The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.6.835
31(6), 835–842

Case study: application of fused sliced average variance estimation to near-infrared spectroscopy of biscuit dough data

Hye Yeon Um[a] · Sungmin Won[a] · Hyoin An[a] · Jae Keun Yoo[a,1]

[a]Department of Statistics, Ewha Womans University

(Received November 22, 2018; Revised November 23, 2018; Accepted November 24, 2018)

Abstract

The so-called sliced average variance estimation (SAVE) is a popular methodology in sufficient dimension reduction literature. SAVE is sensitive to the number of slices in practice. To overcome this, a fused SAVE (FSAVE) is recently proposed by combining the kernel matrices obtained from various numbers of slices. In the paper, we consider practical applications of FSAVE to large p-small n data. For this, near-infrared spectroscopy of biscuit dough data is analyzed. In this case study, the usefulness of FSAVE in high-dimensional data analysis is confirmed by showing that the result by FASVE is superior to existing analysis results.

Keywords: fused approach, inverse regression, large p-small n data, sliced average variance estimation, sufficient dimension reduction

1. 서론

회귀분석에서 충분 차원 축소(sufficient dimension reduction; SDR)는 설명변수가 주어졌을 때 반응변수의 조건부 분포 $Y|\mathbf{X}$에 대해 정보의 손실없이 원 설명변수 $\mathbf{X}$를 저차원의 선형변환 설명변수 $\mathbf{M}^T\mathbf{X}$로 대체하는 것을 목적으로 한다. 이 때 $\mathbf{M}$은 행의 수가 p이고 열의 수가 q인 행렬이고, 일반적으로 q는 p보다 같거나 작다고 가정한다. 이를 조건부 독립 형태로 표현하면 다음과 같다.

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{M}^T\mathbf{X},$$

여기서 $\perp\!\!\!\perp$는 통계적 독립을 의미하고, 위의 식을 만족하는 $\mathbf{M}$의 열공간을 차원 축소 공간이라 부른다.

하나의 회귀문제에 대해 다수의 차원 축소 공간이 존재할 수 있고, 그렇다면 자연히 이 중 가장 작은 차원의 공간을 선택할 것이다. 이 최소 차원의 공간을 중심 부분 공간 $\mathcal{S}_{Y|\mathbf{X}}$이라고 부르고, 충분 차원 축소의 주요 문제는 이 중심 부분 공간의 추정에 있다. 이 중심 부분 공간의 추정은 이 공간을 생성하는 정규 직교 기저(orthonormal basis)와 차원의 추정 두 가지로 나뉘게 된다.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1A2B1004909).

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

충분 차원 축소에서 가장 많이 쓰이는 대표적인 방법론으로 sliced inverse regression (SIR) (Li, 1991)과 sliced average variance estimation (SAVE) (Cook와 Weisberg, 1991)가 있다. SIR와 SAVE은 그 이름에서 암시하듯이, 회귀분석에서 일반적으로 사용되는 순회귀 $Y|\mathbf{X}$보다는 역회귀 $\mathbf{X}|Y$를 이용한다. 이 두 방법론의 가장 큰 차이는 SIR의 경우는 $\mathbf{X}|Y$의 1차 적률 $E(\mathbf{X}|Y)$을 이용하는 반면, SAVE는 2차 적률 $\text{cov}(\mathbf{X}|Y)$을 사용하는 데 있다. 두 방법론의 공통점은 슬라이싱(slicing)이라고 불리우는 반응변수 Y의 범주화 과정이 요구된다는 점이다. 역회귀의 $\mathbf{X}|Y$의 분포를 정확히 모르는 상황에서 중심 부분 공간을 비모수적으로 추정하기 위해 필요한 매우 중요한 과정이다. 이 과정을 통해 범주화된 Y의 수준에서 $\mathbf{X}$의 표본 평균과 분산을 통하여 중심 부분 공간을 추정하는 것이 SIR와 SAVE의 방법론적 핵심이다. 하지만 SIR와 SAVE로 부터의 결과는 슬라이스의 총 수에 따라 민감하게 달라져서, 부적절한 슬라이스 수의 사용은 잘못된 결과를 유도할 수 있음은 이미 잘 알려져 있다.

