The Korean Journal of Applied Statistics
2021, Vol. 34, No. 4, 659–669

DOI: https://doi.org/10.5351/KJAS.2021.34.4.659
Print ISSN1225-066X / Online ISSN 2383-5818

Principal selected response reduction in multivariate regression

Jae Keun Yoo [1,a]

[a]Department of Statistics, Ewha Womans University

Abstract

Multivariate regression often appears in longitudinal or functional data analysis. Since multivariate regression involves multi-dimensional response variables, it is more strongly affected by the so-called curse of dimension that univariate regression. To overcome this issue, Yoo (2018) and Yoo (2019a) proposed three model-based response dimension reduction methodologies. According to various numerical studies in Yoo (2019a), the default method suggested in Yoo (2019a) is least sensitive to the simulated models, but it is not the best one. To release this issue, the paper proposes an selection algorithm by comparing the other two methods with the default one. This approach is called *principal selected response reduction*. Various simulation studies show that the proposed method provides more accurate estimation results than the default one by Yoo (2019a), and it confirms practical and empirical usefulness of the propose method over the default one by Yoo (2019a).

Keywords: model-based response reduction, multivariate regression, principal fitted response reduction, principal response reduction, unstructured principal fitted response reduction

1. 서론

회귀분석은 p차원의 설명변수 $\mathbf{X} \in \mathbb{R}^p$ 가 주어졌을 때 반응변수의 조건부 분포 ($\mathbf{Y} \in \mathbb{R}^r|\mathbf{X}, r \geq 1$) 를 연구하는 것을 목적으로 한다. 하지만 실제로는 분포를 알아가기보다는 일반적으로 반응변수의 평균적 변화에 관심을 갖고, 이에 따라 조건부 평균 $E(\mathbf{Y}|\mathbf{X})$을 통계적 모형으로 추정을 한다. 다변량회귀분석은 반응변수의 차원이 2 이상인 회귀분석을 의미한다. 다차원의 반응변수는 경시적 자료분석이나 반복측정자료 혹은 함수적 자료에서 빈번하게 나타나고 있다. 다차원 반응변수를 가진 통계적 분석은 쉽지 않다.

예로 두 그룹간의 반응변수의 평균비교를 생각해보자. 만약 반응변수가 일변량이라면 단순히 t 검정에 의해 쉽게 해결될 수 있다. 하지만 반응변수가 다차원이라면 단순 t검정이 이보다는 확실히 더 복잡한 Hotelling T^2검정으로 전환되어지고, 표본의 수가 일정할 때 차원이 높아질수록 검정의 정확도 떨어지게 된다. 이 예제에서 그룹을 설명변수로 간주한다면, 이 문제는 회귀분석의 문제로 전환된다. 여기서 설명변수가 추가적 고려를 생각해보자. 그렇다면 추정해야 할 회귀계수의 수는 일변량 회귀분석에 비해 반응변수의 수 만큼 곱해져서 증가하게 된다. 즉 반응변수가 4차원이고 5개의 설명변수를 고려한 상황에서 다변량 선형회귀모형을 적합할 때, 절편을 제외한 기울기에 해당되는 회귀계수의 수는 20개가 된다. 따라서 고차원의 반응변수는 그 자체만으로 통계적 모형을 통한 자료분석을 어렵게 만들 수 있다.

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education (NRF-2021R1F1A1059844).

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

Published 31 August 2021 / journal homepage: http://kjas.or.kr

회귀분석에서 충분차원축소(sufficient dimension reduction; SDR)은 조건부 분포 $\mathbf{Y}|\mathbf{X}$의 다양한 특성 중 특정 부분에 대한 정보의 손실없이 설명변수 $\mathbf{X}$의 차원을 축소하는 방법이다. 예를 들어 $\mathbf{Y}|\mathbf{X}$의 특정 부분으로 회귀분석의 주 관심사인 $E(\mathbf{Y}|\mathbf{X})$를 고려해보자. 이러한 경우 SDR은 다음의 조건을 만족하는 $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}(d \le p)$을 찾는 것이 목적이다.

$$E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{Y}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}).$$

그리고 위의 조건을 만족하는 $\boldsymbol{\eta}$를 찾는 다양한 SDR 방법들이 제시되었다 (Cook, 1998; Cook과 Li, 2002; Yoo 등, 2009).

하지만 실제로 설명변수의 차원을 축소한다고 할 지라도 위에서 언급했듯이 다차원의 반응변수의 존재는 통계적 모형화를 자주 어렵게 만든다. 이러한 문제를 해결하기 위해 Yoo와 Cook (2008)은 SDR 분야에서 최초로 반응변수의 차원을 $E(\mathbf{Y}|\mathbf{X})$의 정보적 손실없이 축소하는 이론적 토대를 마련하였다. 이를 위해 Yoo와 Cook (2008)은 다음의 조건을 만족하는 $\boldsymbol{\Gamma}$에 대한 추정해야 한다고 제안하고 있다.

$$E(\mathbf{Y}|\mathbf{X}) = E\left\{\mathbf{P}^{\mathrm{T}}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_y)}\mathbf{Y}|\mathbf{X}\right\}. \tag{1.1}$$

여기서 $q \le r$이며 $\boldsymbol{\Sigma}_y = \mathrm{cov}(\mathbf{Y})$이다. 그리고, $\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_y)} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Sigma}_y\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Sigma}_y$는 $< v_1, v_2 >_{\boldsymbol{\Sigma}_y} = v_1^{\mathrm{T}}\boldsymbol{\Sigma}_y v_2$을 내적으로 사용하는 직교투영행렬이다. 위의 조건을 만족하는 $\boldsymbol{\Gamma}$를 추정하기 위해 Yoo와 Cook (2008)은 $\boldsymbol{\Sigma}_y^{-1}\,\mathrm{cov}(\mathbf{Y}, \mathbf{X})\boldsymbol{\Sigma}_x^{-1}$을 제안하였다. 이 방법은 모형에 대해 구체적 가정을 하지 않은 비모수적 방법론이다.

