박 성 오 교수 지도

석사학위 청구논문

고차원 결측자료의 다중대체 모형 선택

2024

성신여자대학교 대학원

통 계 학 과

이 윤 아

고차원 결측자료의
다중대체 모형 선택

박 성 오 교수 지도

이 논문을 석사학위논문으로 제출함

2023년 11월

성신여자대학교 대학원
통 계 학 과
이 윤 아

인 준 서

이윤아의 석사학위 논문으로 인준함

2023년 11월

심사위원장 박 만 식 인

심 사 위 원 김 동 하 인

심 사 위 원 정 호 현 인

성신여자대학교 대학원

논 문 개 요

다양한 자료에서 자주 발견되는 결측에 대하여 결측값을 모두 제거하고 분석을 수행하면 정보의 손실과 함께 추정량에 편향이 생길 수 있다. 이를 해결하기 위해 각 결측값에 대하여 여러 개의 '그럴듯한' 값을 생성하여 대체한 후 분석 결과들을 재결합하는 다중대체법이 흔히 사용되어 왔다. 다중대체를 위한 가장 대표적인 알고리즘인 MICE 알고리즘은 결측이 포함된 각 변수를 나머지 변수들로 예측하는 지도학습모형, 즉 대체모형을 학습시켜 결측을 메우는 과정을 변수마다 반복한다. 하지만 대부분의 대체모형은 저차원에서 개발되었기 때문에 고차원 환경에서는 기존의 대체모형들이 학습되지 않는다는 한계가 존재한다. 따라서 대체모형을 위한 변수 선택의 필요성이 대두되었다. 본 연구에서는 그래프 이론의 아이디어로부터 새로운 변수 선택법을 제안한다. 제안 방법은 그래프 라쏘와 역확률 가중치 추정량을 결합함으로써 정규성 가정하에서 이론적 정당성을 가지며, 본 논문에서 주요 비교 방법으로 채택된 기존의 고차원 대체모형 DURR과 IURR보다 계산적으로 효율적이고 유연하다. 모의실험을 통해 제안 방법과 비교 방법의 성능을 비교한 결과, 정규성 가정이 만족되는 상황에서 제안 방법이 비교 방법과 비슷하거나 우수한 성능을 보인다는 것을 검증하였다.

목 차

Ⅰ. 서 론

오늘날 의학이나 사회학 분야에서 고차원 결측자료(high-dimensional missing data)가 흔히 발견된다. 의학 분야에선 특정 약물이나 치료법에 대한 생체표지자(biomaker)[1]의 시간에 따른 변화를 파악하기 위해 동일 개체(observation)의 정보를 여러 시점에서 반복 수집하는 종단 연구(longitudinal study)가 흔히 이루어진다. 이때 각 개체, 즉 환자는 행으로, 모든 시점의 생체표지자는 열로 저장된다. 이로 인해 열의 수가 행의 수보다 큰 고차원 자료가 형성될 뿐만 아니라 특정 시점에 환자가 병원에 방문하지 않거나 검사를 중단할 경우 결측은 쉽게 발생한다. Brini & van den Heuvel(2023)는 이러한 형태의 네덜란드 걸음마라톤 자료(Vier Daagse data)를 사용하여 고차원 결측 모형의 비교 연구를 진행하였다. 사회학 분야에서는 Costantini 등(2022)이 유럽 가치 조사 자료(European Values Study data)를 사용한 유사한 연구를 진행했다. 해당 논문은 대규모 사회 조사에서 조사 대상의 무응답률(non-response rate)이 지난 20년 동안 급격히 증가하였으며, 사회학자들이 결측 문제에 많은 관심을 기울이고 있다고 언급하였다. 이 밖에도 전산상의 오류, 조사 대상의 응답 거부, 인터넷 전송 문제 등 다양한 이유로 인해 대부분의 자료에서 결측이 쉽게 발견된다.

결측자료의 분석은 먼저 결측값을 처리한 후에, 분석자의 원래 관심 과제를 분석하는 순으로 진행된다. 여기서 결측값을 어떻게 처리하는지는 분석 결과와 그 신뢰도를 크게 좌우하는 매우 중요한 문제이다. 가장 쉬운 결측값 처리 방법인 완전 제거법(Complete Case Analysis; CCA)은 결측이 발생한 개체를 모두 제거하고 분석을 수행하는 방법으로, 정보의 손실과 함께

1) 혈압, 맥박, 콜레스테롤, DNA, 단백질, 호르몬 등 생물의 상태를 나타내는 지표로 사용되는 물질을 일컫는다. (출처: 「생체표지자」, 『생화학백과』, 네이버 지식백과.)

분석 결과에 편향(bias)을 일으킨다는 사실이 알려져 있다(Little & Rubin, 2002). 이를 효과적으로 개선하는 방법 중 하나는 결측값을 '그럴듯한' 값으로 대체(imputation)하는 것이다. 다중대체법(multiple imputation; Rubin, 1978)은 결측값에 대한 여러 개의 대체값들(imputed values)을 생성하여 대체한 후 분석 결과들을 재결합하는 방법으로, 정보의 손실을 줄여 비교적 유효한 결과와 좋은 성능을 보이며 최근 널리 사용되고 있다. 다중대체를 수행하는 가장 유명한 알고리즘인 MICE (Multiple Imputation by Chained Equations; Van Buuren, 2000) 알고리즘은 결측이 포함된 각 변수를 나머지 변수들로 예측하는 지도학습모형을 학습하여 결측을 메우는 과정을 반복 수행하는데, 이때 사용되는 지도학습모형을 대체모형(imputation model)[2]이라 한다. MICE 알고리즘에 대한 자세한 내용은 3장에 기술한다.

대체모형에 사용할 변수들을 결정하는 것은 어려운 과제이다. 관련해서 일반적으로 사용하는 두 가지 원칙이 있는데, 그 첫 번째는 분석 단계에서 사용할 변수들은 대체모형에도 포함을 시키는 것이다(Meng, 1994). 예를 들어 분석자의 관심 과제가 '혈압'과 'BMI 수치'의 상관성을 보는 것이라면, 혈압 변수와 BMI 변수는 대체모형에도 포함되어야 한다. 두 번째는 결측변수의 예측에 도움이 되는 중요한 변수들을 빠뜨리지 않도록 가능한 한 많은 변수를 대체모형에 포함시키는 것이다(Meng, 1994; Collins 등, 2001).

하지만 고차원 환경에서는 모든 변수를 대체모형에 사용하는 것이 불가능하다는 문제가 존재한다. 대부분의 지도학습모형은 저차원 환경에서 개발되었기 때문에, 고차원 환경에서는 차원 축소나 변수 선택의 기법을 적용하지 않고서는 기존의 대체모형들이 작동하지 않는다. 고차원에서 발생하는 이러한 문제는 차원의 저주(the curse of dimensionality)라고도 불린다. 저차원 자료라 할지라도 사용하는 변수가 너무 많으면 계산적 한계(computational

2) 엄밀히 말하면, MICE 내의 지도학습모형뿐만 아니라 대체를 수행하기 위해 사용하는 모든 모형을 통틀어 '대체모형'이라 일컫는다.

limitation)에 부딪힐 수 있다. Carpenter & Kenward(2008)는 다중대체 시 너무 많은 변수를 사용하는 것에 대한 계산적인 문제들을 소개하였고, Hardt 등(2012)은 다중대체 시 너무 많은 변수를 사용했을 때 추정치의 편향과 정밀도에 부정적인 영향을 미쳤다고 보고하였다. 결론적으로, 대체를 수행할 때는 결측 변수의 예측에 도움이 되는 중요한 변수들을 빠뜨리지 않으면서 계산적 한계에 부딪히지 않는 균형 잡힌 변수 선택이 필요하다 (Costantini 등, 2022).

MICE 알고리즘의 근본적인 아이디어는 나머지 변수들이 주어졌을 때, 대체할 결측 변수의 조건부 분포를 추정하는 것이다. 따라서 본 연구에서는 조건부 의존인(conditional dependent) 변수들만을 추정하여 대체모형에 사용하는 변수 선택법을 제안한다. 이 방식은 MICE 알고리즘에 대해 이론적 정당성을 가질 뿐만 아니라, 결측 변수와 조건부 의존인 변수들을 정확하게 추정한다면 결측 변수의 예측에 도움이 되는 중요한 변수들을 빠뜨리지 않는 동시에 변수의 수를 효과적으로 줄일 수 있다. 변수들 간의 조건부 독립성(conditional independence)을 추정하는 방식으로 다중대체의 변수 선택을 수행한 연구는 이전에 이루어진 바가 없으며 본 연구에서 처음으로 시도한다.

제안 방법은 정규성 가정하에서 개발된 그래프 이론 모형인 그래프 라쏘 (graphical lasso; Friedman 등, 2008)를 사용하여 조건부 독립성을 추정한다. 그래프 라쏘 알고리즘은 표본 공분산 행렬에 의존하는데, 결측자료에서는 표본 공분산 행렬이 편의 추정량(biased estimator)이 되므로 사용하기에 부적절하다. 이에 이전 연구들(Kolar & Xing, 2012; Park 등, 2021)은 표본 공분산 행렬의 편향을 보정한 역확률 가중치(Inverse Probability Weighting; IPW) 추정량을 그래프 라쏘의 입력으로 사용하였는데, 이처럼 그래프 라쏘와 IPW 추정량을 결합함으로써 고차원 결측자료에서 변수들 간

의 조건부 독립성을 추정할 수 있다.

본 연구는 제안 방법의 성능을 기존 방법과 비교 및 평가하고자 한다. 최근 연구에서 Costantini 등(2022)는 일곱 가지 고차원 다중대체 모형의 성능을 비교하였으며 IURR (Indirect Use of Regularized Regression; Deng 등, 2016) 방법, 주성분 분석(principal component analysis)을 이용한 방법, 그리고 전진 선택법(step-forward selection)을 이용한 방법의 성능이 가장 우수하다고 보고하였다. 본 연구에서는 이 중 주성분 분석은 변수 선택이 아닌 차원 축소 방법이므로 주요 비교 대상이 아니라고 간주하였고, 단계적 선택법과 같은 고전적인 모형 선택법은 대체모형에 대해서는 성능이 떨어진다고 언급한 논문(Zhao 등, 2016)이 있었기 때문에 IURR 방법을 주요 비교 방법으로 채택하였다. 또한 IURR과 유사한 방법인 DURR (Direct Use of Regularized Regression; Deng 등, 2016) 방법도 추가로 함께 비교하였다.

본 논문의 나머지 구성은 다음과 같다. 2장에서는 결측자료 메커니즘(missing data mechanism)과 다중대체법을 포함한 몇 가지 결측값 처리 방법들을 소개한다. 3장에서는 MICE 알고리즘 및 MICE 알고리즘에 기반한 DURR과 IURR을 살펴보고, 비교 방법의 단점을 극복하는 제안 방법에 대해 구체적으로 설명한다. 4장에서는 모의실험을 통해 제안 방법과 기존 방법의 성능을 비교하고, 5장에서는 제안 방법의 장·단점을 요약하고 향후 연구 방향에 대해 논의한다.

Ⅱ. 결측의 발생과 처리

결측은 다양한 이유로 발생하며, 결측의 적절한 처리를 위해서는 결측의 발생 요인을 모형화하여 구분할 필요가 있다. Little & Rubin(2002)은 결측의 발생 요인을 세 가지의 결측자료 메커니즘으로 구분하였으며, 오늘날 결측값 처리 방법들은 기본적으로 특정한 결측자료 메커니즘을 가정하고 있다. 따라서 분석자는 사전정보 등을 이용하여 결측의 원인을 파악한 후 적절한 결측자료 메커니즘을 가정하는 것이 결측자료 분석을 위한 첫 번째 단계이다. 본 장의 1절에서 결측자료 메커니즘에 대하여 알아보고, 2절에서 일반적으로 사용되는 몇 가지 결측값 처리 방법들을 소개한다.

1. 결측자료 메커니즘

p개의 변수들과 n개의 개체들로 구성된 자료 행렬 $\boldsymbol{Z}=\left(z_{ij}\right)_{1\le i\le n,1\le j\le p}$와 자료의 결측을 나타내는 지시행렬 $\boldsymbol{R}=\left(r_{ij}\right)_{1\le i\le n,1\le j\le p}$을 고려하자. $\boldsymbol{R}$의 원소인 r_{ij}는 1 또는 0의 값을 갖는 확률변수로, 그 값이 1이면 같은 위치의 확률변수 z_{ij}가 관측됨을, 0이면 결측임을 나타낸다. 행에 해당하는 벡터 $\boldsymbol{z}_i\in\mathbb{R}^p$와 $\boldsymbol{r}_i\in\mathbb{R}^p$는 각각 $i=1,\ldots,n$에 대해 독립적이고 동일한 분포를 따른다고 가정한다. $\boldsymbol{z}_i$의 관측된 원소들을 $\boldsymbol{z}_{i,obs}$, 결측인 원소들을 $\boldsymbol{z}_{i,mis}$로 나타낸다.

모든 i에 대하여, 자료의 결측 여부 $\boldsymbol{r}_i$은 자료의 값 $\boldsymbol{z}_i=\left(\boldsymbol{z}_{i,obs},\boldsymbol{z}_{i,mis}\right)$에 의존할 수도 있고 자료와 관계가 없을 수도 있는데, 이러한 관계를 설명하는 개념이 결측자료 메커니즘이다. 이에 대한 일반화 식은 $\mathrm{P}(\boldsymbol{r}_i|\boldsymbol{z}_i;\boldsymbol{\psi})$이다. 여기서 $\boldsymbol{\psi}$는 $\boldsymbol{r}_i$의 조건부 분포와 관련된 모수 벡터이다. 결측자료 메커니즘은 $\boldsymbol{r}_i$

와 $\boldsymbol{z}_i$의 관계에 따라 다음의 세 가지로 분류된다.

1) 완전임의결측(Missing Completely At Random; MCAR)

완전임의결측은 자료의 결측 여부 $\boldsymbol{r}_i$가 자료 $\boldsymbol{z}_i = (\boldsymbol{z}_{i,obs}, \boldsymbol{z}_{i,mis})$와 무관한 경우로, 일반화 식이 다음과 같이 축소된다.

$$\mathrm{P}(\boldsymbol{r}_i|\boldsymbol{z}_i;\boldsymbol{\psi}) = \mathrm{P}(\boldsymbol{r}_i;\boldsymbol{\psi}) \quad \Leftrightarrow \quad \boldsymbol{r}_i \perp \boldsymbol{z}_i \quad \text{for all } i.$$

예를 들어, 전산 오류로 인해 몇몇 데이터가 무작위로 누락된 경우이다. 이 메커니즘에서 결측은 완전히 무작위로 발생하므로 결측값를 제거하고 분석을 수행하여도 정보의 손실만 있을 뿐, 추정량의 편향이 발생하지 않는다. 하지만 MCAR을 가정하는 것은 강한 가정이며, 실제 자료에서 결측은 많은 경우 자료의 값에 의존하여 발생하기 때문에 현실적인 가정으로 보기는 어렵다(Van Buuren, 2012).

2) 임의결측(Missing At Random; MAR)

임의결측은 자료의 결측 여부 $\boldsymbol{r}_i$가 관측된 자료 $\boldsymbol{z}_{i,obs}$에만 의존하는 경우로, 일반화 식이 다음과 같이 축소된다.

$$\mathrm{P}(\boldsymbol{r}_i|\boldsymbol{z}_i;\boldsymbol{\psi}) = \mathrm{P}(\boldsymbol{r}_i|\boldsymbol{z}_{i,obs};\boldsymbol{\psi}) \quad \Leftrightarrow \quad \boldsymbol{r}_i \perp \boldsymbol{z}_i|\boldsymbol{z}_{i,obs} \quad \text{for all } i.$$

여기서 핵심은 결측인 자료 $\boldsymbol{z}_{i,mis}$에는 의존하지 않는다는 것이다. 예를 들어, 병원 자료에서 나이가 많을수록 몸무게의 결측이 많이 발생한다고 하자. 이

경우 결측값의 추정치로서 나이별 평균 몸무게를 사용할 수 있듯이 관측된 정보(나이)를 통해 결측값(몸무게)의 추정이 가능해지게 된다. MAR 가정은 MCAR 가정보다 일반화된 가정이며 실제 자료에서 더 일어날법한 가정이다(Van Buuren, 2012).

3) 비임의결측(Not Missing At Random; NMAR)

비임의결측은 자료의 결측 여부 $\boldsymbol{r}_i$가 결측인 자료 $\boldsymbol{z}_{i,mis}$에도 의존하는 경우로, 일반화 식이 더 이상 축소되지 않는다.

$$\mathrm{P}(\boldsymbol{r}_i|\boldsymbol{z}_{i,obs},\boldsymbol{z}_{i,mis};\boldsymbol{\psi}) \quad \text{for all } i.$$

이 경우 결측된 정보 자체에 의해 결측이 발생했으므로 관측된 정보만으로는 결측값을 추정할 수 없다. 예를 들어, 병원 자료에서 다른 변수를 통해 몸무게의 추정이 불가능하다고 가정하고, 몸무게가 많이 나갈수록 몸무게의 결측이 많이 발생한다고 하자. 이 경우 결측값을 추정할 수 있는 방법은 존재하지 않는다. NMAR 가정은 가장 일반화된 가정이다.

이어서 결측자료 메커니즘과 관련된 몇 가지 사실들을 소개한다. 먼저, 실제 자료에서 세 가지 메커니즘이 항상 완전히 구분되는 것은 아니며 혼합된 상황이 일반적이다(고길곤·탁현우, 2016). MAR과 NMAR이 혼합된 상황을 예로 들어 보겠다. 병원 자료에서 다른 변수를 통해 몸무게의 추정이 불가능하다고 가정하고, 나이가 많고 몸무게가 많이 나갈수록 몸무게의 결측이 많이 발생한다고 하자. 이 경우 결측값의 추정치로서 나이별 평균 몸무게를 사용하면 나이에 의한 영향은 추정이 되지만, 큰 몸무게에 의한 영향은 추

정되지 않는다.

