다변량해석특론 I

Final Project

Sparse Sufficient Dimension Reduction

- Sufficient Dimension Reduction and Variable Selection

212STG02 고정욱

212STG18 예지혜

Contents

I. Introduction 3

II. Sparse Sufficient Dimension Reduction 3

1. Methodology 3

2. Algorithm 5

3. Simulation 6

3-1. SIR vs Sparse SIR 7

3-2. PHD vs Sparse PHD 9

4. Real Data Applications 10

4-1. Swiss bank notes data 10

4-2. Wisconsin breast cancer data 11

III. Coordinate-Independent Sparse SDR 12

IV. Result 13

V. Appendix 14

I. Introduction

통계학의 적용에 있어 다양한 분야에서 발생하는 차원의 저주를 피하고자 차원 축소에 대한 많은 연구가 이루어져 왔다. 특히 충분차원축소는 고차원 데이터에서 반응 변수에 필요한 핵심 정보를 모두 담고 있는 설명 변수를 추출해낸다는 점에서 굉장히 강력한 도구이다. 특히 데이터 크기에 비해 차원의 개수가 매우 클 때 유용하게 작용하며 이미지 분류, 시각화, 바이오 분야 등 많은 분야에서 사용되며 좋은 성능을 얻고 있다. 하지만 충분 차원 설명 변수는 보통 대부분의 기존 설명 변수의 선형 결합으로 이루어져 있어 해석하는 데에 어려움이 있다. 충분차원축소를 통해 새로운 설명 변수를 얻어 좋은 성능을 낼 수 있지만, 그 설명 변수가 무엇을 의미하는지, 결과 해석에 어떻게 이용할 수 있는 지는 얻을 수 없는 것이다.

이러한 문제를 해결하고자 차원 축소 방법론에서의 변수 선택에 대한 다양한 연구가 이루어져 왔다. 본 프로젝트에서는 그 중 Sparse Sufficient Dimension Reduction (Sparse SDR) 방법론과 Coordinate-Independent Sparse Sufficient Dimension Reduction (CISE) 방법론에 대한 논 문을 리뷰하고, 시뮬레이션을 통해 결과를 해석해 보고자 한다. Sparse 충분차원축소 방법론은 2007년 Biometrika에 게재된 Lexin Li의 Sparse sufficient dimension reduction을 참고하였으며, CISE 방법론은 2010년 The Annals of Statistics에 게재된 Xin Chen과 Changliang Zou의 Coordinate-Independent Sparse Sufficient Dimension Reduction and Variable Selection을 참고하였다.

II. Sparse Sufficient Dimension Reduction

1. Methodology

Sparse 충분차원축소 방법론은 많은 충분차원축소 방법론이 고유값 분해를 사용한다는 점에서 착안했다. 방법론마다 각기 다른 커널 매트릭스(M)를 정의하지만 결국 다음과 같은 형태로 일반화하여 표현할 수 있다.

$$Mv_i = \rho_i G v_i, \qquad i = 1, \dots, p$$

$$where\ v_i = eigenvectors, \qquad \rho_i = eigenvalues$$

$SIR\ )\ M = cov[E\{X - E(X)|Y\}], \qquad G = \Sigma_x$

$SAVE\ )\ M = \Sigma_x^{\frac{1}{2}} E[\{I - cov(Z|Y)\}^2]\Sigma_x^{\frac{1}{2}}, \qquad where\ Z = \Sigma_x^{-\frac{1}{2}}\{X - E(X)\}, \qquad G = \Sigma_x$

$PHD\ )\ M = \Sigma_x^{\frac{1}{2}}\Sigma_{yzz}\Sigma_{yzz}\Sigma_x^{\frac{1}{2}}, \qquad where\ \Sigma_{yzz} = E\{Y - E(Y)|ZZ^T\}, \qquad G = \Sigma_x$

이러한 고유값 분해 방정식을 회귀 모형 형태로 바꿀 수 있다면 회귀 모형에서와 같이 LASSO penalty 항을 통해 변수 선택이 가능해진다. 회귀 모형의 형태에 LASSO penalty 항을 추가하면 sparse estimation을 통한 beta가 찾아진다. 따라서 Central subspace 추정에 필요한 각 Sufficient Direction을 구성하는 원소들을 sparse estimation을 통해 screening할 수 있게 하며, 이는 각 방향을 구성하는 원소들에 대한 해석력을 높인다. 즉, 고유값 분해를 회귀 분석 문제로 바꾸는 것과, 이를 LASSO 모형으로 바꾸는 것이 이 방법의 핵심이며, 따라서 모든 고유값 분해 형태의 충분차원축소 방법론에 적용할 수 있고, 일반화할 수 있는 방법이라는 점에서 이 방법이 경쟁력을 가진다.

