한국심리학회지: 건강
The Korean Journal of Health Psychology
2022. Vol. 27, No. 3, 513 - 532

잠재프로파일 분석에서 결측값 처리를 위한 최근접이웃 대체법의 활용†

김 수 민
부산대학교
심리학과 석사과정

조 승 빈‡
부산대학교
심리학과 조교수

잠재프로파일분석(latent profile analysis: LPA)은 모집단에 존재하는 비슷한 특성을 공유하는 개인들로 구성된 하위집단을 확인하기 위해 심리학의 여러 분야에서 흔히 사용되는 모형이다. 결측값이 존재하는 자료에 잠재프로파일분석을 적용하기 위해 가장 권장되는 방법은 완전정보최대우도법(full information maximum likelihood: FIML)이다. 본 연구에서는 비교적 간단한 알고리즘으로 이루어진 k-최근접이웃(k-nearest neighbor: kNN) 대체법을 LPA에서 결측값을 처리하기 위한 효율적인 대안으로 제안하고 시뮬레이션 자료를 통해 kNN 대체법의 활용 가능성을 검증하였다. 결측값 생성 메커니즘, 결측률, 하위집단 간 거리, 표본 크기를 변화시켜 자료를 생성하고 생성된 자료에 kNN 대체법 적용하고 LPA를 수행한 결과와 FIML을 통한 결과를 추정한 하위집단 수, 하위집단 평균 프로파일의 정확도, 분류의 품질을 기준으로 비교하였다. 분석 결과, 하위집단 수의 추정은 대부분의 조건에서 kNN 대체법이 FIML과 비슷한 정확도를 보였으며 하위집단 간 거리가 가깝고 표본크기가 작은 조건에서는 더 우수한 결과를 보였다. 하위집단 프로파일의 정확도는 두 방법 간에 일관성 있는 차이를 발견할 수 없었다. 분류의 품질은 거의 모든 조건에서 kNN 대체법을 적용한 결과가 완전자료에서 얻어진 결과에 가까웠다. 본 연구는 LPA를 위한 kNN 대체법의 활용 가능성을 확인한 최초의 연구로서 의의를 가진다. 본 연구의 결과를 통해 FIML을 통한 분석이 어려운 조건에서 결측값 처리를 위한 대안으로서 kNN 대체법의 활용 뿐만 아니라 kNN 대체법과 FIML의 결과를 비교함으로써 LPA 결과의 신뢰도를 확인하는 방식으로 사용할 것을 제안한다.

주요어: 잠재프로파일분석, k-최근접이웃대체법, 결측값, 시뮬레이션

† 본 논문은 김수민의 석사 학위 청구논문을 보완 및 수정한 것임.

‡ 교신저자(Corresponding author): 조승빈, (46241) 부산광역시 금정구 부산대학로 63번길 2(장전동) 심리학과 조교수, Tel: 051-510-2144, E-mail: chosngbin@pusan.ac.kr

심리학 자료에서 결측값의 발생과 처리는 피할 수 없는 문제이다. 특히 사람을 대상으로 하는 심리학 자료의 특성상 응답의 거부, 응답자의 실수, 정제되지 않은 측정 도구, 측정 내용의 민감성, 대상의 탈락 등 다양한 이유로 결측값이 발생할 수 있다. 대부분의 통계적 분석 모형과 모수의 추정 방법은 완전한 자료를 가정하기 때문에 결측값의 발생은 모형의 가정 위배, 모수 추정의 편향, 검정력의 저하 등 자료 분석의 과정에서 여러 가지 문제를 발생시킨다(Newman, 2014; Schafer & Graham, 2002). 따라서 심리학 자료에 존재하는 결측값에 대한 처리 방법은 심리학 자료의 적절한 분석을 위해 많은 관심을 받고 있다.

잠재프로파일분석(Latent Profile Analysis: LPA, Lazarsfeldt & Henry, 1968)은 잠재혼합모형(Latent Mixture Models)의 한 종류로 연속형 지표 변인을 바탕으로 전체 표본으로부터 유사한 평균벡터와 공분산의 프로파일을 가지는 하위집단을 추출한다. LPA를 통해 표본에 존재하는 하위 집단의 평균 프로파일과 공분산 행렬뿐만 아니라 분포 가정에 기반하여 각 관찰값이 추정된 하위집단에 소속될 확률을 추정할 수 있다. 표본에 존재하는 하위집단을 추출하는 방법은 LPA 외에도 위계적 군집분석, k-평균군집분석(K-Means Clustering)등의 다양한 군집분석(cluster analysis)방법이 존재한다. 그러나 이러한 비모수 방법과는 달리 LPA는 연구자가 설정한 모형에 기반한(model-based) 방법으로, 하위집단 내 지표변인의 분포에 대한 가정(일반적으로 다변량 정규분포)을 바탕으로 모수를 추정하고 추정된 모수에 대한 가설의 검정이 가능하다.

이러한 장점으로 인해 건강심리학을 비롯하여 교육, 상담, 산업 및 조직 등 심리학의 다양한 분야에서 표본 내에 이질적인 특성을 가진 하위집단의 존재를 가정할 수 있는 경우 이러한 하위집단을 확인하고 하위집단과 관련된 가설을 검정하기 위해 LPA의 활용이 최근 들어 꾸준히 증가하고 있다. 예를 들어, 복잡한 양상을 보이는 문제행동이나 건강 지표에 따른 하위 집단을 확인하거나(김시형, 신지영, 이동훈, 2019; 조다빈, 심은정, 2021) 이러한 하위 집단과 외부 변인 간의 관계를 확인함으로써 척도의 타당화를 위해 쓰이기도 한다(박경우, 장혜인, 2021). 교육심리학 분야에서는 중, 고등학생의 부모화 유형을 확인하기 위해 LPA를 사용하였으며(최현주, 2021), 최현주와 장은비(2021)는 특수교사가 지각하는 학교조직풍토의 유형을 분류하기 위하여 LPA를 사용하였다. 또한 염소란과 김명소(2017)는 상사에 대한 종업원의 정서조절전략의 하위유형을 분류하고 이러한 하위유형과 관련된 선행변인과 결과변인 간 연관성에 대한 가설 검정을 위해 LPA를 사용하였다.

LPA 또한 결측값으로 인한 문제에서 자유롭지 않다. 결측값의 존재는 잠재변인 모형에서 평균과 공분산의 추정에서 편향을 일으킨다(Enders, 2001). LPA 또한 다른 잠재변인 모형과 마찬가지로 모수의 추정에 있어 지표변인의 평균벡터와 분산-공분산 행렬의 정보를 바탕으로 하며 결측값의 존재는 LPA의 모수 추정에 편향과 검정력의 저하로 이어진다. 결측값이 포함된 자료의 분석에서 발생하는 이러한 문제를 최소화하기 위해 LPA를 비롯한 잠재변인 분석뿐만 아니라 많은 통계적 분석에서 완전정보최대우도법(Full Information Maximum Likelihood: FIML)과 다중

대체법(Multiple Imputation: MI)을 가장 많이 권장한다(신택수, 2014; Lee & Shi, 2021).

따라서 본 연구에서는 결측값을 포함하는 자료에 대한 LPA를 위한 결측값 처리 방법의 대안으로서 k-최근접이웃 대체법(k-Nearest Neighbor Imputation: kNN 대체법)의 가능성을 탐색한다. kNN 대체법의 활용 가능성을 알아보기 위해 현재 결측값을 포함하는 자료에 대한 LPA를 위해 가장 널리 받아들여지고 있는 FIML의 결과와 kNN 대체법을 적용한 결과를 비교한다. 따라서 이 논문의 나머지 부분은 다음과 같이 구성된다. 먼저 kNN 의 알고리즘에 대해 개괄하고, 그 다음 이를 이용해 결측값을 대체하는 방법을 소개한다. 다음으로 시뮬레이션 자료를 통해 kNN 대체법을 적용한 결과와 FIML을 적용한 결과 간 비교를 통해 LPA에서 결측값을 처리하는 방법으로서 kNN 대체법의 가능성에 대해 논의한다.

