# 충분차원축소와 변수 선택

다변량 해석특론 프로젝트

212STG02 고정욱
212STG18 예지혜

# 프로젝트 주제 - 변수 선택의 중요성

SDR의 sufficient predictor는 보통 원 데이터의 **모든** 설명변수들의 선형 결합으로 이루어져 있다.

→ 해석이 복잡하다.

→ 중요 변수 식별이 어렵다.

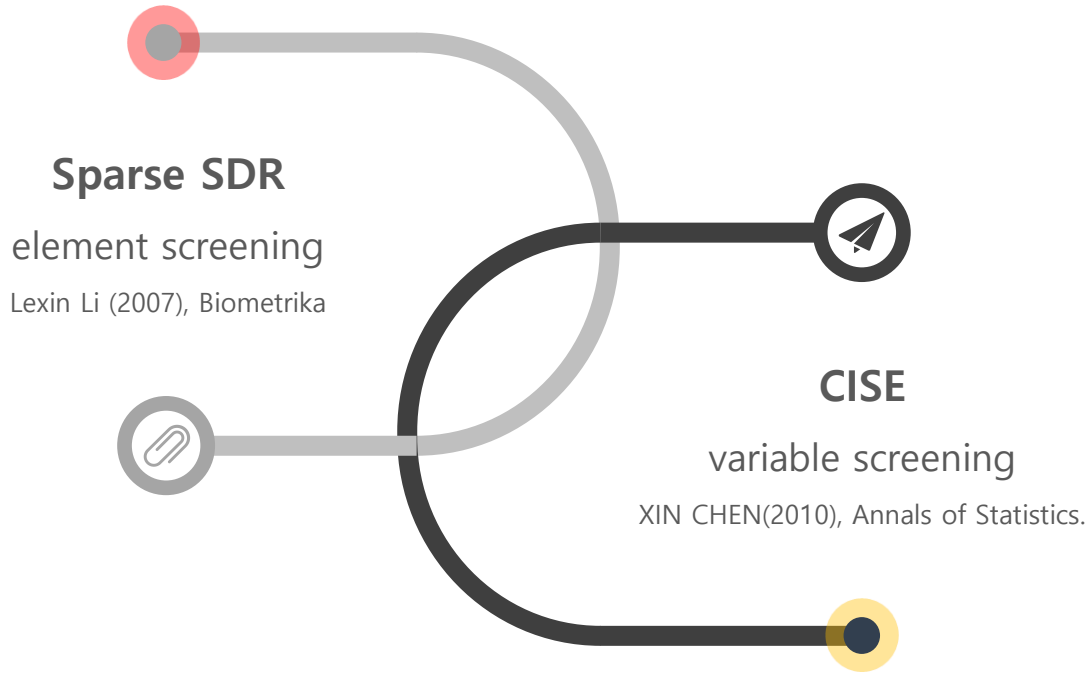
→ less efficient!

**무의미한 설명변수를 제거하면서 유의미한 설명변수들의 선형 결합으로 sufficient predictor를 추정하자.**

# 프로젝트 주제

| 년도 | 저자 | 저널 | 제목 |
|---|---|---|---|
| 2005 | Li, Cook and Nachtsheim | J. R. Stat. Soc. Ser. B Stat. Methodol. | Model-free variable selection |
| 2005 | Ni, Cook and Tsai | Biometrika | A note on shrinkage sliced inverse regression |
| 2006 | Zou, Hastie and Tibshirani | J. Comput. Graph. Statist | Sparse principal component analysis |
| 2006 | Li and Nachtsheim | Technometrics | Sparse sliced inverse regression |
| **2007** | **Li** | **Biometrika** | **Sparse sufficient dimension reduction** |
| 2008 | Zhou and He | Ann. Statist. | Dimension reduction based on constrained canonical correlation and variable filtering |
| 2009 | Leng and Wang | J. Comput. Graph. Statist. | On general adaptive sparse principal component analysis |
| **2010** | **Chen and Zou** | **Annals of Statistics** | **Coordinate-independent sparse sufficient dimension reduction and variable selection** |

# 프로젝트 주제


Sparse SDR
element screening
Lexin Li (2007), Biometrika
CISE
variable screening
XIN CHEN(2010), Annals of Statistics.