두 번째로, 실제 자료의 결측이 어느 메커니즘에 의해 발생했는지는 사전 정보가 없다면 알 수 없다. 분석자가 결측된 정보를 파악할 수 없으므로 MAR과 NMAR 가정은 식별이 불가능하며, MCAR 가정만이 경험적으로 검정하는 몇 가지 방법이 존재한다. Little(1988)은 자료가 MCAR을 따르는지 검정하는 방법을 제안하였으나, 이 방법은 낮은 검정력(low power)을 갖는다고 알려져 있다(Baraldi & Enders, 2010에서 재인용). 이처럼 실제 결측을 발생시키는 변수는 알 수 없지만, 결측값의 추정이 불가능한 NMAR 가정을 완화시키는 방법이 있다. 앞선 두 예시에서 체지방률, 허리둘레, 가족력 등 몸무게와 연관이 있는 변수들을 자료에 추가한다면, 다른 변수를 통해 몸무게의 추정이 가능해지게 되면서 몸무게에 의한 결측값까지도 추정해낼 수 있다. 즉 자료가 NMAR 가정을 따른다고 할지라도, 결측 변수와 연관된 잠재적인 요인들을 자료에 포함시키게 된다면 NMAR 가정이 완화될 수 있다(고길곤·탁현우, 2016). 따라서 NMAR 가정이 완화되도록 가능한 한 많은 변수들을 자료에 포함시키는 것이 좋다.

마지막으로, 대부분의 결측값 처리 방법들은 무시할 수 있는(ignorable) MAR을 가정한다. 이는 $\boldsymbol{r}_i$에 대한 모형을 추정하지 않고도 관측된 자료 $\boldsymbol{z}_{i,obs}$만을 기반으로 추정의 대상인 모수 $\boldsymbol{\theta}$[3]의 추론이 가능하도록 해주는 조건이다. 여기서 '무시할 수 있는'이라는 조건은 자료의 분포와 관련된 모수 $\boldsymbol{\theta}$와 결측자료 메커니즘과 관련된 모수 $\boldsymbol{\psi}$가 별개(distinct)라는 것을 의미한다. 즉, $(\boldsymbol{\theta}, \boldsymbol{\psi})$의 결합모수공간 $\Omega_{\boldsymbol{\theta},\boldsymbol{\psi}}$가 각각의 모수공간의 곱 $\Omega_{\boldsymbol{\theta}} \times \Omega_{\boldsymbol{\psi}}$으로 나타내어지는 경우이다. 이러한 경우 $\boldsymbol{\theta}$의 추정을 위해 $\boldsymbol{\psi}$를 고려할 필요가 없다는 것이 알려져 있다(Little & Rubin, 2002; 송주원·안형진, 2009).

---

3) 이어지는 2.3절과 3장에서 추정의 대상인 모수는 결측 변수의 조건부 분포와 관련된 모수(혹은 변수)를 나타낸다.

2. 결측값 처리 방법

결측자료에서 분석자의 관심 과제는 변수 간의 상관성을 추정하는 것, 평균이나 분산과 같은 자료의 분포를 대표하는 모수를 추정하는 것, 혹은 회귀계수 같은 통계 모형의 모수를 추정하는 것 등으로 다양하다. 이러한 분석의 관심 모수를 Q라고 하자. 자료의 결측은 정보의 손실로 인해 Q의 추정량 $\hat{Q}$에 대한 편향 및 변동성(variation)을 더 크게 만들기 때문에 이를 최대한 방지하는 적절한 결측값 처리 방법을 선택하는 것이 중요하다. 본 절에서는 기존의 결측값 처리 방법들을 소개하고 전술한 측면에서 다중대체법이 가지는 장점에 대해 설명한다.

1) 완전 제거법(Complete Case Analysis; CCA)

결측을 처리하는 가장 간단한 방법은 결측이 발생한 개체를 삭제하고 분석을 진행하는 방법이다. 이는 적용이 매우 간단하지만 결측자료 메커니즘이 MCAR이 아니면 구하고자 하는 모수의 추정치에 편향이 발생하며, 단 하나의 변수에만 결측이 발생하더라도 그 행 자체를 삭제하기 때문에 정보의 손실이 매우 크다. 특히 개체의 수보다 변수의 수가 많은 고차원 자료에서 더욱 그러하다. 따라서 CCA 방법은 분석의 신뢰성을 떨어뜨릴 위험이 크기 때문에 결측을 가진 개체의 비율이 굉장히 낮지 않은 한 추천되지 않는다.

2) 우도에 근거한 추정 방법(likelihood based approach)

자료에 대한 구체적인 통계적 분포를 가정할 경우 해당 분포의 우도 함수

에 근거한 최대우도 추정량으로 모수를 추정할 수 있다. 자료에 결측이 있는 경우에도 관측된 자료에 근거하여 우도 함수를 유도할 수 있으나 그 형태가 매우 복잡하며 결측 패턴(missingness pattern)에 따라 우도 함수가 달라지게 된다. 따라서 최대우도 추정량을 폐쇄형(closed form)으로 나타낼 수 없어 EM 알고리즘과 같은 수치적 방법을 통해 추정량을 계산한다(송주원·안형진, 2009). 이 접근법은 분포 가정이 필요하고 모수 추정량으로 최대우도 추정량만을 사용할 수 있다는 한계가 존재한다.

3) 대체방법(imputation methods)

이를 보완하고자 일반적으로 많이 사용되는 방법이 결측값 대체방법이다. 이 방법은 관측된 정보들로 결측값에 대한 '그럴듯한' 값을 예측하여 대체함으로써, 결측이 없는 유사완전(pseudo-complete) 형태인 대체 데이터 세트(imputed data set)를 생성한 후, 기존에 존재하는 통계 모형을 적용하여 분석을 수행한다. 우도에 근거한 추정 방법과 달리 대체방법은 대체 데이터 세트를 마치 완전한[4] 데이터 세트를 가진 것처럼 취급할 수 있으므로, 여기에 기존에 존재하는 통계 분석 방법론들을 곧바로 적용할 수 있다는 장점이 존재한다. 대체방법은 CCA 방법과 달리 MAR 가정하에서 관심 모수의 불편추정량을 얻을 수 있기 때문에, 기본적으로 무시할 수 있는 MAR을 가정한다(Little & Rubin, 2002).

대체방법은 대체 데이터 세트를 생성하는 횟수에 따라 단일대체(single imputation)와 다중대체(multiple imputation)로 나뉜다. 단일대체는 단 한 개의 대체 데이터 세트를 생성하는 방법으로 대표적으로 연속형 변수의 결측값을 해당 변수의 평균으로 대체하는 평균대체(mean imputation)가 있다.

4) 결측값 대체 분야에서 '완전한(complete)'이라는 용어는 '결측이 없는' 상태를 나타낸다.

하지만 단일대체는 결측값의 불확실성(uncertainty)을 전혀 고려하지 못한다는 단점이 존재한다. 결측자료 분석에서는 결측된 값이 실제로 어떤 값인지 모르기 때문에 이 불확실성이 분석 결과에 반영되는 것이 바람직하다. 따라서 완전히 관측된 데이터 세트(completely observed data set)에서보다 결측 데이터 세트에서 구한 추정량의 분산이 더 크게 추정되어야 한다. 하지만 단일대체는 단 한 번 대체된 값을 마치 관측값처럼 취급하므로 추정량의 분산이 과소추정되고 신뢰구간이 좁아지는 등의 문제가 발생한다(송주원·안형진, 2009).

단일대체의 문제점을 해결하기 위해 Rubin(1978)은 여러 개의 대체 데이터 세트를 생성하는 다중대체를 제안하였다. 다중대체의 과정은 다음의 세 단계로 이루어진다.

(1) 대체 단계: 적절한 대체모형을 선택하여 $M(M>1)$개의 대체 데이터 세트를 생성한다.

(2) 분석 단계: 각 대체 데이터 세트에 대하여 개별적인 통계분석을 수행한다. 이때 관심 모수 Q에 대하여 $m(m=1,...,M)$번째 대체 데이터 세트에 계산된 점 추정량을 Q_m라 하고, 추정량 Q_m의 분산을 U_m으로 표기하자.

(3) 결합 단계: Rubin's rule으로 알려진 다음식에 의하여 m개의 추정량을 결합함으로써 최종추정량을 정의한다.

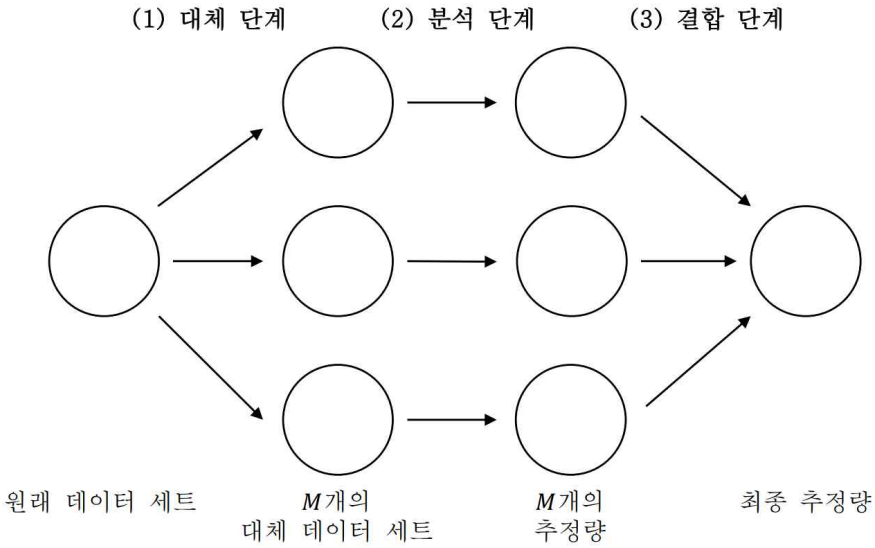
$$\overline{Q}=\frac{1}{M}\sum_{m=1}^{M}Q_m, \tag{2.1}$$

$$T=\overline{U}+\left(1+\frac{1}{M}\right)B,\ \text{이때}\ \overline{U}=\frac{1}{M}\sum_{m=1}^{M}U_m,\ B=\frac{1}{M-1}\sum_{m=1}^{M}\left(Q_m-\overline{Q}\right)^2, \tag{2.2}$$

$$\overline{Q}\pm t_v(\alpha/2)\sqrt{T},\ \text{이때}\ v=\frac{M-1}{\left(T-\overline{U}/T\right)^2}. \tag{2.3}$$

여기서 결합된 모수의 최종추정량 $\overline{Q}$, 결합된 추정량의 최종분산 T, $\overline{Q}$의 $(1-\alpha)100\%$ 신뢰구간 $\overline{Q} \pm t_v(\alpha/2)\sqrt{T}$은 각각 식 (2.1), 식 (2.2), 식 (2.3)에 해당한다. 식 (2.2)에서 확인할 수 있듯이 다중대체는 추정량의 분산인 대체 내 분산 $\overline{U}$에 대체 간 분산 B까지 합하여 최종적인 추정량의 분산을 정의함으로써 분산이 과소추정되는 문제를 해결한다. [그림 1]은 다중대체의 과정을 도식화한 것이다.

Van Buuren(2012)은 더 큰 M을 설정할수록 최종추정량이 더 정확해지지만 $M=5$이상에서는 값의 변화가 크지 않다고 하였다. 따라서 분석자는 계산과 저장 공간을 고려하여 5 이상의 적당한 M을 설정하는 것이 추천된다.



[그림 1] 다중대체의 과정

다중대체의 단계 중 (1) 대체 단계에서 M개의 대체 데이터 세트를 생성하는 방법에 대해 구체적으로 알아보겠다. 자료 행렬 $\mathcal{Z}\in\mathbb{R}^{n\times p}$의 관측된 원소들을 $\mathcal{Z}_{obs}$, 결측인 원소들을 $\mathcal{Z}_{mis}$로 나타낸다. 다중대체법의 목표는 관측값 $\mathcal{Z}_{obs}$이 주어졌을 때 결측값 $\mathcal{Z}_{mis}$의 조건부 예측 분포[5] $\mathrm{P}(\mathcal{Z}_{mis}|\mathcal{Z}_{obs})$를 추정하여 그로부터 여러 번 랜덤 추출한 $\mathcal{Z}_{mis}$의 예측값을 대체값으로 사용하는 것이다. $\mathrm{P}(\mathcal{Z}_{mis}|\mathcal{Z}_{obs})$를 풀어서 표현하면 다음과 같다.

$$\mathrm{P}(\mathcal{Z}_{mis}|\mathcal{Z}_{obs})=\int \mathrm{P}(\mathcal{Z}_{mis}|\mathcal{Z}_{obs},\theta)\mathrm{P}(\theta|\mathcal{Z}_{obs})d\theta. \tag{2.4}$$

여기서 $\mathrm{P}(\mathcal{Z}_{mis}|\mathcal{Z}_{obs},\theta)$는 $\mathcal{Z}_{obs}$와 θ가 주어졌을 때 $\mathcal{Z}_{mis}$의 조건부 예측 분포, $\mathrm{P}(\theta|\mathcal{Z}_{obs})$는 $\mathcal{Z}_{obs}$이 주어졌을 때 θ의 사후분포이며, θ는 모집단의 특성을 나타내는 변수이다. 식 (2.4)를 이용하여 $\mathrm{P}(\mathcal{Z}_{mis}|\mathcal{Z}_{obs})$에서 $m\,(m=1,\ldots,M)$번째 대체값을 생성하는 구체적인 과정은 다음과 같다.

[알고리즘 1] 다중대체 알고리즘

Step 1) $\hat{\theta}^{(m)}\sim\mathrm{P}(\theta|\mathcal{Z}_{obs})$

Step 2) $\mathcal{Z}_{mis}^{(m)}\sim\mathrm{P}(\mathcal{Z}_{mis}|\mathcal{Z}_{obs},\hat{\theta}^{(m)})$

먼저 추정된 $\mathrm{P}(\theta|\mathcal{Z}_{obs})$로부터 $\hat{\theta}^{(m)}$을 추출한 후, 이를 조건에 대입한 $\mathrm{P}(\mathcal{Z}_{mis}|\mathcal{Z}_{obs},\hat{\theta}^{(m)})$로부터 $\mathcal{Z}_{mis}^{(m)}$을 추출한다. 이렇게 생성된 표본 $\mathcal{Z}_{mis}^{(1)},\ldots,\mathcal{Z}_{mis}^{(M)}$가 결측값 $\mathcal{Z}_{mis}$에 대한 대체값이 된다. 여기서 분포 $\mathrm{P}(\theta|\mathcal{Z}_{obs})$를 획득하기 위해

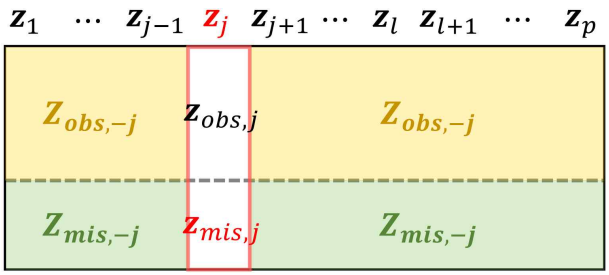
5) 예측 분포란 예측 모형에서 주어진 입력에 대한 예측된 결과의 확률분포로, 모형의 불확실성을 내포한다.

서는 $P(\boldsymbol{Z}|\boldsymbol{\theta})$를 위한 통계적 모형을 가정해야 한다. 많은 경우, $\boldsymbol{Z}$가 다변량 정규분포를 따른다고 가정한다(Zhao 등, 2016).

대체값 $\boldsymbol{Z}_{mis}^{(m)}$을 동시에 생성하는지, 각 변수별로 생성하는지에 따라 다중대체를 수행하는 방법은 크게 JM (Joint Modeling) 접근법과 FCS (Fully Conditional Specification; Van Buuren 등, 2006) 접근법으로 나뉜다. JM 접근법은 기본적으로 식 (2.4)에 근거하여 $\boldsymbol{Z}_{mis}$의 결측 원소들을 결합확률분포로부터 동시에 생성하는 방식으로 다변량 분포의 가정이 필요하며, 대체를 위해 보통 MCMC (Markov Chain Monte Carlo)와 같은 베이지안 통계 방법을 사용한다. 이는 자료가 가정한 분포를 따르면 이론적 정당성을 갖지만, 실제 자료가 가정한 분포를 정확히 따르기란 불가능하고 차원이 증가하면 결합확률분포의 추정이 어려워 일반적으로는 사용되지 않는다(김현태·장가영, 2020; Deng 등, 2016). 실제로 다중대체를 수행하는 데 흔히 사용되는 접근법은 각 변수별로 일변량 조건부 분포를 반복적으로 추정하여 대체값을 생성하는 FCS 접근법으로, 다음 장에서 자세히 알아본다.

Ⅲ. 방법론

먼저, 본 장에서 사용될 기호를 재정의하겠다. 데이터 구조와 관련하여 [그림 2]를 참고한다. 자료 행렬 $\mathcal{Z}=(z_{ij})_{1\le i\le n,1\le j\le p}$의 열에 해당하는 각 변수를 벡터 $\boldsymbol{z}_j\in\mathbb{R}^n\,(j=1,\ldots,p)$로 나타내자. 일반성을 잃지 않고, p개의 변수 중 첫 $l\,(l\le p)$개의 변수에서 결측이 발생함을 가정한다. 결측 변수 $\boldsymbol{z}_j\,(j=1,\ldots,l)$에 대하여 관측된 원소 o_j개를 벡터 $\boldsymbol{z}_{obs,j}\in\mathbb{R}^{o_j}$, 결측인 원소 $n-o_j$개를 벡터 $\boldsymbol{z}_{mis,j}\in\mathbb{R}^{n-o_j}$로 나타낸다. 변수 $\boldsymbol{z}_j$를 제외한 나머지 변수들을 행렬 $\mathcal{Z}_{-j}=[\boldsymbol{z}_1,\ldots,,\boldsymbol{z}_{j-1},\boldsymbol{z}_{j+1},\ldots,\boldsymbol{z}_p]\in\mathbb{R}^{n\times(p-1)}$로 표현할 때, $\mathcal{Z}_{-j}$에서 $\boldsymbol{z}_{obs,j}$와 같은 행에 위치하는 블록 행렬을 $\mathcal{Z}_{obs,-j}\in\mathbb{R}^{o_j\times(p-1)}$, $\boldsymbol{z}_{mis,j}$와 같은 행에 위치하는 블록 행렬을 $\mathcal{Z}_{mis,-j}\in\mathbb{R}^{(n-o_j)\times(p-1)}$로 표기한다.



