Midterm Report

By Liu Zeyang, 11911325

Abstract

This report will mainly investigate the paper $Sufficient\ dimension\ reduction\ and\ prediction\ in\ regression$ by Kofi P. Andragni and R. Dennid Cook. In this paper, it gave a broad overview of the methods in dimension reduction like inverse reduction and principal components (PCs) and it described the corresponding methods for prediction in regressions with many predictors. These will be included in the summary part of my report. Practical implementation issues like the $choice\ of\ d$ in the paper will be excluded in my report since it is uncorrelated the topic in my report. Besides, I will compare the estimators in this paper with NW estimator and figure out the relationship between them. In the end, I will propose my own estimator.

1 Summary

1.0 Pre-knowledge

- **Dimension reduction and sufficient dimension reduction**

Dimension reduction is used in the process to determine a rule $m(\mathrm{x})$ for predicting a future observation of a univariate response variable Y at the given value x of a $p \times 1$ vector X of continuous predictors. When Y is quantitative, minimize $E(Y - m(\mathrm{x}))^2$ can directly lead to $m(\mathrm{x}) = E(Y|\mathrm{X} = \mathrm{x})$. As a result, the problem can usually be specified to be the task of estimating $E(Y|\mathrm{X})$. Besides, when Y is categorical with sample $S_Y = \{C_1, \ldots, C_h\}$, the predicted category C_* is usually taken to be $\arg\max \Pr(C_k|\mathrm{X} = \mathrm{x})$, where the maximization is over S_Y. When pursuing the $E(Y|\mathrm{X})$ or $\Pr(C_k|\mathrm{X})$, it is worthwhile to consider predictions based on a function $R(\mathrm{X})$ with dimension less than p, provided that it captures all of the information that X contains about about Y so that $E(Y|\mathrm{X}) = E(Y|R(\mathrm{X}))$. The action of replacing X with a lower-dimensional function $R(\mathrm{X})$ is called $dimension\ reduction$; it is called $sufficient\ dimension\ reduction$ when $R(\mathrm{X})$ retains all the relevant information about Y. Here is the definition of $sufficient\ dimension\ reduction$:

Definition 1.1. A reduction $R : \mathbb{R}^p \to \mathbb{R}^q, q < p$, is sufficient if it satisfies one of the following three statements:

1. inverse reduction, $\mathrm{X}|(Y, R(\mathrm{X})) \sim \mathrm{X}|R(\mathrm{X})$,
2. forward reduction, $Y|\mathrm{X} \sim Y|R(\mathrm{X})$,
3. joint reduction, $\mathrm{X} \perp\!\!\!\perp Y|R(\mathrm{X})$

Besides, for a sufficient reduction R, if for any sufficient reduction T, we can find a function $f\ s.t.\ R = f(T)$, then R is called $minimal\ sufficient$.

- **Dimension-reduction space and central space**

Most often, we encountered multidimensional reduction with the form $R(\mathrm{X}) = \eta^T\mathrm{X}$, where η is an unknown $p \times q$ matrix with $q \leq p$. When $\eta^T\mathrm{X}$ is a sufficient linear reduction, then so is $(\eta\mathrm{A})^T\mathrm{X}$ for any $q \times q$ full-rank matrix A. Consequently, only the subspace $\mathrm{span}(\eta)$ spanned by the columns of η can be identified. There we call $\mathrm{span}(\eta)$ a $dimension - reduction\ subspace$.

Moreover, there are some properties of dimension-reduction space.

1. If span(η) is a dimension-reduction space, then so is span(η, η_2) for any $p \times q$ matrix η_2.
2. If span(η_1) and span(η_2) are both dimension-reduction space, then under mild conditions so is $\mathrm{span}(\eta_1) \cap \mathrm{span}(\eta_2)$ (Cook 1996, 1998).

Therefore, the inferential target in sufficient dimension reduction is often taken to be the central space $S_{Y|X}$, defined as the interaction of all dimension-reduction subspaces (Cook 1994, 1996, 1998). And a $minimal\ sufficient\ linear\ reduction$ is then of the form $R(\mathrm{X}) = \eta^T \mathrm{X}$, where the columns of η now form a basis for $S_{Y|X}$.