최근의 SDR의 연구 (Cook, 2007)에 따르면 비모수적 방법론 보다는 모형기반의 반모수적 방법론이 더 정확하게 설명변수의 차원을 축소의 가능성을 제시하였다. 최근 Yoo (2018)은 Yoo와 Cook (2008)의 이론적 토대에 반응변수의 차원축소를 위해 주 반응변수 차원축소(principal response reduction)와 주적합 반응변수 차원축소(principal fitted response reduction)라는 두 가지의 모형기반 반모수적 방법론을 제시하였다. 이후 Yoo (2019a)에서는 필요한 공분한 행렬을 제약조건이 없다는 가정에서 추정하는 모형기반 반응변수 차원축소 방법론인 비구조적 주적합반응변수차원축소(unstructured principal fitted response reduction)을 최근에 제시하고 있다. 그리고 Yoo (2019b)에서는 위에서 언급된 Yoo와 Cook (2008), Yoo (2018)과 Yoo (2019a)에 대해 이론적인 비교 및 차이점과 장단점에 대해 설명을 하고 있다.

Yoo (2019a)에서 개발된 비구조 주적합 반응변수 차원축소 방법론은 안정적인 차원검정을 제시하고 있다. 하지만 차원이 알려진 상황에서 $\boldsymbol{\Gamma}$의 추정의 정확도는 Yoo (2018)에서 제시된 주반응변수차원축소와 주적합 반응변수차원축소보다 더 뛰어나지는 않고, 두 방법 중 더 뛰어난 것에 근접하다는 것이 모의실험을 통해 발견하였다.

이러한 발견을 기반으로 본 논문에서는 비구조 주적합 반응변수 차원축소 방법론과 결과를 주 반응변수 차원축소와 주적합 반응변수 차원축소와 각각 비교한 후 더 나은 방법을 선택하여 제시하는 방법을 고려하고자 한다. 이는 새로운 방법론의 개발은 아니지만, 기존의 세가지 모형기반 방법론의 자료기반 비교를 통해 가장 적절한 방법을 제시하는 것이다. Yoo (2018)과 Yoo (2019a)에서 제시된 다양한 모형에서의 모의실험을 통해 선택의 우수성을 증명하고, 이러한 증명을 바탕으로 실제 자료분석에 적용하고자 한다.

본 논문의 차례는 다음과 같다. 2장에서는 모형기반 반응변수 차원축소에 대해 설명을 할 것이고, 3장은 주선택 반응변수 차원축소에 대한 알고리듬을 소개할 것이다. 이후 4장은 모의실험과 실제 자료분석을 통해 주선택 반응변수 차원축소의 유용성을 살펴볼 것이고, 5장에는 결론이 제시된다.

2. 모형기반 반응변수 차원축소

2.1. 주 반응변수 차원축소(principal response reduction)

모형기반 반응변수 차원축소를 위해서 우선 다음의 모형을 가정한다:.

$$\mathbf{Y} = \boldsymbol{\Gamma}\boldsymbol{\nu}_{\mathbf{X}} + \boldsymbol{\varepsilon}. \tag{2.1}$$

식 (2.1)에서 다음의 가정을 한다.

(a) $\boldsymbol{\Gamma}$는 $\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma} = \mathbf{I}_d$을 만족하는 $r \times d$ ($d \le r$)행렬이다. $\boldsymbol{\Gamma}_0$는 $r \times (r-d)$ 행렬이고 다음의 두 성질을 만족한다. $\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Gamma}_0 = \mathbf{I}_{r-d}$; $\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Gamma} = \mathbf{0}$.

(b) $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma} \in \mathbb{R}^{r\times r})$.

(c) $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} + \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^{\mathrm{T}}$.

(d) $\boldsymbol{\nu}_{\mathbf{X}}$는 알려지지 않은 양정치 공분산 행렬을 갖는 d차원의 설명변수 $\mathbf{X}$의 함수이며, $\boldsymbol{\nu}$에 대해서 $\sum_{\mathbf{X}} \boldsymbol{\nu}_{\mathbf{X}} = \mathbf{0}$ 이다.

(e) $\mathrm{cov}(\boldsymbol{\nu}_{\mathbf{X}}, \boldsymbol{\varepsilon}) = \mathbf{0}$이다.

식 (2.1)에서 제시된 모형에서 $\boldsymbol{\nu}_{\mathbf{X}}$를 $\mathbf{X}$로 대체하면, 이 식은 다변량 선형회귀모형이 되고, $\boldsymbol{\Gamma}$는 회귀모수행렬이 된다. 식 (2.1)에서 제시된 모형이 위에서 제시된 조건 (a)–(e)를 만족한다는 가정 하에, (2.1)에 있는 $\boldsymbol{\Gamma}$는 식 (1.1)을 만족하게 되어, 반응변수의 차원을 축소할 수 있다. Yoo (2018)은 $\boldsymbol{\Gamma}$에 대한 최대가능도 추정량이 $\hat{\boldsymbol{\Sigma}}_y$의 가장 큰 처음의 d개의 고유값에 대응하는 고유벡터들임을 증명하였고, 이 방법을 주 반응변수 차원축소(principal response reduction; PRR)로 명명하였다. PRR에서 $\boldsymbol{\Sigma}$은 과적합으로 인해 추정 불가능로 인해 $\boldsymbol{\Gamma}$의 차원에 대한 가능도 검정은 가능하지 않다.

2.2. 주적합 반응변수 차원축소(principal fitted response reduction)

실제로 PRR의 경우 회귀분석에서 반응변수의 차원을 축소하는 과정에서 설명변수의 정보가 개입되지 않는데, 그 이유는 알려지지 않은 $\boldsymbol{\nu}_{\mathbf{X}}$의 최대 가능도 추정량이 $\boldsymbol{\Gamma}\mathbf{Y}$가 되기 때문이다. $\boldsymbol{\Gamma}$의 추정에 설명변수의 정보를 활용하기 위해서 다음과 같이 식 (2.1)에 있는 $\boldsymbol{\nu}_{\mathbf{X}}$을 $\boldsymbol{\psi}\mathbf{f}_{\mathbf{X}}$로 변형한다.

$$\mathbf{Y} = \boldsymbol{\Gamma}\boldsymbol{\psi}\mathbf{f}_{\mathbf{X}} + \boldsymbol{\varepsilon}. \tag{2.2}$$

