Ewha Womans University

Dimension Reduction in Regression with a Notion of Sufficiency

Jae Keun Yoo

November 6, 2015

Contents

Why dimension reduction in regression?

TOY regression of $Y|\mathbf{X} = (X_1, \ldots, X_p)^T$ is constructed as follows:

$$Y|X = X_1 + \varepsilon,$$

where $X_i (i = 1, \ldots, p) \overset{iid}{\sim} N(0,1) \perp\!\!\!\perp \varepsilon \sim N(0,1)$ and $\perp\!\!\!\perp$ indicates independence.

Predict Y in the following two ways.

Way 1: Fit usual multiple linear regression fit, and then do prediction.

Way 2: Fit $Y|\hat{\beta}^T\mathbf{X}$, where $\hat{\beta}$ is the estimated coefficient vector from the multiple linear regression of Way 1, and then do prediction (Later we will see that why $\hat{\beta}^T\mathbf{X}$ can replaces $\mathbf{X}$)

The difference between Ways 1-2 is clear.

The dimension of $\mathbf{X}$ in Way 1 is p.

The dimension of $\hat{\beta}^T\mathbf{X}$ in Way 2 is 1.

Prediction confidence intervals are summarized in the following figure with varying $p = 5, 10, 20, 30, 40, 50, 60, 70, 80, 90$. The sample sizes in all cases are $n = 100$.

As we can see, as p gets higher, the differences between the two prediction confidence intervals get larger.

The dimension reduction should be inevitable in high-dimensional regression.

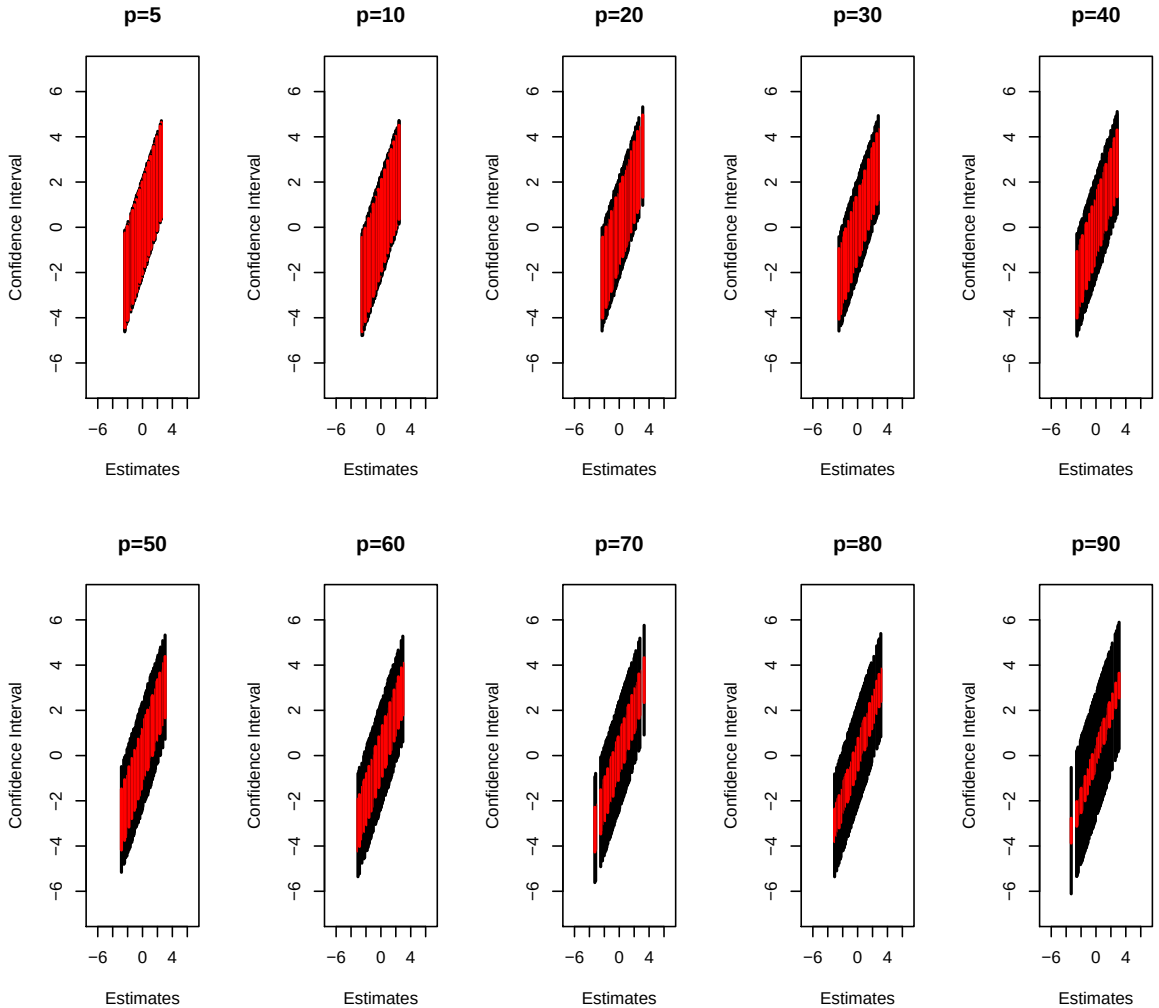

Figure: Prediction confidence interval in TOY regression; Black, usual multiple linear regression fit; Red, dimension reduction linear regression

If dimension reduction is done **without loss of information**, it is GOOD.

Sufficiency

a. no loss of information on parameters of interest
b. replacing the original n-dimensional data by lower-dimensional statistics
c. reduction in # of observations

Example
$\mathbf{x} = (x_1, \ldots, x_n) \overset{iid}{\sim} N(\mu, 1)$ and μ is the parameter of interest.
Then, without loss of information on μ, the original n-dimensional $(x_1, \ldots, x_n)$ are replaced by one-dimensional $\sum_{i=1}^{n} x$ or $\mathbf{1}^T\mathbf{x}$, where $\mathbf{1} = (1, \ldots, 1)^T$.

Philosophy of dimension reduction in regression

Consider a regression of $Y|\mathbf{X} = (X_1, \ldots, X_p)$.

Goal: **dimension reduction of $\mathbf{X} = (X_1, \ldots, X_p)$**, that is, reduction of the dimension of PREDICTORS, **without loss of information** in some sense

In regression, the primary interest is given to the conditional distribution of $Y|\mathbf{X}$, notated as $F_{Y|\mathbf{X}}(\cdot)$.

Want to find lower-dimensional function of $\mathbf{X}$, $g(\mathbf{X})$, than the original p-dimensional predictors $\mathbf{X}$ such that

$$F_{Y|\mathbf{X}}(\cdot) = F_{Y|g(\mathbf{X})}(\cdot).$$

Then, without loss of information on $Y|\mathbf{X}$, the original p-dimensional predictors $\mathbf{X}$ are replaced by lower-dimensional predictors $g(\mathbf{X})$.

$g(\cdot)$ is unknown and has too many possibilities.
$\Rightarrow$ try to make our life simpler.

As a form of $g(\mathbf{X})$, consider linear transformation of $\mathbf{X}$ such that $\mathbf{B}^T\mathbf{X}$, where $\mathbf{B} \in \mathbb{R}^{p\times q}$ with $q < p$.

Finally, we need to try to find $\mathbf{B}$ such that $F_{Y|\mathbf{X}} = F_{Y|\mathbf{B}^T\mathbf{X}}$.

This type of dimension reduction of $\mathbf{X}$ in regression is called **Sufficient Dimension Reduction**(SDR).

SDR in regression pursues a lower-dimensional linear projection of the original p-dimensional predictors without loss of information on $Y|\mathbf{X}$.

Keep in mind the following equivalence:

$$F_{Y|\mathbf{X}} = F_{Y|\mathbf{B}^T\mathbf{X}} \Leftrightarrow Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^T\mathbf{X}.$$

Examples of SDR

$\mathbf{X} = (X_1, \ldots, X_5)^T \overset{iid}{\sim} N(0,1)$. $\varepsilon \sim N(0,1) \perp\!\!\!\perp \mathbf{X}$.

Example 1: $Y|\mathbf{X} = \sum_{i=1}^{5} X_i + \varepsilon$.

$F_{Y|\mathbf{X}} = F_{Y|\sum_{i=1}^{5} X_i} = F_{Y|\mathbf{B}_1^T\mathbf{X}}$, $\mathbf{B}_1 = (1,1,1,1,1)^T$.

Example 2: $Y|\mathbf{X} = X_1(X_1 + X_2) + \varepsilon$.

$F_{Y|\mathbf{X}} = F_{Y|X_1,X_2} = F_{Y|\mathbf{B}_2^T\mathbf{X}}$, $\mathbf{B}_2 = \{(1,0,0,0,0),(0,1,0,0,0)\}^T$.

Example 3: $Y|\mathbf{X} = X_1 + \exp(X_2)\varepsilon$.

$F_{Y|\mathbf{X}} = F_{Y|X_1,X_2} = F_{Y|\mathbf{B}_3^T\mathbf{X}}$, $\mathbf{B}_3 = \{(1,0,0,0,0),(0,1,0,0,0)\}^T$.

Example 4: $Y|\mathbf{X} \sim B(m,p)$, where $p = \frac{\exp(\sum_{i=1}^{3} X_i)}{1+\exp(\sum_{i=1}^{3} X_i)}$.

$F_{Y|\mathbf{X}} = F_{Y|\sum_{i=1}^{5} X_i} = F_{Y|\mathbf{B}_4^T\mathbf{X}}$, $\mathbf{B}_4 = (1,1,1,0,0)^T$.

Dimension reduction subspace

Recall Examples 2-3.

We have $F_{Y|\mathbf{X}} = F_{Y|X_1,X_2}$.

Does the equivalence hold for $case_1 = (X_1, X_1 + X_2)$,
$\mathbf{B}_{case_1} = \{(1,0,0,0,0)(1,1,0,0,0)\}^T$?