[그림 2] 데이터 구조와 기호

본 장의 1절에서는 현실에서 다중대체를 가능하게 하는 FCS 접근법을 소개하고, 2절에서 본 논문이 제안하는 새로운 변수 선택법에 대하여 설명한다. 제안 방법은 JM이 아닌 FCS 접근법일 때 작동한다.

1. FCS (Fully Conditional Specification) 접근법

FCS 접근법은 $m=1,...,M$에 대하여 대체값 $\mathcal{Z}_{mis}^{(m)}=\left(\boldsymbol{z}_{mis,1}^{(m)},...,\boldsymbol{z}_{mis,l}^{(m)}\right)$을 동시에 생성하는 대신 $\boldsymbol{z}_{mis,1}^{(m)}$부터 시작하여 $\boldsymbol{z}_{mis,l}^{(m)}$로 한 변수씩 순차적으로 대체값을 생성한다. 이를 위한 FCS의 목표는 각 변수의 일변량 조건부 분포

$$\mathrm{P}(\boldsymbol{z}_j|\mathcal{Z}_{-j}),\ j=1,...,l, \tag{3.1}$$

를 추정하는 것이다. FCS 접근법은 변수별로 각기 다른 대체모형을 적용할 수 있으며, 자료 $\mathcal{Z}$에 대한 다변량 분포의 가정이 필요하지 않다. 이 방법은 통계적 특성을 확립하기는 어렵지만 JM 접근법보다 유연하고 현실적이며 경험적으로 양호한 성능을 보여왔다(Deng 등, 2016).

이어지는 1.1절에서 MICE 알고리즘에 대하여 설명한 후, 1.2절에서 고차원 환경에서 MICE가 작동하기 위하여 변수 선택이 필요함을 설명한다. 1.3절과 1.4절에서는 기존의 고차원 대체모형인 DURR과 IURR의 변수 선택법을 알아본다.

1) MICE (Multiple Imputation by Chained Equations) 알고리즘

MICE 알고리즘은 FCS 접근법을 구현한 알고리즘 중 가장 대표적인 알고리즘으로, 작동하는 방식은 다음과 같다. 먼저, $\boldsymbol{z}_1$에서 시작하여 $\boldsymbol{z}_l$까지 결측인 부분을 평균값 등으로 초기 대체한다. 초기 대체된 벡터들을 $\left(\tilde{\boldsymbol{z}}_1^{(0)},...,\tilde{\boldsymbol{z}}_l^{(0)}\right)$로 나타낸다. 그리고 다시 $\boldsymbol{z}_1$으로 돌아가서, 주어진 나머지 변수들 $\tilde{\mathcal{Z}}_{-1}^{(0)}=\left[\tilde{\boldsymbol{z}}_2^{(0)},...,\tilde{\boldsymbol{z}}_l^{(0)},\boldsymbol{z}_{l+1},...,\boldsymbol{z}_p\right]$로 $\boldsymbol{z}_1$의 결측인 부분($\boldsymbol{z}_{mis,1}$)을 예측해 대체값

을 업데이트 한다. 같은 방식으로 z_1부터 z_l까지 차례대로 대체값을 업데이트하고, 다시 z_1으로 돌아가는 것을 반복(iteration)한다. t번째 반복에서 대체된 벡터들을 $\left(\tilde{z}_1^{(t)}, ..., \tilde{z}_l^{(t)}\right)$로 표기한다. 대체값들이 수렴하여 반복이 종료될 때까지의 반복 횟수를 T라고 할 때 최종 대체된 벡터들 $\left(\tilde{z}_1^{(T)}, ..., \tilde{z}_l^{(T)}\right)$가 다중대체의 $m(m=1,...,M)$번째 대체 $\left(z_1^{(m)}, ..., z_l^{(m)}\right)$가 된다. 전체 과정을 M번 수행하여 다중 대체된 $\left(z_1^{(1)}, ..., z_l^{(1)}\right), ..., \left(z_1^{(M)}, ..., z_l^{(M)}\right)$를 얻으면 이들 각각은 $\mathrm{P}(z_1|Z_{-1}), ..., \mathrm{P}(z_l|Z_{-l})$로부터의 랜덤 추출로 취급될 수 있다. 반복 횟수 T에 대하여, 이론적으로는 수렴할 때까지 반복해야 하지만 소프트웨어 구현에서는 계산 효율을 위해 T를 사전에 지정한다. 일반적인 상황에서 5에서 10 정도의 작은 수로도 알고리즘이 잘 수렴한다고 알려져 있다(Van Buuren, 2012).

$t\,(t=1, ..., T)$번째 반복에서 $j\,(j=1, ..., l)$번째 결측 변수 z_j의 대체값 $\tilde{z}_{mis,j}^{(t)}$은 다음 식으로부터의 랜덤 추출이다.

$$\mathrm{P}\left(z_{mis,j}|z_{obs,j}, \tilde{Z}_{-j}^{(t-1)}\right) = \int \mathrm{P}\left(z_{mis,j}|\tilde{Z}_{mis,-j}^{(t-1)}, \theta_j\right)\mathrm{P}\left(\theta_j|z_{obs,j}, \tilde{Z}_{obs,-j}^{(t-1)}\right)d\theta_j. \quad (3.2)$$

여기서 행렬 $\tilde{Z}_{-j}^{(t-1)} = \left[\tilde{z}_1^{(t)}, ..., \tilde{z}_{j-1}^{(t)}, \tilde{z}_{j+1}^{(t-1)}, ..., \tilde{z}_l^{(t-1)}, z_{l+1}, ..., z_p\right]$는 t번째 반복에서 z_j를 제외한 나머지 변수로 구성된 유사완전 형태이다. $\tilde{Z}_{obs,-j}^{(t-1)}$는 $\tilde{Z}_{-j}^{(t-1)}$에서 $z_{obs,j}$와 같은 행에 위치하는 블록 행렬, $\tilde{Z}_{mis,-j}^{(t-1)}$는 $\tilde{Z}_{-j}^{(t-1)}$에서 $z_{mis,j}$와 같은 행에 위치하는 블록 행렬이다. z_j에 대해서 결측인 부분은 $Z_{mis} = \{z_{mis,j}\}$, 관측된 부분은 $Z_{obs} = \left\{z_{obs,j}, \tilde{Z}_{obs,-j}^{(t-1)}, \tilde{Z}_{mis,-j}^{(t-1)}\right\}$이므로 식 (2.4)를 재표현하면 식 (3.2)가 된다. 식 (3.2)의 θ_j는 z_j의 조건부 분포와 관련된 변수로 식 (2.4)의

$\boldsymbol{\theta}$와는 별개이다(Deng 등, 2016).

식 (3.2)를 이용하여 대체값을 생성하는 구체적인 과정은 다음과 같다.

[알고리즘 2] MICE 알고리즘

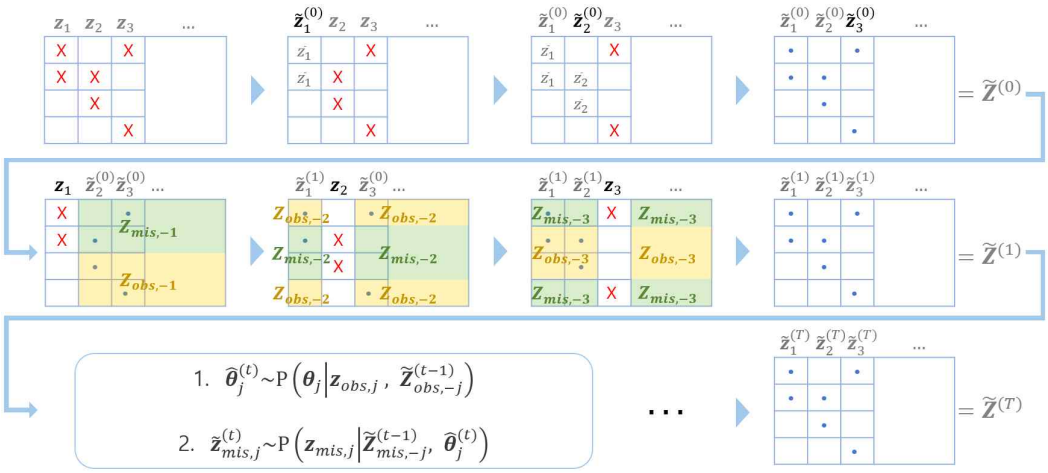
$t\,(t=1,...,T)$번째 반복, $j\,(j=1,...,l)$번째 변수에 대하여 다음을 반복적으로 수행:
Step 1) $\hat{\boldsymbol{\theta}}_j^{(t)} \sim \mathrm{P}\left(\boldsymbol{\theta}_j \mid \boldsymbol{z}_{obs,j}, \tilde{\boldsymbol{Z}}_{obs,-j}^{(t-1)}\right)$
Step 2) $\tilde{\boldsymbol{z}}_{mis,j}^{(t)} \sim \mathrm{P}\left(\boldsymbol{z}_{mis,j} \mid \tilde{\boldsymbol{Z}}_{mis,-j}^{(t-1)}, \hat{\boldsymbol{\theta}}_j^{(t)}\right)$

먼저 추정된 $\mathrm{P}\left(\boldsymbol{\theta}_j \mid \boldsymbol{z}_{obs,j}, \tilde{\boldsymbol{Z}}_{obs,-j}^{(t-1)}\right)$로부터 $\hat{\boldsymbol{\theta}}_j^{(t)}$을 추출한 후, 이를 조건에 대입한 $\mathrm{P}\left(\boldsymbol{z}_{mis,j} \mid \tilde{\boldsymbol{Z}}_{mis,-j}^{(t-1)}, \hat{\boldsymbol{\theta}}_j^{(t)}\right)$로부터 대체값 $\tilde{\boldsymbol{z}}_{j,mis}^{(t)}$을 추출한다. 식 (3.2)는 식 (2.4)와 달리 추정하고자 하는 조건부 분포가 일변량 분포이므로 $\boldsymbol{z}_j$를 타겟변수로 하는 지도학습모형을 대체모형으로 이용할 수 있다. 이 경우 Step 1은 예측변수 $\tilde{\boldsymbol{Z}}_{obs,-j}^{(t-1)}$로부터 타겟변수 $\boldsymbol{z}_{obs,j}$를 학습하는 과정이 된다. 추정된 대체모형의 모수를 $\hat{\boldsymbol{\theta}}_j^{(t)}$라 하면, 이는 $\mathrm{P}\left(\boldsymbol{\theta}_j \mid \boldsymbol{z}_{obs,j}, \tilde{\boldsymbol{Z}}_{obs,-j}^{(t-1)}\right)$로부터의 랜덤 추출로 간주될 수 있다. Step 2는 학습된 대체모형을 사용하여 $\tilde{\boldsymbol{Z}}_{mis,-j}^{(t-1)}$로부터 $\boldsymbol{z}_{mis,j}$를 예측하는 과정이 된다. 예를 들어 대체모형으로 선형회귀모형을 상정한 경우, MICE 알고리즘을 다음과 같이 구체화할 수 있다(Zhao 등, 2016).

Step 1) $\boldsymbol{z}_{obs,j} = \alpha_0 + \tilde{\boldsymbol{Z}}_{obs,-j}^{(t-1)}\boldsymbol{\alpha} + \epsilon$로부터 회귀계수 $\hat{\boldsymbol{\theta}}_j = \left(\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}\right)^T$를 추정한다.

Step 2) $\tilde{\boldsymbol{z}}_{mis,j}^{(t)} = \hat{\alpha}_0 + \tilde{\boldsymbol{Z}}_{mis,-j}^{(t-1)}\hat{\boldsymbol{\alpha}}$을 통해 대체값을 생성한다.

[그림 3]은 MICE 알고리즘을 도식화한 것이다. 그림에는 편의를 위해 세 개의 결측 변수만 나타내었지만, 결측이 발생한 $l(l \leq p)$개의 변수로 확장하여 생각한다.



[그림 3] MICE 알고리즘의 도식화

MICE 알고리즘은 다양한 통계 소프트웨어를 통해 구현되어있다. 그 중 가장 널리 사용되는 R의 'mice' 패키지는 연속형 변수에 대해서는 PMM (Predictive Mean Matching), 범주형 변수에 대해서는 로지스틱 회귀모형을 기본(default) 대체모형으로 제공하고 있으며, 그 외에도 포아송 회귀나 베이지안 선형회귀 등 다양한 지도학습모형을 대체모형으로 제공한다. MICE의 기본 대체모형으로 널리 사용되는 PMM은 4장의 모의실험에서 저차원 환경의 비교 방법으로 사용된다. 이는 기본적으로 선형회귀모형을 적합시켜 예측값을 구한 뒤, 예측값 주변의 실제값들(true values) 중 하나를 랜덤하게 선택하여 대체값으로 사용하는 모형이다.

2) 고차원 환경

고차원 환경에서 변수 선택의 필요성과 목표에 대하여 설명하겠다. 원래 대체 시에는 결측값 예측에 도움이 되는 중요한 변수들을 빠뜨리지 않도록 가능한 한 많은 변수를 모형에 포함시키는 것이 좋다고 알려져있다(Meng, 1994; Collins 등, 2001). 실제 결측자료 메커니즘이 MAR이 아니더라도 중요한 변수가 자료에 포함되면 결측값의 예측이 가능해진다. 다시 말해, NMAR 가정을 완화시켜 자료가 MAR 가정을 따르는 것 같은 효과를 가져온다. 따라서 대체모형에 모든 변수를 포함시키는 것이 일반적이다.

결측자료 $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$가 $n > p$로 저차원일 때는 모든 변수를 사용하더라도 대체모형이 문제없이 잘 학습된다. 하지만 저차원에서 개발된 대부분의 대체모형들은 $n \leq p$인 고차원 상황에서는 제대로 학습이 이루어지지 않는 '차원의 저주' 문제를 겪는다. 즉, [알고리즘 2]의 Step 1에서 다중공선성 등으로 $\hat{\boldsymbol{\theta}}_j^{(t)}$가 식별되지 않는다. 따라서 고차원 상황에서 대체를 위해서는 결측의 예측에 도움이 되지 않는 변수들은 걸러내고, 도움이 되는 중요한 변수만을 선택하는 균형 잡힌 변수 선택이 필요하다.

FCS에서 결측 변수 $\boldsymbol{z}_j (j = 1, ..., l)$를 예측하는 것의 근본적인 아이디어는 식 (3.1)의 조건부 분포 $\mathrm{P}(\boldsymbol{z}_j | \boldsymbol{Z}_{-j})$를 추정하는 것이므로 $\boldsymbol{z}_j$의 예측에 필요 없는 변수들은 결국 $\boldsymbol{z}_j$와 조건부 독립인 변수들이다. 여기서 조건부 독립이란, 두 변수를 제외한 나머지 변수들 $\boldsymbol{Z}_{-\{j,k\}}$이 주어졌을 때 $\boldsymbol{z}_j$와 독립인 $\boldsymbol{z}_j \perp \boldsymbol{z}_k | \boldsymbol{Z}_{-\{j,k\}}$의 관계를 만족하는 변수 $\boldsymbol{z}_k (k \neq j)$로 정의된다. 즉 고차원 대체모형의 목표는 다음의 조건부 분포를 추정하는 것이다.

$$\mathrm{P}(\boldsymbol{z}_j | \boldsymbol{Z}_{S_j}), \ j = 1, ..., l. \tag{3.3}$$

여기서 집합의 크기가 p보다 훨씬 작은 S_j는 $\boldsymbol{z}_j$와 조건부 의존인 변수들의

집합을 가리킨다. 이를 본 논문에서는 실제 중요 변수 집합(true active set; Zhao 등, 2016)이라 부르겠다.[6] 따라서 S_j를 추정하는 과정은 곧 변수 선택 과정이 되고, 추정된 중요 변수 집합 $\hat{S}_j$이 $\boldsymbol{z}_j$의 예측을 위한 대체모형에 사용된다. 본 논문에서는 S_j을 추정하기 위한 새로운 변수 선택법을 제안한다. 기존의 고차원 대체모형인 DURR과 IURR은 벌점화 회귀를 사용해 식 (3.3)을 추정한다. 이어지는 절에서 두 방법에 대해 간단하게 살펴본 후, 이들의 한계를 극복하는 제안 방법을 소개한다.

3) DURR (Direct Use of Regularized Regression)

DURR의 기본 아이디어는 MICE 알고리즘의 대체모형으로 벌점화 회귀모형을 사용하는 것으로, 실제 중요 변수 집합 S_j와 모수 $\boldsymbol{\theta}_j$의 추정이 동시에 이루어진다. DURR의 알고리즘은 다음과 같다.

[알고리즘 3] DURR 알고리즘

MICE의 $t\,(t=1,...,T)$번째 반복, $j\,(j=1,...,l)$번째 변수에 대하여 다음을 반복적으로 수행:

Step 1) 현재 상태의 데이터 세트 $\left[\boldsymbol{z}_j, \tilde{\boldsymbol{Z}}_{-j}^{(t-1)}\right]$에서 n개의 표본을 랜덤하게 복원 추출하여 붓스트랩(bootstrap) 데이터 세트 $\left[\boldsymbol{z}_j^*, \tilde{\boldsymbol{Z}}_{-j}^{*(t-1)}\right]$를 구성한다.

