Intensive comparison of semi-parametric and non-parametric dimension reduction methods in forward regression

Minju Shin\textsuperscript{a}, Jae Keun Yoo\textsuperscript{1,a}

\textsuperscript{a}Department of Statistics, Ewha Womans University, Korea

Abstract

Principal Fitted Component (PFC) is a semi-parametric sufficient dimension reduction (SDR) method, which is originally proposed in Cook (2007). According to Cook (2007), the PFC has a connection with other usual non-parametric SDR methods. The connection is limited to sliced inverse regression (Li, 1991) and ordinary least squares. Since there is no direct comparison between the two approaches in various forward regressions up to date, a practical guidance between the two approaches is necessary for usual statistical practitioners. To fill this practical necessity, in this paper, we newly derive a connection of the PFC to covariance methods (Yin and Cook, 2002), which is one of the most popular SDR methods. Also, intensive numerical studies have done closely to examine and compare the estimation performances of the semi- and non-parametric SDR methods for various forward regressions. The founding from the numerical studies are confirmed in a real data example.

Keywords: covariance methods, principal fitted component, semi-parametric dimension reduction, sufficient dimension reduction

1. Introduction

Sufficient Dimension Reduction (SDR) is a reduction of the predictor dimension in a regression of \(y|X\) without losing the information on the regression. This can be achieved by replacing the original \(p\)-dimensional predictor \(X\) with its lower-dimensional linearly transformed predictor to satisfy

\[
y \perp X|\eta^T X,
\]

where \(\perp\) stands for the statistical independence, and \(\eta\) is a \(p \times d\) matrix.

If satisfying (1.1), we have that \(F_{y|X}() = F_{y|\eta^T X}()\), where \(F()\) is a distribution function. Naturally, the main goal of SDR is to estimate \(\eta\) to have the minimum dimension \(d\), which is called the central subspace \(S_{Y|X}\).

Many SDR methods (Li, 1991; Cook and Weisberg, 1991; Li, 1992; Yin and Cook, 2002) estimate \(S_{Y|X}\) non-parametrically under mild conditions. For more details of the methods and methodological

\textsuperscript{1}Corresponding author: Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-Gu, Seoul 03760, Korea. E-mail: peter.yoo@ewha.ac.kr

Published 30 September 2022 / journal homepage: http://csam.or.kr
© 2022 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.
comparison, readers are recommended to read Yoo (2016). Among them, two methods by Li (1991) and Yin and Cook (2002) are widely used in SDR context, and the methods are called sliced inverse regression (SIR) and covariance methods (covK), respectively. According to Yoo (2018b), the two methods often yield similar estimation results. The two methods are extended to regularization in Li and Yin (2008) and are adopted in seeded dimension reduction (Cook et al., 2007). Also their fused methods are developed in Cook and Zhang (2014) and Yoo (2018a).

Different from these non-parametric approaches, a semi-parametric sufficient dimension reduction was proposed in Cook (2007). The proposed inverse model has two sub-models called principal component and principal fitted component models depending on the setting between the predictor X and the response y. The principal component model does not directly consider the relation between X and y, while the principal fitted model can allow users to choose various functions between X and y. Also, depending on the covariance structure of a random error, each semi-parametric model has three variations of isotonic, which has a diagonal covariance matrix, and unstructured models. The structured model releases the restricted form of the isotonic one, and the unstructured model generalizes the structured one. There should be the following six combination models of isotonic principal component model, isotonic principal fitted component, structured principal component model, structured principal fitted component model, unstructured principal component model, and unstructured principal fitted component model. This is why the inverse model by Cook (2007) is semi-parametric. However, only four models, which are isotonic principal component, isotonic principal fitted component, and unstructured fitted component models, among the six will be considered in Section 2, following Cook (2007)'s guidance. As principal fitted component models are classified by covariance matrix structure of a random error, starting by understanding of isotonic principal component and fitted component model, the simplest models, will help researchers to grasp the general concept of principal fitted component models. We only use unstructured principal fitted component models in numerical study to find connections between model-based methods and model-free methods, which is the main goal of this paper. Especially focusing on the unstructured fitted component model, Cook and Forzani (2009) derived various theoretical results. More on Cook’s approach will be discussed in later section.

Up to date, the comparison between the non-parametric methods and the four variations of Cook (2007) has been limited to SIR and the ordinary least squares under considering the inverse regression setting, which should be more favorable to Cook (2007). This may not be enough for the semi-parametric approach to have potential advantages over the usual non-parametric SDR methods in practice. The main goal of this paper is to examine and compare SIR and covK with the four variations of Cook (2007) in various forward regression settings, which were widely studied in the non-parametric methods. This will fill a current gap in the fair competition between the two approaches, and will provide more complete guidance to statistical practitioners and researchers.