How about the followings?
$case_2 = (X_2, X_1 + X_2)$, $\mathbf{B}_{case2} = \{(0,1,0,0,0)(1,1,0,0,0)\}^T$
$case_3 = (X_1/2, X_1 - X_2)$, $\mathbf{B}_{case3} = \{(0.5,0,0,0,0)(1,-1,0,0,0)\}^T$
$case_4 = (-X_1, X_1 + 2X_2)$, $\mathbf{B}_{case4} = \{(-1,0,0,0,0)(1,2,0,0,0)\}^T$

In all cases, we have $F_{Y|\mathbf{X}} = F_{Y|X_1,X_2} = F_{Y|case_i}$, $i = 1,2,3,4$.

There are more cases such that $F_{Y|\mathbf{X}} = F_{Y|X_1,X_2} = F_{Y|cases}$.

If so, what should we choose among that MANY?

Let us define $\mathcal{S}(\mathbf{B})$ as a subspace spanned by the columns of a $p \times q$ matrix $\mathbf{B}$. It is often called column subspace of $\mathbf{B}$.

It must be noted that $\mathcal{S}(\mathbf{B}_2) = \mathcal{S}(\mathbf{B}_{cases})$.

Considering column subspaces, all of them are the same.

We define **dimension reduction subspace** as follows:
Definition A subspace spanned by the columns of $\mathbf{B}$ such that $Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^T\mathbf{X}$ is called *dimension reduction subspace*.

Then, the choice problems turn out to be that of basis of $\mathcal{S}(\mathbf{B})$.

Any basis of $\mathcal{S}(\mathbf{B})$ should be fine, because they all spans the same subspace.

$\mathbf{B}^T\mathbf{X}$ is called *sufficient predictor*.

Central subspace, $\mathcal{S}_{Y|\mathbf{X}}$

Recall Example 1: $Y|\mathbf{X} = \sum_{i=1}^{5} X_i + \varepsilon$.

In the examples, a dimension reduction subspace can be easily seen, which is $\mathcal{S}\{(1,1,1,1,1)^T\}$.

How about $(\sum_{i=1}^{4} X_i, X_5)$, $\mathbf{B}_\dagger = \{(1,1,1,1,0),(0,0,0,0,1)^T\}$?

The matrix $\mathbf{B}_\dagger$ satisfies that $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}_\dagger^T\mathbf{X}$. So, $\mathcal{S}(\mathbf{B}_\dagger)$ is a dimension reduction subspace.

This means that there can exist many dimension reduction subspaces in a single regression problem.

What should be chosen among all possible dimension reduction subspaces?

The minimal subspace among all possible dimension reduction subspaces are desirable, because the minimal subspace still has full information on $Y|\mathbf{X}$.

Definition If the intersection of all possible dimension reduction subspace is a dimension reduction subspace, the intersection is called THE *central subspace*, $\mathcal{S}_{Y|\mathbf{X}}$.

Remarks on $\mathcal{S}_{Y|\mathbf{X}}$

Naturally, $\mathcal{S}_{Y|\mathbf{X}}$ is the primary target subspace in SDR.

$\mathcal{S}_{Y|\mathbf{X}}$ is unique and minimal, as long as $\mathcal{S}_{Y|\mathbf{X}}$ exits.

This means that $\mathcal{S}_{Y|\mathbf{X}}$ may not exists. There are certain conditions to guarantee the existence of $\mathcal{S}_{Y|\mathbf{X}}$. Rigorous investigation of the conditions is out of focus. Under soft conditions, $\mathcal{S}_{Y|\mathbf{X}}$ exists, so it is NEVER any issue. For details for it, read Cook (1998a; Section 6.6)

Central mean subspace, $\mathcal{S}_{E(Y|\mathbf{X})}$

Example 1: $Y|\mathbf{X} = \sum_{i=1}^{5} X_i + \varepsilon$.
Example 2: $Y|\mathbf{X} = X_1(X_1 + X_2) + \varepsilon$.
Example 4: $Y|\mathbf{X} \sim B(m, p)$, where $p = \frac{\exp(\sum_{i=1}^{3} X_i)}{1+\exp(\sum_{i=1}^{3} X_i)}$.

Consider the conditional mean of $E(Y|\mathbf{X})$:
Example 1: $E(Y|\mathbf{X}) = E(Y|\sum_{i=1}^{5} X_i)$
Example 2: $E(Y|\mathbf{X}) = E(Y|X_1, X_2)$
Example 4: $E(Y|\mathbf{X}) = E(Y|\sum_{i=1}^{3} X_i)$

That is, $\sum_{i=1}^{5} X_i$, (X_1, X_2) and $\sum_{i=1}^{3} X_i$ in Examples 1,2,4, respectively, are sufficient predictors for $E(Y|\mathbf{X})$ such that replacing the original predictors does not cause loss of information on $E(Y|\mathbf{X})$.

Often, the primary focus in many regression is given in $E(Y|\mathbf{X})$.

If so, $\mathcal{S}_{Y|\mathbf{X}}$ is too much information. In Examples 1,2,4, sufficient predictors for $E(Y|\mathbf{X})$ are enough to $\mathcal{S}_{Y|\mathbf{X}}$.

Why not seeing $E(Y|\mathbf{X})$ instead of $\mathcal{S}_{Y|\mathbf{X}}$?

Definition A subspace spanned by the columns of $\mathbf{B}$ such that $E(Y|\mathbf{X}) = E(Y|\mathbf{B}^T\mathbf{X})$ equivalently $Y \perp\!\!\!\perp E(Y|\mathbf{X})|\mathbf{B}^T\mathbf{X}$ is called *mean subspace*.

Definition If the intersection of all possible mean subspaces is a mean subspace, the intersection is called THE *central mean subspace*, $\mathcal{S}_{E(Y|\mathbf{X})}$.

The condition of the existence of $\mathcal{S}_{E(Y|\mathbf{X})}$ is the same as that of $\mathcal{S}_{Y|\mathbf{X}}$.

Central kth moment subspace $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$

Recall **Example 3**: $Y|\mathbf{X} = X_1 + \exp(X_2)\varepsilon$.

Note that $E(Y|\mathbf{X}) = E(Y|X_1)$.

To summarize the regression fully, X_1 is not sufficient due to $\text{var}(Y|\mathbf{X}) = \exp(2X_2)$.

The second moment is also needed.

Define that $M^{(k)}(Y|\mathbf{X}) = E[\{Y - E(Y|\mathbf{X})\}^k|\mathbf{X}]$. For $M^{(1)}(Y|\mathbf{X})$, $E(Y|\mathbf{X})$ will be used.

If $E(Y|\mathbf{X}) = E(Y|\mathbf{B}^T\mathbf{X})$ is not enough, how about considering higher conditional moments $M^{(i)}(Y|\mathbf{X}) = M^{(i)}(Y|\mathbf{B}^T\mathbf{X})$ for $i = 2, \ldots, k$?

In Example 3, considering the first two conditional moments of $Y|\mathbf{X}$ TOGHETHER, we have

$$E(Y|\mathbf{X}) = E(Y|X_1, X_2) \text{ and } M^{(2)}(Y|\mathbf{X}) = M^{(2)}(Y|X_1, X_2).$$

Definition A subspace spanned by the columns of $\mathbf{B}$ such that $M^{(i)}(Y|\mathbf{X}) = M^{(i)}(Y|\mathbf{B}^T\mathbf{X})$ equivalently $Y \perp\!\!\!\perp \{E(Y|\mathbf{X}), \ldots, M^{(k)}(Y|\mathbf{X})\}|\mathbf{B}^T\mathbf{X}$ is called *kth moment subspace*.

Definition If the intersection of all possible kth moment subspaces is a kth moment subspace, the intersection is called THE *central kth moment subspace*, $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$.

The condition of the existence of $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ is the same as that of $\mathcal{S}_{Y|\mathbf{X}}$.

Non-singular transformation of $\mathbf{X}$

Results:

Assume that $\mathbf{A}$ is a $p \times p$ non-singular matrix. Let $\mathbf{Z} = \mathbf{A}^T\mathbf{X}$. Then, $\mathcal{S}_{\mathbf{X}} = \mathbf{A}\mathcal{S}_{\mathbf{Z}}$, where $\mathcal{S}_{\mathbf{X}}$ represents one of $\mathcal{S}_{Y|\mathbf{X}}$, $\mathcal{S}_{E(Y|\mathbf{X})}$ or $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ and $\mathcal{S}_{\mathbf{Z}}$ does one of $\mathcal{S}_{Y|\mathbf{Z}}$, $\mathcal{S}_{E(Y|\mathbf{Z})}$, or $\mathcal{S}^{(k)}_{Y|\mathbf{Z}}$.

Suppose that $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}\{\mathbf{X} - E(\mathbf{X})\}$, where $\mathbf{\Sigma} = \text{cov}(\mathbf{X})$ and $\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}^{-1}$. Then $\mathbf{Z} \sim (0, \mathbf{I}_p)$.

In most SDR methodologies, $\mathbf{Z}$ is often used for computational stability, and then back-transformed to the original scale.

In this case, $\mathcal{S}_{\mathbf{X}} = \mathbf{\Sigma}^{-1/2}\mathcal{S}_{\mathbf{Z}}$.

It should be noted that the dimension of $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{S}_{\mathbf{Z}}$ are equal, although bases of $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{S}_{\mathbf{Z}}$ are different.

Relation of $\mathcal{S}_{E(Y|\mathbf{X})}$, $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$, $\mathcal{S}_{Y|\mathbf{X}}$

If $\mathcal{S}_{E(Y|\mathbf{X})}$, $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ and $\mathcal{S}_{Y|\mathbf{X}}$ exist for a regression of $Y|\mathbf{X}$, we have

$$\mathcal{S}_{E(Y|\mathbf{X})} \subseteq \mathcal{S}^{(k)}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}.$$

In SDR, $\mathcal{S}_{E(Y|\mathbf{X})}$, $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$, and $\mathcal{S}_{Y|\mathbf{X}}$ are the primary target subspaces. Hereafter, its dimension, often called *structural dimension*, and an orthonormal basis will be denoted as d and $\boldsymbol{\eta}$ (or $\boldsymbol{\eta}_z$ for $\mathcal{S}_{\mathbf{Z}}$).

