The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.1.029
31(1), 29–40

Imputation method for missing data based on clustering and measure of property

Sunghyun Kim[a] · Dongjae Kim[a,1]

[a]Department of Biomedicine · Health Science, The Catholic University of Korea

(Received August 22, 2017; Revised October 23, 2017; Accepted December 15, 2017)

Abstract

There are various reasons for missing values when collecting data. Missing values have some influence on the analysis and results; consequently, various methods of processing missing values have been studied to solve the problem. It is thought that the later point of view may be affected by the initial time point value in the repeated measurement data. However, in the existing method, there was no method for the imputation of missing values using this concept. Therefore, we proposed a new missing value imputation method in this study using clustering in initial time point of the repeated measurement data and the measure of property proposed by Kim and Kim (*The Korean Communications in Statistics*, **30**, 463–473, 2017). We also applied the Monte Carlo simulations to compare the performance of the established method and suggested methods in repeated measurement data.

Keywords: imputation of missing value, clustering, measure of property, initial time point

1. 서론

빅데이터 및 정보화 시대에 맞춰 데이터도 여러 분야에서 점점 개인에게 구체적으로 초점이 맞춰지고 있다. 하지만 데이터에 담겨야 할 정보가 많아지므로 결측치도 자연스럽게 증가할 수밖에 없게 되어 결측치 문제를 해결하기 위한 방안이 필요하다. 결측치는 분석을 위한 관측이나 실험과 같은 자료 생성 과정에서 관측치를 얻지 못하는 경우를 말한다. 결측치의 발생에는 다양한 이유가 존재하는데 예를 들어, 기계의 오류로 인한 결측의 발생하는 경우, 임상시험에서 피험자 본인의 의사에 의해 임상시험 참여를 중단하여 결측이 발생하는 경우, 설문조사에서 무응답 하는 경우 등이 있다. 결측치의 발생은 분석의 어려움과 편의를 발생시켜 결과에 영향을 미치므로 결측치의 처리는 분석에 앞서 해결해야 할 중요한 과제이다 (Kang, 2013). 결측치를 처리하는 기존 방법 중 하나인 완전사례분석(complete case analysis)은 결측이 없는 완벽한 자료만을 이용하여 분석하는 것이므로 검정력의 손실과 편의가 발생할 수 있는 단점이 있다. 또 다른 결측치 대체 방법으로는 평균 대체법(mean imputation), 핫덱 대체법(Hot-Deck imputation), Kim과 Kim (2017)이 제안한 특성도 대체법(measure of property imputation) 등이 있다. 평균 대체법은 측정된 자료들의 평균으로 대체하는 방법으로 결측값이 반복적으로 평균값으로 대체되므로 통계량의 표준오차가 과소 추정되는 문제가 있다 (Kim과 Kim, 2017). 핫덱 대체법은 결측

[1]Corresponding author: Department of Biomedicine · Health Science, The Catholic University of Korea, 222 Banpo-dero Seocho-gu, Seoul 06591, Korea. E-mail: djkim@catholic.ac.kr

Table 2.1. Structure of repeated measure data

N	Time			
	1	2	$\cdots$	t
1	y_{11}	y_{12}	$\cdots$	y_{1t}
2	y_{21}	y_{22}	$\cdots$	y_{2t}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_{n1}	y_{n2}	$\cdots$	y_{nt}

이 발생한 변수에 대해 그 변수가 가질 수 있는 값들 중 임의로 하나를 선택하여 대체하는 방법으로 표준오차를 구하기 어렵다는 문제가 있다 (Kim과 Kim, 2017). 특성도 대체법은 개체의 특성을 나타내는 특성도를 이용하여 비슷한 특성을 갖는 개체의 관측값을 이용하여 결측치를 대체하는 방법이다 (Kim과 Kim, 2017). 하지만, 일반적으로 임상시험에 있어 초기 시점의 관측값에 따라 개체의 특성이 다를 것으로 예상된다. 예를 들어, 고혈압 환자를 대상으로 실시하는 임상시험의 경우 초기 혈압이 200이상인 대상자가 200보다 낮은 대상자에 비해 치료제로 인한 감소량이 더 크며, 이 경우 특성도가 개체의 특성을 충분히 설명하는데 한계가 존재할 것이다.

군집분석이란 각 관측치들 간의 유사성을 측정하여 유사성이 높은 관측치들끼리 묶어 군집을 나누는 통계적 분석방법으로 크게 계층적 군집화 방법(hierarchical clustering method)과 비계층적 군집화 방법(nonhierarchical clustering method)으로 나눌 수 있다 (Shin, 2010). 계층적 군집화 방법이란, 각각 관측치를 하나의 군집으로 놓거나 전체 관측치를 하나의 군집으로 놓고, 어떤 기준에 따라서 그 군집들을 묶거나 나누어 가는 방법으로 단일 연결법, 최장 연결법, 평균 연결법, Ward 연결법 등이 있다 (Shin, 2010). 비계층적 군집 방법은 이전 단계에서 집단화된 군집을 분해하여 분해된 것과 더불어 새로운 군집을 형성하는 방법으로 K-평균 알고리즘 등과 같은 방법이 있다.