그러나 이 범주화 과정에서 최적의 범주 개수에 대한 기준을 정하기가 매우 어렵기 때문에 슬라이스의 개수를 정하는 문제는 SIR와 SAVE의 취약점이라 볼 수 있다. 이 문제는 1차 적률을 사용하는 SIR에 비해 2차 적률을 사용하는 SAVE에서 더 심각하게 나타난다.

이를 해결하기 위하여 Cook과 Zhang (2014)은 SIR에서 다양한 슬라이스의 개수를 통해 얻어진 $E(\mathbf{X}|Y)$의 추정값을 모두 결합하는 접근법을 제안하였다. 그들은 SIR에 대해 슬라이스 개수를 3에서 15까지 변화시키면서 결합시키는 접근이 슬라이스의 개수에 강건하면서 기존의 SIR보다 더 정확하게 중심 부분 공간을 추정함을 모의실험을 통해 보였다. 그리고 최근 Yoo와 Cho (2018)에서는 Cook과 Zhang (2014)가 제안한 접근법이 중심 부분 공간의 차원을 강건하게 추정함을 보여주고 있다.

An 등 (2017)에서는 Cook과 Zhang (2014)처럼 다양한 개수의 슬라이스로 부터 얻어진 결과를 결합하는 접근법을 SAVE에 적용한 fused SAVE (FSAVE) 방법론을 제시하였다. 이러한 결합 접근법을 통해 슬라이스의 수에 민감한 SAVE의 문제점을 완화하였고, 또한 중심 부분 공간의 차원 역시 보다 강건하게 추정함을 보였다.

SIR에 비해 SAVE는 지금까지 소위 large p-small n 자료에 대해 실제 적용되어 분석된 사례가 없다. 여기서 large p-small n 자료란 변수의 수가 관측수 보다 더 많은 자료를 의미한다. 이에 대한 이유로는 SAVE가 2차 적률을 사용해야 한다는 점과 슬라이스의 수에 SIR보다 민감하게 반응한다는 점을 생각할 수 있다. 본 논문에서는 최근 개발된 FASVE를 large p-small n 자료에 적용하여 분석하는 실증사례를 살펴보고자 한다. FSAVE의 실제적 구현에는 설명변수의 공분산 행렬이 추정되어야 하기 때문에, large p-small n 자료로의 직접적인 적용은 어렵다. 이를 해결하기 위해 우선 설명변수를 주성분 분석을 통해 일차적인 차원 축소를 하고, 이 축소된 자료에 대해 FSAVE를 적용하여 추가적인 보다 정밀한 차원 축소를 실시할 것이다. 두 단계에 걸쳐 축소된 설명변수를 이용하여, 회귀식을 적합하여 기존의 방법론의 결과와 비교를 할 것이다.

본 논문의 순서는 다음과 같다. 2장에서는 SAVE 이론에 대한 내용과 전제 조건을 설명함과 동시에 FASVE을 소개할 것이다. 3장에서는 FSAVE를 비스킷 반죽의 근적외분광분석법 분석 자료에 적용하고 자료를 분석하고자 한다. 마지막 4장에서는 결론과 본 연구의 시사점을 기술한다. 지금부터 $\mathcal{S}_{Y|\mathbf{X}}$의 정규 직교 기저와 그 차원을 $\boldsymbol{\eta}$와 d로써 나타낼 것이다.

2. Fused sliced variance estimation의 소개

2.1. Sliced average variance estimation

SAVE를 설명하기 위해, 원 설명 변수 $\mathbf{X}$는 다음과 같이 표준화 한다.

$$\mathbf{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - E(\mathbf{X})),$$

여기서 $\mathbf{\Sigma}$은 설명변수 $\mathbf{X}$의 공분산 행렬 $\mathrm{cov}(\mathbf{X})$를 의미한다.

다음의 회귀 $Y|\mathbf{Z}$의 중심 부분 공간과 그 정규 직교 기저를 각각 $\mathcal{S}_{Y|\mathbf{Z}}$와 $\boldsymbol{\eta}_z$로 정의하자. 그렇다면 $\mathcal{S}_{Y|\mathbf{X}}$와 $\mathcal{S}_{Y|\mathbf{Z}}$ 그리고 $\boldsymbol{\eta}$와 $\boldsymbol{\eta}_z$에 대해 다음과 같은 관계가 성립한다.

$$\mathcal{S}_{Y|\mathbf{X}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}_{Y|\mathbf{Z}} \;\Leftrightarrow\; \boldsymbol{\eta} = \mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{\eta}_z. \tag{2.1}$$

위의 관계는 Cook (1998)의 Proposition 6.3에 입증되어 있다.

