SDR 과 머신러닝을 이용한 바디 퍼포먼스 등급 예측

-Order Determination 을 중심으로

다변량 해석 특론

Final Project

212STG07 노기정

목차

1. 프로젝트 목표

1) 다양한 order determination 방법론을 정리한 후, Body Performance Data에 적용해 각 방법론의 결과를 비교한다. 그리고 충분차원축소를 위한 최적의 dimension "d"를 추정한다.

2) 충분차원축소 방법론 중 SIR, SAVE, DR, IHT, PHD을 Body Performance Data에 적용한 후, 머신러닝 예측모형을 만들어 예측 성능을 비교한다.

2. Order Determination method 정리

충분차원축소에서 차원 축소를 위한 order를 결정하는 것은 중요한 일이다. 따라서, 본 프로젝트에서는 수업시간에 배운 sequential test와 ladle estimator 이외에도 BIC-type criteria와 Ye-weiss method을 사용하여 order를 추정하고 비교하였다.

데이터에 적용하기 전, 네 가지 Order Determination method를 간단히 정리하고 알고리즘을 살펴보았다.

2-1) Sequential Test

eigenvalue 크기에 의존해 카이제곱검정을 하는 방법이다.

<Algorithm>

1. Sufficient Dimension Reduction 방법론을 적용해 candidate matrix를 구한다.
2. Candidate matrix에서 eigen-decomposition을 통해 eigenvalue를 구한다. 구해진 eigenvalue로 candidate matrix의 Rank를 계산한다.
3. Rank(M)= d0이므로, true eigenvalue의 개수가 d0개이다. 따라서, eigenvalue가 크다가 작아지는 부분에 주목해서 가설검정을 실시한다.

$$L_r(F_n) = n\sum_{i=r}^{p}\lambda_i[\Lambda(F_n)]. \qquad r = 0, 1, \ldots, p-1,$$

[test statistic]

$$H_0^{(r)} : \ \mathrm{rank}[\Lambda(F_0)] = r, \quad r = 0, \ldots, p-1.$$

[귀무가설]

4. Test statistic이 크면 귀무가설을 기각한다.
5. 귀무가설을 기각할 수 없는 첫번째 d0 dimension을 찾는다.

$$\hat{d} = \min\{r : \ H_0^{(r)} \ is \ accepted, \ r = 0, \ldots, p-1\}$$

2-2) BIC-type Criteria

파라미터 수에 벌점을 주면서 information을 최대화하는 방법으로 sequential test와 비슷하게 eigenvalue 크기에 의존한다.

<Algorithm>

1. Sufficient Dimension Reduction 방법론을 적용해 candidate matrix를 구한다.
2. Candidate matrix에서 eigen-decomposition을 통해 eigenvalue를 뽑아낸다.
3. 구해진 eigenvalue를 가지고 다음과 같이 BIC-type criterion을 정의한다.

$$B_n(k) = \rho_k(\lambda_1, \ldots, \lambda_p) + c_1(n)c_2(k)$$

4. BIC-type Criterion인 $B_n(k)$을 가장 크게하는 k를 찾는 것이 목표이다.

$$\hat{r} = \mathrm{argmax}\{B_n(k) : k = 0, \ldots, p\}.$$

2-3) Ye-Weiss Method

Bootstrapped Eigenvector Variation에 의존하는 방법이다.

<Algorithm>

1. Sufficient Dimension Reduction 방법론을 적용해 candidate matrix를 구한다.

2. Candidate matrix에서 eigen-decomposition을 통해 eigenvector를 뽑아낸다. (full sample)

3.

$$B_b^*(k) = (v_{1b}^*, \ldots, v_{kb}^*).$$

여기서 bootstrap을 통해 candidate matrix를 m개 더 생성하고 eigenvector 또한 m개 더 뽑아낸다.

4. 그 후, criterion $C(k)$를 평가한다.

$$C(k) = m^{-1} \sum_{b=1}^{m} [1 - \det(B_k^{\top} B_{kb}^*)]$$

full sample의 eigenvector와 bootstrap을 통한 eigenvector 사이의 거리를 비교한다. 거리가 비슷하면 $\det(B_k^T B_{kb}^*)$ 부분이 1에 가까워지고, 결국 $C(k)$는 작은 값이 나온다.

따라서, $C(k)$ 가 작다가 커지는 부분을 찾아야한다.

2-4) Ladle Estimator

Ye-Weiss method에서 한 단계 더 나아간 방법론으로, Eigenvalue 크기와 bootstrapped eigenvector의 변동을 둘 다 가져와 결합하는 방법이다.