본 연구에서는 군집분석을 이용하여 집단을 나누고, 군집화된 집단별로 Kim과 Kim (2017)이 제안한 특성도를 이용한 결측치 대체 방법을 제안하였다. 이는 초기 시점의 값에 따른 변화량의 크기의 차이로 인한 문제를 군집화를 통해 해결함으로써 개체의 특성을 더욱 잘 보일 것으로 예상된다. 또한 각 집단이 특성도를 이용한 결측치 대체 방법을 이용하기에 충분한 표본수를 유지하기 위해 집단수를 2개로 제한할 것이며, 계층화 군집화 방법 중 자료를 군집화할 때 군집 간 정보 손실을 최소화하는 Ward (1963)의 군집 방법을 이용하여 초기 시점을 군집화하였다. 3장에서는 반복 측정 자료 중 intermittent missing 상황에서의 모의실험을 통하여 기존 방법들과 제안하는 방법의 결측치 대체의 성능을 비교하였다.

2. 제안하는 방법

2.1. 자료 형태

연구에서 다루는 자료의 형태는 y_{ij} $(i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, t)$가 n개의 개체와 t개의 반복수를 갖는 반복 측정 자료이다 (Table 2.1).

$$y_{ij} = \mu + \tau_j + \epsilon_{ij}, \quad (i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, t),$$

여기서 μ는 전체 평균, τ_j는 j번째 시점의 효과, ϵ_{ij}는 오차항을 나타낸다.

2.2. 군집화

Ward 연결법은 각 개체를 하나의 군집으로 간주함을 시작으로 군집들을 묶어 단계적으로 그 수를 하나가 될 때까지 줄여나가는 방법으로 각 군집의 평균 벡터로부터 유클리디안 거리의 제곱합을 이용하였다

Table 2.2. Measure of property of repeated measure data after clustered

Cluster	N	Time				x_{hi}	a_{hi}	m_{hi}	p_{hi}
		1	2	$\cdots$	t				
1	1	y_{111}	y_{112}	$\cdots$	y_{11t}	x_{11}	a_{11}	m_{11}	p_{11}
	2	y_{121}	y_{122}	$\cdots$	y_{12t}	x_{12}	a_{12}	m_{12}	p_{12}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n_1	y_{1n_11}	y_{1n_12}	$\cdots$	y_{1n_1t}	x_{1n_1}	a_{1n_1}	m_{1n_1}	p_{1n_1}
	mean	c_{11}	c_{12}	$\cdots$	c_{1t}				
2	1	y_{211}	y_{212}	$\cdots$	y_{21t}	x_{21}	a_{21}	m_{21}	p_{21}
	2	y_{221}	y_{222}	$\cdots$	y_{22t}	x_{22}	a_{22}	m_{22}	p_{22}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n_2	y_{2n_21}	y_{2n_22}	$\cdots$	y_{2n_2t}	x_{2n_2}	a_{2n_2}	m_{2n_2}	p_{2n_2}
	mean	c_{21}	c_{22}	$\cdots$	c_{2t}				

(Jeon, 2012). h번째 군집에서 j번째 시점에 대한 군집 평균을 $\bar{y}_{hj}$라 할 때, h번째 군집에서 편차제곱합(error sum of squares; ESS)은

$$\mathrm{ESS}_h = \sum_{i=1}^{n_h}\sum_{j=1}^{2}(y_{hij} - \bar{y}_{hj})^2, \quad (h = 1, 2) \tag{2.1}$$

로 나타낼 수 있고 총 편차제곱합은

$$\mathrm{ESS} = \sum_{h=1}^{2}\mathrm{ESS}_h \tag{2.2}$$

와 같이 정의할 수 있다 (Choi와 Jeong, 2003). 군집화 과정의 첫 단계인 군집당 표본 수가 1일 때는 ESS_h가 0의 값을 가지나 군집을 만들어 감에 따라 ESS는 증가하게 된다 (Choi와 Jeong, 2003). 다시 말해, 모든 가능한 군집 쌍에 대해 두 군집의 병합으로 인한 ESS의 증가가 최소가 되도록 군집을 병합해 나가는 방법을 Ward 연결법이라 한다 (Choi와 Jeong, 2003). 시점 1과 시점 2 변수를 대상으로 한 Ward 연결법 군집분석을 실시하여 $n\,(= n_1 + n_2)$개의 개체들을 두 개의 군집으로 나눈다 (Table 2.2).

2.3. 특성도

초기 두 시점을 이용해 Ward 연결법으로 두 개의 군집으로 구분한 반복 측정 자료에서 군집별로 Kim과 Kim (2017)가 정의한 일치 지수, 유지 지수, 특성도를 이용한다.

2.3.1. 자료의 이분화 각 군집의 시점별 관측치의 평균을 c_{hj}라 하고

$$c_{hj} = \frac{1}{n}\cdot\sum_{i=1}^{n_h} y_{hij}, \quad (h = 1, 2,\ i = 1, 2, \ldots, n_h,\ j = 1, 2, \ldots, t) \tag{2.3}$$

와 같이 나타낼 때, c_{hj}를 절단값으로 이용해 관측치를 이분화한다. 여기서 y_{hij}는 h번째 군집, i번째 개체, j번째 시점의 관측치이다. 이분화된 관측치는 x_{hij}라 하고

$$x_{hij} = \begin{cases} 1, & \text{if } y_{hij} \geq c_{hj}, \\ 0, & \text{if } y_{hij} < c_{hj}, \end{cases} \quad (h = 1, 2,\ i = 1, 2, \ldots, n_h,\ j = 1, 2, \ldots, t) \tag{2.4}$$

로 정의하였다. 이때, h번째 군집의 i번째 개체의 이분된 관측치들 중 값이 1인 관측치의 수인 x_{hi}를

$$x_{hi} = \sum_{j=1}^{t} x_{hij}, \quad (h = 1, 2,\ i = 1, 2, \ldots, n_h) \tag{2.5}$$

라 하면, 값이 0인 이분된 관측치의 수는 $t - x_{hi}$가 된다 (Kim과 Kim, 2017).

