The Korean Journal of Applied Statistics
2022, Vol. 35, No. 5, 657–666

DOI: https://doi.org/10.5351/KJAS.2022.35.5.657
Print ISSN1225-066X / Online ISSN 2383-5818

Note on the estimation of informative predictor subspace and projective-resampling informative predictor subspace

Jae Keun Yoo[1,a]

[a]Department of Statistics, Ewha Womans University

Abstract

An informative predictor subspace is useful to estimate the central subspace, when conditions required in usual suffcient dimension reduction methods fail. Recently, for multivariate regression, Ko and Yoo (2022) newly defined a projective-resampling informative predictor subspace, instead of the informative predictor subspace, by the adopting projective-resampling method (Li *et al.* 2008). The new space is contained in the informative predictor subspace but contains the central subspace. In this paper, a method directly to estimate the informative predictor subspace is proposed, and it is compapred with the method by Ko and Yoo (2022) through theoretical aspects and numerical studies. The numerical studies confirm that the Ko-Yoo method is better in the estimation of the central subspace than the proposed method and is more efficient in sense that the former has less variation in the estimation.

Keywords: clustering mean method, informative predictor subspace, multivariate regression, projective-resampling informative predictor subspace, projective-resampling mean method, suffcient dimension reduction

1. 서론

설명 변수 $\mathbf{X} \in \mathbb{R}^p$가 주어졌을 때 다변량 반응 변수 $\mathbf{Y} \in \mathbb{R}^r$의 조건부 분포인 $\mathbf{Y}|\mathbf{X}$를 연구하는 것을 다변량 회귀분석이라고 한다. 다변량 반응변수는 경시적 자료분석이나 반복측정자료 혹은 함수적 자료에서 빈번하게 나타나고 있다. 이러한 다변량 회귀분석에서 충분차원축소(sufficient dimension reduciton)은 조건부 분포 $\mathbf{Y}|\mathbf{X}$에 대한 정보의 손실 없이 p차원의 설명 변수 $\mathbf{X}$를 저차원의 $\mathbf{X}$의 선형결합 형태인 $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$로 대체하는 것이다. 여기서 $\boldsymbol{\eta}$는 $p \times d$ 행렬이다. 이를 조건부 독립식으로 표현하면 다음과 같다:

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}.$$

여기서 $\perp\!\!\!\perp$는 통계적 독립을 의미한다.

다변량 회귀에서 이러한 충분차원축소는 $\boldsymbol{\eta}$에 대한 추정을 그 목적으로 하고, 위의 식을 만족하는 많은 $\boldsymbol{\eta}$들 중에서 최소 열의 수을 가지는 $\boldsymbol{\eta}$의 열에 의해 생성되는 공간을 중심 부분 공간(central subspace)이라고 부른다. 최근 Setodji와 Cook (2004), Yin와 Bura (2006), Yoo와 Cook (2007), Yoo (2008), Li 등 (2008) 그리고 Lee 등 (2019) 등에서 중심 부분 공간을 추정하는 다양한 방법론을 제시하고 있다. 하지만 이러한 방법론들은

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education (NRF2019R1F1A1050715).

[1] Department of Statistics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: peter.yoo@ewha.ac.kr

Published 31 October 2022 / journal homepage: http://kjas.or.kr

소위 선형성(linearity) 조건, 등분산성(constant variance) 조건 그리고 범위(coverage) 조건 등이 만족되어야 한다. 하지만 Yoo (2016)에 따르면 이러한 조건들이 만족하는지 실제로 진단하는 것은 불가능하고, 이러한 조건이 만족되지 않을 경우 차원축소의 결과는 잘못된 분석을 야기할 수 있음을 지적하고 있다. 해당 조건에 대한 자세한 내용은 Yoo (2016)과 Yoo (2018)을 참고하길 바란다.

반응 변수의 차원이 1인 일변량 회귀에서 이러한 조건적 한계를 극복하기 위해 Yoo (2016)은 정보적 설명 변수 공간(informative predictor subspace)을 통한 중심 부분 공간의 추정을 제시하고 있다. 추정 방법론은 우선 설명 변수 $\mathbf{X}$을 K평균 군집화 방법을 이용하여 자료를 범주화 한 후, 이 군집에 해당되는 자료에 대해 반응 변수 Y을 다시 범주화 한후 최종 범주에서 $\mathbf{X}$의 평균을 추정하여, 이를 이용하여 중심 부분 공간을 추정하는 방법이다. 정보적 설명 변수 공간과 추정 방법에 대한 자세한 설명은 이후 절에서 기술할 것이다.

최근 Ko와 Yoo (2022)는 다변량 회귀분석에서 정보적 설명 변수 공간을 추정하기 위해 Li 등 (2008)이 제시한 투영-재표본 방법론을 적용한다. 적용 결과 Ko와 Yoo (2022)에서는 정보적 설명 변수 공간이 아닌 정보적 설명 변수 공간에 속하고 중심 부분 공간을 포함하는 투영-재표본 정보적 설명 변수 공간을 새롭게 정의하고, 이를 추정하는 방법을 제시하고 있다.

하지만 Ko와 Yoo (2022)는 정보적 설명 변수 공간을 직접적으로 추정하는 방법을 제안하지 않았고, 이에 따라 정보적 설명 변수 공간을 직접적으로 추정하는 방법과 Ko와 Yoo (2022)에서 제시된 투영-재표본 정보적 설명 변수 공간을 통하여 중심 부분 공간을 추정하는 방법과의 비교도 제시되지 않았다. 실제로 Ko와 Yoo (2022)에서 정의되는 투영-재표본 정보적 설명 변수 공간의 유용성을 보다 객관적으로 입증하기 위해서는 정보적 설명 변수 공간의 추정 방법과의 비교는 대단히 중요하다고 할 수 있다. 본 논문의 목적은 다변량 회귀분석에서 정보적 설명 변수 공간을 추정하는 방법을 제시하고, 이 추정 방법과 Ko와 Yoo (2022)에 제시하는 투영-재표본 정보적 설명 변수 공간의 추정 방법을 이론적으로 유사점과 차이점을 설명하고자 한다. 또한 모의실험을 통하여 각 방법의 장단점을 파악하고자 한다.