$$\hat{\beta} = argmin_{\beta} \sum_{i=1}^{p} \left|\left| G^{-1} m_i - \beta \beta^t m_i \right|\right|_G^2 \quad (1)$$

$$where\ m_i = ith\ column\ of\ M^{1/2}, \beta^T G \beta = I_d$$

고유값 분해 문제를 회귀 문제로 바꾸면 (1)과 같이 표현할 수 있고, 이때 $\hat{\beta_j} = v_j,\ j = 1, \ldots, d$ 이며 $\hat{\beta_j}$는 $\hat{\beta}$의 j번째 column이다. 즉, 위와 같은 RSS형태의 값을 최소화하는 $\hat{\beta}$을 찾으면 그것이 곧 충분차원축소에서의 고유벡터가 되는 것이다.

회귀 분석 형태의 식에 Lasso penalty 항을 부여하면 (2)와 같이 표현할 수 있다.

$$min_{\beta} \sum_{i=1}^{p} \left|\left| G^{-1} m_i - \beta \beta^t m_i \right|\right|_G^2,\ subject\ to\ \beta^t G \beta = I_d,\ and\ \left|\beta_j\right|_1 \leq \tau_j \quad (2)$$

$$\text{Lasso regression)}\ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

여기서 Lasso 회귀 분석에 대해 잠깐 설명하겠다. Lasso 회귀 분석은 선형 회귀 모델에서 Mean Squared Error를 최소화하는 식에 각 계수의 절대값의 합을 penalty 항으로 추가하는 모델이다. MSE와 함께 계수의 전체 크기도 최소화하기 때문에 덜 중요한 설명 변수의 계수들이 0으로 수렴하게 된다. 여기서 계수의 크기 합산은 penalty 항으로 작용하고, 이때 penalty 얼마나 부여할지를 결정하는 척도로 λ를 파라미터로 사용한다. 제약 조건을 통해 일반화된 모형을 찾을 수 있고, 이는 변수 선택의 결과를 가져다주어 모델 해석력을 높일 수 있다. 이렇게 0으로 수렴한 계수가 많은 모델을 sparse model이라 하며 이 방법론의 이름이 Sparse 충분차원축소인 이유이기도 하다.

다시 Sparse 충분차원축소로 돌아가보면, penalty 항을 통해 변수 선택이 가능하게 타겟 함수를 변형하였지만, 실제로 (2)번 식 형태의 최적화 문제를 푸는 것은 매우 복잡하다. 따라서 이 논문에서는 Zou가 2006년 Sparse principal component analysis 논문에서 제안한 아이디어를 사용하여 타겟 함수를 다음과 같이 수정한다.

$$min_{\alpha,\beta} \sum_{i=1}^{p} \left|\left| G^{-1} m_i - \alpha \beta^t m_i \right|\right|_G^2 + \lambda_2 tr(\beta^t G \beta) + \sum_{j=1}^{d} \lambda_{1j} \left|\beta_j\right|_1 \quad (3)$$

(3) 식의 λ_1항은 지금까지 설명한 Lasso penalty 항에 해당한다. Lasso의 penalty 항은 절대값이기 때문에, feature selection 효과를 얻을 수는 있지만, 미분이 불가능하여 closed form solution을 얻을 수 없다. 따라서 Ridge penalty 형태인 $\lambda_2 tr(\beta^T G\beta)$ 항을 식에 추가해 optimization 문제를 해결한다. 해당 항은 타겟 함수를 convex function으로 만들어 주어 global minimum을 찾기 쉽게 만들어 준다고 알려져 있다.

Sparse 충분차원축소의 최종 목적 함수 식은 (3)과 같으며, 이를 최소화하는 β를 찾는 것이 알고리즘의 목적이다. $\hat{\beta}$는 Sparse 충분차원축소 방법론의 추정치가 되며, 일반 충분차원축소 추정치들과는 달리, 중요하지 않은 변수들의 계수는 0이 되어 중요한 변수들만의 선형 결합으로 최종 충분 차원 설명 변수를 만들어낸다.

2. Algorithm

최종 목적 함수인 식 (3)를 풀기 위한 알고리즘을 살펴보면 다음과 같다. α와 β 모두 최적화해야 하기 때문에 우선적으로 하나의 파라미터를 고정하고 다른 파라미터를 구하는 과정을 서로 교차 반복하여, 수렴할 때까지 진행한다. 모든 과정을 정리하면 다음과 같다.