The organization of the paper is as follows. In Section 2 the semi-parametric models proposed by Cook (2007) are reviewed focusing on isotonic principal component and principal fitted component models and unstructured fitted component models. Section 3 is devoted to connecting Cook’s approach to the existing sliced inverse regression and covariance methods. Intensive numerical studies in various forward regressions will be done in Section 4. A real data example will be presented in Section 5. Our work will be summarized in Section 6.

For notational conveniences throughout the rest of the paper, for $B \in \mathbb{R}^{p \times q}, S_B$ stands for a subspace spanned by the columns of $B$, and $P_B$ represents an orthogonal projection operation onto $S_B$. 
2. Review: Semi-parametric inverse model

2.1. Isotonic principal component and principal fitted component models

We start with a principal component (PC) model in Cook (2007) with an isotonic error model:

\[ X_y = \mu + \Gamma y + \sigma \varepsilon, \]  

(2.1)

where \( \mu \in \mathbb{R}^p, \Gamma \in \mathbb{R}^{p \times d}, d < p, \Gamma^T \Gamma = I_p, \sigma \geq 0 \) and \( \varepsilon \sim N(0, I_p) \) and \( d \) is assumed to be known. The vector \( y \in \mathbb{R}^d \) is an unknown function of \( y \) that is assumed to have a positive definite sample covariance matrix and is centered to have a mean 0, \( \sum_y y = 0 \). Under model (2.1), predictors be in the same scale and be independent with the same variance as PC are not invariant or equivariant under full rank linear transformations of predictors. According to Cook (2007), after translation by intercept \( \mu \), this model indicates that the conditional means fall in the \( d \)-dimensional subspace \( S_{FP} \). According to Cook (2007), the columns of \( \Gamma \) span \( S_{FP} \).

Naturally \( \Gamma \) is estimated via maximizing the likelihood function of (2.1) over the Grassman manifold \( G_{pod} \). The maximum likelihood estimator of \( S_{FP} \) can be constructed by holding \( \Gamma \) and \( \sigma^2 \) fixed at possible values \( G \) and \( s^2 \) and maximizing the log-likelihood over \( \mu \) and \( y \). Then we can get partially maximized log-likelihood \( \ell_{PC}(\Gamma) \) depends only on \( S_{FP} \):

\[ \ell_{PC}(\Gamma) = -\frac{n p}{2} \log s^2 - \frac{n}{2s^2} \text{trace} \left( \hat{\Sigma} Q_G \right), \]  

(2.2)

where \( Q_G = I_p - P_G \), \( s^2 \) is the sample variance of \( y \), and \( \hat{\Sigma} \) stands for usual moment estimator of \( \Sigma = \text{cov}(X) \). The first \( d \) principal components of \( \hat{\Sigma} \) maximizes (2.1), and \( S_{FP} \) is the estimate of \( S_{FP} \) under model (2.1).

In the isotonic PC model in (2.1), there is no direct use of the response \( y \). For this Cook (2007) suggests to replace \( y \) with \( \beta f_y \) in (2.1):

\[ X_y = \mu + \Gamma \beta f_y + \sigma \varepsilon, \]  

(2.3)

where \( \beta \) is an \( d \times r \) unknown matrix and \( f_y \in \mathbb{R}^r \) is a known vector-valued function of the response with \( \sum_y f_y = 0 \). The model in (2.3) is called isotonic principal fitted component (PFC) model. Under the isotonic PFC model, \( f_y \) could be constructed in several ways. Graphical guidance such as inverse plots of \( X_y \) versus \( y \) for \( j = 1, \ldots, p \) as described by Cook and Weisberg (1994, Chapter 8) and Cook (1998, Chapter 10) is practical when the response is bivariate or univariate. According to Cook (2007), when \( y \) is univariate and graphical guidance is not available, \( f_y \) could be constructed by first partitioning the range of \( y \) into \( h = r + 1 \) "slices" or bins \( H_k \) and then setting the \( k \)th coordinate \( f_{yk} \) of \( f_y \) to

\[ f_{yk} = J(y \in H_k) - \frac{n_k}{n}, \quad k = 1, \ldots, r, \]  

(2.4)

where \( J \) is the indicator function and \( n_k \) is the number of observations falling in \( H_k \).