본 논문의 차례는 다음과 같다. 2장에서는 정보적 설명 변수 공간과 투영-재표본 정보적 설명 변수 공간에 대해 설명을 할 것이다. 3장은 다변량 회귀분석에서 정보적 설명 변수 공간을 추정하는 이중 군집 평균 방법을 제시하고, 투영-재표본 정보적 설명 변수 공간의 추정 방법과 비교를 할 것이다. 이후 4장은 모의실험을 통하여 두 방법론을 비교하고, 5장에는 결론이 제시된다.

2. 정보적 설명 변수 공간과 투영-재표본 정보적 설명 변수 공간

2.1. 정보적 설명 변수 공간

다변량 회귀 $\mathbf{Y} \in \mathbb{R}^r | \mathbf{X} \in \mathbb{R}^p$에서 정보적 설명 변수 공간은 다음과 같이 정의 한다.

$$\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}^{-1}\mathcal{S}\left\{E\left(\mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\right) - E\left(\mathbf{X}\right)\right\}.$$

여기서 $\mathcal{S}\{E(\mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}) - E(\mathbf{X})\}$는 $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$의 값이 변화하면서 $E(\mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X})$에 의해 생성되는 공간을 의미한다. $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$는 중심 부분 공간의 직교 정규 기저이다. 물론 $\boldsymbol{\eta}$ 대신에 차원축소부분 공간의 생성하는 다른 기저를 사용해도 되지만, 최소 정보적 설명 변수 공간을 정의하기 위해서 중심 부분 공간의 기저를 사용하고자 한다.

위의 정의는 반응 변수 $\mathbf{Y}$에 대한 부분이 없는 설명 변수만으로 생성되는 공간이기 때문에 설명 변수가 주어졌을 때 반응 변수의 조건부 분포를 알고자 하는 회귀분석의 목적과 일치하지 않는다. 그래서 Yoo (2016)은 $E(\mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}) - E(\mathbf{X})$을 다음과 같이 대체하였다.

$$\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}^{-1}\mathcal{S}\left\{E\left(E\left(\mathbf{X}|\mathbf{Y},\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\right)|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\right) - E\left(\mathbf{X}\right)\right\}.$$

정보적 설명 변수 공간은 $\boldsymbol{\eta}$를 이미 알고 있는 상황에서 생성되는 공간이다. 하지만 $\boldsymbol{\eta}$를 안다면 충분차원축소의 목적이 달성되었기 때문에 정보적 설명 변수 공간의 추정은 의미가 없게된다. 그렇기에 $\boldsymbol{\eta}$를 모르는 상황에서

정보적 설명 변수 공간을 알아낼 수 있는 우회적 방법이 필요하고, Yoo (2016)은 다음을 제시하고 있다.

$$\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}^{-1}\mathcal{S}\left\{E\left(E\left(\mathbf{X}|\mathbf{Y}, C_{\mathbf{X}}\right)|C_{\mathbf{X}}\right) - E\left(\mathbf{X}\right)\right\}.$$

여기서 $C_{\mathbf{X}}$는 $\mathbf{X}$의 군집(cluster)를 의미하고, 이를 위해 Yoo (2016)은 K-평균 군집방법을 사용하고, 이에 대한 이유에 대해서는 다음 절에서 설명하기로 한다.

그럼 $E(\mathbf{X}|\mathbf{Y}, C_{\mathbf{X}})$는 $C_{\mathbf{X}}$에 따른 군집내에서의 평균이고, $C_{\mathbf{X}} = c$일 때, $E(\mathbf{X}_c|\mathbf{Y}_c)$로 나타낼 수 있고, 위의 식을 다음과 같이 나타낼 수 있다.

$$\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}^{-1}\mathcal{S}\left\{E\left(E\left(\mathbf{X}_c|\mathbf{Y}_c\right)|C_{\mathbf{X}} = c\right) - E\left(\mathbf{X}\right)\right\}, \ \ c = 1, \ldots, k. \tag{2.1}$$

식 (2.1)에서 $E(\mathbf{X}_c|\mathbf{Y}_c)$의 추정은 Setodji와 Cook (2004)의 K-평균군집 역회귀 방법론(K-means inverse regression)으로 구할 수 있게 되고, 이에 대해 다음 절에서 자세하게 설명할 것이다.

그리고, Yoo (2016)에 따르면, 평균이 0이고 분산이 $\mathbf{I}_p$로 정규화 한 설명 변수 $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - E(\mathbf{X}))$와 $\mathbf{X}$에 대해 다음이 성립한다:

$$\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{Z}}.$$

실제로 $\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}}$을 추정할 때 연산 과정에서 발생할 수 있는 계산 오차를 줄이기 위해 $\mathbf{X}$대신에 보다는 정규화된 $\mathbf{Z}$을 사용한다. 정규환된 설명 변수 $\mathbf{Z}$에 대한 $\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{Z}}$에 대해 식 (2.1)은 다음과 같이 변형된다:

$$\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{Z}} = \mathcal{S}\left\{E\left(E\left(\mathbf{Z}_c|\mathbf{Y}_c\right)|C_{\mathbf{X}} = c\right)\right\}, \ \ c = 1, \ldots, k.$$