# Sparse SDR - 논문 소개

- Lexin Li (2007), Biometrika

Regression-type formulation of SDR

Shrinkage estimation (Lasso)

# Sparse SDR

1. sufficient dimension reduction formula

$$M\upsilon_i = \rho_i G\upsilon_i, \quad \text{for } i = 1, \ldots, p,$$

1. 1번 식을 regression-type 으로 재구성

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{p} \|G^{-1}m_i - \beta\beta^{\mathsf{T}}m_i\|_G^2,$$

1. sparse 추정치로 만들기 위해 L1 penalty 부여

$$\min_{\beta} \sum_{i=1}^{p} \|G^{-1}m_i - \beta\beta^{\mathsf{T}}m_i\|_G^2, \text{ subject to } \beta^{\mathsf{T}}G\beta = I_d, \text{ and } |\beta_j|_1 \leqslant \tau_j,$$

1. 최종 최적화 문제

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

# Sparse SDR - 알고리즘

최종 최적화 문제 :

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

**Step 1**
원하는 SDR estimator를 α의 초기 값으로 사용한다.

**Step 2**
고정된 α 값에 대해 β를 계산한다.

**Step 3**
고정된 β 값에 대해 α 값을 계산한다.

**Step 4**
Step 2와 Step 3를 β가 수렴할 때까지 반복한다.

**Step 5**
수렴한 β를 정규화 시킨다.

# Sparse SDR - Simulation

- true model : $$Y_1 = \text{sign}(\beta_1^T X)\log(|\beta_2^T X + 5|) + 0{\cdot}2\varepsilon$$

  (i) $\beta_1 = (1,1,1,1,0,\ldots,0)^T$, $\beta_2 = (0,\ldots,0,1,1,1,1)^T$

  (ii) $\beta_1 = (1,1,0.1,0.1,0,\ldots,0)^T$, $\beta_2 = (0,\ldots,0,0.1,0.1,1,1)^T$

  (iii) $\beta_1 = (1,\ldots,1,0,\ldots,0)^T$, $\beta_2 = (0,\ldots,0,1,\ldots,1)^T$

  $\Rightarrow$ n = 200, p = 20, true 0 : (i), (ii) 16개, (iii) 10개

- 평가 지표
  - 추정된 $\beta_1$, $\beta_2$의 number of zero components
  - $\beta X$의 추정치와 실제값의 절대값 상관계수
  - mean squared error
  - vector correlation coefficient (vcc)

# Sparse SDR - Simulation

(i) $\beta_1 = (1,1,1,1,0,\ldots,0)^T$, $\beta_2 = (0,\ldots,0,1,1,1,1)^T$ : 유의미 4개, 불필요 16개

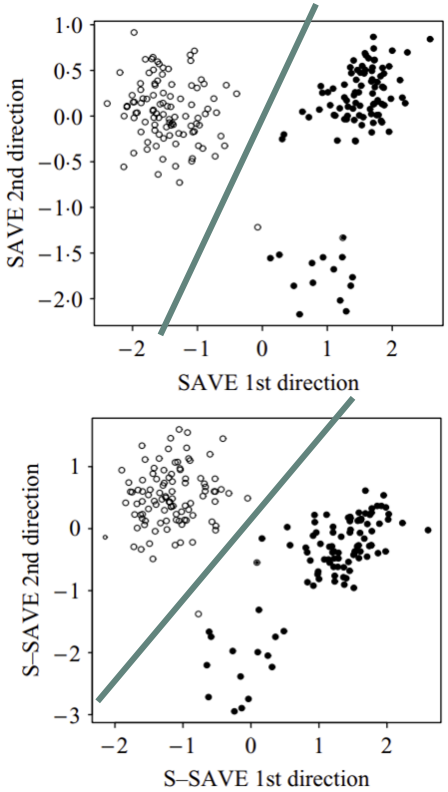
(ii) $\beta_1 = (1,1,0.1,0.1,0,\ldots,0)^T$, $\beta_2 = (0,\ldots,0,0.1,0.1,1,1)^T$ : 유의미 4개, 불필요 16개

