Ensemble Fractional Imputation for Incomplete Categorical Data with a Graphical Model

Yonghyun Kwon [1] **Jae-Kwang Kim** [1]

Abstract

Missing data is common in practice, and standard statistical inference can be biased when missingness is related to the outcome of interest. We present a frequentist approach using a graphical model and fractional imputation, which can handle missing data for multivariate categorical variables under missing at random assumption. To avoid the problem due to the curse of dimensionality in multivariate data, we adopt the idea of a random forest to fit multiple reduced models and then combine multiple models using model weights. The model weights are computed from the novel method, double projection, where the observed likelihood is projected to the class of a graphical mixture model. The performance of the proposed method is investigated through an extensive simulation study.

1. Introduction

Missing data are encountered in many scientific areas, including survey sampling, social science, and clinical research. It is widely known that improper handling of missing values may cause biased estimates or efficiency loss and hinder rigorous statistical analysis. Imputation is a popular approach to handle item nonresponse. When the data are released to the public, by filling in the missing values using imputation, different data users can obtain the same results. Such consistency is particularly important for government agencies that produce official statistics.

Different types of imputation methods have been developed, including multiple imputation (MI) and fractional imputation (FI). MI, initially proposed by Rubin (2004), replaces missing data with multiple plausible values to create several datasets and has become a popular tool for addressing missing data. FI, originally proposed by Kalton & Kish (1984), on the other hand, creates a single completed dataset with additional information called fractional weights that reflect the probabilities of the candidate imputed values. FI has been developed further by Kim & Fuller (2004), Kim (2011), She & Wu (2019) and Sang et al. (2022).

Early efforts in MI include using parametric models such as multivariate normal distribution (Honaker et al., 2011). MICE is a popular algorithm for imputing incomplete datasets using chained equations with various regression methods (Van Buuren & Groothuis-Oudshoorn, 2011). More recently, state-of-the-art machine learning methods have been proposed to address missing data based on generative adversarial networks (GAN, Yoon et al. (2018)) or deep latent variable models (DLVM, Mattei & Frellsen (2019)). While MI is a common practice for dealing with item nonresponse, existing MI methods can be computationally intensive, especially when the dataset is high-dimensional.

In this article, we are mainly interested in developing fractional imputation with multivariate categorical data. In practice, many datasets contain various types of categorical variables, and continuous variables can also be summarized as categorical variables. For example, annual income can be reported into several categories of income groups. In developing fractional imputation for categorical data, the main difficulty is to handle the curse of dimensionality problems associated with high-dimensional categorical variables. Since the number of parameters increases exponentially with increasing dimension, imputation of high-dimensional categorical data remains challenging.

One remedy for such a problem is bagging, or ensemble learning method, where multiple bootstrap samples are extracted from the data, and each sample is used to train a separate model. The predictions from all fitted models are then averaged to obtain the bagged prediction. Stekhoven & Bühlmann (2012) proposed a seminal algorithm to employ random forests for single imputation: missForest. Although missForest can be applied to high-dimensional categorical data, experimental studies show that missForest may yield biased results when the missing pattern is nonmonotone, where there is no nested missing pattern of missingness, since the algorithm depends on the initial guess of missing

[1] Department of Statistics, Iowa State University, Ames, IA, USA. Correspondence to: Jae-Kwang Kim <jkim@iastate.edu>.

DMLR workshop at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA, 2023.

values and incorporates sorting of the features according to the missing proportions.

Unlike ordinary unweighted random forests, model averaging combines predictions from multiple models by assigning higher weights to the more reliable models. Recently, Xie et al. (2021) applied multiple imputation sure independence screening to the model averaging, particularly in the context of high-dimensional data. Following Zhang et al. (2015) and Rigollet (2012), our approach involves selecting model weights by utilizing the Kullback-Leibler loss of the aggregated estimator.

In this paper, we wish to fill this important research gap by proposing the so-called ensemble fractional imputation (EFI), where the "ensemble" means that we combine many imputation models. In order to overcome the challenge associated with high dimensional data, we propose a novel use of random forest where each tree uses a small subset of the variables, and the selected subset of variables is considered to build the relationship in the tree. To combine the multiple models efficiently, we apply information projection innovatively and develop a double projection method.

2. Basic Setup

Let $\boldsymbol{Y} = (Y_1, \cdots, Y_p)$ be a p-dimensional categorical random vector with support $\mathcal{Y}$. Define $\pi_{\mathbf{y}} = \mathbb{P}(\boldsymbol{Y} = \mathbf{y})$ for $\mathbf{y} \in \mathcal{Y}$ such that $\sum_{\mathbf{y}\in\mathcal{Y}} \pi_{\mathbf{y}} = 1$. Random variables $\boldsymbol{Y}$ are subject to missingness, and we assume an arbitrary pattern of missing data so that the missing pattern can be nonmonotone. Let $\boldsymbol{y}_i$ be the identical and independent i-th realization of $\boldsymbol{Y}$. Instead of observing $\boldsymbol{y}_i$, we only observe a subset of $\boldsymbol{y}_i = (\boldsymbol{y}_{i,\text{obs}}, \boldsymbol{y}_{i,\text{mis}})$, where $\boldsymbol{y}_{i,\text{obs}}$ and $\boldsymbol{y}_{\text{i,mis}}$ are the observed part and the missing part of $\boldsymbol{y}_i$, respectively. We define the response indicator functions $\boldsymbol{\delta}_i = (\delta_{i1}, \cdots, \delta_{ip})$ by

$$\delta_{ij} = \begin{cases} 1 & Y_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \cdots, n$. Since $\boldsymbol{y}_{i,\text{mis}}$ is not observed, we can develop fractional imputation using the conditional distribution of $\boldsymbol{y}_{i,\text{mis}}$ given $\boldsymbol{y}_{i,\text{obs}}$ and $\boldsymbol{\delta}_i$. We assume that the response mechanism is missing at random (MAR) in the sense that the missing mechanism does not depend on the missing variables after conditioning on the observed ones: $\boldsymbol{Y}_{\text{mis}} \perp \boldsymbol{\delta} \mid \boldsymbol{Y}_{\text{obs}}$.

Define $G(\boldsymbol{Y}, \boldsymbol{\delta})$ to be the mapping from $\boldsymbol{Y}$ to $\boldsymbol{Y}_{\text{obs}}$ based on $\boldsymbol{\delta}$. For each unit i with observed $\boldsymbol{y}_{i,\text{obs}} = G(\boldsymbol{y}_i, \boldsymbol{\delta}_i)$, fractional weights $w_{i\mathbf{y}}$ are assigned to each possible value of $\mathbf{y} \in \mathcal{Y}$. For categorical data, under MAR, the fractional weight of unit i assigned to the imputed value of $\mathbf{y}$ is constructed by

$$w_{i\mathbf{y}} = \mathbb{P}(\boldsymbol{Y} = \mathbf{y} \mid G(\boldsymbol{Y}, \boldsymbol{\delta}_i) = \boldsymbol{y}_{i,\text{obs}}) \tag{1}$$

