Information Theoretic Approaches for Testing Missingness in Predictive Modeling

Shreyas Bhave[1], Rajesh Ranganath[2], Adler Perotte[1]

[1]Columbia University Department of Biomedical Informatics

[2]New York University, The Courant Institute

COLUMBIA | COLUMBIA UNIVERSITY DEPARTMENT OF BIOMEDICAL INFORMATICS

What are we *missing* by making assumptions about missing data?

R := **missingness pattern**
X := **data matrix**
Y := **outcome, fully observed**

X_{obs} := **observed portion of data**
X_{mis} := **missing portion of data**

	Assumptions	What can be done?	Challenges
MCAR	$R \perp\!\!\!\perp X_{obs}\ ,\ R \perp\!\!\!\perp X_{mis} \mid X_{obs}$	mean impute, marginal sampling	how do you know data is MCAR?
MAR	$R \perp\!\!\!\perp X_{mis} \mid X_{obs}$	multiple imputation e.g. MICE, MissForest	comp expensive, how do you know data is MAR?
MNAR	any data that violates MAR	model missingness process e.g. graphical modeling	biased models, poor inference, dataset shift

Data is often in this category but MAR is assumed anyway

Intuition and domain knowledge about data generation process are often valuable but are there more *rigorous, general* ways to *test* assumptions?

MI-MCAR: Mutual Information for Missing Completely at Random

MCAR: $\boxed{R \perp\!\!\!\perp X_{obs}}$, $R \perp\!\!\!\perp X_{mis} \mid X_{obs}$

OAR (Observed at Random)

$$\hat{I}(R, X_{obs}) = \hat{H}(R) - \hat{H}(R \mid X_{obs})$$

$$\hat{H}(R) = -\frac{1}{N}\sum_{i=1}^{N} log(p_R(r_i))$$

$$\hat{H}(R \mid X_{obs}) = -\frac{1}{N}\sum_{i=1}^{N} log(p_{R|X_{imp}}(r_i \mid x_i))$$

$$\hat{ct} = \sum_{b=1}^{B} \mathbb{1}\left(\hat{I}(R, X_{obs}) \leq \hat{I}^b\right)$$

$$\hat{p} = \frac{1}{B+1}\left(1 + \hat{ct}\right)$$

Little's Test for MCAR	• assumes data are continuous, normal • comparing means within a missingness pattern to some true estimated population mean • *only continuous data, limiting parametric assumptions*
MI-MCAR (ours)	• use mutual information (MI) to build test statistic for independence • randomization test • MI is *robust to transformations, nonparametric* • can accommodate continuous and categorical data