SAVE가 중심 부분 공간을 추정하기 위해서는 다음의 두 조건의 만족이 필요하다.

A1. $E(\mathbf{Z}|\boldsymbol{\eta}_z^T\mathbf{Z})$는 $\boldsymbol{\eta}_z^T\mathbf{Z}$에 대해 선형이다.

A2. $\mathrm{cov}(\mathbf{Z}|\boldsymbol{\eta}_z^T\mathbf{Z})$는 $\boldsymbol{\eta}_z^T\mathbf{Z}$ 일정하다.

위의 두 조건 A1과 A2는 선형성과 등분산성 조건이라고 부른다. 조건부 2차 적률 $\mathrm{cov}(\mathbf{Z}|\boldsymbol{\eta}_z^T\mathbf{Z})$은 주어진 $\boldsymbol{\eta}_z^T\mathbf{Z}$의 함수인 것은 명확하다. 하지만 조건 A2가 만족된다면 $\mathrm{cov}(\mathbf{Z}|\boldsymbol{\eta}_z^T\mathbf{Z})$에 주어진 $\boldsymbol{\eta}_z^T\mathbf{Z}$에 의존하지 않고, $\boldsymbol{\eta}_z^T\mathbf{Z}$에 상관없이 일정한 값을 갖는다. 만약 설명변수 $\mathbf{Z}$가 타원형의 분포(elliptical distribution)를 따른다면 조건 A1은 만족하는 반면 조건 A2는 만족하지 않을 수 있다. 만약 $\mathbf{Z}$가 다변량 정규분포를 따른다면 두 조건 A1과 A2 모두 충족된다. 두 조건이 만족되지 않을 경우 일반적으로 $\mathbf{X}$에 대해 정규성 만족을 위한 일대일 변환을 실시한다.

두 조건 A1과 A2가 만족된다는 가정 하 $\mathrm{cov}(\mathbf{Z}|Y)$에 대해 다음 관계가 성립한다.

$$\mathcal{S}(I_p - \mathrm{cov}(\mathbf{Z}|Y)) \subseteq \mathcal{S}_{Y|\mathbf{Z}}, \tag{2.2}$$

여기서 행의 수가 p이고 열의 수가 q인 행렬 $\mathbf{M}$에 대해 $\mathcal{S}(\mathbf{M})$은 $\mathbf{M}$의 열공간을 의미한다.

식 (2.1)의 관계에 따라서 다음이 성립한다.

$$\mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}(\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y)) \subseteq \mathcal{S}_{Y|\mathbf{X}}. \tag{2.3}$$

수식 (2.3)에서 $\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y)$의 열공간은 다음의 관계를 갖는다.

$$\mathcal{S}(\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y)) = \mathcal{S}\left(E(\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y))^2\right). \tag{2.4}$$

식 (2.2)–(2.4)를 통하여 다음의 관계가 최종적으로 성립한다.

$$\mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}\left(E(\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y))^2\right) \subseteq \mathcal{S}_{Y|\mathbf{X}}. \tag{2.5}$$

위의 수식 (2.5)에서 $\mathbf{M}_{\mathrm{SAVE}} = E(\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y))^2$라고 정의하자. $\mathbf{M}_{\mathrm{SAVE}}$를 이용하여 $\mathcal{S}_{Y|\mathbf{X}}$를 추정하는 방법이 SAVE이다. Ye와 Weiss (2003, 2.2절)는 SAVE가 $E(\mathbf{Z}|Y)$를 사용하는 SIR가 생성하는 공간을 포함함을 증명하였고, 따라서 SAVE가 SIR보다 더 포괄적으로 $\mathcal{S}_{Y|\mathbf{X}}$를 추정한다고 논의하였다.

