STAT714 Final Report

MCMC 알고리즘과 Non-MCMC 알고리즘을 사용한 다중 대체(MI) 방법 비교 및 대체간 반복(Between-Imputation Iteration)의 효과

2020021218 통계학과 이소연

1. 개요

1.1. 주제 소개

자료를 수집하는 과정에서 결측치는 다양한 원인으로 발생한다. 하지만, 이러한 결측치를 처리하기 위한 최적의 해법은 알려지지 않았기 때문에, 이를 어떻게 처리해야 할 지 결정하는 것은 쉽지 않다. 본 연구는 Rubin(1987)이 제시한 다중 대체 (Multiple Imputation)의 대체 모형 중 MCMC(Markov Chain Monte Carlo) 알고리즘을 사용한 모델과, 사용하지 않은 모델을 중심으로 비교하며, 'Confidence Proper'(van Buuren, 2018)한 대체 방법을 탐색하고자 하였다. 전반적인 연구 계획은 Takahashi(2017)[1]의 논문을 참조했다.

1.2. 모델 소개

본 절에서는 EMB 대체법(Expectation-Maximization with Bootstrapping), MCMC 대체법(MCMC for joint modeling), 그리고 FCS 대체법(Fully Conditional Specification, 완전조건부 대체)을 중심으로 다룰 것이다. 이 중 EMB 대체법은 MCMC 알고리즘을 사용하지 않은 모델이며, MCMC 대체법과 FCS 대체법은 MCMC 알고리즘을 사용한 모델로 정의한다.

먼저, EMB 대체법은 붓스트랩 방법을 통해 생성된 m개의 대체 데이터셋(multiple set) 각각의 최대 우도 추정치(MLE)를 추정하는 방법이다. 결측치를 대체하기 위해 적용되는 EM 알고리즘은 E(expectation)와 M(maximization)의 단계를 거친다. E 단계에서는 (1) 균등분포에서 임의로 추출된 초기값, 혹은 관측된 값으로 결측치의 로그 우도(Log Likelihood) 기댓값을 추정하고, M 단계에서는 (2) 대체값의 로그 우도 조건부 기댓값이 최대화될 때까지(혹은 수렴할 때까지) 앞선 단계를 반복한다. 이를 수식으로 나타내면 아래와 같다.

$$Q(\theta|\theta^t) = \int l(\theta|Y) f(Y_0|Y_1, \theta^t) dY_1 \qquad (1)$$

$$\theta^{t+1} = argmax_\theta Q(\theta|\theta^t) \qquad (2)$$

Dempster, Laird, Rubin(1977)은 MAR 가정 아래 모든 변수들이 다변량정규분포를 따를 때 이러한 EM 알고리즘은 타당한 결과가 도출됨을 밝혔다. 또한, 결측 발생 매커니즘이 MAR인 경우, EM 대체법은 대체로 비편향된(unbiased) 평균 추정치를 제공한다고 알려져 있다 (Peng, 2006).

1) Takahashi, M. (2017). *Statistical inference in missing data by MCMC and non-MCMC multiple imputation algorithms: Assessing the effects of between-imputation iterations*. Data Science Journal, 16.

다음으로, MCMC 대체법은 베이지안 추정방법을 바탕으로 한 MCMC 알고리즘을 사용하여 사후분포를 만든 뒤, 결측치를 대체하는 방법이다. 이는 I(imputation) 단계와 P(posterior)단계로 이루어져 있으며, I 단계에서는 (1) 분포의 모수를 알고 있다는 가정하에, 결측치들의 조건부 분포에서 EM 알고리즘을 통해 결측값의 대체값을 산출한다. P 단계에서는 (2) 산출된 대체값을 통해 모수를 다시 추정한다. 모수의 초기치는 EM 알고리즘의 추정치를 사용하며, 사후분포가 수렴할 때까지 앞선 두 단계를 반복한다. 수식으로 표현하면 다음과 같다.

$$Y_1^{t+1} \sim P(Y_1|Y_0, \theta^t) \qquad (1)$$

$$\theta^{t+1} \sim P(\theta|Y_0, Y_1^{t+1}) \qquad (2)$$

MCMC 대체법이 사용하는 Data augmentation의 경우, 200번의 burn-in단계를 거친 후. 각 대체 데이터셋의 독립성을 보장하기 위해 한 데이터셋이 생성된 뒤 다음 데이터셋까지 일련의 between-imputation iteration을 수행한다. 해당 Iteration 수는 경험에 입각하여(rule of thumb), EM 알고리즘의 iteration 수의 두 배를 사용한다(Schafer and Olsen, 1998).