여기서 $\boldsymbol{\psi}$는 알려지지 않은 $d \times q$ 행렬이고, $\mathbf{f}_{\mathbf{X}}$는 q차원의 알려진 $\mathbf{X}$의 벡터형태의 함수이고 $\sum_{\mathbf{X}} \mathbf{f}_{\mathbf{X}} = 0$를 가정한다. Yoo (2018)에서 $\mathbf{f}_{\mathbf{X}}$의 후보 형태로 $\mathbf{X}$, $\mathbf{X}^2$, $\exp(\mathbf{X})$와 이들의 조합 그리고 K-평균군집으로 만들어지는 $\mathbf{X}$의 군집지시행렬(cluster indicator matrix)를 제안하고 있다. Yoo (2018)은 모의실험을 통하여 $\boldsymbol{\nu}_{\mathbf{X}} = \boldsymbol{\psi}\mathbf{f}_{\mathbf{X}}$ 관계가 성립되지 않더라도 이 후보군의 사용으로 $\boldsymbol{\Gamma}$가 잘 추정됨을 보여주었다. 식 (2.2)에서 정의된 모형에서 $\boldsymbol{\Gamma}$에 대한 최대가능도 추정량을 위해서 다음에 대한 기호를 정의한다.

$\mathbb{Y}$: $n \times r$ 반응변수 자료 행렬; $\mathbb{X}$: $n \times p$ 설명변수 자료 행렬,

$\mathbb{F}$: q차원의 $\mathbf{f}_{\mathbf{X}}$을 행으로 쌓아서 만든 $n \times q$ 행렬,

$\mathbf{P}_{\mathbb{F}} = \mathbb{F}(\mathbb{F}^{\mathrm{T}}\mathbb{F})^{-1}\mathbb{F}^{\mathrm{T}}$;. $\hat{\boldsymbol{\Sigma}}_{\mathrm{fit}} = \frac{1}{n}\mathbb{Y}^{\mathrm{T}}\mathbf{P}_{\mathbb{F}}\mathbb{Y}/n$; $\hat{\boldsymbol{\Sigma}}_{\mathrm{res}} = \hat{\boldsymbol{\Sigma}}_y - \hat{\boldsymbol{\Sigma}}_{\mathrm{fit}}$.

식 (2.2)에서 정의된 $\boldsymbol{\Gamma}$ 역시 식 (1.1)을 만족하고, $E(\mathbf{Y}|\mathbf{X})$의 정보손실없이 반응변수의 차원을 축소할 수 있다. Yoo (2018)에 따르면 식 (2.2) $\boldsymbol{\Gamma}$의 최대가능도 추정량이 PRR과 달리 해석적으로 표현할 수 없고, $\hat{\boldsymbol{\Sigma}}_y$와 $\hat{\boldsymbol{\Sigma}}_{\mathrm{res}}$에 의존하여 수치적으로 추정량을 구해야 한다. 이를 위해 $\hat{\boldsymbol{\Sigma}}_y$, $\hat{\boldsymbol{\Sigma}}_{\mathrm{fit}}$와 $\hat{\boldsymbol{\Sigma}}_{\mathrm{res}}$의 고유벡트터들을 Cook (2007)에서 제시된 축차적 선택 알고리듬을 이용하여 $\boldsymbol{\Gamma}$을 추정하였다. 식 (2.2)에서 축차적 선택 알고리듬을 이용하여

$\boldsymbol{\Gamma}$을 추정하는 방법을 주적합 반응변수 차원축소(principal fitted response reduction; PFRR)라고 한다. PFRR은 $\boldsymbol{\Gamma}$의 차원인 d에 대한 $\mathrm{H}_0 : d = m$ versus $\mathrm{H}_1 : d = r, m = 0, 1, \ldots, (r-1)$,의 검정이 가능하다. H_0하에서 검정통계량은 $\chi_{q(r-m)}$분포를 따른다.

2.3. 비구조적 주적합 반응변수 차원축소(unstructured principal fitted response reduction)

PFRR에서 $\boldsymbol{\Gamma}$는 $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} + \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^{\mathrm{T}}$의 조건에 대한 제약상황 아래서 추정이 된다. 이러한 제약조건으로 인해 PFRR에서 $\boldsymbol{\Gamma}$의 최대가능도 추정량은 해석적 식이 아닌 수치적으로 구해진다. 실제로 $\boldsymbol{\Sigma}$에 대한 위의 조건은 $\boldsymbol{\Gamma}$가 $E(\mathbf{Y}|\mathbf{X})$에 대한 정보의 손실없이 $\mathbf{Y}$의 차원을 축소하는 데 필요한 조건이지 $\boldsymbol{\Gamma}$를 추정하는 데 반드시 필요한 조건은 아니다. 따라서, $\boldsymbol{\Sigma}$에 대한 조건은 그대로 유지하되, $\boldsymbol{\Gamma}$를 가능도 함수를 이용하여 추정하는 데 이 조건을 고려하지 않는다면 $\boldsymbol{\Gamma}$에 대한 해석적 해를 찾을 수 있다. 이를 기술하기 위해 다음의 기호를 정의한다.

$\mathbf{E}_d$: 정방행렬 $\mathbf{E}$의 가장 큰 처음의 d개의 고유값에 대응하는 고유벡터들

$\mathbf{B} = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\Sigma}}_{\mathrm{fit}}\hat{\boldsymbol{\Sigma}}^{-1/2}$, $\mathbf{B}_{\mathrm{res}} = \hat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\hat{\boldsymbol{\Sigma}}_{\mathrm{fit}}\hat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}$ 그리고 $\mathbf{B}_y = \hat{\boldsymbol{\Sigma}}_y^{-1/2}\hat{\boldsymbol{\Sigma}}_{\mathrm{fit}}\hat{\boldsymbol{\Sigma}}_y^{-1/2}$

$\mathbf{B}_{\mathrm{res}}$에 대한 스테트럼분해: $\mathbf{B}_{\mathrm{res}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_r)\,\mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_r)(\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_r)^{\mathrm{T}}$ 그리고 $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_r \geq 0$

$\hat{\mathbf{V}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_a)$, $\hat{\mathbf{K}}_d = \mathrm{diag}(0, \ldots, 0, \hat{\lambda}_{d+1}, \ldots, \hat{\lambda}_a)$ 그리고 $a = \min(q, r)$

Yoo (2019a)는 다음을 증명하였다

(1) $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{\mathrm{res}} + \hat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\hat{\mathbf{V}}\hat{\mathbf{K}}_d\hat{\mathbf{V}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2} = \hat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}(\mathbf{I}_r + \hat{\mathbf{V}}\hat{\mathbf{K}}_d\hat{\mathbf{V}}^{\mathrm{T}})\hat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}$.