Step 1. 일반적인 충분차원축소 방법을 이용해 해당 추정치를 α의 초기값으로 설정한다.
Step 2. 고정된 α값에 대해 타겟 함수를 최소화하는 β를 계산한다.
Step 3. 고정된 β값에 대해 $G^{-1/2}M\beta$를 특이값 분해(SVD)하여 $\alpha = G^{-1/2}UV^T$를 계산한다.
Step 4. 2 ~ 3 단계를 반복하며 β를 수렴시킨다.
Step 5. β를 정규화한다.

Table 1. Alternating minimization algorithm for solving (5)

Step 2에서의 β 계산 식은 다음과 같다.

$$\hat{\beta}_{\alpha j} = argmin_{\beta}\{\beta_j{}^T(M+\lambda_2 G)\beta_j - 2\alpha_j{}^T M\beta_j + \lambda_{1j}|\beta_j|_1\} \quad (4)$$

$$\hat{\theta}_{\alpha j} = argmin_{\theta_j}\{\|u^* - m^*\theta_j\|^2 + \lambda_{1j}|\theta_j|_1\} \quad (5)$$

$$where \;\; m^* = \left(\frac{M^{1/2}}{\sqrt{\lambda_2}G^{1/2}}\right)_{2p\times p},\; u^* = \begin{pmatrix} M^{1/2}\alpha_j \\ 0 \end{pmatrix}_{2p\times 1}$$

즉, $u*$를 반응변수로, $m*$를 설명변수로 취급하여 Lasso 회귀 분석을 진행한 $\hat{\theta}_{\alpha j}$가 $\hat{\beta}_{\alpha j}$값이 된다. 이를 계산하기 위해서는 어떤 Lasso 알고리즘이든 사용할 수 있고, 해당 논문에서는 least angle regression method (Lars) 알고리즘을 사용하였지만, 본 보고서의 simulation 및 방법론 구현 과정에서는 Lars 알고리즘에 더해, glmnet 패키지의 lasso 방법론을 함께 활용하여 진행하였다.

이 알고리즘을 진행하는 데에 있어 Lasso, Ridge 두 penalty 항에 대한 규제 정도를 의미하는 척도들이 필요하다. 특히 λ_1에 해당하는 Lasso penalty 항은 d개의 각 설명 변수에 대해 각각 적용되기 때문에 d개의 λ_1이 필요하며, λ_2까지 총 $d+1$개의 파라미터가 필요하다. 이 논문에서는 λ_1과 λ_2 값을 결정하기 위해 AIC 척도를 사용하였으며, d개의 λ_1 ($\lambda_{11} = \lambda_{12} = \ldots = \lambda_{1d}$)은 같다고 놓고 진행하였다. AIC 척도는 다음과 같다.

$$\sum_{i=1}^{p} \left|\left| G^{-1} m_i - \hat{\beta}_\lambda \hat{\beta}_\lambda^T m_i \right|\right|_G^2 + 2p_\lambda / n \tag{6}$$

첫번째 항은 RSS로, 모형의 적합도를 설명해주며 sparse estimator가 적은 모델을 더 선호한다. 두번째 항 Lasso penalty가 없는 p_λ는 0이 아닌 계수 $\hat{\beta}_\lambda$ 의 개수로, 모형의 복잡도를 설명해주며 sparse estimator가 많은 모델을 더 선호한다. 따라서 이 AIC 척도는 계수를 모형의 적합도를 고려하면서도, 불필요한 변수들의 계수를 적당히 0으로 보내어 복잡하지 않은 모델을 선택하게 한다. 본 보고서에서는 다양한 λ_1 후보들 중 AIC 가 가장 작은 경우를 모델 학습에 이용했다.

