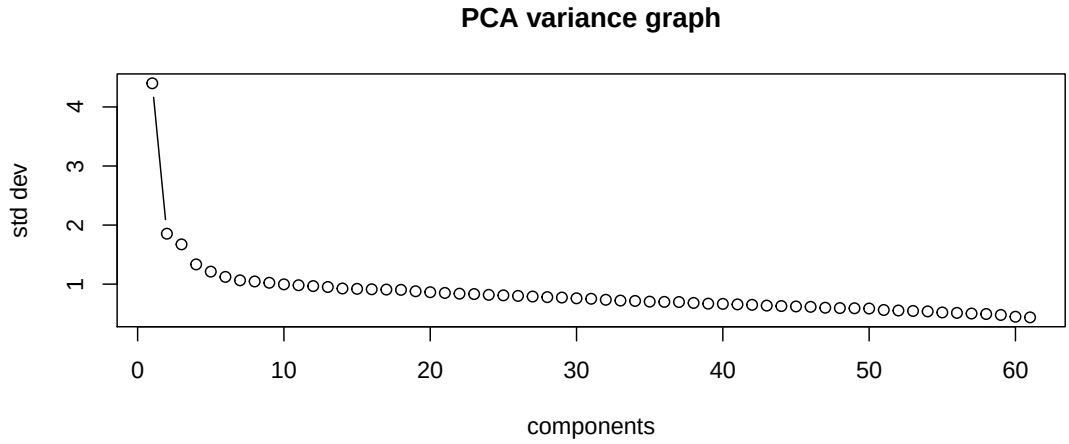
Here I show an example under the kernel method. As stated before, one can select latent factor models based on data structure and substantive knowledge. Appendix section B presents an example under latent utility model and the clustering method, using same survey dataset in this section. I offer more discussion to the advantages and limitations of both approaches later in the paper.

4.1 Sensitive survey datasets

I use the 2017 Chinese Netizen Survey (Ma, 2017), in which the proposed method helps to impute missing values in sensitive questions. The survey contains 61 sensitive questions such as "what is your opinion towards Chairman Mao?" There are 2379 observations and 1314 of them are complete. MNAR is a reasonable suspicion towards this dataset due to the wordings of these questions and furthermore, most of its respondents reside inside mainland China.

I conducted the proposed method step by step. I first conduct a PCA analysis on the binary missing matrix of the dataset. Then, I calculate a pairwise distance using a Gaussian kernel. The kernel was fitted using PCA result and observed values of the dataset. I inverse and standardize

Figure 3



the distance into weights, by which I impute all the missing entries. I cross-validated to select the best bandwidth for the Gaussian kernel.

Figure 3 shows the relative variance for the all 61 components. As we can see, most of the variance concentrates on the first several components.

Figure 4 shows the pairwise correlation (with 95% confidence interval) between the covariates and the first component of PCA model. A significant correlation indicates that the variable covaries with this component. Due to the nature of PCA method and matrix rotation procedure, it is hard to interpret the sign of the correlation. Figure 4 shows that higher political salience (such as higher frequency of political discussion, higher political interests, more frequent usage of vpn, higher level of satisfaction towards society and being a government worker) significantly correlates to refusal rate. Also, gender is a significant predictor.

4.2 Imputation results

Figure 5 presents the imputation result after a Guassian kernel. The variance term σ is selected by cross validation. The white bars plot the observed data (with missing values). The gray bar shows a naive multiple imputation approach, where we assume MAR and use the whole dataset to impute missing values. The purple bars shows the imputation results by proposed method. The proposed method imputes more missing answers to be people with a more extreme ideology.

Figure 4

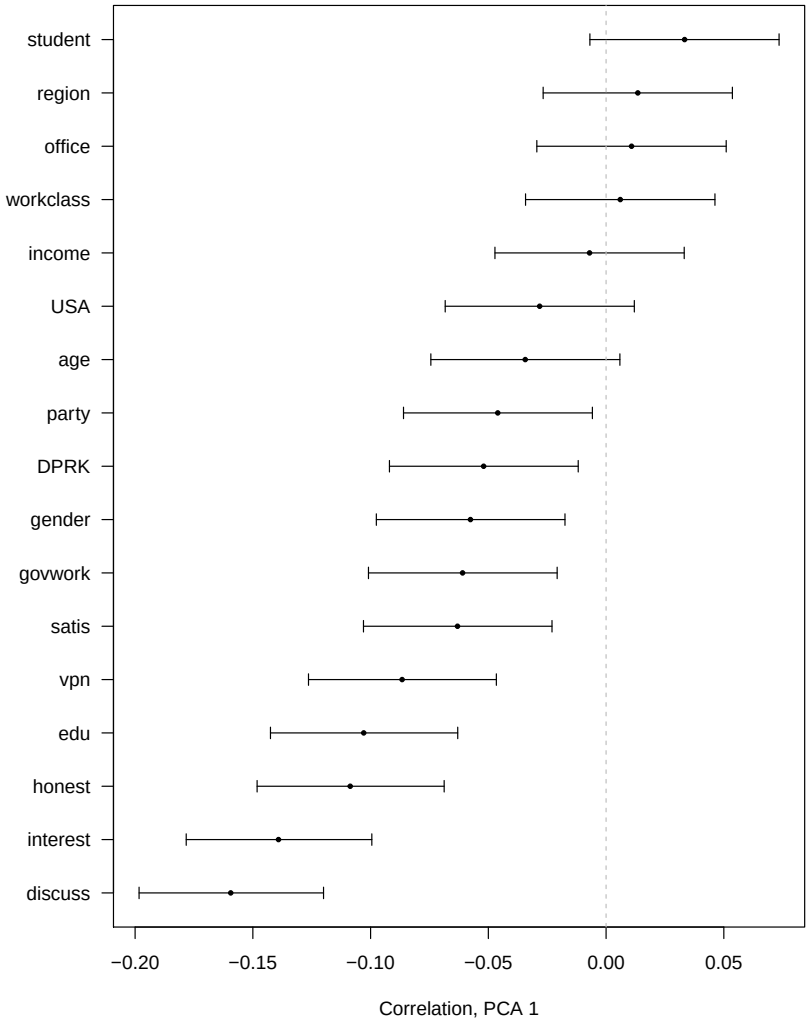
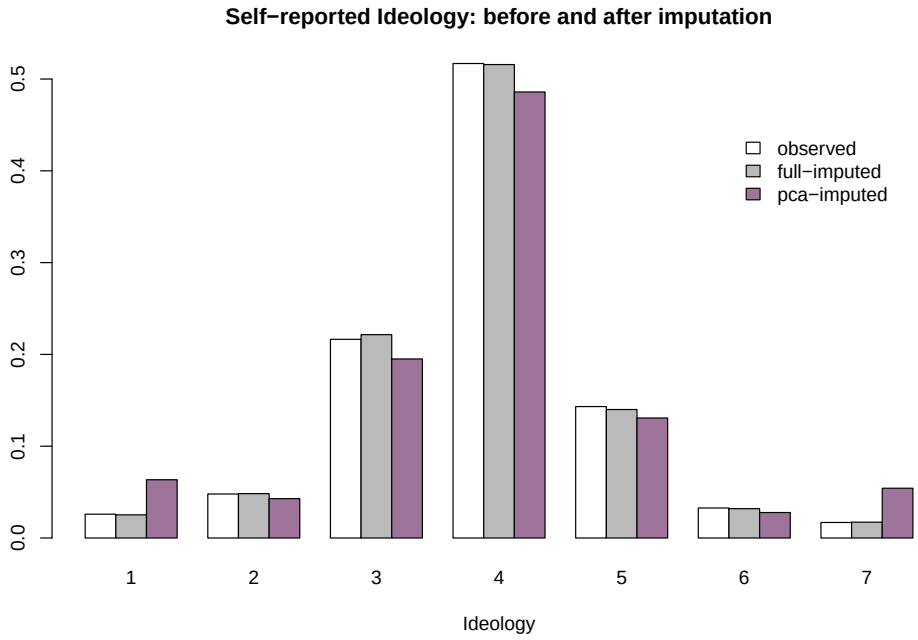