완전정보최대우도법

FIML은 결측값을 포함하는 자료를 통해 모수를 추정하기 위해 사용되는 방법이다. FIML은 결측값을 적절한 값으로 대체하거나 삭제하는 등 자료에 변형을 가하는 것이 아니라 무선결측(Missing At Random, MAR) 가정을 바탕으로 결측값을 포함하는 주어진 자료에 대한 최적의 모수를 추정한다. FIML은 자료 전체의 가능도(likelihood)를 계산하기 위해 각 관찰값이 가진 가용한 정보를 이용한다. 결측값을 포함하는 관찰값의 경우 남아 있는 정보를 통해 계산된 가능도가 전체 모형의 가능도의 계산에 기여한다. 결측값이 존재하는 자료에 대해 FIML 방법을 통해 추정된 모수는 가장 편향이 적은 것으로 알려져 있으며 (Lee & Shi, 2021), 같은 자료에 대한 분석에서 도출되는 값의 편차가 적다는 장점이 있다(Enders, 2001; Enders & Bandalos, 2001). 또한 FIML은 일반적으로 정규분포를 가정하지만 자료가 정규성에서 어느 정도 벗어나는 경우에도 편향되지 않은 결과를 내놓는 것으로 확인되었다(Enders, 2001). 특히 결측값을 포함하는 자료에 잠재 변인 모형을 적용할 때 FIML을 사용할 경우 완전제거법(listwise deletion)이나 대응제거법(pairwise deletion)과 같은 제거 방법에 비해 모형 수렴률, 모수 추정 편향 및 정확도, 최적 모형 적합도 등에서 더 나은 결과를 보이는 것으로 알려졌다(Enders & Bandalos, 2001). 이러한 이유로 FIML은 결측값이 있는 자료에 LPA를 적용할 때 일반적으로 권장하는 방법이다(Collins & Lanza, 2009; Enders, 2001; Hagenaars & McCutchoen, 2002).

MI 또한 FIML과 같이 MAR 가정을 바탕으로 하지만, FIML이 반복 과정을 통해 수렴되는 최적의 모수를 찾는 과정인 반면, MI는 관찰된 자료를 바탕으로 추정된 변인의 분포에서 추출된 값으로 결측값을 대체하여 완전한 자료를 생성한다. MI는 그 이름이 의미하듯이 이러한 완전한 자료를 여러 개 생성하여 각각의 자료로부터 주어진 모형의 모수를 추정한 후 생성된 모든 자료에서 분석된 결과를 통합한다(Little, 2014). 다중대체법으로 추정된 모형의 모수 또한 편향이 적고 추정치의 표준오차가 적은 것으로 알려져 있다(Enders, 2001; Enders & Bandalos, 2001; Lee & Shi, 2021). 이러한 장점으로 인해 MI 방법 또한 결측값이 존재하는 자료의 분석을 위해 FIML과

함께 가장 많이 사용하는 방법이다.

결측값이 존재하는 자료의 분석을 위한 FIML과 MI 방법은 위에서 열거한 바람직한 특성에도 불구하고 LPA를 위한 결측값 처리 방법으로서 사용을 제한할 수 있는 몇 가지 특성이 존재한다. 다중대체법의 경우, 생성된 각 자료에서 추정된 모수를 통합하는 과정을 거치는데(Little, 2014), 다중대체법을 LPA에 적용할 경우, 생성된 자료들에서 추정한 군집의 수나 프로파일이 일치하지 않을 수 있다. 이러한 경우에 서로 다른 자료에서 얻어진 추정치를 논리적으로 통합할 방법이 없다는 결정적인 문제를 가지고 있다(Collins & Lanza, 2009). FIML은 모형과 결측률에 따라 모수의 추정이 영향을 받게 되며 특히 결측률이 높을수록 모수 추정의 수렴률이 낮아진다(Yung & Zhang, 2011). 또한, 모수의 추정에 상대적으로 많은 시간이 필요한데, 이러한 특성은 결측값을 대체한 자료를 생성하는 것이 아닌 매 분석마다 모수 추정의 과정을 거쳐야 하는 FIML의 특성과 맞물려 같은 자료에 대해 여러 번의 분석이 이루어질 때 단점으로 두드러질 수 있다. 또한 FIML을 적용하기 위해서는 상대적으로 큰 표본크기가 필요하다(Enders, 2001).

k-최근접이웃법

k-최근접이웃법(k-Nearest Neighbor: kNN)은 예측과 분류를 위한 기계 학습 방법 중 하나이다. 선형회귀와 같은 전통적인 통계적 예측방법과는 다르게 kNN은 예측값을 도출하기 위해 모형을 설정하지 않는다. 그 대신, 주어진 값 주변의 k 개의 관찰값의 정보를 이용한다. 근접 이웃을 찾기 위해서는 관찰값 간의 거리를 계산하는데, 연속변인으로 이루어진 관찰값의 경우 가장 흔히 이용되는 거리 척도(distance measure)는 유클리드 거리(Euclidean distance)이다. p 개의 변인 $X_1 \cdots X_p$ 으로 이루어진 관찰값 i 와 j 간 유클리드 거리 d_{ij}는 다음과 같이 계산한다.

$$d_{ij} = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 \cdots (X_{ip} - X_{jp})^2}$$
$$= \sqrt{\sum_{o=1}^{p} (X_{io} - X_{jo})^2}$$

kNN은 주어진 관찰값으로부터 계산된 거리를 바탕으로 k 개의 주변 값들의 평균이나 중앙값을 통해 예측값을 추정한다(James, Witten, Hastie, & Tibshirani, 2013). 즉, $X_{1i}, \ldots X_{pi}$ 를 통해 Y_1 을 예측하기 위해 $X_{1i}, \ldots X_{pi}$ 의 값을 기준으로 거리 척도 상 가장 가까운 k 개의 이웃을 찾고, 그 이웃들의 Y 값의 중앙값 또는 평균으로 Y_1을 예측한다. kNN은 이러한 단순한 알고리즘으로 이루어져 있지만 예측과 분류에서 상당한 정확성을 보이는 것이 확인 되었다(예를 들어, Ali et al., 2021; Goyal, Chauhan, & Parveen, 2016; Park, Kim, & Kwon, 2014).

k-최근접 이웃 대체법: 결측값 대체를 위한 kNN의 활용

인접한 관찰값을 토대로 결과값을 예측하는 kNN의 알고리즘을 결측값의 대체를 위해 사용할 수 있는데 이를 k-최근접이웃 대체법(k-Nearest Neighbor Imputation: kNN 대체법)이라 한다. 결측값의 처리를 위해 kNN을 사용할 때 예측의 대

상은 주어진 관찰값에서 결측된 변인이다. 결측값의 예측을 위해서는 예측이나 분류를 위해 kNN을 사용할 때와 마찬가지로 참가자의 결측되지 않은 변인을 기반으로 가장 가까운 k 개의 이웃을 결정하고 그 이웃들이 가진 결측값에 해당하는 변인의 평균이나 중앙값으로 주어진 관찰값의 결측 변인의 값을 대체한다(Beretta & Santaniello, 2016).

결측값의 처리를 위한 방법으로서 kNN 대체법의 가능성을 확인한 연구들이 존재한다. kNN 대체법은 단일대체법의 일종이지만 kNN에 기반한 대체값은 평균대체법이나 회귀대체법과 같은 다른 단일대체법에서 추정한 대체값에 비해 편향이 적은 것으로 알려졌다. 전체 자료에 기반 하는 평균대체법이나 회귀대체법의 경우 대체값이 평균으로 회귀하는 편향이 발생하거나 변인 간 상관계수를 과대추정 하는 편향이 발생하기도 하지만(신택수, 2014) kNN 대체법의 경우 인접한 일부의 자료만을 활용하기 때문에 이러한 편향이 상대적으로 덜하다(Chen & Shao, 2000; Jonsson & Wohlin, 2004). 또한 자료 내에 존재하는 값 중 가장 가까운 관찰값들을 토대로 결측값을 추정하기 때문에 결측값이 자료에 실제로 존재할 수 있는 값으로 대체된다는 장점이 있다(Beretta & Santaniello, 2016). kNN 대체법은 비모수 방법이기 때문에 자료에 대해 특정 분포를 가정하지 않으며 분석 모형의 영향을 받지 않는다. 또한 연산이 복잡하지 않으며, 차원이 큰 자료에도 사용 가능하다는 장점을 가진다(노민정, 유진은, 2019). kNN 대체법과 다른 결측값 처리 방법의 수행을 비교한 연구에서 kNN 대채법은 다른 방법에 비해 우수한 수행을 보여주었다. 예를 들어 정규화 회귀(regularized regression) 모형에서 kNN 대체법은 EM(expectation maximization) 알고리즘, 제거법 등에 비해 대체 정확도, 변수 선택, 예측 정확도에서 우수한 것으로 나타났으며(노민정, 유진은, 2019), 기계 학습을 위한 결측값 처리 방법으로서 kNN 대체법은 의사결정 나무 기반의 대체법이나 다른 단일 대체법에 비해 우수한 수행을 보여주었다(Batista & Monard, 2003). 그러나 LPA에서 위한 결측값의 처리를 위한 kNN 대체법의 수행을 확인한 연구는 아직까지 이루어지지 않았다.