3. Simulation

이 논문에서는 Sparse SDR의 성능을 검증하기 위해 두 가지의 충분차원축소 방법에 대한 simulation과 두 개의 데이터를 분석한다. Simulation에서는 Sparse model을 true model로 가정하고 데이터를 생성하여, 기존 충분차원축소 방법론과 Sparse 충분차원축소 방법론의 결과를 비교한다. 비교 척도로는 (1) 차원 축소 방향인 β의 계수가 0이 된 sparse estimator의 개수와 (2) $corr(\hat{\beta}_j^T X, \beta_j^T X)$, $j = 1, 2$, 추정된 차원 축소 설명 변수와 실제 모델의 설명 변수의 상관 계수의 절댓값, (3) $\hat{E}(\hat{\beta}_j^T X - \beta_j^T X)^2$, $j = 1,2$, 추정된 차원 축소 설명 변수와 실제 모델 설명 변수의 평균 제곱 오차(mean squared error)를 사용한다. 0이 된 계수의 개수가 true model의 개수와 일치할수록 sparse 충분차원축소 방법론이 불필요한 설명 변수들을 잘 가려낸다고 해석할 수 있다. 또한 상관계수는 높을수록, 평균 제곱 오차는 낮을수록 true model에 가깝다고 볼 수 있다. 이 척도들을 각 시뮬레이션 상황에 대해 200번씩 계산하여 그 평균을 비교하였다.

3-1. SIR vs Sparse SIR

$$Y_1 = sign(\beta_1^T X)\log(|\beta_2^T X + 5|) + 0.2\varepsilon$$

첫번째 시뮬레이션은 이와 같은 true model을 가정하며, X는 p=20 차원이고, 모든 X와 ε는 표준 정규분포를 독립적으로 따른다. 시뮬레이션 데이터 Y_1은 n=200개를 사용하였으며, 이를 위해 X와 ε도 200개의 난수를 발생시켰다. 비교 방법론으로는 SIR방법을 사용하였고, Sparse model에 대해서도 SIR 모델을 초기값으로 사용하였다. true model에 대해 true beta는 총 3가지 버전으로 시도하였다. (1) 20개의 변수 중 4개만이 의미 있는 변수이 경우, (2) 4개만이 의미 있으나 그 중 2개의 회귀 계수는 매우 작은 경우, (3) 20개의 변수 중 10개의 변수가 의미 있는 경우이다. 하나씩 살펴보겠다.

(1) $\boldsymbol{\beta_1} = (\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0})^{\boldsymbol{T}}$,

$\boldsymbol{\beta_2} = (\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0})^{\boldsymbol{T}}$

앞서 설명했듯, 하이퍼 파라미터 λ_1과 λ_2를 결정해야 한다. 이를 위해 논문에서 언급한 듯 AIC 기준을 사용하여 비교해보았으나, 주로 매우 작은 λ값들이 선택되어 sparse estimator가 적은 모델이 선택되는 경향이 있었다. 따라서 직접 몇 가지 λ를 시도하여 적절한 값을 찾은 결과 모든 λ에 대해 0.01을 사용하였다. 다음은 200번의 시뮬레이션을 통해 SIR와 Sparse SIR를 비교한 결과이다.

$method$	p	$corr\ of\ \beta_1$	$corr\ of\ \beta_2$	$mse\ of\ \beta_1$	$mse\ of\ \beta_2$
sir	0.00	0.7496574	0.7423155	2.641265	2.541920
$sparse\ sir$	31.67	0.9831867	0.9542988	2.907466	2.639175

Table 2. Simulations comparing SIR and Sparse SIR.

SIR 방법론의 경우 0이 된 계수가 없었고 Sparse SIR는 평균 31.7개의 계수가 0이 되었다. True model의 개수가 32개인 점을 고려하면 매우 유사하게 나온 것이라 볼 수 있다. True model과의 상관계수는 sparse sir이 0.98, 0.95 정도로 항상 매우 높았으며 sir 방법론은 0.75 정도로 낮았다. 0이 되어야 하는 계수들이 0이 되지 못해 상관 계수가 낮게 나온 것으로 볼 수 있다. 반면 mse는 sir이 약간 더 낮았다.

(2) $\boldsymbol{\beta_1} = (\mathbf{1},\mathbf{1},\mathbf{0.1},\mathbf{0.1},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0})^{\boldsymbol{T}}$,

$\boldsymbol{\beta_2} = (\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0.1},\mathbf{0.1},\mathbf{1},\mathbf{1})^{\boldsymbol{T}}$

이번 시뮬레이션은 4개의 의미 있는 변수 중 2개는 다른 2개 변수에 비해 계수가 $\frac{1}{10}$로 상대적으로 덜 중요한 효과를 주는 변수이다. 이 경우에 대해서도 적절한 λ를 찾은 결과 모든 값에 대해 0.1을 사용하였다. 다음은 200번 시뮬레이션의 결과이다.

$method$	p	$corr\ of\ \beta_1$	$corr\ of\ \beta_2$	$mse\ of\ \beta_1$	$mse\ of\ \beta_2$
sir	0.00	0.9261164	0.8989056	2.899740	1.413201
$sparse\ sir$	35.53	0.9878816	0.9859833	2.884633	1.317157

Table 3. Simulations comparing SIR and Sparse SIR.