(2) $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{B}_d$, $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\mathbf{B}_{\mathrm{res}d}$ 혹은 $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\Sigma}}_y^{-1/2}\mathbf{B}_{y_d}$.

위의 결과로 $\boldsymbol{\Gamma}$를 추정하는 방법을 비구조적 주적합 반응변수 차원축소(unstructured principal fitted response reduction; UPFRR)라고 한다. Yoo (2019a)에서는 위의 (2)에 있는 세가지 해중에서 $\hat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1/2}\mathbf{B}_{\mathrm{res}d}$를 디폴트로 사용한다. PFRR과 마찬가지로 UPFRR 역시 $\boldsymbol{\Gamma}$에 대한 차원검정을 할 수 있고, 검정통계량은 H_0하에서 $\chi_{(q-m)(r-m)}$분포를 따른다.

위의 세가지 모형기반 반응변수 차원축소에 대한 심도있는 논의를 위해서 Yoo (2019b)를 참고하길 바란다.

3. 주선택 반응변수 차원축소(principal selected response reduction)

Yoo (2019a)은 모형기반 반응변수 차원축소에서 제시된 세 가지 방법론의 $\boldsymbol{\Gamma}$ 추정의 다양한 모의실험 결과를 제시하고 있다. 고려된 모형에 따라 PRR 혹은 PFRR이 가장 좋은 추정결과를 보여주고 있고, UPFRR은 PRR와 PFRR 중 더 좋은 추정을 하는 방법보다는 다소 못하고, 더 나쁜 추정을 하는 방법보다는 더 우수한 추정결과를 보여준다. 따라서 UPFRR은 Yoo (2019a)에서 고려된 모든 모의실험 모형에서 극단적으로 나쁘거나 좋은 추정을 하기 보다는 PRR와 PFRR 중 더 나은 추정을 하는 방법과 유사하게 추정을 한다. 또한 $\boldsymbol{\Gamma}$의 차원 검정에서는 PFRR보다는 다소 더 안정적인 결과를 제시하고 있다. 이러한 이유로 Yoo (2019a)에서는 UPFRR을 세 가지 방법 중 기본방법론으로 사용하기를 권장하고 있다. 하지만 이는 PRR과 PFRR이 UPFRR보다 더 나은 추정결과를 보여주는 것을 간과하고 있다. 따라서, 실제로는 UPFRR의 추정결과를 기반으로 PRR과 PFRR 중 하나를 선택하는 것이 더 바람직할 수 있다. 이러한 논의를 바탕으로 다음의 방법 선택 알고리듬을 제안하고자 한다. 아래의 알고리듬으로 $\boldsymbol{\Gamma}$를 추정하는 방법을 주선택 반응변수 차원축소(principal selected response reduction; PSRR)라고 한다.

• **모형기반 방법 선택 알고리듬**

1. 주어진 자료에 대해 UPFRR을 적합하고, $\boldsymbol{\Gamma}$에 대한 차원 검정을 실시한다. 이때 검정을 통해 추정된 $\boldsymbol{\Gamma}$의 차원을 $\hat{d}$라고 하자.
2. UPFRR에서 $\boldsymbol{\Gamma}$에 대해 다음의 추정량을 얻는다: $\hat{\boldsymbol{\Gamma}}_{\text{UPFRR}} = \hat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\mathbf{B}_{\text{res}\hat{d}}$.
3. $\hat{d}$가 주어진 상황에서 PRR과 PFRR로 $\boldsymbol{\Gamma}$을 추정하고 이를 각각 $\hat{\boldsymbol{\Gamma}}_{\text{PRR}}$ 과 $\hat{\boldsymbol{\Gamma}}_{\text{PFRR}}$ 이라고 하자.
4. $\mathbf{P}_{\text{PRR.U}} = \hat{\boldsymbol{\Gamma}}_{\text{PRR}}(\hat{\boldsymbol{\Gamma}}_{\text{PRR}}^{\text{T}}\hat{\boldsymbol{\Gamma}}_{\text{PRR}})^{-1}\hat{\boldsymbol{\Gamma}}_{\text{PRR}}^{\text{T}}\hat{\boldsymbol{\Gamma}}_{\text{UPFRR}}(\hat{\boldsymbol{\Gamma}}_{\text{UPFRR}}^{\text{T}}\hat{\boldsymbol{\Gamma}}_{\text{UPFRR}})^{-1}\hat{\boldsymbol{\Gamma}}_{\text{UPFRR}}^{\text{T}}$와
   $\mathbf{P}_{\text{PFRR.U}} = \hat{\boldsymbol{\Gamma}}_{\text{PFRR}}(\hat{\boldsymbol{\Gamma}}_{\text{PFRR}}^{\text{T}}\hat{\boldsymbol{\Gamma}}_{\text{PFRR}})^{-1}\hat{\boldsymbol{\Gamma}}_{\text{PFRR}}^{\text{T}}\hat{\boldsymbol{\Gamma}}_{\text{UPFRR}}(\hat{\boldsymbol{\Gamma}}_{\text{UPFRR}}^{\text{T}}\hat{\boldsymbol{\Gamma}}_{\text{UPFRR}})^{-1}\hat{\boldsymbol{\Gamma}}_{\text{UPFRR}}^{\text{T}}$를 계산한다.
5. $\mathbf{P}_{\text{PRR.U}}$와 $\mathbf{P}_{\text{PFRR.U}}$의 가장 큰 $\hat{d}$개의 고유값의 산술평균의 제곱근을 각각 $r_{\text{PRR.U}}$와 $r_{\text{PFRR.U}}$라고 하자. 이렇게 계산된 $r_{\text{PRR.U}}$와 $r_{\text{PFRR.U}}$은 trace correlation coefficient (Hooper, 1959)이다. 계산된 $r_{\text{PRR.U}}$ 값이 1에 가까울 수록 PRR과 UPFRR의 동일성은 증가하게 된다. 즉 만약 $r_{\text{PRR.U}} > r_{\text{PFRR.U}}$ 라면, UPFRR은 PRR과 더 근접하다고 할 수 있다.
6. 계산된 $r_{\text{PRR.U}}$와 $r_{\text{PFRR.U}}$을 바탕으로 다음과 같이 $\boldsymbol{\Gamma}$를 추정한다.
   만약 $r_{\text{PRR.U}} > r_{\text{PFRR.U}}$ 라면, $\hat{\boldsymbol{\Gamma}}_{\text{PRR}}$으로 $\boldsymbol{\Gamma}$를 추정한다.
   만약 $r_{\text{PRR.U}} < r_{\text{PFRR.U}}$ 라면, $\hat{\boldsymbol{\Gamma}}_{\text{PFRR}}$으로 $\boldsymbol{\Gamma}$를 추정한다.
   만약 $r_{\text{PRR.U}} = r_{\text{PFRR.U}}$ 라면, $\hat{\boldsymbol{\Gamma}}_{\text{UPFRR}}$으로 $\boldsymbol{\Gamma}$를 추정한다.