2.3.2. 일치 지수 각 개체의 이분된 관측치 중 1의 개수인 x_{hi}를 가지고 Kim과 Kim (2017)이 제안한 일치 지수를 군집별로 적용하였다. 여기서 h번째 군집, i번째 개체의 불일치 지수(index of disagreement; d_{hi})를

$$d_{hi} = \frac{2}{t^2} x_{hi}(t - x_{hi}), \quad (h = 1, 2,\ i = 1, 2, \ldots, n_h). \tag{2.6}$$

일치 지수(index of agreement; a_{hi})를

$$a_{hi} = 1 - d_{hi}, \quad (h = 1, 2,\ i = 1, 2, \ldots, n_h) \tag{2.7}$$

와 같이 정의한다. 이때 $x_{hi}(t - x_{hi})$의 최댓값이 $t^2/4$이므로 d_{hi}는 0과 0.5사이의 값을 가진다. 결론적으로 a_{hi}는 t의 값에 상관없이 0.5와 1사이의 값을 가지고, 일치도가 낮을수록 0.5에 가까워지며, 일치도가 높을수록 1에 가까운 값을 가진다 (Kim과 Kim, 2017).

2.3.3. 유지 지수 Kim과 Kim (2017)은 개체의 시점에 따른 전반적인 추세 및 유지도를 나타내는 유지 지수를 정의하였다. 여기에 이 유지 지수를 군집별로 나누어 다음과 같이 유지 지수(index of maintenance; m_{hi})를 재정의하였다.

$$m_{hi} = \frac{\sum_{j=1}^{t-1} I(x_{hij+1},\ x_{hij}) \cdot j}{\binom{t}{2}}, \quad (i = 1, 2, \ldots, n_h),\ I(a, b) = \begin{cases} 1, & \text{if } a = b = 1, \\ -1, & \text{if } a = b = 0, \\ 0, & \text{if } a \neq b. \end{cases} \tag{2.8}$$

이때 유지 지수 m_{hi}는 -1과 1사이의 값을 갖는다 (Kim과 Kim, 2017).

2.3.4. 특성도 각 개체의 시점에 따른 방향성과 연속성을 나타내는 유지 지수와 절단값을 기준으로 일치성을 나타내는 일치 지수가 각 개체의 특성을 설명하고 있으므로 두 지수를 곱해준 것을 특성도라 Kim과 Kim (2017)이 정의하였다. 이를 군집별로 나누어 계산해 특성도(measure of property; p_{hi})를 다음과 같이 재정의하였다.

$$p_{hi} = a_{hi} \cdot m_{hi} = \left\{1 - \frac{2}{t^2} x_{hi}(t - x_{hi})\right\} \frac{\sum_{j=1}^{t-1} I(x_{hij+1}, x_{hij}) \cdot j}{\binom{t}{2}}, \quad (i = 1, 2, \ldots, n_h). \tag{2.9}$$

기존 자료를 두 개의 군집으로 군집화하고, 군집별로 자료의 시점별 절단값부터 시점별로 이분화한 자료를 개체별로 합한 값, 일치 지수, 유지 지수, 특성도까지 새로 정의한 값들을 포함한 자료는 다음과 같다.

Table 3.1. Basic Statistics about original data

Data	Mean	STD	CV	p-value
A1	18.33	1.52	0.083	0.0090
A2	19.94	0.99	0.050	0.0090
B1	14.38	3.19	0.222	0.0349
B2	14.54	0.58	0.040	0.0349
C1	14.29	4.00	0.280	0.0674
C2	14.78	0.58	0.040	0.0674
D1	14.42	2.90	0.201	0.1334
D2	14.38	0.57	0.040	0.1334

STD = standard deviation; CV = coefficient of variation.

2.4. 대체 방법

군집화 및 특성도를 이용한 결측치 대체 방법의 단계는 다음과 같다.

Step 1: 반복 측정 자료에서 1 시점과 2 시점은 결측이 발생하지 않는다고 가정한다.

Step 2: 1 시점과 2 시점의 값으로 계층적 군집분석의 Ward 연결법을 이용하여 두 개의 군집으로 나눈다.

Step 3: $m\,(m \geq 3)$ 시점에서 첫 결측이 발생했을 때 군집별로 $m-1$ 시점까지의 자료를 이용해 각 개체의 일치 지수, 유지 지수, 특성도를 구하고, 결측치를 결측치의 특성도와 가장 가까운 특성도를 가진 값으로 대체한다.

Step 4: 결측치의 특성도와 가장 가까운 특성도가 2개 이상일 경우 해당 관측치의 평균을 이용하여 대체한다.

Step 5: m시점 이후 결측이 발생한 경우 대체된 자료를 포함하여 단계 3, 4를 반복하여 결측치를 대체한다.