CLASSICAL estimation methods

$\mathcal{S}_{Y|\mathbf{X}}$: sliced inverse regression/ sliced average variance estimation

$\mathcal{S}_{E(Y|\mathbf{X})}$: ordinary least squares/ principal Hessian direction/ iterative Hessian transformation

$\mathcal{S}^{(k)}_{Y|\mathbf{X}}$: polynomial ordinary least square; covariance method

Inference on $\mathcal{S}_{\mathbf{X}}$ has two components:

a. Estimation of the true dimension d

b. Estimation of a basis $\boldsymbol{\eta}$

For the inference, under certain conditions, not discussed here, normally construct kernel matrices $\mathbf{M} \in \mathbb{R}^{p \times p} \geq 0$ such that

$$\mathcal{S}(\mathbf{M}) = \mathcal{S}_{\mathbf{X}}.$$

Next, $\mathbf{M}$ is spectral-decomposed as follows:

$$\mathbf{M} = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^T,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$ and $\gamma_i^T \gamma_i = 1$ and $\gamma_i^T \gamma_j = 0$, $i \neq j$.

Estimation of d

The d is determined via testing a sequence of hypothesis:
Beginning with $m = 0$, test the hypothesis of

$$\mathrm{H}_0 : d = m \ \textit{versus} \ \mathrm{H}_1 : d > m.$$

If $\mathrm{H}_0 : d = m$ is rejected, increment m by 1 and redo the test. The test is stopped for the first time $\mathrm{H}_0 : d = m$ is not rejected, and set $\hat{d} = m$.

Test statistics is as follows:

$$\Lambda_m = n \sum_{m+1}^{p} \lambda_i, \ \ m = 0, 1, \ldots, (p-1).$$

Estimation of η

Once d is determined to $\hat{d}$, $\hat{\boldsymbol{\eta}}$ becomes a set of $(\gamma_1, \ldots, \gamma_{\hat{d}})$.

Sliced inverse regression

Recall d and $\boldsymbol{\eta}_z$, which are the true dimension and an orthonormal basis matrix for $\mathcal{S}_{Y|\mathbf{Z}}$.

Under certain conditions, not discussed here, we have

$$\mathcal{S}\{E(\mathbf{Z}|Y)\} = \mathcal{S}_{Y|\mathbf{Z}} \Leftrightarrow \boldsymbol{\Sigma}^{-1}\mathcal{S}\{E(\mathbf{X}|Y)\} = \mathcal{S}_{Y|\mathbf{X}}.$$

Inferring $\mathcal{S}_{Y|\mathbf{Z}}$ through constructing $E(\mathbf{Z}|Y)$ is called *sliced inverse regression* (SIR; Li, 1991).

Construction of $E(\mathbf{Z}|Y)$ in Population

If Y is categorical with h level:

Computation of $E(\mathbf{Z}|Y = s), s = 1, \ldots, h$, is straightforward.

If Y is continuous:

1. Categorize Y with h levels, called *slicing*, to try to have equal numbers of observations, that is, $Y \to \tilde{Y}$.
2. Then compute $E(\mathbf{Z}|\tilde{Y} = s), s = 1, \ldots, h$.

Sample structure of SIR: Algorithm

We assume that n iid data observations $\{(X_i, Y_i, i = 1, \ldots, n)\}$ throughout the rest of Tutorial.

1. Obtain $\tilde{Y}$ by slicing the range of Y into h non-overlapping intervals. Let n_s be the number of observations for $\tilde{Y} = s$.

2. Standardize $\mathbf{X}$:
$$\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}}), \ i = 1, \ldots, n.$$

3. Compute sample inverse mean with each slice $s = 1, \ldots, h$,:
$$\hat{E}(\mathbf{Z}|\tilde{Y} = s) = \frac{1}{n_s} \sum_{\tilde{Y}=s} \hat{\mathbf{Z}}_i (\in \mathbb{R}^p).$$

4. Construct sample covariance estimator
$$\hat{\mathbf{M}}_{SIR} = \hat{\text{cov}}\{E(\mathbf{Z}|\tilde{Y})\} = \sum_{s=1}^{h} \frac{n_s}{n} \hat{E}(\mathbf{Z}|\tilde{Y} = s)\hat{E}(\mathbf{Z}|\tilde{Y} = s)^T$$

Sample structure of SIR: Algorithm continued...

5. Perform spectral decomposition

$$\hat{\mathbf{M}}_{SIR} = \sum_{i=1}^{p} \hat{\lambda}_j \hat{\gamma}_j \hat{\gamma}_j^T,$$

where $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p \geq 0$.

6. Determine the structural dimension $d = \dim(\mathcal{S}_{Y|\mathbf{Z}})$. Let $\hat{d}$ denote an estimate of d.

7. Form an orthonormal basis estimate $(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}})$ for $\mathcal{S}_{Y|\mathbf{Z}}$.

8. Back-transform to obtain a sample basis estimate $\hat{\mathbf{\Sigma}}^{-1/2}(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}})$ for $\mathcal{S}_{Y|\mathbf{X}}$.

Dimension test in SIR

Test statistic

$$\hat{\Lambda}_m = n \sum_{j=m+1}^{p} \hat{\lambda}_j, \text{ for } m = 0, 1, \ldots, \min(p-1, h-1).$$

Asymptotic distribution of $\hat{\Lambda}_m$ (Bura and Cook, 2001)

$$\hat{\Lambda}_m \overset{d}{\sim} \sum_{k=1}^{(p-m)(h-m)} \omega_k \chi^2_k(1),$$

where $\chi^2_k(1)$s are independent χ^2 with one degree of freedom and h indicates the total number of slices.

The weights ω_k can be estimated consistently for use in practice

Implementation in R

```
> library(dr)
> data(ais)
```

Fitting SIR

```
> sir <- dr(LBM~log(SSF)+log(Wt)+log(Hg)+log(Ht)
+ log(WCC)+log(RCC)+log(Hc)+log(Ferr), nslice=3,
+ method="sir", data=ais)
> summary(sir)
```

For dimension test

```
> summary(sir)$test
```

For basis estimation with X-scale

```
> summary(sir)$evectors
```

Failure of SIR

$n = 100,\ p = 5,\ \eta = (1, -1, 0, 0, 0)^T/\sqrt{2},$

$$Y = (0.5\eta^T\mathbf{X})^2 + 0.1\varepsilon,$$

where $\mathbf{X} = (X_1, \ldots, X_5)^T \perp\!\!\!\perp \varepsilon \overset{iid}{\sim} N(0, 1)$.

SIR estimates: $\hat{\eta} = (0.29, 0.49, 0.53, 0.44, .0.45)^T$ (with $h = 10$).

Evaluations: $\text{corr}(\eta^T\mathbf{X}, \hat{\eta}^T\mathbf{X}) = 0.08$.

What happened?:
At population level:

$$E(\mathbf{X}|Y) = 0.$$

So, $\text{cov}\{E(\mathbf{X}|Y)\}$ can NOT provide information about $\mathcal{S}_{Y|\mathbf{X}}$. But, SAVE, which is followed, can do this.

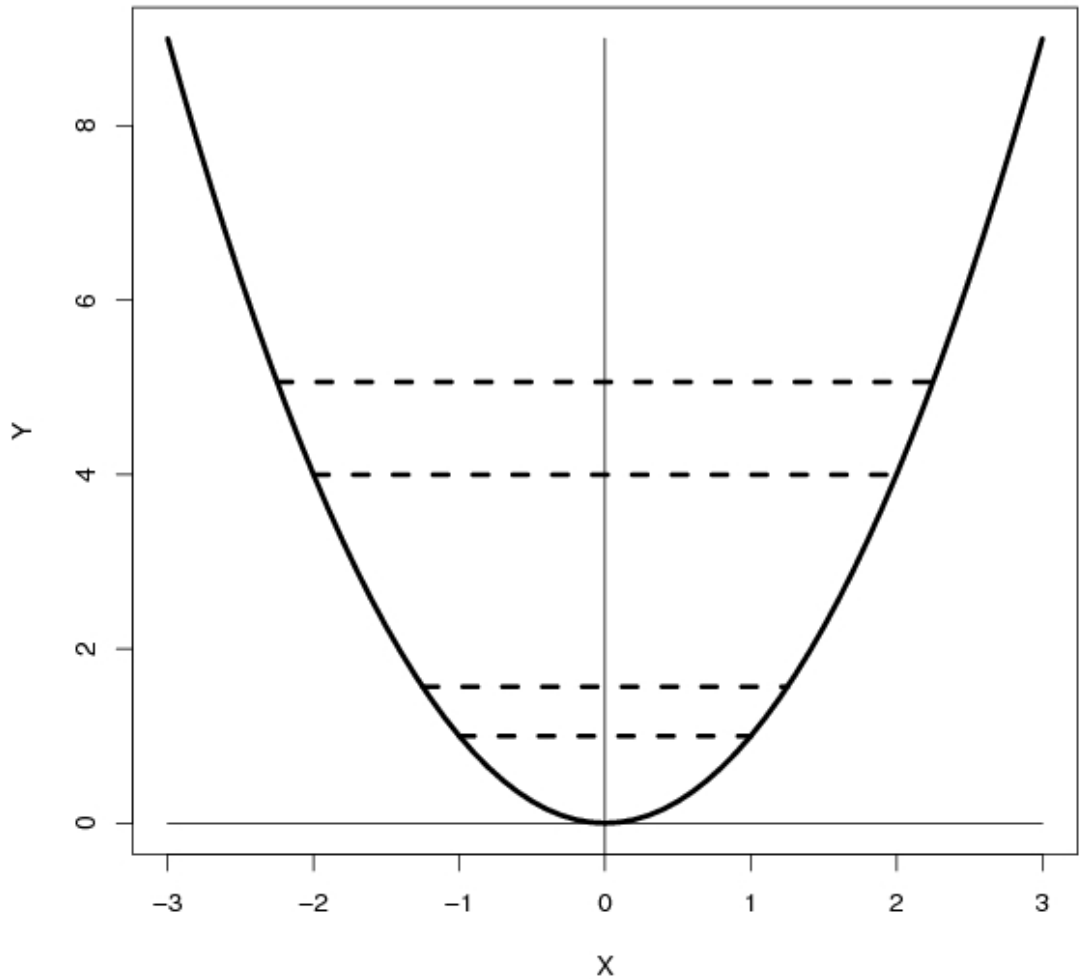

Figure: Symmetric regression when SIR fails

Sliced average variance estimation

Under certain conditions, not discussed here, we have

$$\mathcal{S}\{\mathbf{I}_p - \text{cov}(\mathbf{Z}|Y)\} = \mathcal{S}\{E\{\mathbf{I}_p - \text{cov}(\mathbf{Z}|Y)\}^2\} = \mathcal{S}_{Y|\mathbf{Z}}.$$

To recover $\mathcal{S}_{Y|\mathbf{Z}}$, a methodology to construct $E\{\mathbf{I}_p - \text{cov}(\mathbf{Z}|Y)\}^2$ is called *sliced average variance estimation* (SAVE; Cook and Weisberg, 1991)

Construction of $\text{cov}(\mathbf{Z}|Y)$ **in Population**
$\text{cov}(\mathbf{Z}|Y)$ is constructed by the same way of $E(\mathbf{Z}|Y)$.