관찰값 간의 거리를 기반으로 가까운 관찰값들을 바탕으로 결측값을 추정하는 kNN 대체법의 특성은(James et al., 2013), 관찰값, 즉 참가자 간의 관계에 기반하는 잠재프로파일분석(Latent Profile Analysis: LPA)이나 잠재계층모형(Latent Class Analysis: LCA, Collins & Lanza, 2009; Hagennars & McCutcheon, 2002)와 같은 참가자 중심 접근(person-oriented approach)의 특성과 일치한다. LPA의 경우 관찰된 지표변인을 바탕으로 표본 내에 존재하는 하위집단을 추정한다. 자료 내에 하위집단이 존재할 때 같은 하위 집단에 속하는 참가자들이 가까운 거리에 존재할 가능성이 높다. 또한 LPA의 경우, 지표 변인들에 대한 분포 가정을 바탕으로 참가자들을 하위집단으로 판단하는데 가까운 거리의 이웃일수록 같은 분포 모형을 따를 확률이 높다(Steinley & Brusco, 2011). 따라서 가까운 거리에 있는 참가자일수록 같은 집단에 속할 확률이 높기 때문에, 관찰값 간의 거리를 통해 측정한 유사성을 바탕으로 대체값을 생성하는 kNN 대체법이 LPA를 위한 결측값의 처리에서 좋은 성능을 보일 것이라고 기대

할 수 있다. 실제로 kNN 대체법은 분류(classification) 모형을 위한 결측값 처리에서 좋은 결과를 보여주었다(García-Laencina, Sancho-Gómez, Figueiras-Vidal, & Verleysen, 2009; Pujianto, Wibawa, & Akbar, 2019). Troyanskaya 등(2001)은 결측이 있는 유전체 자료를 처리하기 위해 kNN 대체법과 특이값 분해 방법(Singular Value Decomposition) 기반 결측값 대체법을 사용하였는데 kNN 대체법을 통해서 결측값을 처리했을 때 대체값의 정확도가 높았으며 이러한 정확도는 결측값의 비율의 영향을 적게 받는 것으로 나타났다. 또한 kNN 대체법으로 생성된 자료는 유전체의 군집 특성을 잘 유지하였다. Keerin, Kurutach와 Boongoen(2012)은 유전체 자료에서 군집을 활용한 kNN을 활용하여 결측값을 대체하는 방법을 개발하고 그 방법을 평가하였는데, kNN 대체법으로 추정한 대체값이 다른 대체법에 비해 실제값에 더 가까운 것으로 확인되었다.

또한, 비교적 단순하고 직관적인 알고리즘으로 이루어진 비모수 방법인 kNN 대체법은 결측값을 포함하는 LPA에서 기존에 권장되는 FIML과 MI와 같은 모수적 결측값 처리 방법이 가지는 몇 가지 한계에서 자유롭다. 전술한 모수 통합의 문제로 인해 결측값을 포함하는 자료에 대한 LPA에서 가장 권장되는 방법은 FIML이며 FIML은 잠재변인 모형에서 대체로 좋은 수행을 보인다(Enders, 2001; Enders & Bandalos, 2001). 그러나 FIML은 지표변인들의 다변량 정규분포의 가정이 필요하며 모수의 추정이 모형의 영향을 받는다. 비모수 방법인 kNN 대체법은 FIML이 가지는 지표변인의 분포에 대한 가정을 가지지 않기 때문에(Enders, 2001) 분포 가정의 위배에 의한 영향을 상대적으로 덜하다. 또한 kNN 대체법은 분석 모형에 기반하지 않기 때문에 결측값의 대체값이 분석 모형에 따라 달라지지 않는다. 뿐만 아니라 kNN 대체법은 그 결과로 결측값이 없는 완전한 데이터 한 벌이 생성되기 때문에 매 분석 시 마다 결측값을 처리하는 과정을 반복할 필요가 없으며(Yung & Zhang, 2011) 복수의 생성된 자료에서 추정된 모수를 통합하는 과정을 거치지 않으므로 다중대체법이 가지는 문제에서도 자유로울 수 있다. 이러한 사실들은 결측값이 있는 자료에 LPA를 적용하기 위한 결측값 처리 방법으로 kNN 대체법이 일반적으로 권장되는 FIML에 대한 효율적인 대안이 될 수 있으며 이를 검증해 볼 필요가 있음을 시사한다. 그러나 kNN 대체법을 기존의 결측값 처리 방법과 비교한 연구는 아직 이루어지지 않았다. 따라서, 본 연구에서 LPA를 위한 자료의 결측값 처리 방법으로서 kNN 대체법의 가능성을 확인하고자 한다. 구체적으로, 결측값이 존재하는 자료에 대해 kNN 대체법을 적용하여 완전자료를 생성하고 이러한 완전자료에 대한 LPA를 실행한 결과와 같은 자료에 FIML을 적용한 LPA의 결과를 비교한다. 이러한 비교를 통해 일반적으로 권장되는 방법인 FIML의 대안으로서 kNN 대체법의 가능성을 확인하고자 한다.

방 법

본 연구에서는 몇 가지 조건을 변화시키면서 자료를 생성하고 이러한 자료에 의도적으로 결측값을 발생시켜 결측 자료를 만들었다. 이러한 결측 자료에 kNN 대체법으로 대체값을 생성하고 LPA를 적용한 결과와 FIML을 통해 LPA를 적용

한 결과를 비교한다. 또한 생성된 자료에 결측값을 발생시키지 않은 완전 자료에 LPA를 적용한 결과를 kNN 대체법과 FIML의 결과와 비교한다. 연구 방법의 개요는 그림1에 요약되어 있으며 이어지는 내용에서 구체적인 절차가 상술되어 있다.

시뮬레이션 데이터의 생성

kNN 대체법과 FIML의 결과를 비교하기 위해 두 방법을 적용시킬 자료를 생성한다. 생성된 데이터는 8개의 지표변인으로 이루어져 있으며 하위집단의 수는 3개, 하위집단의 크기는 동일하도록 설정되었다. 하위집단 내 분산-공분산 행렬은 단위행렬로 설정되었다. 데이터 생성에서 변화된 조건은 하위집단 간 거리, 결측값의 비율, 데이터의 크기이다. 하위집단 간 거리는 하위집단 간 겹침의 정도와 관련되며 분류의 결과에 영향을 미치는 중요한 파라메터이다(Tein, Coxe, & Cham, 2013). 하위집단의 범위가 동일할 경우, 즉 정규분포를 가정한 하위집단의 분산이 동일한 경우에 하위집단 간 거리가 가까울수록 겹침의 정도가 크며(송주원, 2017) 오분류율(misclassification