3. 모의실험 및 결과

제안하는 방법을 기존 방법인 평균 대체법, 핫덱 대체법, 특성도 대체법, Markov chain Monte Carlo (MCMC) 방법과 결측치 대체 성능을 비교하기 위하여 Monte Carlo 모의실험을 시행하였다. MCMC 방법은 정상분포에 안정화하기 위해 충분히 긴 마코브 연쇄를 만든 후 자료가 다변량 정규분포를 따른다는 가정하에 베이지안 추론에 자료증대 방법을 통하여 1개의 완전한 자료집합을 만든 후 완전한 자료집합으로부터 구한 추정치들의 평균을 이용하여 결측치를 보정하는 방법이다 (Lee, 2008). 30개의 개체($n = 30$)에 대해 8회 반복 측정($t = 8$)한 자료의 형태로 반복 요인이 하나인 반복 측정 분산분석을 실시하여, 구형성 가정이 만족되면 univariate-ANOVA 결과를 이용하였고, 구형성 가정이 만족되지 않으면 multivariate analysis of variance (MANOVA) 검정의 결과를 이용하였다. p-값이 서로 다른 4개가 되고, 각 p-값마다 변동 계수가 다른 데이터가 2개씩 되도록 총 8개의 데이터셋을 정규분포 가정 하에 임의로 구축하였다. 또한 자료 생성시 증감 여부나 기울기 정도를 나타내는 자료의 형태와 초기 시점과의 연관성 정도를 다르게 설정하여 자료의 형태나 연관성에 따라서도 결측 대체 성능을 비교할 수 있도록 하였다. 각 데이터의 기초통계량 및 p-값은 Table 3.1과 같다.

자료 A1, A2는 p-값이 0.009로 가장 낮게 설정하였고, 자료의 형태가 증가-감소-일정의 패턴을 보였다. 여기서 A1은 변동계수(coefficient of variation; CV)가 0.083, 초기 시점과의 연관성이 0.7이 되도록 하

였고, A2는 변동계수가 0.05, 초기 시점과의 연관성이 0.3이 되도록 설정하였다. 자료 B1, B2는 p-값이 0.0349, 자료의 형태는 가파르게 증가하다가 완만하게 증가하는 패턴을 보였다. 여기서 B1은 변동계수가 0.222, 초기 시점과의 연관성이 0.8이 되도록 하였고, B2는 변동계수가 0.040, 초기 시점과의 연관성이 0.5가 되도록 설정하였다. 자료 C1, C2는 p-값이 0.0674, 자료의 형태가 기울기 변화가 거의 없는 감소하는 패턴을 보였다. 여기서 C1은 변동계수가 0.280, 초기 시점과의 연관성이 0.8이 되도록 하였으며, C2는 변동계수가 0.040, 초기 시점과의 연관성은 C1과 마찬가지로 0.8로 설정하였다. 자료 D1, D2는 p-값이 0.1334로 가장 높게 설정하였으며, 자료의 형태는 기울기 변화가 거의 없이 증가하는 패턴을 보였다. 여기서 D1은 변동계수가 0.201, 초기 시점과의 연관성이 0.7이 되도록 하였으며, D2는 변동계수가 0.040, 초기 시점과의 연관성이 0.1이 되도록 설정하였다. 그리고 8개의 데이터에 각각 완전임의결측 가정 하에 5%, 10%, 20%의 결측을 발생 시킨 후 결측 비율별로 평균 대체법, 핫덱 대체법, 특성도 대체법, 제안하는 방법을 통해 결측치를 대체하였다. 결측치를 대체한 자료들을 이용하여 한 집단, 한 반복 요인의 반복 측정 분산분석을 실시하는 과정을 1,000회씩 반복하였다.

대체 방법과 제안 방법을 비교하기 위한 지수인 원자료의 검정 결과의 기각 여부에 따른 대체된 자료의 검정결과가 원자료와 같은 비율(rate of rejected or adopted; RRA), 대체된 자료의 p-값과 원자료의 p-값 간 차이의 제곱합(sum of square of p-value's difference; SSP), 정규화 제곱근 평균오차(normalized root mean squared error; NRMSE)를 다음과 같이 정의하고 모의실험 결과를 각각 Table 3.2, Table 3.3, Table 3.4로 정리하였다 (Kim과 Kim, 2017).

$$\text{RRA} = \frac{[\text{원자료가 기각일 때 대체된 자료가 기각한 횟수(반대의 경우도 포함)}]}{1000}, \tag{3.1}$$

$$\text{SSP} = \sum_{k=1}^{1000} (p_k - p)^2, \tag{3.2}$$

$$\text{NRMSE} = \frac{1}{y'_{\max} - y'_{\min}} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(y_{ij} - y'_{ij})^2}{M} \right\}^2, \tag{3.3}$$

여기서 p_k는 대체된 자료의 p-값, p는 원자료의 p-값, y_{ij}는 실제값, y'_{ij}는 추정값, M은 결측치 수, $y'_{\max}$는 추정치 중 최댓값, $y'_{\min}$는 추정치 중 최솟값이다 (Kim과 Kim, 2017). RRA는 그 값이 클수록 대체 성능이 좋음을 의미하며, SSP와 NRMSE는 그 값이 작을수록 대체 성능이 좋음을 의미한다 (Kim과 Kim, 2017).