<Algorithm>

$$f_n^0(k) = \begin{cases} 0, & k = 0, \\ m^{-1} \sum_{b=1}^{m} \{1 - |\det(\hat{B}_k^{\top} B_{k,b}^*)|\}, & k = 1, \ldots, p-1. \end{cases}$$

$$f_n : \{0, \ldots, p-1\} \to \mathbb{R}, \quad f_n(k) = f_n^0(k) \,/\, \{1 + \textstyle\sum_{i=0}^{p-1} f_n^0(i)\},$$

1. Ye-Weiss method에서 구한 criterion $C(k)$와 비슷하게 $f_n^0(k)$ 도 full sample의 eigenvector와 bootstrap을 통한 eigenvector 사이의 거리를 비교한다. 거리가 비슷하면 $\det(B_k^T B_{kb}^*)$ 부분이 1에 가까워지고, 결국 $f_n(k)$는 작은 값이 나온다.

$$\phi_n : \{0, \ldots, p-1\} \to \mathbb{R}, \quad \phi_n(k) = \hat{\lambda}_{k+1} / (1 + \sum_{i=0}^{p-1} \hat{\lambda}_{i+1}),$$

2. 다음으로 full sample에서 eigenvalue의 값만 고려해 $\emptyset_n(k)$ 함수를 정의한다. λ값은 true dimension 일수록 크기 때문에 $\emptyset_n(k)$는 true dimension일수록 크다.

$$g_n : \{0, \ldots, p-1\} \to \mathbb{R}, \quad g_n(k) = f_n(k) + \phi_n(k),$$

3. Ladle estimator는 $f_n(k)$ + $\emptyset_n(k)$가 가장 작아지는 k를 찾는 것이 목표이다.

($f_n(k)$ 는 작고, $\emptyset_n(k)$는 크다가 어느 순간 $f_n(k)$ 와 $\emptyset_n(k)$가 둘 다 작아지는 부분이 생긴다.)

3. 데이터 분석

3-1) 데이터 소개 및 전처리

<데이터 소개>

kaggle에서 Body Performance Data를 사용해 분석을 진행하였다.

(출처: https://www.kaggle.com/kukuroo3/body-performance-data)

Variables	Type	description
age	Numerical	나이
gender	categorical	성별
height_cm	Numerical	신장(cm)
weight_kg	Numerical	체중(kg)
body.fat_.	Numerical	체지방률(%)
diastolic	Numerical	이완기 혈압
systolic	Numerical	수축기 혈압
gripForce	Numerical	악력
sit.and.bend.forward_cm	Numerical	앉아 윗몸 앞으로 굽히기(cm)
sit.ups.counts	Numerical	의자에 앉았다 일어서기 (회)
broad.jump_cm	Numerical	멀리뛰기(cm)
class	categorical	등급(A/B/C/D)

나이, 신체적 특징, 운동 성취도에 따른 performance grade를 분류한 데이터이며, column은 12개, row는 13393개로 이루어져있다.

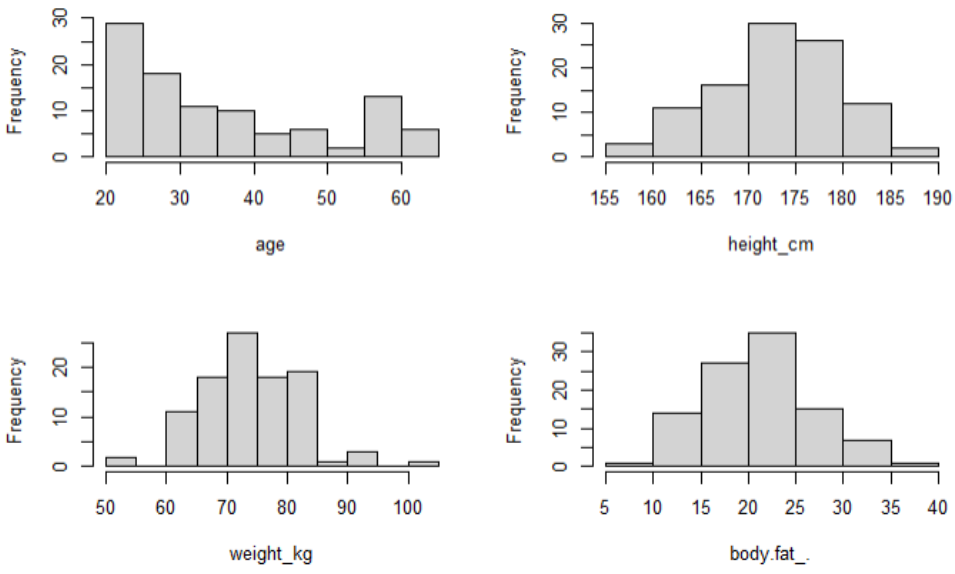
본 프로젝트에서는 충분차원축소를 이용한 예측모형의 정확도 향상을 보는 것이 목표이기 때문에 전체 데이터를 다 사용하지 않고 차원의 저주가 생길 정도로 데이터의 행을 줄였다. 따라서, gender변수에서 "male"에 해당하는 데이터 100개를 random하게 뽑아서 "male"데이터로 저장하고 이를 분석 데이터로 사용하였다. 아래 그림은 male data의 일부이다.