where

$$\mathbb{P}(\boldsymbol{Y} = \mathbf{y} \mid \boldsymbol{G}(\boldsymbol{Y}, \boldsymbol{\delta}_i) = \boldsymbol{y}_{i,\text{obs}}) = \frac{\mathbb{P}(\boldsymbol{Y} = \mathbf{y})\mathbb{I}\left\{G(\boldsymbol{Y}, \boldsymbol{\delta}_i) = \boldsymbol{y}_{i,\text{obs}}\right\}}{\sum_{\mathbf{y}\in\mathcal{Y}} \mathbb{P}(\boldsymbol{Y} = \mathbf{y})\mathbb{I}\left\{G(\boldsymbol{Y}, \boldsymbol{\delta}_i) = \boldsymbol{y}_{i,\text{obs}}\right\}}$$

for each $i = 1, \cdots, n$ and $\mathbf{y} \in \mathcal{Y}$, following Ibrahim (1990) and Kim (2011). The fractional weight is the conditional probability of missing part $\boldsymbol{y}_{i,\text{mis}}$, given the observed part $\boldsymbol{y}_{i,\text{obs}}$. The fractional weights (1) satisfy

$$\sum_{\mathbf{y}\in\mathcal{Y}} w_{i\mathbf{y}} = 1, \qquad i = 1, \cdots, n,$$

and

$$\sum_{\mathbf{y}\in\mathcal{Y}} w_{i\mathbf{y}}\mathbb{I}\left\{y_j = k\right\} = \mathbb{P}\left[Y_j = k \mid G(\boldsymbol{Y}, \boldsymbol{\delta}_i) = \mathbf{y}_{i,\text{obs}}\right]$$

for $i = 1, \cdots, n$, $\delta_{ij} = 0$, and $k \in \mathcal{Y}_j$.

To compute the conditional probability in (1), we need to estimate the joint probabilities $\pi_{\mathbf{y}} = \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})$ from the partial observations, where $\mathbf{y} \in \mathcal{Y}$. Suppose that the joint probability under model $\mathcal{M}_k, k = 1, \cdots, K$ can be written as

$$\pi_{\mathbf{y}} = \pi_{\mathbf{y}}(\boldsymbol{\theta}_k)$$

where $\boldsymbol{\theta}_k$ is the model parameter for model $\mathcal{M}_k$. The dimension of $\boldsymbol{\theta}_k$ determines the level of sparsity in the model. The observed log-likelihood function can be written as

$$l_{\text{obs}}(\boldsymbol{\theta}_k) = \sum_{i=1}^{n} \sum_{d\in\mathcal{D}(\boldsymbol{\delta}_i)} \mathbb{I}(\boldsymbol{y}_{i,\text{obs}} = d) \log\{\pi_d(\boldsymbol{\theta}_k)\} \tag{2}$$

where $\pi_d(\boldsymbol{\theta}_k) = \mathbb{P}\left\{G(\boldsymbol{Y}, \boldsymbol{\delta}_i) = d; \boldsymbol{\theta}_k\right\} = \sum_{\{y\in\mathcal{Y}\}} \pi_{\boldsymbol{y}}(\boldsymbol{\theta}_k)\mathbb{I}\{G(\boldsymbol{y}, \boldsymbol{\delta}_i) = d\}$ and $\mathcal{D}(\boldsymbol{\delta}_i)$ is the support of $\boldsymbol{y}_{i,\text{obs}} = G(\boldsymbol{Y}, \boldsymbol{\delta}_i)$. To find the maximizer of $l_{\text{obs}}(\boldsymbol{\theta}_k)$ in (2), we can use the EM algorithm of Dempster et al. (1977). The details of the EM algorithm for categorical data under the parametric model are elucidated in the Appendix.

In practice, the true model is unknown. To find the best model, we may use

$$AIC(k) = -2l_{\text{obs}}(\hat{\boldsymbol{\theta}}_k) + 2p_k \tag{3}$$

to choose the best model, where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}_k$ and p_k is the dimension of $\boldsymbol{\theta}_k$. The model with the smallest value of $AIC(k)$ will be chosen.

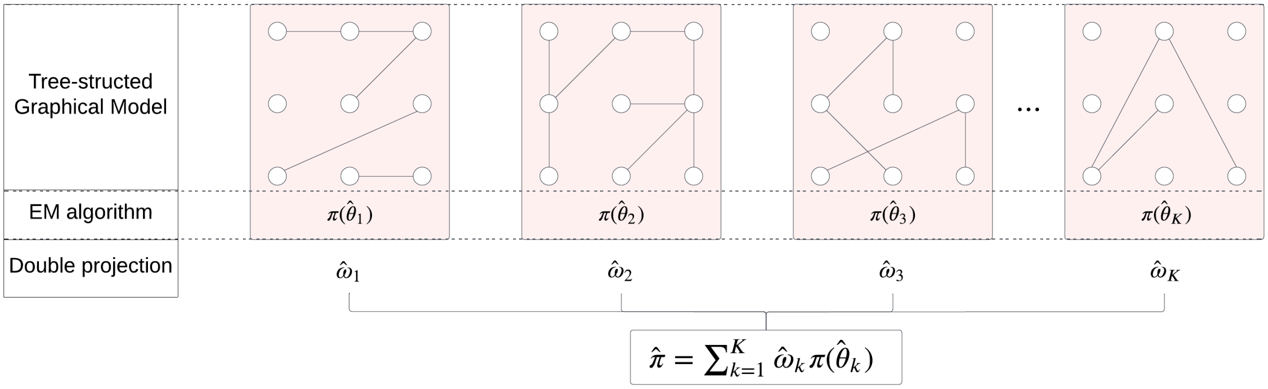

Figure 1. Graphical Forest for Ensemble Fractional Imputation

If the dimension p of $\boldsymbol{Y}$ is large, then the computational burden associated with the AIC can be huge. One cannot compute the MLEs for all possible models and compare the AIC for all possible models. It is a fundamental problem associated with high-dimensional data.

To overcome the aforementioned issues, we propose an ensemble model-averaging approach that attempts to integrate multiple individual models under general missing mechanisms that lead to improved prediction accuracy. The basic idea of the proposed method is illustrated in Figure 1.

3. Double projection

In this study, we focus our attention on tree models. An undirected graphical model, $\mathcal{T} = (V, E)$ with nodes V and edges E, is a tree if it is connected and has no cycles. If a probability measure P is Markov with respect to the tree $\mathcal{T}$, P can be factorized as

$$\mathbb{P}(\boldsymbol{y}) = \prod_{i\in V} \mathbb{P}(y_i) \prod_{(i,j)\in E} \frac{\mathbb{P}(y_i, y_j)}{\mathbb{P}(y_i)\mathbb{P}(y_j)}. \tag{4}$$

Expression (4) can be seen as a special case of the exponential graphical model or log-linear model (Loh & Wainwright, 2013). Given a finite set of samples, the maximum likelihood estimator of a tree is the maximum weight spanning tree that maximizes the sum of empirical mutual information of the edges, as described by Chow & Liu (1968).