이번 결과에 대해서는 앞선 결과와 비교했을 때 훨씬 많은 계수가 0이 되었다. 35.5개 정도가 0이 되었으므로 결국 계수 1에 해당하던 4개의 변수들 만이 주로 선택되었다고 볼 수 있다. 상관계수는 sir 방법론이 첫번째 경우에 비해 훨씬 높아진 것을 확인할 수 있고, 여전히 sparse sir의 결과가 더 좋다. 이번에는 mse 또한 sparse sir model이 더 좋은 결과를 보였다.

(3) $\boldsymbol{\beta_1} = (\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0})^{\boldsymbol{T}}$,

$\boldsymbol{\beta_2} = (\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1})^{\boldsymbol{T}}$

이번 시뮬레이션은 20개의 변수 중 10개의 변수가 유의미한 변수이다. 상대적으로 true model 자체가 덜 sparse한 모델이라고 볼 수 있다. 이 경우에도 모든 λ에 대해 0.01을 사용하였다.

$method$	p	$corr\ of\ \beta_1$	$corr\ of\ \beta_2$	$mse\ of\ \beta_1$	$mse\ of\ \beta_2$
sir	0.000	0.8505567	0.8452100	3.583607	2.417704
$sparse\ sir$	21.655	0.8210612	0.8068314	3.532310	2.327421

Table 4. Simulations comparing SIR and Sparse SIR.

시뮬레이션 결과, 약 21.7개의 계수가 0이 되었다. True model이 20개인 것을 고려할 때 이전의 시뮬레이션에 비해 적은 변수들을 0으로 만든 것을 확인할 수 있다. 이번 시뮬레이션에서는 sparse sir보다 sir의 상관계수가 더 좋았다. 또한 Sparse sir의 상관계수는 0.8 정도로 앞의 시뮬레이션에 비해 낮아졌다.

다양한 true model에 대해 시뮬레이션을 진행한 결과를 요약해보면, 논문에서와 마찬가지로, sparse model이 실제 모델일 경우 sparse 충분차원축소 방법론이 그 효과를 훨씬 잘 보여주었다. 유의미한 변수가 4개일 때는 sparse sir 방법의 상관계수가 0.98 정도로 sir보다 더 높고, 유의미한 변수가 10개로 늘었을 때는 0.82 정도로 sir보다 더 낮았다. 이를 통해 Lasso 알고리즘의 자체의 성질이기도한, sparse model에서 더 성능이 좋고 유용하다는 특징을 확인할 수 있다.

3-2. PHD vs Sparse PHD

$$Y_2 = cos(2\beta_1^T X) - \cos(\beta_2^T X) + 0.5\varepsilon$$

(1) $\boldsymbol{\beta_1} = (\mathbf{1},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0})^{\boldsymbol{T}}$,

$\boldsymbol{\beta_2} = (\mathbf{0},\mathbf{1},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0})^{\boldsymbol{T}}$

이번 시뮬레이션은 잔차를 기반으로 한 principal Hessian directions 방법론을 활용한 결과이다. 10개의 변수 중 1개의 변수가 유의미한 변수인 모델로, true model이 상대적으로 sparse한 모델이라고 볼 수 있다. 샘플의 크기는 $n = 100$, $n = 200$ 두 경우로 비교 분석을 진행했으며, $n = 100$인 경우의 $\lambda_1 = 0.2$로, $n = 200$인 경우의 $\lambda_1 = 0.06$을 사용했으며, 두 경우 모두 $\lambda_2 = 0.01$을 사용했다.

	$method$	$selected\ var.$	$corr\ of\ \beta_1$	$corr\ of\ \beta_2$	$mse\ of\ \beta_1$	$mse\ of\ \beta_2$
$n = 100$	$phdres$	0.000	0.6378344	0.6029761	0.9935561	2.474934
	$sparse\ phdres$	2.21	0.7614605	0.7175787	0.6856594	2.724228
$n = 200$	$phdres$	0.000	0.9142757	0.9480889	0.1849264	1.117345
	$sparse\ phdres$	2.17	0.9907722	0.999917	2.235337	1.077499