(iii) $\beta_1 = (1,\ldots,1,0,\ldots,0)^T$, $\beta_2 = (0,\ldots,0,1,\ldots,1)^T$ : 유의미 10개, 불필요 10개

| | | $\hat{\beta}_1$ | | | $\hat{\beta}_2$ | | | $(\hat{\beta}_1, \hat{\beta}_2)$ |
|---|---|---|---|---|---|---|---|---|
| | | NUM | COR | MSE | NUM | COR | MSE | VCC |
| Case (i) | SIR | 0·000 | 0·926 | 1·604 | 0·000 | 0·911 | 1·245 | 0·934 |
| | S-SIR | 15·16 | 0·975 | 1·352 | 15·38 | 0·974 | 1·026 | 0·946 |
| Case (ii) | SIR | 0·000 | 0·884 | 0·551 | 0·000 | 0·856 | 0·544 | 0·932 |
| | S-SIR | 17·67 | 0·984 | 0·245 | 17·68 | 0·986 | 0·205 | 0·968 |
| Case (iii) | SIR | 0·000 | 0·916 | 4·793 | 0·000 | 0·942 | 4·168 | 0·917 |
| | S-SIR | 9·220 | 0·877 | 5·006 | 9·630 | 0·908 | 4·329 | 0·816 |

# Sparse SDR - 스위스 은행 위조지폐 데이터

| SAVE | |
|---|---|
| SD1 | −0.033 × Length − 0.200 × Left + 0.250 × Right<br>+ 0.594 × Bottom + 0.571 × Top − 0.466 × Diagonal |
| SD2 | −0.284 × Length − 0.055 × Left − 0.158 × Right<br>+ 0.505 × Bottom + 0.333 × Top + 0.725 × Diagonal |
| **Sparse SAVE** | |
| SD1 | 0 × Length + 0 × Left + 0 × Right<br>+ 0.785 × **Bottom** + 0.619 × **Top** + 0 × Diagonal |
| SD2 | 0 × Length + 0 × Left + 0 × Right<br>+ 0.400 × **Bottom** + 0 × Top + 0.917 × **Diagonal** |




첫번째 방향은 bottom과 top의 길이,

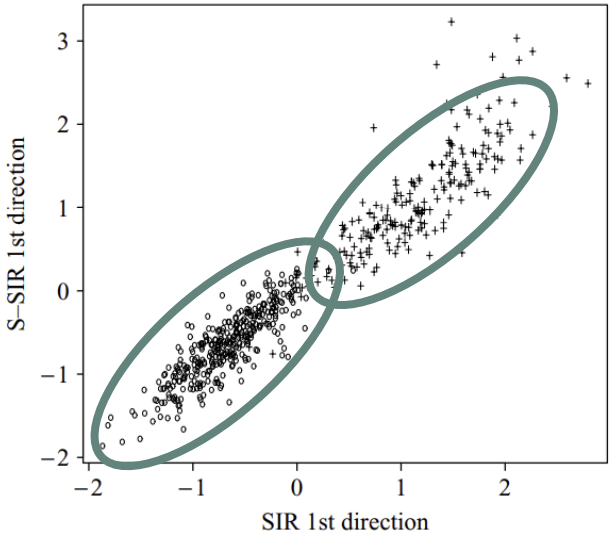
두번째 방향은 bottom과 diagonal의 길이를 의미

⇒ 여전히 위조 지폐를 잘 분류하며, 해석이 간단하다.