Table 3.2는 결측 대체된 자료의 RRA를 나타낸 표이다. p-값이 매우 작고 변동계수가 상대적으로 높으며, 증가-감소-일정의 패턴을 보이고 초기 시점과의 연관성이 높은 자료에서는 5%, 10%, 20% 결측에서 제안 방법과 MCMC 방법이 0.98 이상으로 가장 우수하게 나타났으며, 평균 대체와 핫덱 대체의 경우 5% 결측에서 0.933, 0.912을 기록하고 결측비율이 높아질수록 RRA 값이 더 떨어져 상대적으로 우수함이 떨어지는 것으로 나타났다. 반면, p-값이 매우 작고 증가-감소-일정의 패턴을 보이나 변동계수와 초기 시점과의 연관성이 낮은 자료에서는 평균 대체와 MCMC 방법에 비해 제안 방법이 약간 우수함이 떨어지는 것으로 나타났다. p-값이 약간 유의한 값을 가지고, 가파르게 증가했다가 완만하게 증가하는 형태를 띄는 자료에서는 대부분 MCMC 방법, 제안 방법 순으로 우수함이 나타났다. 여기서 변동계수가 높고, 초기 시점과의 연관성이 큰 자료에서는 평균 대체, 핫덱 대체, 특성도 대체에 비해 결측 비율이 커질수록 제안 방법이나 MCMC 방법이 약 0.3 이상 차이의 우수함이 나타났다. 변동계수가 높고, 초기 시점의 연관성이 작은 자료에서는 제안 방법과 MCMC 방법이 가장 좋았고 기존 방법과의 대체 성능 차이가 결측이 커질수록 차이가 줄어들었으며, 특히 특성도 대체법과의 우수성 차이가 결측이 커질수록 거의 나지 않았다. p-값이 약간 유의하지 않은 값을 가지고, 기울기의 변화가 거의 없이 감소하는 형태의

Table 3.2. Rate of rejected or adopted (RRA, $\alpha = 0.05$)

Data	Missing	Mean	HD	Pro	MCMC	Sub
A1	5%	0.933	0.912	0.971	1.000	0.984
	10%	0.893	0.867	0.977	0.999	0.990
	20%	0.884	0.791	0.962	0.999	0.992
A2	5%	0.995	0.977	0.979	1.000	0.985
	10%	0.992	0.980	0.988	0.997	0.984
	20%	0.999	0.991	0.997	0.999	0.997
B1	5%	0.782	0.791	0.749	0.990	0.872
	10%	0.569	0.555	0.638	0.986	0.903
	20%	0.584	0.461	0.651	0.989	0.922
B2	5%	0.847	0.870	0.955	0.984	0.980
	10%	0.834	0.853	0.957	0.983	0.973
	20%	0.917	0.891	0.990	0.990	0.991
C1	5%	0.603	0.619	0.587	0.559	0.610
	10%	0.798	0.838	0.859	0.189	0.830
	20%	0.634	0.787	0.727	0.053	0.830
C2	5%	0.412	0.543	0.594	0.551	0.658
	10%	0.573	0.748	0.666	0.210	0.670
	20%	0.620	0.808	0.662	0.050	0.634
D1	5%	0.881	0.925	0.936	0.901	0.917
	10%	0.880	0.925	0.890	0.486	0.870
	20%	0.793	0.908	0.780	0.169	0.680
D2	5%	0.949	0.960	0.953	0.913	0.945
	10%	0.918	0.902	0.907	0.405	0.868
	20%	0.658	0.776	0.664	0.179	0.628

HD = Hot-Deck; Pro = Measure of Property; MCMC = Markov chain Monte Carlo; Sub = suggest method.

자료에서는 MCMC 방법의 성능이 낮았으며, 특히 결측의 비율이 증가함에 따라 큰 폭으로 대체 성능이 낮아짐을 보였다. 변동계수가 크며, 연관성이 높은 경우 결측 비율이 증가함에 따라 기존 방법에 비해 제안 방법이 조금 더 우수함을 보였고, 변동 계수가 작으며, 연관성이 높은 경우에는 제안 방법이 우수함을 나타내나 결측 비율이 증가함에 따라 상대적으로 기존 방법보다 점점 우수성이 약간씩 떨어짐을 나타내었다. p-값이 매우 높고, 기울기의 변화가 거의 없이 증가하는 형태의 자료에서는 MCMC 방법이 결측 비율이 증가함에 따라 대체 성능이 큰폭으로 떨어졌다. 그리고 기존 방법인 평균 대체, 핫덱 대체, 특성도 대체보다 제안 방법이 상대적으로 RRA 값이 좋지 않았다. 초기 시점과의 연관성 및 변동계수를 고려 하였을 경우에는 제안 방법의 상대적 대체 성능 순위에는 큰 변화가 없었다.