Consider a discrete probability measure P and a set of candidate tree models $\mathcal{T}_k : (V, E_k)$ characterized by an edge E_k such that $E_k \neq E_\ell$, $\forall k \neq \ell$. We define the mixture graphical models $\mathcal{M}(P)$ for a given distribution P:

$$\begin{aligned}\mathcal{M}(P) &:= \Pi\left(P \mid \mathcal{T}_1\right) + \Pi\left(P \mid \mathcal{T}_2\right) + \cdots + \Pi\left(P \mid \mathcal{T}_k\right) \\ &:= \left\{\omega_1 \Pi\left(P \mid \mathcal{T}_1\right) + \cdots + \omega_k \Pi\left(P \mid \mathcal{T}_k\right) : \omega_k \in \Omega^+\right\}\end{aligned}$$

where $\Omega^+ = \left\{\omega_k : \sum_{k=1}^K \omega_k = 1, \omega_k \geq 0\right\}$,

$$\Pi\left(P \mid \mathcal{T}_k\right) = \operatorname*{arg\,min}_{Q\in\mathcal{T}_k} D_{\mathrm{KL}}(P \mid Q), \tag{5}$$

and

$$\begin{aligned}D_{\mathrm{KL}}(P \mid Q) &= \sum_{\boldsymbol{y}\in\mathcal{Y}} P(\boldsymbol{y}) \log \frac{P(\boldsymbol{y})}{Q(\boldsymbol{y})} \\ &= E_P\left\{\log P(Y)\right\} - E_P\left\{\log Q(Y)\right\}\end{aligned}$$

is the Kullback-Leibler divergence which is the expectation of the logarithmic difference between the probabilities P and Q evaulated at P. Note that $Q_k = \Pi\left(P \mid \mathcal{T}_k\right)$ is a projection of P onto $\mathcal{T}_k$ in the sense that the KL divergence of $Q \in \mathcal{T}_k$ evaluated at P is minimized at Q_k. That is, Q_k is an element in $\mathcal{T}_k$ that approximates P as closely as possible in terms of KL divergence.

Now, we are interested in finding an element in $\mathcal{M}(P)$ that approximate P as closely as possible. To achieve this goal, we apply the projection of P onto $\mathcal{M}(P)$ indexed by ω. We define the *double-projection* of P onto the mixture model $\mathcal{M}(P)$ as follows:

$$\begin{aligned}\Pi\left(P \mid \mathcal{M}(P)\right) &= \operatorname*{arg\,min}_{\sum_{k=1}^K \omega_k Q_k \in \mathcal{M}(P)} D_{\mathrm{KL}}\left(P \mid \sum_{k=1}^K \omega_k Q_k\right) \qquad (6) \\ &= \omega_1^* Q_1 + \cdots + \omega_K^* Q_K \in \mathcal{M}(P)\end{aligned}$$

Note that

$$\begin{aligned}f(\boldsymbol{\omega}) &:= -D_{\mathrm{KL}}\left(P \mid \sum_{k=1}^K \omega_k Q_k\right) \\ &= \sum_{\boldsymbol{y}\in\mathcal{Y}} P(\boldsymbol{y}) \log\left(\sum_{k=1}^K \omega_k Q_k(\boldsymbol{y})\right) + \text{const} \qquad (7)\end{aligned}$$

is a strictly convex function of $\boldsymbol{\omega}$ and $\Omega^{+} = \left\{\boldsymbol{\omega} : \sum_{k=1}^{K} \omega_k = 1, \omega_k \geq 0\right\}$ is compact. Therefore, the solution to (6) always exists and is unique. The choice of weights via a Kullback-Leibler distance was first proposed by Rigollet (2012). They aggregated a collection of fixed components of function to achieve the performance of the best model under a given class.

Lemma 3.1 (Hardy–Littlewood inequality)**.** *Suppose that $H(P, Q_1) \geq \cdots \geq H(P, Q_k)$ where $H(\cdot, \cdot)$ is a cross entropy defined by $H(P, Q) := -\sum_{y \in \mathcal{Y}} P(y) \log Q(y)$. If $\omega_1 \leq \cdots \leq \omega_k$, the following rearrangement inequality holds:*

$$H\left(P, \sum_{k=1}^{K} \omega_k Q_k\right) \leq H\left(P, \sum_{k=1}^{K} \omega_k Q_{\sigma(k)}\right) \tag{8}$$

where $\sigma(k)$ is a permutation of $\{1, \cdots, K\}$.

Lemma 3.1 implies that the solution to (6) preserves the order of $H(P, Q_k)$. That is,

$$\begin{aligned} &\omega_1^* \leq \omega_2^* \leq \cdots \leq \omega_K^* \\ &\quad \iff H(P, Q_1) \geq H(P, Q_2) \geq \cdots \geq H(P, Q_K). \end{aligned}$$

Furthermore, if Q_k has a tree structure, as described by (Chow & Liu, 1968),

$$D_{\mathrm{KL}}(P \mid Q_k) = -\sum_{(i,j) \in E_k} I(Y_i, Y_j) + \sum_{i \in V} H(Y_i) - H(\boldsymbol{Y}) \tag{9}$$

where $I(Y_i, Y_j)$ is the mutual information between Y_i and Y_j, and $H(Y_i)$ and $H(\boldsymbol{Y})$ and the marginal and joint entropies under P.

(8) and (9) imply that the followings are equivalent

$$\begin{aligned} &\omega_1^* \leq \hat{\omega}_2 \leq \cdots \leq \omega_K^* \\ &\quad \overset{(8)}{\iff} KL\left(P \mid Q_1\right) \geq \cdots \geq KL\left(P \mid Q_K\right) \\ &\quad \overset{(9)}{\iff} \sum_{(i,j) \in E_1} I(Y_i, Y_j) \leq \cdots \leq \sum_{(i,j) \in E_K} I(Y_i, Y_j) \end{aligned}$$

In fact, by Jensen's inequality, we have

$$KL\left(P \mid \sum_{k=1}^{K} \omega_k Q_k\right) \leq \sum_{k=1}^{K} \omega_k KL(P \mid Q_k) \tag{10}$$

and the RHS of (10) is minimized when the probability mass of $\boldsymbol{\omega}$ is concentrated on the minimum $KL(P \mid Q_k)$.

Proposition 3.2. *Let $E_k \subset V \times V$ be a singleton so that $|E_k| = 1$. If P is a distribution of a tree model associated with a tree $\mathcal{T} = (V, E)$, and $\boldsymbol{\omega}^*$ is the solution to (6), then*

$$\omega_k^* = 0 \tag{11}$$

if $E_k \not\subset E$.

The proposition above states that the model weights are sparse and can be nonzero only if the corresponding nodes of the underlying tree model are connected. Indeed, as supported by simulation studies in Appendix, edges recovered from double projection can be much sparser than the true tree model $\mathcal{T} = (V, E)$.

Algorithm 1 Ensemble Fractional Imputation

```
Input: data $\boldsymbol{y}_i$ and response indicator $\boldsymbol{\delta}_i$, $i = 1, \cdots, n$.
1. EM algorithm
for $k = 1$ to $K$ do
  Construct a sparse tree model $\mathcal{T}_k$.
  Estimate $\hat{\theta}_k$ using the EM algorithm.
end for
2. Double projection
Estimate $\hat{\boldsymbol{\omega}} = (\hat{\omega}_1, \cdots, \hat{\omega}_K)$ by the Kullback-Leibler aggregation. The estimated probability is
  $$\hat{\mathbb{P}}(\boldsymbol{y}) = \sum_{k=1}^{K} \hat{\omega}_k \mathbb{P}(\boldsymbol{y}; \hat{\theta}_k)$$
3. Fractional Imputation
for $i = 1$ to $n$ do
  for $\boldsymbol{z} =$ possible values of $\boldsymbol{y}_{i,\mathrm{mis}}$ do
    $w_{i\boldsymbol{z}} \leftarrow \hat{\mathbb{P}}\left(\boldsymbol{y}_{i,\mathrm{mis}} = \boldsymbol{z} \mid \boldsymbol{y}_{i,\mathrm{obs}}\right)$
  end for
end for
```

4. Proposed method

Our proposed method involves (1) preparing candidate models and estimating parameters from candidate models, (2) computing model weights assigned to candidate models, and (3) implementing fractional imputation from the aggregated model. We first construct K different models, each of which is a sparse tree model $\mathcal{T}_k = (V, E_k)$ whose parameter estimation is computationally efficient. For example, $\mathcal{T}_k$ can be constructed by selecting a fixed number of edges $\tilde{E}_k$ from all possible sets of edges and rejecting $\tilde{E}_k$ if $(V, \tilde{E}_k)$ does not form a tree structure. We repeat random selection until $(V, \tilde{E}_k)$ is a tree and set $\mathcal{T}_k = (V, \tilde{E}_k)$. Let $\boldsymbol{\theta}_k$ be the parameter associated with the tree model $\mathcal{T}_k$. Then $\boldsymbol{\theta}_k$ can be estimated by applying the EM algorithm independently within each model $\mathcal{T}_k$.