# Sparse SDR - Wisconsin breast cancer data

- 유방암 분류 문제
- n = 569, p = 30
- 두 방법 모두 잘 분류하며, 추정치 간 상관계수가 0.959로 매우 높다.

| SIR | |
|---|---|
| first direction | $(-0.508, 0.013, 0.382, 0.074, 0.001, -0.147, 0.074, 0.055, 0.002, 0.000, 0.080, -0.002, -0.030, -0.028, 0.031, 0.001\ -0.071, 0.043, 0.009, -0.013, 0.624, 0.029, -0.054, -0.381, 0.008, 0.007, 0.053, 0.020, 0.023, 0.051)^T$ |
| **Sparse SIR** | |
| first direction | $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, \mathbf{0.685}, \mathbf{0.294}, 0, 0, 0, 0, 0, \mathbf{0.667}, 0, 0)^T$ |



# Sparse SDR - discussion

## Sparse SDR의 장점

1. eigen decomposition을 사용하는 대부분의 차원 축소 방법론에 동일하게 적용할 수 있다.
2. 바이오 분야와 같이 sparse model이 많은 분야에서 유용하다.

## 한계점

eigen decomposition을 사용하지 않는 차원 축소 방법론에 바로 적용할 수 없다.

# CISE (COORDINATE-INDEPENDENT SPARSE ESTIMATION)

- *XIN CHEN(2010), Annals of Statistics.*

CISE (COORDINATE-INDEPENDENT SPARSE ESTIMATION)

sparse sufficient dimension reduction

screen out irrelevant and redundant variables efficiently

# CISE (COORDINATE-INDEPENDENT SPARSE ESTIMATION)

A new SDR penalty function

- invariant under orthogonal transformation
- targets the removal of row vectors from the basis matrix

$$\text{(2.2)} \qquad \widehat{\mathbf{V}} = \arg\min_{\mathbf{V}} \sum_{i=1}^{p} \|\mathbf{N}_n^{-1}\mathbf{m}_i - \mathbf{V}\mathbf{V}^T\mathbf{m}_i\|^2_{\mathbf{N}_n} \qquad \text{subject to } \mathbf{V}^T\mathbf{N}_n\mathbf{V} = \mathbf{I}_d,$$

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{V}}_s) = \min_{\boldsymbol{\alpha},\mathbf{V}} \left\{ \sum_{i=1}^{p} \|\mathbf{N}_n^{-1}\mathbf{m}_i - \boldsymbol{\alpha}\mathbf{V}^T\mathbf{m}_i\|^2_{\mathbf{N}_n} + \tau_2 \operatorname{tr}(\mathbf{V}^T\mathbf{N}_n\mathbf{V}) + \sum_{i=1}^{d} \tau_{1,j}\|\mathbf{V}_j\|_1 \right\},$$

$$\text{(2.7)} \qquad \tilde{\mathbf{V}} = \arg\min_{\mathbf{V}} \{-\operatorname{tr}(\mathbf{V}^T\mathbf{M}_n\mathbf{V}) + \rho(\mathbf{V})\} \qquad \text{subject to } \mathbf{V}^T\mathbf{N}_n\mathbf{V} = \mathbf{I}_d,$$

$$\rho(\mathbf{V}) = \sum_{i=1}^{p} \theta_i \|\mathbf{v}_i\|_2.$$

# CISE (COORDINATE-INDEPENDENT SPARSE ESTIMATION)

**A Coordinate-Independent penalty function.**

$$\phi(\mathbf{V}) = \sum_i \theta_i h_i(\mathbf{q}_i^T \mathbf{V}\mathbf{V}^T \mathbf{q}_i),$$

$$\rho(\mathbf{V}) = \sum_{i=1}^{p} \theta_i \|\mathbf{v}_i\|_2.$$

**Coordinate-Independent sparse Estimation**

$$\tilde{\mathbf{V}} = \arg\min_{\mathbf{V}} \{-\operatorname{tr}(\mathbf{V}^T \mathbf{M}_n \mathbf{V}) + \rho(\mathbf{V})\} \qquad \text{subject to } \mathbf{V}^T \mathbf{N}_n \mathbf{V} = \mathbf{I}_d, \tag{2.7}$$

**A Coordinate-Independent penalized procedure**

- incorporate many model-free and model-based SDR approaches into a simple and unified framework to implement variable selection within SDR.