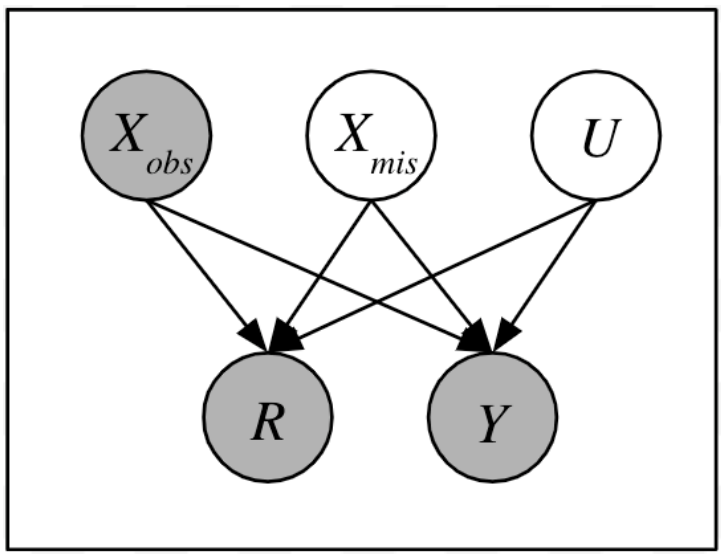
Algorithm 1 MI-MCAR

Input: $X \in \mathbb{R}^{NxP}$, $R \in \{0,1\}^{NxP}$
Output: p, the p-value where null hypothesis is $R \perp\!\!\!\perp X_{obs}$
Use multiple imputation to get X_{imp} from X
Fit p_R using density estimation
Fit $p_{R|X_{imp}}$ using some conditional model
Compute $\hat{I}(R, X_{obs})$ using p_R and $p_{R|X_{imp}}$
for $j \in [1, 2, \ldots, B]$ **do**
  Sample R^j from p_R
  Fit $p_{R^j|X_{imp}}$ using same conditional model specification
  Compute $\hat{I}_j := \hat{I}(R^j, X_{obs})$ using p_R and $p_{R_j|X_{imp}}$
end for
Compute $p := \frac{1}{B+1}\left(1 + \sum_{j=1}^{B} \mathbb{1}\left(\hat{I}(R, X_{obs}) \leq \hat{I}_j\right)\right)$

MI-US: Mutual Information for Unobserved Sources

how to test this condition?

MAR: $R \perp\!\!\!\perp X_{obs}$, $\boxed{R \perp\!\!\!\perp X_{mis} \mid X_{obs}}$



Idea: *we can use Y as a surrogate for information in the missing data*

MI-US: Conditional randomization test (CRT) as in Candes et al.[1] with conditional mutual information as test statistic.

Null Hypothesis: $R \perp\!\!\!\perp Y \mid X_{obs}$

To obtain samples from the null we can directly model $P(R \mid X_{obs})$

$$I(R, Y \mid X_{obs}) = H(Y \mid X_{obs}) - H(Y \mid X_{obs}, R)$$
$$I_{null}(\tilde{R}, Y \mid X_{obs}) = H(Y \mid X_{obs}) - H(Y \mid X_{obs}, \tilde{R})$$

Algorithm 2 MI-US

Input: $X \in \mathbb{R}^{NxP}$, $R \in \{0,1\}^{NxP}$, Y
Output: p, the p-value where null hypothesis is $R \perp\!\!\!\perp \mathrm{Y}|\mathrm{X}_{obs}$
Use multiple imputation to get X_{imp} from X
Fit $P_{Y|X_{imp},R}$ using some conditional model
Fit $P_{R|X_{imp}}$ using some conditional model
Compute $\hat{H}(Y|X_{obs}, R) := -\frac{1}{N}\sum_{i=1}^{N} log P_{Y|X_{imp},R}(y_i|x_i, r_i)$
for $j \in [1, 2, \ldots, B]$ **do**
  Sample $\tilde{R}^j$ from $p_{R|X_{imp}}$
  Fit $P_{Y|X_{imp},\tilde{R}^j}$ using same conditional model
  Compute $\hat{H}_j := -\frac{1}{N}\sum_{i=1}^{N} log P_{Y|X_{imp},\tilde{R}^j}(y_i|x_i, r_i^j)$
end for
Compute $p := \frac{1}{B+1}\left(1 + \sum_{j=1}^{B} \mathbb{1}\left(\hat{H}(Y|X_{obs}, R) \geq \hat{H}_j\right)\right)$

[1] Candes, E., Fan, Y., Janson, L., and Lv, J. Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Experiments & Discussion

- MI-MCAR Simulated Data
  - mixture of continuous normal and binary data
  - missingness simulated
  - logistic models, MADE for density estimation
- MI-US Simulated Data
  - binary outcome Y simulated with random logistic
  - continuous features from multivariate normal
  - used logistic to estimate conditional model
- MI-US Semi-Simulated MNIST
  - simple CNN model specification
  - missingness simulated with masking approach