Table 3.3은 결측 대체법과 자료에 따른 SSP를 나타낸 표이다. p-값이 매우 작고 변동계수가 상대적으로 높으며, 증가-감소-일정의 패턴을 보이고 초기 시점과의 연관성이 높은 자료에서는 MCMC 방법, 제안 방법 순으로 우수성이 좋았으며, 핫덱 대체가 0.53, 0.84, 1.66으로 가장 좋지 못하였다. p-값이 매우 작고 변동계수가 상대적으로 낮으며, 증가-감소-일정의 패턴을 보이고 초기 시점과의 연관성이 낮은 자료에서는 MCMC 방법이 가장 좋았으며, 제안 방법이 0.138에서 0.084값을 가져 평균 대체, 특성도 대체보다 상대적으로 우수성이 떨어졌다. p-값이 약간 유의한 값을 가지고, 가파르게 증가했다가 완만하게 증가하는 형태이며, 변동계수가 높고 연관성이 높은 자료에서는 제안 방법이 0.53, 0.57, 0.72로 가장 우수함을 보였으며, 결측 비율이 높아질 수록 SSP 값이 상대적으로 더 낮아졌다. p-값이 약간 유의한 값을 가지고, 가파르게 증가했다가 완만하게 증가하는 형태이며, 변동계수가 작고 초기 시점과의 연관성

Table 3.3. Sum of square of p-value's difference (SSP)

Data	Missing	Mean	HD	Pro	MCMC	Sub
A1	5%	0.4023	0.5306	0.1924	0.0405	0.1468
	10%	0.6274	0.8435	0.2183	0.0690	0.1426
	20%	0.9345	1.6670	0.2588	0.0750	0.1261
A2	5%	0.1237	0.2032	0.1658	0.0375	0.1386
	10%	0.1193	0.1769	0.1340	0.0735	0.1544
	20%	0.0726	0.1139	0.0831	0.0782	0.0847
B1	5%	1.2303	1.1433	0.6818	0.3067	0.5310
	10%	3.9360	4.1997	1.3221	0.6392	0.5781
	20%	4.3677	5.8678	2.4496	1.0752	0.7245
B2	5%	0.3988	0.4592	0.4199	0.2993	0.3527
	10%	0.4849	0.5716	0.5762	0.5998	0.4908
	20%	0.5905	0.6219	0.7834	1.0485	0.7428
C1	5%	21.3325	28.3074	19.4034	1.0293	12.3243
	10%	48.8470	63.5364	58.2752	2.4372	55.4713
	20%	39.4746	63.7684	44.3036	3.8769	64.1713
C2	5%	5.3191	10.6705	4.9378	1.0417	2.3425
	10%	29.2339	43.9918	20.0055	2.5614	4.6760
	20%	39.2474	73.9760	36.0639	3.9235	10.3874
D1	5%	25.9181	43.0497	42.9028	22.4960	34.1016
	10%	66.5847	94.7286	57.4027	13.5570	27.2226
	20%	78.7035	132.0466	66.1604	17.2738	27.7840
D2	5%	35.2248	34.9153	35.9353	22.0225	39.8248
	10%	29.5110	31.8152	26.1706	13.0360	28.7773
	20%	18.3082	31.1842	13.8431	17.3759	16.6074

HD = Hot-Deck; Pro = Measure of Property; MCMC = Markov chain Monte Carlo; Sub = suggest method.

이 낮은 자료에서는 결측 비율이 낮을 때는 MCMC 방법, 제안 방법, 평균 대체 순으로 좋았으나, 결측 비율이 증가 함에 따라 평균 대체, 제안 방법, MCMC 방법 순으로 우수함을 나타내었다. p-값이 약간 유의하지 않은 값을 가지고, 기울기의 변화가 거의 없이 감소하는 형태, 변동계수가 크고, 초기 시점과의 연관성이 높은 자료에서는 MCMC 방법이 가장 좋았다. 제안 방법의 경우 결측 비율이 낮을 때는 상대적으로 우수하였으나, 결측 비율이 높아짐에 따라 SSP 값이 커져 우수성이 떨어졌다. 위와 같은 조건에서 변동 계수만 작아지는 경우에는 MCMC 방법, 제안 방법 순으로 우수성을 보였다. 제안 방법은 변동계수가 큰 경우보다 작은 경우가 기존 방법에 비해 SSP의 값이 우수하였으며, 결측 비율이 증가할수록 상대적으로 우수함이 더 잘 나타났다. p-값이 매우 높고, 기울기의 변화가 거의 없이 증가하는 형태의 자료에서는 MCMC 방법이 대부분의 경우에서 가장 우수하였다. 여기서, 변동계수가 크고, 연관성이 높은 경우 결측 비율이 커짐에 따라 제안 방법이 평균 대체보다 더 우수하였다. 같은 조건에서 변동계수가 낮아진 경우에는 결측 비율이 낮을 때는 제안 방법이 가장 좋지 않았으나 20% 결측에서는 제안 방법이 16.6으로 기존 방법보다 우수함을 보였다. 이는 결측 비율이 커짐에 따라 제안 방법이 더 우수해지는 것을 보여준다.

Table 3.4는 결측 대체법과 자료에 따른 NRMSE를 나타낸 표이다. p-값이 매우 작고 변동계수가 상대적으로 높으며, 증가-감소-일정의 패턴을 보이고 초기 시점과의 연관성이 높은 자료에서는 제안 방법이 0.085, 0.069, 0.067로 MCMC 방법, 제안 방법, 특성도 대체 순으로 우수함을 보였다. p-값이 매우 작고 변동계수가 상대적으로 작으며, 증가-감소-일정의 패턴을 보이고 초기 시점과의 연관성이 낮은 자료

Table 3.4. Normalized root mean squared error (NRMSE)

