欠測データに対する種々の多重補完法の性能比較

M2021SS004 田中一州

指導教員：松田眞一

1　はじめに

経時的測定データでは, 同じ対象に対して繰り返しデータを測定することで, 研究開始時点で想定していたサンプルサイズより小さくなってしまうということがある. この現象は欠測が発生したことが原因である. 欠測が存在するデータを解析する場合の問題点は, サンプルサイズが小さくなることで解析精度が低下してしまうことである. 本研究で経時的測定データを統計解析するときの問題点を解決したいと考え, 欠測データへのアプローチ方法について研究したいと考えた.

近年, 深層学習モデルに基づく欠測データ補完法が開発され, 小規模な研究において有望な結果が得られている.

2　先行研究

Wang *et al.*[5] において, 4 つの機械学習ベースの多重補完法の反復抽出特性を比較するために, 米国地域社会調査 (ACS) のサブサンプルに基づくシミュレーション研究を行なっている. 4 つの補完法のうち深層学習を用いる 2 つの補完法は, 敵対的生成補完ネットワーク (GAIN: Generative Adversarial Imputation Network), ノイズ除去自己符号化器による多重代入法 (MIDA: Multiple Imputation using Denoising Autoencoders) である. 先行研究では, 実践している技術者はハイパーパラメータ値に頼る可能性が高いということから, 機械学習ベースの補完法のデフォルトを採択し, 微調整をしているが, 機械学習を用いる補完法の性能はハイパーパラメータの選択に大きく依存していると分かった.

3　シミュレーション

本研究では, 先行研究の 2 つの補完法 (GAIN, MIDA) を取り上げ, 先行研究での結果と応答曲面法を用いて最適なハイパーパラメータを選び, シミュレーションを行なって得られた結果を比較する. また, 最適なハイパーパラメータに設定した GAIN と MIDA の 2 つの補完法と従来の補完法であるフラクショナルホットデック補完法 (FHDI: Fractional Hot Deck Imputation) を比較する. 先行研究では, Python を用いてシミュレーションを行なっているのだが, 本研究では深層学習を用いる補完法を Python で, 従来の補完法を統計ソフト R でシミュレーションを行ない, 補完法の評価を Python で行ない, 比較検討をする.

3.1　ハイパーパラメータ

本研究では, GAIN, MIDA で用いる損失関数の値を目標変数として, D 最適基準の応答曲面法 (RSM) を用いて, 最適なハイパーパラメータを選定した. (金子 [2] 参照)

GAIN では, バッチサイズ (`batch_size`), イテレーション数 (`iteration`) は固定して, 説明変数をヒント率 (`hint_rate`), 式 (1) の β (`alpha`) とする. (先行研究のプログラムコードが `alpha` だったためそのまま使う.) MIDA では, バッチサイズ (`batch_size`), 第 1 相, 微調整相のエポック数 (`num_steps_phase1`, `num_steps_phase2`) は固定して, 説明変数を学習率 (`learning_rate`), 隠れ層 θ (`theta`) とする. その結果を表 1, 2 で示す.

表 1　GAIN のハイパーパラメータ

	先行研究		RSM	
	MCAR	MAR	MCAR	MAR
`batch_size`	512	512	512	512
`iteration`	200	200	200	200
`hint_rate`	0.3	0.13	0.1	0.1
`alpha`	100	100	60	60

表 2　MIDA のハイパーパラメータ

	先行研究		RSM	
	MCAR	MAR	MCAR	MAR
`num_steps_phase1`	100	100	100	100
`num_steps_phase2`	2	2	2	2
`batch_size`	512	512	512	512
`learning_rate`	0.001	0.001	0.01	0.01
`theta`	7	7	10	10

ここで, n 個のユニットがあり, それぞれが p 個の変数を持っている標本 $\mathbf{Y}$ を考える. ユニット i の変数 j の値を Y_{ij} $(i = 1, \cdots, n,\ j = 1, \cdots, p)$ とする. また, $\mathbf{M}$ は $\mathbf{Y}$ の観測を示す $M_{ij} \in \{0, 1\}$ のマスク行列である.

