Communications for Statistical Applications and Methods
2024, Vol. 31, No. 2, 191–202

https://doi.org/10.29220/CSAM.2024.31.2.191
Print ISSN 2287-7843 / Online ISSN 2383-4757

Applications of response dimension reduction in large p-small n problems

Minjee Kim[a], Jae Keun Yoo[1,a]

[a]Department of Statistics, Ewha Womans University, Korea

Abstract

The goal of this paper is to show how multivariate regression analysis with high-dimensional responses is facilitated by the response dimension reduction. Multivariate regression, characterized by multi-dimensional response variables, is increasingly prevalent across diverse fields such as repeated measures, longitudinal studies, and functional data analysis. One of the key challenges in analyzing such data is managing the response dimensions, which can complicate the analysis due to an exponential increase in the number of parameters. Although response dimension reduction methods are developed, there is no practically useful illustration for various types of data such as so-called large p-small n data. This paper aims to fill this gap by showcasing how response dimension reduction can enhance the analysis of high-dimensional response data, thereby providing significant assistance to statistical practitioners and contributing to advancements in multiple scientific domains.

Keywords: high-dimensional data analysis, large p-small n data, model-based reduction, multivariate regression, response dimension reduction

1. Introduction

In the domain of statistical analysis, sufficient dimension reduction (SDR) has gained prominence as an effective tool for dimension reduction in regression contexts. SDR focuses on transforming predictor variables into a lower-dimensional linear format while preserving essential information pertinent to regression. Moreover, the application of multivariate regression, denoted as $\mathbf{Y} \in \mathbb{R}^r | \mathbf{X} \in \mathbb{R}^p$ with $r \geq 2$, is becoming widespread across numerous disciplines, as evidenced by the works of Li *et al.* (2011), Im *et al.* (2015), Lee *et al.* (2019), and Ko and Yoo (2022).

Principal component analysis (PCA) stands as a notable method for dimension reduction, frequently utilized in regression to diminish the dimensions of regressors. However, PCA's limitation in not accounting for response variables can lead to significant information loss during the dimension reduction process. As a result, various SDR techniques like sliced inverse regression (Li, 1991), sliced average variance estimation (Cook and Weisberg, 1991), contour regression (Li *et al.*, 2005), and directional regression (Li and Wang, 2007) have been developed within the realm of multivariate regression to adeptly capture the variability in $\mathbf{Y}$ while compressing the dimensions of $\mathbf{X}$. Further advancements and applications of these techniques in multivariate regression have been extensively

For Jae Keun Yoo and Minjee Kim, this work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education (RS-2023-00240564 and RS-2023-00217022).

[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 31 March 2024 / journal homepage: http://csam.or.kr

explored by Cook and Setodji (2003), Yoo and Cook (2007), Yoo *et al.* (2010), Yoo (2013), and Yoo (2018).

Nevertheless, the drive to reduce the dimensions of response variables in multivariate regression should be motivated. When the sample sizes (n) of data are considerably smaller than the dimensions of predictors (p) or responses (r), it exacerbates the complexity of statistical analysis and leads to an exponential increase in the number of parameters, often resulting in the curse of dimensionality. In addressing the challenges of large p-small n problems, this study draws upon seminal works such as those by Hung and Huang (2019) and Yin and Hilafu (2015), which have contributed to the development and application of response dimension reduction techniques in such contexts.

Recognizing the complexities inherent in multivariate regression analysis, this paper endeavors to demonstrate the effective use of response dimension reduction techniques. It focuses on comparing predictor and response dimensions with sample sizes, illustrating how these variables interact in different analytical scenarios. The paper blends practical application with theoretical advancements, highlighting the evolution and pivotal role of response dimension reduction methodologies in this field.

Yoo and Cook (2008) established a foundation for response dimension reduction in multivariate regression, ensuring no loss of information on $E(\mathbf{Y}|\mathbf{X})$. This approach delineated two forms of response dimension reduction—linear and conditional—and introduced a non-parametric technique for estimating linear response reduction. Yoo (2013) explored the theoretical relationship between linear and conditional response reduction in the envelope model context, as proposed by Cook *et al.* (2010). This inquiry set the stage for model-based response dimension reduction. Building upon this, Yoo (2018) proposed two novel model-based methods for response dimension reduction, showcasing their advantages over previous methods through both numerical studies and real data applications. Further advancing in this field, Yoo (2019) expanded upon the covariance structure from Yoo (2018) and developed new theoretical frameworks for the reduction subspace and its estimation, also supported by detailed simulation studies that demonstrate the efficacy of these advancements.

The structure of this paper is designed to provide a comprehensive understanding of response dimension reduction in multivariate regression. Section 2 introduces four key methodologies in response dimension reduction, followed by Section 3, which details the implementation strategies considering variable dimensions and sample sizes. Section 4 demonstrates the practical application of these methodologies through the analysis of real datasets, and the paper concludes in Section 5 with a summary of the findings and insights gained from this research.

Throughout this paper, specific notation conventions are adopted for clarity and consistency. A random variable with p dimensions, represented as $\mathbf{X} \in \mathbb{R}^p$ without any explicit mention. For random variables $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^r$, their covariances are denoted as $\text{cov}(\mathbf{X}) = \mathbf{\Sigma}_x$ and $\text{cov}(\mathbf{Y}) = \mathbf{\Sigma}_y$, respectively. It is assumed that both $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_y$ are positive-definite matrices. Furthermore, the notation $\mathcal{S}(\cdot)$ is used to denote the subspace spanned by the columns of a given matrix.