Under model (2.3), the columns of \( \Gamma \) still span \( S_{FP} \), and \( \Gamma \) is estimated via maximum likelihood method. Let \( \bar{X} \) and \( \bar{Y} \) denote the matrices with stacking \( (X_y - \bar{X})^T \) and \( f_y^T \) as rows, respectively. Also, we define that \( \bar{X} = P_G \bar{X} \) the \( n \times p \) matrix of centered fitted values from the multivariate linear regression of \( X \) on \( f_y \). Holding \( \Gamma \) and \( \sigma \) fixed at \( G \) and \( \sigma \) respectively, and maximizing the likelihood over \( \mu \) and \( \beta \), we obtain the partially maximized log likelihood:

\[ \ell_{PFC}(G) = -\frac{n p}{2} \log \left( \sigma^2 \right) - \frac{n}{2\sigma^2} \left( \text{trace} \left( \hat{\Sigma} - P_G \hat{\Sigma}_{int} \right) \right), \]  

(2.5)
where \( \hat{\Sigma}_{\text{fit}} = \frac{1}{n} \hat{X} \hat{X}^T \) is the sample covariance matrix of the fitted values. The log-likelihood function in (2.5) is maximized at the first \( d \) principal components of \( \hat{\Sigma}_{\text{fit}} \), which is the maximum likelihood estimator of \( \Gamma \).

### 2.2. Structured principal fitted component model

Now we consider model (2.6) with relevance of information. Let \( \Gamma_0 \in \mathbb{R}^{p \times (p-d)} \) denote a completion of \( \Gamma \); that is, \((\Omega, \Gamma) \in \mathbb{R}^{p \times p} \) is an orthogonal matrix. Extending the covariance matrix of \( \epsilon \) in model (2.3), while preserving the form of the relevant and irrelevant information, consider the PFC model with heterogeneous errors,

\[
X_f = \mu + \Gamma \beta f + \Gamma_0 \Omega \epsilon_0 + \Gamma \Omega \epsilon,
\]

where \( \mu, \Gamma, \Gamma_0, \) and \( f \) are defined previously. The error vectors \( \epsilon_0 \in \mathbb{R}^{p-d} \) and \( \epsilon \in \mathbb{R}^d \) are assumed to be independent and normally distributed, each with zero mean vectors and the identity covariance matrices. The full-rank matrices \( \Omega \in \mathbb{R}^{d \times d} \) and \( \Omega_0 \in \mathbb{R}^{(p-d) \times (p-d)} \) convert the normal errors to appropriate scales and, without loss of generality, are assumed to be symmetric and must be independent. Since \((\Gamma_0, \Gamma)^T X \) contains two independent components, this extended model preserves the independence property that \((Y, \Gamma)^T I \Gamma^T X\), so \( \Gamma^T X \) attains sufficient reduction.

According to Cook (2007), the maximum likelihood estimator of \( \Gamma \) can be obtained from partially maximizing the log-likelihood \( \ell_{\text{PFC}} \):

\[
\ell_{\text{PFC}}(G) = - \frac{n}{2} \log \| G^T \hat{\Sigma} G_0 \| - \frac{n}{2} \log \| G^T \hat{\Sigma}_{\text{res}} G \|,
\]

where \( \hat{\Sigma}_{\text{res}} = \hat{\Sigma} - \hat{\Sigma}_{\text{fit}} \) is the sample covariance matrix of the residuals from the fit of \( X \) on \( f \).

Different from the previous likelihoods, \( \Gamma \) in (2.7) is associated not only with \( \hat{\Sigma} \) but also with \( \hat{\Sigma}_{\text{res}} \). This indicates that there is no closed form for \( \Gamma, \hat{\Gamma} \) should be obtained by maximizing numerically over the Grassman manifold \( \mathcal{G}_{p,d} \). Alternatively, Cook (2007) suggests a sequential evaluation of \( \ell_{\text{PFC}} \) in (2.7) from three candidate sets constructed by PC directions from \( \hat{\Sigma} \), PFC directions from \( \hat{\Sigma}_{\text{fit}} \), and residual component directions from \( \hat{\Sigma}_{\text{res}} \). Proposition 4 of Cook (2007) gives the population versions of three matrices of \( \hat{\Sigma}, \hat{\Sigma}_{\text{fit}} \) and \( \hat{\Sigma}_{\text{res}} \) in the partially maximized likelihood function (2.7).