Once the estimators $\hat{\boldsymbol{\theta}}_k$ are estimated from the EM algorithm, we construct the model weights using double projection (6). If $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ are fully observed, and $\hat{P}$ denotes its empirical distribution without missing values, we approxi-

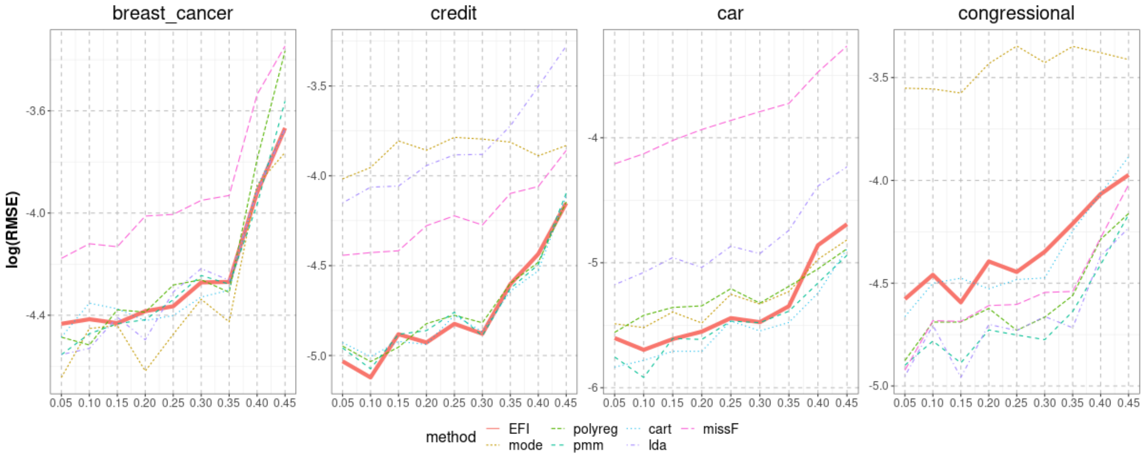

Figure 2. RMSE under nonmonotone MAR

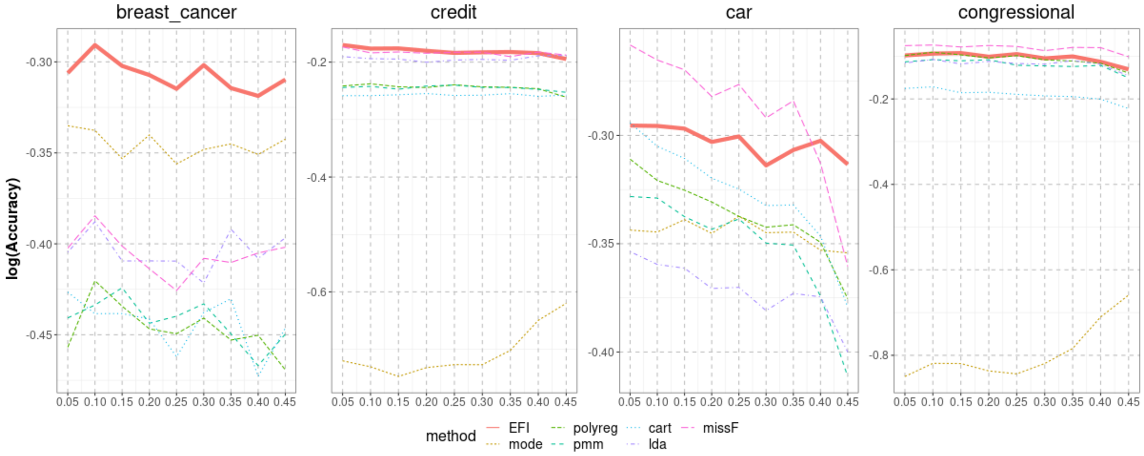

Figure 3. Classification accuracy under nonmonotone MAR

mate (4) by double projection (6):

$$\min_{\boldsymbol{\omega}\in\Omega^+} D(\hat{P} \mid \mathcal{M}(\hat{P})) = \max_{\boldsymbol{\omega}\in\Omega^+} \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \omega_k \hat{Q}_k(\boldsymbol{y}_i; \hat{\boldsymbol{\theta}}_k)\right), \tag{12}$$

where $\hat{Q}_k = \Pi\left(\hat{P} \mid \mathcal{T}_k\right)$ and

$$\Omega^+ = \left\{(\omega_1, \cdots, \omega_K)^T : \sum_{k=1}^{K} \omega_k = 1, \omega_k \geq 0\right\}.$$

(12) is related to the synthetic control method (Abadie et al., 2010). (12) can be regarded as a special case of the maximum likelihood aggregate suggested by (Rigollet, 2012) if Q_k's are fixed distributions.

When missing data are present, we first estimate $\hat{\boldsymbol{\theta}}_k$ and $\hat{Q}_k = \Pi\left(\hat{P}_{\text{obs}} \mid \mathcal{T}_k\right)$ using the EM algorithm. Then the optimal model weights are computed by solving

$$\hat{\boldsymbol{\omega}} = \operatorname*{arg\,max}_{\boldsymbol{\omega}\in\Omega^+} \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \omega_k \hat{Q}_k(\boldsymbol{y}_{i,\text{obs}}; \hat{\boldsymbol{\theta}}_k)\right). \tag{13}$$

The joint probability of $\boldsymbol{Y}$ is now approximated by $\hat{Q}$ as follows

$$\hat{Q}(\cdot) = \sum_{k=1}^{K} \hat{\omega}_k \hat{Q}_k(\cdot; \boldsymbol{\theta}_k).$$

One of the great advantages of the proposed method is that the missing data pattern does not need to be MCAR and can be MAR. The following theorem explains why the ensemble method works in the MAR setup if the underlying tree model is sparse. According to the theorem, the MAR condition under the full model can imply the MAR condition under the reduced model when the reduced model is true.