PCR 방법에서 trace correlation coefficient 대신에 vector correlation coefficient (Hotelling, 1936) 를 사용할 수 있다. 모의실험 결과 trace correlation coefficient와 vector correlation coefficient 모두 동일한 결과를 제공하기 때문에, 두 행렬간의 유사성을 계산하는 방법으로 계산이 더 간편한 trace correlation coefficient를 사용하고자 한다.

모의실험에 대한 절에서 상세하게 결과를 제공하지는 않지만, PSRR의 방법 선택 결과는 $\mathbf{f}_{\mathbf{X}}$의 선택과 $\hat{d}$의 추정값에 따라 차이가 날 수 있다. 동일한 $\hat{d}$가 주어졌을 때 $\mathbf{f}_{\mathbf{X}} = \mathbf{X}$와 $\mathbf{f}_{\mathbf{X}}$를 $\mathbf{X}$의 클러스터 지시지행렬로 했을 때, PRR와 PFRR의 선택이 바뀔 수 있다. 하지만, $\mathbf{X}$, $\mathbf{X}^2$, exp($\mathbf{X}$) 혹은 이들의 조합에서는 방법 선택의 차이가 거의 없다. Yoo (2019a)에 따르면 $\mathbf{f}_{\mathbf{X}}$에 대한 기본선택으로 $\mathbf{X}$를 제시하고 있고, 모의실험 결과 $\mathbf{f}_{\mathbf{X}}$를 $\mathbf{X}$로 잘못 선택하는 경우에도 PFRR과 UPFRR 모두 $\boldsymbol{\Gamma}$를 잘 추정하기 때문에 이는 크게 문제가 되지 않는다.

동일한 $\mathbf{f}_{\mathbf{X}}$가 주어졌을 때, $\hat{d}$의 추정된 값에 따라 PSRR의 선택이 바뀔 수 있지만, 모의실험 결과 선택의 차이는 매우 미비하다. 그리고 $\boldsymbol{\Gamma}$의 실제 차원의 추정은 Yoo (2019a)에 따르면 UPFRR에 의한 추정이 가장 안정적이기 때문에 $\hat{d}$는 PSRR에서 크게 문제되지 않는다.

4. 모의 실험

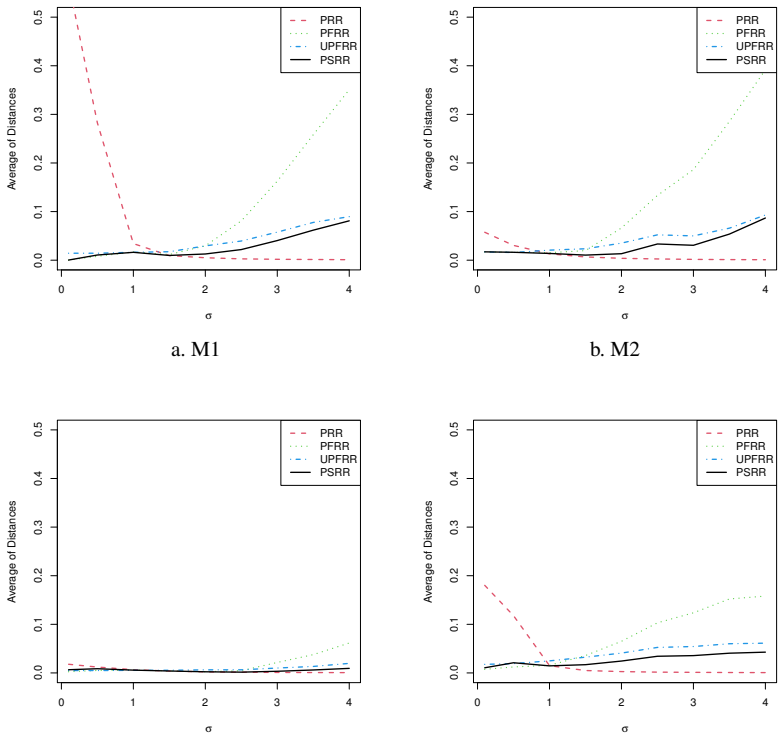
Yoo (2018)과 Yoo (2019a)에서 공통적으로 다루어진 5차원의 설명변수가 주어진 4차원의 반응변수의 아래의 다변량 회귀 모형 $\mathbf{Y} = (Y_1, \ldots, Y_4)|\mathbf{X} = (X_1, \ldots, X_5)$ 을 고려하고자 한다

$$\mathbf{Y} = \boldsymbol{\Gamma}\boldsymbol{\nu}_x + \boldsymbol{\varepsilon}.$$

위의 모형에서 설명변수와 임의오차는 다음의 분포에서 생성되었다.

$$\mathbf{X} \overset{\text{iid}}{\sim} N\left(\mathbf{0}, \sigma_x^2\mathbf{I}_5\right) \perp\!\!\!\perp \boldsymbol{\varepsilon} \overset{\text{iid}}{\sim} N\left(\mathbf{0}, \sigma^2\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\text{T}} + \sigma_0^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^{\text{T}}\right).$$