방법론적으로 SAVE는 $\mathbf{X}|Y$에 대한 분포적 가정을 하지 않기 때문에, $\mathbf{M}_{\mathrm{SAVE}}$에서 $\mathrm{cov}(\mathbf{Z}|Y)$의 모수적 추정은 매우 어렵다. 하지만 만약 Y가 범주형 변수라면 $\mathrm{cov}(\mathbf{Z}|Y=y)$ 추정은 가능하다. 왜냐하면 $Y=y$의 범주내에서 $\mathbf{Z}$의 공분산 행렬이 $\mathrm{cov}(\mathbf{Z}|Y=y)$이 되기 때문이다. 이를 이용하면 Y가 연속형일 경우 Y를 범주화 하면 $\mathrm{cov}(\mathbf{Z}|Y)$에 대한 추정량을 구할 수 있다. 이 반응변수의 범주화 과정을 슬라이싱이라고 부른다. 우선 Y를 전체 슬라이스 개수 h만큼 범주화한 $\tilde{Y}$을 얻은 후에 각 슬라이스 내에서 공분산 행렬을 다음과 같이 구하면 될 것이다.

$$\widehat{\mathrm{cov}}\left(\mathbf{Z}|\tilde{Y}=s\right) = \frac{1}{n_s}\sum_{\tilde{Y}_i=s}\left(\hat{\mathbf{Z}}_{i\in s} - \bar{\mathbf{Z}}_s\right)\left(\hat{\mathbf{Z}}_{i\in s} - \bar{\mathbf{Z}}_s\right)^T,$$

여기서 n_s는 s의 범주 내 표본 크기이다. 또한 $\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$이고, $\bar{\mathbf{Z}}_s = (1/n_s)\sum_{\tilde{Y}=s} \hat{\mathbf{Z}}_i$이다. 이를 통해 $\mathbf{M}_{\text{SAVE}}$의 추정량은 다음과 같이 유도된다.

$$\hat{\mathbf{M}}_{\text{SAVE}} = \sum_{s=1}^{h} \frac{n_s}{n} \left(I_p - \widehat{\text{cov}}\left(\hat{\mathbf{Z}}|\tilde{Y}=s\right)\right)\left(\mathbf{I}_p - \widehat{\text{cov}}\left(\hat{\mathbf{Z}}|\tilde{Y}=s\right)\right). \tag{2.6}$$

이후 $\hat{\mathbf{M}}_{\text{SAVE}}$에 대해 스펙트럼 분해를 시행한 후 0이 아닌 고유값에 대응하는 고유벡터들이 $\boldsymbol{\eta}_z$의 추정량이 된다.

2.2. Fused sliced average variance estimation

이번 절에서는 기존 SAVE를 결합하는 FSAVE 방법을 소개하고자 한다. 이를 위해 먼저 $\mathbf{M}^{(h)}_{\text{SAVE}}$를 h개의 슬라이스를 사용한 SAVE의 커널 행렬이라고 정의하자. 그리고 다음의 행렬을 정의한다.

$$\mathbf{M}^{(h)}_{\text{FSAVE}} = \left(\mathbf{M}^{(2)}_{\text{SAVE}}, \ldots, \mathbf{M}^{(h)}_{\text{SAVE}}\right) \tag{2.7}$$

수식 (2.7)에서 $h = 1$인 경우 영행렬이 도출되므로 고려하지 않으며, $h = 2$인 경우 또한 $\mathbf{M}^{(2)}_{\text{FSAVE}}$와 $\mathbf{M}^{(2)}_{\text{SAVE}}$ 두 행렬이 일치하기 때문에 배제한다. 앞절에서 언급한대로 $h = 2, \ldots, k$에 대해서 $\mathbf{M}^{(h)}_{\text{SAVE}} \subseteq \mathcal{S}_{Y|\mathbf{Z}}$이 성립하기 때문에 다음의 관계를 유도할 수 있다.

$$\mathcal{S}\left(\mathbf{M}^{(h)}_{\text{SAVE}}\right) \subseteq \mathcal{S}\left(\mathbf{M}^{(h)}_{\text{FSAVE}}\right) \subseteq \mathcal{S}_{Y|\mathbf{Z}} \;\Leftrightarrow\; \mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}\left(\mathbf{M}^{(h)}_{\text{FSAVE}}\right) \subseteq \mathcal{S}_{Y|\mathbf{X}}. \tag{2.8}$$

위의 식 (2.8)를 통해 $\mathbf{M}^{(h)}_{\text{FSAVE}}$는 $\mathcal{S}_{Y|\mathbf{Z}}$를 추정할 수 있는 커널 행렬로 사용이 가능하며, $\mathbf{M}^{(h)}_{\text{FSAVE}}$이 기존의 SAVE보다는 많은 정보를 갖고 있음을 알 수 있다. FSAVE은 $\mathbf{M}^{(h)}_{\text{FSAVE}}$를 이용하여 $\mathcal{S}_{Y|\mathbf{X}}$를 추정하는 방법이다. 지금부터 $\mathcal{S}_{Y|\mathbf{X}}$을 완전히 추정하기 위해 $\mathbf{\Sigma}^{-1/2}\mathcal{S}(\mathbf{M}^{(h)}_{\text{FSAVE}}) = \mathcal{S}_{Y|\mathbf{X}}$를 가정한다.