마지막으로, FCS 대체법은 MCMC 대체법과 유사하지만, 결측치 대체에 있어 사전 분포에 대한 가정을 생략한다는 점이서 차이가 있다. 이는 분포에 대한 가정 없이, 연속된 회귀방정식을 통해 단일적으로(univariately) 값을 대체해 나가는 방식이다. 따라서 FCS 알고리즘은 변수의 순서대로 (1) 모수를 추정하고 (2) 추정된 모수를 통해 대체값을 찾는 과정을 수렴할 때까지 반복한다. 이를 수식으로 표현하면 다음과 같다.

$$\widetilde{\theta_j^t} \sim P(\theta_j^t|Y_{j(0)}, \widetilde{Y_{-j}^t}) \qquad (1)$$

$$\widetilde{Y_j^t} \sim P(Y_{j(1)}|Y_{j(0)}, \tilde{Y}_{-j}^t, \widetilde{\theta_j^t}) \qquad (2)$$

FCS 대체법 또한 MCMC 알고리즘을 기반으로 한 방법론이므로, 앞선 MCMC 대체법과 같이 between-imputation iteration을 수행한다. 본 프로젝트는 이 between-imputation iteration의 효과를 검증하고자, 이 두 대체법에 대해 between-imputation iteration을 수행하지 않은 경우(MCMC1, FCS1)와 수행한 경우(MCMC2, FCS2)를 나누어 결과를 비교해보았다.

각 대체 데이터셋 별 추정치($\bar{\beta}$)와 표준오차를 결합하는 과정은 다음과 같다. 이는 위 세 가지 대체법에 동일하게 적용된다.

$$\bar{\beta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_i$$

$$\mathrm{SE}(\bar{\beta})^2 = \frac{1}{m}\sum_{i=1}^{m} SE(\beta_i)^2 + S_\beta^2\left(1+\frac{1}{m}\right), \qquad S_\beta^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\beta_i - \bar{\beta})^2$$

가능한 여러 대체 알고리즘의 비교를 위해, 결측치가 존재하지 않는 Complete Data를 사용했을 경우와, 결측이 존재하는 변수를 모두 제외하는 완전제거법(Listwise deletion), 그리고 대표적인 단일대체법(Single Imputation)인 확률적 회귀대체법(Stochastic Regression Imputation)을 앞서 소개한 방법과 함께 비교하였다.

1.3. 데이터 소개

본 프로젝트에선 가상의 데이터에 대한 몬테카를로 모의실험을 진행하였다. 즉, 일련의 난수를 바탕으로 자료의 생성과 분석을 500번 반복한 후, 그 결과를 종합하여 임의의 목표 변수인 β_1을 확률적으로 추론하고자 하였다.

시뮬레이션 데이터를 만들기 위한 각 초모수는 Takahashi(2017)의 원문과 같이, 유사한 주제를 다룬 20여개의 논문에서 설정한 해당 초모수의 약 75번째 백분위수로 지정하였다. 그 결과, 1000개의 관측치수(n)와 2개-10개의 변수(p)를 선택하였다. 단, 대체 데이터셋 개수(number of imputation sets, m)는 시간적 제약으로 인해 10번에서 5번으로 수정하였다. 이때 p-1개의 설명변수 X는 $N_{p-1}(0,1)$의 다변량 정규분포를 따르며, $U(-1,1)$으로부터 임의로 추출된 수로 구성된 상관행렬을 만족시키도록 만들어졌다. 마지막으로, $\beta_j \sim U(-2,2)$, $\varepsilon \sim N(0,\sigma)$, $\sigma \sim U(0.5,2)$ 의 가정 아래 y는 다음과 같이 생성되었다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots \beta_{p-1} x_{p-1} + \varepsilon_i$$