위에서 " $\perp\!\!\!\perp$ "는 통계적 독립을 의미한다. 그리고 σ_x, σ 그리고 σ_0의 값들은 각각 $\mathbf{X}$와 $\boldsymbol{\Gamma}$와 $\boldsymbol{\Gamma}_0$에 의한 $\boldsymbol{\varepsilon}$의 변동성을 측정한다.



a. M1 b. M2

c. M3 d. M4

Figure 1: *Estimation results of* $\mathbf{\Gamma}$ *varying* σ*; PRR, principal response reduction; PFRR, principal fitted response reduction; UPFRR, unstructured principal response reduction; PSRR, principal selected response reduction.*

$\mathbf{\Gamma}$에 대해서는 다음의 두 가지 중 하나를 모형에 따라 고려하였다.

$$\mathbf{\Gamma}_1 = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)^{\mathrm{T}} \in \mathbb{R}^{4\times 1} \; ; \; \mathbf{\Gamma}_2 = \left\{\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0\right)(0,0,1,0)\right\}^{\mathrm{T}} \in \mathbb{R}^{4\times 2}.$$

위에서 기술된 회귀 모형과 변수 설정을 기반으로 다음의 네 개의 모형을 고려하였다.

M1 $\mathbf{\Gamma} = \mathbf{\Gamma}_1$ and $\boldsymbol{\nu}_x = X_1$.

Under $\boldsymbol{\psi} = (1,0,0,0,0)$ and $\mathbf{f}_{\mathbf{X}} = (X_1, 0, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\nu}_{\mathbf{X}} = \boldsymbol{\psi}\mathbf{f}_{\mathbf{X}}$.

$E(\mathbf{Y}|\mathbf{X}) = (1/2)(X_1, X_1, X_1, X_1)^{\mathrm{T}}$.

M2 $\mathbf{\Gamma} = \mathbf{\Gamma}_1$ and $\boldsymbol{\nu}_x = X_1 + X_1X_2$.

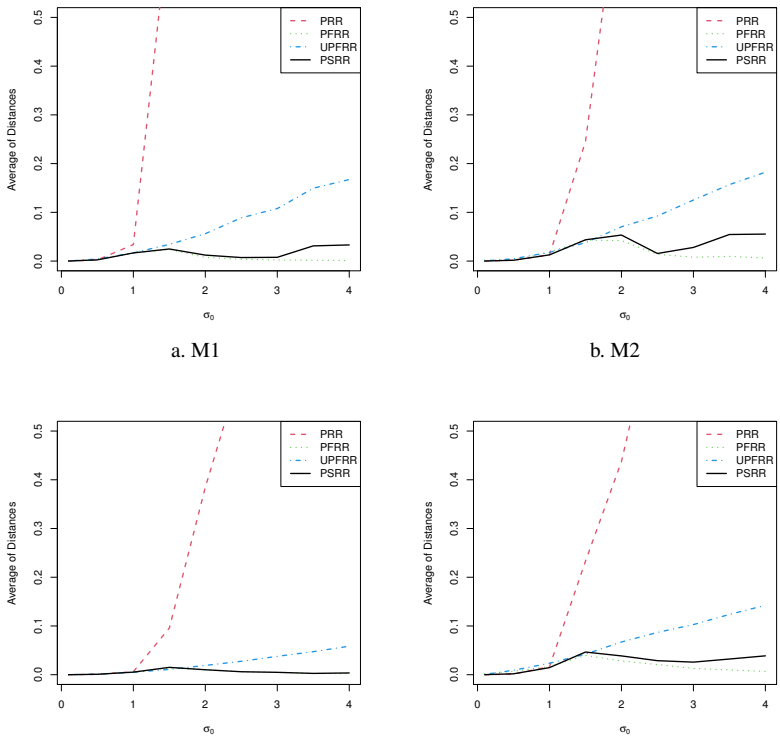
Under $\boldsymbol{\psi} = (1,1,0,0,0)$ and $\mathbf{f}_{\mathbf{X}} = (X_1, X_1X_2, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\nu}_{\mathbf{X}} = \boldsymbol{\psi}\mathbf{f}_{\mathbf{X}}$.

$E(\mathbf{Y}|\mathbf{X}) = (1/2)(X_1 + X_1X_2, X_1 + X_1X_2, X_1 + X_1X_2, X_1 + X_1X_2)^{\mathrm{T}}$.

M3 $\mathbf{\Gamma} = \mathbf{\Gamma}_2$ and $\boldsymbol{\nu}_x = (X_1 + X_2, \exp(X_3))^{\mathrm{T}}$.

Under $\boldsymbol{\psi} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$ and $\mathbf{f}_{\mathbf{X}} = (X_1, X_2, \exp(X_3), 0, 0)^{\mathrm{T}}$, $\boldsymbol{\nu}_{\mathbf{X}} = \boldsymbol{\psi}\mathbf{f}_{\mathbf{X}}$.

$E(\mathbf{Y}|\mathbf{X}) = ((1/\sqrt{2})(X_1 + X_2), (1/\sqrt{2})(X_1 + X_2), \exp(X_3), 0)^{\mathrm{T}}$.



a. M1 b. M2

c. M3 d. M4

Figure 2: *Estimation results of* $\boldsymbol{\Gamma}$ *varying* σ_0*; PRR, principal response reduction; PFRR, principal fitted response reduction; UPFRR, unstructured principal response reduction; PSRR, principal selected response reduction.*

M4 $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_2$ and $\boldsymbol{\nu}_x = (X_1 + X_2, X_1 + X_1X_2)^{\mathrm{T}}$.