Need to slice Y first, and then compute $\text{cov}(\mathbf{Z}|\tilde{Y})$.

Since $p \times p$ covariance matrices should be computed, relatively larger n_s should be considered per slice (less number of slices).

Sample structure of SAVE: Algorithm

1. Obtain $\tilde{Y}$ by slicing Y with h levels.
2. Standardize $\mathbf{X}$:

$$\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}}),\ i = 1, \ldots, n.$$

3. Compute sample inverse covariance with each slice:

$$\hat{\text{cov}}(\mathbf{Z}|\tilde{Y} = s) = \frac{1}{n_s} \sum_{\tilde{Y}_i \in \text{slice } s} (\hat{\mathbf{Z}}_{i \in s} - \bar{Z}_s)(\hat{\mathbf{Z}}_{i \in s} - \bar{Z}_s)^T.$$

4. Compute sample estimator of the kernel matrix:

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

Sample structure of SAVE: Algorithm continued...

5. Perform spectral decomposition:

$$\hat{\mathbf{M}}_{SAVE} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\gamma}_j \hat{\gamma}_j^T,$$

where $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p \geq 0$.

6. Determine the structural dimension $d = \dim(\mathcal{S}_{Y|\mathbf{Z}})$. Let $\hat{d}$ denote an estimate of d.

7. Form a sample basis estimate $(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}})$ for $\mathcal{S}_{Y|\mathbf{Z}}$.

8. Back-transform to obtain a sample basis estimate $\hat{\mathbf{\Sigma}}^{-1/2}(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}})$ for $\mathcal{S}_{Y|\mathbf{X}}$.

Inferences on d

1. FAIRLY MILD conditions, not discussed, are required.

2. Reformulizing $\mathbf{M}_{SAVE}$: $\mathbf{M}_{SAVE} = \sum_{s=1}^{h} \mathbf{A}_s^2$, where $\mathbf{A}_s = f_s^{1/2}(\mathbf{\Sigma}_s - \mathbf{I}_p)$ and $f_s = n_s/n$.

3. Let $\hat{\theta}_{p-m} = (\hat{\gamma}_{m+1}, \ldots, \hat{\gamma}_p)$:

$$T_n(\hat{\theta}_{p-m}) = n \sum_{k=1}^{h} \mathrm{tr}\{(\hat{\theta}_{p-m}^T \hat{\mathbf{A}}_k \hat{\theta}_{p-m})^2\}.$$

Inferences on d continued...

4. **Test statistics for** $\mathrm{H}_0 : d = m$: $T_n(\hat{\theta}_{p-m})$.

5. Under $\mathrm{H}_0 : d = m$,

$$T_n(\hat{\theta}_{p-m}) \overset{d}{\sim} \sum_{j=1}^{(p-m)(p-m+1)/2} \delta_j \chi^2_j(h-1),$$

   where $\chi^2_j(h-1)$s are independent χ^2 with $(h-1)$ degrees of freedom.

6. Refer Shao, Cook, and Weisberg (2007) on this matter.

Implementation in R

```
> library(dr)
```

Fitting SAVE

```
> save <- dr(LBM~log(SSF)+log(Wt)+log(Hg)+log(Ht)
+ log(WCC)+log(RCC)+log(Hc)+log(Ferr), nslice=4,
+ method="'save", data=ais)
> summary(save)
```

For dimension test

```
> summary(save)$test
```

For basis estimation with X-scale

```
> summary(save)$evectors
```

Relation between SIR and SAVE

a. In general, by Ye and Weiss (2003) and Li and Wang (2007),

$$\mathcal{S}(\mathbf{M}_{SIR}) \subseteq \mathcal{S}(\mathbf{M}_{SAVE}).$$

b. If Y is categorical, (Cook and Critchley, 2000)

$$\mathcal{S}(\mathbf{M}_{SAVE} = \mathcal{S}(\mathbf{M}_{SIR}) \oplus \mathcal{S}_{\Delta_{\mathbf{Z}|Y}},$$

where $\mathcal{S}_{\Delta_{\mathbf{Z}|Y}} = \mathcal{S}\{\text{cov}(\mathbf{Z}|Y = s+1) - \text{cov}(\mathbf{Z}|Y = s)\}$, $s = 1, \ldots, h-1$.

c. SIR tends to detect linear trend, while SAVE does nonlinear and quadratic trends.

principal Hessian direction

Supposing that $\mathcal{S}_{E(Y|\mathbf{Z})} = \mathcal{S}(\boldsymbol{\eta}_z)$, consider the $p \times p$ Hessian matrix of the regression function

$$H(\mathbf{Z}) = \frac{\partial^2 E(Y|\mathbf{Z})}{\partial \mathbf{Z} \partial \mathbf{Z}^T} = \boldsymbol{\eta}_z^T \frac{\partial E(Y|\boldsymbol{\eta}_z^T \mathbf{Z})}{\partial(\boldsymbol{\eta}_z^T \mathbf{Z})\partial(\mathbf{Z}^T \boldsymbol{\eta}_z)} \boldsymbol{\eta}_z.$$

Therefore, we have $\mathcal{S}[E\{H(\mathbf{Z})\}] = \mathcal{S}_{E(Y|\mathbf{Z})}$.

Let $\boldsymbol{\Sigma}_{yzz} = E[\{Y - E(Y)\}\mathbf{Z}\mathbf{Z}^T]$. Under normality of $\mathbf{Z}$, Li (1992) showed that $\mathcal{S}[E\{H(\mathbf{Z})\}] = \mathcal{S}(\boldsymbol{\Sigma}_{yzz})$.

Under weaker conditions (Cook, 1998b; not discussed here) than the normality, it can be established that $\mathcal{S}(\boldsymbol{\Sigma}_{yzz}) = \mathcal{S}_{E(Y|\mathbf{Z})}$.

This approach to recover $\mathcal{S}_{E(Y|\mathbf{Z})}$ through $\boldsymbol{\Sigma}_{yzz}$ is called *principal Hessian directions* (pHd; Li, 1992).

Residual-based pHd

The inference procedure associated with pHd can be greatly simplified if $\text{cov}(\mathbf{Z}, Y) = 0$.

Residual-based pHd (Cook, 1998b):

1. Obtain OLS residual $r = Y - E(Y) - \beta_z^T \mathbf{Z}$, so $\text{cov}(\mathbf{Z}, r) = 0$, where β_z is the ordinary least square coefficient vector in $Y|\mathbf{Z}$.
2. Construct the kernel matrix $\mathbf{\Sigma}_{rzz} = E(r\mathbf{Z}\mathbf{Z}^T)$.
3. Estimate $\mathcal{S}_{E(r|\mathbf{Z})}$ via spectral decomposition of $\mathbf{\Sigma}_{rzz}$.

Relation between $\mathcal{S}_{E(Y|\mathbf{Z})}$ and $\mathcal{S}_{E(r|\mathbf{Z})}$

Normally, we have $\mathcal{S}_{E(Y|\mathbf{Z})} = \mathcal{S}_{E(r|\mathbf{Z})} + \mathcal{S}(\beta_z)$.

The performance of $\{E(r\mathbf{Z}\mathbf{Z}^T), \beta_Z\}$ often turns out to be better than $E(y\mathbf{Z}\mathbf{Z}^T)$.

Sample structure of r-based pHd: Algorithm

1. Get the sample residuals

$$\hat{r}_i = Y_i - \bar{Y} - \hat{\beta}_z^T \hat{\mathbf{Z}}_i, \ i = 1, \ldots, n$$

where $\hat{\beta}_z = \frac{1}{n} \sum_{i=1}^n Y_i \hat{\mathbf{Z}}_i$.

2. Construct a sample version of $\mathbf{\Sigma}_{rzz}$

$$\hat{\mathbf{\Sigma}}_{rzz} = \frac{1}{n} \sum_i^n r_i \mathbf{Z}_i \mathbf{Z}_i^T.$$

3. Obtain eigensystem of $\hat{\mathbf{\Sigma}}_{rzz} \hat{\mathbf{\Sigma}}_{rzz}^T$:

$$(\hat{\gamma}_1, \ldots, \hat{\gamma}_p) \ \ \& \ \ (\hat{\lambda}_1 \geq, \ldots, \hat{\lambda}_p \geq 0).$$

Sample structure of r-based pHd: Algorithm continued ...

4. Sequential dimension tests under $\mathrm{H}_0 : d = m$ with

$$\frac{n\sum_{j=m+1}^{p}\hat{\lambda}_j}{2\hat{\mathrm{var}}(\hat{r})} \sim \frac{1}{2\,\mathrm{var}(r)}\sum_{j=1}^{(p-m)(p-m+1)/2}\omega_j\chi_j^2(1),$$

   where $\chi_j^2(1)$s stands for independent χ^2s with 1 degree of freedom.

5. $\hat{\mathcal{S}}_{E(r|\mathbf{Z})} = \mathcal{S}(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}})$ and
   $\hat{\mathcal{S}}_{E(Y|\mathbf{Z})} = \mathcal{S}\{\hat{\mathbf{\Sigma}}^{-1/2}(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}}, \hat{\beta}_z)\}$.