1.1 Reduction in forward linear regression

This part only consider the standard linear regression model $Y = \beta_0 + \beta^T \mathrm{X} + \epsilon$ with $\epsilon \perp\!\!\!\perp \mathrm{X}$ and $E(\epsilon) = 0$. In this case, we can figure out that $S_{Y|X} = \mathrm{span}(\beta)$ and $R(\mathrm{X}) = \beta^T \mathrm{X}$ is minimal sufficient. So the crux of this model is to estimate β. This part reviews tow methods used to improve on the OLS estimator of β and to deal with $n < p$ regressions. But we should note that the performance of the two methods is bad when when the model is not accurately linear.

- **G-reduction approach**

This approach consists of regressing Y on X in two steps. The first is to reduce X linearly to $\mathrm{G}^T\mathrm{X}$, where $\mathrm{G} \in \mathbb{R}^{p\times q}$, $q \leq p$, produced by some methodology. The second is to estimate $E(Y|\mathrm{G}^T\mathrm{X})$ by using OLS. We will address some notations firstly. Let $\mathbb{Y}$ be the vector of centered responses, $\bar{\mathrm{X}} = \sum_{i=1}^n \mathrm{X}/n$ denote the sample mean vector, $\mathbb{X}$ be the $n \times p$ matrix with rows $(\mathrm{X}_i - \bar{\mathrm{X}})^T, i = 1, \ldots, n$, $\hat{\Sigma} = \mathbb{X}^T\mathbb{X}/n$ denote the usual estimator of $\mathrm{var}(\mathrm{X})$, and $\hat{\mathrm{C}} = \mathbb{X}^T\mathbb{Y}/n$. Then we can summary the estimators:

$$\hat{\beta}_{\mathrm{G}} = \mathrm{G}(\mathrm{G}^T\hat{\Sigma}\mathrm{G})^{-1}\mathrm{G}^T\hat{\mathrm{C}},$$
$$\hat{E}(Y|\mathrm{X}) = \bar{\mathrm{Y}} + \hat{\beta}_{\mathrm{G}}^T(\mathrm{X} - \bar{\mathrm{X}}).$$

Note that if $\mathrm{span}(\mathrm{G})$ is a consistent estimator of a dimension-reduction subspace S, then $\hat{\beta}_{\mathrm{G}}$ may be a reasonable estimator of β since $\mathrm{span}(\beta) \subseteq S \subseteq \mathbb{R}^p$.

- **Penalizing approach**

This approach is based on estimating β by using a penalized objective function like that for the lasso:

$$\hat{\beta}_{\mathrm{lasso}} = \arg\min_{\beta}\{\sum_{i=1}^{n}(Y_i - \bar{Y} - \beta^T(\mathrm{X}_i - \bar{\mathrm{X}}))^2 + \lambda\sum_{j=1}^{p}|\beta_j|\}$$

where β_j is the j th element of β.

1.2 Invese reduction

Inverse regression provides an alternative approach to estimating a sufficient reduction. Inverse reduction by itself does not require the response Y to be random, and it is perhaps the only reasonable reductive route when Y is fixed by design. There are two general paradigms for determining a sufficient reduction inversely: moment-based inverse reduction and model-based inverse reduction. We should note that model accuracy is nearly always an issue in the model-based approach, while efficiency is worrisome in the moment-based approach.

- **Moment-based inverse reduction**

In moment-based, derived moment relations are used to estimate a sufficient reduction by way of the central subspace. The first two moment-based methods were sliced inverse regression (SIR; Li 1991) and sliced average variance estimation (SAVE; Cook & Weisberg 1991). Here are some details of this two methods.

Both SIR and SAVE estimate $S_{Y|X}$ under two key conditions: (i) $E(\mathrm{X}|\eta^T\mathrm{X})$ is a linear function of X ($linearity\ condition$) and (ii) $\mathrm{var}(\mathrm{X}|\eta^T\mathrm{X})$ is a non-random matrix ($constant\ covariance\ condition$). Under the linearity condition, $E(\mathrm{X}|Y) - E(\mathrm{X}) \in \Sigma S_{Y|X}$. Under the linearity and constant covariance conditions, $\mathrm{span}(\Sigma - \mathrm{var}(\mathrm{X}|Y)) \in \Sigma S_{Y|X}$. When the response is categorical, let $\bar{\mathrm{X}}_k$ denote the average predictor vector in category C_k. Then the SIR estimator of $S_{Y|X}$, which requires $n > p$, is the span of the first d eigenvectors of $\hat{\Sigma}^{-1}\mathrm{M}\mathrm{M}^T$, where M is the $p \times h$ matrix with columns $\bar{\mathrm{X}}_k - \bar{\mathrm{X}}$. Continuous responses are treated by slicing the observed range of Y into h categories C_k and then applying the same process above.