FSAVE와 SAVE 모두 2차 적률을 추정해야 하기 때문에 많은 슬라이스를 사용할 경우 자연히 $\text{cov}(\mathbf{Z}|\tilde{Y})$의 추정이 부정확해진다. An 등 (2017)에 따르면 다소 큰 슬라이스의 수에 $\mathbf{M}^{(h)}_{\text{FSAVE}}$의 추정 결과에 다소 차이가 있지만, 기존의 SAVE에 비해 매우 강건하고 보다 정확하게 $\mathcal{S}_{Y|\mathbf{X}}$를 추정함을 보였다.

$\mathbf{M}^{(h)}_{\text{FSAVE}}$의 추정은 기존의 SAVE와 동일하다. 이를 통한 $\mathcal{S}_{Y|\mathbf{Z}}$추정은 다음의 스펙트럼 분해를 통해 추정된다.

$$\hat{\mathbf{K}}^{(h)}_{\text{FSAVE}} = \hat{\mathbf{M}}^{(h)}_{\text{FSAVE}}\hat{\mathbf{M}}^{(h)T}_{\text{FSAVE}} \tag{2.9}$$

$\mathcal{S}_{Y|\mathbf{Z}}$의 정규 직교 기저는 $\hat{\mathbf{K}}^{(h)}_{\text{FSAVE}}$에서 0이 아닌 고유값들에 대응하는 고유 벡터이며, 0이 아닌 고유값의 개수는 $\mathcal{S}_{Y|\mathbf{Z}}$의 차원 추정량인 $\hat{d}$에 해당한다. $\mathcal{S}_{Y|\mathbf{Z}}$의 차원의 추정은 다음의 연속적 가설 검정 (Rao, 1965)에 의해 결정된다. 연속적 가설검정은 $m = 0$일 때 두 가설 $H_0 : d = m, H_1 : d > m$을 검정하는 것으로부터 시작한다. 귀무가설이 기각되면 m값을 1씩 증가시키고, H_0가 최초로 기각되지 않을 때까지 검정을 반복하게 되며 이 때 귀무가설의 m를 d의 추정치로 한다. 이 검정을 위해 다음의 검정 통계량이 사용된다.

$$\hat{\Lambda}_m = n\sum_{i=m+1}^{p} \hat{\lambda}_i.$$

위 수식에서 $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$은 $\hat{\mathbf{K}}^{(h)}_{\text{FSAVE}}$에서 얻어진 내림차순의 고유값이며 n은 표본의 크기를 의미한다.

$\hat{\Lambda}_m$에 대한 p-value를 구하기 위해서, An 등 (2017)은 다음 순환 검정(permutation test)을 이용하였다.

1. $\hat{\mathbf{K}}_{\text{FSAVE}}^{(h)}$에 대해 귀무가설 $H_0 : d = m$ 하에서 $\hat{\Lambda}_m$과 다음의 분할된 고유벡터 행렬을 구성한다.

$$\hat{\mathbf{\Gamma}}_1 = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_m) \quad \text{and} \quad \hat{\mathbf{\Gamma}}_2 = (\hat{\boldsymbol{\gamma}}_{m+1}, \ldots, \hat{\boldsymbol{\gamma}}_p).$$

   여기서 $\hat{\boldsymbol{\gamma}}_i$는 고유값 $\hat{\lambda}_i$에 상응하는 고유벡터이다.

2. 두 벡터 $\hat{\mathbf{V}}_i = \hat{\mathbf{\Gamma}}_1^T \hat{\mathbf{Z}}_i \in \mathbb{R}^{m\times 1}$와 $\hat{\mathbf{U}}_i = \hat{\mathbf{\Gamma}}_2^T \hat{\mathbf{Z}}_i \in \mathbb{R}^{(p-m)\times 1}$를 구한다.
3. $\hat{\mathbf{U}}_i$에 한해서 i를 임의로 교환하면서 순환 집합(permuted set) $\hat{\mathbf{U}}_i^*$을 얻어낸다.
4. $(\hat{\mathbf{V}}_i \hat{\mathbf{U}}_i^*)$를 설명변수로 하여 FASVE를 적용한 후 귀무가설 $H_0 : d = m$ 하에서 검정통계량 $\hat{\Lambda}_m^*$을 계산한다.
5. (3)과 (4)의 과정을 N(총 순환 횟수)차례 반복하여 $\hat{\Lambda}_m$보다 큰 $\hat{\Lambda}_m^*$의 비율을 계산하고, 이 비율이 가설검정을 위한 p-value가 된다.