Step 2) 벌점화 회귀모형을 사용하여, $\hat{\boldsymbol{\theta}}_j^{(t)} \sim \mathrm{P}(\boldsymbol{\theta}_j | \boldsymbol{z}_{obs,j}^*, \tilde{\boldsymbol{Z}}_{obs,-j}^{*(t-1)})$

Step 3) $\tilde{\boldsymbol{z}}_{mis,j}^{(t)} \sim \mathrm{P}(\boldsymbol{z}_{mis,j} | \tilde{\boldsymbol{Z}}_{mis,-j}^{(t-1)}, \hat{\boldsymbol{\theta}}_j^{(t)})$

---

6) 엄밀히 말하면, S_j는 '실제 중요 변수 집합을 나타내는 색인(index) 집합'이지만 색인 집합과 변수 집합을 구분하지 않아도 이해하는 데 혼란이 없어 본 논문에서는 구분하지 않고 사용한다.

Step 1에서 행해지는 붓스트랩 단계는 다중대체의 불확실성을 부여하기 위함이다. Step 2에서는 $\boldsymbol{z}^{*}_{obs,j}$를 타겟변수, $\tilde{\boldsymbol{Z}}^{*(t-1)}_{obs,-j}$를 예측변수로 간주하고 벌점화 회귀모형을 학습한다. $\boldsymbol{z}^{*}_{obs,j}$는 $\boldsymbol{z}^{*}_{j}$ 중 관측값에 해당하는 원소들이고, $\tilde{\boldsymbol{Z}}^{*(t-1)}_{obs,-j}$는 행렬 $\tilde{\boldsymbol{Z}}^{*(t-1)}_{-j}$에서 $\boldsymbol{z}^{*}_{obs,j}$와 같은 행에 위치하는 블록 행렬이다. 추정된 회귀계수 $\hat{\boldsymbol{\theta}}^{(t)}_{j}$에서 0으로 축소되지 않은 원소에 대응되는 변수 집합이 추정된 중요 변수 집합 $\hat{S}^{(t)}_{j}$가 된다. Step 3에서 $\hat{\boldsymbol{\theta}}^{(t)}_{j}$를 예측변수 $\tilde{\boldsymbol{Z}}^{(t-1)}_{mis,-j}$에 적용하여 대체값 $\tilde{\boldsymbol{z}}^{(t)}_{mis,-j}$를 생성한다. 여기서 $\tilde{\boldsymbol{Z}}^{(t-1)}_{mis,-j}$는 붓스트랩 데이터 세트가 아닌 기존의 데이터 세트에 해당한다.

4) IURR (Indirect Use of Regularized Regression)

IURR은 DURR과 달리 실제 중요 변수 집합 S_j의 추정을 위해서만 벌점화 회귀모형를 사용한다. 이 같은 측면에서 벌점화 회귀모형을 간접적(indirect)으로 이용한다고도 한다. IURR의 알고리즘은 다음과 같다.

[알고리즘 4] IURR 알고리즘

MICE의 $t\,(t=1,...,T)$번째 반복, $j\,(j=1,...,l)$번째 변수에 대하여 다음을 반복적으로 수행:

Step 1) 학습 데이터 세트 $\left[\boldsymbol{z}_{obs,j}, \tilde{\boldsymbol{Z}}^{(t-1)}_{obs,-j}\right]$에서 벌점화 회귀모형을 학습하여 선택된 변수 집합 $\hat{S}^{(t)}_{j}$를 얻는다.

Step 2) $\hat{\boldsymbol{\theta}}^{(t)}_{j} \sim \mathrm{P}\left(\boldsymbol{\theta}_j \mid \boldsymbol{z}_{obs,j}, \tilde{\boldsymbol{Z}}^{(t-1)}_{obs,\hat{S}^{(t)}_{j}}\right)$

Step 3) $\tilde{\boldsymbol{z}}^{(t)}_{mis,j} \sim \mathrm{P}\left(\boldsymbol{z}_{mis,j} \mid \tilde{\boldsymbol{Z}}^{(t-1)}_{mis,\hat{S}^{(t)}_{j}}, \hat{\boldsymbol{\theta}}^{(t)}_{j}\right)$

Step 2에서 $\tilde{Z}^{(t-1)}_{obs,\hat{S}_j^{(t)}}$는 행렬 $\tilde{Z}^{(t-1)}_{obs,-j}$에서 선택된 변수 $\hat{S}_j^{(t)}$만으로 이루어진 블록 행렬이고, Step 3에서 $\tilde{Z}^{(t-1)}_{mis,\hat{S}_j^{(t)}}$는 행렬 $\tilde{Z}^{(t-1)}_{mis,-j}$에서 선택된 변수 $\hat{S}_j^{(t)}$만으로 이루어진 블록 행렬이다. IURR은 Step 1에서 변수 선택으로 차원이 축소되기 때문에, Step 2와 3에서 기존의 저차원 대체모형들도 적용이 가능하다는 장점이 있다. 하지만 DURR과 IURR 모두 MICE의 반복마다 벌점화 회귀모형을 학습함으로써 S_j를 '반복적으로' 추정한다. 따라서 알고리즘의 계산 시간이 오래 걸린다는 한계점이 있다.

2. 제안 방법

제안하는 방법은 그래프 라쏘와 IPW 추정량을 사용하여 S_j를 '한 번에' 추정한다. 이를 위해 본 연구는 그래프 이론을 사용하는 시도를 하였다. 그래프 이론 분야에서는 변수들 간의 조건부 독립성을 추정하는 방법들이 존재하는데, 그래프 라쏘는 그 중 대표적인 알고리즘이다. 그래프 라쏘 알고리즘은 정규성 가정하에서 작동하며, 입력값으로 표본 공분산 행렬을 사용한다. 완전한 자료에서 불편 추정량인 표본 공분산 행렬은 결측자료에서 편의 추정량이 되기 때문에, 이를 입력으로 사용하면 변수들 간의 조건부 독립성이 제대로 추정되지 않는다. 본 연구에서는 결측자료에서 공분산 행렬의 불편추정량인 IPW 추정량을 사용하여 결측자료에서도 조건부 독립성의 추정이 가능하도록 한다. 결과적으로, 이러한 방식은 MICE 알고리즘 내에서 S_j를 반복적으로 추정할 필요가 없이 사전에 변수 선택을 가능하게 하여 알고리즘의 계산 시간을 효율적으로 단축시킨다.

이어지는 2.1절과 2.2에서는 그래프 라쏘와 IPW 추정량에 대해 각각 자세히 설명한 후, 마지막 2.3절에서 제안 방법의 알고리즘을 정리한다.

1) 그래프 라쏘(graphical lasso)

본 절에서는 임의의 p차원 확률벡터를 $(X_1, ..., X_p)^T$로, 이 확률벡터의 공분산 행렬을 $\boldsymbol{\Sigma} = (\sigma_{jk})_{1 \le j,k \le p}$로, $\boldsymbol{\Sigma}$의 역행렬을 의미하는 정밀행렬(precision matrix)을 $\boldsymbol{\Omega} = (\omega_{jk})_{1 \le j,k \le p}$로 표기한다.

그래프 이론에서 무방향 그래프 모형(undirected graphical models)은 확률변수 간의 조건부 독립 관계를 네트워크 그래프(network graph)로 나타내는 확률 모형이다. $G=(V,E)$로 표기되는 무방향 그래프는 노드(nodes)와 엣지(edges)로 구성되는데 각 노드는 하나의 확률변수를, 무방향 엣지는 확률변수 간의 상관관계를 나타낸다. $V=\{X_1, ..., X_p\}$는 노드의 집합을 나타내는데, 특히 V를 구성하는 확률변수들이 다변량 정규분포(Multivariate Normal Distribution; MVN)를 따른다는 가정하에서의 그래프 모형을 가우시안 그래프 모형(gaussian graphical models)이라 한다. $E \subset V \times V$ 는 엣지의 집합이다. 서로 다른 두 노드 $X_j, X_k \in V$ $(j,k=1, ..., p)$를 잇는 엣지는 $(X_j - X_k) \in E$로 표기한다.

두 노드를 연결하는 엣지가 존재하지 않는다는 것은 두 확률변수가 나머지 변수들이 주어졌을 때 조건부 독립임을 의미한다. 한편 정규분포에서는 정밀행렬의 비대각 원소 ω_{jk} $(j \neq k)$가 0이면, j와 k번째 확률변수는 나머지 확률변수들이 주어졌을 때 조건부 독립이며, 그 역도 성립함이 알려져 있다. 정리하면, 가우시안 그래프 모형에서는 다음의 관계가 성립한다.

$$(X_j - X_k) \notin E \Leftrightarrow X_j \perp X_k \mid X_{-\{j,k\}} \Leftrightarrow \omega_{jk} = 0, \quad \text{if } (X_1, ..., X_p)^T \sim MVN. \quad (3.4)$$

여기서 $X_{-\{j,k\}}$는 X_j와 X_k를 제외한 나머지 확률변수 집합을 나타낸다.

그래프 라쏘는 정규성 가정하에서 성긴(sparse)[7] 정밀행렬을 추정하는 대표적인 알고리즘이다. 평균이 $\mathbf{0} \in \mathbb{R}^p$, 공분산 행렬이 $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$인 다변량 정규분포를 따르는 n개의 확률표본(random sample)들로 이루어진 자료 행렬 $\boldsymbol{X} \in \mathbb{R}^{n \times p}$에 대하여 정밀행렬 $\boldsymbol{\Omega}$의 로그 우도 함수는 다음과 같다.

$$l(\boldsymbol{\Omega}|\boldsymbol{X}) = \frac{n}{2}\left\{\log|\boldsymbol{\Omega}| - \mathrm{tr}(\boldsymbol{\Omega}\hat{\boldsymbol{\Sigma}}^{sample})\right\}.$$

여기서 $\hat{\boldsymbol{\Sigma}}^{sample}$는 $\boldsymbol{X}$의 표본 공분산 행렬이다. 이 로그 우도 함수를 최대화하는 $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$의 최우추정량은 이론적으로 표본 공분산 행렬의 역행렬 $\left(\hat{\boldsymbol{\Sigma}}^{sample}\right)^{-1}$이다. 하지만 고차원 환경에서는 $\left(\hat{\boldsymbol{\Sigma}}^{sample}\right)^{-1}$의 계산이 불가능하다.

그래프 라쏘는 로그 우도 함수에 l_1-노음 벌점(l_1-norm penalty)을 추가하여 고차원 환경에서도 $\boldsymbol{\Omega}$의 추정이 가능하게 하면서, 동시에 몇몇 원소를 0으로 만들어 성긴 정밀행렬의 추정치를 도출한다.

$$\hat{\boldsymbol{\Omega}}_\lambda = \underset{\boldsymbol{\Omega} > 0}{\mathrm{argmin}} \left\{-\log|\boldsymbol{\Omega}| + \mathrm{tr}(\boldsymbol{\Omega}\hat{\boldsymbol{\Sigma}}^{sample}) + \lambda \| \boldsymbol{\Omega} \|_1\right\}. \tag{3.5}$$

여기서 $\| \boldsymbol{\Omega} \|_1 = \sum_{j,k} |\omega_{jk}|$은 행렬의 l_1-노음 벌점, $\lambda \geq 0$는 벌점의 크기를 조절하는 조절 모수이다. λ가 클수록 더욱 성긴 정밀행렬이 추정된다. 최적의 λ를 선택하기 위하여 Abbruzzo 등(2019)이 제안한 다음의 GBIC (Generalized Bayesian Information Criterion) 기준을 고려하였다.

---

7) 행렬이 성기다는 것은 행렬이 0의 원소를 많이 가진다는 것을 의미한다. 즉 행렬이 성길수록 0인 원소가 많다.

$$\mathrm{GBIC}(\lambda) = -n\log|\hat{\boldsymbol{\Omega}}_\lambda| + n\,\mathrm{tr}(\hat{\boldsymbol{\Omega}}_\lambda \hat{\boldsymbol{\Sigma}}^{sample}) + \log(n) \times \mathrm{df}(\lambda).$$

여기서 $\mathrm{df}(\lambda) = \sum_{j,k} \mathrm{I}\left(\left(\hat{\boldsymbol{\Omega}}_\lambda\right)_{jk} \neq 0\right)$는 $\hat{\boldsymbol{\Omega}}_\lambda$의 0이 아닌 원소들의 개수로 정의되며 모형의 복잡도를 나타낸다. 일반적으로 GBIC 기준이 작은 λ값이 적합하다.

그래프 라쏘 알고리즘을 이용하여 성긴 정밀행렬을 추정하는 것은 식 (3.4)에 따라 가우시안 그래프 모형을 추정하는 것이며, 고차원 환경에서도 적용 가능한 방법이다. 따라서 본 연구에서 집중하고 있는 '고차원 환경에서 변수들 간의 조건부 독립성을 추정'하기 위해 사용하기 적합하다.

2) 역확률 가중치(Inverse Probability Weighting; IPW) 추정량

그래프 라쏘의 목적함수 식 (3.5)은 표본 공분산 행렬 $\hat{\boldsymbol{\Sigma}}^{sample}$에 의존하는데, 이때 자료가 결측이면 $\hat{\boldsymbol{\Sigma}}^{sample}$가 편향되는 문제가 발생한다. 이전의 연구들(Kolar & Xing, 2012; Park 등, 2021)에서는 $\hat{\boldsymbol{\Sigma}}^{sample}$의 편향을 보정한 추정량인 IPW 추정량을 그래프 라쏘의 입력으로 넣는 방법을 다루었다.

자료 행렬 $\boldsymbol{X}=\left(x_{ij}\right)_{1 \le i \le n, 1 \le j \le p}$와 자료의 결측을 나타내는 지시행렬 $\boldsymbol{R} = \left(r_{ij}\right)_{1 \le i \le n, 1 \le j \le p}$을 고려하자. $\boldsymbol{X}$의 $i\,(i = 1, ..., n)$번째 행에 해당하는 벡터 $\boldsymbol{x}_i \in \mathbb{R}^p$는 평균이 $\mathbf{0} \in \mathbb{R}^p$, 공분산 행렬이 $\boldsymbol{\Sigma} = \left(\sigma_{jk}\right)_{1 \le j,k \le p}$인 다변량 분포의 확률표본을 가정한다. $\boldsymbol{x}_i$에서 관측된 원소들을 $\boldsymbol{x}_{i,obs}$, 결측인 원소들을 $\boldsymbol{x}_{i,mis}$로 나타낸다. $\boldsymbol{R}$의 원소인 r_{ij}는 관측 확률 $\pi_{ij} = \mathrm{P}\left(r_{ij} = 1 | \boldsymbol{x}_i\right)\,(0 \le \pi_{ij} \le 1)$을 갖는 베르누이 확률변수로, 1(관측) 또는 0(결측)의 값을 갖는다. $\boldsymbol{X}$에서 계산한 표본 공분산 행렬 $\hat{\boldsymbol{\Sigma}}^{sample} = \left(\hat{\sigma}_{jk}^{sample}\right)_{1 \le j,k \le p}$의 각 원소는 다음과 같이 정의된다.

$$\hat{\sigma}_{jk}^{sample} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}. \tag{3.6}$$

하지만 자료에 결측이 있을 때 위 추정량은 바로 계산할 수 없다. 이때 사용할 수 있는 방법이 자료의 결측 원소를 0으로 대체한 후, 이로 인해 발생하는 편향을 보정해주는 방법이다. $\boldsymbol{X}^{*} := \left(r_{ij}x_{ij}\right)_{1 \le i \le n,\, 1 \le j \le p}$는 $\boldsymbol{X}$의 결측을 0으로 대체한 행렬에 해당한다. $\boldsymbol{X}^{*}$에서 식 (3.6)을 적용해 계산한 표본 공분산 행렬 $\hat{\boldsymbol{\Sigma}}^{*} = \left(\hat{\sigma}_{jk}^{*}\right)_{1 \le j,k \le p}$의 원소는 다음과 같다.

$$\hat{\sigma}_{jk}^{*} = \frac{1}{n}\sum_{i=1}^{n} r_{ij}r_{ik}x_{ij}x_{ik}. \tag{3.7}$$

MCAR 가정하에서 정의된 추정량의 기댓값은 다음과 같이 계산된다.

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}_{jk}^{*}\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[r_{ij}r_{ik}x_{ij}x_{ik}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[r_{ij}r_{ik}\right]\mathbb{E}\left[x_{ij}x_{ik}\right] \\
&= \pi_{jk} \cdot \sigma_{jk}.
\end{aligned} \tag{3.8}$$

여기서 두 확률변수 x_{ij}와 x_{ik}의 동시 관측 여부를 나타내는 확률변수 $r_{ij}r_{ik}$는 동시 관측 확률이 $\pi_{i,jk} = \mathrm{P}(r_{ij} = r_{ik} = 1 | \boldsymbol{x}_i) = \mathbb{E}\left[r_{ij}r_{ik}\right]$인 베르누이 분포를 따르며, r_{ij}와 r_{ik}의 독립 조건이 없다면 $\pi_{i,jk} \neq \pi_{ij}\pi_{ik}$이다. MCAR 가정하에서는 결측의 발생이 자료값과 상관없기 때문에 모든 i에 대해 $r_{ij}r_{ik} \perp x_{ij}x_{ik}$와 $\pi_{i,jk} = \pi_{jk}$가 성립한다. 이때 π_{jk}는 i번째 자료에 의존하지 않는 두 확률변수

X_j와 X_k의 고유한 동시 관측 확률값이다. 결론적으로, $\hat{\sigma}_{jk}^{*}$의 기댓값은 π_{jk}에 의한 편향이 존재한다. 즉, 0으로 대체된 결측값들에 의해 결측이 더 많이 발생한 j,k번째 확률변수의 공분산 값이 더 작게 편향되는 것이다.