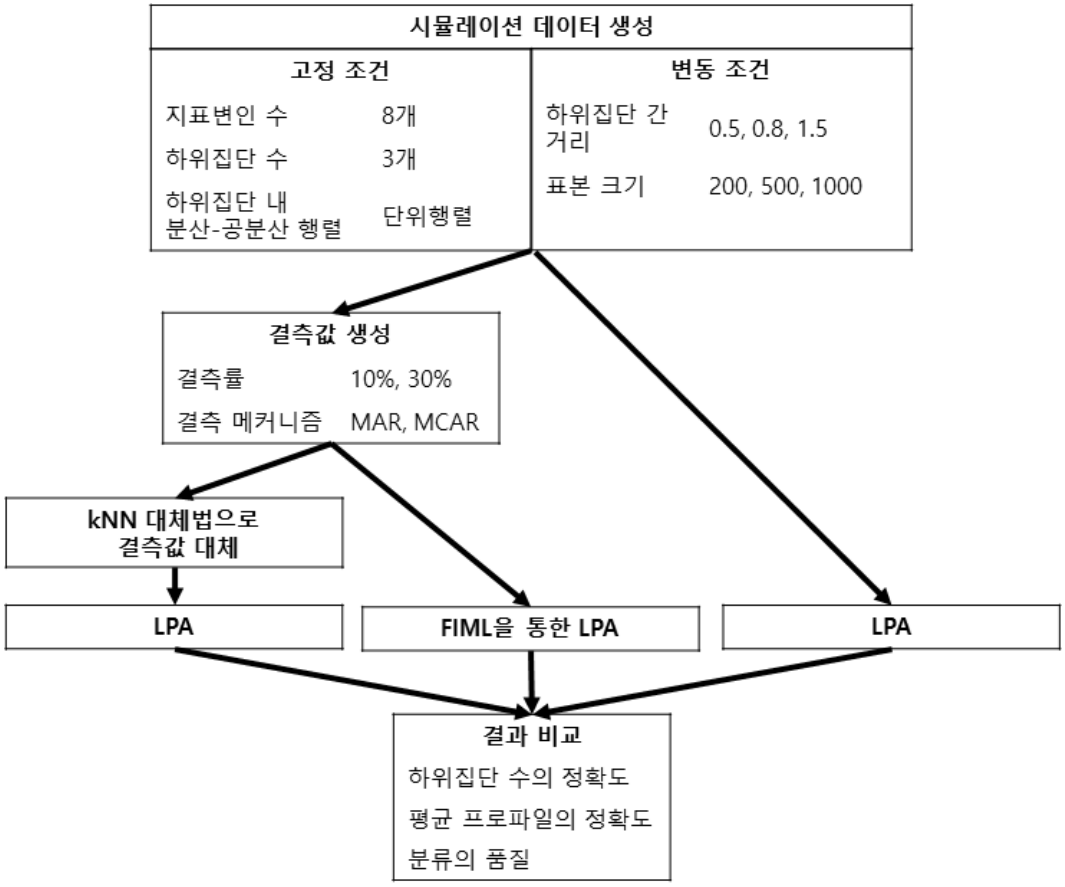

그림 1. 연구 방법의 개요

rate)이 높아진다. Tein 등(2013)의 시뮬레이션 연구에서 하위집단 간 거리가 LPA의 결과에 가장 큰 영향을 미치는 파라메터로 나타났으며 하위집단의 수가 3개이고 하위집단 내 분산-공분산행렬이 단위행렬일 때, 하위집단의 평균 프로파일이 (-0.5, 0, 0.5), (-0.8, 0, 0.8), (-1.5, 0, 1.5)인 경우를 각각 효과크기가 “보통”, “높음”, “매우높음”으로 정의하였다. 본 연구에서는 이러한 결과를 참고하여 하위집단 중심점 간 거리를 효과크기를 기준으로 보통(-0.5, 0, 0.5), 높음(-0.8, 0, 0.8), 매우 높음(-1.5, 0, 1.5)으로 설정하였다. 표본 크기 또한 LPA의 결과에 영향을 미치는 변수이며 본 연구에서는 Tein 등(2013)과 Nylund, Asparouhov와 Muthén(2007)에서 사용한 조건을 참고하여 200, 500, 1000으로 설정하였다.

결측값의 생성을 위해 결측값의 비율과 결측값 생성 메커니즘의 두 가지 조건을 변화시켰다. 결측값 처리를 위한 FIML의 수행을 확인한 연구들에서 5-25%의 결측률을 사용하였고(Enders, 2001, Lee & Shi, 2021), kNN 대체법의 수행을 확인한 연구들에서 5-30%의 결측률을 사용한 것을 참고하여(노민정, 유진은, 2019; Beretta & Santaniello, 2016; Troyanskaya et al., 2001) 결측값의 비율은 10%와 30%의 두 가지 조건으로 설정하였다. 결측값의 생성 메커니즘은 MCAR(missing at completely random)과 MAR(missing at random)의 두 가지의 가정에 기반하였다. MCAR 가정으로 결측값을 생성하는 경우에는 결측값이 주어진 비율에 따라 자료의 어느 부분에서든 동일한 확률로 발생되었다. MAR 가정에 따른 결측값의 생성을 위해 첫 번째 지표 변인의 값에 따라 나머지 지표 변인의 결측 확률이 결정되도록 설정하였다. 구체적으로, 주어진 관찰값의 첫 번째 변인의 값이 그 중앙값보다 크거나 작을 경우 나머지 변인에 결측이 발생하며 첫 번째 변인의 값이 중앙값보다 큰 경우의 결측 확률과 작은 경우의 결측 확률의 비율은 1에서 10배까지 매 반복마다 무선적으로 변화하도록 설정하였다. 각 조건에 대해 100회의 반복 데이터를 생성하고 분석하였다.

분석 및 결과의 비교

변화된 조건에 따라 총 36개의 조합으로 자료가 생성되었으며 각 조건에서 100회의 반복이 이루어졌다. 각 반복마다 FIML과 kNN 대체법을 적용하여 분석한 결과들을 비교하였다. 구체적으로, FIML을 적용한 경우에는 결측값을 생성시킨 자료의 모수를 FIML 추정 방법을 통해 추정한다. FIML을 통한 모수의 추정은 Mplus의 최대우도추정 방법(analysis 명령어의 type=MIXTURE)을 사용하여 이루어졌다. 같은 자료에 대해 kNN 대체법을 적용하는 경우에는 kNN 대체법을 통해 완전한 자료를 생성하고 이 자료를 Mplus에서 최대우도추정 방법을 통해 모수를 추정하였다. kNN 대체법을 사용하는 경우 대체값을 계산하기 위한 이웃의 개수를 나타내는 파라메터인 k와 관찰값 간 거리 계산을 위한 거리 척도를 사용자가 설정하여야 하는데 본 연구에서는 Jonsson과 Wohlin(2004)의 제안에 따라 매 반복에서 결측이 발생하지 않은 관찰값의 수의 제곱근을 k로 설정하였으며, 유클리드 거리를 활용하여 관찰값 간의 거리를 계산하였다. kNN 대체법을 통한 완전자료의 생성은 R의 VIM 패키지(Kowarik & Templ, 2016)를 사용하여 이루어졌다.

두 방법으로 얻어진 결과의 비교를 위한 주요 준거는 추정된 하위 집단의 수와 하위 집단의 평균 프로파일의 정확성이다. 이러한 준거에 있어 kNN 대체법과 FIML 두 방법의 수행을 비교하였을 뿐만 아니라 결측값을 생성하지 않은 완전자료에서 얻은 결과와 비교를 통해 두 방법의 수행을 평가하였다. 엔트로피(Entropy)를 통해 두 방법을 통해 얻어진 결과의 분류의 불확실성 또한 비교하였다. 각 반복에서 집단의 수를 결정하기 위해 하위집단의 수가 2개에서 4개인 모형을 적합하고 Bayesian Information Criteria(BIC, Schwarz, 1978)와 Bootstrap Likelihood Ratio Test(BLRT, McLachlan & Peel, 2004) 의 두 가지 준거를 사용하여 최적의 하위집단 수를 결정하였다. BIC와 BLRT는 이전 연구에서 하위집단의 수를 결정하는데 있어 다른 방법에 비해 안정적인 수행을 보인 것으로 알려졌다(Nylund et al., 2007). 본 연구에서는 각 준거에 따라 결정된 하위집단의 수가 자료 생성시에 설정한 3개로 나오는 빈도를 완전한 자료의 경우, FIML을 사용한 경우, kNN 대체법을 사용한 경우에 대해 계산하였다. BIC를 기준으로 하위집단의 수를 결정할 경우 2개에서 4개의 하위집단을 가지는 LPA 모형 중에서 3 집단 모형의 BIC가 가장 낮을 때 3 집단 모형을 채택하였다. BLRT의 경우 k개와 k-1개의 하위집단의 로그 가능도(log likelihood)를 비교하며 k개의 하위집단을 가지는 모형이 k-1개의 하위집단을 가지는 모형에 비해 가능도의 향상이 95% 유의수준에서 유의미하지 않은 경우, 즉 BLRT의 p 값이 0.05 이상일 때 하위집단의 증가가 더 이상의 적합도 향상으로 이어지지 않는다고 결론 내린다. 따라서, 본 연구에서는 2 집단 모형과 3 집단 모형 간 BLRT의 p 값이 0.05 미만이고 3집단 모형과 4집단 모형 간의 p 값이 0.05 이상일 때 3 집단 모형을 채택하였다. 엔트로피의 경우 1에 가까운 값을 가질수록 분류의 불확실성이 낮음을 의미한다(Celeux & Soromenho, 1996). 하위집단의 평균벡터의 정확성은 완전자료에서 추정된 하위집단의 평균 벡터와 추정된 하위집단 평균 벡터 간 RMSE(root mean squared error)를 통해 확인하였다. RMSE는 지표 변인의 수가 k개라고 할 때 다음과 같이 계산되며 작은 값을 가질수록 완전자료에서 추정한 평균 벡터와 유사함을 의미한다.

$$RMSE(V_1, V_2) = \sqrt{\frac{\sum_{i=1}^{k}(V_{1i} - V_{2i})^2}{k}}$$

여기서 V_1은 완전자료에서 추정된 하위집단의 평균벡터이고 V_2는 추정된 하위집단의 평균벡터이다. RMSE의 계산은 BIC와 BLRT를 기준으로 추정된 하위집단의 수와 상관없이 3개 하위집단 모형을 기준으로 계산하였다.