Theorem 4.1. *[MAR under the reduced model (Lemma 7.2. in (Kim & Shao, 2021))] Suppose $E \subset E_k(\mathcal{M} \subset \mathcal{M}_k)$ for some k and the MAR condition holds in the sense that*

$$\boldsymbol{Y}_{\text{mis}} \perp \boldsymbol{\delta} \mid \boldsymbol{Y}_{\text{obs}}$$

and $\mathcal{M}_k$ selects A_k variables. That is,

$$\mathcal{M}_k : \mathbb{P}(\boldsymbol{Y}) = \mathbb{P}(\boldsymbol{Y}(A_k)) \prod_{j \notin A_k} \mathbb{P}(Y_j).$$

Then MAR condition holds under $\mathcal{M}_k$ in the sense that

$$\boldsymbol{Y}_{\text{mis}}(A_k) \perp \boldsymbol{\delta}(A_k) \mid \boldsymbol{Y}_{\text{obs}}(A_k)$$

The idea of bootstrapping in random forests can still be applied in the double projection. To control the estimation error related to the double use of data when estimating $\hat{\boldsymbol{\theta}}_k$ and $\hat{\omega}_k$, Chernozhukov et al. (2018) suggested the idea of cross-fitting, where the sample is randomly divided into two, and each of them is used to estimate different sets of models to avoid the bias due to nonparameteric estimation of the model parameters. Building on this idea, we first take bootstrap samples and use them to estimate the model parameters $\hat{\boldsymbol{\theta}}_k$. Then the model weights $\hat{\omega}_k$ are estimated using the double projection, where its observed entropy is estimated with the out-of-bag samples.

Ultimately, the parameter of interest, say $p = \mathbb{P}(\boldsymbol{Y}(A) = \mathbf{y}_0)$ for a subvector $\boldsymbol{Y}(A) := \{Y_a\}_{a \in A}$, can be estimated using the estimated model parameters $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots \hat{\theta}_K)$ and the estimated mixture weights $\hat{\boldsymbol{\omega}} = (\hat{\omega}_1, \cdots, \hat{\omega}_K)$. We can estimate p by

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} w_{i\mathbf{y}} \mathbb{I}(\boldsymbol{y}(A) = \boldsymbol{y}_0) \tag{14}$$

where the fractional weights $w_{i\mathbf{y}}$ are defined by

$$w_{i\mathbf{y}} = \frac{\hat{Q}(\boldsymbol{y})\mathbb{I}\{G(\boldsymbol{Y}, \boldsymbol{\delta}_i) = \boldsymbol{y}_{i,\text{obs}}\}}{\sum_{y \in \mathcal{Y}} \hat{Q}(\boldsymbol{y})\mathbb{I}\{G(\boldsymbol{Y}, \boldsymbol{\delta}_i) = \boldsymbol{y}_{i,\text{obs}}\}} \tag{15}$$

The entire algorithm of the Ensemble Fractional Imputation is presented in Algorithm 1.

5. Real data experiments

5.1. Downstream tasks

The performance of the EFI method in downstream tasks is assessed together with various machine learning algorithms. We consider the imputation of categorical datasets from the UCI Machine Learning Repository: `breast_cancer`, `credit`, `car`, and `congressional (voting record)` datasets. Missing values are generated in the datasets by missing at random(MAR). The results for MCAR case is available on the Appendix. The missing rate varies from 5% to 45%.

For each realized sample, we apply the following methods to estimate the parameters and impute missing values:

- (EFI) The proposed Ensemble Fractional Imputation method. In each tree, $[\sqrt{p}]$ number of variables(or nodes) are randomly selected.
- (mode) mode imputation in which the missing values are replaced with the sample mode
- (polyreg, pmm, cart, lda) MICE using the corresponding regression methods. The imputation size is set to $M = 5$.
- (missFst) missForest algorithm proposed by Stekhoven & Bühlmann (2012).

The parameter of interest is the probability of a variable equal to its mode and is computed from the population data without missing values. The root mean squared error for the parameter of interest and the classification accuracy of the imputed values are evaluated.

5.2. Imputation performance

Figure 2 and Figure 3 demonstrate the influence of the selected imputation method on the performance of downstream tasks. Each method is evaluated using various data sets and across varying degrees of missing data. It can be seen from the summarized figure that our proposed Ensemble Fractional Imputation method gives better results when it comes to both estimation(RMSE) and prediction(classification error). Table 5.1 and Table 5.1 show that our proposed method provides the lowest RMSE and the highest accuracy in many datasets. It is worth noticing that mode imputation can perform as well as, or even better than, other imputation methods for some categorical datasets.

5.3. Graphical Structure

We also considered the graphical interpretation after double projection using `congressional voting record` data, assuming that all values are observed. As we noticed in Proposition 3.2, the aggregated forests are sparse, as the estimated model weights are also sparse. Fine gray edges are the collection of all edges of candidate trees, whereas the colored edges are selected edges after the double projection. Edges of identical colors represent that they are built from the same tree. A variable of interest, *party*, is connected to five nodes by eight edges.

6. Conclusion

We propose ensemble fractional imputation as a tool for general-purpose estimation for multivariate categorical data under item nonresponse. The basic idea is to use a weighted

	breast_cancer	credit	car	congressional
EFI	1.247(1.018)	**0.724(0.685)**	0.388(0.336)	1.235(0.880)
mode	**0.987(0.990)**	2.112(1.024)	0.416(0.397)	3.232(1.700)
polyreg	1.243(1.162)	0.804(0.788)	0.477(0.346)	0.982(0.747)
pmm	1.206(1.164)	0.774(0.776)	0.365(0.348)	**0.885(0.885)**
cart	1.202(1.134)	0.717(0.718)	**0.332(0.319)**	1.081(0.960)
lda	1.115(1.118)	1.939(0.619)	0.649(0.337)	0.909(0.852)
missF	1.810(1.689)	1.387(0.848)	1.957(1.463)	0.996(0.803)

Table 1. RMSE (and its standard error) of the imputation estimators over four datasets multiplied by 100 when the missing rate is 20%.

	breast_cancer	credit	car	congressional
EFI	**0.736(0.062)**	**0.835(0.029)**	0.739(0.027)	0.904(0.044)
mode	0.712(0.053)	0.481(0.036)	0.708(0.024)	0.433(0.061)
polyreg	0.640(0.073)	0.783(0.033)	0.718(0.026)	0.901(0.039)
pmm	0.642(0.072)	0.785(0.031)	0.709(0.029)	0.897(0.046)
cart	0.643(0.068)	0.775(0.034)	0.726(0.030)	0.832(0.056)
lda	0.664(0.070)	0.819(0.027)	0.690(0.027)	0.894(0.044)
missF	0.661(0.079)	0.831(0.024)	**0.754(0.035)**	**0.928(0.035)**

Table 2. Classification accuracy (and its standard error) of the imputation estimators over four datasets when the missing rate is 20%.