### 2.3. Unstructured principal fitted component model

Suppose that \( X_f \) follows the following PFC model with \( \text{cov}(\epsilon) = \Lambda \), which is unknown:

\[
X_f = \mu + \Gamma_0 \Omega \epsilon_0 + \Lambda \Omega \epsilon.
\]

According to Cook (2007), a sufficient reduction for model (2.8) is can be done via \((\Lambda^{-1} \Gamma)^T X\). Under model (2.8), the estimator of \( \Lambda \) is defined first, and then we can specify the estimate of the reductive subspace. The maximum likelihood estimator of the reductive space can be constructed through following steps. Maximize the log-likelihood over \( \beta \) and \( \mu \), and get partially maximized log likelihood which is a function of possible values \( \mathbf{D} \) and \( \mathbf{G} \) for \( \Lambda \) and \( \Gamma \). With \( \mathbf{D} \) fixed, this function is maximized by choosing \( \mathbf{D}^{-1/2} \mathbf{G} \) to be a basis for the span of the first \( d \) eigenvectors of \( \mathbf{D}^{1/2} \hat{\Sigma}_{\text{fit}} \mathbf{D}^{-1/2} \), yielding another partially maximized log-likelihood \( \mathbf{K(D)} \),

\[
\mathbf{K(D)} = - \frac{n}{2} \log |\mathbf{D}| - \left( \frac{1}{2} \right) \text{trace} [\mathbf{D}^{-1} \hat{\Sigma}_{\text{res}}] + \sum_{i=d+1}^{p} \lambda_i (\mathbf{D}^{-1} \hat{\Sigma}_{\text{fit}}).
\]
The first two terms are maximized by $\hat{A} = \hat{\Sigma}_{\text{res}}$. Once $\hat{A}$ is determined, the reductive subspace $\hat{S}\hat{A}^{-1}$ is estimated by the span of $\hat{A}^{-1/2}\hat{\Sigma}_{\text{res}}\hat{A}^{-1/2}$.

3. Connecting to covariance methods and sliced inverse regression

3.1. Before connecting to model-free sufficient dimension reduction methods

From now we focus on the unstructured PFC model. The reason is because $\Lambda$ is not estimable in the structured PFC models. Usual SDR methods require the inverse of cov$(X)$, which is crucial to recover $S_{XY}$. This brings that the usual SDR methods can be done through various linear transformation of the predictors to make the estimation of $S_{XY}$ simpler and more accurate. However, in the structured PFC model, this is not possible, because $\Lambda$ cannot be estimated. So, it would not be plausible to make theoretical connection between the structured PFC model and the usual SDR methods.

Depending on the choices of $f$, in the unstructured PFC model, it can be connected to two different SDR methods of cov$K$ and SIR. That is, a choice of $f$, corresponds to a SDR method. This implies that the unstructured PFC model is a class of SDR methods and has more flexibility than usual SDR methods.

3.2. Covariance methods

Define that $\hat{w} = (y - \bar{y})/s_y$ and $\hat{W}_k = (\hat{w}, \hat{w}^2, \ldots, \hat{w}^k)$ and that $\hat{U}_k$ is the mean-centered $\hat{W}_k$ to have zero means, where $s_y$ is the sample standard deviation of $y$. Also we define that $\hat{M}_k = 1/n \sum_{i=1}^n \hat{Z}_i \hat{W}_{ki}^T$, $\hat{M}_k^U = 1/n \sum_{i=1}^n \hat{Z}_i \hat{U}_{ki}^T$, $\hat{C}_k = 1/n \sum_{i=1}^n (\hat{X}_i - \bar{X}) \hat{W}_{ki}$, and $\hat{\text{cov}}(X, \hat{W}_k) = 1/n \sum_{i=1}^n (\hat{X}_i - \bar{X})(\hat{W}_{ki} - \bar{W}_k)^T$, where $\hat{Z}_i = \hat{\Sigma}^{-1/2}(X_i - \bar{X})$. Then Yin and Cook (2002) show that $S_{\hat{\text{cov}}(X, \hat{w})} = \Sigma^{-1/2}S_{M_k}$, and the following lemma is established.

**Lemma 1** The two spaces of $\hat{\Sigma}^{-1}S_{\hat{\text{cov}}(X, \hat{w})}$ and $\hat{\Sigma}^{-1}S_{C_k}$ are the same.

**Proof:** Since $\hat{M}_k$ is equal to $\hat{M}_k^U$, we have that $S_{M_k} = S_{M_k^U}$. Then we have that $\hat{\Sigma}^{-1}S_{\hat{\text{cov}}(X, \hat{w})} = \Sigma^{-1/2}S_{M_k} = \Sigma^{-1/2}S_{M_k^U} = \hat{\Sigma}^{-1}S_{C_k}$. This completes the proof.