결 과

시뮬레이션 자료에 kNN 대체법과 FIML을 적용한 결과는 표 1(하위집단 간 거리 0.5), 표 2(하위집단 간 거리 0.8), 표 3(하위집단 간 거리 1.5)에 정리되어 있다. 표 1에서 3의 kNN, FIML 열은 각각 kNN 대체법과 FIML을 적용한 결과를 나타내며 BIC와 BLRT 행은 각 조건 별 반복 중에서 최적 하위집단 수를 3개로 추정한 빈도, Entropy 행은 각 조건의 3개 하위집단 모형에서

계산한 엔트로피의 반복 간 평균값이다. BIC, BLRT, Entropy 행은 완전자료 분석의 결과 또한 포함한다(완전자료 열). RMSE 행은 각 조건별 반복에서 추정한 하위집단 평균 프로파일과 완전자료가 추정한 하위집단 평균 프로파일 간 RMSE의 반복 간 평균이다.

표 1은 효과크기가 보통인 경우, 즉 하위집단 간 거리가 0.5일 때 하위집단의 수에 대한 추정의 결과이다. 전체적으로, 효과크기가 작은 경우에는 kNN 대체법 또는 FIML 중 어떤 방법을 사용해도 정확한 하위집단의 수를 추정하기가 어려운 것으로 나타났지만 kNN 대체법을 적용한 자료에 BLRT를 기준으로 하위집단의 수를 추정한 경우의 정확도가 비교적 높은 것으로 나타났다. 표 1에서 완전자료를 사용하였을 때에도 하위집단의 수를 정확하게 추정하는 경우는 최소 0, 최대 16에 불과했으며, 하위집단 수의 추정이 정확한 빈도가 가장 높았던 조건, 즉 kNN 대체법 사용, n = 500, 10% 결측률, MAR 메커니즘, BLRT 기준의 경우에도 하위집단의 수를 정확히 추정한 빈도는 총 100회의 반복 중에서 34회에 불과했다. BIC를 기준으로 한 하위집단 수의 추정이 정확한 경우는 최대 8회로 대부분의 조건에서 정확한 하위집단 수의 추정에 실패했으며, BLRT를 기준으

표 1. 하위집단 간 거리가 0.5일 때 BIC, BLRT, Entropy 및 RMSE 계산 결과

하위집단 간 거리 0.5										
	완전자료	결측률 10%				완전자료	결측률 30%			
		MAR		MCAR			MAR		MCAR	
		kNN	FIML	kNN	FIML		kNN	FIML	kNN	FIML
n = 200										
BIC	0	0	0	0	0	0	0	0	2	0
BLRT	9	**13**	7	**20**	8	3	**31**	4	**30**	2
Entropy	0.70	0.70	0.68	0.69	0.67	0.69	0.73	0.62	0.71	0.62
RMSE		0.48	0.45	**0.52**	0.56		**0.71**	0.71	0.76	0.76
n = 500										
BIC	0	0	0	0	0	0	6	0	1	0
BLRT	16	**34**	17	**31**	13	10	6	7	**21**	8
Entropy	0.65	0.66	0.64	0.66	0.63	0.64	0.69	0.57	0.64	0.58
RMSE		0.40	0.38	**0.45**	0.46		**0.61**	0.65	**0.62**	0.72
n = 1000										
BIC	0	0	0	0	0	0	8	0	7	0
BLRT	14	**25**	13	**28**	16	7	1	5	**11**	6
Entropy	0.60	0.61	0.59	0.61	0.58	0.59	0.66	0.54	0.59	0.51
RMSE		0.34	0.33	0.49	0.43		**0.56**	0.57	**0.54**	0.59

BIC: Bayesian information criteria; BLRT: Bootstrap likelihood ratio test; RMSE: Root mean squared
MAR: Missing at random; MCAR: Missing completely at random
kNN: k-nearest neighbor imputation;
FIML: Full information maximum likelihood method

표 2. 하위집단 간 거리가 0.8일 때 BIC, BLRT, Entropy 및 RMSE 결과

하위집단 간 거리 0.8										
	완전자료	결측률 10%				완전자료	결측률 30%			
		MAR		MCAR			MAR		MCAR	
		kNN	FIML	kNN	FIML		kNN	FIML	kNN	FIML
n = 200										
BIC	7	**8**	5	**10**	3	2	**16**	2	**16**	0
BLRT	48	**58**	44	**52**	43	53	**31**	27	**26**	22
Entropy	0.73	0.73	0.72	0.75	0.72	0.73	0.78	0.71	0.77	0.70
RMSE		0.35	0.34	**0.41**	0.45		**0.60**	0.68	0.63	0.62
n = 500										
BIC	40	**43**	26	**50**	21	36	**71**	4	**69**	1
BLRT	80	79	82	76	77	88	10	63	18	62
Entropy	0.67	0.67	0.65	0.67	0.64	0.66	0.71	0.60	0.69	0.60
RMSE		**0.24**	0.27	**0.30**	0.34		0.48	0.47	**0.43**	0.53
n = 1000										
BIC	100	**100**	89	**100**	93	94	**89**	25	**99**	9
BLRT	81	71	87	78	90	94	0	89	4	85
Entropy	0.66	0.66	0.63	0.67	0.63	0.65	0.69	0.56	0.69	0.55
RMSE		0.32	0.28	**0.31**	0.32		0.43	0.42	**0.46**	0.48

BIC: Bayesian information criteria; BLRT: Bootstrap likelihood ratio test; RMSE: Root mean squared
MAR: Missing at random; MCAR: Missing completely at random
kNN: k-nearest neighbor imputation;
FIML: Full information maximum likelihood method

로 하위집단의 수를 추정한 경우 kNN 대체법을 사용했을 때 하위집단의 수를 정확하게 추정한 경우의 수가 BIC를 기준으로 한 조건보다 전체적으로 높음을 확인할 수 있다. 결측값 발생 기제에 따른 일관적인 양상은 발견하지 못했다.

높은 효과크기 수준에 해당하는 하위집단 간 거리가 0.8일 때의 하위집단 수 추정의 결과는 표 2에 정리되어 있다. 먼저, 정확한 하위집단의 수를 추정한 빈도가 하위집단 간 거리 0.5인 조건에 비해 증가한 것을 확인할 수 있다. 또한 BIC를 기준으로 하위집단의 수를 결정한 경우에는 모든 조건에서 kNN 대체법을 사용한 경우가 FIML을 사용한 경우에 비해 정확한 하위집단의 수를 추정한 빈도가 더 높았다. 그에 비해 BLRT를 기준으로 하위집단의 수를 추정한 경우에는 표본크기가 가장 작은(n = 200) 조건에서만 kNN 대체법의 수행이 FIML에 비교해 우수한 것으로 나타났다. BLRT를 기준으로 하위집단의 수를 결정했을 때 kNN 대체법의 수행은 결측률에 따라 상반된 모습을 보였다. 결측률이 10%인 조건에서는 kNN 대체법을 적용했을 때 정확한 하위집단 수를 추정하는 빈도는 완전자료의 경우와 비슷하거나 약간 낮은 정도였다. 그러나 결측률이 30%일 때는 n = 500일 때 하위집단 수 추정의 정확도가 급격

히 떨어졌으며 n = 1000일 때는 kNN 대체법을 BLRT와 함께 사용했을 때 정확한 하위집단 수를 추정한 빈도가 0 또는 4로 매우 저조하였다. 표 1에서와 마찬가지로, 결측값 발생 기제에 따른 하위집단 수 추정의 정확도의 일관성 있는 차이는 확인할 수 없었다.