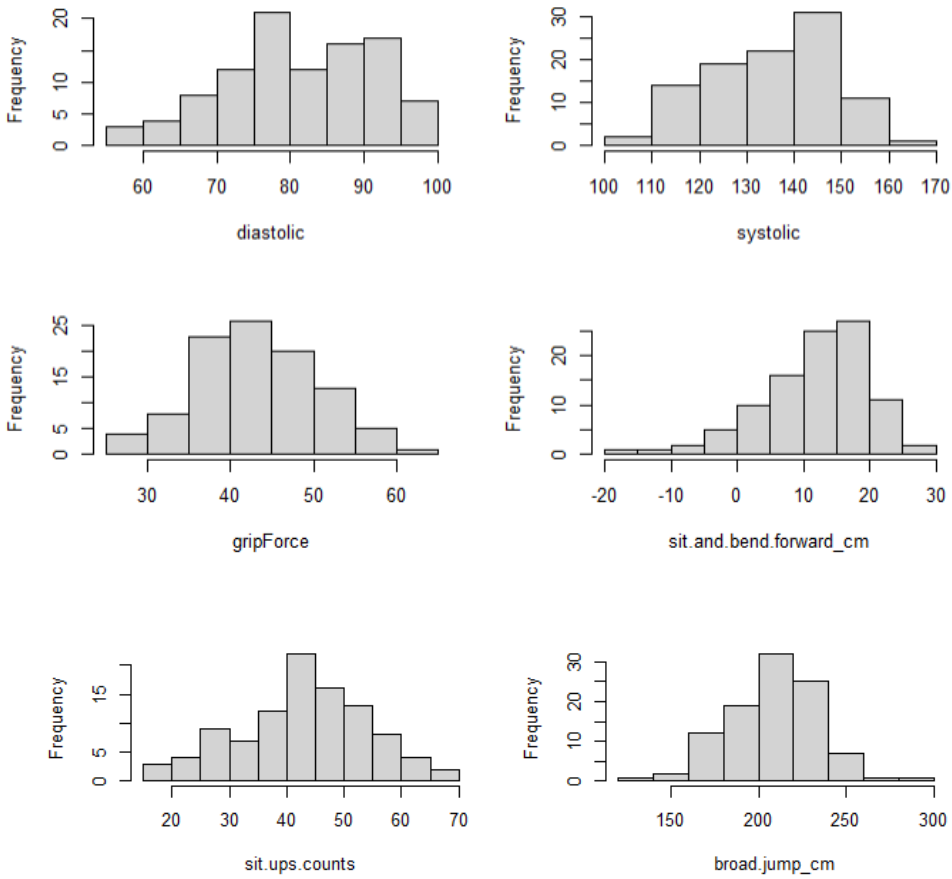
	age	height_cm	weight_kg	body.fat_.	diastolic	systolic	gripForce	sit.and.bend.forward_cm	sit.ups.counts	broad.jump_cm	class
1744	55	162.4	64.38	21.2	85	143	39.3	10.1	28	124	C
6839	26	171.2	74.10	13.2	75	145	46.4	16.0	65	249	A
2859	32	178.1	88.00	21.6	74	138	58.1	22.5	54	230	A
9075	32	183.0	90.70	26.5	96	152	57.2	20.8	47	217	D
10679	58	178.0	72.80	24.3	62	129	39.0	23.5	30	168	D
7035	50	175.0	81.30	18.5	80	123	51.1	18.3	45	202	A
7289	21	181.9	76.60	24.4	77	146	49.2	21.6	47	226	D
163	45	170.0	70.00	20.0	73	137	34.2	8.5	36	219	D
593	30	175.2	71.50	12.6	96	145	45.2	18.1	47	224	A
12163	21	181.8	72.20	17.0	84	143	48.5	11.8	59	240	B
4781	49	173.8	63.04	19.3	78	115	43.5	4.6	31	202	C
1932	56	164.0	68.66	22.3	92	144	41.2	10.8	44	199	B
7527	56	164.0	73.80	24.3	95	129	48.7	18.6	40	188	B
6795	24	181.3	74.90	12.0	89	140	53.3	11.3	52	178	C

분석에 사용한 male data는 column이 11개, row가 100개인 데이터이다.

<EDA & 전처리>

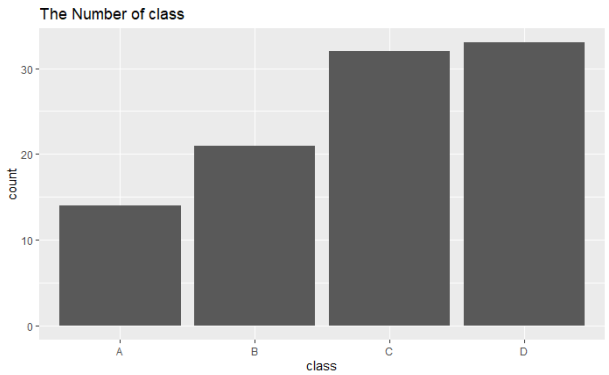
먼저 male data에서 연속형 변수의 특성을 살펴보기 위해 히스토그램을 그렸다.