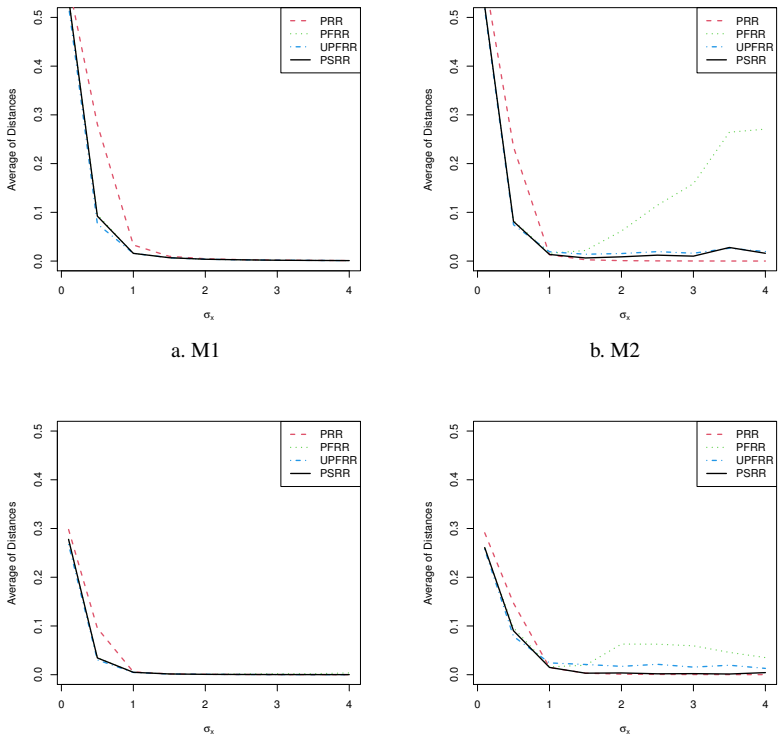
Under $\boldsymbol{\psi} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}$ and $\mathbf{f}_{\mathbf{X}} = (X_1, X_2, X_1X_2, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\nu}_{\mathbf{X}} = \boldsymbol{\psi}\mathbf{f}_{\mathbf{X}}$.

$E(\mathbf{Y}|\mathbf{X}) = ((1/\sqrt{2})(X_1 + X_2), (1/\sqrt{2})(X_1 + X_2), X_1 + X_1X_2, 0)^{\mathrm{T}}$.

M1과 M2는 $\boldsymbol{\Gamma}$의 차원은 1인 반면, M3와 M4에서는 $\boldsymbol{\Gamma}$의 차원은 2이다. M1을 제외한 M2–M4 모두 $E(\mathbf{Y}|\mathbf{X})$에 설명변수간에 비선형 관계가 존재한다.

M1–M4을 100개의 표본수에 대해 500번의 자료를 생성하였고, 본문에서 기술된 네 가지의 모형기반 반응변수 차원축소 방법인 PRR, PFRR, UPFRR과 PCRR이 적용되었다. 그리고 σ, σ_0와 σ_x의 값이 바뀜에 따라 $\boldsymbol{\Gamma}$의 추정이 어떻게 바뀌는지 살펴보았다. 이들에 대한 값으로 $(0.1, 0.5, 1, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0)$이 고려되었다. 또한 σ, σ_0와 σ_x^2의 모든 조합을 고려하지 않고, 한 번에 이중 하나만을 변화 시켰다. 그리고 나머지는 모두 1로 고정을 하였다. 예를 들어 σ_x의 값을 변화시킬 때, σ과 σ_0는 모두 1의 값에 고정이 된다. 그리고, M1–M4 모형들에 대해 $\mathbf{f}_{\mathbf{X}}$로 $\mathbf{X}$을 사용하였다. 이는 M1을 제외한 나머지 모형에서 실제 $\mathbf{f}_{\mathbf{X}}$를 잘못 선택한 것이다.

그리고, 실제 $\boldsymbol{\Gamma}$와 각각의 방법론의 추정값이 얼마나 근접한지를 살펴보기 위해 실제 $\boldsymbol{\Gamma}$와 각각의 추정값 간에 PSRR 알고리듬에서 언급된 trace correlation coefficient 1에서 뺀 값의 평균을 제시한다. Trace correlation coefficient에서 1을 뺀 이유는 상관성을 거리개념으로 전환하기 위해서이다. 1에서 trace correlation coefficient



a. M1 b. M2

c. M3 d. M4

Figure 3: *Estimation results of* $\boldsymbol{\Gamma}$ *varying* σ_x*; PRR, principal response reduction; PFRR, principal fitted response reduction; UPFRR, unstructured principal response reduction; PSRR, principal selected response reduction.*

에서 1을 뺀 값이 작을 수록 $\boldsymbol{\Gamma}$를 잘 추정하는 것이다.

위의 모의실험 결과는 Figures 1–3에 정리되어 있다.

Figures 1–3를 살펴보면 σ, σ_0, σ_x의 값의 변화에 따른 M1–M4에서의 $\boldsymbol{\Gamma}$의 추정 결과가 매우 비슷한 형태임을 확인할 수 있다. 즉 σ가 커짐에 따라 PFRR만 매우 급격하게 $\boldsymbol{\Gamma}$의 추정이 나빠지고, 반대로 σ^2가 커지면 PRR만 $\boldsymbol{\Gamma}$를 매우 부정확하게 추정한다. 그리고 σ_x이 커지면서 PFRR이 PRR보다 다소 부정확하게 $\boldsymbol{\Gamma}$를 추정함을 확인할 수 있다. 또한 UPFRR은 모든 경우에서 PRR와 PFRR 중 더 좋게 $\boldsymbol{\Gamma}$를 추정하는 방법과 유사하게 변화함을 확인할 수 있다. 이러한 상황에서 PSRR은 Yoo (2019a)가 기본 방법으로 제안하는 UPFFRR보다 더 정확하게 $\boldsymbol{\Gamma}$를 추정하고, 가장 잘 추정하는 방법과 매우 비슷하게 $\boldsymbol{\Gamma}$를 추정함을 확인할 수 있다.

그리고, Table 1에 각각의 모형에 대해 σ, σ_0와 σ_x가 변화할 때 PRR의 선택 퍼센트가 정리되어 있다. Table 1에 따른 PSRR은 모든 경우에 있어 더 바람직한 방법을 잘 선택하는 것을 확인할 수 있다. 이러한 다양한 모의실험 결과를 통해 PSRR이 UPFRR보다 실제적으로 또 경험적으로 유용함을 확인할 수 있다.