IPW 추정량은 이러한 편향을 보정하여 동시 관측 확률이 더 작은 두 변수에 대하여 계산된 공분산 값을 더 크게 키워준다. 편향 π_{jk}를 반비례하게 조정하여 정의된 IPW 추정량 $\hat{\boldsymbol{\Sigma}}^{IPW}=\left(\hat{\sigma}_{jk}^{IPW}\right)_{1\le j,k\le p}$의 원소는 다음과 같다.

$$\hat{\sigma}_{jk}^{IPW}:=\begin{cases}\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\dfrac{r_{ij}r_{ik}x_{ij}x_{ik}}{\pi_{jk}}\text{ , if }\mathit{MCAR},\\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\dfrac{r_{ij}r_{ik}x_{ij}x_{ik}}{\pi_{i,jk}}\text{ , if }\mathit{MAR}.\end{cases} \tag{3.9}$$

MCAR 구조에서 정의된 IPW 추정량이 불편추정량임은 식 (3.8)을 통해 쉽게 확인할 수 있다. MAR 구조에서 정의된 IPW 추정량이 불편추정량임은 다음의 식을 통해 보인다.

$$\begin{aligned}\mathbb{E}\left[\hat{\sigma}_{jk}^{IPW}\right]&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{r_{ij}r_{ik}x_{ij}x_{ik}}{\pi_{i,jk}}\right]\\&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{1}{\pi_{i,jk}}\mathbb{E}\left[r_{ij}r_{ik}x_{ij}x_{ik}|\boldsymbol{x}_{i,obs}\right]\right]\\&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{1}{\pi_{i,jk}}\mathbb{E}\left[r_{ij}r_{ik}|\boldsymbol{x}_{i,obs}\right]\mathbb{E}\left[x_{ij}x_{ik}|\boldsymbol{x}_{i,obs}\right]\right]\\&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[x_{ij}x_{ik}\right]\\&=\sigma_{jk}.\end{aligned}$$

여기서 확률변수 $r_{ij}r_{ik}$는 동시 관측 확률이 $\pi_{i,jk}=\mathrm{P}(r_{ij}=r_{ik}=1|\boldsymbol{x}_{i,obs})$

$=\mathbb{E}\left[r_{ij}r_{ik}|\boldsymbol{x}_{i,obs}\right]$인 베르누이 분포를 따른다. MAR 가정하에서는 모든 i에 대해 $r_{ij}r_{ik}\perp x_{ij}x_{ik}|\boldsymbol{x}_{i,obs}$가 성립하며, $\pi_{i,jk}$는 i번째 확률표본에서 관측된 자료 $\boldsymbol{x}_{i,obs}$에 의존한다.

식 (3.9)의 IPW 추정량은 결측자료 메커니즘과 동시 관측 확률에 대한 정보를 이미 알고 있을 때 정의되는 추정량이다. 하지만 실제 상황에서 이러한 정보들을 알 수 없으므로 결측자료 메커니즘에 대한 가정 및 동시 관측 확률에 대한 추정이 필요하다. MCAR 가정하에서는 경험적(empirical) 추정량 $\hat{\pi}_{jk}=\sum_{i=1}^{n}r_{ij}r_{ik}/n$ 을 사용할 수 있다. 이는 전체 n개의 표본 중 j,k번째 변수가 동시에 관측된 표본의 비율을 나타낸다. 따라서 MCAR 가정하에서 사용하기 적합한 경험적 IPW 추정량 $\hat{\boldsymbol{\Sigma}}^{EMP}=\left(\hat{\sigma}_{jk}^{EMP}\right)_{1\le j,k\le p}$의 원소는 다음과 같이 정의된다.

$$\hat{\sigma}_{jk}^{EMP}:=\frac{\sum_{i=1}^{n}r_{ij}r_{ik}x_{ij}x_{ik}}{\sum_{i=1}^{n}r_{ij}r_{ik}}.$$

MAR 가정하에서는 관측된 자료 $\boldsymbol{x}_{i,obs}$에 기반하여 동시 관측 확률 $\pi_{i,jk}$을 예측하는 모형 기반의(model based) 추정량을 사용할 수 있다. 하지만 $\pi_{i,jk}$ 추정을 위해 모든 변수 조합별로 $\left({}_pC_2+p\right)$개의 모형을 학습하는 것은 계산 비용이 너무 크므로 변수 r_{ij}와 r_{ik}가 독립이라는 가정을 추가하여 $\hat{\pi}_{i,j}$, $\hat{\pi}_{i,k}$를 각각 추정한 후 이를 곱한 $\hat{\pi}_{i,jk}=\hat{\pi}_{i,j}\hat{\pi}_{i,k}$를 추정량으로 사용한다. $\hat{\pi}_{i,j}$를 추정하기 위해 현재 자료에서 변수 $\boldsymbol{x}_j=(x_{1j},...,x_{nj})^T$의 결측 여부 $\boldsymbol{r}_j=(r_{1j},...,r_{nj})^T$를 타겟변수로 설정하고, 나머지 변수[8]

8) 이에 대한 논의 및 한계를 Ⅴ.결론에 서술한다.

$\boldsymbol{X}_{-j} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, ..., \boldsymbol{x}_p] \in \mathbb{R}^{n \times (p-1)}$를 예측변수로 설정한 지도학습모형을 학습하여 $\hat{\pi}_j = (\pi_{1,j}, ..., \pi_{n,j})^T$를 예측한다. 본 연구에서 사용한 지도학습모형은 고차원 데이터에서 사용 가능한 지도학습모형인 벌점화 로지스틱 회귀모형이다. $\hat{\pi}_k = (\pi_{1,k}, ..., \pi_{n,k})^T$도 같은 방식으로 예측한다. MAR 가정하에서 사용하기 적합한 모형 기반의 IPW 추정량 $\hat{\boldsymbol{\Sigma}}^{MB} = \left(\hat{\sigma}_{jk}^{MB}\right)_{1 \le j,k \le p}$의 원소는 다음과 같이 정의된다.

$$\hat{\sigma}_{jk}^{MB} := \frac{1}{n} \sum_{i=1}^{n} \frac{r_{ij} r_{ik} r_{ij} r_{ik}}{\hat{\pi}_{i,j} \hat{\pi}_{i,k}}.$$

3) 제안 방법

변수 선택을 하는 고차원 대체모형의 핵심은 본 장의 1.2절에 서술한 바와 같이 대체할 결측 변수 $\boldsymbol{z}_j (j = 1, ..., l)$의 실제 중요 변수 집합 S_j을 추정하는 것이다. S_j는 $\boldsymbol{z}_j$와 조건부 의존인 변수 집합으로, 이를 추정하는 것이 변수 선택의 과정이다. 기존의 고차원 대체모형들이 S_j를 반복적으로 추정한 것과 달리, 본 연구에서는 가우시안 그래프 모형을 이용한 접근방식을 취한다. 가우시안 그래프 관점에서 S_j은 변수 $\boldsymbol{z}_j$와 연결된 이웃 노드 집합에 해당한다. 제안하는 방법은 정규성 가정하에서 그래포 라쏘와 IPW 추정량을 결합하여 변수 선택을 먼저 한 후, MICE 알고리즘을 작동시키는 것으로 구현된다. 구체적인 과정을 [알고리즘 5]로 나타낸다.

[알고리즘 5] 제안 방법 알고리즘

Step 1) 변수 선택 단계

Step 1-1) 결측자료로부터 $\hat{\boldsymbol{\Sigma}}^{IPW}$를 계산한다.

Step 1-2) 그래프 라쏘를 사용하여 성긴 정밀행렬의 추정치

$$\hat{\boldsymbol{\Omega}}_{\lambda} = \underset{\boldsymbol{\Omega} > 0}{\operatorname{argmin}} \left\{ -\log|\boldsymbol{\Omega}| + \operatorname{tr}(\boldsymbol{\Omega}\hat{\boldsymbol{\Sigma}}^{IPW}) + \lambda \| \boldsymbol{\Omega} \|_1 \right\}$$를 얻는다.

Step 1-3) 실제 중요 변수 집합 $S_j (j = 1, ..., l)$의 추정치

$$\hat{S}_j = \left\{ k \neq j \mid \left(\hat{\boldsymbol{\Omega}}_{\lambda}\right)_{jk} \neq 0 \right\} \ (k = 1, ..., p)$$를 얻는다.

Step 2) MICE 단계

MICE의 $t\,(t = 1, ..., T)$번째 반복, $j\,(j = 1, ..., l)$번째 변수에 대하여 다음을 반복적으로 수행:

Step 2-1) $\hat{\boldsymbol{\theta}}_j^{(t)} \sim \mathrm{P}\left(\boldsymbol{\theta}_j \mid \boldsymbol{z}_{obs,j}, \tilde{\boldsymbol{Z}}_{obs,\hat{S}_j}^{(t-1)}\right)$

Step 2-2) $\tilde{\boldsymbol{z}}_{mis,j}^{(t)} \sim \mathrm{P}\left(\boldsymbol{z}_{mis,j} \mid \tilde{\boldsymbol{Z}}_{mis,\hat{S}_j}^{(t-1)}, \hat{\boldsymbol{\theta}}_j^{(t)}\right)$

Step 1-1에서 실제 결측자료의 모평균은 0이 아니므로, 자료에서 변수별 표본평균을 빼서 중심화시킨 행렬 $\boldsymbol{C} = \left(c_{ij} = x_{ij} - \bar{x}_j\right)_{1 \le i \le n, 1 \le j \le p}$를 사용한다. $\hat{\boldsymbol{\Sigma}}^{IPW}$ 계산 시 모평균 추정에 의한 자유도를 고려하여 n이 아닌 $n-1$으로 나누어주어야 한다. 즉, 실제 구현에서는 $\hat{\sigma}_{jk}^{EMP} \leftarrow \frac{1}{n-1} \frac{\Sigma_{i=1}^{n} r_{ij} r_{ik} c_{ij} c_{ik}}{\Sigma_{i=1}^{n} r_{ij} r_{ik} / n}$ 및 $\hat{\sigma}_{jk}^{MB} \leftarrow \frac{1}{n-1} \sum_{i=1}^{n} \frac{r_{ij} r_{ik} c_{ij} c_{ik}}{\hat{\pi}_{i,j} \hat{\pi}_{i,k}}$을 계산한다. 자료의 MCAR을 가정할 때는 $\hat{\boldsymbol{\Sigma}}^{EMP}$, MAR을 가정할 때는 $\hat{\boldsymbol{\Sigma}}^{MB}$가 추천된다. Step 1-2에서 그래프 라쏘는 입력으로 양정치 행렬(positive definite matrix)인 공분산 행렬의 추정량을 받는다.

하지만 계산된 $\hat{\boldsymbol{\Sigma}}^{IPW}$는 양정치 행렬이 아닐 수 있으므로 $\hat{\boldsymbol{\Sigma}}^{IPW} \leftarrow \hat{\boldsymbol{\Sigma}}^{IPW} + \left(\left|\lambda_{\min}(\hat{\boldsymbol{\Sigma}}^{IPW})\right| + \epsilon\right)\boldsymbol{I}_p$를 그래프 라쏘의 입력으로 사용한다. 여기서 $\lambda_{\min}(\hat{\boldsymbol{\Sigma}}^{IPW})$는 $\hat{\boldsymbol{\Sigma}}^{IPW}$의 최소 고윳값을, ϵ은 아주 작은 양의 실수 나타낸다. Step 2에서 대체모형이 선택된 변수만을 받을 수 있도록 R의 'mice' 패키지는 mice() 함수의 'predictorMatrix'라는 인수를 제공한다. 이를 통해 변수별로 대체를 위해 사용할 예측변수를 명시할 수 있다.

제안하는 방법은 MICE 단계 '이전에' 변수 선택을 함으로써 기존의 DURR과 IURR 방법에 비하여 알고리즘의 계산 시간을 효율적으로 단축시킨다. 또한 DURR 방법은 대체모형으로 벌점화 회귀모형만을 사용할 수 있는 것과 달리 제안 방법은 변수 선택으로 미리 차원을 축소시키기 때문에 기존의 어떠한 대체모형도 적용이 가능한 유연한 변수 선택법이다.

Ⅳ. 모의실험

모의실험을 통해 정규성 가정하에서 제안하는 방법의 성능을 기존의 방법과 비교하여 입증하고자 한다. 본 장의 1절에서 모의실험 설계에 관해 설명한 후, 2절에서 모의실험 결과를 확인한다.

1. 모의실험 설계

모의실험 설계의 전반적인 절차는 Deng 등(2016)을 따랐다. 본 실험에서 설정한 분석 과제는 선형회귀 분석으로, 분석자의 최종 관심 모수는 선형회귀계수 $\boldsymbol{\beta}$이다. 실험 데이터의 불확실성을 고려하기 위하여 모든 실험 조건(scenario)마다 총 100번의 몬테카를로 실험을 수행하였다. 분석 소프트웨어는 R이며 사용한 주요 패키지는 'mice', 'huge'이다.

1.1절에서 실험 데이터의 생성 절차 및 실험 조건을 소개한 후, 1.2에서 세부적인 실험 설정 사항들에 대해 알아본다. 1.3절에서는 성능 확인을 위해 고려한 세 가지 평가지표에 대해 소개한다. 1.4절에서 비교 방법들을 정리한다.

1) 데이터 생성

선형회귀모형을 적합시킬 실험 데이터 세트 $[\boldsymbol{y}, \boldsymbol{Z}] \in \mathbb{R}^{n \times (1+p)}$를 생성하고자 한다. 타겟변수 $\boldsymbol{y} \in \mathbb{R}^{p}$는 완전히 관측되었고, 나머지 변수들 $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_p] \in \mathbb{R}^{n \times p}$ 중에는 곳곳에 결측이 포함된 상황을 고려한다. $\boldsymbol{Z}$의 생성, $\boldsymbol{y}$의 생성, $\boldsymbol{Z}$에 결측 발생순으로 실험 데이터를 생성하였다.

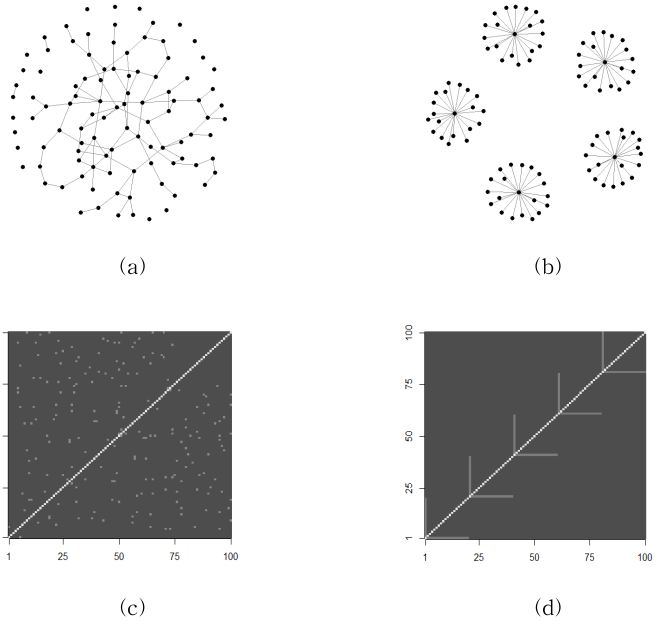
① $\boldsymbol{Z}$의 생성

평균이 $\mathbf{0} \in \mathbb{R}^p$, 공분산 행렬이 $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1} \in \mathbb{R}^{p \times p}$인 다변량 정규분포를 따르는 확률벡터 $(Z_1, ..., Z_p)^T$로부터 n개의 표본을 추출하여 행렬 $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_p] \in \mathbb{R}^{n \times p}$를 구성한다. 본 연구에서는 $n = 100$으로 고정하고, 저차원 환경과 고차원 환경에서의 성능을 모두 확인하기 위하여 $p = 50$, $p = 100$, $p = 200$의 다양한 차원 수를 고려하였다.

정밀행렬 $\boldsymbol{\Omega}$는 변수 간의 조건부 독립성 관계에 대한 그래프 구조를 반영한다. 실험에서 고려한 그래프 구조는 랜덤 구조와 허브 구조이다. 랜덤 구조는 노드들이 무작위로 연결되어 있는 구조로, 본 실험에서는 ${}_pC_2$개의 각 엣지에 대한 발생확률을 0.02로 설정하였다. 허브 구조는 허브라고 불리는 몇몇 중요 노드들을 중심으로 노드 간의 그룹이 형성되어 있는 구조이다. 이를 위해 p개의 변수들을 처음부터 20개씩 순서대로 그룹화하여 서로 다른(disjoint) $\lceil p/20 \rceil$개 그룹을 형성한 후, 각 그룹의 첫 번째 변수$(Z_{21}, Z_{41}, ...)$를 허브로 선택하였다. 그래프 구조를 행렬로 나타내는 방법은 두 노드를 연결하는 엣지가 존재할 경우 1, 존재하지 않을 경우 0의 값을 갖는 대칭행렬인 인접행렬(adjacency matrix)를 이용하는 것이다. 인접행렬로부터 정밀행렬을 도출하는 식은 Liu & Wang(2017)에 제시된 다음의 식을 채택한다.

$$\boldsymbol{\Omega} = \boldsymbol{A} + (|\lambda_{\min}(\boldsymbol{A})| + 0.2)\boldsymbol{I}_p.$$

여기서 $\boldsymbol{A} \in \mathbb{R}^{p \times p}$는 인접행렬이며, $\lambda_{\min}(\boldsymbol{A})$는 $\boldsymbol{A}$의 최소 고윳값을 나타낸다. [그림 4]는 랜덤 구조와 허브 구조 각각의 네트워크 그래프와 정밀행렬에 대한 히트맵을 보여준다.



[그림4] 네트워크 그래프 및 정밀행렬 히트맵

(a): 랜덤 구조의 네트워크 그래프 (b): 허브 구조의 네트워크 그래프

(c): 랜덤 구조의 정밀행렬 히트맵 (d): 허브 구조의 정밀행렬 히트맵

② y의 생성

참 선형회귀모형을 다음과 같이 설정한다.

$$Y=\beta_0+\beta_1 Z_1+\beta_2 Z_2+\beta_3 Z_3+\beta_4 Z_4+\beta_5 Z_5+\epsilon,\ \ \epsilon\sim N(0,1),$$

Y를 설명하는 참 설명변수는 $Z_1,...,Z_5$인 상황이며, ϵ은 $Z_j(j=1,...,5)$와 독립인 랜덤 노이즈 변수이다. 모회귀계수는 $\boldsymbol{\beta}=(\beta_0,\beta_1,...,\beta_5)^T$는 모두 1로 설정하였다. 이에 따라 자료로부터 $\boldsymbol{y}=\mathbf{1}+\boldsymbol{z}_1+\cdots+\boldsymbol{z}_5+\boldsymbol{\epsilon}\in\mathbb{R}^n$를 생성하였다.