2. Response dimension reduction methodologies

2.1. Non-parametric method

Consider a multivariate regression scenario with $\mathbf{Y} \in \mathbb{R}^r|\mathbf{X} \in \mathbb{R}^p$. We propose the existence of a $r \times q$ matrix $\mathbf{L}$, which has the lowest rank while fulfilling the relation for $E(\mathbf{Y}|\mathbf{X})$ as follows:

$$E\left(\mathbf{Y} \mid \mathbf{X}\right) = E\left\{\mathbf{P}^{\mathrm{T}}_{\mathbf{L}(\mathbf{\Sigma}_y)}\mathbf{Y} \mid \mathbf{X}\right\}, \tag{2.1}$$

where $q \leq r$ and $\mathbf{P}_{\mathbf{L}(\boldsymbol{\Sigma}_y)} = \mathbf{L}(\mathbf{L}^\intercal\boldsymbol{\Sigma}_y\mathbf{L})^{-1}\mathbf{L}^\intercal\boldsymbol{\Sigma}_y$ serves as an orthogonal projection operator relative to the inner product $< \omega_1, \omega_2 >_{\boldsymbol{\Sigma}_y} = \omega_1^\intercal\boldsymbol{\Sigma}_y\omega_2$.

Equation (2.1) implies that the impact of predictors $\mathbf{X}$ on the conditional mean $E(\mathbf{Y}|\mathbf{X})$ is mediated only through $\mathbf{P}_{\mathbf{L}(\boldsymbol{\Sigma}_y)}$. Hence, the transformed $\mathbf{P}^\intercal_{\mathbf{L}(\boldsymbol{\Sigma}_y)}\mathbf{Y}$ effectively replaces the original response $\mathbf{Y}$, without loss information about $E(\mathbf{Y}|\mathbf{X})$. This concept is referred to as linear response reduction in Yoo and Cook (2008).

Further, suppose a $r \times k$ matrix $\mathbf{K}$ exists, satisfying the equivalences in:

$$E(\mathbf{Y} \mid \mathbf{X}) = E\{E(\mathbf{Y} \mid \mathbf{X}, \mathbf{K}^\intercal\mathbf{Y}) \mid \mathbf{X}\} = E\{E(\mathbf{Y} \mid \mathbf{K}^\intercal\mathbf{Y}) \mid \mathbf{X}\} = E\{g(\mathbf{K}^\intercal\mathbf{Y}) \mid \mathbf{X}\}, \tag{2.2}$$

where $k \leq r, \mathbf{K} \neq \mathbf{I}_r$ and $g(\cdot)$ represents an unknown function.

By equation (2.2), another dimension reduction of $\mathbf{Y}$ can be done, and this response reduction is called conditional response reduction when $k < r$. As detailed in Yoo and Cook (2008), the column spaces of matrices $\mathbf{L}$ and $\mathbf{K}$ are identified as a response dimension reduction subspace.

Yoo and Cook (2008) demonstrate $\mathcal{S}(\mathbf{K}) \subseteq \mathcal{S}(\mathbf{L})$ and $\mathcal{S}(\mathbf{K}) = \mathcal{S}(\mathbf{L})$ under certain conditions: $E(\mathbf{Y}|\mathbf{K}^\intercal\mathbf{Y} = a)$ is linear in a. This condition holds when $\mathbf{Y}$ follows an elliptical distribution. Under the condition, the quantity $\boldsymbol{\Sigma}_y^{-1}\text{cov}(\mathbf{Y}, \mathbf{X}^\intercal)\boldsymbol{\Sigma}_x^{-1}$ is proposed in Proposition 3 of Yoo and Cook (2008) to estimate $\mathbf{L}$ and $\mathbf{K}$. We propose to refer to such non-parametric dimension reduction method as the YC-method, and it needs the inverses of $\hat{\boldsymbol{\Sigma}}_x$ and $\hat{\boldsymbol{\Sigma}}_y$.

2.2. Principal response reduction

In this section, we analyze a multivariate regression model under the assumption that both $E(\mathbf{Y})$ and $E(\mathbf{X})$ are zero without loss of generality:

$$\mathbf{Y} = \boldsymbol{\Gamma}v_x + \boldsymbol{\varepsilon}. \tag{2.3}$$

Here, $\boldsymbol{\Gamma}$ is an orthogonal matrix belonging to $\mathbb{R}^{r\times d}$, where $d \leq r$, $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$, and $\text{cov}(v_x, \boldsymbol{\varepsilon}) = 0$. The function v_x represents a d-dimensional unknown random function of the predictors $\mathbf{X}$, characterized by a positive definite sample covariance, and $\boldsymbol{\Sigma}_x v_x = 0$. Notably, if $v_x = \mathbf{X}$, this model aligns with the standard multivariate linear regression.

A key premise of the model (2.3) is that the subspace $\mathcal{S}(\boldsymbol{\Gamma})$ is invariant and acts as a reducing subspace for $\boldsymbol{\Sigma}$. This condition ensures the decomposition of $\boldsymbol{\Sigma}$ into $\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^\intercal + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^\intercal$, with $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{r\times(r-d)}$, $\boldsymbol{\Gamma}_0^\intercal\boldsymbol{\Gamma}_0 = \mathbf{I}_{r-d}$, and $\boldsymbol{\Gamma}_0^\intercal\boldsymbol{\Gamma} = 0$. Here, $\boldsymbol{\Omega}$ represents $\boldsymbol{\Gamma}^\intercal\boldsymbol{\Sigma}\boldsymbol{\Gamma}$, and $\boldsymbol{\Omega}_0$ is $\boldsymbol{\Gamma}_0^\intercal\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0$.

As per Yoo (2018), within this model (2.3), it is established that $E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{P}^\intercal_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_y)}\mathbf{Y}|\mathbf{X})$. Essentially, this implies that the response $\mathbf{Y}$ can be effectively condensed using $\boldsymbol{\Gamma}$, without compromising the informational value of $E(\mathbf{Y}|\mathbf{X})$.