3. 실증 예제: 비스킷 반죽의 근적외분광분석법 분석 자료

앞 장에서 소개된 FSAVE을 이용하여 large p-small n 자료에 대한 실제 분석을 위해 비스킷 반죽의 근적외분광분석법 분석에 대한 자료 (Brown 등, 2001)를 살펴보고자 한다. 이 데이터는 음식, 음료, 제약품, 석유화학품 등 다양한 물질의 구조를 분석할 때 가장 많이 사용되는 근적외분광분석법을 이용하여 비스킷 반죽의 구성 요소들을 측정한 자료이다. 본 논문에서 사용되는 자료는 통계 패키지 R 프로그램(www.r-project.org)의 ppls 패키지에 포함된 cookie라는 데이터이며 총 72개의 표본이 포함되었다.

cookie 자료는 비스킷 반죽의 성분과 지방, 자당, 밀가루, 물의 네 가지 구조의 측량을 포함하고 있다. 비스킷 반죽의 성분은 1100에서 2498 나노미터(nm)까지 2nm의 간격으로 근적외분광분석법에 의해 측정된 700개의 자료점들로 구성되어 있다. Brown 등 (2001)은 유용한 정보가 거의 포함되지 않았다고 여겨지는 처음 140개와 마지막 49개의 파장을 제거한 후, 2nm에서 4nm으로 단계를 증가시켰다. 그들은 이렇게 구해진 1380nm부터 2400nm에 걸쳐 4nm 단계로 구성된 256개의 점들이 설명 변수($\mathbf{X} \in \mathbb{R}^{256}$)로 사용하였다. 본 논문에서도 Brown 등 (2001)에서 사용된 설명 변수를 사용한다. 실제 자료의 수가 72인점을 고려한다면 설명 변수의 차원은 그 보다 약 3배 가량 더 큼을 알 수 있다. 그리고 반응 변수(Y)로는 밀가루의 비율이 고려되었다.

Brown 등 (2001)의 연구에 따라 72개의 관측값을 차원 축소와 모형개발을 위한 40개의 training 자료와 모형의 적합도를평가하기 위해 나머지 32개의 test 자료로 나누었다. 본 논문에서는 Brown 등 (2001)이 나눈 training 자료와 test 자료를 그대로 이용하였다. Training 자료의 23번째 관측치와 test 자료의 21번째 관측치는 이상치로 판단하여 분석 전에 제거하였다. 또한 training 자료의 $\mathbf{X}$를 $\mathbf{X}_R$, test 자료에 대한 $\mathbf{X}$를 $\mathbf{X}_E$로 각각 정의하였다.

FSAVE는 설명 변수 $\mathbf{X}$를 표준화 하는 과정이 필요하고, 이를 위해서는 설명 변수의 표본 공분산 행렬의 역행렬을 계산하여야 한다. 하지만 본 논문에서 사용하는 training 자료의 경우 $p = 256 > n = 39$이므로 $\mathbf{X}_R$의 공분산 행렬에 대한 역행렬이 존재하지 않는다. 따라서 본 자료에는 FSAVE의 직접적인 적용은 불가능하다.

따라서 FSAVE를 적용하기 전에 $\mathbf{X}_R$의 차원을 $n = 39$보다 작게 되도록 먼저 축소시켜 필요가 있다. 이를 위해 주성분 분석(principal component analysis)을 고려한다. 주성분 분석은 반응 변수와

Table 3.1. Dimension estimation in cookie data

	$H_0 : d = 0$	$H_0 : d = 1$	$H_0 : d = 2$
p-value	0.044	0.027	0.885

설명 변수간의 상호관계를 무시하고, 설명 변수의 주변 관계에 의해 차원을 축소하는 방법이기에 주성분 성분의 개수를 선택함에서 보수적 입장을 취해야 한다. 여기서 보수적 입장이란 높은 누적 기여율(cumulative proportion)에 해당되는 주성분을 추출하는 것을 의미한다. Training 자료의 설명 변수 $\mathbf{X}_R$만을 고려하였을 때, 4개의 주성분을 선택했다. 이 네 개 성분의 누적 기여율은 전체의 99.8%에 해당되며, training 자료의 원 설명변수을 4개의 주성분으로 대체할 수 있음을 의미한다.