student
region
office
workclass
income
USA
age
party
DPRK
gender
govwork
satis
vpn
edu
honest
interest
discuss
−0.20
−0.15
−0.10
−0.05
0.00
0.05
Correlation, PCA 1

Figure 5



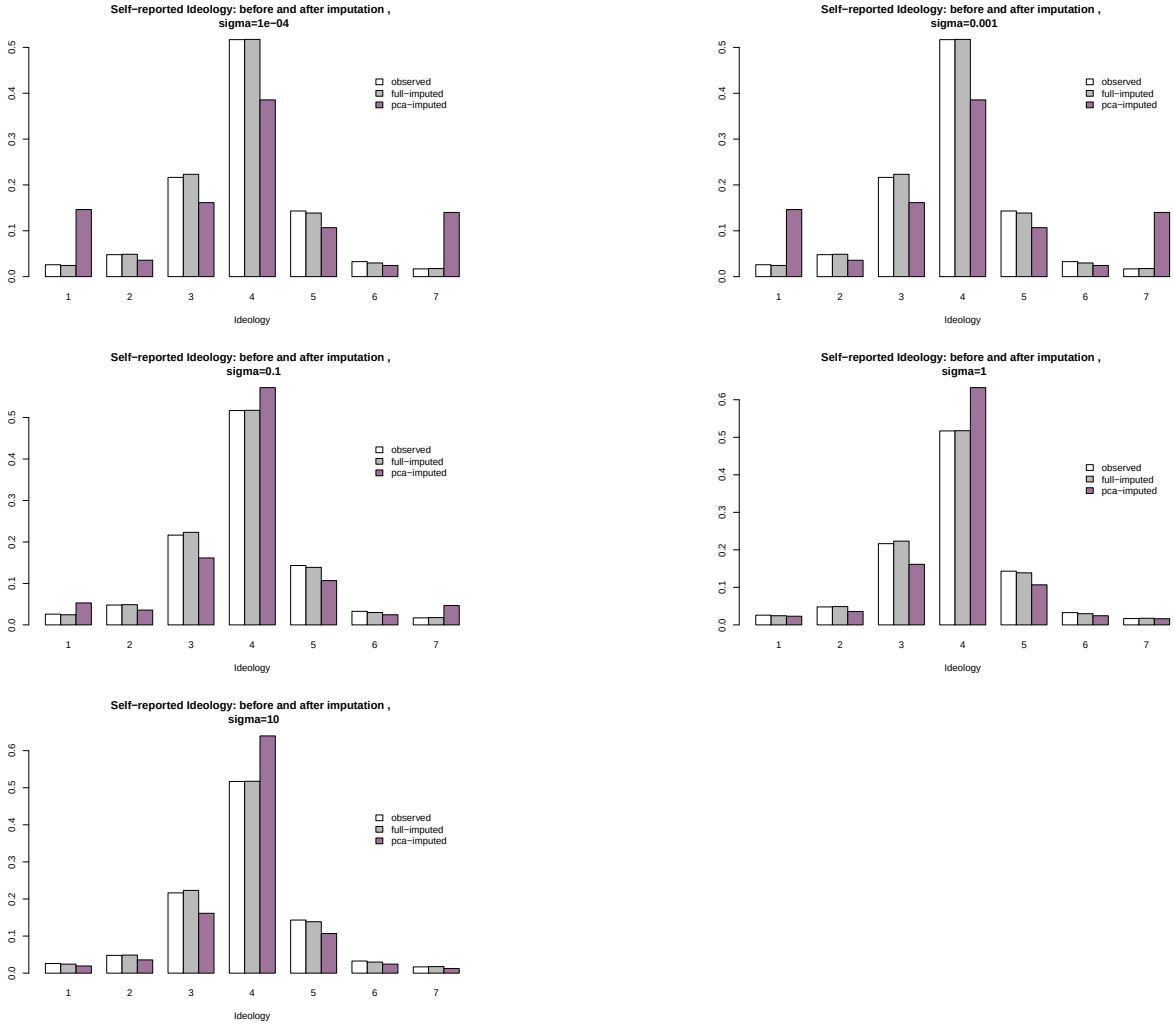
4.3 Selection of bandwidth

Figure 6 shows imputation result of the same variable with different σ under Gaussian kernel. The term governs how precise a Gaussian kernel will be. A higher σ means a more flat Gaussian curve, by which we assign more equal weight to everyone. A lower σ means a tighter curve and we discriminate more against observations "farther" in the dataset.