3.1.1　GAIN の損失関数

生成器 Generator の損失関数は

$$L(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}, \hat{\mathbf{M}}) = L_G(\mathbf{M}, \hat{\mathbf{M}}) + \beta L_M(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}) \tag{1}$$

である. ここで, 生成器損失 $L_G(\mathbf{M}, \hat{\mathbf{M}})$ は, 識別器 Discriminator が補完値を誤って観測値として識別した場合に最小化され, 再構成損失 $L_M(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M})$ は, 予測値が観測値に近い場合に最小となり, ハイパーパラメータ β で重み付けされる.

3.1.2　MIDA の損失関数

Lu *et al.*[3] に従い, MIDA を 2 つの相である第一相と微調整相で学習する. 第一相では, 最初に補完されたデータを MIDA に与え, N_{prime} 個のエポックについて学習する. 微調整相では, MIDA は第一相での出力に対して N_{tune} 個

のエポックについて学習を行い, 結果を生成する. 以下の損失関数は 2 つの相で使用されている.

$$L\left(Y_{ij_0}, \hat{Y}_{ij}, M_{ij}\right) = \begin{cases} (1-M_{ij})(Y_{ij_0}-\hat{Y}_{ij})^2 & \text{if } Y_{ij} \text{ is continuous} \\ -(1-M_{ij})Y_{ij_0}\log\hat{Y}_{ij} & \text{if } Y_{ij} \text{ is categorical} \end{cases}$$

ここで, 連続変数については平均値, カテゴリ変数については最頻値のラベルを用いて欠測値に対する最初の補完を行い, 完成した最初の補完データ $\mathbf{Y_0}$ のユニット i の変数 j の値を Y_{ij_0} とする.

3.2 データ

先行研究でも使用されている 2018 年の ACS の 1 年間の Public Use Microdata Sample を使用する. 2018 年の ACS データには, 世帯水準の変数 (持ち家か賃貸かなど) と個体水準の変数 (各世帯内の個人の年齢, 所得, 性別など) の両方が含まれている. データを使いやすくするために加工した結果, 1,257,501 個のユニットで, 18 個のバイナリ変数, 20 個の 3 から 9 の水準のカテゴリ変数, および 8 個の連続変数が含まれている.

3.3 ビン化連続変数

先行研究は, ACS では離散変数が一般的であるため, バイナリ変数とカテゴリ変数の周辺確率に注目している. 例えば, K 個のカテゴリを持つカテゴリ変数は $K-1$ 個の推定値を持つ. 補完法が多変量分布特性をどの程度保持しているかを評価するために, Akande *et al.*[1] と同様に, バイナリ変数とカテゴリ変数におけるカテゴリのすべての二者択一の組み合わせの二変量確率も考慮している. カテゴリ変数と連続変数の結果を有意義に比較するために, 各連続変数を標本の分位数に基づいて K 個のカテゴリに離散化すること (これをビン化と呼ぶ) を提案している. そして, これらのビン化連続変数を, 前述の周辺確率と二変量確率の推定値に基づいて, カテゴリ変数として評価している.

3.4 補完法の評価基準

先行研究と同様に, 欠測メカニズム（Little and Rubin [4] 参照）に従って完全なデータセットから欠測値を作成し, 補完法によって欠測値を補完する. 次に, Rubin's MI combination rules を用いて各推定値の点推定値と区間推定値を構築し, 3 つの基準に基づいてこれらの補完値と元の「真の」値とを比較する.

3.4.1 Rubin's MI combination rules

母集団における目標推定値を Q とし, $q^{(l)}$ と $u^{(l)}$ をそれぞれ l 番目の補完データセットに基づく Q の点推定値と分散推定値とする. Q の多重補完 (MI) 点推定値は

$$\bar{q}_L = \frac{1}{L}\sum_{l=1}^{L} q^{(l)}$$

であり, それに対応する分散の推定値は

$$T_L = \left(1+\frac{1}{L}\right)b_L + \bar{u}_L$$

に等しく, ここで代入間分散と代入内分散は

$$b_L = \frac{1}{L-1}\sum_{l=1}^{L}\left(q^{(l)}-\bar{q}_L\right)^2,$$

$$\bar{u}_L = \frac{1}{L}\sum_{l=1}^{L} u^{(l)}$$

である. Q の信頼区間は $(\bar{q}_L - Q)/\sqrt{T_L} \sim t_\nu$ を用いて構築され, t_ν は自由度

$$\nu = (L-1)\left(1+\frac{\bar{u}_L}{\left(1+\frac{1}{L}\right)b_L}\right)^2$$

を持つ t 分布である.