# CISE - Simulation Studies

**4 Studies**

- 3 with forward reg models
- 1 using inverse reg model

**SDR methods**

- C3(alpha = 0.01 / 0.005)
- SSIR (BIC, RIC used to select the tuning parameters)
- CISE (use SIR and PFC to generate Mn and Nn for Cise selection) = CIS-SIR / CIS-PFC

# CISE - Simulation Studies

**simulation data**

- 2500 datasets / n=60, n = 120
- C3 = quadratic spline w/ 4 internal knots
- h=6 For SSIR
- calculate Mn in PFC, using f(y) = (|y|, y, y^2)T

**summary statistics - r1, r2, r3 (how well the methods select variables)**

- r1 = average fraction of nonzero rows of V (relevant predictors)
- r2 = average fraction of zero rows of V (irrelevant predictors)
- r3 = fraction of runs (both relevant and irrelevant predictors)

# CISE - Simulation Studies

**simulation results**

**STUDY 1**

with 24 predictors

X= (x1,...,x24)

true beta

b = (1,1,1,0,0,...,0)'

beta with 21 zero coefficients

$$y = x_1 + x_2 + x_3 + 0.5\epsilon,$$

TABLE 2
*Summary of Study 1*

| Method: | CIS-SIR | CIS-PFC | $C^3$ | | SSIR | |
|---|---|---|---|---|---|---|
| Criterion: | BIC | BIC | $\alpha = 0.01$ | $\alpha = 0.005$ | BIC | RIC |
| Sample size | | | $n = 60$ | | | |
| $r_1$ | 0.991 | 1.000 | 1.000 | 1.000 | 0.993 | 0.974 |
| $r_2$ | 0.999 | 1.000 | 0.999 | 0.999 | 0.997 | 0.999 |
| $r_3$ | 0.970 | 1.000 | 0.978 | 0.991 | 0.939 | 0.914 |
| Sample size | | | $n = 120$ | | | |
| $r_1$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $r_2$ | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 1.000 |
| $r_3$ | 1.000 | 1.000 | 1.000 | 1.000 | 0.994 | 1.000 |

# CISE - Simulation Studies

**oracle property?**

**CISE and C3**

- CISE is a unified method (can be applied to many popular sdr methods

  (PCA, PFC, SIR, SAVE, DR)

- C3 is based on one specified sdr method = canonical correlation
- on r3 measure, CISE >>> C3 ( only in Table 1, C3 did slightlyl better than CISE)
- simpler, easily implemented

# CISE - Simulation Studies

TABLE 4
*Summary of Study 3*

| **Method:** | **CIS-SIR** | **CIS-PFC** | $C^3$ | | **SSIR** | |
|---|---|---|---|---|---|---|
| **Criterion:** | **BIC** | **BIC** | $\alpha = 0.01$ | $\alpha = 0.005$ | **BIC** | **RIC** |
| Sample size | | | $n = 60$ | | | |
| $r_1$ | 0.789 | 0.906 | 0.770 | 0.742 | 0.934 | 0.888 |
| $r_2$ | 0.965 | 0.979 | 0.948 | 0.955 | 0.633 | 0.828 |
| $r_3$ | 0.344 | 0.588 | 0.229 | 0.226 | 0.000 | 0.004 |
| Sample size | | | $n = 120$ | | | |
| $r_1$ | 0.948 | 0.995 | 0.839 | 0.781 | 0.994 | 0.983 |
| $r_2$ | 0.992 | 0.998 | 0.956 | 0.963 | 0.664 | 0.865 |
| $r_3$ | 0.838 | 0.973 | 0.309 | 0.245 | 0.001 | 0.027 |

# CISE - Real Data Example with Boston Housing data

**Variable screening**

- 506 obs (census tracts)
- response y = median value of owner-occupied homes
- 13 predictors
- x1 - x13 (table)

* as suggestion of previous studies, remove obs with crime rate greater than 3.2 (used 374 obs)

- in PFC model **f** = (√y, y, y2)T / did not standardize since PFC is a scale-invariant method
- pick up 2 Directions to estimate the central subspace

source - http://lib.stat.cmu. edu/datasets/boston_corrected.txt.