Figure 6 provides us with two take away points. First of all, kernel imputation is sensitive to model selection. When we allow a σ too high, we basically ignore the minority groups in the dataset by assigning almost zero weights to them. In this case, the neutral ideology category is the dominant group and will hence have the highest weight for imputation. When we allow a σ term too low, we may not take the best use of the data. The most extreme case is to find the nearest person and assign a weight of 1 on her and zero for everyone else. A principled way of model selection such as cross validation is strongly recommended for this reason.

Secondly, among all the specification of σ, $\sigma = 1$ best replicate the multiple imputation result. The MAR assumptions for multiple imputation is different from the proposed method. It is possible that a σ is selected so that we simulate a MAR situation and thus produce similar results to multiple imputation. Again, the selection of σ should be statistically justified.

Figure 6


Self−reported Ideology: before and after imputation ,
sigma=1e−04
observed
full−imputed
pca−imputed
Ideology
Self−reported Ideology: before and after imputation ,
sigma=0.001
observed
full−imputed
pca−imputed
Ideology
Self−reported Ideology: before and after imputation ,
sigma=0.1
observed
full−imputed
pca−imputed
Ideology
Self−reported Ideology: before and after imputation ,
sigma=1
observed
full−imputed
pca−imputed
Ideology
Self−reported Ideology: before and after imputation ,
sigma=10
observed
full−imputed
pca−imputed
Ideology

5 Discussion

It is not the most straightforward way to model missing patterns with latent factor model. Instead, one may suggest adding missing indicators in the data as additional regressors. There is one major limitation to the missing indicator method. First of all, if number of variables with missing data is relatively high, one might suffer from loss of degree of freedom in a regression setting. Since researchers will not know the true missing mechanism, a safe way would be interact all the missing indicators with all variables in the dataset. Penalized regression may be able to select meaningful variables in this situation. However, we cannot guarantee that variables of interests to be selected. Furthermore, regression models such as Lasso and Ridge produces biased point estimate. In appendix 3, I show simulation results with this alternative method. And as a matter of fact, some major variables in the original dataset were dropped after cross validation.

5.1 Limitation of the method

Wang and Blei (2019) proposed "deconfounder" method to deal with assignment of multiple treatments. They propose a hierarchical approach, in which they first detect the confounding variables among multiple treatment assignments via a latent factor model, then calculate treatment effect with the help of latent factor step. This approach shares great similarity with the proposed method. Both methods are interested in extracting the latent structure to control for unobserved confounders. Deconfounder received several criticism and I discuss their implications on the proposed method in this section.

D'Amour (2019) argues that for certain distributions, one may not have a unique factorization of the treatment assignment. It is problematic for the purposes of a unique treatment effect estimation. Suppose that both V and U factorizes the distribution of T.

$$\int_v P(Y \mid T = t, V = v)P(V = v)dv = \int_u P(Y \mid T = t, U = u)P(U = u)du$$

The proposed method also suffers from this critique. Different factorization of missing pattern may lead to different cluster assignments in the imputation step, and thus a different imputed dataset. One way to mitigate the problem is to try different specifications of latent factor model and compare among imputation results. If the missing pattern is well concentrated in finite dimensions and with a sample large enough, one should get reasonably similar imputation results.

Furthermore, Ogburn, Shpitser and Tchetgen (2020) provides counter-examples that the factorization of treatment assignment may not lead to ignorability. In other words, conditioning on the latent factor may give us independence between each treatment assignment and the potential outcome. But this does not guarantee that all treatment assignments are independent of the

outcome variable.

$$T_1 \perp\!\!\!\perp Y(t_1, t_2) \mid U, \quad T_2 \perp\!\!\!\perp Y(t_1, t_2) \mid U, \quad T_1 \perp\!\!\!\perp T_2 \mid U$$

This **does not** guarantee:

$$T \perp\!\!\!\perp Y(t_1, t_2) \mid U \tag{6}$$

The proposed method focuses on data imputation, instead of estimation of a certain treatment effect. As a result, ignorability assumption is not required for the targeted imputation procedure. As for the estimation of modified propensity score, ignorability assumption requires that $T \perp\!\!\!\perp Y \mid X_{\text{obs}}, Z$ and this is different from the critique above in equation (6).

Furthermore, positivity assumption might be violated (Imai and Jiang, 2019; Ogburn, Shpitser and Tchetgen, 2020). When number of possible treatment assignments increases, suppose one finds a U that can factorize the distribution of T. T lies in disjoint regions of the space partitioned by U. For some T_m (m very large) and some $U = u$, one will have measure zero on some factor space:

$$P(\hat{U}(T_m) \mid U = u) \to 0 \tag{7}$$

Again, targeted imputation does not suffer from the violation of positivity assumption. There is no inverse weighting involved in any step and thus one should not be worried about the measure zero problem. As for the estimation of modified propensity score, positivity assumption requires $P(T \mid X_{\text{obs}}, Z) > 0$ and this is different from the critique above in equation (7).

Finally on the estimation part, Grimmer, Knox and Stewart (2020) point out that with U being a function of T, it will be extremely hard to estimate linear models due to perfect collinearity. For non-linear models, bias may occur if one were to use any form of penalized regressions (Wang and Blei, 2019). This concern does not apply to the imputation method and it does not apply to modified propensity score either.

6 Conclusion

In conclusion, the paper proposes a method to deal with missing data caused by unobserved confounders. Researchers often fail to utilize missing pattern to obtain more information regards to missing mechanism. The proposed method adopts dimension reduction techniques on missing matrix to extract useful information on unobserved confounders. As a result, the method can be applied to situation where missingness is nonignorable, and multiple imputation with observed values does not suffice. Being able to relax the missing at random assumption, the proposed

method can be applied to potentially a broad range of datasets for imputation purposes.

Potential applications include but not limited to sensitive survey questions, nonignorable missing values in report datasets by organizations / governments, and any other scenarios with unobserved confounders for missing values. Furthermore, for missing data with known structures (such as prior distribution of missing probabilities, certain assumption regarding respondent's rationality) one should be able to apply the method with specific choice of latent factor models. By doing so, researchers will be able to obtain imputation results with higher interpretability. As a demonstration, I provide a latent utility model version of imputation result using the same Chinese survey dataset.