以下では, h 回目のシミュレーションにおける Q の MI 点推定値として $\bar{q}_L^{(h)}$ を用いる.

3.4.2 the Absolute Standardized Bias

1 つ目の指標は, バイアスに着目したものである. カテゴリ変数の確率に多く見られるゼロに近い推定値に対応するため, Q の各推定値の絶対標準化バイアス（ASB : the Absolute Standardized Bias）を考慮する.

$$ASB = \frac{1}{H}\sum_{h=1}^{H}\frac{|\bar{q}_L^{(h)} - Q|}{Q} \tag{2}$$

ASB は比較する補完法と値を比べたときに値が小さいほど補完がうまくできているとわかる.

3.4.3 the Relative Mean Squared Error

2 つ目の指標は, 相対平均二乗誤差（Rel.MSE : the Relative Mean Squared Error）であり, これは補完されたデータから Q を推定する際の MSE と欠測データ導入前のサンプリングデータから Q を推定する際の MSE との比である.

$$Rel.MSE = \frac{\sum_{h=1}^{H}(\bar{q}_L^{(h)} - Q)^2}{\sum_{h=1}^{H}(\tilde{Q}^{(h)} - Q)^2} \tag{3}$$

ここで, $\tilde{Q}^{(h)}$ は Q のプロトタイプ推定値, すなわち h 回目のシミュレーションのサブ標本の完全データからの点推定値である.

Rel.MSE は 1 に近いほど誤差が少なく上手く補完ができたことが分かる.

3.4.4 Coverage Rate

3 つ目の指標は, 被覆率 (Coverage Rate) で, $CI_h^\alpha (h = 1, \cdots, H)$ で示される $100\alpha\%$ (例えば 95%) 信頼区間が, H 回のシミュレーションの中で, 真の Q を含む割合である. 信頼区間は 3.4.1 項の分布に基づいて構成される.

$$Coverage = \frac{1}{H}\sum_{h=1}^{H}\mathbf{1}\left\{Q \in CI_h^\alpha\right\} \tag{4}$$

4　先行研究 vs. RSM

サンプルサイズ $n = 10000$, 補完回数 $L = 10$, シミュレーション回数 $H = 100$, 3.1 節で示したハイパーパラメータを用いて, シナリオごとに実行した結果は以下のようになった. このとき, 先行研究と RSM で用いた GAIN, MIDA を使い分けるために, 先行研究の方の結果を GAINp, MIDAp としている.

4.1　MCAR シナリオ (vs. 先行研究)

GAINp と比較すると, ASB のビン化連続変数のみ良い結果となった. また ASB では, カテゴリ変数の周辺確率は 5%, 二変量確率は 10% 付近までに 8 割ほどだが, ビン化連続変数の周辺確率は 30%, 二変量確率は 50% 付近で 8 割ほどあり, 上手く補完できていないと考える. Rel.MSE から GAIN は他と比べて補完がうまくいっていないことが確認された.

MIDAp と比較すると, MIDA では被覆率は改善できたが ASB や Rel.MSE から若干 MIDAp の方が良い結果となった. そのため, 差はほとんどないと考える. MIDA は 3 つの指標の結果から可能な補完値を満遍なく補完しているように感じる.

表 3　ABS, Rel.MSE の分布 (vs. 先行研究, MCAR)