The estimation of $\boldsymbol{\Gamma}$ in this model involves maximizing the likelihood function, under the assumption of $\boldsymbol{\varepsilon}$ being normally distributed. Here, $\hat{\boldsymbol{\Sigma}}_y$ is utilized as the conventional moment estimator for $\boldsymbol{\Sigma}_y$. Yoo (2018) demonstrates that the maximum likelihood estimator (MLE) for $\boldsymbol{\Gamma}$ is derived from the eigenvectors corresponding to the first d largest eigenvalues of $\hat{\boldsymbol{\Sigma}}_y$. This method of dimension reduction is known as principal response reduction (PRR).

2.3. Principal fitted response reduction

In PRR, the estimation of $\boldsymbol{\Gamma}$ does not consider the predictors $\mathbf{X}$. To integrate $\mathbf{X}$ into the model, we assume $v_x = \boldsymbol{\psi}\mathbf{f}_x$:

$$\mathbf{Y} = \boldsymbol{\Gamma}\boldsymbol{\psi}\mathbf{f}_x + \boldsymbol{\varepsilon}. \tag{2.4}$$

Here, $\boldsymbol{\psi}$ represents an unknown $d \times q$ matrix, and $\mathbf{f}_x$, a q-dimensional known vector function of $\mathbf{X}$, satisfies $\boldsymbol{\Sigma}_x \mathbf{f}_x = 0$. For clarity, we introduce the following notations:

$\mathbb{Y}$: an $n \times r$ matrix representing response data.

$\mathbb{X}$: an $n \times p$ matrix representing predictor data.

$\mathbb{F}$: a $n \times q$ matrix formed by stacking $\mathbf{f}_x^\intercal$, with $\mathbf{P}_\mathbb{F} = \mathbb{F}(\mathbb{F}^\intercal\mathbb{F})^{-1}\mathbb{F}^\intercal$.

$\hat{\boldsymbol{\Sigma}}_{\text{fit}} = \mathbb{Y}^\intercal \mathbf{P}_\mathbb{F} \mathbb{Y}/n$ and $\hat{\boldsymbol{\Sigma}}_{\text{res}} = \hat{\boldsymbol{\Sigma}}_y - \hat{\boldsymbol{\Sigma}}_{\text{fit}}$.

The MLE for $\boldsymbol{\Gamma}$ in model (2.4) does not have a closed-form solution. The likelihood function for $\boldsymbol{\Gamma}$, as per Yoo (2018), is given by:

$$\mathcal{L}(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) = -\frac{n}{2} \log \left| \boldsymbol{\Gamma}_0^\intercal \hat{\boldsymbol{\Sigma}}_y \boldsymbol{\Gamma}_0 \right| - \frac{n}{2} \log \left| \boldsymbol{\Gamma}^\intercal \hat{\boldsymbol{\Sigma}}_{\text{res}} \boldsymbol{\Gamma} \right|.$$

Thus, the estimation of $\boldsymbol{\Gamma}$ is influenced by both $\hat{\boldsymbol{\Sigma}}_y$ and $\hat{\boldsymbol{\Sigma}}_{\text{res}}$. A sequential selection algorithm, which involves choosing from the eigenvectors of $\hat{\boldsymbol{\Sigma}}_y$, $\hat{\boldsymbol{\Sigma}}_{\text{fit}}$, and $\hat{\boldsymbol{\Sigma}}_{\text{res}}$, as recommended by Cook (2007), is employed. This method of estimating $\boldsymbol{\Gamma}$ is termed principal fitted response reduction (PFRR), and it necessaries only the inverse of $\mathbb{F}^\intercal\mathbb{F}$ for reducing the response dimensions.

2.4. Unstructured principal fitted response reduction

In this model, we assume that $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma} > 0)$, and $\text{cov}(v_x, \boldsymbol{\varepsilon}) = 0$:

$$\mathbf{Y} = \boldsymbol{\Gamma} v_x + \boldsymbol{\varepsilon}. \tag{2.5}$$

A key distinction between this model (2.5) and the previous model (2.3) lies in the structure of $\boldsymbol{\Sigma}$. In model (2.5), the condition $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^\intercal + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^\intercal$ is no longer a requisite.

Yoo (2019) illustrates that for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_y$ to meet the invariant condition, $\mathcal{S}(\boldsymbol{\Sigma}\boldsymbol{\Gamma}) \subseteq \mathcal{S}(\boldsymbol{\Gamma})$ if and only if $\mathcal{S}(\boldsymbol{\Sigma}_y\boldsymbol{\Gamma}) \subseteq \mathcal{S}(\boldsymbol{\Gamma})$. Consequently, for model (2.5), it is deduced that $E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{P}^\intercal_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_y)}\mathbf{Y}|\mathbf{X})$ if $\mathcal{S}(\boldsymbol{\Gamma})$ is invariant for $\boldsymbol{\Sigma}_y$. Henceforth, this invariant condition for $\boldsymbol{\Gamma}$ relative to $\boldsymbol{\Sigma}_y$ is assumed.

To integrate the predictor $\mathbf{X}$ into the estimation of $\boldsymbol{\Gamma}$, the model is augmented as:

$$\mathbf{Y} = \boldsymbol{\Gamma}\boldsymbol{\psi}\mathbf{f}_x + \boldsymbol{\varepsilon}. \tag{2.6}$$

The following quantities are defined for this model:

$\mathbf{E}_d$ and $\mathcal{S}_d(\mathbf{E})$: The first d largest eigenvectors of a matrix $\mathbf{E}$ and the column subspace of $\mathbf{E}_d$.