Finally, more future research on the topic is needed. Potential topics include but not limited to sensitivity test on missing assumptions, robustness check on imputation results for different model selections. And more importantly, equal attention should be drawn to missing not at random with small datasets (low dimensionality datasets), where latent factor model does not work well.

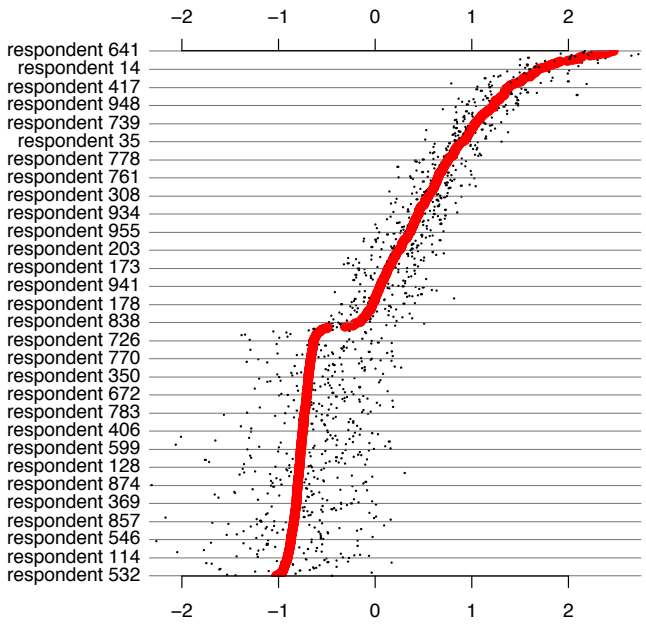


Figure 7: Pew (2017) 11 Questions on Political Knowledge, 1002 Respondents

A Uni-dimension example: Pew Political Knowledge Survey

During the data collection period, data points go missing due to attrition, refusal to answer, etc. In survey questions designed to test respondent's knowledge, a "don't know" answer is always provided. While most of the studies treat "don't know" as missing values and then conduct listwise deletion, researchers have been trying to get a better inference out of the "don't know" answers in a non-MAR setting (Kuha et al., 2018). Below I demonstrate how the proposed model can help with the problem.

Pew research conducts political knowledge survey (Pew, 2017), where respondents are asked to answer questions regarding to current political affairs. Figure 7 shows a preliminary result of applying the above model using the survey dataset. The data has 11 questions testing people's political knowledge and 1002 respondents participated. In the survey, they were given the option of "Don't know" or "Refuse to answer". These two choices are coded as 1 (missing) and all other answers are coded as 0 (observed).

Each data point in figure 7 is the location (posterior mean) of one respondent and the red line is the smoothed line. After examination, "respondent 641" (who is on very top of the graph) answered "Don't know" for all 11 questions. One can interpret the result as one's willingness to share their answers. On the extreme of 2 at the spectrum, we have the most discreet respondents, i.e they have the lowest threshold to refusal. On the other side of negative locations, we have the more open respondents who tend to feel more comfortable answering these questions.

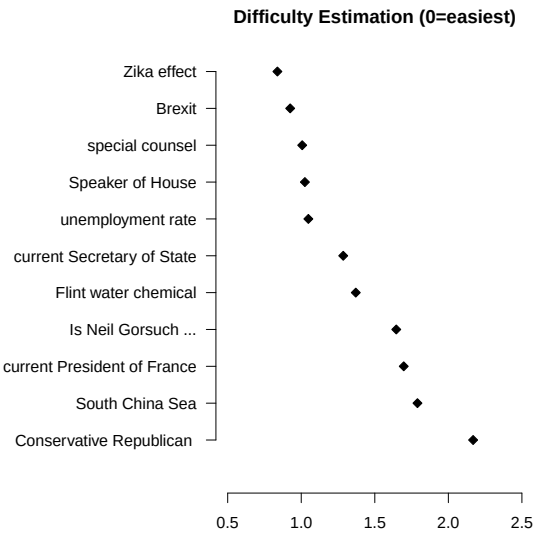


Figure 8

Furthermore, this model can also produce a parameter of "difficulty" for each question. As shown in figure 8, the easiest question (the one with highest probability of being answered) is "According to the CDC, humans are infected with the Zika virus primarily by...". Only 5 people didn't answer and most respondents gave the correct answer (mosquitoes). While the hardest question (the one with least probability of being answered) is "Many conservative Republicans in the House of Representatives are members of which of the following groups?", with "The Freedom Caucus" being the correct answer.

B Application using Latent Utility Model

This section presents an application using latent utility model, which is a kind of latent factor model. I start by briefly introducing latent utility model. I then use the model to deal with missing data in a sensitive survey question, where utility provides researchers with reasonable imputation result and more interpretability.

For each respondent r and each question q, we have the quadratic utility function:

$$u_{r,q} = -\left(\alpha_r - \delta_q\right)^T W_q \left(\alpha_r - \delta_q\right) + \epsilon_{r,q} \tag{8}$$

where $\epsilon_{r,q}$ follows a standard normal distribution.

A respondent refuses to answer a question when the thresold is reached $u_{r,q} \leq \bar{u}_q$:

$$\epsilon_{r,q} \leq \bar{u}_q + \left(\alpha_r - \delta_q\right)^T W_q \left(\alpha_r - \delta_q\right) \tag{9}$$

This leads to the following:

$$P(\text{refusal} = 1) = \Phi\left(\bar{u}_q + (\alpha_r - \delta_q)^T W_q (\alpha_r - \delta_q)\right) \tag{10}$$

$$P(\text{refusal} = 0) = 1 - \Phi\left(\bar{u}_q + (\alpha_r - \delta_q)^T W_q (\alpha_r - \delta_q)\right) \tag{11}$$

Intuitively, a respondent would deliberate on the decision whether to answer a survey question or not by assessing the answer he / she has in mind and the consequence of revealing it. The parameter α_r and δ_q represents the location for the respondent r and question q. When the distance between these two parameters is trivial, respondent should be willing to share his answer. When the distance reaches a certain threshold point, the respondent will be reluctant to answer. For example, for some sensitive questions, α_r can be the location of respondent's answer and δ_q represents a socially desirable answer. If the respondent speculates his answer will be too far from the socially desirable answer, he may refuse to answer. This then, will lead to the social desirability bias.