결측 발생 메커니즘이 MAR가정을 만족시키도록, y_i는 결측치가 없는 완전한 변수이며, x_i은 y_i과 임의의 난수 $u_{ij} \sim Unif(0,1)$의 값에 의존하여 결측치를 갖는 변수라고 가정하였다. 조건은 아래와 같다.

$$x_{ji} \begin{cases} \text{missing} & \text{if } y_i < \text{median}(y_i) \text{ and } u_{ij} < 0.5 \text{ or } y_i > \text{median}(y_i) \text{and } u_{ij} > 0.9 \\ \text{not missing} & \text{otherwise} \end{cases}$$

위 결측 발생 메커니즘에 따르면, 각 x_j는 약 30%의 결측확률을 갖게 되었다.

1.4. 평가 지표 소개

1.4.1. 가정

이제 분석에 사용한 가정을 명시하겠다. 먼저 MAR 가정이다. 앞서 언급하였듯, 결측이 발생할 확률은 오직 관측된 데이터인 Y에 의존하기 때문에 결측 발생 매커니즘은 MAR 가정을 만족한다. 다음으로, Ignorability 가정이다. Rubin(1976)은 (1) MAR가정 아래 (2) 결측 발생 매커니즘의 모수 θ와 통계적 모형의 모수 φ가 독립이라면(Parameter Distinctness), 결측치가 최대우도법으로 추정되는 모수φ에 영향을 주지 않기 때문에 무시할 수 있다고(Ignorable) 주장했다. 이론상으로, 위 MAR과 Ignorability가정이 만족된다면, 앞선 세 알고리즘의 추정치는 수렴 시 최대 우도 추정치(MLE)가 되므로, 일치성(consistency)과 불편성(unbiasedness)을 충족시킨다.

1.4.2. Confidence Proper

Schafer(1997)은 대체 값들이 $\Pr(Y_{mis}|Y_{obs})$에서 독립적으로 추출된 추청치를 기반으로 도출되었다는 'Proper Imputation'의 개념을 소개했다. 이를 간소화한 것이 van Burren(2012)의 'Confidence proper'이다. 특정 대체 방법이 'Confidence Proper'하려면, 다음의 세 가지 사항을 만족해야 한다. 여기서 D는 주어진 데이터, $\bar{V}$는 대체내(within) 분산의 평균, $\hat{V}$는 이론상 전체

데이터의 분산, 그리고 B는 대체간(between) 분산이다.

$$(1)\ E\left(\bar{\beta}|D\right) = \hat{\beta}$$

$$(2)\ E(\bar{V}|D) = \hat{V}$$

$$(3)\ \left(1+\frac{1}{M}\right)E(B|D) \geq V(\bar{\beta})$$

(1)과 (2)는 불편성을 내포하며, (3)은 대체 과정에 있어 결측치 발생에 의한 추가적인 불확실성(uncertainty)이 반영되어야 한다는 의미이다. 본 프로젝트에선 결측 대체 방법의 'Confidence Properness'를 판단하기 위한 지표로 'Bias,' 'RMSE,' 'Coverage Rate,' 그리고 'CI length'를 설정하였다.

2. 분석 결과

Biased results (Bias > 0.010.) are in boldface.

	The Number of Predictors $(p-1)$								
	1	2	3	4	5	6	7	8	9
CD	-0.0041	-0.0043	-0.0040	-0.0049	-0.0055	-0.0060	-0.0057	-0.0088	-0.0072
LD	-0.0059	-0.0073	-0.0043	-0.0080	**-0.0110**	-0.0097	**-0.0103**	**-0.0146**	**0.0119**
SSI	-0.0024	-0.0075	-0.0040	-0.0044	-0.0049	0.0027	-0.0032	**-0.0325**	**-0.0859**
EMB	-0.0046	-0.0038	-0.0039	-0.0062	-0.0071	-0.0082	-0.0048	**-0.0486**	**-0.0480**
MCMC1	-0.0024	-0.0038	**-0.0132**	**-0.0178**	**-0.0173**	**-0.0207**	**-0.0191**	**-0.0110**	-0.0087
MCMC2	-0.0004	-0.0006	-0.0006	-0.0008	0.0001	-0.0005	-0.0003	0.0088	0.0131
FCS1	-0.0016	-0.0010	**-0.0102**	-0.0095	**-0.0106**	**-0.0115**	**-0.0118**	-0.0053	-0.0067
FCS2	-0.0033	0.0002	-0.0014	-0.0019	-0.0024	-0.0020	-0.0001	-0.0050	**0.0399**