$\mathbf{B} = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\Sigma}}_{\text{fit}}\hat{\boldsymbol{\Sigma}}^{-1/2}$, $\mathbf{B}_{\text{res}} = \hat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\hat{\boldsymbol{\Sigma}}_{\text{fit}}\hat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}$, and $\mathbf{B}_y = \hat{\boldsymbol{\Sigma}}_y^{-1/2}\hat{\boldsymbol{\Sigma}}_{\text{fit}}\hat{\boldsymbol{\Sigma}}_y^{-1/2}$.

$\hat{\boldsymbol{\Lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_q)$ and $\hat{\mathbf{V}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_q)$: Ordered eigenvalues and corresponding eigenvectors of $\mathbf{B}_{\text{res}}$.

$\hat{\mathbf{K}}_d$: $\text{diag}(0, \ldots, 0, \hat{\lambda}_{d+1}, \ldots, \hat{\lambda}_q)$.

Yoo (2019) derives the following results under model (2.6):

(1) $\hat{\mathcal{S}}(\boldsymbol{\Gamma}) = \hat{\boldsymbol{\Sigma}}^{1/2}\mathcal{S}_d(\mathbf{B})$ or equivalently, $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\Sigma}}^{1/2}\mathbf{B}_d$.

Table 1: Feasible response dimension reduction methods according to p, r and n

	Relation between p, r, and n	Methods
Case 1	$\max(p, r) < n$	YC-method, PRR, PFRR, UPFRR
Case 2	$p \geq n$, but $r < n$	PRR
Case 3	$p < n$, but $r \geq n$	PRR, PFRR
Case 4	$\min(p, r) \geq n$	PRR

(2) $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{\text{res}} + \hat{\boldsymbol{\Sigma}}_{\text{res}}^{1/2}\hat{\mathbf{V}}\hat{\mathbf{K}}_d\hat{\mathbf{V}}^{\top}\hat{\boldsymbol{\Sigma}}_{\text{res}}^{1/2} = \hat{\boldsymbol{\Sigma}}_{\text{res}}^{1/2}(\mathbf{I}_r + \hat{\mathbf{V}}\hat{\mathbf{K}}_d\hat{\mathbf{V}}^{\top})\hat{\boldsymbol{\Sigma}}_{\text{res}}^{1/2}$.

(3) $\mathcal{L}^d_{\mathcal{UPFRR}} = (-n/2)\log|\hat{\boldsymbol{\Sigma}}_{\text{res}}| + (n/2)\Sigma_{i=d+1}^{q}\log(1 + \hat{\lambda}_i)$.

(4) $\hat{\mathcal{S}}(\boldsymbol{\Gamma}) = \hat{\boldsymbol{\Sigma}}^{1/2}\mathcal{S}_d(\mathbf{B}) = \hat{\boldsymbol{\Sigma}}_{\text{res}}^{1/2}\mathcal{S}_d(\mathbf{B}_{\text{res}}) = \hat{\boldsymbol{\Sigma}}_{y}^{1/2}\mathcal{S}_d(\mathbf{B}_y)$.

Thus, it needs the inverses of $\hat{\boldsymbol{\Sigma}}_{\text{res}}$ and $\mathbb{F}^{\top}\mathbb{F}$ to reduce the response dimensions, related to $\mathbf{B}_{\text{res}}$. This method in model (2.6) is termed unstructured principal fitted response reduction (UPFRR).

3. Implementation along with sample sizes

The selection of an appropriate response dimension reduction method hinges on the dimensions of the predictors (p) and responses (r), relative to the sample size (n). Notably, the YC-method and UPFRR require the inverses of $\hat{\boldsymbol{\Sigma}}_x$ and $\hat{\boldsymbol{\Sigma}}_y$, and the inverses of $\hat{\boldsymbol{\Sigma}}_{\text{res}}$ and $\mathbb{F}^{\top}\mathbb{F}$, respectively. Therefore, these methods are only feasible when both p and r are smaller than n. In contrast, the PFRR method demands the invertibility of $\mathbb{F}^{\top}\mathbb{F}$ alone, necessitating $p < n$ to satisfy full rank. The PRR method, due to the spectral decomposition of $\hat{\boldsymbol{\Sigma}}_y$, does not impose specific constraints onto p, r, and n. Table 1 summarizes the feasible methods based on the relationships between p, r, and n.

To align scenarios in **cases 2–4** with the more generalized model of **case 1** (where both predictors and responses are smaller than the sample size), initial dimension reduction steps are necessary. These preliminary steps are crucial in establishing conditions where methods such as the YC-method, PFRR, and UPFRR become applicable.

In **case 2**, where $p \geq n$ and $r < n$, reducing the dimensions of predictors is essential. This can be achieved through multivariate seeded dimension reduction, as suggested by Yoo and Im (2014), thereby enabling the application of the YC-method, PFRR, and UPFRR.

For **case 3**, with $p < n$ and $r \geq n$, the initial reduction of response dimensions is imperative. Employing PRR and PFRR for this purpose allows the subsequent use of the YC-method and UPFRR.

In **case 4**, where both p and r are greater than or equal to n, PRR is the only applicable method. Starting with PRR for initial response reduction, reducing the dimensions of predictors similar to **case 2** facilitates the application of the YC-method, PFRR, and UPFRR.

Since **cases 2–4** necessitate initial reduction steps, it's crucial to meticulously examine and compare the final response dimension reduction outcomes to assure their validity and accuracy.