③ Z에 결측 발생

마지막으로 Z의 일부에 결측을 발생시킨다. 실제 설명변수 중 일부가 결측인 상황을 고려하기 위하여 변수 z_1, z_2, z_3에 결측을 발생시켰다. 결측자료 메커니즘은 본 연구에서 전제로 한 MAR 메커니즘 및 민감도 분석(sensitivity analysis)을 위한 NMAR 메커니즘을 고려하였다. 먼저, MAR 가정하에서의 결측자료 메커니즘은

$$\begin{aligned}
\mathrm{P}(R_1 = 0 | Z_5, Z_{10}, Y) &= \mathrm{sigmoid}(k - Z_5 + 2Z_{10} - Y), \\
\mathrm{P}(R_2 = 0 | Z_{20}, Z_{30}, Y) &= \mathrm{sigmoid}(k - Z_{20} + 2Z_{30} - Y), \\
\mathrm{P}(R_3 = 0 | Z_{40}, Z_{45}, Y) &= \mathrm{sigmoid}(k - Z_{40} + 2Z_{45} - Y).
\end{aligned}$$

이고, NMAR 가정하에서의 결측자료 메커니즘은

$$\begin{aligned}
\mathrm{P}(R_1 = 0 | Z_5, Z_1, Y) &= \mathrm{sigmoid}(k - Z_5 + 2Z_1 - Y), \\
\mathrm{P}(R_2 = 0 | Z_{20}, Z_2, Y) &= \mathrm{sigmoid}(k - Z_{20} + 2Z_2 - Y), \\
\mathrm{P}(R_3 = 0 | Z_{40}, Z_3, Y) &= \mathrm{sigmoid}(k - Z_{40} + 2Z_3 - Y).
\end{aligned}$$

이다. 여기서 $R_j (j = 1,2,3)$는 변수 Z_j에 대한 결측 지시변수이고, 시그모이드 함수 $\mathrm{sigmoid}(t) = \dfrac{e^t}{1+e^t} \in (0,1), t \in \mathbb{R}$는 실수를 0과 1 사이의 값으로 보내주는 역할을 한다. 모수 k를 통해 결측 비율(missing rate)을 조정할 수 있는데, 본 연구에서는 전체 표본 중 결측이 발생한 표본의 비율이 40% 55% 70%가 되도록 다양한 결측 비율을 고려하였다.

2) 실험 설정

다중대체를 위한 대체 횟수는 $M=10$으로, MICE 알고리즘의 최대 반복 횟수는 $T=10$으로 지정하였다.

제안 방법을 실행할 때 적용한 두 가지 사항들을 설명하겠다. 먼저, 서론에서 소개한 바와 같이 분석 단계에서 사용할 변수는 대체모형에도 포함시키는 것이 대체에 유리하다. 고차원 데이터 세트에서는 설명변수 후보 $\boldsymbol{Z}$ 중 어느 변수가 (대체 후) 변수 선택이 되어 $\boldsymbol{y}$의 예측을 위해 사용될지 모르기 때문에, 현재 상황에서는 $\boldsymbol{y}$만이 분석 단계에서 사용할 것으로 알려진 유일한 변수이다. 추가로, 대체모형에 반응변수 $\boldsymbol{y}$를 항상 포함시키는 것이 대체에 유리하다는 연구 결과들(Sterne 등, 2009; Van Ginkel 등, 2020)이 존재한다. 따라서 본 모의실험에서는 이러한 원칙에 따라 제안 방법을 수행할 때 $\boldsymbol{Z}$ 에서만 조건부 독립성을 추정하여 변수 선택을 한 후 선택된 변수 집합에 $\boldsymbol{y}$를 추가하여 대체를 수행하였다. 두 번째 고려사항은 모형 기반의 IPW 추정량 $\hat{\boldsymbol{\Sigma}}^{MB}$의 계산에서 사용한 모형에 관한 것이다. 결측 벡터인 $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3$에 대한 관측 확률의 추정치 $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3$를 예측하기 위하여 각각 $\boldsymbol{z}_j (j=1,2,3)$의 결측 지시벡터 $\boldsymbol{r}_j \in \mathbb{R}^n$를 타겟변수로, 관측된 변수 집합 $[\boldsymbol{z}_4, ..., \boldsymbol{z}_p] \in \mathbb{R}^{n \times (p-3)}$를 예측변수로 설정한 벌점화 로지스틱 회귀모형을 사용하였다.

본 모의실험에서는 서로 다른 평가지표를 확인하기 위하여 두 종류의 분석을 수행하였다. 첫 번째로, 대체 데이터 세트에서 모수 $\beta_0, ..., \beta_5$가 얼마나 잘 추정되는지 확인하기 위하여 대체 데이터 세트에서 변수 $[\boldsymbol{z}_1, ..., \boldsymbol{z}_5] \in \mathbb{R}^{n \times 5}$만을 사용하여 $\boldsymbol{y}$에 대한 선형회귀모형을 적합시켰다. 이를 통해 상수항을 포함하여 추정된 회귀계수를 $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, ..., \hat{\beta}_5\right)^T$를 얻는다. 두 번

째로, '대체 후 변수 선택'이 얼마나 잘 되는지 확인하였다. 실제 상황에서는 $\boldsymbol{y}$를 설명하는 참 설명변수가 $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_5$인지 알 수 없으므로 이 또한 변수 선택을 통해 추정해야 한다. 따라서 대체 데이터 세트에서 참 설명변수가 잘 선택되는지를 확인하는 것 또한 중요한 분석 과제이다. 이를 위하여 대체 데이터 세트에서 $\boldsymbol{Z}$로부터 $\boldsymbol{y}$에 대한 라쏘 모형을 적합시켰다. 라쏘 모형은 R의 'glmnet' 패키지의 cv.glmnet() 함수를 사용하였다. 획득된 $\hat{\boldsymbol{\beta}}$와 '대체 데이터 세트에서 선택된 변수들'로 성능을 확인하기 위한 평가지표는 이어지는 2.3절에서 자세히 소개한다.

3) 평가지표

본 모의실험에서는 제안 방법의 성능을 확인하기 위하여 총 세 가지의 평가지표를 사용한다. 모수 $\boldsymbol{\beta}$의 추정 성능을 확인하기 위한 '편향' 및 '95% 신뢰구간의 포함률(Coverage Rate of the 95% confidence interval; CR)'과, 대체 후 변수 선택 성능을 확인하기 위한 'ROC 커브(Receiver Operating Characteristic curve)'를 확인한다. 다중대체에서 각 대체 데이터 세트마다 얻어지는 추정치들은 식 (2.1)-(2.3)의 Rubin's Rule에 의해 결합된다.

① 편향: $\| Bias \|_2$

모수 $\boldsymbol{\beta}$의 추정에 대한 첫 번째 지표로는 추정된 모수가 참 모수와 얼마나 가까운지를 나타내는 지표인 편향을 확인한다. 본 연구에서 관심 모수는 $\beta_0, \ldots, \beta_5$로 총 6개이므로 이들의 편향을 하나의 값으로 확인하기 위하여 편향 벡터의 l_2-노음인 $\| Bias \|_2$을 제시하였다.

$$\| Bias \|_2 = \sqrt{\sum_{j=0}^{5} (Bias_j)^2}\ ,\ \ Bias_j = \hat{\beta}_j - \beta_j\ \ (j = 0, ..., 5),$$

여기서 추정된 회귀계수 $\hat{\beta}_j$는 식 (2.1)에 의하여 결합된 최종추정치며, 이를 참 모수 $\beta_j = 1$과 비교한다.

② 95% 신뢰구간의 포함률(Coverage Rate of the 95% confidence interval; CR)

모수 β의 추정에 대한 두 번째 지표로 추정된 회귀계수가 유효한지 확인하기 위하여 95% 신뢰구간의 포함률을 확인한다. 최종추정치 $\hat{\beta}_j\,(j = 0, ..., 5)$의 95% 신뢰구간은 식(2.3)으로부터 계산된다. 100번의 몬테카를로 실험 중 계산된 95% 신뢰구간에 참 모수가 포함된 비율이 신뢰구간 포함률 CR이 된다.

③ ROC 커브(Receiver Operating Characteristic curve)

대체 후 변수 선택을 통해 선택된 변수와 참 설명변수를 비교하기 위한 지표로 ROC 커브를 확인한다. ROC 커브는 Y축이 TPR (True Positive Rate), X축이 FPR (False Positive Rate)인 그래프 위에서 그려진다.

먼저, 변수 선택에서 TPR와 FPR의 의미를 살펴보겠다.

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{TN + FP}.$$

여기서 TP (True Positive)는 참 설명변수가 변수 선택된 경우, FN (False Negative)는 참 설명변수가 변수 선택되지 않은 경우. FP (False Positive)는 참 설명변수가 아닌 변수($z_6, ..., z_p$)가 변수 선택된 경우, TN (True Negative)는 참 설명변수가 아닌 변수가 변수 선택되지 않은 경우를 나타낸다. 즉 TPR은 참 설명변수가 변수 선택을 통해 제대로 선택된 비율이며, FPR은 참 설명변수가 아닌 변수가 변수 선택을 통해 잘못 선택된 비율이다. TPR은 1에 가까울수록, FPR은 0에 가까울수록 변수 선택이 잘된 것이다.

다음으로 커브에 대한 의미를 살펴보겠다. 라쏘 모형은 조절 모수인 $\lambda(\lambda \geq 0)$가 0이면 모든 변수를 선택하고, 값이 커질수록 더 적은 변수를 선택한다. 따라서 λ의 경로(path)에 따라 변수 선택 결과로 (X,Y)=(FPR, TPR) 좌표에 점을 찍고, 그 점들을 이어 커브를 그릴 수 있다. 본 실험에서는 0부터 적당히 큰 실수까지 100개의 λ들로 λ 경로를 지정하였다. 다중대체에서 동일한 λ별로 얻어진 M개의 변수 선택 결과를 다수결 투표(majority vote)하여 $M/2$번 이상 선택된 변수들의 집합이 (λ별) 최종 선택된 변수 집합이 된다.

4) 비교 방법

본 모의실험에서 비교할 방법들을 [표 1]에 제시하였다. GS (Gold Standard) 방법은 결측을 발생시키기 전의 실제 데이터 세트에서 분석을 수행하는 것으로, 정답을 알고 있는 경우이다. CCA는 결측이 발생한 개체를 모두 제거하고 분석을 수행하는 완전 제거법이다. TAS (True Active Set)는 대체 시 실제 중요 변수 집합을 사용해 변수 선택한 방법이다. 모의실험에서는 실제 그래프 구조를 알고 있으므로 변수 z_j의 대체 시 z_j와 연결된

실제 이웃 노드만을 변수 선택할 수 있다. PMM은 대체 시 변수 선택하지 않은 방법이다. 이는 저차원($p=50$)인 상황에서만 비교 가능했으며, 대체모형으로 PMM을 사용했다. 서론에 서술한 바와 같이 저차원이라도 변수가 너무 많은 경우 변수 선택을 하는 것이 유리하다는 연구 결과(Hardt 등, 2012)가 존재한다. DURR과 IURR은 제안 방법의 주요 비교 방법이다. P-EMP 과 P-MB은 각각 $\hat{\boldsymbol{\Sigma}}^{EMP}$와 $\hat{\boldsymbol{\Sigma}}^{MB}$를 이용해 계산한 제안 방법이다.

[표 1] 비교 방법

대체 유무	적용 가능성	비교 방법
대체 X	이상적	GS
	현실적	CCA
대체 O	이상적	TAS
	현실적 (저차원)	PMM
	현실적 (고차원)	DURR
		IURR
		P-EMP
		P-MB

2. 모의실험 결과

모의실험의 결과는 100번의 몬테카를로 실험에 대해 요약된 결과이다. 차원 수($p=50$, $p=100$, $p=200$)와 결측 비율(40%, 55%, 70%)의 변화에 따른 실험 결과를 비교하기 쉽도록 하나의 그림에 요약하였다. 각 평가지표마다 랜덤 구조와 허브 구조, 두 종류의 그림이 제시된다. 2.1절에서는 대체모형의 기본 가정인 MAR 메커니즘으로부터 결측자료가 생성되었을 때의 분석 결과를 확인하고, 2.2절에서는 MAR 가정을 위반했을 때의 민감도 분석 결과를 확인한다.

차원별 알고리즘의 평균 계산 시간은 [표 2]와 같다. 기존의 방법은 분 단위로 오래 걸리는 반면 제안하는 방법의 계산 시간은 초 단위로 빠르게 계산되는 것을 확인할 수 있다. 시간 차가 가장 큰 $p=200$에서는 제안 방법이 기존 방법에 비하여 약 35배 빨랐다.

[표 2] 알고리즘의 평균 계산 시간

	$p=50$	$p=100$	$p=200$
DURR	2.39 *min*	4.50 *min*	22.47 *min*
IURR	2.43 *min*	4.97 *min*	21.04 *min*
P-EMP	**11.72 *sec***	**20.61 *sec***	**37.88 *sec***
P-MB	**12.80 *sec***	**21.84 *sec***	**37.82 *sec***

1) MAR 메커니즘에서의 분석 결과

[그림 5]와 [그림 6]은 $\|Bias\|_2$에 대한 100번의 몬테카를로 실험의 평균과 평균으로부터 1 표본오차까지의 구간을 나타낸 것이다. 랜덤 구조와 허브 구조 모두에서, 모든 경우에서 GS의 편향은 거의 없었고 대부분의 경우에서 결측인 개체를 모두 삭제하는 CCA의 편향이 가장 컸다. $p=50$일 때 CCA 다음으로는 PMM의 편향이 큰 것으로 보아, 저차원 환경에서도 변수 선택을 하는 것이 유리하다는 것을 알 수 있다. 저차원에서 제안 방법(P-EMP, P-MB)과 주요 비교 방법(DURR, IURR)간의 성능 차이는 두드러지지 않았다. 고차원 환경($p=100, 200$)일 때를 중심으로 비교하면, 랜덤 구조에서는 제안 방법이 주요 비교 방법보다 전체적으로 낮은 편향을 보였으며, 이러한 경향성은 차원이 증가하고 결측 비율이 증가할 수로 두드러지게 나타났다. 허브 구조에서는 고차원 대체모형의 성능이 랜덤 구조에 비하여 전체적으로 떨어진다. 제안 방법을 살펴보면, 고차원의 모든 경우에서

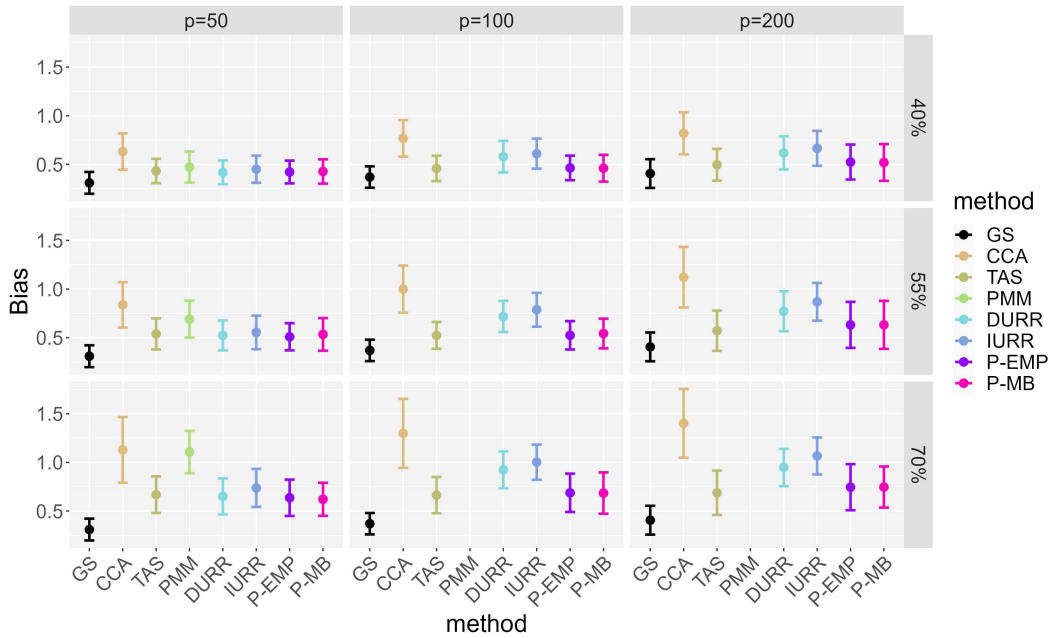
P-EMP는 DURR과 IURR보다 낮은 편향을 보였다. 하지만, P-MB의 경우 차원이 증가하고 결측 비율이 증가할수록 편향이 크게 증가하였다. 특히 $p=200$이고 결측 비율이 55% 또는 70%인 상황에서 CCA보다도 높은 편향을 가졌다. 하지만, 제안 방법의 이상적 경우인 TAS의 편향이 다른 방법들보다 눈에 띄게 낮은 것을 보아 변수 간의 조건부 독립성을 추정해 변수 선택을 하는 제안 방법의 아이디어가 유용함을 확인할 수 있다.