The matrix $\hat{\Sigma}^{-1}\hat{\text{cov}}(X, \hat{W}_k)$ is the kernel matrix suggested by Yin and Cook (2002). Lemma 1 makes a connection between $C_k$, constructed by the unstructured PFC with $f_k = \hat{U}_k$, with $M_k$ or $\hat{\text{cov}}(X, \hat{W}_k)$. This connection is summarized by the following proposition.

With setting $f_k = \hat{U}_k$ in (2.8), we have that

$$
\Sigma_{\text{fit}} \equiv \frac{\mathbf{X}^T \mathbf{F} \mathbf{X}}{n} = \frac{\mathbf{X}^T \hat{U}_k \left( \hat{U}_k^T \hat{U}_k \right)^{-1} \hat{U}_k^T \mathbf{X}}{n} = \left( \frac{\mathbf{X}^T \hat{U}_k}{n} \right) \left( \frac{\hat{U}_k^T \mathbf{X}}{n} \right)^{-1} \left( \frac{\hat{U}_k^T \mathbf{X}}{n} \right),
$$

$$
= \hat{C}_k \left( \frac{\hat{U}_k^T \hat{U}_k}{n} \right)^{-1} \hat{C}_k^T,
$$

$$
= \hat{C}_k \hat{C}_k^T/(n)^{-1/2},
$$

where $\hat{C}_k = \hat{C}_k(\hat{U}_k^T \hat{U}_k/n)^{-1/2}$.
Accordingly, we have that $\hat{\Sigma}_{\text{res}} = \hat{\Sigma} - \hat{\Sigma}_{\text{fit}} = \hat{\Sigma} - \hat{C}_k^T \hat{C}_k$ and that $\hat{\Sigma}_{\text{res}}^{-1} = \hat{\Sigma}^{-1} + \hat{\Sigma}^{-1} \hat{C}_k (I_k - \hat{C}_k^T \hat{\Sigma}^{-1} \hat{C}_k)^{-1} \hat{C}_k^T \hat{\Sigma}^{-1}$. Following Cook (2007), then, the unstructured PFC and covK have the following relation.

**Proposition 1** If setting $f_i = \hat{U}_k$ in the unstructured PFC model in (2.8), we have $\hat{\Sigma}_{\text{res}}^{-1} S_{\text{fit}} = \hat{\Sigma}^{-1} S_{\hat{C}_k}$.

**Proof:** It should be noted that $S_{\hat{C}_k} = S_{\hat{C}_k}$. Then, we have that $\hat{\Sigma}_{\text{res}}^{-1} S_{\text{fit}} = \hat{\Sigma}_{\text{res}}^{-1} S_{\hat{C}_k} = S_{\hat{C}_k}$. Therefore, we have the following equivalences:

\[
\hat{\Sigma}_{\text{res}}^{-1} \hat{C}_k = \left[ \hat{\Sigma}^{-1} + \hat{\Sigma}^{-1} \hat{C}_k \left( I_k - \hat{C}_k^T \hat{\Sigma}^{-1} \hat{C}_k \right)^{-1} \hat{C}_k^T \hat{\Sigma}^{-1} \right] \hat{C}_k,
\]

\[
= \hat{\Sigma}^{-1} \hat{C}_k \left[ I_k + \left( I_k - \hat{C}_k^T \hat{\Sigma}^{-1} \hat{C}_k \right)^{-1} \hat{C}_k^T \hat{\Sigma}^{-1} \hat{C}_k \right],
\]

\[
= \hat{\Sigma}^{-1} \hat{C}_k \left[ I_k + \left( \left( \hat{C}_k^T \hat{\Sigma}^{-1} \hat{C}_k \right)^{-1} - I_k \right) \right],
\]

\[
= \hat{\Sigma}^{-1} \hat{C}_k \left( \frac{\hat{U}_k^T \hat{U}_k}{n} \right)^{-1} \left[ I_k + \left( \left( \hat{C}_k^T \hat{\Sigma}^{-1} \hat{C}_k \right)^{-1} - I_k \right) \right],
\]

\[
= \hat{\Sigma}^{-1} \hat{C}_k C_0,
\]