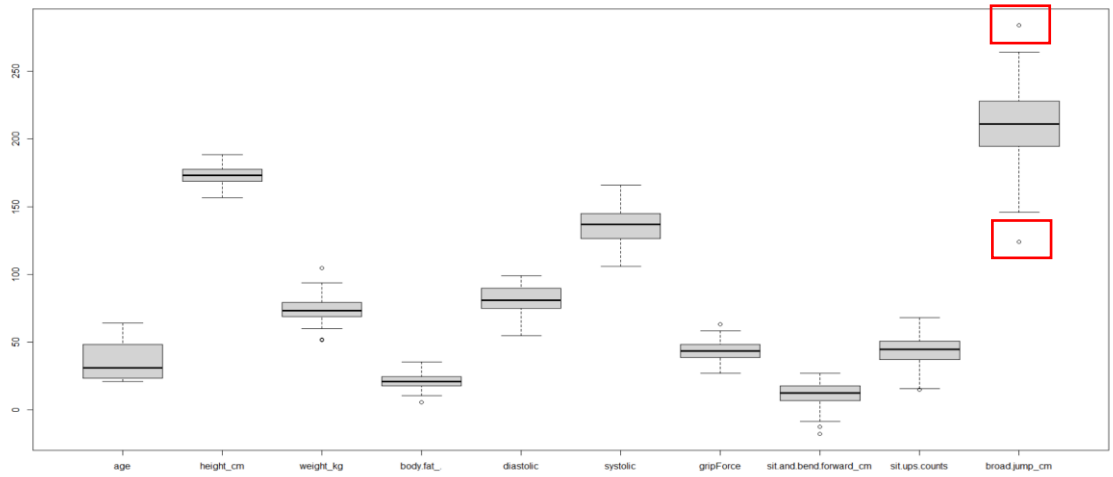
age변수를 제외한 나머지 특성들의 분포가 거의 정규분포에 가깝다.

그 후, 범주형 변수의 특성을 bar plot을 그려 살펴보았다.

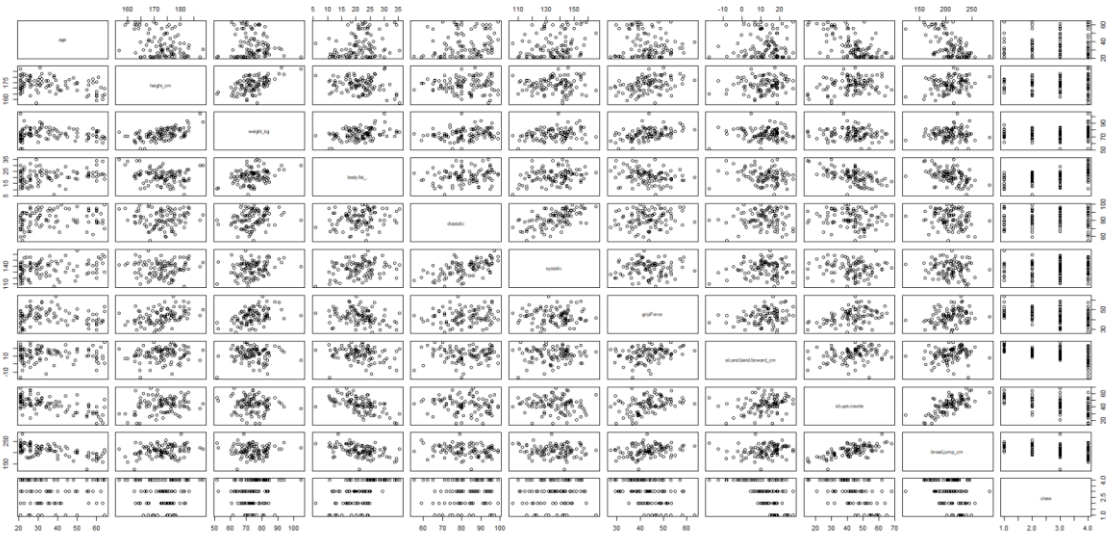


A등급에 속한 사람이 가장 적고, D등급에 속한 사람이 가장 많은 것을 확인할 수 있다.

다음은 각 변수의 이상치를 탐색하고 제거하는 과정이다.



borad.jump_cm의 경우 양 극단에 있는 값을 이상치로 판단하였다. 따라서, 최소값과 최대값에 해당하는 2개의 데이터를 제거하였다. 나머지 98개의 데이터는 scatter matrix를 그려 충분차원축소를 위한 LCM을 만족하는지 확인하였다.



male 데이터는 elliptical distribution이므로 LCM을 만족한다.

따라서, 충분차원축소를 적용할 수 있다.

충분차원축소를 하기 전, train set과 test set을 8:2로 나누었다. train set은 78개, test set은 20개이다.

3-2) Sufficient Dimension Reduction

충분차원축소의 진행 과정은 다음과 같다.

충분차원축소 방법론으로 SIR, SAVE, DR, PHD, IHT를 사용한다.

이 때, order determination을 위한 방법론으로 sequential test, BIC type criteria, Ye-Weiss

method, ladle estimator을 사용해 최적의 dimension을 결정한다. 선택된 dimension으로 train set을 차원축소 한 후, train set에서 얻은 방향으로 test set을 차원축소한다.

<Sequential Test>

SIR, SAVE, DR은 slice 개수를 8로 지정하였다.

SIR, SAVE, DR, PHD, IHT의 sequential test 후 p-value 결과는 아래 표와 같다.

r	0	1	2	3	4	5	6	7
SIR	0.01	0.31	0.81	0.94	0.99	1	0.87	
SAVE	0.03	0.07	0.09	0.1	0.43	0.5	0.38	0.36
DR	0	0.09	0.18	0.43	0.24	0.36	0.58	0.32
PHD	0.07	0.35	0.61	0.77	0.86	0.94	0.88	0.86
IHT	0	1	1	1	1	1	1	1

P-value를 살펴보니 SIR은 r=2에서, SAVE는 r=4에서, DR은 r=3에서, PHD는 r=1에서, IHT는 r=1에서 상당한 jump가 일어났다. 따라서, SIR의 d는 2로 추정되고, SAVE의 d는 4, DR의 d는 3, PHD의 d는 1, IHT의 d는 1로 추정된다.