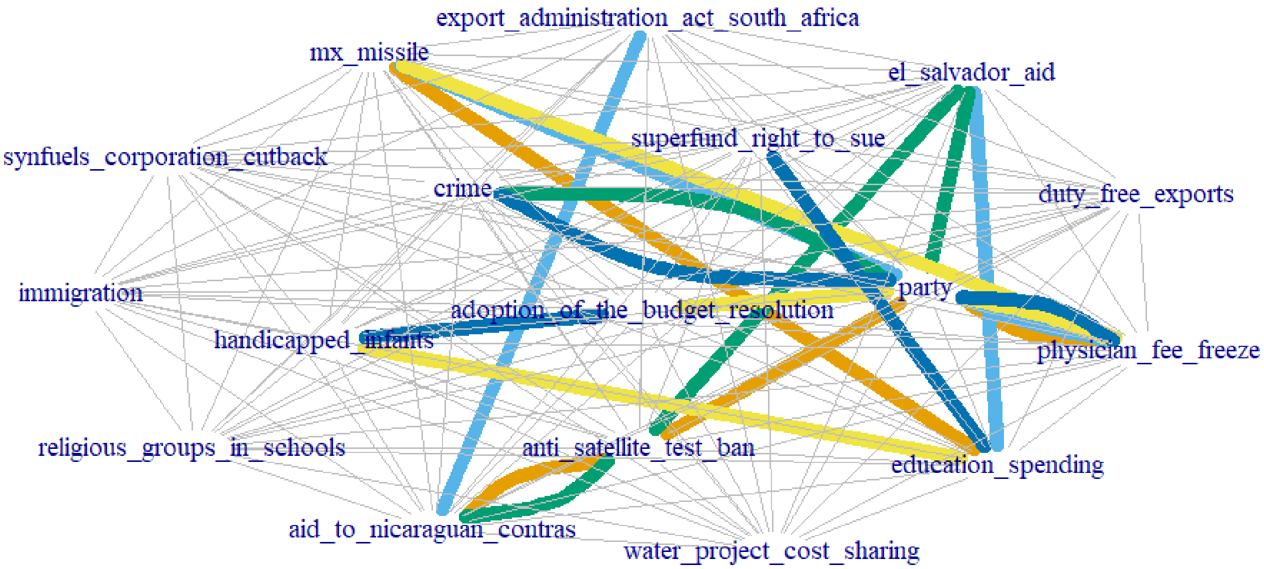
export_administration_act_south_africa
mx_missile
el_salvador_aid
synfuels_corporation_cutback
superfund_right_to_sue
crime
duty_free_exports
immigration
party
adoption_of_the_budget_resolution
handicapped_infants
physician_fee_freeze
religious_groups_in_schools
anti_satellite_test_ban
education_spending
aid_to_nicaraguan_contras
water_project_cost_sharing

Figure 4. Selected edges of `congressional voting record` data example.

random forest for prediction, and the model weights are computed by the double projection. In each tree, the usual EM algorithm and fractional imputation can be developed easily using the selected variables only. The model weights can be included to be a part of fractional weights in fractional imputation. Once the ensemble fractional imputation is established, users can use the fractionally imputed data and estimate various parameters of interest without worrying about the missingness mechanism. Unlike multiple imputation, fractional imputation creates a single imputed data file, and the resulting analysis is relatively simple.

Acknowledgements

This research was partially supported by the National Science Foundation (NSF) grant number OAC-1931380. We thank Sayantani Ghosh and Inho Cho for helpful discussions and code suggestions. We also thank the reviewers for their valuable comments.

References

Abadie, A., Diamond, A., and Hainmueller, J. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal* *of the American statistical Association*, 105(490):493–505, 2010.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1), 2018.

Chow, C. and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Honaker, J., King, G., and Blackwell, M. Amelia ii: A program for missing data. *Journal of statistical software*, 45:1–47, 2011.

Ibrahim, J. G. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411): 765–769, 1990.

Kalton, G. and Kish, L. Some efficient random imputation methods. *Communications in Statistics A*, 13:1919–1939, 1984.

Kim, J. K. Parametric fractional imputation for missing data analysis. *Biometrika*, 98(1):119–132, 2011.

Kim, J. K. and Fuller, W. Fractional hot deck imputation. *Biometrika*, 91(3):559–578, September 2004.

Kim, J. K. and Shao, J. *Statistical methods for handling incomplete data*. Chapman and Hall/CRC, 2021.

Kirshner, S., Smyth, P., and Robertson, A. W. Conditional chow-liu tree structures for modeling discrete-valued vector time series. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 317–324, 2004.

Loh, P.-L. and Wainwright, M. J. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, 41(6): 3022–3049, 2013.

Mattei, P.-A. and Frellsen, J. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pp. 4413–4423. PMLR, 2019.

Rigollet, P. Kullback-leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40 (2):639–665, 2012.

Rubin, D. B. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

Sang, H., Kim, J. K., and Lee, D. Semiparametric fractional imputation using gaussian mixture models for handling multivariate missing data. *Journal of the American Statistical Association*, 117(538):654–663, 2022.

She, X. and Wu, C. Fully efficient joint fractional imputation for incomplete bivariate ordinal responses. *Statistica Sinica*, 29(1):409–430, 2019.

Stekhoven, D. J. and Bühlmann, P. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Xie, J., Yan, X., and Tang, N. A model-averaging method for high-dimensional regression with missing responses at random. *Statistica Sinica*, 31(2):1005–1026, 2021.

Yoon, J., Jordon, J., and Schaar, M. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pp. 5689–5698. PMLR, 2018.

Zhang, X., Zou, G., and Carroll, R. J. Model averaging based on kullback-leibler distance. *Statistica Sinica*, 25: 1583–1598, 2015.

A. Appendix

A.1. EM algorithm

In the E-step, we compute

$$Q\left(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_k^{(t)}\right) \equiv \mathbb{E}\{l_{\text{com}}(\boldsymbol{\theta}_k) \mid \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}_k^{(t)}\},$$

where

$$l_{\text{com}}(\boldsymbol{\theta}_k) = \sum_{i=1}^{n} \sum_{\mathbf{y} \in \mathcal{Y}} \mathbb{I}(\boldsymbol{y}_i = \mathbf{y}) \log\{\pi_{\mathbf{y}}(\boldsymbol{\theta}_k)\}.$$

Note that the E-step in the categorical data is simply computing

$$\begin{aligned} p_i^{(t)}(\boldsymbol{y}) &\equiv \mathbb{E}\{\mathbb{I}(\boldsymbol{Y} = \boldsymbol{y}) \mid G(\boldsymbol{y}, \boldsymbol{\delta}_i) = \boldsymbol{y}_{i,\text{obs}}; \boldsymbol{\theta}_k^{(t)}\} \\ &= \mathbb{I}\left\{\boldsymbol{y}_{\text{obs},i} = G(\boldsymbol{y}, \boldsymbol{\delta}_i)\right\} \cdot \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \mid G(\boldsymbol{y}, \boldsymbol{\delta}_i) = \boldsymbol{y}_{i,\text{obs}}; \boldsymbol{\theta}_k^{(t)}), \end{aligned} \tag{16}$$

where

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \mid G(\boldsymbol{y}, \boldsymbol{\delta}_i) = \boldsymbol{y}_{i,\text{obs}}; \boldsymbol{\theta}_k^{(t)}) = \begin{cases} \frac{\pi_{\mathbf{y}}(\boldsymbol{\theta}_k^{(t)})}{\sum_{\mathbf{y} \in \mathcal{Y}} \pi_{\mathbf{y}}(\boldsymbol{\theta}_k^{(t)}) \mathbb{I}\left\{G(\boldsymbol{y}, \boldsymbol{\delta}_i) = \boldsymbol{y}_{i,\text{obs}}\right\}} & \text{if } \boldsymbol{y}_{i,obs} = G(\boldsymbol{y}, \boldsymbol{\delta}_i) \\ 0 & \text{otherwise.} \end{cases}$$

Using $p_i^{(t)}(\mathbf{y})$ in (16), we obtain

$$Q\left(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_k^{(t)}\right) = \sum_{\mathbf{y} \in \mathcal{Y}} \sum_{i=1}^{n} p_i^{(t)}(\mathbf{y}) \log\{\pi_{\mathbf{y}}(\boldsymbol{\theta}_k)\} = \sum_{\mathbf{y} \in \mathcal{Y}} n_c^{(t)} \log\{\pi_{\mathbf{y}}(\boldsymbol{\theta}_k)\},$$

where $n_c^{(t)} = \sum_{i=1}^{n} p_i^{(t)}(\mathbf{y})$. In the M-step, we update parameters by

$$\boldsymbol{\theta}_k^{(t+1)} \leftarrow \arg\max_{\theta_M} Q\left(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_k^{(t)}\right). \tag{17}$$

Thus, in the M-step, we have only to replace n_c by $n_c^{(t)}$ in the formula for MLE of $\boldsymbol{\theta}_k$. We iterate the E-step and M-step until convergence.