where $C_0 = (\hat{U}_k^T \hat{U}_k/n)^{-1/2} [I_k + \left( \left( \hat{C}_k^T \hat{\Sigma}^{-1} \hat{C}_k \right)^{-1} - I_k \right) \]$. Since $C_0$ is a $k \times k$ non-singular matrix. So, we have that $S_{\hat{C}_k} = S_{\hat{C}_k} C_0 = \hat{\Sigma}^{-1} S_{\hat{C}_k} C_0$. Since the column space of a matrix is not changed by post-multiplying the non-singular matrix, we have that $S_{\hat{C}_k} = S_{\hat{C}_k}$, which induces that $\hat{\Sigma}_{\text{res}}^{-1} S_{\hat{C}_k} = \hat{\Sigma}^{-1} S_{\hat{C}_k}$, and it completes the proof.

Combining Lemma 1 and Proposition 1, the unstructured PFC model in (2.8) with $f_i = \hat{U}_k$ yields the equivalent dimension reduction with the covK by Yin and Cook (2002). According to Yin and Cook (2002), the covariance method is potentially useful to detect the information on var($y$)$. X). This indicates that a proper choice of $f_i$ can sufficiently reduce the dimension of the predictors in a forward regression setting with var($y$) depending on $X$ in practice, although the assumed inverse regression model in (2.8) does not have heteroscedasticity for $X$ and $y$.

### 3.3. Sliced inverse regression

According to Cook (2007), $f_i$ must be constructed using the slice basis function (2.4) to relate SIR estimator of the reductive subspace $S_{\hat{A}^{-1}P}$ and the unstructured PFC model using $\hat{A} = \hat{\Sigma}_{\text{res}}$. Cook (2007) demonstrated the following relation between $\hat{\Sigma}_{\text{fit}}$ and $\hat{\Sigma}_{\text{fit}}$.

\[
\hat{\Sigma}_{\text{fit}} = \sum_{k=1}^h \frac{n_k}{n} (\bar{X}_k - \bar{X})(\bar{X}_k - \bar{X})^T,
\]

\[
= \sum_{k=1}^h \frac{n_k}{n} \hat{\Sigma}^{-\frac{1}{2}} (\bar{X}_k - \bar{X})(\bar{X}_k - \bar{X})^T \hat{\Sigma}^{-\frac{1}{2}},
\]

\[
= \hat{\Sigma}^{-\frac{1}{2}} \hat{\Sigma}_{\text{fit}} \hat{\Sigma}^{-\frac{1}{2}}.
\]
Intensive comparison of semi- and non-parametric dimension reduction methods in forward regression

(a) $X_1$ from non-normal distribution
(b) $X_2 + X_3$ from normal distribution
(c) $X_1$ from non-normal distribution
(d) $X_2 + X_3$ from non-normal distribution

Figure 1: Summary boxplots for model 1.

Under model (2.8), the reductive space is spanned by $\hat{\Sigma}^{-1/2} \left( \hat{\Sigma}_{\text{res}}^{-1/2} \hat{\Sigma}_{\text{fit}} \right)$ and the reductive space under SIR is spanned by $\Sigma^{-1/2} \left( \Sigma_{\text{res}}^{-1/2} \Sigma_{\text{fit}} \right)$. According to Cook and Forzani (2009), these two spaces give the same estimated subspace.

4. Numerical examples

To investigate performances of PFC and non-parameteric SDR methods, the following numerical models were considered.

1. $Y = 1.5 (X_1 + 5) (X_2 + X_3 + 2) + 0.5 \varepsilon$,
2. $Y = X_1 / (1 + X_1 + X_2 + X_3) + 0.5 \varepsilon$,
3. $Y = X_1 (X_1 + X_2 + X_3 + 1) + 0.5 \varepsilon$,
4. $Y = 0.5X_1 + (X_2 + X_3 + 1.5)^2 + 0.5 \varepsilon$,
5. $Y = X_1 + X_1 (X_2 + X_3) + 0.5 \varepsilon$,
6. $Y = X_1 + \exp (X_2 + X_3) 0.5 \varepsilon$.

Every model was considered with six dimension reduction methods, structured principal fitted component with slicing $f_y$ (sPFCh), structured principal fitted component with poly $f_y$ (sPFCp), unstructured principal fitted component with slicing $f_y$ (uPFCh), unstructured principal fitted component with poly $f_y$ (uPFCp), sliced inverse regression (SIR), and the covariance methods (covK).
For each model, it was generated 500 times and reported its arithmetic means. Without mentioning specific values, we set $p = 5$, $d = 2$, and $n = 100, 200$. The predictors were constructed as follows. First, we generated two random vectors $V = (V_1, V_2, V_3)^T$ and $W = (W_1, W_2)^T$. We then generated the predictors as $X_1 = W_1, X_2 = V_1 + W_2/2, X_3 = -V_1 + W_2/2, X_4 = V_2 + V_3, X_5 = -V_2 + V_3$.