<BIC-type Criteria>

SIR, SAVE, DR의 slice 개수는 8개로 지정하고, "lal" criterion을 사용하였다.

r	0	1	2	3	4	5	6	7	8	9	10
SIR	0.01	0.36	0.32	0.2	0.04	-0.18	-0.48	-0.80	-1.13	-1.45	-1.78
SAVE	0	1.12	1.8	2.3	2.54	2.52	2.35	2.12	1.81	1.4	0.96
DR	0	3.36	5.66	6.85	7.77	8.59	9.08	8.61	8	6.95	5.75
PHD	0.	0.42	0.68	0.77	0.65	0.53	0.24	-0.14	-0.78	-1.57	-2.5
IHT	0	3.46	0.35	-2.74	-5.84	-8.94	-12.04	-15.14	-18.24	-21.35	-24.45

[r curve table]

BIC-type Criterion인 $B_n(k)$를 크게 하는 k를 찾으면 된다. 따라서, SIR은 dimension은 1, SAVE는 4, DR은 6, PHD는 3, IHT는 1로 추정된다.

<Ye-Weiss Method>

SIR, SAVE, DR의 slice 개수는 8개로 지정하고, nbootstrap은 200으로 지정하였다.

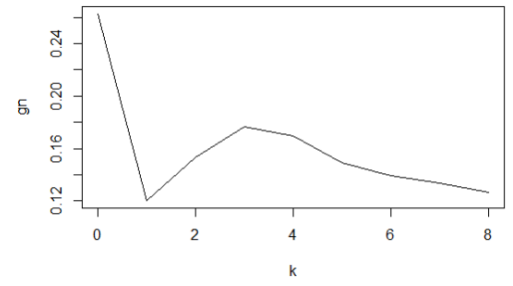
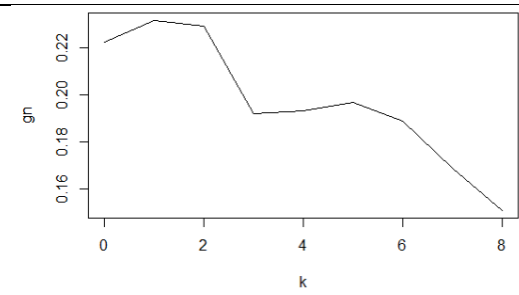
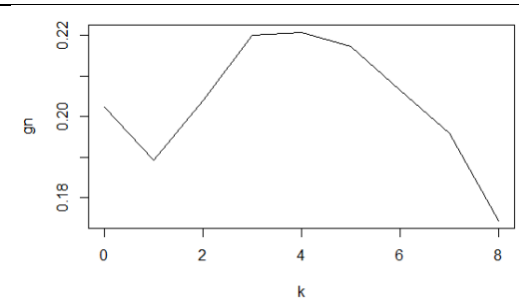
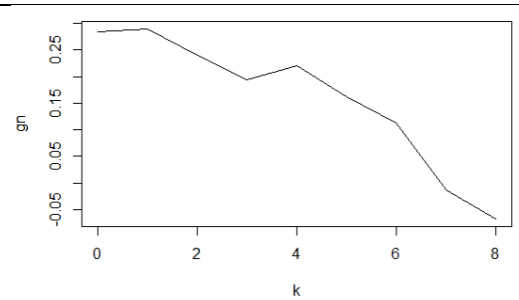
k	1	2	3	4	5	6	7
SIR	0.081	0.835	0.784	0.886	0.917	0.93	
SAVE	0.514	0.722	0.787	0.825	0.858	0.858	0.854

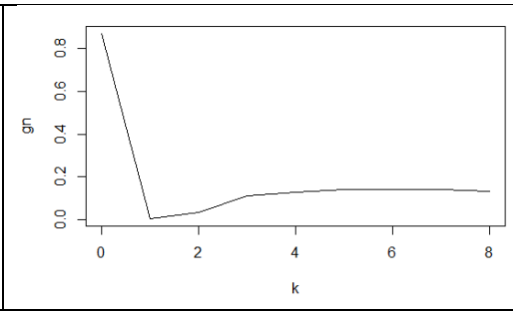
DR	0.33	0.573	0.768	0.813	0.829	0.891	0.832
PHD	0.293	0.327	0.397	0.595	0.574	0.519	0.388
IHT	0.048	0.195	0.726	0.853	0.924	0.932	0.912

Bootstrap eigenvector variation을 본 결과, SIR, SAVE, DR은 k가 1에서 2로 갈 때 상당히 값이 크게 jump한다. 따라서, true rank는 1이다. 또한, PHD는 k가 3에서 4로 갈 때 값이 크게 jump하므로 true rank는 3이다. 마지막으로, IHT는 k가 2에서 3으로 갈 때 값이 크게 jump 하므로 true rank는 2이다.

<Ladle Estimator>

SIR, SAVE, DR의 slice 개수는 8개로 지정하고, nbootstrap은 200으로 지정하였다.

SIR	
SAVE	
DR	
PHD	

IHT	

Ladle estimator는 $f_n(k)$ + $\emptyset_n(k)$ 가 가장 작아지는 k를 찾으면 된다. 따라서, SIR의 dimension은 1로 추정되고, SAVE는 3, DR은 1, PHD는 3, IHT는 1로 추정된다.