매우 높은 효과크기 조건인 하위집단 간 거리가 1.5로 설정된 경우의 하위집단 수 추정의 결과는 표 3에 요약되어 있다. 표 3에서 하위집단 간 거리가 멀어질 경우 정확한 하위집단 수를 추정하는 빈도가 높아지는 것을 확인할 수 있다. kNN 대체법을 적용하고 BIC를 기준으로 하위집단의 수를 추정하는 경우, 정확한 하위집단의 수를 추정한 가장 낮은 빈도가 80으로 대부분의 조건에서 높은 정확도를 보였다. 그러나 표본 크기가 상대적으로 크고(n = 500, 1000) 결측률이 높을 때(30%)에는 BIC를 사용한 경우에도 kNN 대체법의 하위집단 수 추정의 정확도가 FIML에 비해 상대적으로 낮았다. BLRT를 기준으로 하위집단의 수를 결정하는 경우는 kNN 대체법을 적용하였을 때 정확한 하위집단의 수를 추정한 빈도가 FIML을 적용한 경우에 비해 저조했다. kNN 대체법과 BLRT 조합을 통한 하위집단 수 추정의 정확도는 결측률이 높을수록 큰 폭으로 낮아졌으며, 특히 n = 1000, 결측률 30%일 때는 정확한 하위집단의 수를 추정하지 못했다. 하위집단 간 거리가 1.5인 경우에도 결측값 발생 기제에 따른 일관성 있는 차이는 발견할 수 없었다.

하위집단 수의 추정에 대한 결과를 종합하자면, kNN 대체법의 하위집단 수 추정의 정확도는
조건과 하위집단 수 추정의 준거에 따라 달라졌다. BLRT를 기준으로 하위집단의 수를 추정하는 경우, 하위집단 간 거리가 가까울 때(하위집단 간 거리 0.5) FIML에 비해 높은 정확도를 보였다. 하위집단 간 거리가 0.8, 1.5일 때는 BIC를 기준으로 한 추정의 정확도가 높았다. kNN 대체법과 BIC의 조합을 통한 하위집단 수의 추정은 대부분의 조건에서 완전 자료의 정확도와 비슷한 정확도를 보였다. kNN과 BLRT의 조합은 표본의 크기가 커지고(≥500) 결측률이 높을 때(30%) 정확도가 큰 폭으로 떨어졌다.

표 1에서 3은 각 조건에서 하위집단의 수가 3개인 모형에서 계산된 엔트로피와 RMSE의 평균 또한 포함한다(Entropy 행, RMSE 행). 분류의 품질을 나타내는 엔트로피의 경우 거의 모든 조건에서 kNN 대체법을 적용한 결과가 FIML의 결과에서 얻어진 엔트로피보다 높았을 뿐만 아니라 완전 자료에서 얻어진 값에 가까웠다. 그러나 하위집단 간 거리가 0.5와 0.8인 경우에는 완전자료를 사용하였을 때에도 엔트로피의 값이 0.8 미만으로 엔트로피를 통해 판단한 분류의 불확실성이 낮다고 판단하기는 어려운 값을 보였다. 반면에 하위집단 간 거리가 1.5일 때는 모든 조건에서 1에 가까운 엔트로피를 보였다. 표 1에서 3의 각 조건 마지막 행은 하위집단 프로파일의 정확도를 확인하기 위해 계산된 RMSE이다. RMSE 기준으로 한 하위집단 프로파일의 정확도는 kNN 대체법과 FIML 간에 주목할 만한 차이는 없는 것으로 나타났으며 대부분의 조건에서 둘 중 한 방법이 눈에 띄는 우위를 보였다고 하기는 어렵다.

kNN 대체법을 BLRT와 함께 사용했을 때 정확한 하위집단 수의 추정에 실패한 조건은 주로 표본크기가 크고 결측률이 높은 조건이다. 본 연구에서 BLRT를 이용하여 하위집단을 추정하는 방

법은 2개와 3개 하위집단 모형 비교가 유의미하고 3개와 4개 하위집단 모형 비교가 유의미하지 않을 때 하위집단의 수를 3개로 추정하는 것이다. 따라서 kNN 대체법으로 생성한 완전자료에 BLRT를 기준으로 정확한 하위집단 수의 추정에 실패한 경우는 BLRT가 3개 이상의 하위집단을 지지했을 가능성이 높다. 이러한 가능성을 확인하기 위해 kNN 대체법과 BLRT 조합의 하위집단 수 추정의 정확도가 가장 낮았던 두 조건, 즉 (하위집단 간 거리 0.8, 결측률 30%, 표본크기 1000), (하위집단 간 거리 1.5, 결측률 30%, 표본크기 1000)에서 하위집단의 수를 9개까지 증가시키면서 BLRT의 결과를 확인하였으며 그 결과를 표 4에 요약하였다. 표 4의 최적 하위집단 수는 하위집단 수를 3개에서 9개까지 변화시키면서 BLRT를 실시했을 때 4개에서 8개 사이의 하위집단 수가 도출되는 경우 도출된 하위집단 수의 반복 간 평균이다. 3-9개까지 하위집단의 수를 변화시키면서 실시한 BLRT의 p 값이 모두 0.05 미만인 반복의 횟수는 “모두 유의미” 행에 표시되었다. 표 4에서 kNN 대체법으로 생성된 완전자료에 BLRT를 실시할 경우 4개보다 훨씬 많은 약 7-8개 또는 그 이상의 하위집단을 수를 추정하는 것을 확인할 수 있다.

표 3. 하위집단 간 거리가 1.5일 때 BIC, BLRT, Entropy 및 RMSE 결과

하위집단 간 거리 1.5										
	결측률 10%					결측률 30%				
	완전 자료	MAR		MCAR		완전 자료	MAR		MCAR	
		kNN	FIML	kNN	FIML		kNN	FIML	kNN	FIML
n = 200										
BIC	100	100	100	100	100	99	99	100	97	99
BLRT	82	73	89	68	86	85	25	96	16	89
Entropy	0.95	0.95	0.94	0.95	0.93	0.95	0.94	0.88	0.93	0.88
RMSE		**0.53**	0.59	0.55	0.53		**0.76**	0.79	0.83	0.75
n = 500										
BIC	100	100	100	100	100	100	92	100	98	100
BLRT	93	78	88	80	89	95	2	93	13	91
Entropy	0.95	0.95	0.93	0.95	0.93	0.95	0.94	0.88	0.94	0.87
RMSE		**0.56**	0.59	**0.59**	0.62		**0.67**	0.80	**0.70**	0.72
n = 1000										
BIC	100	100	100	100	100	98	81	98	80	100
BLRT	81	65	81	72	93	91	0	93	0	95
Entropy	0.95	0.95	0.93	0.95	0.93	0.95	0.94	0.88	0.94	0.87
RMSE		0.52	0.48	0.63	0.56		0.60	0.53	**0.56**	0.62

BIC: Bayesian information criteria; BLRT: Bootstrap likelihood ratio test; RMSE: Root mean squared
MAR: Missing at random; MCAR: Missing completely at random
kNN: k-nearest neighbor imputation;
FIML: Full information maximum likelihood method

표 4. 하위집단 간 거리 0.8 또는 1.5, 결측률 30, 표본크기 1000일 때 3-9 집단 BLRT 결과

하위 집단 간 거리	0.8		1.5	
결측률	0.3		0.3	
표본크기	1000		1000	
결측 메커니즘	MAR	MCAR	MAR	MCAR
최적 하위집단 수	7.74	7.05	7.86	7.19
모두 유의미*	65	20	45	14

MAR: Missing at random; MCAR: Missing completely at random
* 하위집단 수 3개에서 9까지 변화시키면서 시행한 BLRT의 p 값이 모두 유의미했던 반복의 수

논 의

본 연구에서 결측값을 포함하는 자료에 잠재프로파일분석(LPA)을 적용할 때 결측값 처리방법으로서 kNN 대체법의 가능성을 알아보았다. 이를 위해 결측값이 존재하는 자료에 kNN 대체법을 적용한 결과와 결측값을 포함하는 자료의 분석을 위해 현재 가장 널리 쓰이는 FIML을 적용한 결과를 비교하였다. 하위집단 간 거리, 결측률, 표본크기, 결측값 생성 메커니즘의 조건을 변화시켜 자료를 생성한 자료를 1) kNN 대체법을 적용하여 완전자료를 생성하고 분석한 결과와, 2) 같은 자료를 FIML로 분석한 결과를 정확한 하위집단 수의 추정, 분류의 품질, 추정된 하위집단 프로파일의 정확도를 기준으로 비교하였다. 또한 두 방법의 결과를 결측값을 생성하지 않은 완전자료에 대해 LPA를 적용한 결과와도 비교하였다. 결론적으로, 대부분의 조건에서 kNN 대체법을 적용한 결과는 FIML의 결과와 비슷한 수준의 수행을 보였으며, 일부 조건에서는 kNN 대체법을 적용한 결과가 FIML을 적용한 결과에 비해 나은 결과를 보여주었다.