6. Refer Cook (1998b) for more about the asymptotics of the test statistics.

Implementation in R

```
> library(dr)
```

Fitting residual-based pHd

```
> rphd <- dr(LBM~log(SSF)+log(Wt)+log(Hg)+log(Ht)
+ log(WCC)+log(RCC)+log(Hc)+log(Ferr),
+ method="phdres", data=ais)
> summary(rphd)
```

For dimension test

```
> summary(rphd)$test
```

For basis estimation with X-scale

```
> summary(rphd)$evectors
```

Ordinary least square

It is well known that $\beta_z \in \mathcal{S}_{E(Y|\mathbf{Z})}$, where β_z is the ordinary least square (OLS) coefficient vector of $Y|\mathbf{Z}$.

Pros and Cons of OLS General definition of OLS:
$\beta = \mathbf{\Sigma}^{-1}\,\mathrm{cov}(\mathbf{X}, Y)$ (**Z-scale**: $\beta_z = E(Y\mathbf{Z})$).

Pros:

a. Very simple to use.
b. It does not imply that any associated model is true or even provides an adequate fit of the data!
c. In the TOY example, this OLS was used.

Cons

a. It can still miss the target in some situations (symmetric).
b. At most one direction! (if $d \geq 2$, not so good!).

Iterative Hessian transformation

Let's enhance the power of β_z.

Under mild conditions, not discussed here, we have

$$\mathcal{S}[\mathbf{H}_r = \{\beta_z, \mathbf{\Sigma}_{rzz}\beta_z, \ldots, \mathbf{\Sigma}_{rzz}^{(p-1)}\beta_z\}] = \mathcal{S}_{E(Y|\mathbf{Z})},$$

where $\mathbf{\Sigma}_{rzz}^k$ stands for multiplication of $\mathbf{\Sigma}_{rzz}$ itself k times.

This approach through $\mathbf{H}_r$ to recover $\mathcal{S}_{E(Y|\mathbf{Z})}$ is called *Iterative Hessian transformation* (IHT; Cook and Li, 2002).

(1) IHT is nothing but OLS acquired from transformed response variables.

(2) $\mathbf{H}_r$ clearly are more informative than β_z.

(3) For IHT to be practically useful, the regression should have linear trend . IHT fails in a symmetric regression again.

Sample structure of IHT: Algorithm

1. $\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}}),\ i = 1, \ldots, n.$
2. $\hat{r}_i = Y_i - \bar{Y} - \hat{\beta}_z^T \hat{\mathbf{Z}}_i,\ i = 1, \ldots, n,$ where $\hat{\beta}_z = \frac{1}{n}\sum_{i=1}^n Y_i \hat{Z}_i.$
3. Construct a sample version of $\mathbf{\Sigma}_{rzz}$

$$\hat{\mathbf{\Sigma}}_{rzz} = \frac{1}{n}\sum_i^n r_i \mathbf{Z}_i \mathbf{Z}_i^T.$$

4. Compute sample estimator of the kernel matrices:

$$\hat{\mathbf{H}}_{r(k)} = \{\hat{\beta}_z, \hat{\mathbf{\Sigma}}_{rzz}\hat{\beta}_z, \ldots, \hat{\mathbf{\Sigma}}_{rzz}^{(k)}\hat{\beta}_z\}\ \&$$

$$\hat{\mathbf{M}}_{IHTr}^{(k)} = \hat{\mathbf{H}}_{r(k)}\hat{\mathbf{H}}_{r(k)}^T.$$

Sample structure of IHT: Algorithm continued...

5. Perform spectral decomposition:

$$\hat{\mathbf{M}}_{IHTr}^{(k)} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\gamma}_j \hat{\gamma}_j^T,$$

where $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p \geq 0$.

6. Determine the structural dimension. Let $\hat{d}$ denote an estimate of d.

7. Form a sample basis estimate $(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}})$ for $\mathcal{S}_{E(Y|\mathbf{Z})}$.

8. Back-transform to obtain a sample basis estimate $\{\hat{\mathbf{\Sigma}}^{-1/2}(\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}})\}$ for $\mathcal{S}_{E(Y|\mathbf{X})}$.

Statistics for dimension determination

As usual, under $H_0 : d = m$, the sum of eigenvalues $\hat{\Lambda}_m = n \sum_{i=m+1}^{p} \hat{\lambda}_i$ of $\hat{\mathbf{M}}_{IHTr}^{(k)}$ is proposed as statistics for dimension determination.

Due to difficulty in derive the asymptotics of $\hat{\Lambda}_m$, a permutation test is employed, although Cook and Li (2004) derive it.

Pros of permutation tests in SDR:

a. It is easily implemented and does not require the asymptotics.

b. The test algorithm is applicable to all SDR methods.

Cons of permutation tests in SDR:

a. The tests require an addition condition of $(Y, \boldsymbol{\Gamma}_1^T \mathbf{Z}) \perp\!\!\!\perp \boldsymbol{\Gamma}_2^T \mathbf{Z}$.

b. It takes LONGER time than tests performed by the asymptotics.

Permutation tests: Algorithm

1. Compute the sample kernel matrix $\hat{\mathbf{M}}_{IHTr}$, and, under $H_0 : d = m$, obtain $\hat{\Lambda}_m$ and partition eigenvector matrices

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

2. Construct the vectors $\hat{V}_i = \hat{\boldsymbol{\Gamma}}_1^T \hat{\mathbf{Z}}_i$ and $\hat{U}_i = \hat{\boldsymbol{\Gamma}}_2^T \hat{\mathbf{Z}}_i$,
3. Randomly permute the indices i of the $\hat{U}_i$ to obtain the permuted set $\hat{U}_i^*$.
4. Construct the test statistic $\hat{\Lambda}_m^*$ based on the original data Y_i and $\hat{V}_i$ along with the permuted data $\hat{U}_i^*$.
5. Repeat steps (3) - (4) N times, where N is the total number of permutations. The p-value of the hypothesis testing is the fraction of $\hat{\Lambda}_m^*$ that exceed $\hat{\Lambda}_m$.

Permutation tests of IHT in R

```
> source("IHT.txt"); source("IHT_perm.txt")
> data(ais); attach(ais)
> X.dat<-data.frame(lSSF=log(SSF),lWt=log(Wt),lHg=log(Hg),
+ lHt=log(Ht),lWCC=log(WCC),lRCC=log(RCC),
+ lHc=log(Hc),lFerr=log(Ferr))
> X<-as.matrix(X.dat); y<-LBM
```

For dimension test

```
> IHT.perm <- perm.test.IHT(X, y, 2, npermute=1000)
> IHT.perm$summary
```

For basis estimation with X-scale

```
> IHT.fit <- IHT(X,y,2)
> IHT.fit$B.x
```

Polynomial Least Squares: covariance method (cov_k)

As usual, we consider $\mathbf{Z}$-scale.

Under certain conditions, again not discussed here,

$$\mathcal{S}\{E(\mathbf{Z}Y^\ell), \ell = 1, \ldots, k\} = \mathcal{S}^{(k)}_{Y|\mathbf{Z}}.$$

Letting $\mathbf{M}_{\text{cov}_k} = \{E(\mathbf{Z}Y), E(\mathbf{Z}Y^2), \ldots, E(\mathbf{Z}Y^k)\}$, we have $\mathcal{S}(\mathbf{M}_{\text{cov}_k}) = \mathcal{S}^{(k)}_{Y|\mathbf{Z}}$.

This approach through $\mathbf{M}_{\text{cov}_k}$ to recover $\mathcal{S}^{(k)}_{Y|\mathbf{Z}}$ is called *Polynomial Least Squares* or *covariance method* (cov_k; Yin and Cook, 2002).

Some comments on covariance method

a. $E(\mathbf{Z}Y^\ell)$ is nothing but the OLS coefficient vector on $Y^\ell|\mathbf{Z}$.

b. In Example 2: $Y|\mathbf{X} = X_1(X_1 + X_2) + \varepsilon$,
$E(\mathbf{Z}Y^2) = (0, 2, 0, 0, 0)^T$.

c. In Example 3: $Y|\mathbf{X} = X_1 + \exp(X_2)\varepsilon$,
$E(\mathbf{Z}Y^2) = (2, 2e^2, 0, 0, 0)^T$.

d. Let $U = \{Y - E(Y)\}/\text{var}(Y)$ and let $\mathbf{M}_{U,\text{cov}_k}$ represent cov_k for $U|\mathbf{Z}$. Then we have $\mathcal{S}(\mathbf{M}_{U,\text{cov}_k}) = \mathcal{S}(\mathbf{M}_{\text{cov}_k})$.

e. In practice, to avoid numerical instability, the standardized response U is used over Y.

A view of relation between cov_k and SIR

Results

Let $\theta = \mathbf{\Sigma}^{-1} E\{t(Y)\mathbf{X}\}$, where $t(Y)$ denotes a function of Y with $E\{t(Y)\} = 0$. Under certain conditions, we have $\mathcal{S}(\theta) \subseteq \mathcal{S}_{Y|\mathbf{X}}$.

Implications

a. SIR takes $t(Y) = J_s(Y)$, slice indicators.
b. cov_k takes $t(Y) = Y^s$, $s = 1, \ldots, k$.

Roles of k

a. For any value of k, $\mathcal{S}(\mathbf{M}_{\text{cov}_k}) \subseteq \mathcal{S}_{E(\mathbf{Z}|Y)}$.
b. $\mathcal{S}(\mathbf{M}_{\text{cov}_k})$ provides lower bounds on $\mathcal{S}_{E(\mathbf{Z}|Y)}$, that is SIR.
c. We may wish to conduct analysis with different numbers of slices, focusing first on low order covariances (few slices) and eventually on high order covariances (many slices) to gain a more detailed understanding of $Y|\mathbf{Z}$.

Sample structure of cov_k: Algorithm

Define $Y(k) \in \mathbb{R}^k = (Y, Y^2, \ldots, Y^k)^T$.