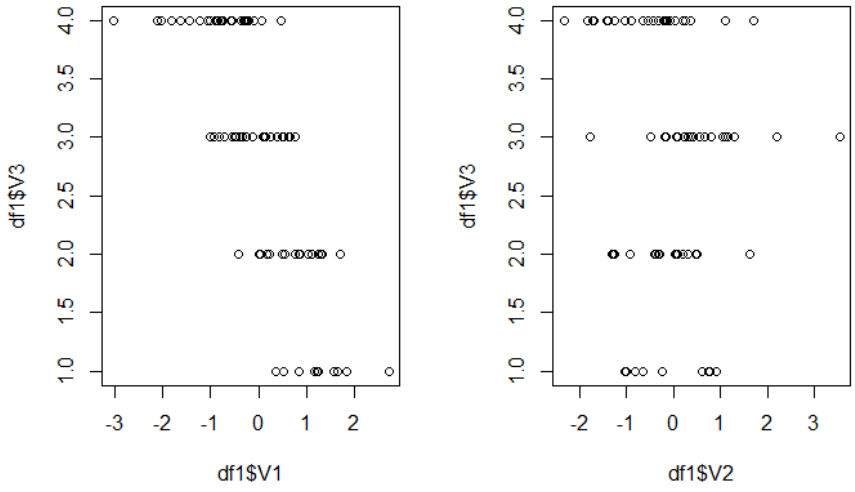
따라서, 각 order determination 방법론 별로 추정된 dimension을 다음 표에 정리하였다.

	Sequential Test	**BIC-type Criteria**	**Ye-Weiss method**	**Ladle Estimator**
SIR	2	1	1	1
SAVE	4	4	1	3
DR	3	6	1	1
PHD	1	3	3	3
IHT	1	1	2	1

[추정된 dimension]

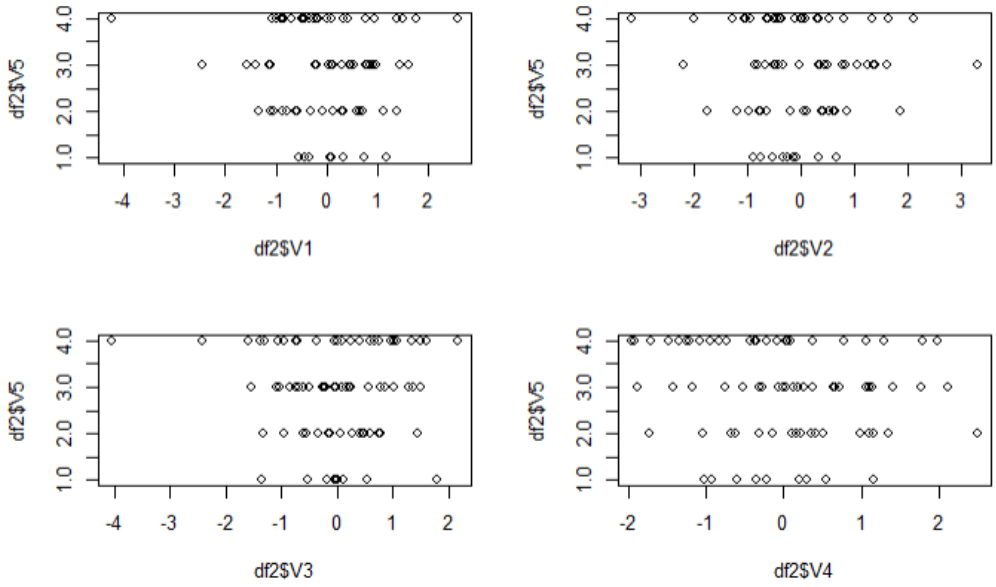
method별로 dimension이 다르게 추정되었기 때문에 충분차원축소 후 sufficient predictors와 y와의 관계를 plot으로 그려 확인한 후 최종 차원을 선택하고자 한다.

<SIR>



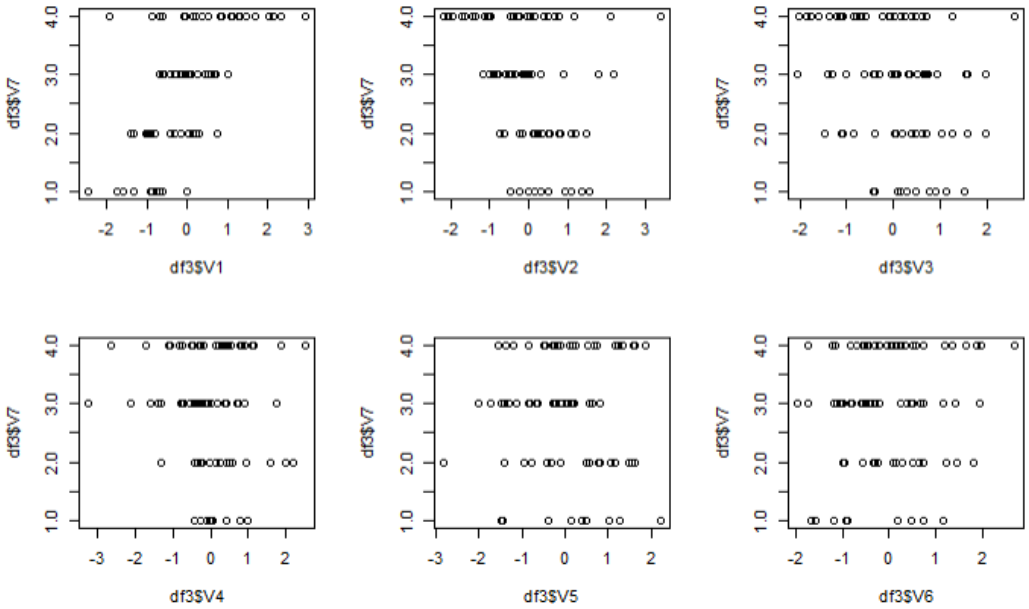
y와 SIR의 1-d sufficient predictor, 2-d sufficient predictor의 관계를 살펴보았다. y와 1-d sufficient predictor는 sufficient predictor의 크기가 커질수록 y는 작아지는 강한 음의 관계가 나타나는데, 2-dimensional에서는 딱히 관계가 보이지 않는다. 따라서, sir의 최종 dimension은 1로 선택한다.