2.2. 투영-재표본 정보적 설명 변수 공간

다변량 반응 변수 $\mathbf{Y}$에 대해 1차원 선형변환 $t^{\text{T}}\mathbf{Y}$을 고려한다. 여기서 t는 길이가 1인 확률 벡터이다. 그리고, 단변량 회귀인 $t^{\text{T}}\mathbf{Y}|\mathbf{X}$에서 중심 부분 공간을 생성하는 $p \times p$ 정칙 행렬을 $\mathbf{M}(t)$라고 정의하자:

$$\mathcal{S}_{t^{\text{T}}\mathbf{Y}|\mathbf{X}} = \mathcal{S}\left\{\mathbf{M}\left(t\right)\right\}.$$

그리고, $t^{\text{T}}\mathbf{Y}|\mathbf{X}$ 회귀에서 $\mathbf{M}(t)$에 의해 정의되는 정보적 설명 변수 공간을 생성하는 $p \times p$ 정칙 행렬을 $\boldsymbol{\phi}(t)$라고 정의하자:

$$\mathcal{S}^{\text{IPS}}_{t^{\text{T}}\mathbf{Y}|\mathbf{X}} = \mathcal{S}\left\{\boldsymbol{\phi}\left(t\right)\right\}.$$

Ko와 Yoo (2022)에 따르면 다음의 관계가 성립한다:

$$\mathcal{S}\left\{\mathbf{M}\left(t\right)\right\} \subseteq \mathcal{S}\left\{E\left(\mathbf{M}\left(\mathbf{T}\right)\right)\right\} = \mathcal{S}_{\mathbf{Y}|\mathbf{X}}.$$

여기서 $E(\mathbf{M}(\mathbf{T}))$는 $\mathbf{M}(t)$에서 확률 벡터 t에 대한 기대값을 강조하여, $\mathbf{M}(t)$와 달리 t에 의존하지 않는다는 것을 강조하기 위한 표현이다. 위의 관계에서 중요한 점은 $\mathbf{M}(t)$는 $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$의 부분공간을 생성하지만 $E(\mathbf{M}(t))$는 $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$를 정확하게 생성한다는 것이다. 그리고 Ko와 Yoo (2022)는 $E(\boldsymbol{\phi}(\mathbf{T}))$에 대해서 다음의 관계가 성립한다:

$$\mathcal{S}_{\mathbf{Y}|\mathbf{X}} \subseteq \mathcal{S}\left\{E\left(\boldsymbol{\phi}\left(\mathbf{T}\right)\right)\right\} \subseteq \mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}}.$$

위의 식에 따르면 $E(\boldsymbol{\phi}(\mathbf{T}))$는 $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$를 완전추정할 수 있고, $\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}}$보다 작은 $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$의 상위한계 공간을 생성한다. 여기서 $\mathcal{S}\{E(\boldsymbol{\phi}(\mathbf{T}))\}$을 투영-재표본 정보적 설명 변수 공간이라고 부른다. 그리고 $E(\boldsymbol{\phi}(\mathbf{T}))$의 형태는 다음과 같다:

$$E_{\mathbf{T}}\left[\text{cov}\left(E\left(E\left(\mathbf{X}_c|\mathbf{T}^{\text{T}}\mathbf{Y}_c\right)|C_{\mathbf{X}} = c\right) - E\left(\mathbf{X}\right)\right)\right].$$

$E(\mathbf{X}_c|\mathbf{T}^{\text{T}}\mathbf{Y}_c)$에서 $\mathbf{T}^{\text{T}}\mathbf{Y}$는 일변량이기 때문에 Li (1991)에 제안한 sliced inverse regression을 이용하여 추정을 한다. 또한 $\mathbf{T}$는 다변량 표준 정규 분포를 따른 확률 변수로 길이가 1이 되도록 $\mathbf{T}/\|\mathbf{T}\|$으로 변환한다. 투영-재표본 정보적 설명 변수 공간의 추정에 대해서는 다음 절에서 상세히 기술할 것이다.

3. 추정 방법론적 비교

3.1. 정보적 설명 변수 공간의 추정: 이중 군집 평균 방법

다변량 회귀에서 정보적 설명 변수의 공간의 추정은 반응 변수의 차원이 1일때 Yoo (2016)에서 제안한 군집 평균 방법을 그대로 적용할 수 있다. Yoo (2016)에서 슬라이싱(slicing)이라고 불리는 일차원 반응 변수의 범주화가 요구된다. 하지만 다변량 회귀에서 이러한 슬라이싱을 적용하기 위해서는 최종 범주의 수가 반응 변수의 차원에 지수적으로 증가하는 단점이 있다. 이러한 다차원 반응 변수의 범주화에서 발생하는 차원의 저주를 회피하기 위해 Setdoji와 Cook (2004)는 K-평균 군집법을 적용하여 반응 변수에 대한 군집을 만들고, 이 군집을 다차원 반응 변수의 범주로 사용하였다. 차원의 저주를 피하기 위해 다차원 반응 변수에 대한 주성분분석 혹은 설명 변수와 정준상관분석을 이용한 저차원의 반응 변수에 대해 슬라이싱을 사용할 수 있지만, Setodji와 Cook (2004)에 따르면 이 과정에서 발생하는 반응 변수의 손실로 인해 K-군집 방법론에 비해 차원 축소의 결과가 부정확해짐을 지적하고 있다. 본 논문에서는 이에 따라 다변량 회귀에서 정보적 설명 변수 공간 $\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}}$를 추정하기 위해 다변량 반응 변수를 K-군집 방법론으로 범주화를 시도하고, $C_{\mathbf{Y}}$가 군집화에 따른 범주를 의미한다고 하자. 그리고 반응 변수의 군집화는$\mathbf{Y}_c$, $c = 1, \ldots, k$에 대해 각각 이루어진다. 또한 $\mathbf{X}$를 직접으로 사용하여 $\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}}$의 기저를 추정하기 보다는 $\hat{\mathbf{Z}} = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X} - \bar{\mathbf{X}})$을 이용하여 $\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{Z}}$에 대한 기저를 추정한 후, $\hat{\boldsymbol{\Sigma}}^{-1/2}$을 곱하여 $\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}}$에 대한 기저를 추정하고자 한다.