1. Transform Y to $\hat{U} = (Y - \bar{Y})/\hat{\sigma}_Y^2$.

2. Suppose that we need to test $\text{H}_0 : d = m$. Fix $k = m + 1$. We do not have to go beyond $m + 1$. Then construct a sample version $\hat{\mathbf{M}}_{\hat{U},\text{cov}_{m+1}}$ of $\mathbf{M}_{\hat{U},\text{cov}_{m+1}}$.

3. Spectral decompose $\hat{\mathbf{M}}_{\hat{U},\text{cov}_{m+1}} \hat{\mathbf{M}}^T_{\hat{U},\text{cov}_{m+1}}$:

$$\{\hat{\mathbf{\Gamma}}_1 = (\hat{\gamma}_1, \ldots, \hat{\gamma}_m), \hat{\mathbf{\Gamma}}_2 = (\hat{\gamma}_{m+1} \ldots \hat{\gamma}_p)\} \quad \& \quad (\hat{\lambda}_1 \geq, \ldots, \hat{\lambda}_p \geq 0).$$

Sample structure of cov_k: Algorithm continued...

4. Permutation-test $\text{H}_0 : d = m$ with $[\{\hat{U}(m+1), \mathbf{Z}\hat{\boldsymbol{\Gamma}}_1\}; \mathbf{Z}\hat{\boldsymbol{\Gamma}}_2]$ to have p-values.

5. If H_0 is rejected, add one to m and repeat (1)-(4).

6. Starting $m = 0$, repeat (1)-(5) until $\text{H}_0 : d = m$ is not rejected for the first time.

7. Then set $\hat{d} = m$ and have $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_m)$ from $\hat{\mathbf{M}}_{\hat{U}, \text{cov}_m}$.

8. $\hat{\mathcal{S}}^{(k)}_{Y|\mathbf{Z}} = \mathcal{S}(\hat{\boldsymbol{\gamma}})$ and $\hat{\mathcal{S}}^{(k)}_{Y|\mathbf{X}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\mathcal{S}(\hat{\boldsymbol{\gamma}})$.

Permutation tests of cov_k in R

```
> source("covk.txt"); source("covk_perm.txt")
> data(ais); attach(ais)
> X.dat<-data.frame(lSSF=log(SSF),lWt=log(Wt),lHg=log(Hg),
+ lHt=log(Ht),lWCC=log(WCC),lRCC=log(RCC),
+ lHc=log(Hc),lFerr=log(Ferr))
> X<-as.matrix(X.dat); y<-LBM
```

For dimension test

```
> covk.perm <- perm.test.covk(X, y, 2, npermute=1000)
> covk.perm$summary
```

For basis estimation with X-scale

```
> covk.fit <- covk(X,y,2)
> covk.fit$B.x
```

Setup for numerical study

a. $\mathbf{X} = (X_1, \ldots, X_5)^T \overset{iid}{\sim} N(0,1) \perp\!\!\!\perp \varepsilon \sim N(0,1)$.

b. $n = 100$ and the total number of iterations are 1000.

c. To summarize the dimension estimation, the percentages of $\hat{d} = m, m = 0, 1, 2$ and $\hat{d} \geq 3$ are computed.

d. To measure how close $\mathcal{S}(\boldsymbol{\eta})$ and $\mathcal{S}(\hat{\boldsymbol{\eta}})$, Trace Correlation Distance (TRD) are computed:

$$TRD = 1 - r,$$

where $r = \sqrt{\frac{1}{2} trace(\{\boldsymbol{\eta}(\boldsymbol{\eta}^T\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^T\}\{\hat{\boldsymbol{\eta}}(\hat{\boldsymbol{\eta}}^T\hat{\boldsymbol{\eta}})^{-1}\hat{\boldsymbol{\eta}}^T\})}$.

Smaller TRD, better estimation of $\boldsymbol{\eta}$.

Example 2: $Y|\mathbf{X} = X_1(X_1 + X_2) + \varepsilon$

$Y|\mathbf{X} = Y|(X_1, X_2)$ so that $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{E(Y|\mathbf{X})}$ with $d = 2$.

Table: Dimension test summary: Percentage of $\hat{d}$

	SIR5	SIR10	SAVE2	SAVE4	rpHd	IHT	COVk
$d = 0$	4.9	9.6	52.8	100	17.2	0.0	0.2
$d = 1$	72.4	70.4	41.7	0.00	31.8	25.1	16.0
$d = 2$	21.2	19.2	5.0	0.00	48.0	72.5	77.4
$d \geq 3$	1.5	0.8	0.5	0.00	2.8	2.4	6.4

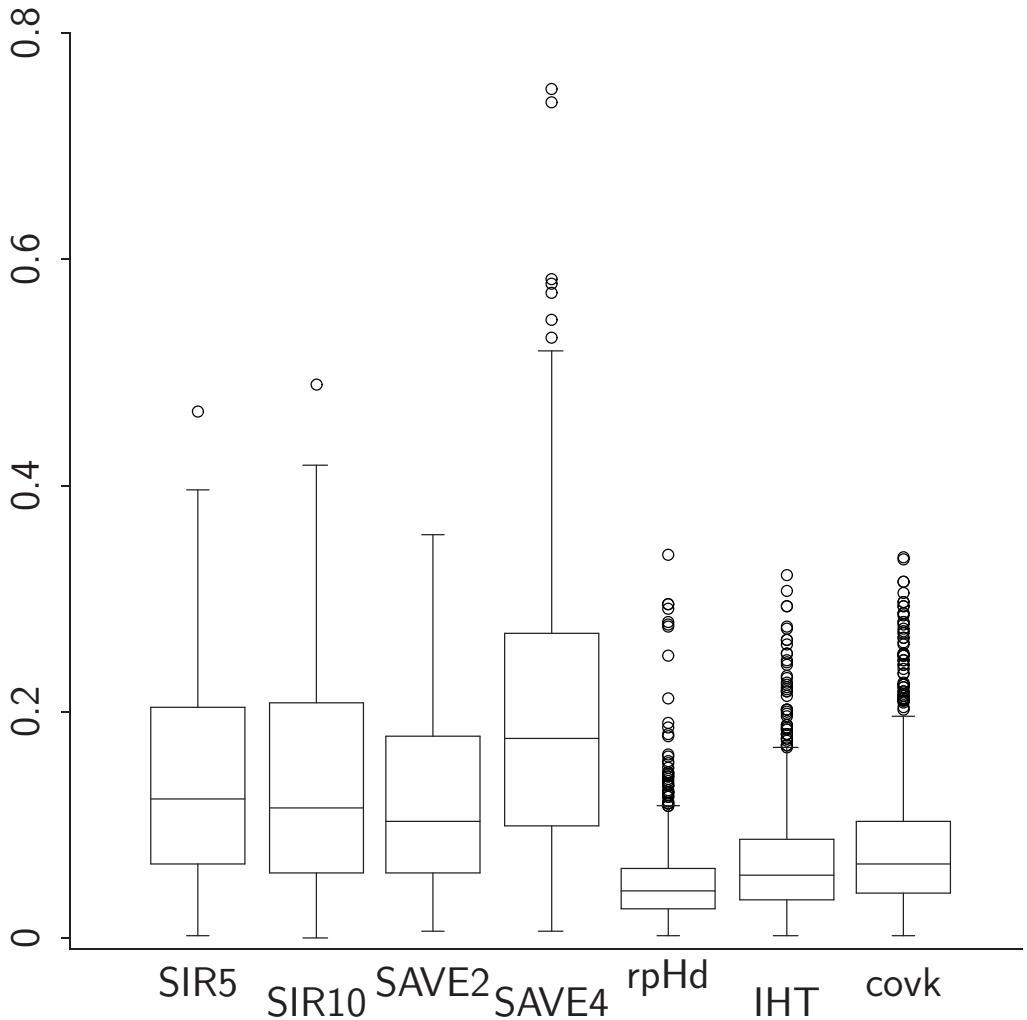

Figure: Direction estimation: trace correlation distance

Example 3: $Y|\mathbf{X} = X_1 + \exp(X_2)\varepsilon$

$Y|\mathbf{X} = Y|(X_1, X_2)$ so that $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}^{(2)}_{Y|\mathbf{X}}$ with $d = 2$.

$\mathcal{S}_{E(Y|\mathbf{X})}$ detects only X_1 and is spanned by $(1, 0, 0, 0, 0)^T$.

Table: Dimension test summary: Percentage of $\hat{d}$

	SIR5	SIR10	SAVE2	SAVE4	rpHd	IHT	COVk
$d = 0$	0.1	0.6	79.3	76.9	99.4	33.2	10.4
$d = 1$	30.2	33.7	19.8	21.9	0.5	44.5	1.6
$d = 2$	66.1	63.3	0.9	1.2	0.1	21.6	48.0
$d \geq 3$	3.6	2.4	0.00	0.00	0.00	0.7	40.0