The **seedCCA** package, as discussed in Kim *et al.* (2021), offers two datasets: Near-infrared (NIR) spectroscopy of biscuit doughs and Nutrimouse data, exemplifying **cases 2** and **3**, respectively. In **case 2**, the scenario converges with **case 1** after reducing the dimensions of predictors. Thus, a distinct example for **case 1** is omitted in this paper, as it is extensively covered in Yoo and Cook

Table 2: UPFRR and PFRR results for `cookie` data

(a) UPFRR test summary

Comparison	Stat	df	p-value
0D vs ≥ 1D	340.9	32	<0.0001
1D vs ≥ 2D	159.4	21	<0.0001
2D vs ≥ 3D	27.39	12	0.0068
3D vs ≥ 4D	9.508	5	0.0904

(b) PFRR test summary

Comparison	Stat	df	p-value
0D vs ≥ 1D	340.9	32	<0.0001
1D vs ≥ 2D	196.1	24	<0.0001
2D vs ≥ 3D	63.80	16	<0.0001
3D vs ≥ 4D	11.38	8	0.1809

Stat, df, p-value are from Asymp. Chi-square tests for dimension.

Table 3: Cumulative variance of 1–4 dimensions with each response dimension reduction method

Method	1D	2D	3D	4D
UPFRR	0.1136	0.3470	0.6676	1.0000
PFRR	0.1915	0.4281	0.7059	1.0000
PRR	0.7085	0.9902	1.0000	1.0000

Cumulative variance values are calculated based on the proportion of eigenvalues.

(2008) and Yoo (2018). **Case 4**, while not explicitly exemplified, can be understood as a sequential process—initially applying PRR as in **case 3**, followed by reducing the dimensions of predictors akin to **case 2**.

4. Real data applications

4.1. Near-infrared spectroscopy of biscuit doughs data (**case 2**: $p \geq n$ and $r < n$)

The `cookie` dataset from the **seedCCA** package (Kim *et al.*, 2021) consists of quantitative near-infrared (NIR) spectroscopy measurements from biscuit dough experiments, which aims to test the feasibility of NIR spectroscopy. This dataset focuses on four key ingredients: Fat ($Y1$), sucrose ($Y2$), dry flour ($Y3$), and water ($Y4$), whose percentages vary from the standard recipe and constitute the response variables. The predictors are wavelengths measured by spectroscopy, spanning from 1380 to 2400nm at 4nm intervals, resulting in 256 dimensions.

In the `cookie` dataset, the first 40 samples form the training set, excluding the 23^{rd} one identified as an outlier. The remaining 32 samples serve as the test set, which also excludes an outlier, the 21^{st} observation. Thus, the dimensions of the training and test sets for the predictors are 39×256 and 31×256, respectively, while for the response variables, the dimensions are 39×4 for the training set and 31×4 for the test set.

Given the higher number of predictors compared to the sample size in the training data, dimension reduction of predictors is necessary for applying UPFRR and PFRR. This reduction is performed using multivariate seeded reduction (Yoo and Im, 2014). During this process and subsequent analysis, the response variables are standardized to have zero mean and unit variance, ensuring numerical stability in the predictor dimension reduction and the partial least squares fit. The `pseudoinverse` function from the **corpcor** package (Schafer *et al.*, 2017) is utilized for computing a general inverse of matrices. The analysis shows that four iterations of projections are required, reducing the 256-dimensional predictors to four dimensions, denoted as `Rx`.

With these dimension-reduced predictors, UPFRR, PFRR, and PRR are applied. In this analysis, UPFRR and PFRR both suggest reducing the response to three dimensions, with a 5% significance level as shown in Table 2. In contrast, the PRR analysis reveals a variance distribution that differs from the results of UPFRR and PFRR. The first two components of PRR account for approximately 70.85%

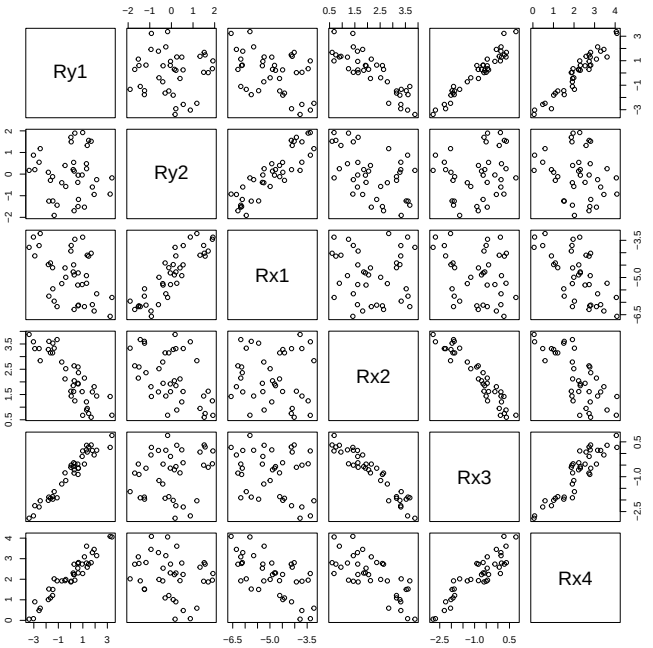


Figure 1: *Scatter plot matrix of (`Ry,Rx`) for `cookie` data.*

and 99.02% of the cumulative variance as shown in Table 3, implying that UPFRR and PFRR may overestimate the number of necessary dimensions. To investigate, the trace correlation coefficients (Hooper, 1959) are calculated between the first two and three components of UPFRR, PFRR, and PRR. Here, given two matrices A_1 and A_2, and a parameter q representing the number of principal components to consider, the trace correlation $\bar{r}$ can be calculated using the formula:

$$\bar{r} = \sqrt{\frac{\sum_{i=1}^{q} \lambda_i}{q}}, \tag{4.1}$$

where λ_i are the eigenvalues of the matrix $P = A_1^{\top} A_2 A_2^{\top} A_1$. The trace correlation coefficients of the first two components of UPFRR and PFRR, UPFRR and PRR, PFRR and PRR are 0.9247, 0.9218, and 1, respectively, and for the first three components, all of them equal to 1. This similarity suggests that UPFRR, PFRR, and PRR yield comparable reduction results, indicating that a two-dimensional response might suffice for regression, as implied by PRR's cumulative eigenvalue proportions. Thus, the two-dimensional responses derived from PFRR are used in place of the original four-dimensional responses.