[그림 7]과 [그림 8]은 $\beta_0, ..., \beta_5$별로 100번의 몬테카를로 실험 중 95% 신뢰구간에 참 모수가 포함되는 비율인 CR을 나타낸다. GS는 거의 1에 가까운 CR을 보이므로 거의 모든 경우에 신뢰구간 안에 참 모수가 포함된다고 해석할 수 있다. CCA는 모든 상수항에서 0에 가까운 낮은 CR을 보였다. 랜덤 구조를 먼저 보면, CCA의 성능을 나타내는 노란 점선보다 아래에 위치하는 선은 PMM과 IURR이다. 두 모형 모두 결측 비율이 증가할수록 $\beta_1, \beta_2, \beta_3$의 CR이 눈에 띄게 감소하였다. 반면, 제안 방법은 모든 경우에서 안정적으로 우수한 CR을 보였다. 다음으로 허브 구조를 확인하면, 랜덤 구조보다 전체적으로 모형들의 성능이 떨어진다. 이로써 모수 추정의 유효성을 대변하는 CR은 변수 간 네트워크 구조에 민감함을 알 수 있다. 저차원의 경우 TAS, DURR, P-EMP · P-MB, IURR, PMM 순으로 성능이 우수한 것으로 관찰되었다. 고차원의 경우, 결측 변수와 관련된 회귀계수 $\beta_1, \beta_2, \beta_3$에 대해서는 제안 방법(P-EMP, P-MB)이 주요 비교 방법(DURR, IURR)보다 우수한 성능을 보였으며, 회귀계수 β_4, β_5에 대해서는 반대의 결과가 나왔다. 변수별로 다른 CR 패턴이 나타나는 이유에 대해 추가적인 연구가 필요하다.

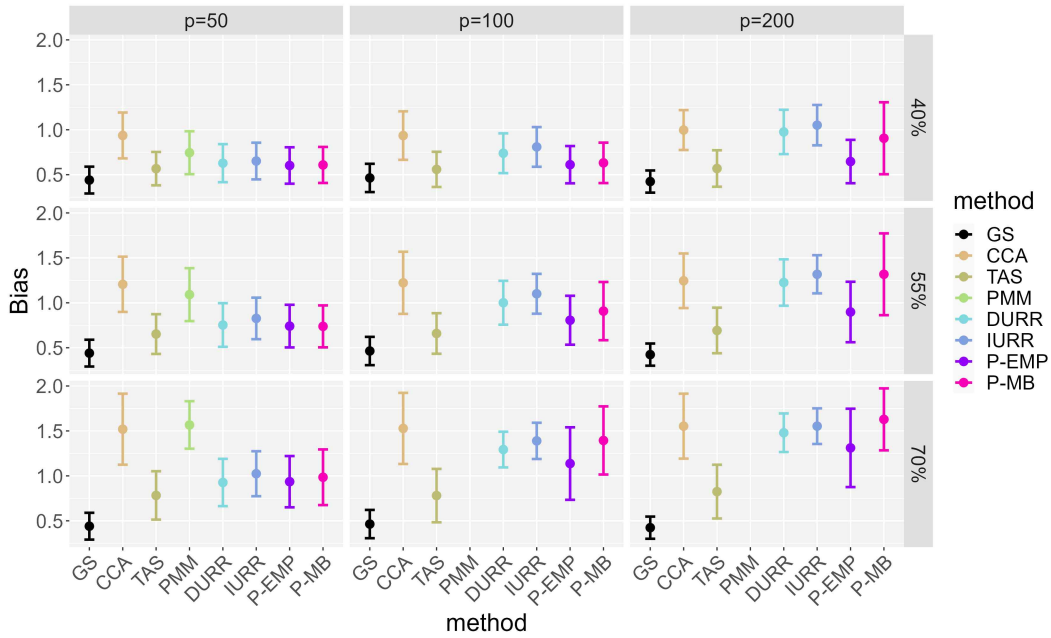
[그림 9]와 [그림 10]은 대체 후 변수 선택의 결과를 나타내는 ROC 커브를 100번의 몬테카를로 실험에 대하여 평균 커브로 그린 것이다. TPR은 1에 가까울수록, FPR은 0에 가까울 우수한 성능을 나타내기 때문에 커브가

좌측 상단에 붙어 커브 아래 면적이 넓을수록 이상적이다. 랜덤 구조를 확인하면, GS는 거의 가장 이상적인 형태의 커브가 그려졌으며 TAS, P-EMP · P-MB, DURR · IURR, CCA 순으로 대체 후 변수 선택 성능이 우수한 것을 확인할 수 있다. 차원이 증가하고 자료의 결측 비율이 증가할수록 DURR, IURR, CCA의 성능은 떨어지는 반면, 제안 방법의 성능은 안정적으로 유지된다. 확인 결과, 이러한 현상은 반응변수인 $\boldsymbol{y}$를 대체모형에 항상 포함시켰을 때 나타났다. 따라서 반응변수를 대체모형에 항상 포함시키는 것이 대체 후 변수 선택에 유리하다는 것을 실험적으로 확인하였다. 허브 구조에서는 전체적으로 커브가 좌측 상단에서 멀어졌지만 모형들의 성능 순위는 랜덤 구조와 동일했다.

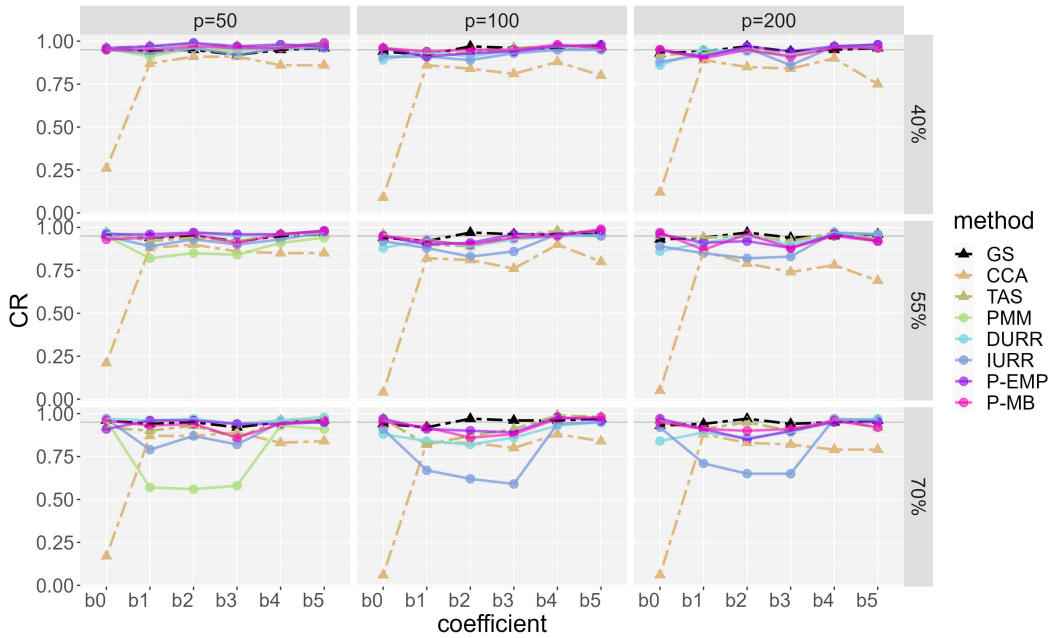
본 모의실험을 통해 MAR 가정 및 정규성 가정을 만족할 때, 세 가지 지표 모두에서 제안하는 방법의 성능이 DURR과 IURR보다 전체적으로 우수하게 나타났다. 특히 변수 간의 실제 네트워크 구조를 알고 있는 이상적인 경우인 TAS의 성능이 GS 다음으로 좋았는데, 이는 그래프 모형을 통해 변수 간의 네트워크 구조를 잘 추정하면 다중대체를 위한 변수 선택에 도움이 된다는 점을 시사한다.



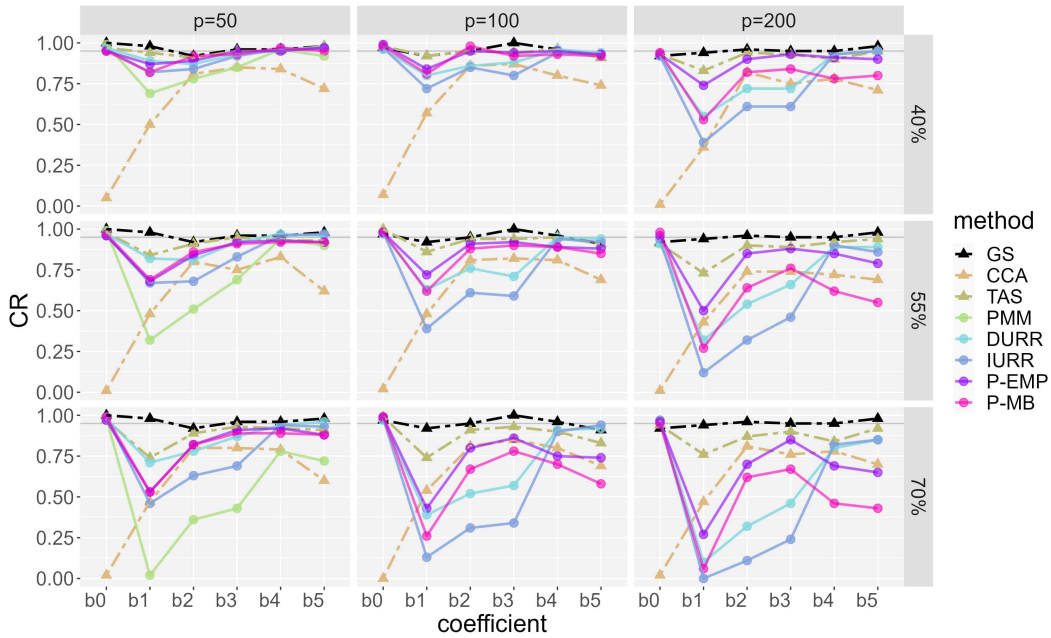
[그림 5] 랜덤 구조의 편향: $\| Bias \|_2$



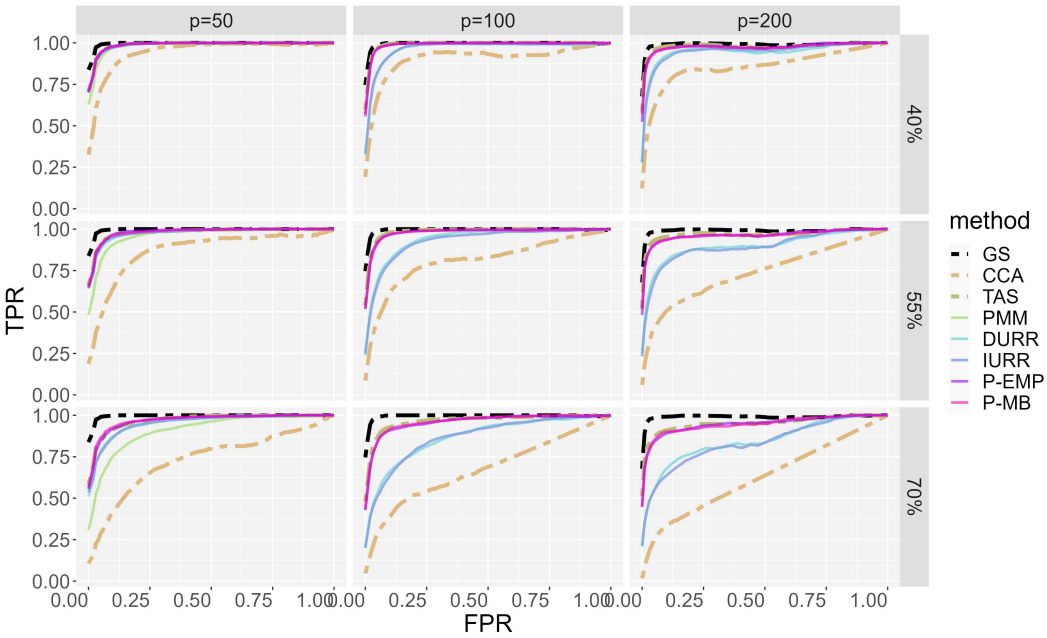
[그림 6] 허브 구조의 편향: $\| Bias \|_2$



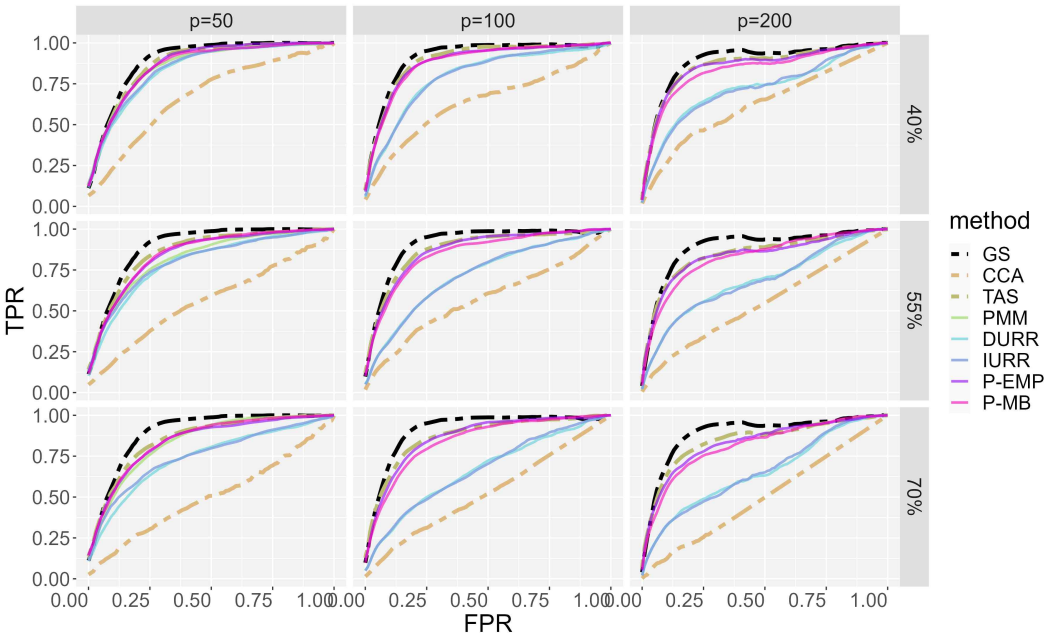
[그림 7] 랜덤 구조의 CR



[그림 8] 허브 구조의 CR



[그림 9] 랜덤 구조의 ROC 커브



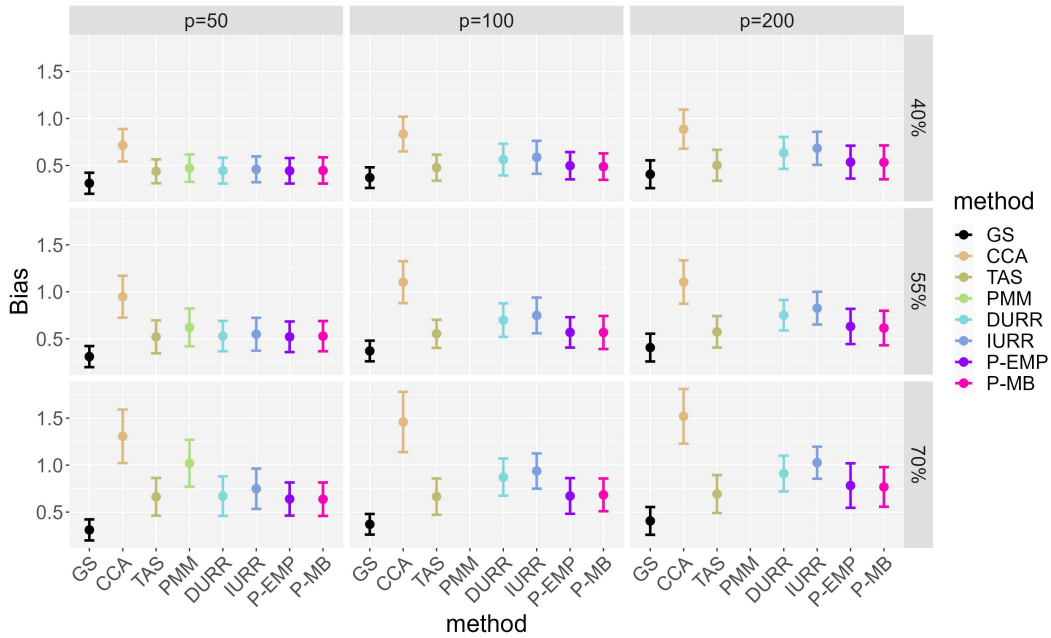
[그림 10] 허브 구조의 ROC 커브

2) 민감도 분석 결과

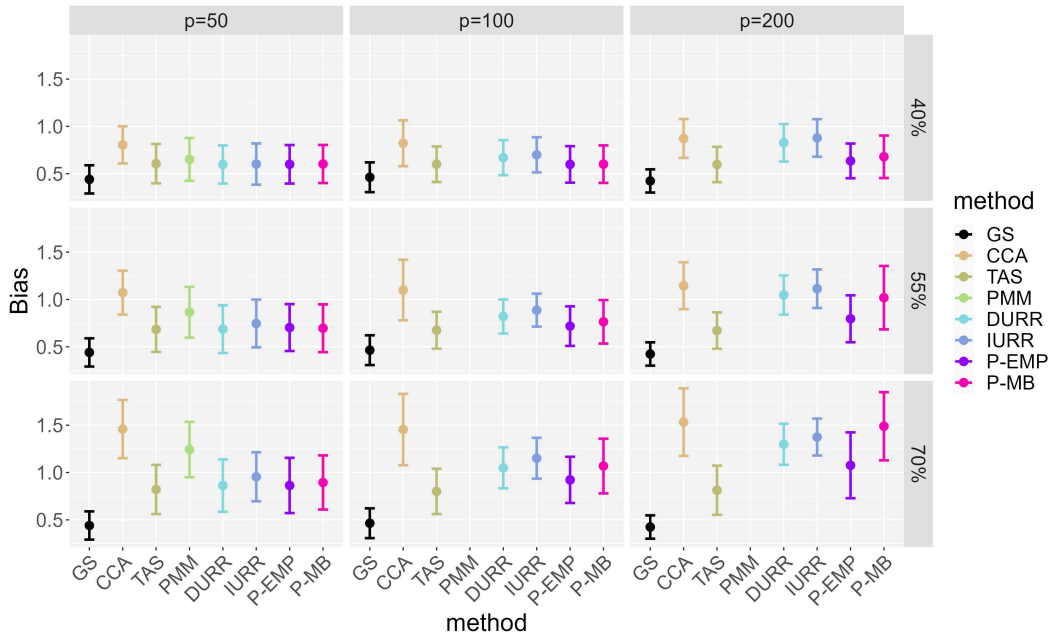
1장에 서술한 바와 같이 주어진 결측자료가 실제로 어느 메커니즘을 따르는지는 알 수 없다. 따라서 대체모형이 가정하고 있는 MAR 가정이 위배되었을 때 대체모형의 성능이 얼마나 나빠지는지를 확인하기 위하여 민감도 분석을 수행하였다. NMAR 메커니즘으로 결측을 발생시켰음에도 불구하고 전체적으로 MAR 메커니즘에서의 결과와 비슷한 결과가 나왔다. 제안 방법의 성능 또한 민감하게 떨어지지 않았다.