CD 는 결측치가 없는 Complete Data, LD 는 완전제거법(Listwise deletion), 그리고 SSI 는 확률적 회귀대체법(Stochastic Regression Imputation)이다. 우선 방법 별 추정치의 편향(bias)를 알아보겠다. 편향이 0.01 보다 작으면서 CD 와 유사하다면, 비편향 추정치라고 판단하였다. 위 결과표를 통해, LD, MCMC1, FCS1 방법이 편향이 있음을 확인할 수 있다. 특히, Between-Imputation Iteration 이 없는 MCMC 와 FCS 대체법의 편향이 눈에 띄게 컸다.

Inefficient results (RMSE > 0.2.) are in boldface.

	The Number of Predictors $(p-1)$								
	1	2	3	4	5	6	7	8	9
CD	0.0606	0.0608	0.0714	0.0719	0.0723	0.0754	0.0758	0.1573	**0.3088**
LD	0.0774	0.1189	0.1910	**0.2543**	**0.3000**	**0.3377**	**0.3710**	**0.4850**	**0.7989**
SSI	0.0742	0.0815	0.1080	0.1115	0.1194	0.1353	0.1239	**0.4124**	**0.5199**
EMB	0.0596	0.0670	0.0911	0.1001	0.1071	0.1126	0.1131	**0.5220**	**0.4874**
MCMC1	0.0884	0.1020	0.1630	0.1797	0.1853	**0.2002**	**0.2019**	**0.2301**	**0.2262**
MCMC2	0.0349	0.0377	0.0475	0.0517	0.0547	0.0559	0.0564	**0.2015**	**0.4773**
FCS1	0.0442	0.0627	0.1174	0.1331	0.1403	0.1616	0.1632	**0.2061**	**0.2076**
FCS2	0.0431	0.0479	0.0638	0.0683	0.0743	0.0782	0.0787	**0.2948**	**0.6016**

RMSE 의 결과 또한 위 결과와 비슷했다. 이때 변수가 많아질 수록($\geq$8), 대부분 방법의 bias 와 RMSE 가 커졌다. 이는 설명변수의 개수가 늘어남에 따라 생긴 다중공선성의 문제라고 예측하였다. 실제로, 8 번째와 9 번째 변수의 상관계수는 상대적으로 높은 경우(>0.6)가 많았다. 위 두 결과를 통해, SSI, EMB, MCMC2, 그리고 FCS2 는 기저 이론에 부합하는 비편향추정치를 도출하며, RMSE 가 작은 효율적인(Efficient) 추정치라고 판단하였다.

Confidence invalid results (outside of 0.931 and 0.969) are in boldface.

	The Number of Predictors $(p-1)$								
	1	2	3	4	5	6	7	8	9
CD	0.946	0.942	0.942	0.940	0.942	0.932	0.940	0.940	0.938
LD	0.924	**0.824**	**0.708**	**0.600**	**0.524**	**0.526**	**0.532**	**0.818**	**0.906**
SSI	**0.882**	**0.820**	**0.766**	**0.752**	**0.730**	**0.670**	**0.754**	**0.430**	**0.352**
EMB	0.950	0.934	**0.914**	**0.912**	**0.912**	**0.918**	**0.908**	**0.884**	**0.874**
MCMC1	**0.986**	**0.974**	**0.928**	**0.906**	**0.874**	**0.842**	**0.862**	**0.704**	**0.738**
MCMC2	0.962	**0.974**	0.960	0.966	0.950	0.954	0.954	0.952	0.962
FCS1	0.944	**0.876**	**0.670**	**0.628**	**0.632**	**0.600**	**0.604**	**0.522**	**0.506**
FCS2	0.934	0.934	0.944	0.936	**0.910**	**0.898**	**0.930**	**0.874**	**0.862**