Interpreting the reduced responses through their estimated coefficients, the first component appears to be a linear combination of sucrose ($Y2$), dry flour ($Y3$), and water ($Y4$), while the second component primarily represents fat ($Y1$).

Four models are fitted for prediction using the test data. Initially, two multivariate linear regressions are performed: One with both dimension-reduced responses and predictors (`Ry,Rx`), and another with original responses alongside dimension-reduced predictors (`Y,Rx`). The scatter plot matrix for

Table 4: MSPE results for `cookie` data

Model	$Y1$	$Y2$	$Y3$	$Y4$
(Ry,Rx)	0.0901	0.0462	0.0452	0.0471
(Y,Rx)	0.0937	0.0463	0.0506	0.0470
(Ry,X)	0.0397	0.0418	0.0462	0.0400
(Y,X)	0.0396	0.0418	0.0495	0.0392

Table 5: Cumulative variance of 1–6 dimensions with PRR method for `nutrimouse` data

Dimension	1D	2D	3D	4D	5D	6D
Variance	0.4038	0.6949	0.8623	0.9554	0.9726	0.9835

(Ry,Rx) is included in Figure 1. For comparison, Ry are subsequently converted back to their four-dimensional from two-dimensional form.

Next, multivariate partial least squares regression is conducted in two variations:
One with dimension-reduced responses and original predictors (Ry,X), and the other with original responses and predictors (Y,X). For these latter two fits, the `mvr` function from the **pls** package (Liland *et al.*, 2021) is utilized, with cross-validation suggesting that 6 components are optimal.

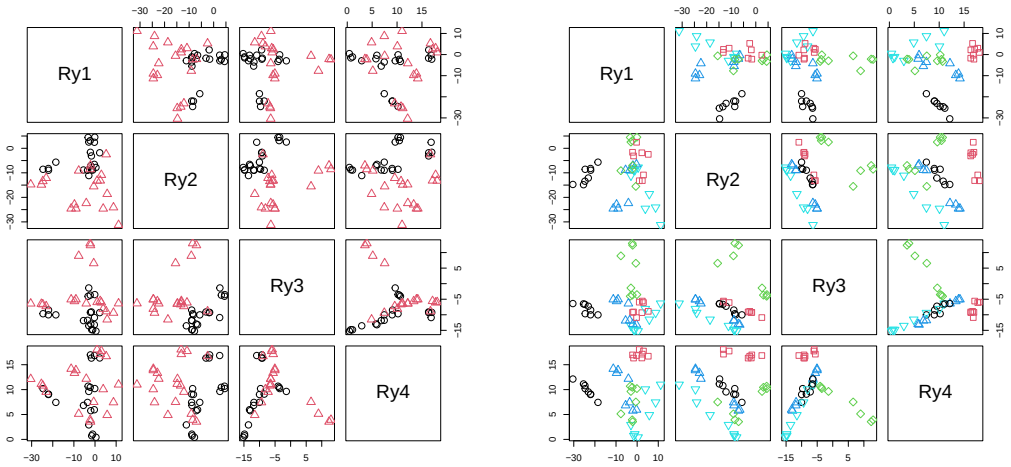
To evaluate the prediction accuracy in the test data, the mean squared prediction error (MSPE = $(1/(n \times r)) \sum_{i=1}^{n} \sum_{j=1}^{r} (y_{ij} - \hat{y}_{ij})^2$) is calculated for each of the models ((Ry,Rx), (Y,Rx), (Ry,X), (Y,X)) and compared (Table 4). It is observed that using dimension-reduced predictors slightly compromises accuracy, particularly for $Y1$ (fat). However, this reduction in predictor dimensions does not significantly impact the accuracy of response dimension reduction. Comparing the fits of (Ry,Rx) with (Y,Rx) and (Ry,X) with (Y,X), no substantial differences are noted in prediction accuracy. Interestingly, for $Y3$ (dry flour), using dimension-reduced responses actually enhances prediction accuracy in both cases, albeit marginally. This suggests that response reduction in high-dimensional data can streamline modeling processes without sacrificing prediction accuracy compared to using original responses.

4.2. Nutrimouse data (**case 3** : $p < n$ and $r \geq n$)

The nutrimouse dataset, integral to a nutritional study on mice, is accessible through `nutrimouse` in the **seedCCA** package (Kim *et al.*, 2021). This dataset encompasses data from 40 mice, featuring measurements of 120 genes and 21 lipids. The mice are classified into two genotypes: Wild type and PPARalpha. Additionally, they are subjected to five different diets: COC (hydrogenated coconut oil for a saturated FA diet), FISH (corn, colza, enriched fish oils), LIN (linseed oil for an ω3-rich diet), REF (corn and colza oils), and SUN (sunflower oil for an ω6 FA-rich diet), with each diet administered to 8 distinct mice.