Note that, moment-based inverse reduction methods provide estimate of the minimal sufficient linear reduction, but they are not designed specifically for prediction and do not produce predictive methods.

- **Model-based inverse reduction**

In model-based inverse reduction, $\mathrm{X}|Y$ can itself be inverted to provide a method for estimating the forward mean function $E(Y|\mathrm{X})$ without specifying a model for the full joint distribution of (X, Y). Here is the inspiration of this method:

$$\begin{aligned} E\{Y|\mathrm{X} = \mathrm{x}\} &= \frac{E\{Yg(\mathrm{x}|Y)\}}{E\{g(\mathrm{x}|Y)\}} \\ &= E(R(\mathrm{x})) \\ &= \frac{E\{Yg(R(\mathrm{x})|Y)\}}{E\{g(R(\mathrm{x})|Y)\}} \end{aligned}$$

where g denotes the density function of the corresponding variable. Based on the equations, we can derive the method to estimate $E(Y|\mathrm{X})$:

$$\begin{aligned} \hat{E}\{Y|\mathrm{X} = \mathrm{x}\} &= \sum_{i=1}^{n} w_i(\mathrm{x})Y_i \\ w_i(\mathrm{x}) &= \frac{\hat{g}(\hat{R}(\mathrm{x})|Y_i)}{\sum_{i=1}^{n}\hat{g}(\hat{R}(\mathrm{x})|Y_i)} \end{aligned}$$

where $\hat{g}$ denotes an estimated density and $\hat{R}$ is the estimated reduction. From the expression, we can see that the performance of this method depends on the obtaining good estimators of the reduction and of its conditional density. In the nest part, we will talk about this topic elaborately.

1.3 Normal inverse models

In this part, we will assume $\mathrm{X}_y \equiv \mathrm{X}|(Y = y) \sim N(\mu_y, \Delta)$, where $\Delta > 0$. Then let $\bar{\mu} = E(\mathrm{X})$, $\Gamma \in \mathbb{R}^{p\times d}$ denote a basis matrix whose columns form a basis for the d-dimensional subspace $S_\Gamma = \mathrm{span}\{\mu_y - \bar{\mu}|y \in S_Y\}$, where S_Y denotes the sample space of Y. Then we can write (Cook 2007)

$$\begin{gathered} \mathrm{X}_y = \bar{\mu} + \Gamma v_y + \varepsilon, \\ \text{where } \varepsilon \perp\!\!\!\perp Y, \varepsilon \sim N(0, \Delta), \\ v_y = (\Gamma^T\Gamma)^{-1}\Gamma^T(\mu_y - \bar{\mu}) \in \mathbb{R}^d, \text{ and we assume } \mathrm{var}(v_Y) > 0. \end{gathered}$$

Here Γ is not identifiable and span(Γ) is identifiable and estimable, so WLOG we can assume that $\Gamma^T\Gamma = \mathrm{I}_d$.

Since $R(\mathrm{X}) = \Gamma^T\Delta^{-1}\mathrm{X}$ is minimal sufficient (Cook & Frozani 2009*b*), then the goal is to estimate $\Delta^{-1}S_\Gamma \equiv \{\Delta^{-1}z : z \in S_\Gamma\}$. Later we will summary a few models considering this problems. Before that, we will state some notations. Let $S_d(\mathrm{A}, \mathrm{B})$ denote the span of $\mathrm{A}^{-1/2}$ times the first d eigenvectors of $\mathrm{A}^{-1/2}\mathrm{B}\mathrm{A}^{-1/2}$, where A and B are symmetric matrices and A is non-singular. The subspace $S_d(\mathrm{A}, \mathrm{B})$ can be also described as the span of A^{-1} times the first d eigenvectors of B. We refer to errors having $\Delta = \sigma^2\mathrm{I}_p$ as $isotonic$ and X_{y_i} as X_i.

