



Dealing with big data: comparing dimension reduction and shrinkage regression methods

Hamideh D. Hamedani & Sara Sadat Moosavi

To cite this article: Hamideh D. Hamedani & Sara Sadat Moosavi (2016): Dealing with big data: comparing dimension reduction and shrinkage regression methods, Journal of Applied Statistics

To link to this article: <http://dx.doi.org/10.1080/02664763.2016.1177498>



Published online: 22 May 2016.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

Dealing with big data: comparing dimension reduction and shrinkage regression methods

Hamideh D. Hamedani and Sara Sadat Moosavi

Statistics Department, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran

ABSTRACT

In the past decades, the number of variables explaining observations in different practical applications increased gradually. This has led to heavy computational tasks, despite of widely using provisional variable selection methods in data processing. Therefore, more methodological techniques have appeared to reduce the number of explanatory variables without losing much of the information. In these techniques, two distinct approaches are apparent: 'shrinkage regression' and 'sufficient dimension reduction'. Surprisingly, there has not been any communication or comparison between these two methodological categories, and it is not clear when each of these two approaches are appropriate. In this paper, we fill some of this gap by first reviewing each category in brief, paying special attention to the most commonly used methods in each category. We then compare commonly used methods from both categories based on their accuracy, computation time, and their ability to select effective variables. A simulation study on the performance of the methods in each category is generated as well. The selected methods are concurrently tested on two sets of real data which allows us to recommend conditions under which one approach is more appropriate to be applied to high-dimensional data.

ARTICLE HISTORY

Received 10 January 2015
Accepted 8 April 2016

KEYWORDS

Sufficient dimension reduction; central subspace; SPICE method; shrinkage regression; LASSO; Elastic-Net; FLASH; OSCAR; SCAD; Ridge regression

MSC2010 SUBJECT CLASSIFICATIONS

Primary: 62H20; Secondary: 62J07

1. Introduction

Advances in data collection and storage capabilities during the past few decades have allowed government agencies, researchers, and businesses to store much larger volumes of data often referred to as 'high-dimensional data' or 'big data'. Many of the high-dimensional datasets have a large number of predictor variables, aside from the large sample sizes, that makes the computational analysis of the data very cumbersome if not impossible. Also the large number of predictor variables may often lead to collinearity between variables which reduces their explanatory power. Therefore, one needs to effectively reduce the number of predictor variables without sacrificing the explanatory power of the original ones. Many methodological techniques have appeared to accomplish this goal, within which two distinct approaches are apparent: One approach fits a model to the initial data and then attempts to reduce the number of variables, which are typically referred to as 'shrinkage regression' methods. The other approach, 'sufficient dimension

reduction' (SDR), first selects a subset of effective variables containing most of the information and then an appropriate model is fitted. Different methods within each category have been proposed, but no comparison of methods from these two categories and also within these categories have appeared. Therefore, a question remains as to which method from any of these two categories should be chosen for a given high-dimensional dataset in the first place. In this paper, we attempt to answer this question by analytically and empirically comparing commonly used methods in each category. These analyses allow us to establish under which conditions each of these two category of methods can be more successfully applied to a given big data.

Historically, one of the first attempts in analyzing a moderate to high-dimensional data has been regression analysis. It provides a conceptually simple method for establishing functional relationships among variables. However, sometimes predictor variables are highly collinear, and applying classical regression methods such as ordinary least squares (OLS) to such a large number of variables results in an ill-fitted model with high prediction error. Therefore, once the model is fitted a question remains as to how effective variables should be selected. Along this line, the family of shrinkage regression methods were proposed to deal with this challenge. In this family, the regression coefficients are penalized, so that when fitted to the data, the number of non-zero coefficients in the model are reduced. Different forms of penalty functions have been used in the literature, that vary in their ability to select and group variables and their appropriateness for high-dimensional data.