$\mathbf{L} \in \mathbb{R}^{256\times 4}$을 적재행렬(loading matrix)이라 하고, FSAVE에 적용하고자 하는 설명 변수를 다음과 같이 정의하자.

$$\mathbf{X}_R^{\mathrm{PC}} = \mathbf{X}_R\mathbf{L} \in \mathbb{R}^{32\times 4}.$$

이제 FSAVE $Y|\mathbf{X}_R^{\mathrm{PC}}$인 회귀에 적용한다. $\mathbf{X}_R^{\mathrm{PC}}$의 자료의 수가 32라는 점을 고려한다면, 4이상의 슬라이스를 선택할 때 슬라이스 당 자료가 평균 10개 미만이 된다. FSAVE가 4×4의 행렬을 추정해야 하는 점을 고려한다면 다소 무리가 있는 슬라이스 개수라고 할 수 있다. 따라서 본 논문에서는 슬라이스 개수가 3 ($h = 3$)인 FSAVE를 적용하였다.

먼저 설명 변수의 차원을 4에서 더 줄일 수 있는지 살펴보기 위해 2.2에서 언급한 차원을 추정하기 위한 연속 귀무가설 검정을 실시하였다. 이 연속검정에서 사용된 총 순환의 횟수는 1,000번이었다. 귀무가설 $H_0 : d = 0$, $H_0 : d = 1$, $H_0 : d = 2$에 대해 계산된 p-value는 Table 3.1에 정리되어 있다. Table 3.1에 따르면, 유의수준 5%에서 d는 2로 추정되어, 추가적인 차원축소가 가능함을 확인할 수 있다.

$\hat{\boldsymbol{\eta}} \in \mathbb{R}^{4\times 2}$를 기저 추정량이라고 하면, 다음의 두 변수 W_1과 W_2를 정의할 수 있다.

$$(W_1, W_2) = \mathbf{X}_R\mathbf{L}\hat{\boldsymbol{\eta}}.$$

이제 두 개의 설명을 가진 $Y|(W_1, W_2)$의 회귀를 이용하여 다음의 선형회귀식을 유도하였다.

$$\hat{E}(Y|W_1, W_2) = 48.95 + 1.48W_1 + 24.62W_2 + 6.29W_2^2. \tag{3.1}$$

이어서 training 자료로 부터 얻어진 차원 축소 결과를 test 자료에 적용하여 다음의 변수 W_1^*와 W_2^*를 구한다.

$$(W_1^*, W_2^*) = \mathbf{X}_T\mathbf{L}\hat{\boldsymbol{\eta}}.$$

식 (3.1)에서 구한 회귀식의 모형 적합도를 설명하기 위해, 두 설명 변수 (W_1^*, W_2^*)를 식 (3.1)에서 구한 회귀식에 대입하여 평균 제곱 오차(mean squared error; MSE)를 계산하였다. 기존의 Brown 등 (2001)의 분석결과를 비교하기 위해 5개의 성분을 이용한 부분 최소 제곱법(partial least squares) 및 5개의 주성분을 사용한 주성분 회귀(principal component regression)의 적합 결과와 비교한다 (Varmuza와 Filzmoser, 2009). Brown 등 (2001)은 두 방법에 대해 cross-validation를 통해 5개의 성분을 선택했다. 이는 모두 Table 3.2에 정리되어 있다.

Table 3.2를 통해 근소한 차이지만 FSAVE의 평균 제곱 오차가 가장 작아 기존의 방법보다 우수함을 확인할 수 있다.

Table 3.2. MSEs from FSAVE, partial least squares and principal component regression

Method	FSAVE	Partial least squares	Principal component regression
MSE	0.363	0.375	0.388

MSE = mean squared error; FSAVE = fused sliced average variance estimation.