Table 5. Simulations comparing PHD and Sparse PHD

시뮬레이션 결과, n=100인 경우, 2.21개의 계수가 1의 값을 가져 변수로 선택되었고, n=200인 경우, 2.17개 계수가 1로 추정되었다. True model에서 유의한 변수가 2개인 것을 고려할 때, 서로 다른 샘플의 크기를 가진 두 경우 모두 적절한 변수들을 0으로 만드는 것을 확인할 수 있다. 샘플이 작은 경우, β_1, β_2 상관계수가 sparse phd에서 각 0.76, 0.71 정도로, 일반 phd의 경우 β_1, β_2 상관계수가 각 0.63, 0.60인 것에 비해 높은 값을 가졌다. 한편, 샘플이 큰 경우의 β_1 상관계수가 sparse phd에서 0.99 정도로, 일반 phd의 경우 β_1 상관계수가 0.91인 것에 비해 높은 값을 가졌다. β_2의 경우도 마찬가지로, 일반 phd 모형에서 0.94인 것에 비해, sparse phd 모형에서는 0.99로 더 높은 값을 상관계수 값을 가졌다.

10개의 변수 중, 1개의 변수만이 유의한 sparse true model의 경우, sparse 충분차원축소 방법론이 더 좋은 성능을 보여주었고, 데이터의 개수가 충분히 클수록 모든 충분차원축소 방법론들이 더 좋은 성능을 보여주었다.

4. Real Data Applications

4-1. Swiss bank notes data

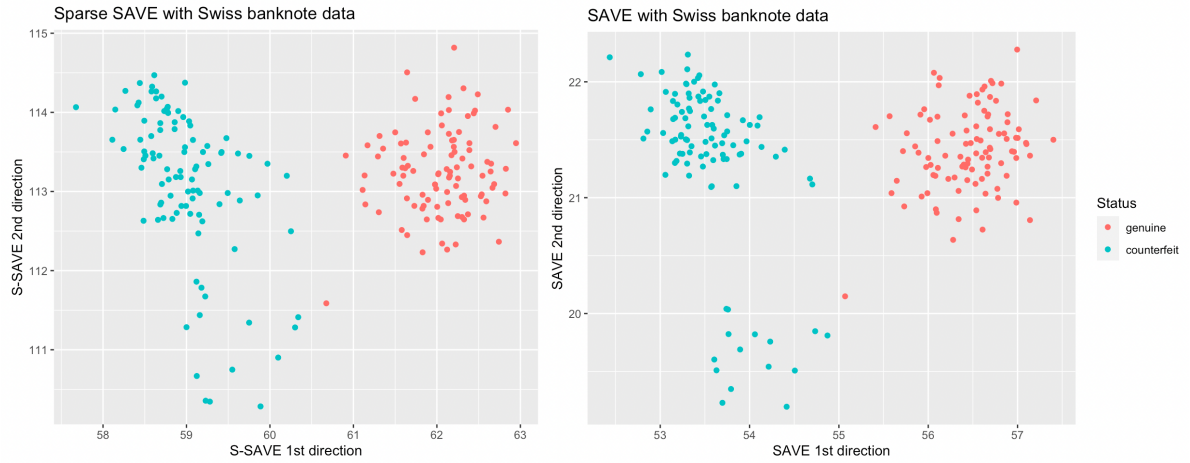
우선적으로 적용해 볼 데이터는 Swiss banknotes data 이다. 정품 100 장, 위조 100 장 등 스위스 지폐 200 장에 대해, 길이, 왼쪽 가장자리 폭, 오른쪽 가장자리 폭, 아래쪽 여백 폭, 위쪽 여백 폭 등의 변수를 제공한다. 측정 단위는 밀리미터(mm)이다. 데이터의 절반이 위조인 상황에서 6 개의 변수를 통해 지폐의 진위여부를 가려내는 분류 모델의 정확도를 높이는 것이 분석 목표이다.

200개의 데이터가 존재했으나, 데이터 특성상 이상치 탐지가 쉽지 않아, box plot을 통해 이상치로 판단되는 관측치를 제거하여 총 194개의 데이터를 분석에 사용하였다. Boxcox 기법을 사용해 변환한 데이터로도 분석을 진행해보았으나, 변환이 필요하지 않다고 판단했다.

논문에 따르면 SAVE 방법론을 적용하며, 두번째 방향까지 사용하였고, λ는 0.01의 값을 사용하였다. 그 결과 SAVE와 S-SAVE 방법을 적용해 추정한 방향과 beta 추정치는 다음과 같다.