민감도 분석에서는 전체 모형의 성능 감소가 일어나는 것이 일반적인 결과이지만, 본 모의실험에서는 MAR과 성능의 차이가 없었다. 확인 결과, 모의실험에서 설계된 결측자료 메커니즘이 결측된 변수 $(Z_1, Z_2, Z_3)^T$와 결측의 발생 $(R_1, R_2, R_3)^T$이 조건부 독립을 만족하지 않는 NMAR에 속했으나, 정확히는 1장에 소개된 MAR과 NMAR이 혼합된 구조였다. 즉 결측의 발생이 온전히 결측 변수에만 의존하지 않고 관측 변수에도 의존하여 발생했기 때문에, 여전히 결측값이 관측값으로 설명되는 부분이 있었다. 따라서 더 정확한 민감도 분석 결과를 확인하기 위해서는 온전한 NMAR 메커니즘을 따르는 자료의 실험을 추가할 필요가 있다.

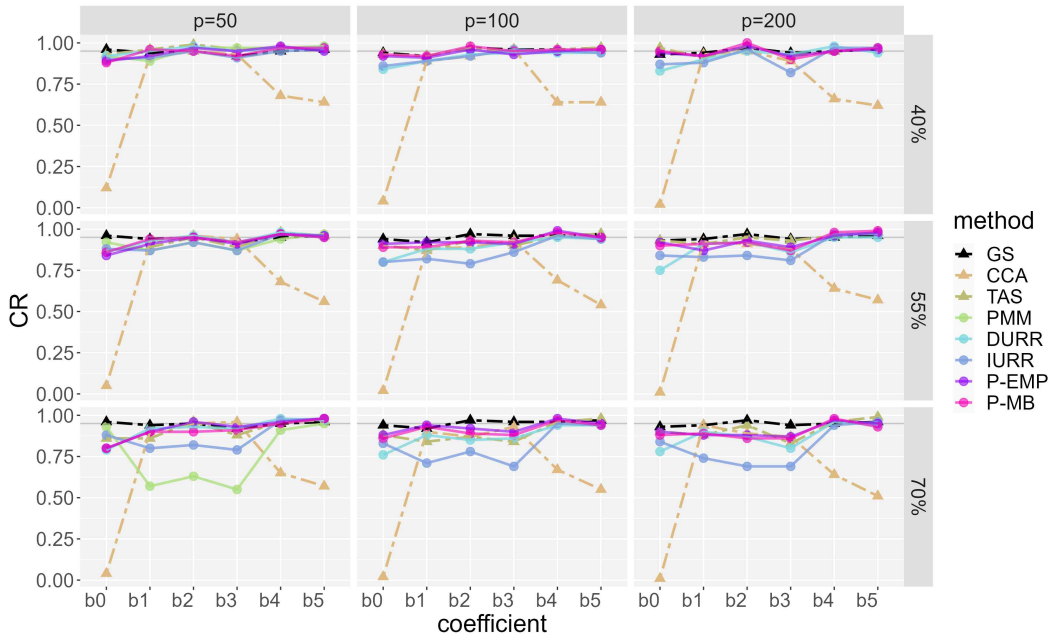
결론적으로, 다른 설계의 결측자료 메커니즘을 확인해볼 필요가 있지만, 현재 설계된 메커니즘 또한 NMAR의 일종이므로 제안 방법이 결측자료 메커니즘에 민감하게 반응하지는 않을 것이라 예상한다. NMAR이라 하더라도 결측 변수를 설명하는데 도움이 되는 변수들 즉, 실제 중요 변수 집합(true active set)을 빠짐없이 잘 포함하면 NMAR 가정이 완화되어 결측값을 예측할 수 있기 때문이다.



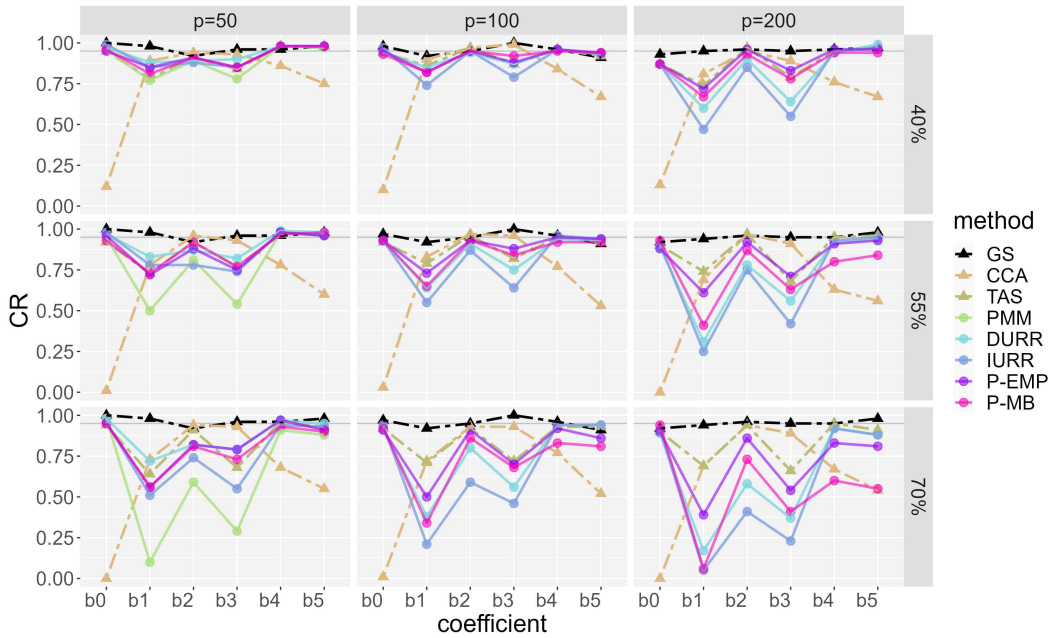
[그림 11] 랜덤 구조의 편향: $\| Bias \|_2$



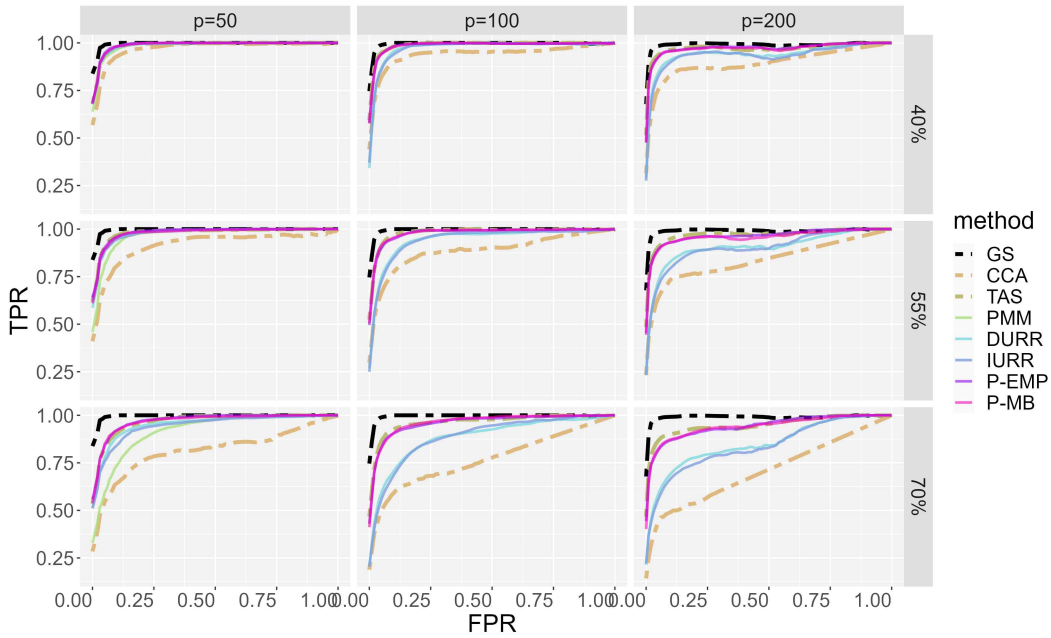
[그림 12] 허브 구조의 편향: $\| Bias \|_2$



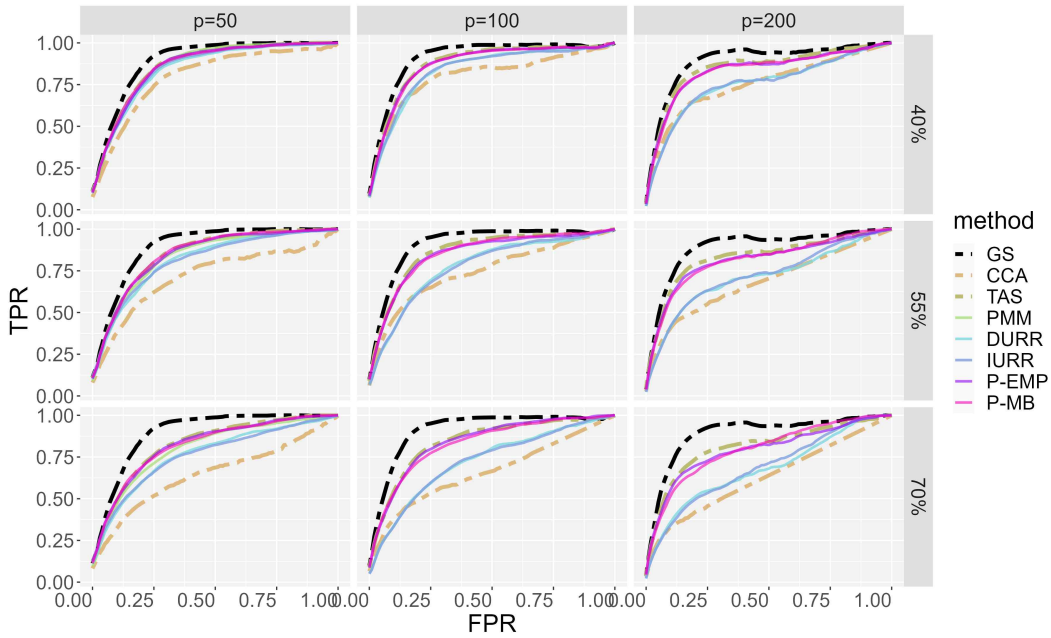
[그림 13] 랜덤 구조의 CR



[그림 14] 허브 구조의 CR



[그림 15] 랜덤 구조의 ROC 커브



[그림 16] 허브 구조의 ROC 커브

Ⅴ. 결 론

본 연구에서는 고차원 결측자료에서 다중대체를 수행하기 이전에 적용 가능한 새로운 변수 선택법을 제안하였다. 제안 방법은 실제 중요 변수 집합의 추정을 위해 변수들 간의 조건부 독립성을 추정하는 그래프 라쏘를 사용하는 측면에서 이론적 정당성을 가진다. 이때 결측자료에서 불편추정량인 IPW 추정량을 정의하여 그래프 라쏘의 입력으로 사용하였다. 제안 방법의 알고리즘은 '변수 선택 단계' 이후 'MICE 단계'로 넘어가기 때문에, 반복적인 변수 선택 과정을 피함으로써 기존 MICE의 고차원 대체모형들보다 훨씬 빠르게 작동하며 변수 선택 이후 저차원 대체모형을 포함해 기존의 어떠한 대체모형도 사용이 가능하다는 장점이 존재한다. 모의실험 결과 본 연구에서 가정하고 있는 정규성 및 MAR이 만족되는 상황에서 제안 방법이 기존 방법들보다 비슷하거나 우수한 성능을 보였다.

본 연구에서는 극복해야 할 몇 가지 한계점 및 향후 연구의 주제들이 있다. 먼저, 모델 기반의 IPW 추정량에 대한 것이다. 이 추정량을 계산하기 위해서는 관측된 변수들로 결측 변수의 관측 확률을 예측한다. 모의실험은 완전히 관측된 변수들을 모형의 예측변수로 사용하였지만, 실제 자료에서는 완전히 관측된 변수가 없거나 적을 수 있으며 어떤 변수에는 한 두 개의 결측값만이 존재할 수도 있다. 따라서 이를 어떻게 처리할지와 같은 현실적인 이슈들이 존재하고, 모델 기반의 IPW 추정량을 실제로 사용할 수 있는 방법에 관한 향후 연구가 필요하다.

두 번째로, 본 논문의 결과들은 본 모의실험에서 설계된 환경으로 제한된다. 보다 일반적인 성능을 확인하기 위해서는 더욱 다양한 설계에서 결과를 확인할 필요가 있다. 특히, 본 연구의 민감도 분석에서는 MAR과 NMAR이 혼합된 메커니즘 설계에 의해 대체모형의 민감도를 정확하게 확인하지 못하

였다. 온전한 MAR과 온전한 NMAR 메커니즘을 따르는 설계에서의 실험 결과를 추가로 확인할 필요가 있다.

마지막으로, 본 연구의 제안 방법은 정규성 가정하에서 그래프 라쏘를 활용한다. 하지만 실제 자료가 정규성 가정을 따르기란 현실적으로 힘들다. 그래프 이론 분야에는 자료가 정규분포가 아닌 다른 분포를 따르는 경우 또는 자료에 이산형 변수가 포함된 경우 등 다양한 경우에서 네트워크 구조를 추정하는 그래프 모형들이 존재한다. 본 연구에서는 그래프 이론 모형 중 이론적으로 탄탄한 그래프 라쏘를 사용했을 때 제안 방법이 잘 작동함을 확인하였다. 따라서 다양한 자료에서도 제안 방법의 사용이 가능하도록 다른 그래프 이론 모형들을 적용한 향후 연구를 통해 제안 방법을 확장시킬 필요가 있다. 본 연구는 다중대체를 위한 변수 선택에 그래프 이론 모형을 사용한 좋은 출발점이 될 것으로 기대한다.

참 고 문 헌

고길곤·탁현우. (2016). 설문자료의 결측치 처리방법에 관한 연구: 다중대체법과 재조사법을 중심으로. 행정논총. 54(4):291-319.

김현태·장가영. (2023) 데이터 가명·익명처리 기법의 현황과 대안: 재현데이터를 중심으로. 보험연구원, 27-31.

송주원·안형진. (2009). 무응답 자료 처리 및 분석. 통계청 통계교육원, 12-17, 69-71, 128-130.

Abbruzzo, A., Vujačić, I., Mineo, A. M., & Wit, E. C. (2019). Selecting the tuning parameter in penalized Gaussian graphical models. Statistics and Computing, 29(3), 559-569.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of school psychology, 48(1), 5-37.

Brini, A., & van den Heuvel, E. R. (2023). Missing data imputation with high-dimensional data. The American Statistician, (just-accepted), 1-19.

Carpenter, J., & Kenward, M. (2008). Brief comments on computational issues with multiple imputation. Unpublished paper retrieved from http://missingdata. lshtm. ac. uk/downloads/mi_comp_issues. pdf.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological methods, 6(4), 330.

Costantini, E., Lang, K. M., Reeskens, T., & Sijtsma, K. (2022). High-dimensional imputation for the social sciences: a comparison of state-of-the-art methods. arXiv preprint arXiv:2208.13656.

Deng, Y., Chang, C., Ido, M. S., & Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. Scientific reports, 6(1), 21689.

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. BMC medical research methodology, 12, 1-13.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3), 432-441.

Kolar, M., & Xing, E. P. (2012, June). Estimating sparse precision matrices from data with missing values. In International Conference on Machine Learning (pp. 635-642).

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American statistical Association, 83(404), 1198-1202.

Little, R. J. A. & Rubin, D. B. (2002) Statistical Analysis with Missing Data, Wiley: New York.

Liu, H., & Wang, L. (2017). Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. Statistical science, 538-558.

Park, S., Wang, X., & Lim, J. (2021). Estimating high-dimensional covariance and precision matrices under general missing dependence. Electronic Journal of Statistics, 15(2), 4868-4915.

Rubin D. B. (1978) Multiple imputation in sample surveys, Proceedings in

Survey Research Methodology, American Statistical Association, 20-34.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338.

Van Buuren, S. (2000). Multivariate imputation by chained equations: MICE V1. 0 user's manual. Leiden: TNO.

Van Buuren, S. (2012). Flexible imputation of missing data. CRC press.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. Journal of statistical computation and simulation, 76(12), 1049-1064.

Van Ginkel, J. R., Linting, M., Rippe, R. C., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. Journal of personality assessment, 102(3), 297-308.

Zhao, Y., & Long, Q. (2016). Multiple imputation in the presence of high-dimensional data. Statistical Methods in Medical Research, 25(5), 2021-2035.

ABSTRACT

Multiple Imputation Model Selection for High-Dimensional Missing Data

Yoonah Lee
Department of Statistics
Graduate School of
Sungshin Women's University

Handling missing data by simply removing them can result in bias or loss of precision in the results of the analysis. As an alternative, multiple imputation has been commonly used. However, most of the existing imputation models do not work well in high-dimensional data due to the curse of dimensionality. We propose a novel variable selection approach derived from the idea of graph theory. Our method, which combines graphical lasso and inverse probability weighting estimator, is a computationally efficient and flexible framework. We demonstrate that our proposed method performs comparably or even outperforms the existing high-dimensional imputation models, namely DURR and IURR, under the assumption of normality in the simulation studies.