The log likelihood of the model would be:

$$\begin{aligned} L^{r,q}(\alpha, \bar{u}, \delta, W) = \sum_{r=1}^{R}\sum_{q=1}^{Q} \Bigg[& (1\{\text{refusal} = 1\}) \log \Phi\left(\bar{u}_q + (\alpha_r - \delta_q)^T W_q (\alpha_r - \delta_q)\right) \\ & + (1\{\text{refusal} = 0\}) \log\left(1 - \Phi\left(\bar{u}_q + (\alpha_r - \delta_q)^T W_q (\alpha_r - \delta_q)\right)\right) \Bigg] \end{aligned} \tag{12}$$

For the purpose of estimation, I simplify the model into following form:

$$\begin{aligned} P(y_{rq} = 1) &= \Phi\left(\bar{u}_q + (\alpha_r - \delta_q)^T W_q (\alpha_r - \delta_q)\right) \\ &= \Phi\left(\bar{u}_q + W_q^T d_r\right) \end{aligned}$$

where $d_r = \|\alpha_r - \delta_q\|^2$, $y_{rq} = 1$ if refused and 0 otherwise. This model can be estimated through a MCMC and Bayesian regression procedure (Clinton, Jackman and Rivers, 2004; Chalmers et al., 2012). In appendix A, I provide a toy example using Pew Research political knowledge survey (Pew, 2017).

B.1 Multidimension IRT

Then I extend the model into a multi-dimensional setting. For each dimension d:

$$P(y_{rqd} = 1) = \Phi\left(\bar{u}_q + W_{d,q}^T d_{d,r}\right) \tag{13}$$

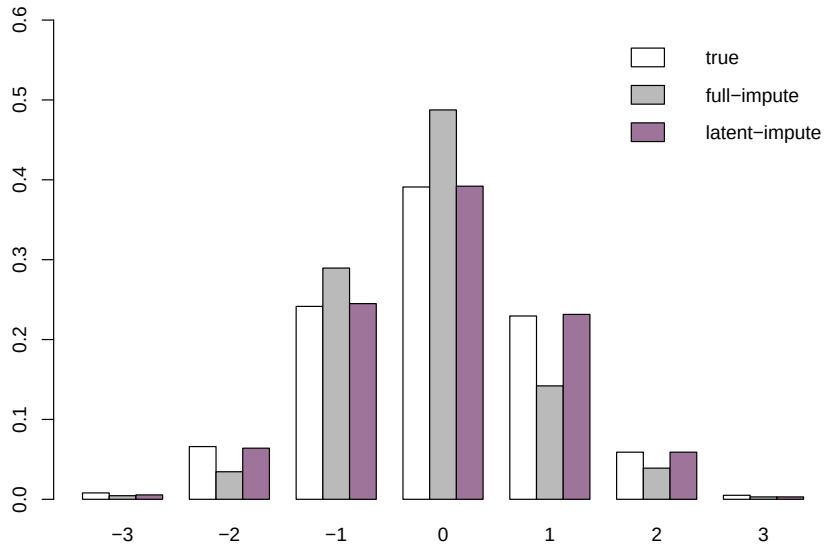


Figure 9: Simulation: Comparison between Naive Imputation and the Latent Utility Approach

For a given survey datasets, questions will focus on many aspects such as political knowledge, opinion and preference, private information and etc. A multi-dimension model will allow questions have different focuses while the one dimension model will assume the refusal to answer are solely controlled by one factor.

B.2 Simulation

The simulated dataset consists of 2000 observations with 20 variables. All 20 variables follow a normal distribution with a correlation of 0.5 among each other. I then transposed 10 variables into ordinal variables from -3 to 3. Missingness were taken follow MNAR mechanism: observations at both of the tails have higher probability of missing.

For the examples hereafter, I use K-means for clustering and `mice` (White, Royston and Wood, 2011) for multiple imputation.

Figure 9 shows the comparison between naive imputation method and the proposed method. Y-axis shows the proportion of each category and X-axis plots the 7 categories. The white bar is the true distribution of one of the ordinal variables in the simulated dataset. The gray bar is the imputation result of feeding the whole dataset into multiple imputation methods. Due to the MAR assumption, naive imputation methods over-impute observations into the middle categories. The proposed method manages to impute most of the observations correctly.

B.3 Sensitive survey datasets

I use the 2017 Chinese Netizen Survey (Ma, 2017), in which the proposed method helps to impute missing values in sensitive questions. The survey contains around 60 sensitive questions such as "what is your opinion towards Chairman Mao?" There are 2379 observations and 1314 of them are complete.

I conducted the proposed method step by step, with a latent utility model of two dimensions. Due to the high correlation (see figure 13 for a 3-D graph) between the two dimensions, I will only show dimension 1 for the descriptive visualizations below.

Figure 10 shows the pairwise correlation (with 95% confidence interval) between the covariates and scaling result for respondent $d_{d,r}, d = 1$. A positive correlation indicates that the higher level of the variable, the more likely the respondent refuses to answer the sensitive questions. Figure 10 shows that higher political salience (such as higher frequency of political discussion, higher political interests, more frequent usage of vpn, higher level of satisfaction towards society and being a government worker) positively correlates to a higher refusal rate. Also, women are more likely to refuse than men.

Figure 11 shows the "difficulty" level for each question included in the scaling model earlier. A question is easy when very few respondents refused and vice versa. Easy questions in this dataset include: "does national anthem make you proud?" Difficulty questions include "do you trust your village level legislators?"

B.4 Imputation results

Figure 12 shows the imputation result by multiple imputation and the proposed method. The white bars plot the observed data (with missing values). The gray bar shows a naive multiple imputation approach, where we assume MAR and use the whole dataset to impute missing values. The purple bars shows the imputation results by proposed method. The proposed method imputes more missing answers to be people with a more extreme ideology.