# CISE - Real Data Example with Boston Housing data

## the estimated bases of the central subspace for all the considered methods

TABLE 6

*Estimated bases of the central subspace in Boston housing data*

| Method: | CIS-SIR | CIS-PFC | $C^3$ | SSIR-BIC | SSIR-RIC |
|---|---|---|---|---|---|
| $x_1$ | 0 0 | 0 0 | 0 0 | −0.050 −0.131 | −0.041 −0.123 |
| $x_2$ | −0.004 −0.047 | 0 0 | 0 0 | −0.001 −0.002 | −0.001 −0.001 |
| $x_3$ | 0 0 | 0 0 | 0 0 | 0.001 0.005 | 0 0 |
| $x_4$ | 0 0 | 0 0 | 0 0 | −0.033 0.020 | 0 0 |
| $x_5$ | 0 0 | 0 0 | 0 0 | 0.719 −0.882 | 0.543 −0.765 |
| $x_6$ | −0.999 0.034 | −0.999 0.034 | 0.962 −0.645 | −0.684 −0.448 | −0.834 −0.627 |
| $x_7$ | −0.008 −0.139 | −0.003 −0.077 | −0.174 −0.096 | 0.006 −0.001 | 0.005 −0.001 |
| $x_8$ | 0 0 | 0 0 | 0 0 | 0.082 −0.012 | 0.060 −0.010 |
| $x_9$ | 0 0 | 0 0 | 0 0 | −0.019 0.035 | −0.016 0.033 |
| $x_{10}$ | −0.001 −0.01 | −0.002 −0.035 | −0.166 0 | 0.001 −0.001 | 0.001 −0.001 |
| $x_{11}$ | 0.021 −0.361 | 0.018 −0.280 | −0.126 0 | 0.058 −0.033 | 0.055 −0.036 |
| $x_{12}$ | 0.001 0.011 | 0.002 0.035 | 0 0 | −0.000 0.000 | 0 0 |
| $x_{13}$ | −0.044 −0.920 | −0.040 −0.955 | 0 −0.758 | 0.014 −0.043 | 0.017 −0.059 |

# CISE - Real Data Example with Boston Housing data

**results**

- the coeff of C3 is based on a data-specific weighted original dataset (the coeff of other methods are based on the original dataset)
- As suggested by CIS-PFC, explanatory var x6,7,10,11,12,x13 would be important in explaining y.

# CISE - Real Data Example with Boston Housing data

**Bootstrap study**

Bootstrap procedure

- randomly choose w.r. 374 obs for y jointly w/ x6,7,10,11,12,x13
- separately for x1,2,3,4,5,8,9
- combine two bootstrap samples to make one complete bootstrap dataset

  forcing x1,2,3,4,5,8,9 to be irrelevant as with the analysis of orig. data.

- repeated 2500 times (M=2500)

# CISE - Real Data Example with Boston Housing data

**Bootstrap study results (w/o C3)**

- similar to the results in simulation studies
- CISE performed quite well

TABLE 7

*Variable selection in bootstrapping Boston housing data*

| **Method:** | **CIS-SIR** | **CIS-PFC** | **SSIR-BIC** | **SSIR-RIC** |
|---|---|---|---|---|
| $r_1$ | 0.947 | 0.962 | 0.963 | 0.877 |
| $r_2$ | 0.969 | 0.980 | 0.780 | 0.952 |
| $r_3$ | 0.550 | 0.672 | 0.118 | 0.264 |

# Discussion

**limitations**

- by the establishment of the oracle property = simple trace form - tr(VT MnV)
- proof in the Appendix can be extended to more general objective functions