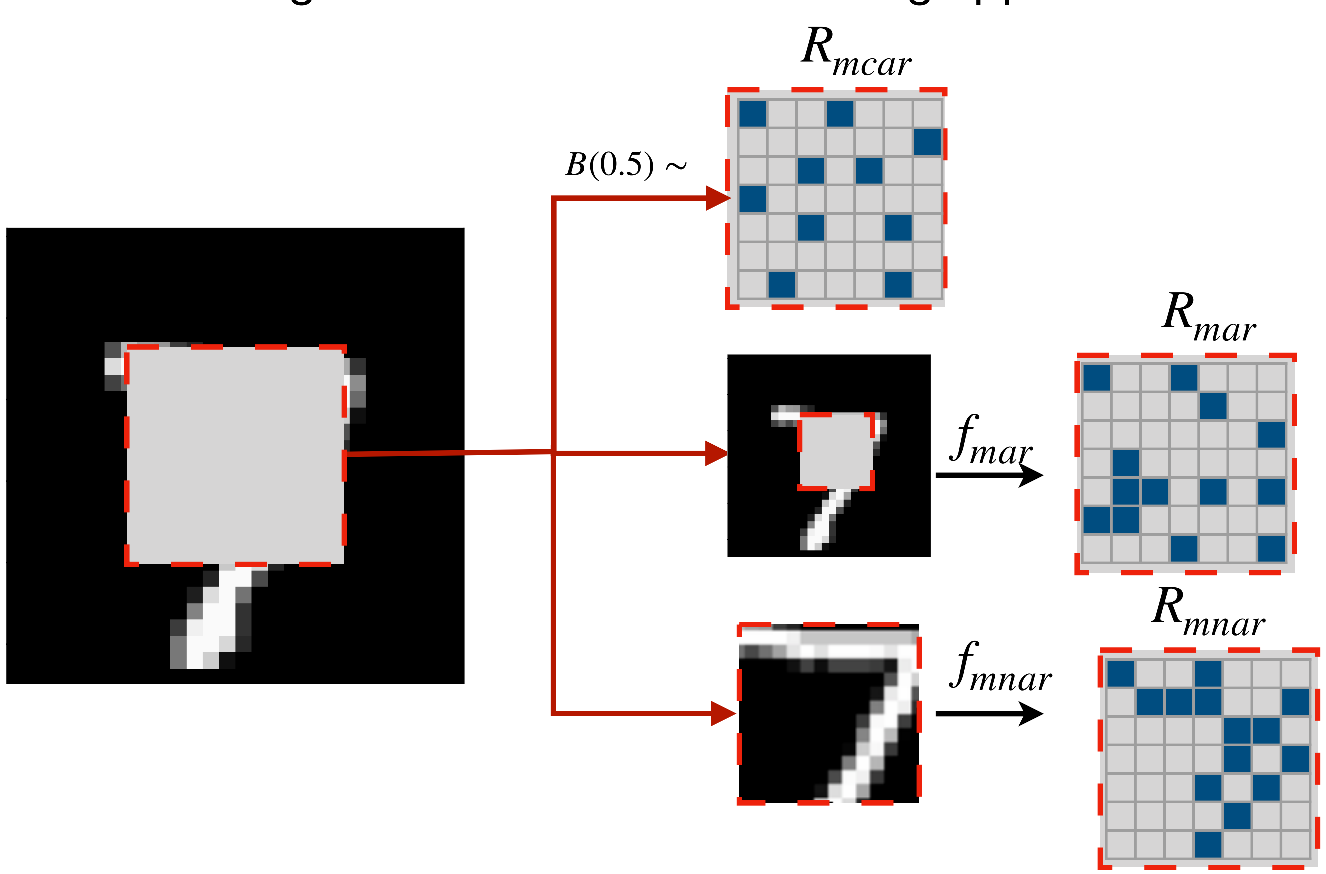


Table 1. MI-MCAR Empirical rejection rate with different numbers of features on heterogeneous data (binary and continuous)

f	MCAR	MAR	MNAR
10	0.02	1.00	0.98
50	0.04	1.00	1.00
100	0.02	1.00	1.00

Table 2. MI-US empirical rejection rate under different missingness simulations with different number of features

f	MCAR	MAR	MNAR
10	0.02	0.06	0.87
50	0.05	0.03	0.96
100	0.03	0.02	0.94

MNIST semi-synthetic p-value distribution