$SAVE$	$(0.0308, 0.2031, -0.2531, -0.5893, -0.5680, -0.4730)^T$ $(-0.2841, -0.0547, -0.1573, 0.5060, 0.3340, 0.7237)^T$
$Sparse\ SAVE$	$(0, 0, 0, -0.6792, -0.1625, 0.7156)^T$ $(0, 0, 0, 0.6492, 0, 0.7605)^T$

SAVE는 모든 변수에 대한 계수를 0이 아닌 수로 추정하였고, Sparse SAVE는 첫번째 방향에 3개의 변수를, 두번째 방향에 2개의 변수만을 남겨두었다. 이러한 결과를 통해 지폐의 진위 여부 판단에 아래쪽 여백, 위쪽 여백, 대각선 길이 등이 중요한 것으로 해석할 수 있다.



타겟 변수인 y에 대해 SAVE와 Sparse SAVE의 각 방향을 비교해보면, 두 방향 모두 위조지폐를 잘 분류해내는 것을 확인할 수 있다. 또한 첫번째 방향의 상관계수는 0.9986로 꽤 높은 수치이며, 두번째 방향의 상관계수도 0.9185로 높은 수치를 보였다. Sparse SDR 방법론이 본래 분석 목적인 분류 문제의 성능을 유지하면서도 해석력을 높이고 있다는 것을 확인할 수 있다.

4-2. Winsconsin breast cancer data

두 번째로 적용해 볼 데이터는 윈스콘신 유방암 데이터이다. 세포핵의 반경, 질감, 면적 등 10가지 정보에 대한 평균, 표준편차, tail 값이 설명변수로 총 30개이며, 종속변수는 유방암 여부이다. 30가지 변수를 통해 유방암을 진단해내는 것이 분석 목표이다. 데이터는 569개가 있으며 원활한 분석을 위해 IQR을 기준으로 이상치는 제거하여 총 545개의 데이터를 사용하였다. SIR 방법론을 적용하기 위해 데이터에 Boxcox transformation을 진행한 후 분석하였다.

논문에 따르면 SIR 방법론을 적용할 때 첫 번째 방향으로도 잘 분류되어 d=1로 사용하였다. λ는 여러 값을 시도하여 논문과 유사한 결과를 내는 값을 선택하였다. 그 결과 얻은 추정된 방향은 다음과 같다.

SIR	$(-0.058, -0.035, 0.346, -0.391, 0.002, 0.0003, -0.217, -0.107, 0.014, 0.077,$ $0.137, 0.009, -0.009, -0.341, -0.020, -0.047, 0.171, -0.077, 0.035, 0.081,$ $-0.242, -0.017, -0.137, 0.563, -0.014, 0.119, -0.072, 0.174, -0.071, -0.162)T$
$Sparse\ SIR$	$(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,$ $0, 0, 0, 0, 0, -0.612, -0.791, 0, 0, 0, 0, 0, 0, 0, 0, 0)T$

SIR은 모든 변수에 대해 계수를 0이 아닌 수로 추정하였고, Sparse SIR은 2개의 변수만을 남겨두었다. 이러한 결과를 통해 유방암 진단에 반경과 질감의 tail이 중요한 것으로 해석할 수 있다.

타겟 변수인 y를 표현하여 SIR과 Sparse SIR의 방향을 비교해보면, 두 방향 모두 유방암을 잘 분류해내는 것을 확인할 수 있다. 또한 두 방향의 상관계수는 0.887로 꽤 높은 수치이다. Sparse SDR 방법론을 통해 해석력을 높이고, 원래의 목적인 분류 문제 또한 잘 해결하는 것을 확인할 수 있다.

III. Coordinate-Independent Sparse Sufficient Dimension Reduction

Sparse 충분차원축소 방법론은 2007 년 Lexin Li 가 소개한 방법으로 sufficient predictor 를 추정할 때 각 direction 을 위한 설명변수를 선택하는 element screening 방법이다. 해당 방법론은 각 방향에 대해서만 sparse estimation 이 진행되기 때문에, 특히 방향이 많아질 경우, 궁극적으로 variable screening 의 효과를 가져다주지는 않는다. CISE 는 Sparse 충분차원축소 방법론의 이런 한계점에서 착안하여 만들어진 방법론으로, Sparse SDR 과 마찬가지로 차원 축소와 동시에 변수 선택을 가능하게 한다.