- **The principal component model**

In this model, we make the assumption that the model is isotonic. The MLE of $\Delta^{-1}S_\Gamma = S_\Gamma$ is $S_d(\mathrm{I_p}, \hat{\Sigma})$ and thus the d components of $\hat{R}(\mathrm{X})$ are simply the first d PCs. So this model is also called the $PC\ model$. Besides the MLE of σ^2 is $\hat{\sigma}^2 = \sum_{j=d+1}^p \hat{\lambda}_j/p$, where $\hat{\lambda_1} > \hat{\lambda_2} > \ldots > \hat{\lambda_{d+1}} \geq \ldots \geq \hat{\lambda_p}$ are the eigenvalues of $\hat{\Sigma}$, and the MLE of $\bar{\mu}$ is simply $\bar{\mathrm{X}}$. If we let $\hat{\Gamma}$ denote the $p \times d$ matrix with columns consisting of the first d eigenvectors of $\hat{\Sigma}$, then the weights to predicted $\hat{E}\{Y|\mathrm{X} = \mathrm{x}\}$ can be written as

$$\begin{aligned} w_i(\mathrm{x}) &\propto exp\{-(2\hat{\sigma}^2)^{-1}||\hat{\Gamma}(\mathrm{x} - \mathrm{X}_i)||^2\} \\ &= exp\{-(2\hat{\sigma}^2)^{-1}||\hat{R}(\mathrm{x}) - \hat{R}(\mathrm{X_i})||^2\} \end{aligned}$$

- **The isotonic principal fitted component model**

In this model, we assume that $v_y = \beta(\mathrm{f}_y - \bar{\mathrm{f}})$, where $\mathrm{f}_y \in \mathbb{R}^r$ is a known vector-valued function of y with linearly independent elements and $\beta \in \mathbb{R}^{d\times r}$, $d \leq \min(r, p)$, is an unrestricted rank-d matrix. And we also assume that the errors are isotonic. Then we let $\mathbb{F}$ denote the $n \times r$ matrix with rows $(\mathrm{f}_i - \bar{\mathrm{f}})^T$, $\hat{\mathbb{X}} = \mathrm{P}_{\mathbb{F}}\mathbb{X},\ \hat{\Sigma}_{\mathrm{fit}} = \mathbb{X}^T\mathrm{P}_{\mathbb{F}}\mathbb{X}/n$, where $\mathrm{P}_{\mathbb{F}} = \mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}\mathbb{F}^T$. Then the MLE of $\Delta^{-1}S_\Gamma$ is $S_d(\mathrm{I}_d, \hat{\Sigma}_{\mathrm{fit}})$ and the sufficient reduction is estimated as $\hat{\Gamma}^T\mathrm{x}$, where the columns of $\hat{\Gamma}$ are the first d eigenvectors of $\hat{\Sigma}_{\mathrm{fit}}$. Additionally, the MLE of β is $\hat{\beta} = \hat{\Gamma}^T\mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}$ and the MLE of σ^2 is $\hat{\sigma}^2 = (\sum_{j=1}^p \hat{\lambda}_j - \sum_{j=1}^d \hat{\lambda}_j^{\mathrm{fit}})/p$, where $\hat{\lambda}_j^{\mathrm{fit}}$ are the ordered eigenvalues of $\hat{\Sigma}_{\mathrm{fit}}$. Let $\mathrm{B}_{\mathrm{ols}} = \mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1},\ \hat{\mathrm{X}}_i = \bar{\mathrm{X}} + \mathrm{B}_{\mathrm{ols}}(\mathrm{f}_i - \bar{\mathrm{f}})$. then the weights to predicted $\hat{E}\{Y|\mathrm{X} = \mathrm{x}\}$ can be written as

$$w_i(\mathrm{x}) \propto exp\{-(2\hat{\sigma}^2)^{-1}||\hat{R}(\mathrm{x}) - \hat{R}(\hat{\mathrm{X}}_i)||^2\}$$