<SAVE>



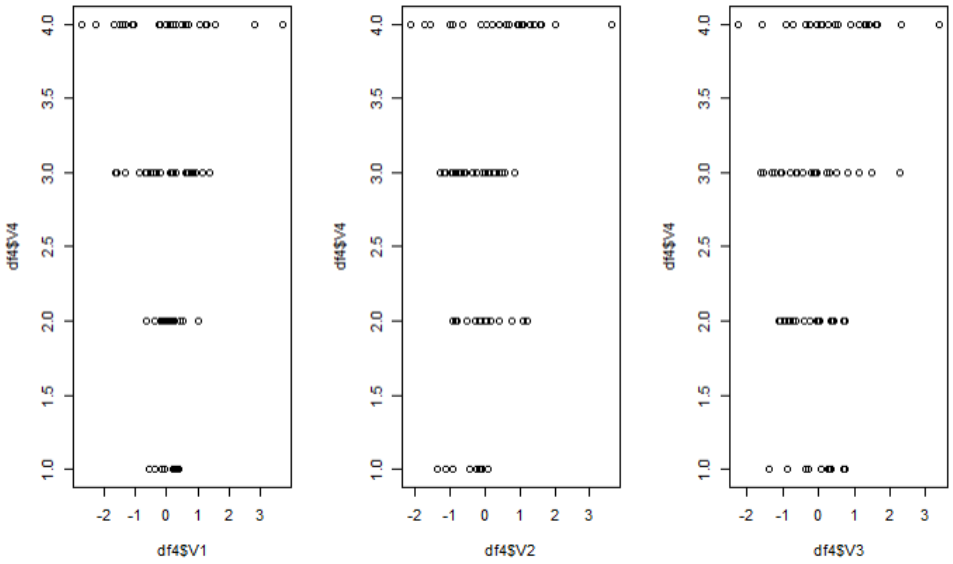
y와 SAVE의 1~4 dimensional sufficient predictors의 관계를 살펴보았다. Y가 모든 차원의 sufficient predictors와 관계가 거의 없는 것으로 보이므로 SAVE의 성능이 좋지 않다고 판단하였다. 따라서, order determination으로 과반수로 추정된 값인 4를 최종 dimension으로 선택한다.

<DR>



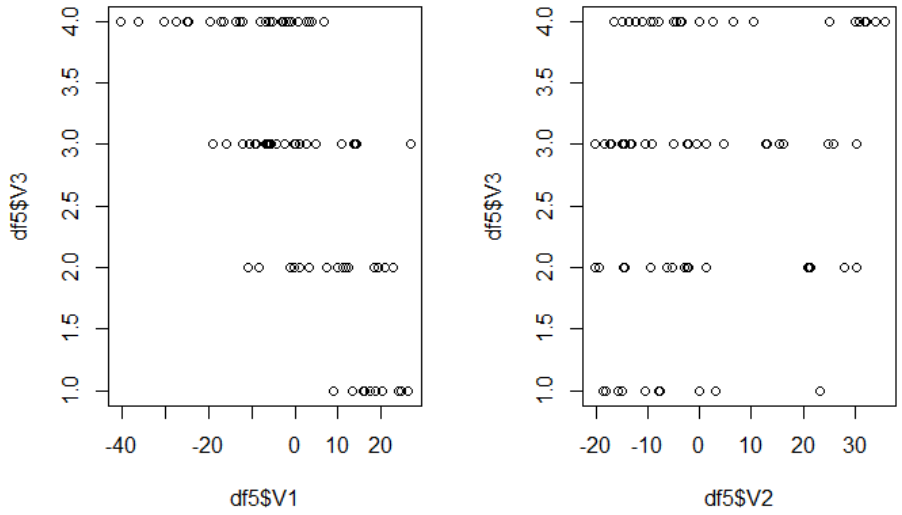
y와 DR의 1~6 dimensional sufficient predictors의 관계를 살펴보았다. y와 1-d sufficient predictor는 양의 관계가 나타나는데, 다른 차원에서는 관계가 거의 보이지 않으므로 DR의 최종 dimension은 1로 선택한다.