B.5 Regression analysis

To show the impact of different imputation procedures, I ran two regression analysis using the imputed data. For both regressions, the main independent variable is seven scale ideology question from the most left to 7 the most right. Outcome variable is people's opinions towards DPRK or USA. The first column of table 4 and table 5 show the results out of naive imputation, that is to use multiple imputation with the whole dataset. The second column of both tables show the results of targeted imputation using the proposed method. Regression results change dramatically between two different imputation procedures. For example, with naive imputation method, the

Figure 10

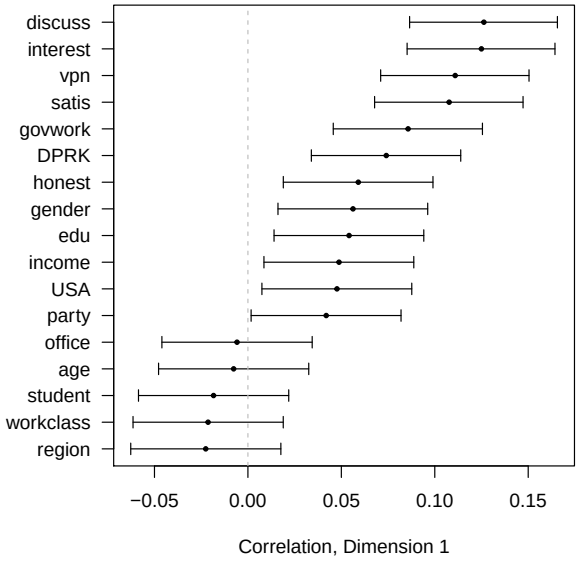


Figure 11

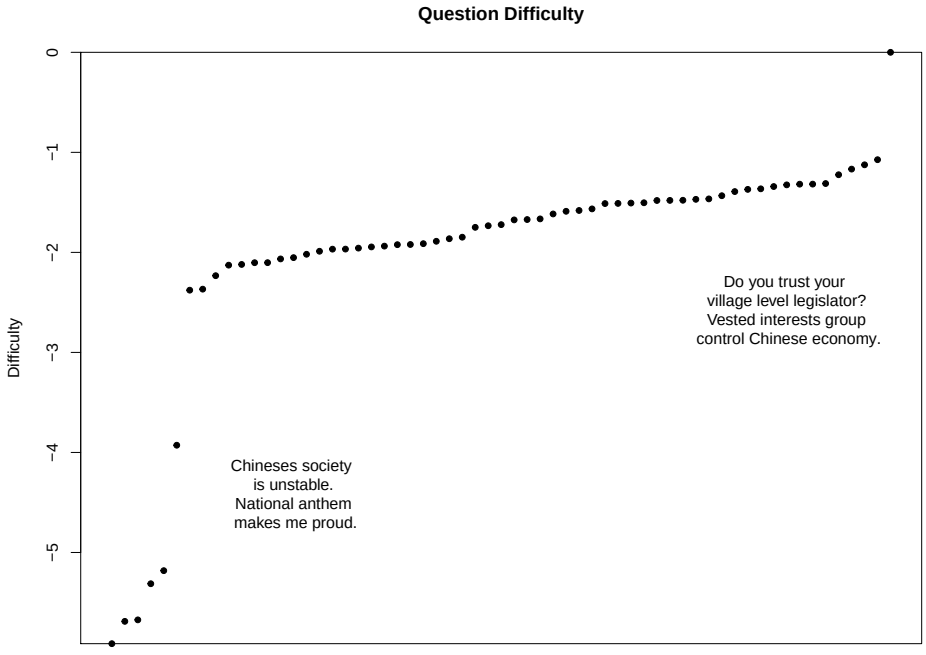
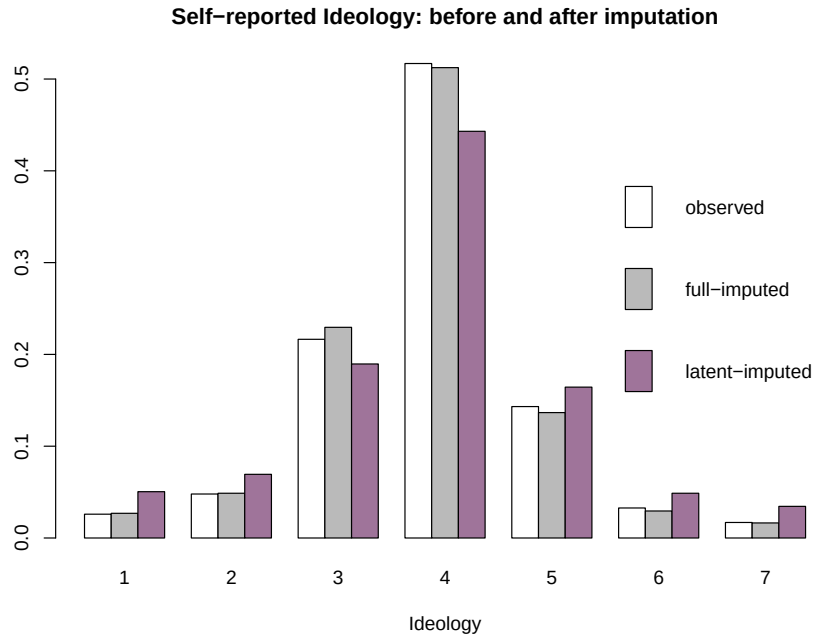


Figure 12



result indicates a significant correlation between ideology and support towards USA. However, the proposed result shows the correlation might not be statistically significant.

Again, the naive imputation method fills in the missing entries in an effort to best imitate the observed data. This might then exacerbate certain regression result. Proposed method avoided doing so by only imputing within clusters.