정보적 설명 변수 공간의 추정에 핵심은 $E(\mathbf{Z}_c|\mathbf{Y}_c)$에 대한 추정에 있고, 이는 Setodji와 Cook (2004)가 제시한 K-평균군집 역회귀 방법의 적용과 동일하다:

$$\hat{E}\left(\mathbf{Z}_c|\mathbf{Y}_c\right) = \bar{\mathbf{Z}}_{c\bullet} = \left\{\bar{Z}_{c1}, \ldots, \bar{Z}_{ch}\right\}.$$

여기서 $\bar{Z}_{cj}$는 $\mathbf{X}$의 c번째 군집에서 상응하는 반응변수의 j번째 군집에 해당되는 $\hat{\mathbf{Z}}$의 관측값들의 표본평균을 의미한다.

추정의 일차적 목표는 $E(E(\mathbf{Z}_c|\mathbf{Y}_c)|C_{\mathbf{X}})$에 있지만, 최종 목표는 $E(E(\mathbf{Z}_c|\mathbf{Y}_c)|C_{\mathbf{X}})$에 의해 생성되는 공간인 $\mathcal{S}\{E(E(\mathbf{Z}_c|\mathbf{Y}_c)|C_{\mathbf{X}})\}$에 있고, 이 공간은 $E(E(\mathbf{Z}_c|\mathbf{Y}_c)|C_{\mathbf{X}})$의 공분산 행렬에 의해 생성되는 공간과 동일하다. 이를 위해 다음의 커널 행렬 $\hat{\boldsymbol{\Omega}}$을 정의한다:

$$\hat{\boldsymbol{\Omega}} = \sum_{c=1}^{k} \frac{1}{n_c} \bar{\mathbf{Z}}_{c\bullet} \bar{\mathbf{Z}}_{c\bullet}^{\mathrm{T}}.$$

그러면 $\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{Z}}$의 기저의 추정량은 $\hat{\boldsymbol{\Omega}}$의 가장 큰 처음의 d개의 고유값에 대응하는 고유 벡터들 $(\hat{\omega}_1, \ldots, \hat{\omega}_d)$이고, $\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}}$의 기저의 추정량은 $\hat{\boldsymbol{\Sigma}}^{-1/2}(\hat{\omega}_1, \ldots, \hat{\omega}_d)$이다.

이렇게 설명 변수와 반응 변수 모두를 K-군집화를 하여 이에 해당되는 설명 변수의 표본 평균을 이용하여 $\mathcal{S}^{\text{IPS}}_{\mathbf{Y}|\mathbf{X}}$를 추정하는 방법을 제안하고, 이 방법을 이중 군집 평균 방법(double-clustering mean method; DCMM)이라고 하고자 한다.

3.2. 투영-재표본 설명 변수 공간의 추정: 투영-재표본 평균 방법

투영-재표본 설명 변수 공간의 추정은 $E_{\mathbf{T}}[\text{cov}(E(E(\mathbf{X}_c|\mathbf{T}^{\mathrm{T}}\mathbf{Y}_c)|C_{\mathbf{X}} = c) - E(\mathbf{X}))]$의 추정이 핵심이다. 여기서도 $\mathbf{X}$ 대신에 $\mathbf{Z}$을 이용하고, 이를 나타다면 다음과 같다:

$$E_{\mathbf{T}}\left[\text{cov}\left[E\left(E\left(\mathbf{Z}_c|\mathbf{T}^{\mathrm{T}}\mathbf{Y}_c\right)|C_{\mathbf{X}} = c\right)\right]\right].$$

위의 식에서 $\text{cov}[E(E(\mathbf{Z}_c|\mathbf{T}^{\mathrm{T}}\mathbf{Y}_c)|C_{\mathbf{X}} = c)$의 형태는 이중 군집 평균 방법과 동일한 반면, 차이는 $\mathbf{T}^{\mathrm{T}}\mathbf{Y}_c$는 일변량이기 때문에 군집 방법론을 적용하지 않고, Li (1991)에서 제시되는 슬라이싱 방법으로 범주화 된다. 그렇기에

다변량 $\mathbf{Y}$의 군집화를 통한 범주화보다는 반응 변수의 차원과 자료수에 덜 민감하게 할 수 있어, 그 과정이 보다 더 편리하고 용이하다:

$$\widehat{\mathrm{cov}}\left[E\left(E\left(\mathbf{Z}_c|\mathbf{T}^{\mathrm{T}}\mathbf{Y}_c\right)|C_{\mathbf{X}}=c\right)\right]=\hat{\mathbf{\Omega}}(\mathbf{T})=\sum_{c=1}^{k}\frac{1}{n_c}\bar{\mathbf{Z}}_{c\bullet}(\mathbf{T})\,\bar{\mathbf{Z}}_{c\bullet}^{\mathrm{T}}(\mathbf{T}).$$

여기서 $\bar{\mathbf{Z}}_{c\bullet}(\mathbf{T})=\{\bar{Z}_{c1}(\mathbf{T}),\ldots,\bar{Z}_{ch}(\mathbf{T})\}$이고, $\bar{Z}_{cj}(\mathbf{T})$는 $\mathbf{X}$의 c번째 군집에서 상응하는 $\mathbf{T}^{\mathrm{T}}\mathbf{Y}$의 j번째 슬라이스에 해당되는 $\hat{\mathbf{Z}}$의 관측값들의 표본평균을 의미한다.
최종적으로 $E_{\mathbf{T}}[\mathrm{cov}[E(E(\mathbf{Z}_c|\mathbf{T}^{\mathrm{T}}\mathbf{Y}_c)|C_{\mathbf{X}}=c)]]$의 추정량은 다음과 같다:

$$\hat{\mathbf{\Psi}}=\frac{1}{m_n}\sum_{i=1}^{m_n}\hat{\mathbf{\Omega}}(\mathbf{T}_i).$$