		周辺確率				二変量確率			
		GAINp	MIDAp	GAIN	MIDA	GAINp	MIDAp	GAIN	MIDA
	分位数	ASB(×100)							
	10%	0.29	1.01	0.32	0.80	0.66	1.84	0.88	1.73
	25%	0.87	3.17	0.73	2.27	1.70	4.88	2.25	4.69
カテゴリ	50%	1.80	5.16	1.82	5.20	3.97	11.59	5.48	11.09
変数	75%	3.36	11.54	5.03	11.18	9.45	20.65	11.88	21.38
	90%	10.05	18.90	10.07	22.14	18.57	27.71	20.98	27.79
	10%	10.88	5.40	8.61	6.36	12.07	5.18	6.71	5.02
	25%	27.57	8.59	14.48	9.59	28.28	12.56	18.23	12.62
ビン化連続	50%	29.90	21.83	28.23	23.75	45.20	25.21	31.81	26.64
変数	75%	30.89	29.85	29.97	29.93	51.22	42.84	50.18	44.46
	90%	81.64	42.79	54.89	43.58	108.58	68.18	88.92	73.96
	分位数	Rel.MSE							
	10%	1.67	1.81	7.92	2.30	2.68	1.79	6.42	1.94
	25%	4.15	4.82	12.76	4.49	5.15	3.30	11.62	3.49
カテゴリ	50%	16.86	16.30	31.21	15.85	11.20	7.74	27.54	7.88
変数	75%	47.79	40.12	96.35	40.19	24.32	17.16	73.22	19.08
	90%	85.19	72.07	335.54	95.65	47.53	38.65	183.21	42.41
	10%	93.04	5.88	93.03	9.46	17.04	1.53	20.81	1.68
	25%	125.14	12.71	131.82	14.56	33.11	4.15	41.00	4.42
ビン化連続	50%	215.26	87.41	344.57	96.10	58.77	13.86	82.64	15.54
変数	75%	644.98	190.43	883.94	188.05	144.14	50.43	223.16	54.09
	90%	1376.70	466.32	1251.38	459.14	541.76	141.65	591.04	148.22

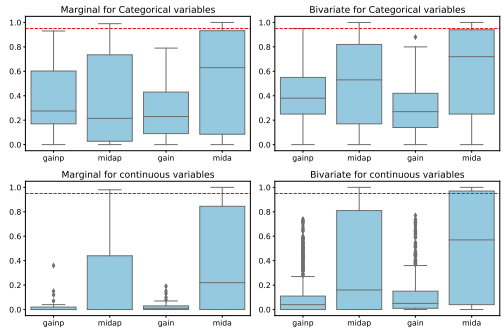

図 1　被覆率 (vs. 先行研究, MCAR)

4.2　MAR シナリオ (vs. 先行研究)

GAINp と比較すると, GAIN では少しだが良い結果となった. MCAR シナリオに比べてビン化連続変数の ASB が全体的に小さくなった.

MIDAp と比較すると, MIDA は 3 つの指標ともに良い結果となった. MCAR シナリオに比べて二変量確率に対する指標が全体的に小さくなり, 精度が良くなる結果となった.

表 4　ABS, Rel.MSE の分布 (vs. 先行研究, MAR)

		周辺確率				二変量確率			
		GAINp	MIDAp	GAIN	MIDA	GAINp	MIDAp	GAIN	MIDA
	分位数	ASB(×100)							
	10%	0.07	0.84	0.06	0.14	0.39	1.30	0.33	0.64
	25%	0.38	1.84	0.30	0.51	1.13	3.38	0.94	1.93
カテゴリ	50%	1.29	3.33	1.11	2.06	2.62	7.76	2.23	5.20
変数	75%	2.85	8.45	2.33	6.14	6.37	15.00	5.45	11.47
	90%	6.32	13.02	5.04	11.50	15.07	23.03	13.69	22.35
	10%	0.36	2.83	0.36	0.29	4.96	3.05	5.63	1.61
	25%	7.77	4.14	7.11	1.96	10.41	7.96	10.48	5.36
ビン化連続	50%	11.23	12.91	11.23	9.00	15.45	16.78	15.81	10.70
変数	75%	13.59	19.09	13.59	11.51	21.91	27.83	21.92	17.00
	90%	31.50	27.50	29.61	16.40	36.22	43.04	36.10	27.18
	分位数	Rel.MSE							
	10%	1.00	1.70	1.00	1.00	1.20	1.63	1.17	1.10
	25%	1.45	3.87	1.39	1.41	1.90	2.66	1.92	1.50
カテゴリ	50%	4.48	10.67	4.03	3.84	5.05	5.47	5.20	2.75
変数	75%	23.93	22.39	24.79	12.08	20.71	11.67	23.26	6.86
	90%	179.24	51.79	208.88	29.75	80.57	26.16	97.33	17.79
	10%	1.00	3.54	1.00	1.00	2.62	1.36	2.37	1.09
	25%	14.84	6.43	15.85	2.50	4.30	2.84	3.92	1.57
ビン化連続	50%	24.90	40.37	24.35	12.81	8.58	7.94	7.88	3.60
変数	75%	69.15	98.64	64.12	35.05	18.98	26.31	18.05	9.17
	90%	241.66	229.77	233.51	63.19	50.84	73.11	47.81	24.19