- **The diagonal principal fitted component model**

In this model we release the restriction of Δ to be $\Delta = diag(\sigma_1^2, \ldots, \sigma_p^2)$. The estimated sufficient reduction for this model is $\hat{R}(\mathrm{x}) = \hat{\Gamma}^T\hat{\Delta}\mathrm{x}$. With this $\hat{R}$, the weights have the same form as the weights for the isotonic PFC models,

$$w_i(\mathrm{x}) \propto exp\{||\hat{R}(\mathrm{x}) - \hat{R}(\hat{\mathrm{X}}_i)||^2\}$$

However, there are no closed-forms of $\hat{\Gamma}$ and $\hat{\Delta}$, and here is an alternating algorithm to calculate:

1. Fit the isotonic PFC model to the original data, getting $\hat{\Gamma}_{(1)}$ and $\hat{\beta}_{(1)}$.
2. For some fixed $\epsilon > 0$, repeat the following steps until $\mathrm{tr}\{\hat{\Delta}_{(j)} - \hat{\Delta}_{(j+1)}\} < \epsilon$.
    1. Calculate $\hat{\Delta}_{(j)} = \mathrm{diag}\{(\mathbb{X} - \mathbb{F}\hat{\beta}_{(j)}^T\hat{\Gamma}_{(j)}^T)^T(\mathbb{X} - \mathbb{F}\hat{\beta}_{(j)}^T\hat{\Gamma}_{(j)}^T)\}$.
    2. Transform $\mathrm{Z} = \hat{\Delta}_{(j)}^{-1/2}\mathrm{X}$.
    3. Fit the isotonic PFC model to Z, yielding estimates $\widetilde{\Gamma}$, and $\widetilde{\beta}$.
    4. set $\hat{\Gamma}_{(j+1)} = \hat{\Delta}_{(j)}^{1/2}\widetilde{\Gamma},\ \hat{\beta}_{(j+1)} = \widetilde{\beta}$

- **The principal fitted component model**

In this model, we just assume that $\Delta > 0$. However, we should be aware of that the methods stated later work best in data-rich regressions where $p \ll n$. Before that, we will summary some notations. Let $\hat{\Sigma}_{\text{res}} = \hat{\Sigma} - \hat{\Sigma}_{\text{fit}} > 0$, $\hat{w}_1 \geq \ldots \geq \hat{w}_p$ and $\hat{\mathrm{V}} = (\hat{\mathrm{v}}_1, \ldots, \hat{\mathrm{v}}_p)$ be the eigenvalues and the corresponding matrix of eigenvectors of $\hat{\Sigma}_{\text{res}}^{-1/2}\hat{\Sigma}_{\text{fit}}\hat{\Sigma}_{\text{res}}^{-1/2}$, and $\hat{\mathrm{K}}$ be a $p \times p$ diagonal matrix with the first d diagonal elements equal to zero and the last $p-d$ diagonal elements equal to $\hat{w}_{d+1}, \ldots, \hat{w}_p$. Then the MLE of Δ is $\hat{\Delta} = \hat{\Sigma}_{\text{res}}^{1/2}\hat{\mathrm{V}}(\mathrm{I}_p + \hat{\mathrm{K}})\hat{\mathrm{V}}^T\hat{\Sigma}_{\text{res}}^{1/2}$. The MLE of β is $\hat{\beta} = \hat{\Gamma}^T \mathrm{P}_{\hat{\Gamma}(\hat{\Delta}^{-1})}\mathrm{B}_{\text{ols}}$, where $\mathrm{P}_{\hat{\Gamma}(\hat{\Delta}^{-1})}\mathrm{B}_{\text{ols}}$ is the projection of B_{ols} onto the span of $\hat{\Gamma}$ in the $\hat{\Delta}^{-1}$ inner product and the MLE of $\Delta^{-1}\Gamma$ is $S_d(\hat{\Delta}, \hat{\Sigma}_{\text{fit}})$. Besides, weights to predicted $\hat{E}\{Y|\mathrm{X} = \mathrm{x}\}$ can be written as

$$w_i(\mathrm{x}) \propto exp\{-\frac{1}{2}(\mathrm{x} - \hat{\mathrm{X}}_i)^T[\hat{\Delta}^{-1}\hat{\Gamma}(\hat{\Gamma}^T\hat{\Delta}^{-1}\hat{\Gamma})^{-1}](\mathrm{x} - \hat{\mathrm{X}}_i)\}.$$