Two random vectors were drawn from normal distribution and non-normal distribution to verify differences between these two cases. Under normal distribution case, both $V_i$ and $W_i$’s were drawn independently from $N(0, 1)$. For non-normal distribution case, $V_i$ were drawn from a $t_6$ distribution and $W_i$ were drawn from $Ga(0.25, 1)$.

Every model has the same $\eta$: $\eta = (1, 0, 0, 0, 0)(0, 1, 1, 0, 0)^T$. To evaluate the estimation performance of $\hat{\eta}$, $R^2$ was computed through linear regression each variable $X_1, X_2 + X_3$ on $\hat{\eta}^T X$.

### 4.1. Model 1

The distribution of $R^2$ of each base obtained by generating $n = 100$ observations from a normal distribution for model 1 can be seen from the first row in Figure 1. For $X_1$ bases, the methods of making secondary $f_i$ performs relatively better than using categorization among the six methods. For the $(X_2 + X_3)$ basis, it can be seen that the basis was well estimated by having the average $R^2$ of all methods above 0.8.

On the other hand, the distribution of $R^2$ of each base obtained by generating $n = 100$ observations from a non-normal distribution can be seen from the second row in Figure 1. It shows a similar trend to the simulation results conducted under the normal distribution, but the results of every method are worse than before. In particular, the performance degradation of sPFCh is noticeable.
4.2. Model 2

From the Figure 2, we can see the distribution of $R^2$ of each base obtained by generating $n = 100$ observations according to distribution conditions while applying six methods to Model 2. We can see that categorized methods show relatively better performance than making secondary $f_i$ among six methods for all $X_1$, $(X_2 + X_3)$ bases. This is because it does not contain a square term for $X_1$, $X_2$, and $X_3$. However, under non-normal distribution, most of methods show lower $R^2$ than the simulation results conducted under normal distribution, especially, the performance of $(X_2 + X_3)$ bases is markedly degraded.

4.3. Model 3

For model 3, the distribution of $R^2$ of each base obtained by generating $n = 100$ observations from each distribution and methods can be seen from Figure 3. Under normal condition, making secondary $f_i$ performs relatively better for the basis of $X_1$ than using categorization among the six methods. Otherwise, for the base $(X_2 + X_3)$ under non-normal condition, the mean $R^2$ was lower than the simulation results conducted under the normal distribution, but for the base $X_1$, $R^2$ was higher.

4.4. Model 4

For model 4, the boxplots in Figure 4 show the distribution of $R^2$ of each base and methods obtained by generating $n = 100$ observations from each distribution. Since there is no quadratic term of $X_1$ in model 4, all methods have well estimated the basis for $X_1$, and the mean $R^2$ is all 0.9 or more. For the $(X_2 + X_3)$ basis, the method of making $cov2$ and quadratic $f_i$ shows relatively good performance. This
is because model 4 contains $1/(X_2 + X_3)^2$, so it is effective to use $f_i$ with squares over categorization. However, under non-normal condition, overall mean of every basis gets lower than the simulation results under the normal distribution.

4.5. Model 5

From the boxplot in Figure 5, we can see the distribution of $R^2$ of each base obtained by generating $n = 100$ observations according to distribution conditions for Model 5. Since there is no quadratic term of $X_1$ in Model 5, under normal condition, all methods well estimated the basis for $X_1$, and the mean $R^2$ is all over 0.9. For $(X_2 + X_3)$ basis, the method of making cov2 and quadratic $f_i$ shows relatively good performance. Otherwise, all bases are not estimated better under non-normal condition than normal condition.

4.6. Model 6

The distribution of $R^2$ of each base obtained by generating $n = 100$ observations from each distribution and six methods for model 6 can be seen from the boxplot in Figure 6. In Model 6, since $X_1$ is a linear term, it can be seen that all six methods estimate the basis of $X_1$ well, and the mean $R^2$ is all 0.95 or more. However, for the $(X_2 + X_3)$ basis, the method of making cov2 and quadratic $f_i$ shows relatively good performance. This is because $\exp(X_2 + X_3)$ in Model 6 is a nonlinear function, so it is effective to use the quadratic $f_i$. Under non-normal distribution, it shows a similar tendency to the simulation results under the normal condition but returns lower $R^2$ than previous result.
4.7. Remarks on numerical studies

From our simulation studies, covK would be better among the other five SDR methods. The unstructured PFC with slicing $f_y$ and polynomial $f_y$ have similar results to SIR and covK, respectively. Although the structured PFC performed well under some models, it had high variability in most of the models. Therefore, the adaptation of the structured PFC in practical way would be possible cause of concern because of its fluctuate performance.