C Figures

D Other choices of latent factor model

Table 4

	Dependent variable:	
	Opinions on DPRK	
	Naive imputation	Proposed method
Ideology	−0.030	0.010
	(0.004)	(0.003)
gender	−0.032	0.001
	(0.008)	(0.009)
age	−0.008	0.002
	(0.002)	(0.002)
interest	−0.007	−0.009
	(0.005)	(0.006)
income	0.0001	−0.0005
	(0.0003)	(0.0003)
edu	−0.030	0.001
	(0.004)	(0.005)
vpn	0.027	0.004
	(0.005)	(0.005)
discuss	0.007	0.011
	(0.006)	(0.007)
satis	0.054	0.008
	(0.005)	(0.005)
Constant	0.028	−0.372
	(0.032)	(0.034)
Observations	2,379	2,379
Adjusted R^2	0.141	0.003
Residual Std. Error (df = 2369)	0.196	0.211
F Statistic (df = 9; 2369)	44.265	1.856
Note:		p<0.1; p<0.05; p<0.01

Table 5

	Dependent variable:	
	Opinions on USA	
	Naive imputation	Proposed method
Ideology	0.011	−0.003
	(0.003)	(0.003)
gender	0.006	0.006
	(0.007)	(0.007)
age	−0.004	−0.001
	(0.002)	(0.002)
income	0.0002	−0.0004
	(0.0002)	(0.0002)
interest	0.003	−0.004
	(0.004)	(0.004)
edu	0.031	−0.003
	(0.003)	(0.004)
vpn	0.016	−0.001
	(0.004)	(0.004)
discuss	0.006	0.006
	(0.005)	(0.005)
satis	−0.018	−0.001
	(0.004)	(0.004)
Constant	−0.408	−0.182
	(0.026)	(0.026)
Observations	2,379	2,379
Adjusted R^2	0.071	-0.0005
Residual Std. Error (df = 2369)	0.159	0.165
F Statistic (df = 9; 2369)	21.166	0.874

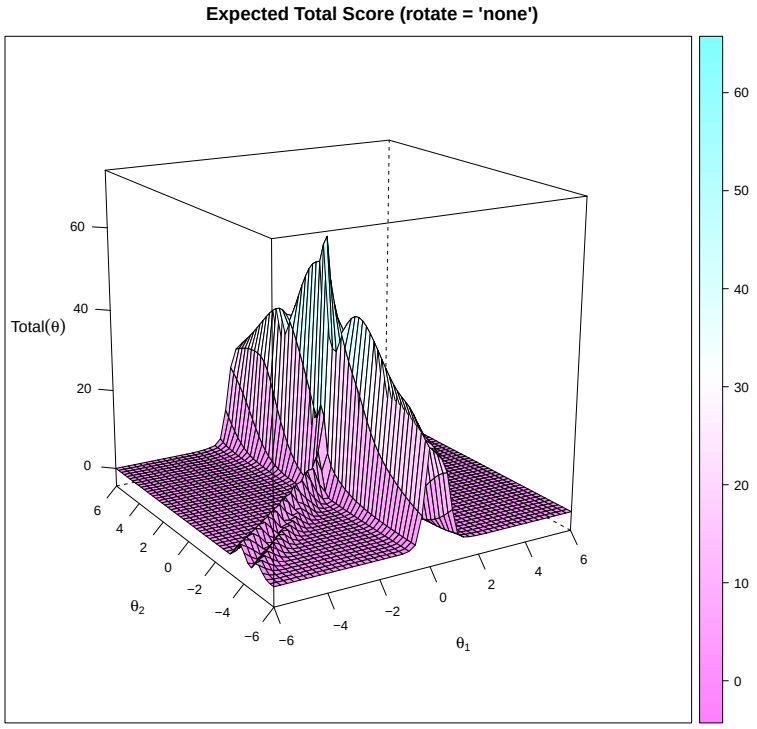
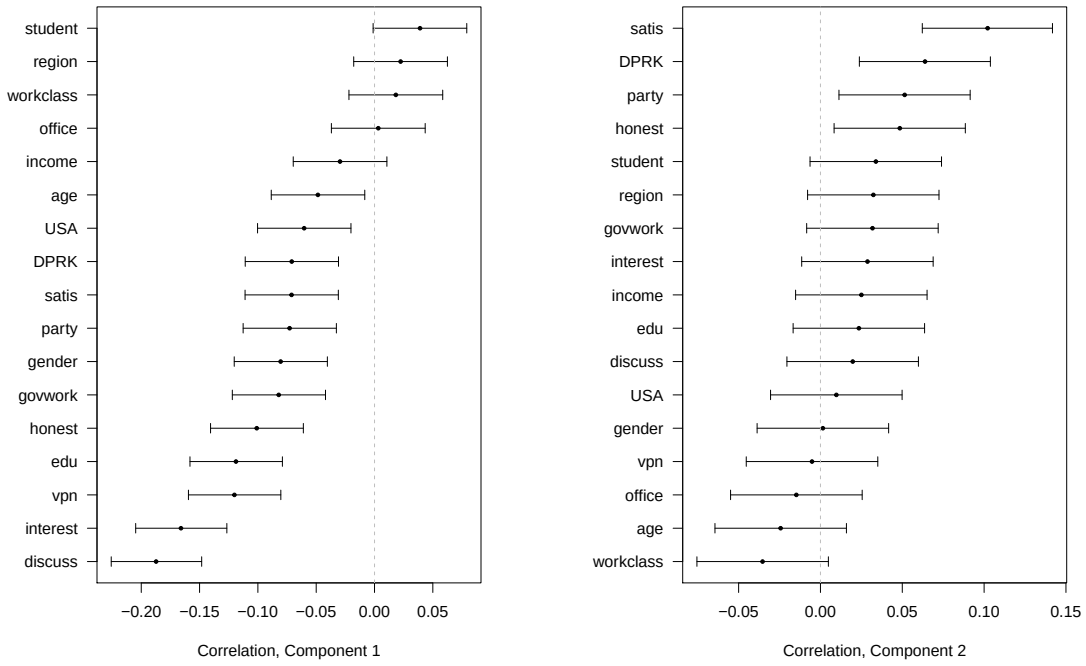
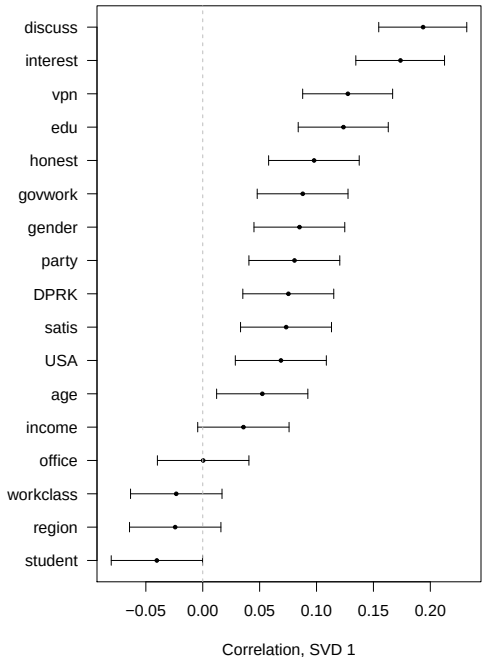
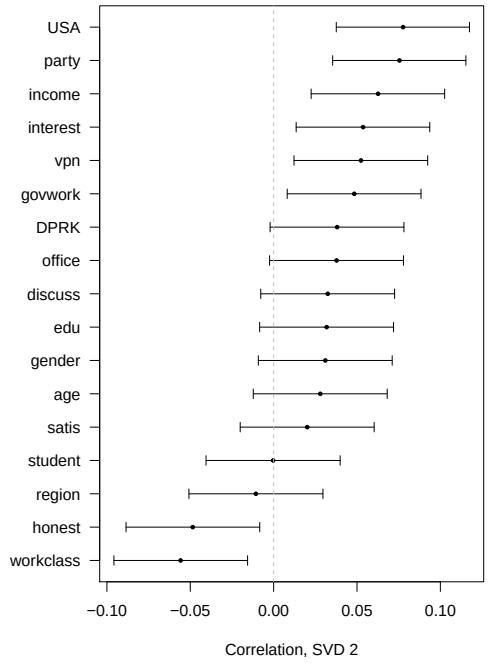