2 Comparison with NW estimator

This paper mainly talks about the estimator in forward linear regression $\hat{E}(Y|\mathrm{X}) = \bar{\mathrm{Y}} + \hat{\beta}_{\mathrm{G}}^T(\mathrm{X} - \bar{\mathrm{X}})$ and the estimator corresponding to the inverse reduction $\hat{E}\{Y|\mathrm{X} = \mathrm{x}\} = \sum_{i=1}^n w_i(\mathrm{x})Y_i,\ w_i(\mathrm{x}) = \frac{\hat{g}(\hat{R}(\mathrm{x})|Y_i)}{\sum_{i=1}^n \hat{g}(\hat{R}(\mathrm{x})|Y_i)}$. The first one and NW estimator are essentially unrelated, so the discussions below will focus on the relationship between the estimator in inverse reduction and the NW estimator.

- **Difference in theoretical basis**

Although the two estimators seem similar and even sometimes have the same expression (the estimator for principal component model based on inverse reduction matches with the NW estimator with a normal product kernel density estimator with bandwidth $\hat{\sigma}$), they are totally two different estimators since they are based on different theories. For both estimators, they start from the same equation:

$$E\{Y|X = x\} = \int y\, f(y|x)\, dy = \frac{\int y f(x,y) dy}{f_X(x)}$$

That's also the reason that why the two estimators have the similar expression. However, in NW estimator, $f(x,y)$ and $f_X(x)$ are directly estimated by $\hat{f}_{h,g}(x,y) = \frac{1}{n}\sum_{i=1}^n K(\frac{x-X_i}{h})K(\frac{y-Y_i}{g})$ and $\hat{f}_X(x) = \frac{1}{n}\sum_{j=1}^n K(\frac{x-X_j}{h})$ respectively. While another estimator goes further in theory. It use the conditional distribution to express $\int y f(x,y)dy$ as $\int y g(x|y)g(y)dy$ and $f_X(x)$ as $\int g(x|y)g(y)dy$ and then use the reduction method and the method that sample mean can be an estimator of population mean to estimate numerator and denominator respectively.

From the statement above, we can see that NW estimator requires the joint distribution of X and Y, while the another does not require Y to be random. Relatively speaking, weights of NW estimator do not depend on the response while that of another estimator do.

- **Difference in performance**

This part is simply the difference in performance for different dimensions d. For the NW estimator, the curse of dimensionality is well-known while another estimator performs well in high dimension problems due to the dimension reduction part $R(x)$.

3 Ideas about new estimators

This part will propose some preliminary ideas of new estimators.

- **Combination of NW estimator and dimension reduction**

To deal with the curse of dimension in NW estimator, a natural idea is to use dimension reduction. So we may use moment-based inverse reduction or PCs to find the sufficient reduction $R(x)$, by which a new version of NW estimator can be proposed to deal with higher-dimensional problems:

$$\hat{m}_h(x) = \frac{1}{n}\sum_{i=1}^{n}(\frac{K_h(R(x)-R(X_i))}{n^{-1}\sum_{j=1}^{n}K_h(R(x)-R(X_j))})Y_i$$

Moreover, we may try just simply combine the NW estimator and the estimator in inverse reduction like this:

$$w_i(x) = \frac{\hat{g}(\hat{R}(x)|Y_i)}{\sum_{i=1}^{n}K_h(x-X_i)} \text{ or } w_i(x) = \frac{K_h(x-X_i)}{\sum_{i=1}^{n}\hat{g}(\hat{R}(x)|Y_i)}$$

- Different choice of the distribution of $\mathbf{X}_y$ inverse reduction

In the paper summarized above, they specified $\mathbf{X}_y$ to follow the normal distribution. However, when we know the range of $\mathbf{X}_y$, we may specify it to follow other models such as multivariate uniform distribution or Dirichlet distribution. Under these distributions we may be able to determine $R(x)$ and $E(Y|X=x)$ that are more suitable for this dataset.

References

Adragni, Kofi P, and R. Dennis Cook. "Sufficient Dimension Reduction and Prediction in Regression." Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical, and Engineering Sciences 367.1906 (2009): 4385-405. Web.