**Further concerns**

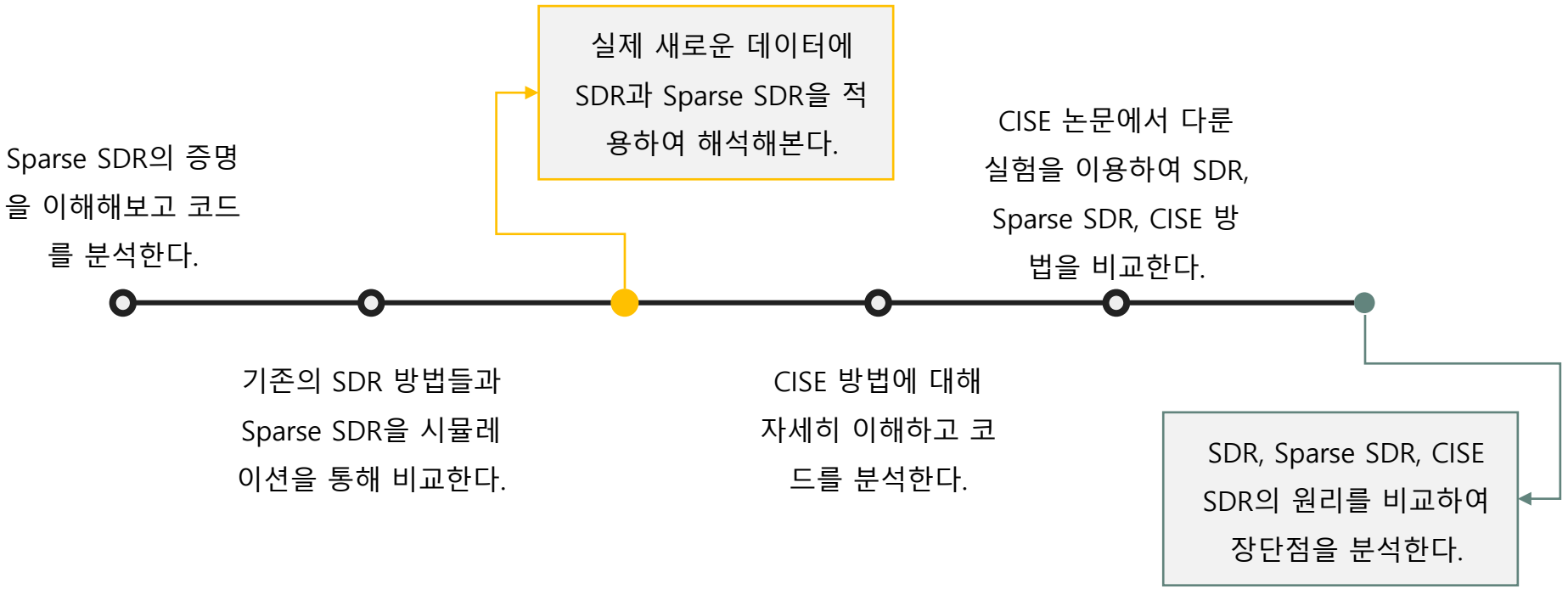
- whether CISE and its oracle property are still valid in high-dimensional setting in which p>n
- Nn usually takes the form of the marginal sample covariance matrix of x
- while, Mn depends on the specific method

  => how to choose Mn for variable selection is an imp issue and merits thorough investigation***

- BIC could identify the true sparse model well if the true model is included in the cand. set

  => also deserves further study

# 프로젝트 계획


Sparse SDR의 증명을 이해해보고 코드를 분석한다.
기존의 SDR 방법들과 Sparse SDR을 시뮬레이션을 통해 비교한다.
실제 새로운 데이터에 SDR과 Sparse SDR을 적용하여 해석해본다.
CISE 방법에 대해 자세히 이해하고 코드를 분석한다.
CISE 논문에서 다룬 실험을 이용하여 SDR, Sparse SDR, CISE 방법을 비교한다.
SDR, Sparse SDR, CISE SDR의 원리를 비교하여 장단점을 분석한다.

# 감사합니다