다음은 Coverage Rate이다. 이는 추정량의 신뢰구간 500개 중 실제 참값을 포함하는 비율을 일컫는다. $SE(CR) = \sqrt{\frac{CR(1-CR)}{500}}$ 를 통해 95% CR에 대한 신뢰구간 [93.6%, 96.4%]을 계산하여, 이 사이에 각 모델의 CR값이 존재해야 Confidence valid 하다고 판단했다. 그 결과, MCMC2 방법의 결과가 가장 confidence valid하였다. FCS2, EMB 기법 또한 나쁘지 않은 결과를 도출했으나, 변수가 커짐에 따라 제시된 신뢰구간을 만족시키진 못했다. SSI 방법의 경우, 앞서 Bias나 RMSE 면에서는 좋은 성능을 보였으나, 상당히 낮은 CR을 가졌다. 이는 아무리 확률오차항(random error term)을 추가하여 변동성을 고려하였지만, 여전히 표본오차를 과소 추정하게 되는 단일대체법의 한계로 볼 수 있었다.

Lengths too short or large (more than 2 times longer or less than 2/3 time shorter) are in boldface.

	The Number of Predictors $(p-1)$								
	1	2	3	4	5	6	7	8	9
CD	0.2186	0.2188	0.2518	0.2568	0.2575	0.2653	0.2669	0.5601	1.0797
LD	0.2593	0.2992	0.3843	0.4323	0.4741	**0.5321**	**0.5830**	**1.2608**	**2.6525**
SSI	0.2186	0.2194	0.2539	0.2606	0.2622	0.2697	0.2728	0.4416	**0.4907**
EMB	0.2575	0.2752	0.3422	0.3677	0.3872	0.4044	0.4104	**1.6975**	**2.7171**
MCMC1	**0.5417**	**0.5301**	**0.6125**	**0.6349**	**0.6287**	**0.6315**	**0.6366**	0.6266	**0.6635**
MCMC2	0.2488	0.2963	0.2475	0.2653	0.2795	0.2935	0.2963	1.0208	**2.4805**
FCS1	0.2796	0.2089	0.2653	0.2863	0.3087	0.3213	0.3257	**0.3480**	**0.3541**
FCS2	0.2493	0.2938	0.2432	0.2555	0.2703	0.2814	0.2820	0.9266	**1.8429**

마지막으로, 신뢰구간 길이(CI length)이다. FCS2, MCMC2, EMB 방법순으로 실제 CD와 가장 유사한 결과를 도출하였다. SSI 방법의 경우, CD와 길이가 거의 동일하여, 대체 과정에서 결측치 발생에 의한 추가적인 불확실성이 반영되지 않았다고 해석하였다. 즉, 'Confidence proper'의 세 번째 조건을 만족시키지 못했다고 판단하였다.

3. 결론

두 MCMC 대체법과 FCS대체법을 비교해 본 결과, Between-Imputation Iteration은 MCMC알고리즘이 'Confidence proper'하기 위해 필수적인 요소임을 확인하였다. Between-Imputation Iteration이 존재할 시, MCMC 방법와 FCS 방법은 대부분의 경우에서 confidence validity를 보였다. EMB 대체법은 비 MCMC 알고리즘이기 때문에 이론상 Between-Imputation Iteration이 존재하지 않음에도, 앞선 두 방법에 비해 조금 낮거나 비슷한 성능을 보였다. 하지만, 실제 데이터 분석에선 계산에 소요되는 시간을 무시할 수 없다. 본 프로젝트 시뮬레이션 과정에서 FCS대체법과 MCMC 대체법은 EMB 대체법에 비해 적게는 약 5배, 크게는 약 100배의 시간이 소요되었다. 따라서, 시간 효율을 중시한다면 EMB 대체법이 우월할 수 있겠다고 결론지었다.