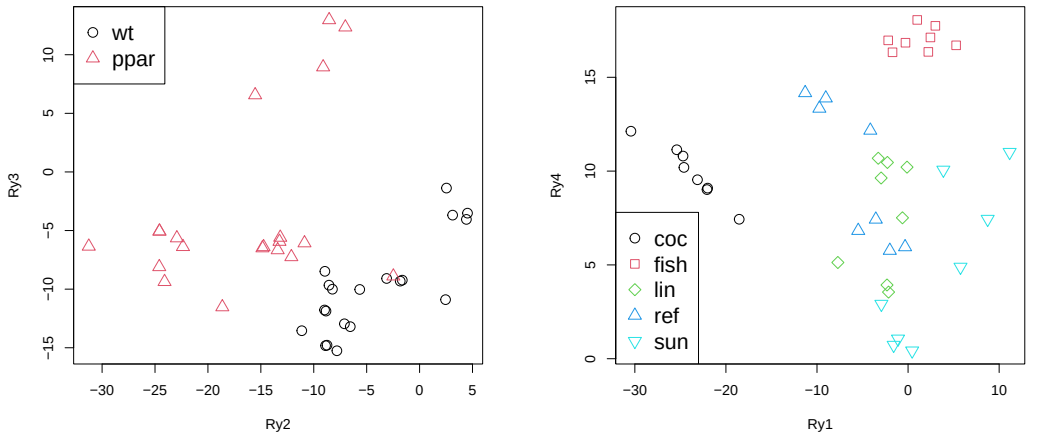
The primary objective is to ascertain if there are significant differences in gene and lipid profiles based on genotype and diet. The response variables are genes and lipids, totaling 141 dimensions, which notably exceed the sample size ($n = 40$). This dimensionality precludes the use of conventional multivariate analysis of variance for examining genotype and diet influences. Hence, dimension reduction of the response variables is necessary. Given that all predictors are categorical, PFRR is not the most suitable option, despite being feasible. Therefore, PRR is the preferable choice in this context. By applying PRR, the first four components capture over 95% of the total variation as shown in Table 5, effectively reducing the responses from 141 dimensions to four.

Scatterplot matrices of the four-dimensionally reduced responses (Ry), labeled by genotype and diet, are illustrated in Figure 2. As depicted in Figure 2(a), Ry2 and Ry3 distinctly separate genotypes,



(a) *Scatter plot matrix marked by genotype* (b) *Scatter plot matrix marked by diet*

Figure 2: *Scatter plot matrices of the reduced responses marked by genotype and diet.*



(a) *Plot of Ry2 and Ry3 marked by genotype* (b) *Plot of Ry1 and Ry4 marked by diet*

Figure 3: *Scatter plots of Ry2 and Ry3 marked by genotype and of Ry1 and Ry4 marked by diet.*

while Ry1 and Ry4 more effectively differentiate diets in Figure 2(b). This suggests that Ry2 and Ry3 are primarily associated with genotype, whereas Ry1 and Ry4 correlate more with diet. Moreover, Figure 3 hints at a potential interaction between genotype and diet, with detailed views of the scatter plots for Ry2 and Ry3 marked by genotype, and Ry1 and Ry4 marked by diet.

Subsequently, a multivariate analysis of variance, incorporating genotype, diet, and their interaction, is conducted using the four-dimensional responses. Table 6 indicates that all effects are significant at a 5% level.

Table 6: MANOVA results for `nutrimouse` data

Effect	df	Pillai's trace	Approx. F	Pr(>F)
Genotype	1	0.9070	65.825	1.574e-13 ***
Diet	4	3.0805	25.125	< 2.2e-16 ***
Genotype:Diet	4	2.0993	8.284	4.339e-13 ***
Residuals	30	-	-	-

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Table 7: Pairwise comparison results for `nutrimouse` data

(a) Pairwise comparison in wild type

Contrast	Estimate	df	t-ratio	p-value
coc-fish	−8.935	15	−23.833	<.0001
coc-lin	−9.714	15	−25.912	<.0001
coc-ref	−3.401	15	−9.072	<.0001
coc-sun	−1.627	15	−4.341	0.0006
fish-lin	−0.779	15	−2.079	0.0552
fish-ref	5.534	15	14.762	<.0001
fish-sun	7.308	15	19.492	<.0001
lin-ref	6.313	15	16.841	<.0001
lin-sun	8.087	15	21.571	<.0001
ref-sun	1.774	15	4.731	0.0003

(b) Pairwise comparison in PPARalpha

Contrast	Estimate	df	t-ratio	p-value
coc-fish	−9.817	15	−15.126	<.0001
coc-lin	−9.370	15	−14.437	<.0001
coc-ref	−2.794	15	−4.305	0.0008
coc-sun	−4.435	15	−6.833	<.0001
fish-lin	0.447	15	0.689	0.5013
fish-ref	7.023	15	10.821	<.0001
fish-sun	5.382	15	8.293	<.0001
lin-ref	6.576	15	10.132	<.0001
lin-sun	4.935	15	<.0001	<.0001
ref-sun	−1.641	15	0.0258	0.0258

P-value adjustments calculated based on fdr method with 10 tests.

To further explore the dietary factor-level means, the dataset is segregated based on genotype, resulting in two distinct groups. A pairwise comparison is conducted in Table 7 to assess the dietary effects. This comparison is carried out using the `test` function in the **lsmeans** package (Lenth, 2016), with the false discovery rate controlled at 10% as per the method suggested by Benjamini and Hochberg (1995). Table 7(a) reveals that in the wild type genotype, all dietary variations are significantly different from each other. Conversely, for the PPARalpha genotype, no notable distinction is observed between the fish and linseed oil diets (Table 7(b)). This disparity in dietary response of genotypes underscores the significance of the interaction between genotype and diet.

5. Discussion

This paper focuses on the challenges presented by the multi-dimensionality of response variables in the context of multivariate regression, specifically where $\mathbf{Y} \in \mathbb{R}^r | \mathbf{X} \in \mathbb{R}^p$. The recent advancements in methodologies for response dimension reduction, which aim to preserve information in $E(\mathbf{Y}|\mathbf{X})$, mark a significant step forward in addressing these complexities. These methodologies serve to simplify complex data structures without losing essential information, a critical aspect in statistical analysis.