이중 군집 평균 방법에서 처럼 일단 $\hat{\mathbf{\Psi}}$을 구하면, $\hat{\mathbf{\Psi}}$의 가장 큰 처음의 d개의 고유값에 대응하는 고유 벡터들 $(\hat{\psi}_1,\ldots,\hat{\psi}_d)$을 계산한다. 그러면, $\mathcal{S}\{\boldsymbol{\phi}(\mathbf{T})\}$의 기저 추정량은 $\hat{\mathbf{\Sigma}}^{-1/2}(\hat{\psi}_1,\ldots,\hat{\psi}_d)$이다. Ko와 Yoo (2022)에서는 이 추정 방법을 $\mathbf{T}$에 대한 반복적인 표본 추출과 $\mathbf{Y}$로의 선형 변환의 중요성을 대표하여 재표본-투영 평균 방법이라고 명명하였다.

3.3. 추정 방법론의 비교

이중 군집 평균 방법은 $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}^{\mathrm{IPS}}$를 추정하기에 $\mathcal{S}\{\boldsymbol{\phi}(\mathbf{T})\}$를 추정하는 재표본-투영 평균 방법보다 더 큰 공간을 추정해야 한다. 하지만 $\mathbf{Y}$에 대한 군집화에 대한 평균이 $\mathbf{T}^{\mathrm{T}}\mathbf{Y}$의 슬라이싱의 $\mathbf{T}$에 대한 평균을 포함한다는 것이 추정 방법론적으로 보장되지는 않다. 실제로 표본을 이용한 추정은 이용한 추정의 경우 군집의 수, 슬라이스의 수, 자료의 수 등에 영향을 받고, 이들의 값에 따라 결과가 차이날 수 있기 때문이다.

이중 군집 추출 방법의 경우 $\mathbf{Y}_c$에서의 군집화가 요구되기 때문에 $\mathbf{Y}_c$에 있는 표본의 수가 적다면 군집화가 제대로 이루어지지 않을 수 있고, 경우에 따라서 한 군집으로의 쏠림에 의한 불균형이 발생할 여지가 충분이 있다. 이러한 불균형이 발생하면 추정의 정확도를 떨어뜨릴 수 있다. 또한 군집화의 신뢰도는 반응 변수의 차원에도 의존하기 때문에, 고차원의 반응 변수에 표본의 수가 적은 경우 군집화의 적정성에도 문제가 제기될 수 있다. 따라서, 실제 자료의 수가 많지 않다면 이중 군집 추출 방법의 한계는 분명하다. 반면에 $\mathbf{T}^{\mathrm{T}}\mathbf{Y}$은 일변량에 대한 범주화이기 때문에 이러한 상황에서 슬라이싱을 이용한 범주화가 더 용이한 측면이 있다. 무엇보다 $\mathbf{T}^{\mathrm{T}}\mathbf{Y}$의 슬라이싱을 통한 범주화는 $\mathbf{Y}$의 차원에 의존하지 않는다. 반응 변수의 차원이 상대적으로 높다면 $\mathbf{T}$의 재표본 수를 증가시켜 정확도를 유지할 수 있는 여지가 있다.

재표본-투영 평균 방법의 단점은 $\mathbf{T}$의 재표본 횟수는 기본적으로 자료의 수에 의존하고, 이에 따라 계산량이 증가하는 단점이 있다. 하지만 $\mathbf{T}$에 대한 인위적 선택에 따라 이를 극복할 수 있다. $\mathbf{T}$에 대해 i번째 요소만 1이고 나머지는 모두 0인 정준 벡터 e_i, $i=1,\ldots,r$를 고려해보자. 그렇다면 $\mathbf{T}^{\mathrm{T}}\mathbf{Y}$은 $Y_1,\ldots,Y_r$이 될 것이고, 이는 반응 변수의 요소별 일변량 회귀의 합으로 표현될 것이다. 실제로 Yoo와 Cook (2007), Yoo 등 (2010) 등에 따르면 반응 변수 요소별 회귀는 중앙부분공간을 추정하는 데 충분한 정보를 제공할 수 있음을 제시하고 있다. 이는 재표본-투영 평균 방법에서 적용하면 재표본의 수를 줄일 수 있어 계산량을 감소시킬 수 있다. 이러한 요소별 회귀의 결합은 이중 군집 추출 방법에서는 적용이 불가하다.

이중 군집 평균 방법과 재표본-투영 평균 방법 모두 $\mathbf{X}$에 군집이 고정되었다고 할때 동일 수의 군집과 슬라이스를 고려한다 할지라도 다른 결과가 계산될 수 있다. 이중 군집 평균 방법의 경우 K-평균 군집 결과가 매 시행마다 동일하게 나오지 않는 것에 기인하고, 재표본-투영 평균방법은 $\mathbf{T}$을 임의로 발생하는데 있다. 이중 군집 평균 방법의 경우 군집 방법 자체에 대한 문제이기에 이를 조정할 수 있는 대안이 없는 반면, 재표본-투영 평균 방법은 반응 변수 요소별 회귀의 정보를 결합해서 차이를 어느 정도 줄일 수 있을 것으로 기대한다.