5. Real data example - wheat data

A real data example is presented to compare performances among the six dimension reduction methods. We considered wheat protein data in Fearn (1983), and six continuous variables were used as predictors for our analysis, which are the log-scaled near infrared reflectance detected by six wavelengths. The response variable is protein percentage of sample ground wheat. Therefore, we have $n = 50$, $p = 6$ for data, also we set the $d = 1$. As a result, Figure 7(a) shows that every methods performed well and all the first sufficient predictors are correlated almost perfectly with each other. Furthermore, from Figure 7(b), linear trends for the second sufficient predictors between uPFCp and covK and between uPFCch and SIR are observed. These results also demonstrate the connection between model-free methods and unstructured PFC by choosing adequate $f_y$. However, structured PFCs seem to have no relation with other methods with respect to the second sufficient predictors. This indicates the concern observed in the numerical studies for the structured PFC.
Figure 6: Summary boxplots for model 6.

(a) $X_1$ from normal distribution
(b) $X_2 + X_3$ from normal distribution
(c) $X_1$ from non-normal distribution
(d) $X_2 + X_3$ from non-normal distribution

Figure 7: Scatterplot matrix of sufficient reductions of six methods for wheat data.

(a) First sufficient reduction $\eta_1^T X$
(b) Second sufficient reduction $\eta_2^T X$

6. Discussion

Cook (2007) proposed a semi-parametric dimension reduction approach and developed models from the isotonic model to general principal fitted component (PFC) models, which is the unstructured PFC model. Cook (2007) also connected the PFC with two types of nonparametric SDR methods of sliced
Intensive comparison of semi- and non-parametric dimension reduction methods in forward regression

In this paper, additionally, a connection between the unstructured PFC and covariance method (Yin and Cook, 2002) are newly established by choosing $f_y$ as polynomials of $y$.

Intensive numerical studies are done to compare performances among six selected methods including SIR, covK, two variations of the structured PFC and the unstructured PFC under forward regression situation for practical adaptation. Numerical studies show that the structured and unstructured PFC with the two choices of $f_y$ yields similar estimation results to SIR and covK. However, it would be covK that returns more stable performances regardless of the simulation settings. Among the PFC models, the unstructured PFC has less variability in the estimation of the central subspace than the structured PFCs. Also, the real data example supports these results.

References

Cook RD (1998). *Regression Graphics*, John Wiley & Sons, New York.
Cook RD (2007). Fisher lecture: dimension reduction in regression, *Statistical Science*, 22, 1–26.
Cook RD and Forzani L (2009). Principal fitted components for dimension reduction in regression, *Statistical Science*, 485, 485–501.
Cook RD, Li B, and Chiaromonte F (2007). Dimension reduction in regression without matrix inversion, *Biometrika*, 94, 569–584.
Cook RD and Weisberg S (1994). *An Introduction to Regression Graphics*, John Wiley & Sons, New York.
Cook RD and Zhang X (2014). Fused estimators of the central subspace in sufficient dimension reduction, *Journal of the American Statistical Association*, 109, 815–827.
Fearn T (1983). A misuse of ridge regression in the calibration of a near infrared reflectance instrument, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 32, 73–79.
Li KC (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein’s lemma, *Journal of the American Statistical Association*, 87, 1025–1039.
Li L and Yin X (2008) Sliced inverse regression with regularizations, *Biometrics*, 64, 124–131.
Yin X and Cook RD (2002). Dimension reduction for the conditional kth moment in regression, *Journal of Royal Statistical Society Series B*, 64, 159–175.
Yoo JK (2016). Tutorial: Methodologies for sufficient dimension reduction in regression, *Communications for Statistical Applications and Methods*, 23, 95–117.
Yoo JK (2018a). Response dimension reduction: model-based approach, *Statistics*, 52, 409–425.
Yoo JK (2018b). Partial least squares fusing unsupervised learning, *Chemometrics and Intelligent Laboratory Systems*, 175, 82–86.

Received June 2, 2022; Revised July 15, 2022; Accepted July 22, 2022