Data	Missing	Mean	HD	Pro	MCMC	Sub
A1	5%	14.9234	5.4005	0.2501	0.0502	0.0857
	10%	12.7308	4.3726	0.2376	0.0252	0.0695
	20%	10.7686	3.7764	0.2624	0.0183	0.0674
A2	5%	2.8827	1.4363	0.4253	0.0862	0.6191
	10%	2.3386	1.0888	0.3206	0.0412	0.4706
	20%	2.1174	0.9079	0.2918	0.0446	0.4563
B1	5%	246.8032	53.4954	1.4587	0.0001	0.1479
	10%	185.1006	44.1211	1.3542	0.0001	0.1462
	20%	133.8119	39.3937	1.3047	0.0002	0.1623
B2	5%	0.3569	0.2574	0.0293	0.0006	0.0178
	10%	0.3226	0.2104	0.0274	0.0004	0.0155
	20%	0.3101	0.1843	0.0281	0.0008	0.0164
C1	5%	650.5083	103.8070	3.6257	0.0000	0.2731
	10%	436.2344	8.5859	3.3396	0.0000	0.2477
	20%	294.2039	82.3199	3.1916	2.9041	0.2432
C2	5%	1.534	0.3210	0.0130	0.0000	0.0007
	10%	1.1581	0.2644	0.0124	0.0000	0.0007
	20%	0.8383	0.2367	0.0125	0.0000	0.0007
D1	5%	197.8760	40.7835	1.2802	0.0002	0.0604
	10%	147.5374	32.7904	1.2193	0.0001	0.0547
	20%	104.0820	30.0526	1.1781	0.0002	0.0600
D2	5%	0.4555	0.2607	0.0414	0.0008	0.0434
	10%	0.3759	0.1971	0.0349	0.0006	0.0366
	20%	0.3518	0.1646	0.0381	0.0008	0.0355

HD = Hot-Deck; Pro = Measure of Property; MCMC = Markov chain Monte Carlo; Sub = suggest method.

에서는 제안 방법이 오히려 특성도 대체보다 NRMSE 값이 좋지 못하였다. p-값이 약간 유의한 값을 가지고, 가파르게 증가했다가 완만하게 증가하는 형태의 자료에서는 MCMC 방법이 가장 우수하였으며 제안 방법이 약 0.15 및 0.015의 값으로 바로 뒤를 따랐다. 여기서 특히 변동계수가 큰 자료에서는 평균 대체의 NRMSE 값이 매우 컸으며, 변동계수가 작아짐에 따라 평균 대체, 핫덱 대체의 값의 감소 폭이 매우 컸다. p-값이 약간 유의하지 않은 값을 가지고, 기울기의 변화가 거의 없이 감소하는 형태의 자료에서는 MCMC 방법이 가장 우수하였다. 뒤이어 제안 방법이 우수함을 나타내었다. 평균 대체 및 핫덱 대체의 경우 변동 계수가 클 때는 매우 큰 값을 가졌다. p-값이 매우 높고, 기울기의 변화가 거의 없이 증가하는 형태의 자료에서는 MCMC 방법이 가장 우수하였으며, 변동계수가 높고 초기 시점과 연관성이 높은 자료에서는 제안 방법이 특성도 대체보다 우수하였으나 변동계수가 작고, 초기 시점과의 연관성이 낮은 자료에서는 제안 방법이 특성도 대체보다 우수하지 못하였다. 평균 대체의 경우 앞의 경우와 마찬가지로 변동계수가 큰 경우에서는 매우 높은 값을 가졌다.

4. 결론 및 고찰

본 논문에서는 군집화 및 특성도를 이용한 새로운 결측치 대체 방법을 제안하였고, Monte Carlo 모의실험을 통해 결측치를 대체한 후 RRA, SSP, NRMSE를 이용하여 제안 방법과 기존 방법들 간 결측치 대체 성능을 비교하였다.

모의실험 결과를 p-값으로 해석을 해보면 p-값이 0.05 근처일 경우가 그렇지 않을 경우보다 RRA, SSP, NRMSE에서 제안 방법이 기존 방법보다 우수한 경우가 많았다. 특히 p-값이 0.05보다 약간 낮을 경우가 제안 방법이 가장 우수하였다. 반면 p-값이 아주 큰 경우에는 다른 경우에 비해 우수성이 떨어지는 것을 보였으며, 특히 RRA를 기준으로 보았을 때 이 현상이 두드러지게 나타났다. 변동계수가 클 때와 작을 때를 기준으로 보면 p-값이 0.05보다 작은 경우와 p-값이 매우 큰 경우에는 변동계수가 클 때가 작을 때 보다 제안 방법이 기존 방법과 대체 성능을 비교하였을 때 훨씬 더 우수함을 보였다. 그러나 p-값이 0.05보다 약간 큰 경우에는 오히려 변동계수가 높을 때보다 낮을 때가 제안 방법의 성능이 기존 방법의 성능보다 우수하였다. 자료의 형태 기준으로 보았을 때 기울기의 변화가 있으면서 증가하는 패턴이 다른 패턴들에 비해 제안 방법에서 우수한 성능을 보였으나 기울기의 변화가 없이 증가하는 패턴에서는 제안 방법이 별로 좋지 못하였다. 또한 감소하는 패턴이 증가와 감소가 섞인 패턴보다는 제안 방법에서 좀 더 우수함을 보였다. 초기 시점과의 연관성을 기준으로 보았을 때 연관성이 상대적으로 높은 경우에는 제안 방법이 기존 방법보다 RRA, SSP, NRMSE 세가지 지표에서 우수함을 훨씬 더 보였으며, 반면 연관성이 낮을 때는 클 때에 비해 p-값이 아주 크거나 작을 경우 기존 방법이 제안 방법보다 더 효율적인 경우가 많음을 보여준다.