지금까지 추정방법론적 비교를 통하여 다변량 회귀에서 정보적 설명 변수를 통해 중심부분공간을 추정하고자 할때, 이중 군집 평균 방법보다는 방법론적 제약이 비교적 적은 재표본-투영 평균 방법을 이용하는 것을 제안하는 바이다.

4. 모의 실험

이중 군집 평균 방법과 재표본-투영 평균 방법의 모의실험을 통한 비교를 위하여 Ko와 Yoo (2022)에서 사용된 다음의 변수 설정을 사용하고자 한다:

1) $U_1 \sim U(0,1) \perp\!\!\!\perp e \sim U(-0.5, 0.5) \perp\!\!\!\perp (W_1, W_2, W_3) \overset{iid}{\sim} N(0,1)$; $U_2 = \log(U_1) + e$. 여기서 $U(a,b)$ a와 b 구간의 균일분포를 의미한다.

2) 차원이 10인 설명 변수 $(X_1, \ldots, X_{10})^{\mathrm{T}}$는 다음과 같이 생성된다: were generated: $X_1 = U_1 + W_1$; $X_2 = U_2 + W_4 + W_5$; $X_3 = W_1 - W_2$; $X_4 = W_2$; $X_5 = W_3$; $(X_6, \ldots, X_{10}) \overset{iid}{\sim} N(0,1)$.

이 설명변수를 이용하여 다음의 다변량 회귀모형 Model1, Model 2, Model 3를 생성하였다.

Model 1

$$Y_1 = \exp\left(0.5\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X} + 1\right) + 0.1\varepsilon_1;\ Y_2 = \boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X} + \left(\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X}\right)^2 + 0.1\varepsilon_2.$$

Model 2

$$Y_1 = \exp\left(0.5\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X} + 1\right) + 0.1\varepsilon_1;\ Y_2 = \boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X} + \left(\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X}\right)^2 + 0.1\varepsilon_2.\ Y_3 = 0.1\varepsilon_3;\ Y_4 = 0.1\varepsilon_4.$$

Model 3

$$Y_1 = \exp\left(0.5\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X} + 1\right) + 0.1\varepsilon_1;\ Y_2 = \boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X} + \left(\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X}\right)^2 + 0.1\varepsilon_2.\ Y_3 = Y_1 + Y_2 + 0.1\varepsilon_3;\ Y_4 = \mid \boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X} \mid + 0.1\varepsilon_4.$$

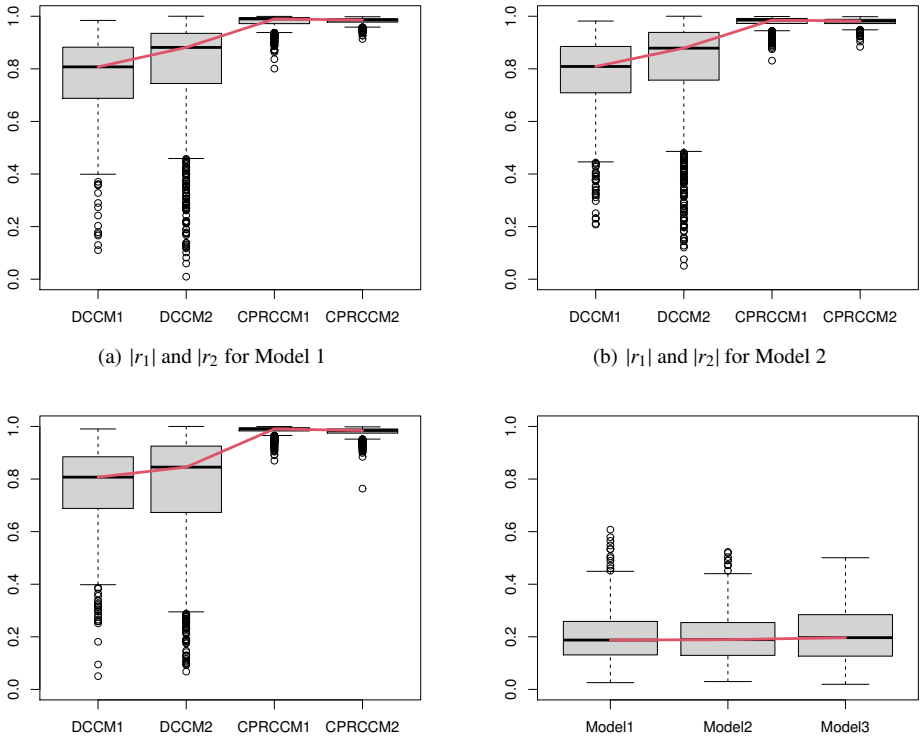
여기서, $\boldsymbol{\eta}_1 \in \mathbb{R}^{10} = (1, -1 - 1, 0, \ldots, 0)^{\mathrm{T}}$이고 $\boldsymbol{\eta}_2 \in \mathbb{R}^{10} = (1, 0, 0, \ldots, 0)^{\mathrm{T}}$. 각각의 모형에서 $\varepsilon 1, \ldots, \varepsilon_4$는 각각 그리고 설명 변수와 독립적으로 표준정규분포에서 생성되었다. 또한 각각의 모형은 100개의 표본수를 가지며 1,000번 반복되었다. 위의 세 모형은 모두 $\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X}$와 $\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X}$에 의존하므로, 중심부분공간은 $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)^{\mathrm{T}}$에 의해 생성되고, 차원은 2이다.

재표본-투영 평균 방법으로는 DCCM과 직접적이고 공평한 비교를 위하여 Ko와 Yoo (2022)이 제안한 방법 중 요소 재표본-투영 평균 방법론(CPRCCM)을 사용한다. 요소 재표본-투영 방법은 3.3절에서 언급한 대로, 요소별 회귀의 정보와 재표본-투영 방법론 일부를 결합한 것이고, 여기서 재표본수는 Ko와 Yoo (2022)에서 사용된대로 100번 추출하였다.