A.2. Proof of Theorem 4.1

Proof. Let $A_k^c = \{1, \cdots, p\} \setminus A_k$. The independence assumption

$$\left(\boldsymbol{Y}_{\text{mis}}(A_k), \boldsymbol{Y}_{\text{obs}}(A_k)\right) \perp \left(\boldsymbol{Y}_{\text{mis}}(A_k^c), \boldsymbol{Y}_{\text{obs}}(A_k^c)\right)$$

implies

$$\boldsymbol{Y}_{\text{mis}}(A_k) \perp \boldsymbol{Y}_{\text{obs}}(A_k^c) \mid \boldsymbol{Y}_{\text{obs}}(A_k)$$

by weak union of CI (conditional independence). Furthermore, the MAR condition

$$\boldsymbol{Y}_{\text{mis}}(A_k) \perp \boldsymbol{\delta}(A_k) \mid \left(\boldsymbol{Y}_{\text{obs}}(A_k), \boldsymbol{Y}_{\text{obs}}(A_k^c)\right)$$

implies

$$\boldsymbol{Y}_{\text{mis}}(A_k) \perp \boldsymbol{\delta}(A_k) \mid \boldsymbol{Y}_{\text{obs}}(A_k)$$

by contraction of CI. □

A.3. Extension to High-dimensional data

When $\boldsymbol{Y}$ is high-dimensional, aggregating different models to predict each component of $\boldsymbol{Y}$ can be unstable. If $\boldsymbol{X}$ is fully observed and we are only interested in estimating $\boldsymbol{Y}$ whose dimension is relatively small, we can use the KL divergence between the conditional distributions. Suppose that the joint probability of $\mathbb{P}(\boldsymbol{Z}) = \mathbb{P}(\boldsymbol{X}, \boldsymbol{Y})$ is known. We try to find the model weights $\hat{\boldsymbol{\omega}}$ that minimizes

$$\mathbb{E}\left[KL\left(\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X}), \sum_{k=1}^{K} \omega_k \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X}^{(k)})\right)\right] = \sum_{\boldsymbol{X} \in \mathcal{X}} \sum_{\boldsymbol{Y} \in \mathcal{Y}} \mathbb{P}(\boldsymbol{X}) \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X}) \log \frac{\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X})}{\sum_{k=1}^{K} \omega_k \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X}^{(k)})}.$$

It follows that

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega} \in \Omega^+}{\arg\max} \sum_{\boldsymbol{X} \in \mathcal{X}} \sum_{\boldsymbol{Y} \in \mathcal{Y}} \mathbb{P}(\boldsymbol{X}, \boldsymbol{Y}) \log \left(\sum_{k=1}^{K} \omega_k \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X}^{(k)}) \right) \tag{18}$$

(18) motivates

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega} \in \Omega^+}{\arg\max} \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \omega_k \mathbb{P}(\boldsymbol{Y}_{\text{obs}} = \boldsymbol{y}_{i,\text{obs}} \mid \boldsymbol{X}_{\text{obs}} = \boldsymbol{x}_{i,\text{obs}}^{(k)}) \right) \tag{19}$$

As noted by Kirshner et al. (2004), we can write

$$\mathbb{E}\left[KL\left(\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X}), \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X}^{(k)})\right)\right] = -\sum_{(i,j) \in E_k} I(Y_i, Y_j) - \sum_{(i,j) \in E_k} I(Y_i, X_j) + \sum_{i \in V} H(Y_i) - H(\boldsymbol{Y} \mid \boldsymbol{X}^{(k)})$$

and develop a similar double projection as in the low-dimensional case.

A.4. Experiments under MCAR

	breast_cancer	credit	car	congressional
EFI	**1.246(1.213)**	**0.562(0.561)**	0.691(0.380)	1.061(0.860)
mode	1.332(1.335)	1.025(1.029)	0.559(0.561)	1.448(1.455)
polyreg	1.493(1.402)	0.676(0.644)	0.424(0.348)	0.906(0.724)
pmm	1.435(1.402)	0.664(0.667)	0.427(0.424)	0.831(0.833)
cart	1.345(1.340)	0.664(0.663)	**0.402(0.381)**	0.941(0.945)
lda	1.349(1.354)	1.280(0.569)	0.874(0.377)	0.856(0.745)
missF	1.751(1.676)	1.158(0.970)	1.536(1.521)	**0.796(0.742)**

Table 3. RMSE (and its standard error) of the imputation estimators over four datasets multiplied by 100 when the missing rate is 20%.

	breast_cancer	credit	car	congressional
EFI	**0.728(0.060)**	**0.833(0.033)**	0.760(0.025)	0.927(0.040)
mode	0.712(0.053)	0.550(0.041)	0.699(0.022)	0.533(0.063)
polyreg	0.627(0.069)	0.776(0.035)	0.746(0.023)	0.919(0.044)
pmm	0.629(0.053)	0.780(0.031)	0.726(0.024)	0.920(0.052)
cart	0.622(0.059)	0.769(0.037)	0.752(0.027)	0.891(0.043)
lda	0.659(0.060)	0.812(0.035)	0.719(0.025)	0.930(0.043)
missF	0.658(0.055)	0.833(0.031)	**0.791(0.026)**	**0.951(0.032)**

Table 4. Classification accuracy (and its standard error) of the imputation estimators over four datasets when the missing rate is 20%.