기존의 Sparse 충분차원축소 방법론들은, sparse solution 을 추정할 때, central subspace 의 basis matrix 를 각 열에 대해 (column by column)으로 즉, stepwise 방식으로 추정을 진행하는데, 이런 경우 각 sufficient direction 을 구성하는 element variable 에 대해 sparse 한 추정을 진행할 수 있으나, 최종적으로 어떤 변수를 가려낼 것인지, 최종 변수 set 에 대해 알려주지는 않는다.

$$(\hat{\alpha}, \hat{V}_s) = min_{\alpha,V}\{ \sum_{i=1}^{p} \left|\left| G^{-1}m_i - \alpha V^t m_i \right|\right|_G^2 + \lambda_2 tr(V^t G V) + \sum_{j=1}^{d} \lambda_{1j} \left\| V_j \right\|_1 \quad (7)$$

$$\rho(V) = \sum_{i=1}^{p} \theta_i \left\| V_i \right\|_2 \quad (8)$$

$$\tilde{V} = argmin_V \{ -tr(\hat{V} M V) + \rho(V) \quad (9)$$

Sparse SDR 방법론의 목적함수이다.(7) L1 penalty term 은 직교 변환에 영향을 받기 때문에, coordinate dependent 하며, 따라서 basis matrix 를 추정 할 때, 어떤 변수에 대한 행 전체를 0 으로 보내는 것이 아니라, 개별 원소들을 0 으로 보낸다. 따라서 앞서 소개됐던 Lexin Li 의 방법론은 "element screening"을 가능하게 한 것이며, "variable screening"을 가능하게 하기 위해, CISE 라는 새로운 변수 선택 방법이 만들어졌다.(9) CISE penalty term (8)은 이는 grouped lasso penalty term 과도 같은 형태를 가진다. CISE 는 moment based SDR 뿐만 아니라 model based SDR 방법에도 적용 가능하다는 장점이 있다.

.IV. Result

일반적인 SDR 방법론은 sufficient predictor 가 보통 데이터의 모든 설명변수들의 선형 결합으로 이루어져 있다. 따라서 해석이 복잡하고 중요한 변수를 식별하기 어렵다는 한계점을 갖는다. 본 보고서에서는 이러한 한계를 해결하기 위해 central subspace 를 추정함에 있어, 무의미한 설명변수들을 제거하면서 sufficient predictor 를 추정하는 변수 선택 방법들을 살펴보았다.

Sparse SDR 방법론들 활용한 simulation 결과들을 보면, 때, data 를 생성한 true model 이 Sparse 할수록 우수한 성능을 보였다. 이는 불필요한 설명변수가 많을수록 우수한 성능을 보이는 Lasso 의 특징이기도 하다. 또한 실제 데이터에 적용한 결과 또한 기존의 방법론과 유사한 성능을 보이는 것을 확인했다.

이 논문에서 소개하는 Sparse SDR 의 장점은, 해석력을 높일 수 있다는 측면에서 해결책을 마련하거나 연구에 도움을 줄 수 있다는 것과, eigen decomposition 을 이용하는 모든 차원 축소 방법론에 동일하게 적용할 수 있다는 것이다. eigen decomposition 식을 변형하여 penalty 를 부여하기 때문에 다양한 방법론에 활용 가능하다. 적용 분야로는 특히 sparse model 이 많은 바이오 분야에서 유용하게 사용할 수 있다. 한계점은, eigen decomposition 을 사용하지 않는 차원 축소 방법론에 대해서는 해당 방법론을 적용할 수 없다는 것이다.

분석을 진행하며, 실제로 논문에 나온 결과 값들을 재현해내는 데 어려움이 있었다는 점이 본 보고서의 가장 큰 한계점이라고 생각한다. λ_1, λ_2를 비롯해 파라미터 값들이 변화하면, 각 방향에 대해 얻어지는 β 에 대해, 계수가 0 으로 수렴하는 변수의 위치들이 달라지는 결과가 나타나 파라미터 지정에 어려움이 있었다. 결과적으로 논문이 보고한 결과 값들과 가장 비슷한 값을 보이는 파라미터를 기준으로 분석을 진행하였지만, Sparse 충분차원축소 방법론은 해석력을 높이기 위한 목적으로도 사용되는 만큼, 수렴하는 변수들의 위치가 파라미터 값에 영향을 많이 받는다는 것은 한계점이라고 생각된다. 추가로 본 보고서에서 CISE 방법론에 대해 깊이 다루지 못한 것과, 프로그래밍을 통해 직접 구현해보지 못한 것에도 아쉬움이 남아, 기회가 된다면 추후에 진행해보고자 한다.

V. Appendix

LI, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613. MR2410011

Chen, X., Zou, C., & Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. The Annals of Statistics, 38(6), 3696–3723. https://doi.org/10.1214/10-AOS826