모의실험에서 $\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X}$와 $\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X}$을 얼마나 잘 추정했느냐를 측정하기 위해 먼저 DCCM과 CPRCCM으로 추정한 2차원의 중심 부분 공간 기저의 추정값 $\hat{\boldsymbol{\eta}}_1$와 $\hat{\boldsymbol{\eta}}_2$를 구한다. 이후 $\hat{\boldsymbol{\eta}}_1^{\mathrm{T}}\mathbf{X}$와 $\hat{\boldsymbol{\eta}}_2^{\mathrm{T}}\mathbf{X}$을 계산하고, 이를 설명 변수로 하고 $\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X}$와 $\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X}$을 반응 변수로 하는 선형회귀모형을 적합하여 결정 계수을 계산하고, 이에 대한 양의 제곱근을 구한다. 이를 각각 $|r_1|$과 $|r_2|$라고 정의한다. 예를 $|r_1|$과 $r_2|$은 모두 0에서 1사이에 있고, 1에 가까울 수록 $\hat{\boldsymbol{\eta}}_1^{\mathrm{T}}\mathbf{X}$와 $\hat{\boldsymbol{\eta}}_2^{\mathrm{T}}\mathbf{X}$ 에 의해서 $\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X}$와 $\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{X}$이 잘 추정됨을 의미한다. 그리고 DCCM과 CPRCCM의 추정값이 얼마나 유사한지를 측정하기 위해 Part 등 (2022)에서 사용된 trace correlation distnace를 사용할 것이다. 이는 다음과 같이 계산된다. $\hat{\boldsymbol{\eta}}_{\mathrm{DC}}$와 $\hat{\boldsymbol{\eta}}_{\mathrm{PR}}$을 각각 DCCM과 CPRCCM에 의해 계산된 $\boldsymbol{\eta}$의 추정값이라고 하자. 그럼 trace correlaton distance는 다음과 같이 계산된다:

$$\sqrt{\frac{1}{2}\mathrm{trace}\left(\hat{\boldsymbol{\eta}}_{\mathrm{DC}}\left(\hat{\boldsymbol{\eta}}_{\mathrm{DC}}^{\mathrm{T}}\hat{\boldsymbol{\eta}}_{\mathrm{DC}}\right)^{-1}\hat{\boldsymbol{\eta}}_{\mathrm{DC}}^{\mathrm{T}}\hat{\boldsymbol{\eta}}_{\mathrm{PR}}\left(\hat{\boldsymbol{\eta}}_{\mathrm{PR}}^{\mathrm{T}}\hat{\boldsymbol{\eta}}_{\mathrm{PR}}\right)^{-1}\hat{\boldsymbol{\eta}}_{\mathrm{PR}}^{\mathrm{T}}\right)}.$$

Trace correlation distance는 거리개념이고 0과 1사이의 값을 갖는다. 0에 가까우면 $\hat{\boldsymbol{\eta}}_{\mathrm{DC}}$와 $\hat{\boldsymbol{\eta}}_{\mathrm{PR}}$이 서로 유사함을 의미하고, 1에 가까우면 직교에 가까워짐을 의미한다.



(a) $|r_1|$ and $|r_2$ for Model 1

(b) $|r_1|$ and $|r_2|$ for Model 2

(c) $|r_1|$ and $|r_2|$ for Model 3

(d) Distances between estimates of DCCM and CPRCCM

Figure 1: *Summary Boxplots for Models 1–3.*

Figure 1에 DCCM과 CPRCCM에 의해 계산된 $|r_1|$과 $|r_2|$와 trace correlation distance의 박스그림을 모의실험에 대한 요약으로 제시한다. Figure 1에서 DCCM1과 DCCM2는 DCCM 의해 계산된 $|r_1|$과 $|r_2|$을 의미하고, CPRCCM1과 CPRCCM2도 이와 동일한 의미를 갖는다.

Figure 1(a)–(c)에서 알 수 있듯이 CPRCCM이 DCCM보다 $\boldsymbol{\eta}_1$과 $\boldsymbol{\eta}_2$을 더 정확하게 추정함을 알 수 있고, 또한 CPRCCM이 DCCM보다 변동이 더 적음을 확인할수 있다. DCCM과 CPRCCM의 변동을 보다 면밀하게 파악하기 위해 범위과 사분위 범위를 Table 1에 정리를 하였다. Table 1에 따르면 DCCM의 경우 사분위 범위가 범위에 비해 매우 짧아짐을 확인할 수 있다. 이는 앞 절에서 언급했듯이 설명 변수와 반응 변수을 모두 군집화를 통환 범주화에서 특정 범주에 적은 표본이 배정됨으로써 발생한 것으로 생각할 수 있다.

Figure 1(d)을 살펴보면 DCCM과 CPRCCM의 추정값에 대한 trace correlation distance의 중앙값이 Models 1–3에 대해 서로 비슷하고, 그 값이 0.2임을 확인할 수 있다. 이를 통해 DCCM과 CPRCCM의 추정값에 있어 다소 차이가 있고, 중심부분공간의 실제 기저인 $\boldsymbol{\eta}$을 기준으로 볼 때, DCCM이 CPRCCM 보다 더 많은 편차를 갖고 있음을 확인할 수 있다.

모의실험 결과 다변량 회귀분석에서 정보적 설명 변수를 통해 중심부분공간을 추정할 경우 이중 군집화를 통한 추정보다는 투영-재표본 방법을 통해 다변량 반응 변수를 일변량 반응 변수로 변환하여 추정하는 것이 추정의 정확도와 효율성을 높일 수 있음을 확인할 수 있다.