먼저, 하위집단 수의 추정에 있어 kNN 대체법은 대부분의 조건에서 FIML을 적용한 결과와 비슷한 수준의 정확도를 보여주었다. kNN 대체법으로 생성한 완전자료에서 추정한 하위집단 수의 정확도는 하위집단 간 거리가 가까울 때(0.5)는 kNN 대체법의 결과가 FIML의 결과보다 우수하였다. 물론, 하위집단 간 거리가 가장 가까운 조건에서는 kNN 대체법과 BLRT를 조합하여 사용한 결과에서도 정확한 하위집단의 수를 추정한 경우가 100회의 반복 중 최대 34회로(MAR, n = 500, 결측률 10% 조건) 받아들일 만한 수준의 결과라고 하기는 어렵지만 표본크기가 작고 하위집단 간 거리가 가까운, 즉 하위집단 간 구분이 어려운 경우에 kNN 대체법을 적용하는 것이 FIML에 비해 나은 결과를 얻을 수 있음을 시사한다.

하위집단들이 상대적으로 멀어짐에 따라, 즉 중심점 간 거리가 0.8일 때와 1.5일 때 kNN 대체법의 하위집단 수의 추정 결과는 추정방법(BIC 또는 BLRT)과 결측률에 따라 달라졌다. 하위집단 간 거리가 0.8일 때는 kNN 대체법과 BIC의 조합이 하위집단 수의 추정에서 높은 정확도를 보여주었으며 BLRT와의 조합은 결측률이 높고(30%) 표본크기가($n \geq 500$) 클 때 정확도가 매우 낮았다. 하위집단 간 거리가 1.5일 때에도 kNN과 BIC의 조합은 대체로 매우 높은 정확도를 보였으나 결측률이 높고(30%) 표본이 커질 때에는($n \geq$

500) FIML에 비해 상대적으로 낮은 정확도를 보였다. 이러한 결과를 통해, 하위집단 간 거리가 멀어질 경우, 즉 하위집단 간 구분이 상대적으로 명확할 경우, kNN 대체법과 BIC를 함께 사용하는 것이 하위집단의 수의 추정에 있어 FIML을 대체할 만큼의 정확도를 보여줌을 알 수 있다.

LPA의 결과로 추정된 하위집단 분류의 품질을 나타내는 엔트로피의 경우 대부분의 조건에서 kNN 대체법을 적용한 자료를 분석한 결과에서 얻어진 엔트로피가 FIML의 결과에서 엔트로피에 비해 높으며 완전자료 분석의 결과와 가까웠다. 이러한 결과는 kNN 대체법을 적용한 결과가 FIML에 비해 하위집단이 명확하게 분리될 뿐만 아니라 완전 자료에서 얻은 결과와 유사함을 의미한다. 하위집단 평균벡터의 정확성을 나타내는 RMSE의 경우 kNN 대체법과 FIML을 적용한 결과 중 어느 한 쪽이 일관적으로 더 나았다고 결론내리기는 어려웠다. 결론적으로, 추정된 하위집단의 수, 하위집단 평균 프로파일의 정확도, 분류의 명확함 등을 모두 고려할 때 kNN 대체법의 수행이 FIML보다 저조하다고 할 수 없으며 일부 조건에서는 더 나은 결과를 보여주었다고 할 수 있다.

표본크기가 크고 결측률이 높을 때 kNN 대체법과 BLRT의 조합은 정확한 하위집단의 수를 추정하는데 실패하는 경우가 많았으며 이러한 경우 실제 하위집단의 수(3개) 보다 과대 추정하는 경향이 있음을 표 4에서 확인할 수 있다. 이러한 결과는 증가된 표본크기로 인해 증가한 통계적 검정력 때문일 수 있다. BLRT의 통계적 검정력에 대한 한 연구에서(Tekle, Gudicha, & Vermunt, 2016) 표본크기가 600 이상인 경우에 실제 하위집단의 수 보다 많은 수의 하위집단 모형에 대한 BLRT에서 매우 높은(≃1.0) 통계적 검정력을 보였다. 이는 표본크기가 커질 때 BLRT가 실제 하위집단의 수보다 많은 하위집단의 수를 추정하는 경향이 있음을 의미하며 본 연구에서 관찰된 표본크기가 커질 때(≧500) kNN 대체법으로 생성된 자료에 대한 BLRT가 하위집단의 수를 과대 추정하는 경향 또한 증가한 통계적 검증력이 이유일 수 있음을 시사한다. 그러나 kNN 대체법과 BLRT의 조합이 실패한 조건이라도 BIC를 통해 추정한 하위집단의 수는 매우 정확했다는 점 또한 주목할 필요가 있다.

본 연구에서 얻은 결과를 일반화하기 위해서는 다음의 두 가지 제한점을 고려할 필요가 있다. 첫째, 본 연구에서 자료의 생성에 사용한 조건들은 실제 자료의 조건을 완벽하게 반영하지 못할 수 있다. 본 연구에서 현실에서도 변화할 수 있는 조건을 생성된 자료에 반영하기 위해 하위집단 간 거리, 결측 메커니즘, 결측률, 표본 크기 등을 변화시켰다. 그러나 하위집단 내 동일한 공분산을 가지는 다변량 정규분포, 동일한 하위집단의 크기 등의 조건은 실제 자료에서 발견하기 어려운 조건일 수 있다. 실제 자료는 분포 가정을 위배하거나 비균형적인 집단 크기, 하위집단 간 서로 다른 공분산 행렬 등의 다양한 조건이 존재한다(예를 들어 염소란, 김명소, 2017; Blevins, Weathers, & Witte, 2014). 따라서 본 연구의 결과를 통해 kNN 대체법의 활용 가능성을 확인하였으나 이러한 결과의 실제 자료에 대한 일반화 가능성을 확인하기 위해서는 추후 연구에서 자료 생성을 위해 분포 가정이나 공분산 가정에서 벗어나는 등의 더욱 다양한 조건을 적용할 필요가 있다. 또한 실제

자료를 사용하여 kNN 대체법의 가능성을 확인하는 것도 유용할 것이다. 실제 자료를 활용하는 직관적인 방법은 실제 자료에 의도적으로 결측값을 발생시키고 kNN 대체법을 적용한 결과를 결측값을 발생시키기 전의 결과 또는 FIML을 적용한 결과와 비교하는 것이다. 실제 자료를 활용하는 또 다른 방법은 실제 자료를 이용하여 시뮬레이션 자료를 생성하는 것이다. 본 연구에서는 연구자가 설정한 분포와 파라메터를 통해 시뮬레이션 자료를 생성하였지만, 연구자가 실제 자료의 다양한 조건을 직접 설정하고 통제하는 것은 한계가 있으며 전술한 바와 같이 설정한 파라메터가 현실적이지 않을 가능성이 항상 존재한다. 이러한 한계를 극복하기 위해 기존 자료 시뮬레이션(extant data simulation, Jaccard & Brinberg, 2021)과 같은 방법을 활용할 수 있다. 기존자료 시뮬레이션은 연구자가 파라메터를 변인의 분포를 설정하고 그 분포에서 자료를 생성하는 것이 아니라 기존의 자료로부터 반복 표집을 통해 자료를 생성함으로써 파라메터 설정의 인위성을 배제하고 시뮬레이션 연구의 외적 타당도를 향상시키는 방법이다.

두 번째 제한점으로, 하위집단 수의 추정과 하위집단 평균 프로파일의 정확성에서 kNN 대체법의 결과가 FIML보다 우수한 조건이 존재하는데 이러한 경우에 대해 일방적으로 kNN 대체법이 FIML에 비해 우수한 수행을 보인다고 결론 내리는 것에는 주의할 필요가 있다. FIML을 사용하여 LPA의 모수를 추정할 때 일반적으로 시작값을 변화시키면서 모형의 수렴 여부를 확인하는 절차를 거친다(Muthén & Muthén, 2017). 그러나 이러한 절차는 각 분석마다 연구자가 직접 조정하여야 하기 때문에 대량의 반복을 실시하는 시뮬레이션에서는 적용이 어렵다. 따라서, 앞서 기술하였듯이 FIML의 가정을 만족시키는 방식으로 자료가 생성되었더라도 결측값으로 인해 각 조건의 반복 중 일부에서 모수 추정의 수렴이 일어나지 않거나 지역최대(local maxima)가 최종결과에 반영되어 하위집단 수와 평균 프로파일의 추정에 문제를 야기하였을 가능성을 배제할 수 없다.