Figure 13: 3D depiction of the two dimensional latent utility model




discuss
interest
vpn
edu
honest
govwork
gender
party
DPRK
satis
USA
age
income
office
workclass
region
student
−0.05
0.00
0.05
0.10
0.15
0.20
Correlation, SVD 1


USA
party
income
interest
vpn
govwork
DPRK
office
discuss
edu
gender
age
satis
student
region
honest
workclass
−0.10
−0.05
0.00
0.05
0.10
Correlation, SVD 2

References

Bang, Heejung and James M Robins. 2005. "Doubly robust estimation in missing data and causal inference models." *Biometrics* 61(4):962–973.

Barabas, Jason, Jennifer Jerit, William Pollock and Carlisle Rainey. 2014. "The question (s) of political knowledge." *American Political Science Review* 108(4):840–855.

Blair, Graeme and Kosuke Imai. 2012. "Statistical analysis of list experiments." *Political Analysis* 20(1):47–77.

Chalmers, R Philip et al. 2012. "mirt: A multidimensional item response theory package for the R environment." *Journal of Statistical Software* 48(6):1–29.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The statistical analysis of roll call data." *American Political Science Review* 98(2):355–370.

D'Amour, Alexander. 2019. "On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives." *arXiv preprint arXiv:1902.10286* .

Gibson, James L and Gregory A Caldeira. 2009. "Knowing the Supreme Court? A reconsideration of public ignorance of the high court." *The Journal of Politics* 71(2):429–441.

Glynn, Adam N. 2013. "What can we learn with statistical truth serum? Design and analysis of the list experiment." *Public Opinion Quarterly* 77(S1):159–172.

Grimmer, Justin, Dean Knox and Brandon Stewart. 2020. "Design Still Trumps Analysis: Limitation of the Deconfounding Methods." *Working paper* .

Imai, Kosuke and Zhichao Jiang. 2019. "Comment: The Challenges of Multiple Causes." *Journal of the American Statistical Association* 114(528):1605–1610.

Kallus, Nathan, Xiaojie Mao and Madeleine Udell. 2018. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in neural information processing systems*. pp. 6921–6932.

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 1998. List-wise deletion is evil: what to do about missing data in political science. In *Annual Meeting of the American Political Science Association, Boston*.

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing incomplete political science data: An alternative algorithm for multiple imputation." *American political science review* 95(1):49–69.

Kuha, Jouni, Sarah Butt, Myrsini Katsikatsou and Chris J Skinner. 2018. "The Effect of Probing "Don't Know" Responses on Measurement Quality and Nonresponse in Surveys." *Journal of the American Statistical Association* 113(521):26–40.

Little, Roderick JA and Donald B Rubin. 2014. *Statistical analysis with missing data*. Vol. 333 John Wiley & Sons.

Luskin, Robert C and John G Bullock. 2011. ""Don't know" means "don't know": DK responses and the public's level of political knowledge." *The Journal of Politics* 73(2):547–557.

Ma, Deyong. 2017. "Chinese Netizen Survey 2017.".

Miller, Joanne M, Kyle L Saunders and Christina E Farhart. 2016. "Conspiracy endorsement as motivated reasoning: The moderating roles of political knowledge and trust." *American Journal of Political Science* 60(4):824–844.

Mondak, Jeffery J and Belinda Creel Davis. 2001. "Asked and answered: Knowledge levels when we will not take "don't know" for an answer." *Political Behavior* 23(3):199–224.

Ogburn, Elizabeth L, Ilya Shpitser and Eric J Tchetgen Tchetgen. 2020. "Counterexamples to" The Blessings of Multiple Causes" by Wang and Blei." *arXiv preprint arXiv:2001.06555* .

Pew. 2017. "Pew Research Center 2017 Weekly Survey." https://www.people-press.org/dataset/june-22-25-2017-weekly-survey.

Pietryka, Matthew T and Randall C MacIntosh. 2013. "An analysis of ANES items and their use in the construction of political knowledge scales." *Political Analysis* 21(4):407–429.

Ripley, Brian, Bill Venables, Douglas M Bates, Kurt Hornik, Albrecht Gebhardt, David Firth and Maintainer Brian Ripley. 2013. "Package 'mass'." *Cran R* 538.

Sportisse, Aude, Claire Boyer and Julie Josse. 2018. "Imputation and low-rank estimation with Missing Non At Random data." *arXiv preprint arXiv:1812.11409* .

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267–288.

Wang, Yixin and David M Blei. 2019. "The blessings of multiple causes." *Journal of the American Statistical Association* pp. 1–71.

White, Ian R, Patrick Royston and Angela M Wood. 2011. "Multiple imputation using chained equations: issues and guidance for practice." *Statistics in medicine* 30(4):377–399.

Wold, Svante, Kim Esbensen and Paul Geladi. 1987. "Principal component analysis." *Chemometrics and intelligent laboratory systems* 2(1-3):37–52.

Yang, Shu, Linbo Wang and Peng Ding. 2019. "Causal inference with confounders missing not at random." *Biometrika* 106(4):875–888.