Table 1: Range and inter-quartile range of $|r_1|$ and $|r_2|$ of DCCM and CPRCCM for Models 1–3: IQR, inter-quartile range

	Model 1		Model 2		Model 3	
	Range	IQR	Range	IQR	Range	IQR
DCCM1	0.874	0.194	0.774	0.176	0.940	0.196
DCCM2	0.990	0.191	0.949	0.182	0.932	0.252
CPRCCM1	0.199	0.023	0.168	0.019	0.129	0.011
CPRCCM2	0.085	0.012	0.115	0.016	0.235	0.015

5. 결론

Ko와 Yoo (2022)는 정보적 설명 변수 공간에 포함되지만 중심부분공간을 포함하는 투영-재표본 정보적 설명 변수 공간을 정의하였고, 이를 추정하기 위해 투영-재표본 평균 방법을 제안하였다. 본 논문에서는 반응 변수가 다차원인 다변량 회귀분석에서 정보적 설명 변수 공간의 추정에 대한 이중 군집 평균 방법론을 제시하고 있다.

이 두 방법론은 설명 변수와 반응 변수를 모두 범주화해야하는 공통점이 있으나, 투영-재표본 평균 방법의 경우 다변량 반응 변수를 일변량 반응 변수로 변환한 후 슬라이싱을 통해 범주화를 하는 반면, 본 논문에서 제안하는 이중 군집 평균 방법은 다변량 반응 변수를 군집화를 통해 범주화하는 차이가 존재한다.

모의실험에 따르면 중심부분공간의 추정에 정확도는 반응 변수의 범주화의 차이에 크게 의존하고 있다. 이중 군집 평균 방법론의 경우 반응 변수를 군집화하면서 최종 범주의 표본수에 심한 불균형이 야기될 수 있어, 추정의 정확도와 효율성을 떨어뜨린다. 반면 투영-재표본 평균 방법의 경우 슬라이싱을 이용한 반응 변수의 범주화에서 이러한 자료의 수에 대한 불균형을 어느 정도 제어할 수 있음을 확인할 수 있다.

따라서 다변량 회귀분석에서 정보적 설명 변수 공간을 통해 중심부분공간을 추정을 할 경우 이중 군집 평균 방법보다는 투영-재표본 평균 방법의 사용이 제안된다.

References

Ko S and Yoo JK (2022). Projective resampling estimation of informative predictor subspace for multivariate regression, *Journal of the Korean Statistical Society*, Available from: https://doi.org/10.1007/s42952-022-00178-0

Lee K, Choi Y, Um HY, and Yoo JK (2019). On fused dimension reduction in multivariate regression, *Chemometrics and Intelligent Laboratory Systems*, **193**, 103828.

Li B, Wen S, and Zhu L (2008). On a projective resampling method for dimension reduction with multivariate responses, *Journal of the American Statistical Association*, **103**, 1177–1186.

Li KC (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistcal Association*, **86**, 316–342.

Park Y, Kim K, and Yoo JK (2022). On cross-distance selection algorithm for hybrid sufficient dimension reduction, *Computational Statistics and Data Analysis*, **176**, 107562.

Setodji CM and Cook RD (2004). *K*-means inverse regression, *Technometrics*, **46**, 421–429.

Yin X and Bura E (2006). Moment-based dimension reduction for multivariate response regression, *Journal of Statististical Planning and Inference*, **136**, 3675–3688.

Yoo JK and Cook RD (2007). Optimal sufficient dimension reduction for the conditional mean in multivariate regression, *Biometrika*, **943**, 231–242.

Yoo JK (2008). A novel moment-based sufficient dimension reduction approach in multivariate regresison, *Com putational Statistics and Data Analysis*, **52**, 3843–3851.

Yoo JK (2016). Sufficient dimension reduction through informative predictor subspace, *Statistics : A Journal of Theoretical and Applied Statistics*, **50**, 1086–1099.

Yoo JK (2018). Tutorial: Dimension reduction in regression with a notion of sufficiency, *Communications for Statistical Applications and Methods*, **23**, 93–103.

Received August 17, 2022; Revised August 22, 2022; Accepted August 23, 2022

다변량회귀에서 정보적 설명 변수 공간의 추정과 투영-재표본 정보적 설명 변수 공간 추정의 고찰

유재근[1,a]

[a]이화여자대학교 통계학과

요 약

정보적 설명 변수 공간은 일반적인 충분차원축소 방법들이 요구하는 가정들이 만족하지 않을 때 중심부분 공간을 추정하기 위해 유용하다. 최근 Ko와 Yoo (2022)는 다변량 회귀에서 Li 등 (2008)이 제시한 투영-재표본 방법론을 사용하여 정보적 설명 변수 공간이 아닌 투영-재표본 정보적 설명 변수 공간을 새로이 정의하였다. 이 공간은 기존의 정보적 설명 변수 공간에 포함되지만 중심 부분 공간을 포함한다. 본 논문에서는 다변량 회귀에서 정보적 설명 변수 공간을 직접적으로 추정할 수 있는 방법을 제안하고, 이를 Ko와 Yoo (2022)가 제시한 방법과 이론적으로 그리고 모의실험을 통해 비교하고자 한다. 모의실험에 따르면 Ko-Yoo 방법론이 본 논문에서 제시한 추정 방법보다 더 정확하게 중심 부분 공간을 추정하고, 추정값들의 변동이 적다는 측면에서 보다 더 효율적임을 알 수 있다.

주요용어: 군집 평균 방법, 다변량 회귀분석, 정보적 설명 변수 공간, 충분차원축소, 투영-재표본 평균 방법, 투영-재표본 정보적 설명 변수 공간

2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아수행된 기초연구사업임 (NRF2019R1F1A1050715).

[1](03760) 서울시 서대문구 이화여대길 52, 이화여자대학교 통계학과. E-mail: peter.yoo@ewha.ac.kr