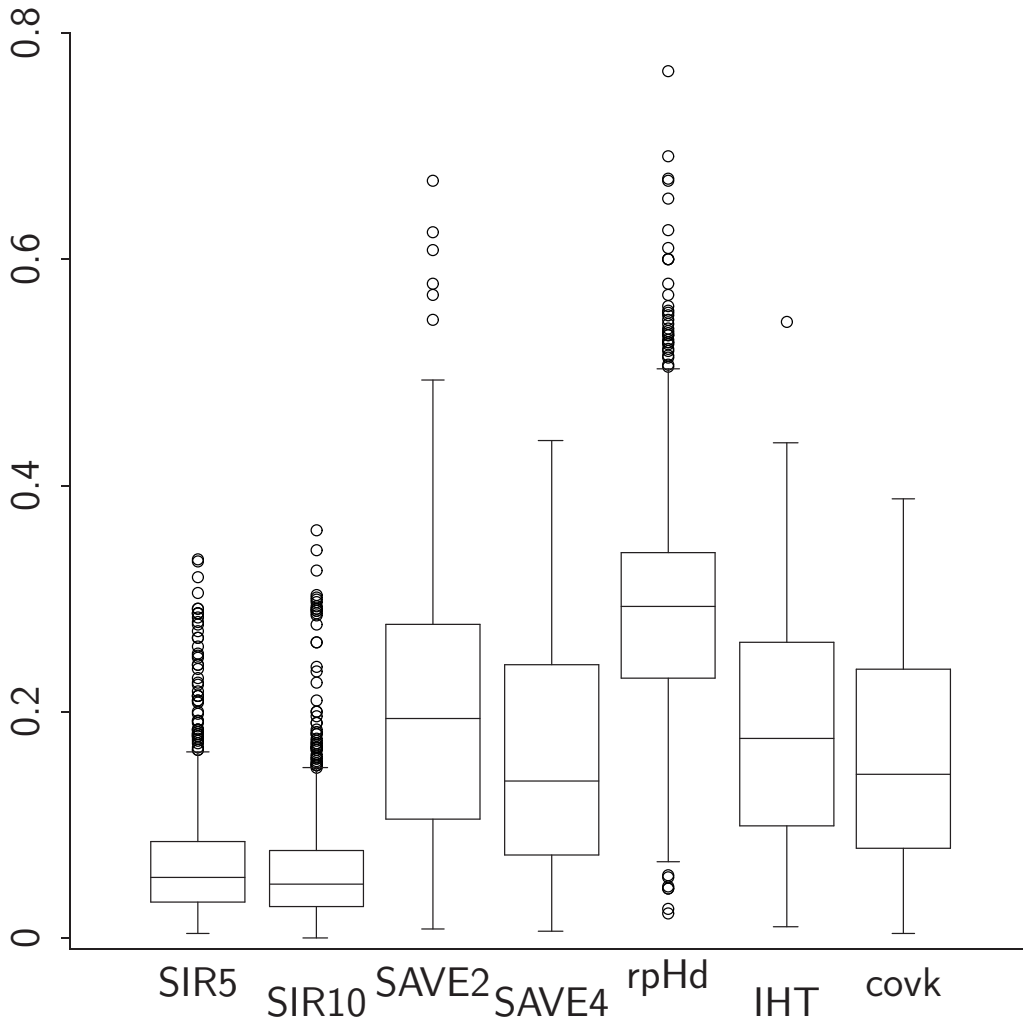

Figure: Direction estimation: trace correlation distance

Dimension reduction summary of AIS data

Dimension estimation

	SIR	SAVE	rpHd	IHT(2)	covk(2)
0D vs >= 1D	0.00	0.00	0.01	0.00	0.00
1D vs >= 2D	0.45	0.04	0.01	0.11	0.21
2D vs >= 3D	N/A	0.12	0.23	0.96	N/A

Basis estimation

	log(SSF)	log(Wt)	log(Hg)	log(Ht)	log(WCC)	log(RCC)	log(Hc)	log(Ferr)
SIR	0.16	-0.77	-0.34	-0.31	-0.03	-0.17	0.36	-0.02
SAVE	0.16	-0.83	-0.43	-0.06	-0.01	-0.17	0.24	-0.00
rpHd	0.04	-0.60	0.36	-0.22	-0.03	0.29	-0.61	-0.02
IHT	0.04	-0.62	0.36	-0.15	-0.02	0.29	-0.61	0.02
COVk	0.14	-0.92	-0.08	0.29	0.02	-0.18	0.16	-0.01

Correlation matrix of the collection of sufficient predictors

	SIR	SAVE	rpHd	IHT	COVk
SIR	1.0000000	0.9961036	0.8957398	0.8995848	0.9273931
SAVE		1.0000000	0.8698745	0.8751002	0.9176177
rpHd			1.0000000	0.9997367	0.9550561
IHT				1.0000000	0.9611313
COVk					1.0000000

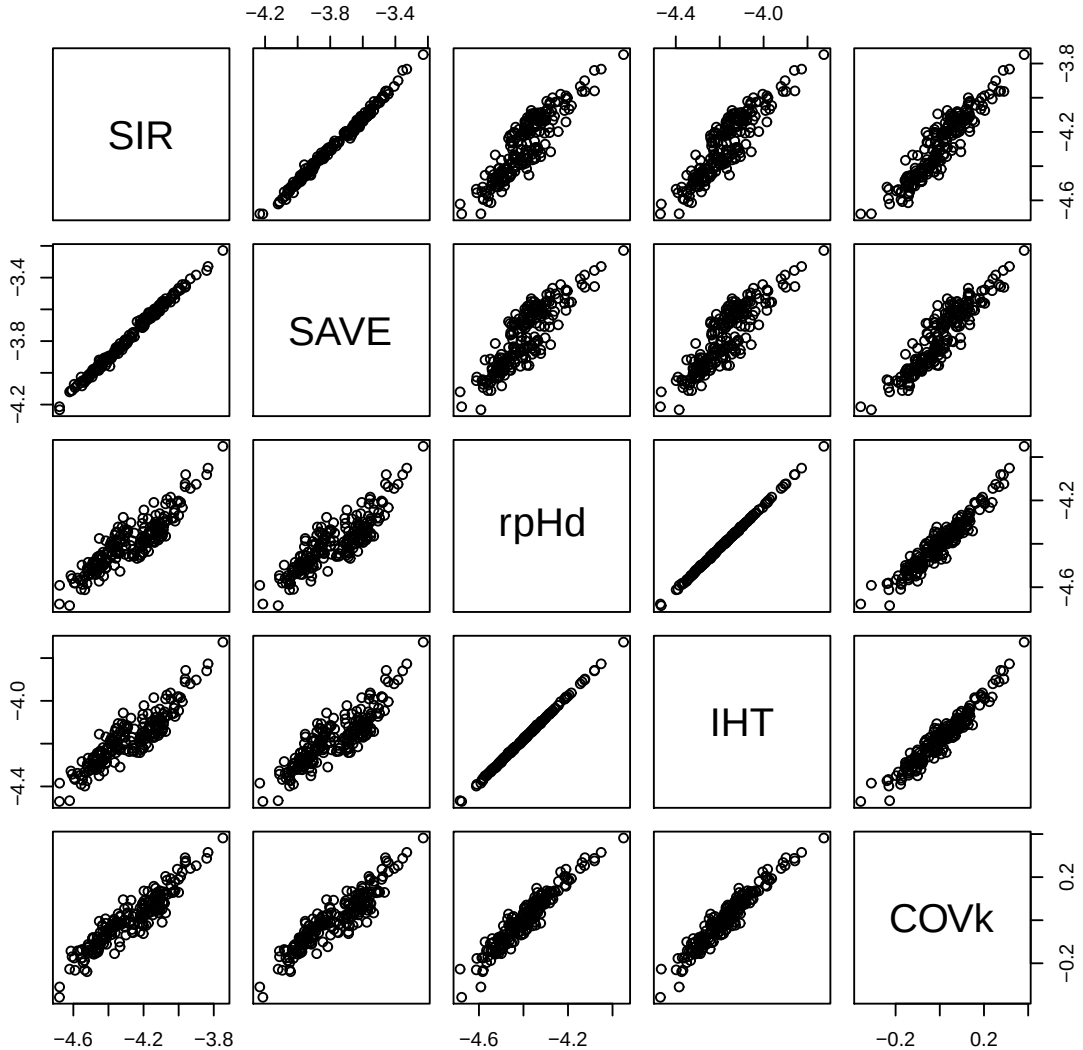

Figure: Scatterplot matrix of the collection of sufficient predictors

Seeded dimension reduction

Have you noticed that the four classical SDR methods require the inverse of the sample covariance matrix $\hat{\mathbf{\Sigma}}$?

If $p > n$, ironically, the classical SDR methods are not plausible in practice!

Let's do something on this, starting with assuming that there exist a matrix $\boldsymbol{\nu} \in \mathbb{R}^{p \times q}$ such that

$$\mathbf{\Sigma}^{-1}\mathcal{S}(\boldsymbol{\nu}) = \mathcal{S}_{Y|\mathbf{X}} \Leftrightarrow \mathcal{S}(\boldsymbol{\nu}) = \mathbf{\Sigma}\mathcal{S}_{Y|\mathbf{X}}.$$

One important requirement for a choice of $\boldsymbol{\nu}$ is that its sample version should be constructed without inverting $\hat{\mathbf{\Sigma}}$.

Let $\mathbf{P}_{\mathbf{R}(\mathbf{\Sigma})} = \mathbf{R}(\mathbf{R}^T\mathbf{\Sigma}\mathbf{R})^{-1}\mathbf{R}^T\mathbf{\Sigma}$ be an orthogonal projection operator onto $\mathcal{S}(\mathbf{R} \in \mathbb{R}^{p \times m})$ such that $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}(\mathbf{R})$.

Since $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}(\mathbf{R})$, the projection of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu} \in \mathcal{S}_{Y|\mathbf{X}}$ onto $\mathcal{S}(\mathbf{R})$ returns itself. So we have:

$$\begin{aligned} \boldsymbol{\Sigma}^{-1}\boldsymbol{\nu} &= \mathbf{P}_{\mathbf{R}(\boldsymbol{\Sigma})}\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu} = \mathbf{R}(\mathbf{R}^T\boldsymbol{\Sigma}\mathbf{R})^{-1}\mathbf{R}^T\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu} \\ &= \mathbf{R}(\mathbf{R}^T\boldsymbol{\Sigma}\mathbf{R})^{-1}\mathbf{R}^T\boldsymbol{\nu}. \end{aligned}$$

THE crucially notable thing is that $\boldsymbol{\Sigma}^{-1}$ is not required in $(\mathbf{R}^T\boldsymbol{\Sigma}\mathbf{R})^{-1}\mathbf{R}^T\boldsymbol{\nu}$, but

$$\mathcal{S}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}) = \mathcal{S}\{(\mathbf{R}^T\boldsymbol{\Sigma}\mathbf{R})^{-1}\mathbf{R}^T\boldsymbol{\nu}\} = \mathcal{S}_{Y|\mathbf{X}}.$$

Instead, the inversion of $(\mathbf{R}^T\boldsymbol{\Sigma}\mathbf{R}) \in \mathbb{R}^{m\times m}$ is required. If $m < n$, hopefully, $m << n$, then it can be invertible.

Then, here come two questions.

a. Choices of $\boldsymbol{\nu}$?
b. Construction of $\mathbf{R}$?