The core of this paper lies in illustrating the practical use of response dimension reduction across different scenarios defined by the relationships between the dimensions of predictors (p), responses (r), and sample sizes (n). Through the analysis of two real datasets - NIR spectroscopy of `cookie` data and `nutrimouse` data - the paper provides insightful outcomes of these methodologies. The study on NIR spectroscopy highlights the necessity of response dimension reduction in cases where the number of predictors considerably surpasses the sample sizes. The analysis indicates that, although dimension reduction may slightly impact the accuracy in certain variables, the overall predictive accuracy is largely maintained. The analysis of the `nutrimouse` data further exemplifies the utility of

response dimension reduction in scenarios where the dimensions of the responses are greater than the sample sizes. The reduction achieved simplifies the model without sacrificing the integrity of the data, enabling clearer interpretations and more efficient statistical modeling.

In both cases, the focus on response dimension reduction streamlines the process of dealing with high-dimensional data, making it more manageable for practitioners across various scientific fields. Particularly, the identification of significant interaction effects in the `nutrimouse` data, which might have been challenging to discern without effective dimension reduction, underscores the value of these methodologies.

While this paper highlights the effectiveness of response dimension reduction methodologies in multivariate regression, it is important to acknowledge limitations. The methodologies may struggle to capture non-linear relationships in scenarios beyond **case 1**, particularly when dealing with high-dimensional data. Additionally, the absence of real data applications for **case 4** and the variable performance of response dimension reduction methods across different datasets point to the need for further research.

Acknowledgements

For Jae Keun Yoo and Minjee Kim, this work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education (RS-2023-00240564 and RS-2023-00217022).

References

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

Cook RD (2007). Fisher lecture: Dimension reduction in regression, *Statistical Science*, **22**, 1–26.

Cook RD, Li B, and Chiaromonte F (2010). Envelope models for parsimonious and efficient multivariate linear regression, *Statistics Sinica*, **20**, 927–960.

Cook RD and Setodji CM (2003). A model-free test for reduced rank in multivariate regression, *Journal of the American Statistical Association*, **98**, 340–351.

Cook RD and Weisberg S (1991). Sliced inverse regression for dimension reduction: Comment, *Journal of the American Statistical Association*, **86**, 328–332.

Hooper JW (1959). Simultaneous equations and canonical correlation theory, *Econometrica: Journal of the Econometric Society*, **27**, 245–256.

Hung H and Huang SY (2019). Sufficient dimension reduction via random-partitions for the large-p-small-n problem, *Biometrics*, **75**, 245–255.

Im Y, Gang H, and Yoo JK (2015). High-throughput data dimension reduction via seeded canonical correlation analysis, *Journal of Chemometrics*, **29**, 193–199.

Kim B, Im Y, and Yoo JK (2021). SEEDCCA: An integrated R-package for canonical correlation analysis and partial least squares, *The R Journal*, **13**, 7–20.

Ko S and Yoo JK (2022). Projective resampling estimation of informative predictor subspace for multivariate regression, *Journal of the Korean Statistical Society*, **51**, 1117–1131.

Lee K, Choi Y, Um HY, and Yoo JK (2019). On fused dimension reduction in multivariate regression, *Chemometrics and Intelligent Laboratory Systems*, **193**, 103828.

Lenth RV (2016). Least-squares means: The R package lsmeans, *Journal of Statistical Software*, **69**, 1–33.

Li B and Wang S (2007). On directional regression for dimension reduction, *Journal of the American*

Statistical Association, **102**, 997–1008.

Li B, Zha H, and Chiaromonte F (2005). Contour regression: A general approach to dimension reduction, *The Annals of Statistics*, **33**, 1580–1616.

Li K-C (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 316–327.

Li X, Gill R, Cooper NG, Yoo JK, and Datta S (2011). Modeling microRNA-mRNA interactions using PLS regression in human colon cancer, *BMC Medical Genomics*, **4**, 1–15.

Liland KH, Mevik B-H, Wehrens R, and Hiemstr P (2021). PLS: Partial Least Squares and Principal Component Regression. R package version 2.8.0. Available from: https://cran.r-project.org/web/packages/pls/pls.pdf

Schafer J, Opgen-Rhein R, Zuber V, Ahdesmaki M, Silva APD, and Strimmer K (2017). corpcor: Efficient Estimation of Covariance and (Partial) Correlation. R package version 1.6.9. Available from: https://cran.r-project.org/web/packages/corpcor/corpcor.pdf

Yin X and Hilafu H (2015). Sequential sufficient dimension reduction for large p, small n problems, *Journal of the Royal Statistical Society, Series B*, **77**, 879–892.

Yoo JK (2023). A theoretical view of the envelope model for multivariate linear regression as response dimension reduction, *Journal of the Korean Statistical Society*, **42**, 143–148.

Yoo JK (2018). Response dimension reduction: Model-based approach, *Statistics*, **52**, 409–425.

Yoo JK (2019). Unstructured principal fitted response reduction in multivariate regression, *Journal of the Korean Statistical Society*, **48**, 561–567.

Yoo JK and Cook RD (2007). Optimal sufficient dimension reduction for the conditional mean in multivariate regression, *Biometrika*, **94**, 231–242.

Yoo JK and Cook RD (2008). Response dimension reduction for the conditional mean in multivariate regression, *Computational Statistics and Data Analysis*, **53**, 334–343.

Yoo JK and Im Y (2014). Multivariate seeded dimension reduction, *Journal of the Korean Statistical Society*, **43**, 559–566.

Yoo JK, Lee K, and Wu S (2010). On the extension of sliced average variance estimation to multivariate regression, *Statistical Methods & Applications*, **19**, 529–540.

Received January 4, 2024; Revised February 5, 2024; Accepted February 5, 2024