<PHD>



y와 PHD의 1~3 dimensional sufficient predictors의 관계를 살펴보았다. Y가 모든 차원의 sufficient predictors와 관계가 거의 없는 것으로 보이므로 PHD의 성능이 좋지 않다고 판단하였다. 따라서, order determination으로 과반수로 추정된 값인 3를 최종 dimension으로 선택한다.

<IHT>



y와 IHT의 1~2 dimensional sufficient predictors의 관계를 살펴보았다. y와 1-d sufficient predictor는 음의 관계가 나타나는데, 2-d에서는 관계가 거의 보이지 않으므로 IHT의 최종 dimension은 1로 선택한다.

따라서, y와의 관계까지 고려해 최종 선택된 dimension을 정리한 표이다.

	d
SIR	1
SAVE	4
DR	1
PHD	3
IHT	1

[최종 선택된 dimension]

3-3) Model Fitting

SDR Data와 Original Data로 모형을 학습시킨 후, classification accuracy를 구한다. Classification accuracy를 비교해보며 최적의 모형을 찾는다.

Classification을 위해 사용한 머신러닝 모형은 Decision Tree와 SVM이다.

각 방법론 별 20개의 test set에 대한 prediction accuracy와 confusion matrix를 아래 표로

정리하였다.

<Decision Tree>

Data	Prediction Accuracy	Confusion Matrix
Original (SDR 전)	0.5	Reference Prediction A B C D A 0 3 0 0 B 0 3 1 0 C 1 4 0 1 D 0 0 0 7
Sufficient Predictor_SIR	0.6	Reference Prediction 1 2 3 4 1 3 0 0 0 2 1 1 2 0 3 0 3 3 0 4 0 0 2 5
Sufficient Predictor_SAVE	0.4	Reference Prediction 1 2 3 4 1 1 0 1 1 2 0 0 2 2 3 0 0 3 3 4 0 0 3 4
Sufficient Predictor_DR	0.5	Reference Prediction 1 2 3 4 1 1 2 0 0 2 0 1 2 1 3 1 1 4 0 4 0 1 2 4
Sufficient Predictor_PHD	0.3	Reference Prediction 1 2 3 4 1 0 1 1 1 2 0 1 0 3 3 0 2 2 2 4 0 0 4 3
Sufficient Predictor_IHT	0.35	Reference Prediction 1 2 3 4 1 2 1 0 0 2 1 1 0 2 3 1 2 1 2 4 0 0 4 3

<SVM>

Data	Prediction Accuracy	Confusion Matrix
Original (SDR 전)	0.45	Reference Prediction A B C D A 0 3 0 0 B 0 1 3 0 C 0 2 3 1 D 0 0 2 5

Sufficient Predictor_SIR	0.55	Reference Prediction 1 2 3 4 1 0 3 0 0 2 0 1 3 0 3 0 1 5 0 4 0 0 2 5
Sufficient Predictor_SAVE	0.2	Reference Prediction 1 2 3 4 1 0 0 1 2 2 0 0 1 3 3 0 0 4 2 4 0 0 7 0
Sufficient Predictor_DR	0.5	Reference Prediction 1 2 3 4 1 1 2 0 0 2 0 1 2 1 3 1 1 4 0 4 0 1 2 4
Sufficient Predictor_PHD	0.3	Reference Prediction 1 2 3 4 1 0 1 1 1 2 0 1 1 2 3 0 1 2 3 4 0 0 4 3
Sufficient Predictor_IHT	0.35	Reference Prediction 1 2 3 4 1 1 2 0 0 2 1 1 2 0 3 0 3 2 1 4 0 0 4 3

3-4) 분석 결과

Decision Tree 와 SVM 모형 모두 Original Data 의 예측 정확도에 비해 SIR 의 예측 정확도가 높았다. 그러나, 다른 SDR 방법론을 사용한 Data 의 예측 정확도는 낮았는데, 이는 SDR sufficient predictor 와 Y 와의 관계를 보았을 때 SIR 의 Sufficient Predictor 만 Y 와 강한 관계가 있었기 때문이라고 판단하였다.

따라서, Body Performance Data 의 최적의 예측모형은 SIR 로 SDR 한 데이터를 Decision Tree 에 적합한 것이다.

4. 느낀점 및 보완점

1. 차원의 저주가 일어나게 하기 위해 데이터의 행을 줄여서 분석을 하였는데, 다음에는 더 변수(p)가 많은 데이터에서 충분차원축소를 적용해보고 싶다.
2. 수업시간에 배운 order determination method 이외에도 BIC-type Criteria 와 Ye-Weiss 방법론을 공부하고 실제 데이터에 적용해보며, 각 방법론의 특징을 비교해볼 수 있었다. 차원축소 후, y 와 sufficient predictor 의 관계를 살펴봤을 때 프로젝트에 사용한 데이터에선 Ladle Estimator 의 성능이 가장 좋은 것을 확인할 수 있었다.
3. 차원축소한 데이터를 이용해 머신러닝 모형을 만든 결과, SIR 을 이용해 차원축소 한 데이터에서만 original data 보다 좋은 결과가 나왔다. SDR 방법론에 따라 성능이 달라지는 것을 알 수 있었다. 다음엔 프로젝트에서 사용하지 못한 MAVE, OPG 를 사용해 차원축소를 해보고 SIR 의 결과와 비교해보고 싶다.