4. 결론

SAVE (Cook와 Weisberg, 1991)는 충분 차원 축소 분야에서 중심 부분 공간을 추정하기 위하여 가장 널리 이용되는 방법론 중 하나지만 슬라이스의 개수에 매우 민감한 것으로 알려져 있다. 이는 중심 부분 공간의 차원을 추정할 때 뿐만 아니라 중심 부분 공간의 기저를 추정하는 데에도 영향을 미칠 수 있기 때문에 실제 적용 시 매우 유의해야 한다. 이러한 SAVE의 취약점을 극복하기 위해 최근 SAVE에 다양한 슬라이스 개수를 적용하여 구성되는 커널 행렬들을 결합하는 방법이 제안 되었고, 이를 FSAVE (An 등, 2017)라 정의하였다. FSAVE에서는 중심 부분 공간의 차원을 추정하기 위해 순환 검정을 제안하였다.

본 논문에서는 소위 large p-small n 자료라고 불리우는 자료의 수가 변수의 수보다 적은 자료에서 FASVE가 어떻게 실제적으로 사용될 수 있을지에 대해 실증적 분석을 하였다. 이를 위해 근적외분광분석을 통해 얻어진 비스킷 자료를 이용하였고, 분석 결과를 기존에 이미 분석된 결과와 비교을 하였다.

FSAVE은 설명변수의 공분산 행렬의 역행렬을 계산하는 과정이 요구되기 때문에, 자료의 수가 변수의 수보다 적은 경우 직접적인 적용이 불가하다. 이를 해결하기 위해 우선 주성분 분석으로 1차 차원축소를 한 후 이를 FSAVE에 적용하여 2차 차원 축소를 실시하였다. 2단계 차원축소를 통해 얻은 결과를 이용하여 모형을 적합하였을 때, 기존의 부분 최소 제곱법과 주성분 회귀분석을 통해 얻어진 결과보다 더 좋은 결과를 얻음을 확인할 수 있었다. 이는 FSAVE가 고차원 자료에서도 유용하게 사용될 수 있음을 확인할 수 있는 좋은 계기가 될 것점에 의의를 갖고자 한다.

References

An, H., Won, S., and Yoo, J. K. (2017). Fused sliced average variance estimation, *Journal of the Korean Statistical Society*, **46**, 623–628.

Brown, P. J., Fearn, T., and Vannucci, M. (2001). Bayesian wavelet regression on curves wth application to a spectroscopic calibration problem, *Journal of the American Statistical Association*, **96**, 398–408.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, Wiley, New York.

Cook, R. D. and Weisberg, S. (1991). Discussion of Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 328–332.

Cook, R. D. and Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction, *Journal of the American Statistical Association*, **109**, 815–827.

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

Rao, C. R. (1965). *Linear Statistical Inference and Its Application*, Wiley, New York.

Varmuza, K. and Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press, New York.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods, *Journal of the American Statistical Association*, **98**, 968–979.

Yoo, J. K. and Cho, Y. (2018). On robustness in dimension determination in fused sliced inverse regression, *Communications for Statistical Applications and Methods*, **25**, 513–521.

Fused sliced average variance estimation의 실증분석: 비스킷 반죽의 근적외분광분석법 분석 자료로의 적용

엄혜연[a] · 원성민[a] · 안효인[a] · 유재근[a,1]

[a]이화여자대학교 통계학과

요 약

충분차원축소의 대표적 방법론 중 하나인 sliced average variance estimation (SAVE)은 슬라이스라고 불리우는 반응변수의 범주화의 총 수에 민감하다고 알려져 있다. 이러한 점을 극복하기 위한 방법으로 최근에 다양한 수의 슬라이스로부터 얻어진 SAVE의 정보를 결합하는 fused SAVE (FSAVE)가 개발되었다. 본 논문에서는 소위 large p-small n 자료라고 불리우는 자료의 수가 변수의 수보다 적은 자료에서 FASVE가 어떻게 실제적으로 사용될 수 있을지에 대해 실증적 분석을 하고자 한다. 이를 위해 근적외분광분석을 통해 얻어진 비스킷 자료를 이용할 것이고, 이러한 자료분석에서 FASVE에 의한 차원축소에 의해 분석된 결과가 기존의 방법론에 비해 우수함을 보고자 한다.

주요용어: 결합접근법, 역회귀, 충분 차원 축소, Large p-small n 자료, sliced average variance estimation

이 연구는 대한민국 교육부의 후원을 받는 한국 국가 연구 재단(NRF)을 통해 기초 과학 연구 프로그램의 지원을 받았다 (NRF-2017R1A2B1004909).

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