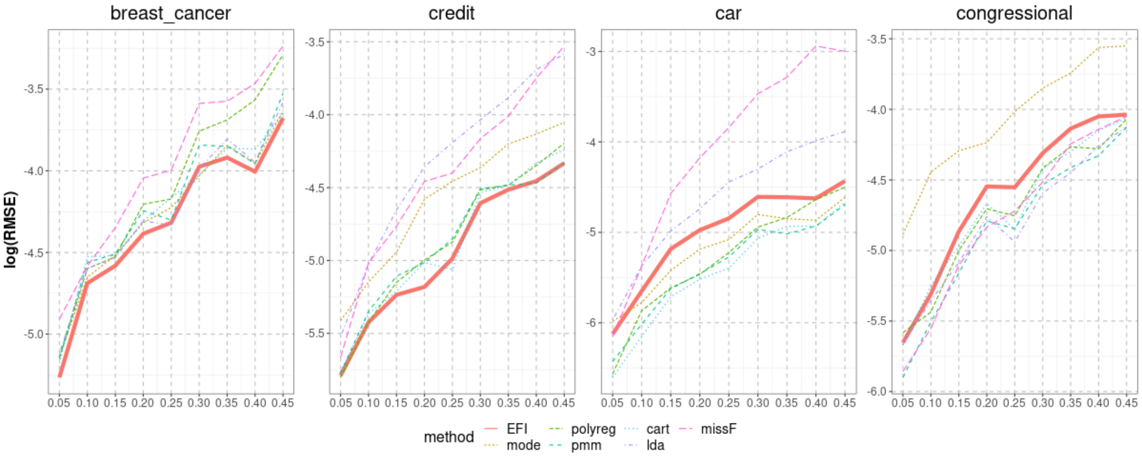

Figure 5. RMSE under nonmonotone MCAR

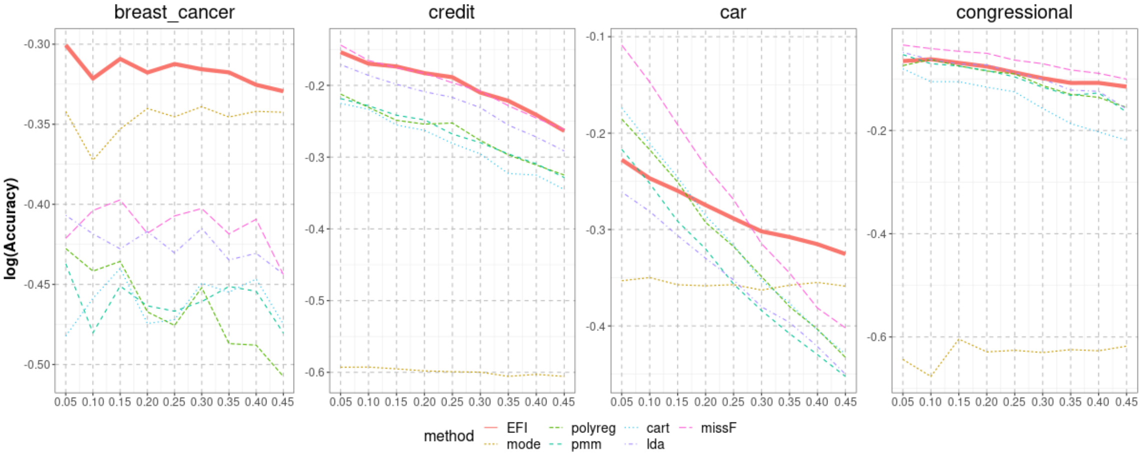

Figure 6. Classification accuracy under nonmonotone MCAR

A.5. Further experiments

We conduct a further simulation study using synthetic data from the Ising model. The symmetric Ising model has the probability mass function

$$f_\beta\left(x^1,\cdots,x^p\right)=\frac{1}{z(\beta)}\exp\left(\sum_{i<j}\beta_{ij}x^ix^j\right),x^i\in\{-1,1\} \tag{20}$$

where $z(\beta)=\sum_{x\in\{-1,1\}^p}\exp\left(\sum_{i<j}\beta_{ij}x^ix^j\right)$. The sample size and the number of variables are chosen to be $n=500$ and $p=10$. The interaction $\beta_{ij}=1$ if $(i,j)=(1,2),(1,5),(2,7),(3,4),(3,7),(3,9),(5,8),(5,9),(6,8),(6,10)$ and $\beta_{ij}=0$ otherwise. It should be noted that (20) is not a tree model but still recovers the graphical structure after the double projection, as can be seen in Figure 7 if all the possible models are included in the set of candidate models. Figure 8 is an example of a recovered graph by applying EFI to a realized dataset under missingness. The missing mechanism is either MCAR or MAR. Under MCAR, each variable Y_{ij} is missing with probability 0.5 independently. Under MAR, we randomly choose κ from $1,2,\cdots,p=10$. If, say, $\kappa=7$, then we observe Y_7,Y_8,Y_9,Y_{10}, and Y_1. The missing rate $\delta_2,\delta_3,\delta_4,\delta_5,\delta_6$ of (Y_2,Y_3,Y_4,Y_5,Y_6) given (Y_7,Y_8,Y_9,Y_{10},Y_1) is determined by

$$P(\delta_2=1|Y_7)=\begin{cases}0 & \text{if } Y_7=-1\\ 0.5 & \text{if } Y_7=1\end{cases},\cdots,P(\delta_6=1|Y_1)=\begin{cases}0 & \text{if } Y_1=-1\\ 0.5 & \text{if } Y_1=1\end{cases}$$

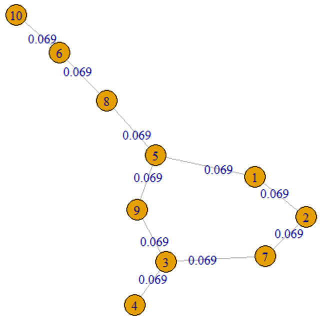

Figure 7. True graphical model that generates $\mathbf{Y}$

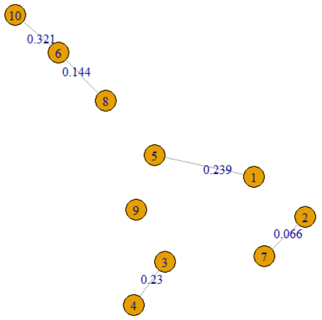

Figure 8. Graphical model estimated by EFI

The parameter of interest is $\theta_1 = \mathbb{P}(Y_1 = Y_2 = Y_4 = Y_5 = -1)$, $\theta_2 = \mathbb{P}(Y_1 = -1, Y_2 = 1)$, and $\theta_3 = \mathbb{P}(Y_1 = 1, Y_6 = -1)$. The simulation results are summarized in Table 5 and Table 6.

θ_1

	BIAS	SE	RMSE
Full(Oracle)	0.0009	0.0130	0.0130
CC	0.0088	0.0481	0.0489
MICE(pmm)	0.0002	0.0170	0.0170
missFst	0.0052	0.0448	0.0451
EFI	-0.0100	0.0118	**0.0154**

θ_2

	BIAS	SE	RMSE
Full(Oracle)	0.0011	0.0207	0.0207
CC	-0.0056	0.0349	0.0353
MICE(pmm)	-0.0055	0.0330	0.0334
missFst	-0.0197	0.0749	0.0775
EFI	0.0175	0.0258	**0.0312**

θ_3

	BIAS	SE	RMSE
Full(Oracle)	0.0007	0.0172	0.0172
CC	0.0007	0.0347	0.0347
MICE(pmm)	0.0000	0.0266	0.0266
missFst	-0.0173	0.0717	0.0738
EFI	0.0234	0.0234	**0.0330**

Table 5. MCAR, $\mathbb{P}(\delta = 0) = 0.5$

θ_1

	BIAS	SE	RMSE
Full(Oracle)	-0.0011	0.0124	0.0124
CC	-0.0083	0.0216	0.0231
MICE(pmm)	-0.0020	0.0144	0.0145
missFst	0.0058	0.0238	0.0245
EFI	-0.0084	0.0118	**0.0144**

θ_2

	BIAS	SE	RMSE
Full(Oracle)	0.0032	0.0187	0.0190
CC	-0.0101	0.0245	0.0265
MICE(pmm)	0.0036	0.0223	**0.0225**
missFst	-0.0067	0.0322	0.0329
EFI	0.0126	0.0222	0.0256

θ_3

	BIAS	SE	RMSE
Full(Oracle)	-0.0037	0.0174	0.0178
CC	-0.0017	0.0233	0.0233
MICE(pmm)	-0.0018	0.0214	**0.0215**
missFst	-0.0278	0.0333	0.0434
EFI	0.0124	0.0204	0.0239

Table 6. MAR, $\mathbb{P}(\delta = 0) = 0.5$