전체적으로 보면 평균 대체의 경우 NRMSE는 변동계수가 큰 자료에서 매우 높게 나타났는데 이는 NRMSE를 정의한 식 (3.3)에서 분모가 다른 대체 방법보다 매우 작아졌기 때문이라 추측할 수 있다. MCMC 방법의 경우 SSP, NRMSE 지표만 보면 대부분의 경우에서 가장 우수한 방법으로 보이며, RRA를 보았을 때 p-값이 0.05보다 작을 경우에는 가장 대체 성능이 좋은 방법으로 보인다. 하지만 p-값이 0.05보다 커지는 경우에는 급격하게 성능이 떨어지는 것을 볼 수있고 이는 결국 p-값을 매우 떨어뜨리는 경향이 짙다고 할 수 있다. 특히 결측비율이 높아질수록 이 현상이 잘 나타났다. 이는 임상적으로 보았을 때 약의 약효가 없는데 약효가 있다고 하는 것과 같이 제 1종 오류를 제어하지 못하는 치명적인 한계가 존재함을 보여준다.

제안 방법은 1종 오류를 제어하지 못한 MCMC 방법을 제외하면 가장 우수한 경우가 많았다. 특히, p-값이 0.05 근처일수록, 변동계수가 상대적으로 큰 값을 가질수록, 기울기의 변화가 있으면서 증가하는 패턴이거나 감소하는 패턴일수록, 초기 시점과의 연관성 정도가 높을 때 결측치 대체 성능이 우수하였다. 자료 형태로 보면 이는 약효가 빠른 시점에서 효과를 보이고, 이 후 시점에서도 오랫동안 약효가 잘 떨어지지 않는 경우에 가장 좋은 결과가 나타났다고 볼 수 있다. 초기 시점과의 연관성을 기준으로 본 경우 연관성이 높은 경우 제안 방법이 우수한 결과를 보였고, 이는 초기 시점의 값이 이 후의 시점에 영향을 미치는 경우, 초기 시점 값에 대해 군집화를 이용하는 것이 기존 결측치 대체 방법보다 결측치 대체에 있어서 더 우수하다는 것을 설명해준다. 또한 대부분의 경우에서 제안 방법이 특성도 대체법보다 우수함을 보였다.

이와 같은 결과들은 결측의 발생이 중도탈락(dropout)인 경우에도 유용할 것으로 여겨진다. 하지만 제안 방법은 초기 시점을 이용하여 군집을 나누어 특성별로 결측을 대체함에 있어 초기 시점이 완전한 자료여야 한다는 한계점이 존재한다. 이를 보완하기 위해 초기 시점의 결측이 발생하였을 경우에 대한 대체 방법의 연구가 필요하며, 표본수가 상대적으로 많은 대규모 임상시험의 경우 군집 개수에 대한 제한이 적으므로 군집 개수의 변화에 따른 대체 성능에 대한 연구도 추후에 필요할 것으로 여겨진다.

References

Choi, Y. and Jeong, K. (2003). *Multivariate Analysis using SAS and Its Application*, Free Academy, Seoul.
Jeon, C. (2012). *Data Mining Techniques and Applications*, Hanarae Academy, Seoul.
Kang, S. (2013). *Medical Statistics for New Medicine Development*, Free Academy, Seoul.

Kim, H. and Kim, D. (2017). Imputation method for missing data based on measure of property, *The Korean Communications in Statistics*, **30**, 463–473.

Lee, S. (2008). Conjugation plan of proc MI, *Industrial Science Research*, **26**, 35–41.

Shin, S. (2010). Model-based cluster analysis of missing data considering outlier, *Korea University Graduate School*.

Ward, J. H. (1963). Hierarchical groupings to optimize an objective function, *Journal of the American Statistical Association*, **58**, 234–244.

군집화 및 특성도를 이용한 결측치 대체 방법

김성현[a] · 김동재[a,1]

[a]가톨릭대학교 의생명· 건강과학과

(2017년 8월 22일 접수, 2017년 10월 23일 수정, 2017년 12월 15일 채택)

요 약

데이터를 수집함에 있어 여러 가지 이유로 결측이 발생하게 된다. 결측치는 분석 및 결과에 적지 않은 영향을 미치므로, 이를 해결하기 위해 결측치를 처리하는 다양한 방법들이 연구되었다. 반복 측정 자료에서 초기 시점의 측정값이 어떠한지에 따라서 뒤의 시점 측정값이 어느 정도 영향을 받을 수도 있을 것으로 생각된다. 하지만 기존 방법에서는 이러한 개념을 이용한 결측치 대체가 없었으므로 본 연구에서는 반복 측정 자료에서 초기 시점을 이용한 군집화 및 Kim과 Kim (2017)이 제안한 특성도를 이용하여 새로운 결측치 대체 방법을 제안하였다. 또한 여러 반복 측정 자료를 이용하여 Monte Carlo 모의실험을 통하여 기존 결측 대체 방법과 제안 방법의 여러 대체 성능을 비교해 보았다.

주요용어: 결측치 대체, 군집화, 특성도, 초기 시점

[1]교신저자: (06591)서울 서초구 반포대로 222, 가톨릭대학교 의생명· 건강과학과. E-mail: djkim@catholic.ac.kr