5. 결론

본 논문은 반응변수가 다차원인 다변량 회귀분석에서 Yoo (2018)과 Yoo (2019a)에서 세 가지 방법론이 제시되고 있다. 이 세 가지 방법론은 주 반응변수 차원축소, 주적합 반응변수 차원축소 그리고 비구조적 주적합

Table 1: Selection percentages of principal response reduction (PRR)

varying σ	0.1	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
M1	0.2	3.6	0.2	75.0	82.2	77.8	77.4	77.4	77.8
M2	2.0	9.6	55.8	77.0	78.8	77.0	79.6	74.6	77.2
M3	38.0	63.4	32.6	44.6	54.2	64.2	68.6	71.2	71.8
M4	6.8	26.4	11.8	27.2	46.8	58.8	67.6	68.8	71.8
varying σ_0	0.1	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
M1	82.8	81.0	0.2	0.2	0.6	0.4	0.6	3.2	3.4
M2	83.2	76.4	51.2	10.0	2.0	0.2	2.2	4.8	5.2
M3	67.8	61.4	30.2	10.8	0.0	0.0	0.0	0.0	0.2
M4	29.8	25.6	13.8	8.6	3.4	1.6	1.6	2.6	3.6
varying σ_x	0.1	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
M1	11.6	4.2	0.6	0.2	0.4	0.0	0.2	0.0	0.0
M2	13.8	4.8	57.2	77.2	80.8	77.6	82.0	80.0	82.0
M3	08.4	2.0	35.2	62.6	65.8	66.4	69.6	70.4	68.8
M4	7.0	9.2	11.2	50.8	72.2	75.8	72.0	74.0	71.4

반응변수 차원축소이다. 특히 Yoo (2019a)에서는 이 세 가지 방법 중에서 비구조적 주적합 반응변수 차원축소를 기본 방법으로 사용하라고 권장하고 있다. 이유는 다양한 모의실험에서 가장 강건한 추정을 하고, 차원축소에 대한 안정적 검정을 제공하기 때문이다. 하지만 모의실험 결과 비구조적 주적합 반응변수 차원축소는 세 가지 모형기반 반응변수 차원축소 방법론 중에서 가장 정확하지는 않다. 다만 가장 정확한 방법론과 매우 비슷하게 움직임을 갖는다. 이를 바탕으로 비구조적 주적합 반응변수 차원축소를 기반으로 주 반응변수 차원축소 주적합 반응변수 차원축소 중 하나를 선택한다면 Yoo (2019a)에서 기본방법으로 제시한 비구조적 주적합 반응변수 차원축소보다는 더 정확한 차원축소를 기대할 수 있을 것이다.

이에 따라 제시된 세 가지의 모형기반 반응변수 방법에서 비구조적 주적합 반응변수 차원축소을 기반으로 하여 주 반응변수 차원축소와 주적합 반응변수 차원축소을 선택하는 알고리듬을 본 논문에서 제시하였다. 그리고 이를 주선택 반응변수 차원축소라고 명명하였다. 다양한 모의실험을 통하여 주선택 반응변수 차원축소의 유용성을 입증하였다.

References

Cook RD (1998). Principal Hessian directions revisited, *Journal of the American Statistical Association*, **93**, 84–10.

Cook RD (2007). Fisher lecture: dimension reduction in regression, *Statistical Science*, **22**, 1–26.

Cook RD and Li B (2002). Dimension reduction for the conditional mean, *Annals of Statististics*, **30** 455–474.

Hooper JW (1959). Simultaneous equations and canonical correlation theory, *Econometrica*, **27**, 245–256.

Hotelling H (1936). Relations between two sets of variates, *Biometrika*, **28**, 321—377.

Yoo JK (2018). Response dimension reduction: model-based approach, *Statistics : A Journal of Theoretical and Applied Statistics*, **52**, 409–425.

Yoo JK (2019a). Unstructured principal fitted response reduction in multivariate regression, *Journal of the Korean Statistical Society*, **48**, 561–567.

Yoo JK (2019b). Note on response dimension reduction for multivariate regression, *Communications for Statis-*

tical Applications and Methods, **26**, 519–526.

Yoo JK and Cook RD (2008). Response dimension reduction for the conditional mean in multivariate regression, *Computational Statistics and Data Analysis*, **53**, 334–343.

Yoo JK, Patterson BS, and Datta S (2009). An OLS-based predictor test for a single-index model for predicting transcription rate from histone acetylation level, *Statistics and Probability Letters*, **79**, 2109–2114.

Received April 18, 2021; Revised May 16, 2021; Accepted May 16, 2021

다변량회귀에서 주선택 반응변수 차원축소

유재근[1,a]

[a]이화여자대학교 통계학과

요 약

다변량 회귀분석은 경시적 자료분석이나 함수적 자료분석 등 다양한 분야에서 빈번하게 사용되는 통계적 방법론이다. 다변량 회귀분석은 설명변수의 차원 뿐만 아니라 반응변수의 차원때문에 일변량 회귀분석에서보다 차원의 저주문제에 더 강한 영향을 받는다. 이러한 문제를 해결하기 위해 최근 Yoo (2018)와 Yoo (2019a)에 세 가지 모형기반 반응변수 차원축소 방법이 제시되었다. 하지만 Yoo (2019a)에서 제시한 기본 방법은 모의실험 결과 모형에 가장 영향을 덜 받지만, 다른 두 방법 중 더 나은 방법보다 더 좋은 추정결과를 제시하지 못한다. 이러한 단점을 극복하기 위해 본 논문에서는 기본 방법의 결과 다른 두 방법의 결과를 비교하여, 자료에 따라 최선의 방법을 제시하는 선택 알고리듬을 제시하고, 이를 주선택 반응변수 차원축소라 명명한다. 다양한 모의실험 결과 주선택 반응변수 차원축소는 Yoo (2019a)의 기본방법보다 더 정확하게 차원을 축소하고, 모든 경우에 있더 더 바람직한 방법을 선택함을 확인할 수 있다. 이러한 결과로 제안한 주선택 반응변수의 차원축소 방법의 실제적 유용성을 확인할 수 있다.

주요용어: 다변량 회귀분석, 모형기반 차원축소, 주 반응변수 차원축소, 주적합 반응변수 차원축소, 비구조 주적합 반응변수 차원축소

2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2021R1F1A1059844).

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