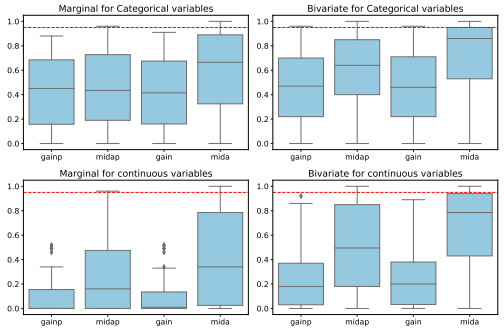

図 2　被覆率 (vs. 先行研究, MAR)

5　FHDI vs. RSM

このシミュレーションを行なうにあたって, FHDI は欠測が 1 つもないユニットをドナーにするために, ここで扱う欠測データセットは欠測値のない完全ユニットが入っていることが前提となるため, 欠測データセットに完全データセットを結合して新たな欠測データセットとした. また, 4 章と同じサンプルサイズで行なうと, FHDI では 1 つの補完データセットを作成するのに 12 時間以上かかってしまうため, サンプルサイズが小さいサブ標本を作成する. サンプルサイズ $n = 1000$ (完全ユニット 700, 欠測ユニット 300), 補完回数 $L = 10$, シミュレーション回数 $H = 50$, 3.1 節で示したハイパーパラメータを用いて, シナリオごとに実行した結果は以下のようになった.

5.1 MCAR シナリオ (vs. FHDI)

MCAR シナリオでは, カテゴリ変数とビン化連続変数で ASB が最小になるものが違った. カテゴリ変数では GAIN, ビン化連続変数では FHDI が一番良いモデルとなった. 被覆率では FHDI と MIDA の中央値がどこでも 9 割付近にある.

表 5 ABS, Rel.MSE の分布 (vs. FHDI, MCAR)

		周辺確率			二変量確率		
		FHDI	GAIN	MIDA	FHDI	GAIN	MIDA
	分位数	ASB(×100)					
	10%	0.26	0.13	0.34	0.55	0.29	0.53
	25%	0.79	0.30	0.79	1.39	0.76	1.29
カテゴリ	50%	1.58	0.73	1.47	3.43	1.74	2.95
変数	75%	3.93	1.59	3.62	6.85	3.63	5.80
	90%	6.64	2.44	5.81	12.08	6.61	9.14
	10%	0.60	0.90	0.89	0.69	1.27	1.47
	25%	1.39	1.56	2.30	1.86	3.36	3.53
ビン化連続	50%	2.20	3.89	5.67	3.55	6.87	7.40
変数	75%	2.97	7.37	8.39	5.73	12.00	13.24
	90%	4.41	9.70	11.83	8.74	25.61	19.95
	分位数	Rel. MSE					
	10%	1.01	1.05	1.01	0.95	1.01	0.97
	25%	1.09	1.11	1.08	1.04	1.09	1.04
カテゴリ	50%	1.22	1.20	1.23	1.17	1.20	1.14
変数	75%	1.41	1.32	1.41	1.33	1.33	1.28
	90%	1.74	1.44	1.73	1.56	1.47	1.48
	10%	0.91	0.97	1.00	0.89	0.94	0.88
	25%	1.08	1.22	1.16	0.97	1.07	0.97
ビン化連続	50%	1.16	1.63	1.84	1.08	1.28	1.17
変数	75%	1.33	3.42	3.41	1.20	1.94	1.76
	90%	1.72	10.53	6.82	1.34	3.48	2.97

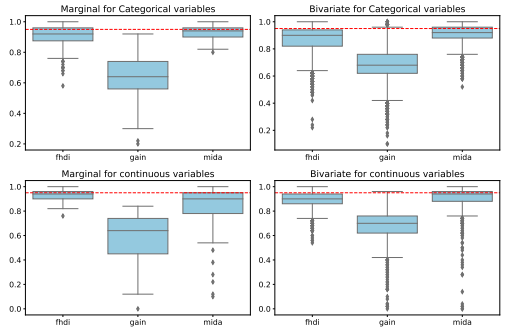

図 3 被覆率 (vs. FHDI, MCAR)

5.2 MAR シナリオ (vs. FHDI)

MAR シナリオでは, ASB に対してカテゴリ変数では GAIN, ビン化連続変数では FHDI が良いモデルとなった. MCAR シナリオとは異なり, 欠測のない変数がある場合は Rel.MSE に対してカテゴリ変数では MIDA, ビン化連続変数では FHDI が良いモデルとなった.