Choices of ν

The matrix ν,called **seed matrix**, is selected among the following quantities:

a. When Y is categorical, $E(\mathbf{X}|Y=y) - E(\mathbf{X}) \in \mathbf{\Sigma}\mathcal{S}_{Y|\mathbf{X}}$ for $y = 1, \ldots, h$.

b. When Y is continuous, Y is categorized into h levels, saying $\tilde{Y}$. Then $E\{\mathbf{X}|\tilde{Y}=\tilde{y}\} - E(\mathbf{X}) \in \mathbf{\Sigma}\mathcal{S}_{Y|\mathbf{X}}$ for $\tilde{y} = 1, \ldots, h$.

c. $\mathrm{cov}(\mathbf{X}, Y) \in \mathbf{\Sigma}\mathcal{S}_{Y|\mathbf{X}}$.

d. $\mathrm{cov}\{\mathbf{X}, U(k)\} \in \mathbf{\Sigma}\mathcal{S}_{Y|\mathbf{X}}$, where $U = \{Y - E(Y)\}/\sqrt{\mathrm{var}(Y)}$ and $U(k) = (U, U^2, \ldots, U^k)$, $k = 1, 2, \ldots$.

Did you notice it? The choices of (a) and (b) correspond to SIR. And (c) and (d) do OLS and PoLS.

Construction of $\mathbf{R}$

The matrix $\mathbf{R}$ is needed to be constructed, so that its column spans a subspace large enough to contain $\mathcal{S}_{Y|\mathbf{X}}$ but reasonably estimable from data.

For this, iterative projections of $\boldsymbol{\nu}$ onto $\boldsymbol{\Sigma}$ is suggested:

$$\mathbf{R}_u \equiv (\boldsymbol{\nu}, \boldsymbol{\Sigma}\boldsymbol{\nu}, \ldots, \boldsymbol{\Sigma}^{u-1}\boldsymbol{\nu}) \in \mathbb{R}^{p\times(q\times u)}, \quad u = 1, 2, \ldots u^*.$$

The letter u above is called a termination index of projections. It is noted that $\mathcal{S}(\mathbf{R}_{u-1}) \subseteq \mathcal{S}(\mathbf{R}_u)$ for any $u \geq 2$.

It is important to make a proper choice of the termination index u, small enough to guarantee that $\mathcal{S}(\mathbf{R}_u) = \mathcal{S}_{Y|\mathbf{X}}$.

Trust me. We can do it through some ways, which will not be discussed here.

The sufficient dimension reduction through the successive projection of seed matrices $\boldsymbol{\nu}$ onto $\boldsymbol{\Sigma}$ is called *seeded dimension reduction*.

In practice, first, $\boldsymbol{\Sigma}$ and $\boldsymbol{\nu}$ are replaced by their sample quantities and then a proper value of u, saying u^*, is determined.

Then the sample version $\hat{\mathbf{R}}_{u^*}$ is constructed, and finally $\mathcal{S}_{Y|\mathbf{X}}$ is estimated by a subspace spanned by the columns of

$$\hat{\boldsymbol{\eta}} = \hat{\mathbf{R}}_{u^*}(\hat{\mathbf{R}}_{u^*}^T \hat{\boldsymbol{\Sigma}} \hat{\mathbf{R}}_{u^*})^{-1} \hat{\mathbf{R}}_{u^*}^T \hat{\boldsymbol{\nu}},$$

where $\hat{\mathbf{R}}_{u^*} \in \mathbb{R}^{p \times (q \times u^*)}$.

a. Tests of predictor effects (Cook, 2004)
Testing $H_0 : \mathbf{P_H}\mathcal{S}_\mathbf{X} = \mathcal{O}_p$, where $\mathbf{H}$ is a $p \times h$ user-selected predictor matrix. For example, if $\mathbf{P_H}\mathcal{S}_\mathbf{X} = \mathcal{O}_p$ holds for $\mathbf{H} = (1, 0, \ldots, 0)$, then the first coordinate variate X_1 in $\mathbf{X}$ do not contribute to $\mathcal{S}_\mathbf{X}$. Then X_1 can be eliminated.

b. Partial sufficient dimension reduction (Chiaromonte, Cook and Li, 2002; Li, Cook, Chiaromonte, 2003)
Inference on $\boldsymbol{\eta}$ such that $Y \perp\!\!\!\perp \mathbf{X} | (\boldsymbol{\eta}^T\mathbf{X}, W)$, where W is a c-level categorical predictor.

c. Minimum discrepancy approach (Cook and Ni, 2005)
Inference on $\boldsymbol{\eta}$ with arguments $\hat{\mathbf{B}}$ that minimizes the objective function, $F_m(\mathbf{B}, \mathbf{C})$, under $H_0 : d = m$: $\text{argmin}_{\mathbf{B},\mathbf{C}} F_m(\mathbf{B}, \mathbf{C})$

$$F_m(\mathbf{B}, \mathbf{C}) = \{\text{vec}(\hat{\mathbf{M}}) - \text{vec}(\mathbf{BC})\}^T \mathbf{V}_n \{\text{vec}(\hat{\mathbf{M}}) - \text{vec}(\mathbf{BC})\},$$

where $\mathbf{B}$ is a $p \times m$ matrix and $\mathbf{C}$ is a $m \times r$ matrix.

d. Sufficient dimension reduction in multivariate regression
Inference of $\mathcal{S}_{\mathbf{X}}$ under multivariate regression. Visit my web.
`http://home.ewha.ac.kr/~yjkstat/publication.html`

e. Response dimension reduction in multivaraite regression (Yoo and Cook, 2008)
Dimension reduction of predictors so far. Why not dimension reduction of responses under multivariate regression?

f. Sparse and regularized sufficient dimension reduction (Li, 2007; Li, Cook and Tsai, 2007; Li and Yin, 2008)
As title indicated ...

g. Model-based sufficient dimension reduction (Cook,2007)
Inference on $\boldsymbol{\eta}$ under the following semi-parametric model:

$$\mathbf{X}_y = \mu + \boldsymbol{\eta}\boldsymbol{\beta}\mathbf{f}_y + \sigma\boldsymbol{\varepsilon}, \ \textit{with } \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Omega}).$$

On-going research by myself

a. Model-based response dimension reduction

b. Sufficient dimension reduction in semi-supervised data

The tutorial slides and codes used in this tutorial can be found below:

`http://home.ewha.ac.kr/~yjkstat/Tutorial_Slide.pdf`

`http://home.ewha.ac.kr/~yjkstat/Tutorial_Codes.zip`

- **Existence of the Central Subspace and Required Conditions**
  a. Cook, R. D. (1998a). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, Wiley, New York.

- **Sliced Inverse Regression**
  a. Li, K-C. (1991). Sliced inverse regression for dimension reduction with discussions. Journal of the American Statistical Association, 86, 316–327.
  b. Bura, E and Cook, R. D. (2001). Extending sliced inverse regression: the weighted chi-squared test. Journal of the American Statistical Association, 96, 996–1003.

- **Sliced Average Variance Estimation**
  a. Cook, R. D., and Weisberg, S. (1991). Discussion of Li (1991). Journal of the American Statistical Association, 86, 328–332.
  b. Cook, R. D. (2000). SAVE: a method for dimension reduction and graphics in regression. Communications in statistics - Theory and methods, 29, 2109-2121.
  c. Shao,Y., Cook, R. D. and Weisberg, S. (2007). Marginal tests with sliced average variance estimation. Biometrika, 94, 285–296.

- **Relation between SIR and SAVE**
  a. Cook, R. D., and Critchley, F. (2000). Identifying regression outliers and mixtures graphically. Journal of the American Statistical Association, 95, 781–794.
  b. 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.
  c. Li, B. and Wang, S. (2007). On directional regression for dimension reduction. Journal of the American Statistical Association, 102, 997–1008.

- **Iterative Hessian Transformation**
  a. Cook, R. D., and Li, B. (2002). Dimension reduction for the conditional mean in regression. The Annals of Statistics, 30, 455–474.
  b. Cook, R.D., and Li, B. (2004). Determining the dimension of iterative Hessian transformation. Annals of Statistics, 32, 2501–2531.

- **Permutation Tests in Sufficient Dimension Reduction**
  a. Cook, R. D., and Weisberg, S. (1991). Discussion of Li (1991). Journal of the American Statistical Association, 86, 328–332.
  b. Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional k-th moment in regression. Journal of the Royal Statistical Society, Series B, 64, 159–176.

- **principal Hessian direction**
  a. Li, K-C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Steinos Lemma. Journal of the American Statistical Association, 87, 1025–1039.
  b. Cook, R. D. (1998b). Principal Hessian directions revisited. Journal of the American Statistical Association, 93, 84–94.
  c. Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. Annals of Statistics, 9, 1135–1151.

- **Polynomial Least Square; covariance method**
  a. Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional k-th moment in regression. Journal of the Royal Statistical Society, Series B, 64, 159–176.

- **Seeded Dimension Reduction**
  a. Cook, R. D., Li, B., and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. Biometrika, 94, 569–584.

- **Tests of Predictor Effects**
  a. Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. Annals of Statistics. 32, 1062-92.

- **Partial Sufficient Dimension Reduction**
  a. Chiaromonte, F., Cook, R. D. and Li, B.(2002). Sufficient dimensions reduction in regressions with categorical predictors. Annals of Statistics. 30, 475–497.
  b. Li, B., Cook, R.D., Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. Annals of Statistics. 31, 1636–1668.

- **Minimum Discrepancy Approach**
  a. Cook, R.D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. Journal of the American Statistical Association. 100, 410–428.

- **Response Dimension Reduction in Multivaraite Regression**
  a. JK. Yoo and Cook, R. D. (2008). Response dimension reduction for the conditional mean in multivariate regression. Computational Statistics and Data Analysis, 53, 334–343.

- **Sparse and Regularized Sufficient Dimension Reduction**
  a. Li, L. (2007). Sparse sufficient dimension reduction. Biometrika. 94, 603–613.
  b. Li, L., Cook, R.D., and Tsai, C.L. (2007). Partial inverse regression. Biometrika. 94, 615–625.
  c. Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations. Biometrics, 64, 124–131.

- **Model-based Sufficient Dimension Reduction**
  a. Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. Statistical Science, 22, 1–26.

MANY THANKS!!