이러한 한계점에도 불구하고 본 연구는 결측값이 존재하는 자료에 대한 LPA를 위한 새로운 대안으로서 kNN 대체법의 활용 가능성을 확인한 최초의 연구로서 의의를 가진다. 본 연구의 결과를 통해 결측값이 존재하는 자료에 대한 잠재프로파일분석을 위해 가장 흔히 권장되는 FIML에 더하여 kNN 대체법을 병행할 것을 제안한다. 본 연구에서 kNN 대체법의 수행이 대체로 FIML과 비슷하며 특히 하위집단 간 거리가 가깝고 표본크기가 작은 경우 FIML에 비해 나은 수행을 보여줌을 확인하였다. 또한, 심리학을 비롯한 사회과학의 자료들은 정규분포의 가정을 위배하는 경우가 많으며(Blanca 등, 2013; Micceri, 1989), 특히 우울이나 불안과 같은 정신건강 척도에서 얻은 자료는 병리적 증상의 상대적인 희소성으로 인해 정규분포를 벗어난 편포를 이루는 경우가 많다(Atkins & Gallop, 2007). kNN 대체법이 FIML과 다르게 지표변인의 분포에 대해 가정하지 않는 특성과 본 연구에서 확인한 하위집단 수와 평균 프로파일의 정확성, 분류의 품질을 고려할 때 건강심리학 분야에서 결측값을 포함하는 자료의 잠재프로파일분석을 위해 기존의 FIML 방법으로 분석이 어려운 경우에 kNN 대체법을 활용할 수 있을 것이다. 또한 kNN 대체법이나 FIML 중 한

가지 방법만 사용하기보다는 같은 자료에 대해 두 가지 방법을 모두 적용하고 그 결과를 비교하는 것을 제안한다. 이러한 전략은 FIML에서 지역최대 등으로 인한 모수 추정의 불안정성을 확인하는 수단이 될 수 있으며 FIML을 통한 모수 추정에 어려움이 있을 때 kNN 대체법의 결과가 모형의 수정이나 시작값의 조정과 같은 대처 전략을 세우는데에도 도움을 줄 수 있을 것이다.

참 고 문 헌

김시형, 신지영, 이동훈 (2019). 사별 이후 지속비애 증상에 대한 잠재프로파일 분석. 한국심리학회지: 건강, 24(2), 371-391.

노민정, 유진은 (2019). 사회과학 대용량 자료 분석을 위한 벌점회귀모형과 결측처리기법의 성능 비교: 몬테카를로 모의실험. 교육평가연구, 32(4), 755-776.

박경우, 장혜인 (2021). 한국어판 강박적 성행동 장애 척도(K-CSBD-19)의 타당화 연구. 한국심리학회지: 건강, 26(5), 859-879.

신택수 (2014). 결측자료 분석방법에 대한 고찰과 활용. 교육평가연구, 27(3), 693-725.

송주원 (2017). 결측자료 분석에서 결측 비율이 결측자료 k-평균 군집분석에 미치는 영향. *Journal of The Korean Data Analysis Society, 19*(3), 1273-1282.

염소란, 김명소 (2017). 상사에 대한 정서노동의 잠재프로파일분석: 정서노동 유형에 따른 선행변인 및 직무효과성 차이. 한국심리학회지: 산업 및 조직, 30(3), 465-489.

조다빈, 심은정 (2021). 분노 경험과 COVID-19 예방수칙 준수행동 및 정신건강 문제의 관계. 한국심리학회지: 건강, 26(1), 55-71.

최현주 (2021). 형제자매의 장애 유무에 따른 청소년의 부모화 유형과 거부민감성 및 우울과의 관계. 한국심리학회지: 상담 및 심리치료, 33(4), 1715-1737.

최현주, 장은비 (2021). 학교조직풍토와 교사 소진의 관계: 특수교사가 지각한 학교조직풍토의 잠재프로파일을 중심으로. 한국심리학회지: 학교, 18(3), 291-316.

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison. *Computers in Biology and Medicine, 136*, 104672.

Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: a tutorial on count regression and zero-inflated models. *Journal of Family Psychology, 21*(4), 726-735.

Batista, G. E., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence, 17*(5-6), 519-533.

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making, 16*(3), 197-208.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9*(2), 78-84.

Blevins, C. A., Weathers, F. W., & Witte, T. K. (2014). Dissociation and posttraumatic stress disorder: A latent profile analysis. *Journal of Traumatic Stress, 27*(4), 388-396.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*, 195-212.

Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official*

Statistics, *16*(2), 113-131.

Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*(4), 352-370.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*(3), 430-457.

Goyal, S., Chauhan, R. K., & Parveen, S. (2016). Spam detection using KNN and decision tree mechanism in social network. *In 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 522-526). IEEE.

García-Laencina, P. J., Sancho-Gómez, J. L., Figueiras-Vidal, A. R., & Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, *72*(7-9), 1483-1493.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge University Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Jaccard, J., & Brinberg, M. (2021). Monte Carlo simulations using extant data to mimic populations: Applications to the modified linear probability model and logistic regression. *Psychological Methods*. *26*(4), 450-465.

Jonsson, P., & Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using likert data. *In 10th International Symposium on Software Metrics, 2004. Proceedings*. (pp. 108-118). IEEE.

Keerin, P., Kurutach, W., & Boongoen, T. (2012). Cluster-based KNN missing value imputation for DNA microarray data. *In 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 445-450). IEEE.

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, *74*(7), 1-16.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mill.

Lee, T., & Shi, D. (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychological Methods, 26*(4), 466-485.

Little, T. D. (Ed.). (2014). *The Oxford handbook of quantitative methods* (Vol. 1). Oxford University Press, USA.

McLachlan, G., & Peel. D. (2004). *Finite mixture models*. New Jersey : Wiley.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156 - 166.

Muthén, L. K., & Muthén, B. (2017). *Mplus user's guide: Statistical analysis with latent variables, user's guide*. Muthén & Muthén.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372-411.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535-569.

Park, J., Kim, K. Y., & Kwon, O. (2014). Comparison

of machine learning algorithms to predict psychological wellness indices for ubiquitous healthcare system design. *In Proceedings of the 2014 International Conference on Innovative Design and Manufacturing (ICIDM)* (pp. 263-269). IEEE.

Pujianto, U., Wibawa, A. P., & Akbar, M. I. (2019). K-Nearest Neighbor (K-NN) based Missing Data Imputation. *In 2019 5th International Conference on Science in Information Technology (ICSITech)*. IEEE.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147-177.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Steinley, D., & Brusco, M. J. (2011). Evaluating mixture modeling for clustering: recommendations and cautions. *Psychological Methods*, *16*(1), 63-79.

Tein, J. Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: a Multidisciplinary Journal*, *20*(4), 640-657.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520-525.

Yung, Y. F., & Zhang, W. (2011). Making use of incomplete observations inthe analysis of structural equation models: The CALIS procedure's fullinformation maximum likelihood method in SAS/STAT®9.3. *SAS Global Forum*, (pp.333-2011), http://support.sas.com/resources/papers/proceedings11/333-2011.pdf).

원고접수일: 2022년 4월 14일
논문심사일: 2022년 4월 16일
게재결정일: 2022년 4월 19일

한국심리학회지: 건강
The Korean Journal of Health Psychology
2022. Vol. 27, No. 3, 513 - 532

Using k-Nearest Neighbor Imputation as a Method to Handle Missing Values in Latent Profile Analysis

Sumin Kim
Department of Psychology,
Pusan National University
Graduate Program

Seung Bin Cho
Department of Psychology,
Pusan National University
Assistant Professor

Latent profile analysis (LPA) is a method commonly used in psychology to identify subgroups of individuals who share common characteristics. To apply LPA on data with missing values, full information maximum likelihood (FIML) and multiple imputation (MI) are commonly recommended. In this study, we propose k-nearest neighbor (kNN) imputation, as an efficient alternative to handle missing data in LPA and examined its potential using simulated datasets. Datasets were generated with varying conditions: missing value generation mechanisms, missing rates, distances between subgroups, and sample sizes. Complete data were generated by kNN imputation from the simulated datasets and were used in LPA. Results were compared to the results from FIML in terms of the number of estimated subgroups, the accuracy of mean profiles, and the quality of classification. The accuracy of the number of subgroups from kNN imputation was comparable to the results from FIML in most conditions, and kNN imputation performed better in some conditions Neither method consistently performed better in terms of the accuracy of mean profiles. The quality of classification from kNN imputation was better in all conditions, and was closer to the results from complete data analyses. From the results, we suggest kNN imputation as an alternative to FIML to handle missing data in LPA, especially in conditions wherein FIML often fails. We also suggest using kNN imputation as well as FIML to compare results to check the stability of parameter estimates.

Keywords: latent profile analysis, k-nearest neighbor imputation, missing values, simulation