6 おわりに

本研究を通して, 深層学習を用いる補完法は欠測率や完全ユニットの数に影響されやすいことが分かった. また, 1 つ目のシミュレーションの結果から, 応答曲面法を用いた最適なパラメータ選定でも不十分に感じた.

表 6 ABS, Rel.MSE の分布 (vs. FHDI, MAR)

		周辺確率			二変量確率		
		FHDI	GAIN	MIDA	FHDI	GAIN	MIDA
	分位数	ASB(×100)					
	10%	0.21	0.21	0.16	0.38	0.37	0.39
	25%	0.43	0.38	0.45	0.98	0.92	1.02
カテゴリ	50%	1.06	0.99	1.14	2.49	2.22	2.46
変数	75%	2.66	2.49	2.68	6.15	5.08	5.35
	90%	6.26	4.66	5.30	12.04	9.85	9.72
	10%	0.23	0.53	0.36	0.46	0.78	0.89
	25%	0.66	1.13	1.05	1.20	1.86	2.24
ビン化連続	50%	1.20	1.98	2.79	2.53	3.91	4.53
変数	75%	1.90	3.08	4.81	4.41	7.37	7.89
	90%	2.86	5.96	7.64	6.48	16.12	12.50
	分位数	Rel. MSE					
	10%	1.00	1.00	1.00	0.95	0.99	0.97
	25%	1.02	1.01	1.00	1.01	1.03	1.01
カテゴリ	50%	1.16	1.17	1.15	1.12	1.14	1.09
変数	75%	1.42	1.44	1.29	1.30	1.32	1.21
	90%	1.83	1.78	1.46	1.67	1.55	1.37
	10%	0.99	1.00	1.00	0.91	0.92	0.92
	25%	1.00	1.05	1.04	0.97	1.00	0.99
ビン化連続	50%	1.05	1.22	1.31	1.03	1.11	1.08
変数	75%	1.13	1.80	1.91	1.11	1.37	1.30
	90%	1.23	3.70	2.66	1.19	2.29	1.88

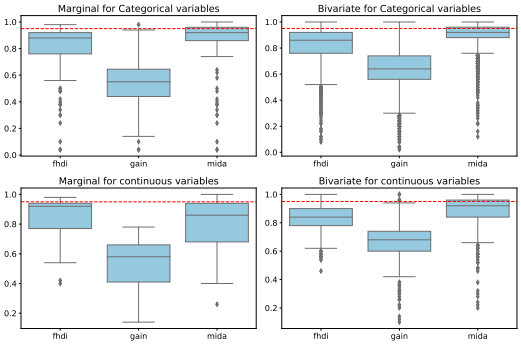

図 4 被覆率 (vs. FHDI, MAR)

参考文献

[1] Akande, O., Li, F., and Reiter, J., An empirical comparison of multiple imputation methods for categorical data, *The American Statistician*, **71** (2), 162−170, 2017.

[2] 金子 弘昌, 『Python で気軽に化学 · 化学工学』, 丸善出版, 2021.

[3] Lu, H.-m., Perrone, G., and Unpingco, J., Multiple imputation with denoising autoencoder using metamorphic truth and imputation feedback, *arXiv* preprint arXiv:2002.08338, 2020.

[4] Little, R. J., and Rubin, D. B., Statistical analysis with missing data, Hoboken, NJ, John Wiley & Sons., 2014.

[5] Wang, Z., Akande, O., Poulos, J., and Li, F., 2021, Are deep learning models superior for missing data imputation in large surveys? Evidence from an empirical comparison, https://arxiv.org/abs/2103.09316 (2022/09 閲覧)