The first shrinkage regression method, referred to as Ridge regression, was proposed by Hoerl and Kennard [25]. Applying the Ridge regression penalty has the effect of shrinking the estimates of regression coefficients toward zero which introduces bias in their estimation. However, the resulting mean-squared error from Ridge regression tends to be smaller than that of OLS. Ridge regression cannot select effective variables, which is overcome by the introduction of the least absolute shrinkage and selection operator (LASSO) method [35]. However, LASSO method remains biased and cannot select more than n (sample size) variables when the number of predictor variables (p) are much larger than sample size (i.e. $p > n$). Therefore, LASSO estimates often perform poorly for high-dimensional data.

Many methods have focused on addressing various possible shortcomings of the LASSO method, specially when there is dependence or collinearity between predictors. Most commonly used smoothly clipped absolute deviation penalty (SCAD) [2,19,37], Elastic-Net method [40] and octagonal shrinkage and clustering algorithm for regression (OSCAR) [1]. All of these methods use penalties to shrink regression coefficients but have different degree of success in reducing the number of variables. In more detail, SCAD can both select effective variables and obtain unbiased estimates. Elastic-Net removes the restriction faced by LASSO on the number of effective variables chosen and has a strong ability to group predictor variables together (grouping property). Therefore, it is a good candidate method for use with high-dimensional data. OSCAR method performs similar to Elastic-Net method but has a stronger grouping property than Elastic-Net, such that it has the ability to group negatively as well as positively correlated predictors that is sometimes desirable.

One last commonly used method is Forward-Lasso Adaptive Shrinkage (FLASH) [31], which can adaptively adjust the level of shrinkage necessary instead of using a penalty function. Therefore, it can be effectively used to perform variable selection in high-dimensional classification problems ($p > n$). This property significantly expands the range of problems that FLASH can be applied.

From 1990s, an alternative family of methods called SDR appeared [33], which unlike shrinkage regression methods, attempts to sufficiently reduce the number of predictor variables before fitting a model. The aim is to obtain a reduced-size data, in which relevant important information is preserved, that can then be fed into standard statistical models such as regression. Therefore, the SDR methodology offers an effective means to facilitate regression analysis for high-dimensional data. The SDR family of methods reduces the set of variables to a smaller set of either the original variables or new ones, where the new variables are linear combinations or even nonlinear functions of the original ones. However, the majority of SDR methods focus on linear reductions of variables, which arise naturally in many contexts. This allows the sufficiently reduced predictor variables to form a subspace known as ‘central subspace’, onto which the original data are projected.

Many properties of the central subspace have been developed resulting in a number of successfully applied dimension reduction methods. Two widely used approaches to find the central subspace are inverse moment based, and kernel smoothing-based approaches. The inverse moment-based approach is very easy and fast in computation and requires a relatively large sample size, while kernel smoothing-based approach works well for a small sample size and is slow in computation. The most famous methods founded on the inverse moment-based approach are sliced inverse regression (SIR) [26,28] and sliced average variance estimation (SAVE) [14]. Minimum average variance estimation (MAVE) [38] is also the most commonly used method that relies on the kernel smoothing-based approach. One method might be selected over another depending on the characteristics of data and the aim of data analysis, but overall SIR method is the most widely used one.

Many recent developments on the SDR methods have been made that tend to extend the previous methods with fewer assumptions, develop likelihood-based methods, and/or adapt and fine-tune methods for specific applications (see, e.g. [6,7,8,10,11,13,16,20–22,26,27,39]). Almost all of these methods require the computation of the inverse of a $p \times p$ sample covariance matrix, which easily becomes problematic for high-dimensional data, specially the ones that are not sparse or have $p > n$. More recently, Cook *et al.* proposed a rather different methodology which integrates recent work on the estimation of high-dimensional covariance matrix, but circumvents the problems faced when the high-dimensional data are abundant and $p > n$. The main idea is to start from the SIR technique, and to develop several population weight matrices that can be used for estimation. These weight matrices play a major role in dimension reduction as the reduction estimator corresponding to each weight matrix forms a basis for central subspace. The weight matrices are estimated by using sparse permutation invariant covariance estimation (SPICE) [32] and Moore–Penrose generalized inverse technique.

In this paper, we elaborate more on shrinkage regression and SDR families of methods. We select Ridge, LASSO, SCAD, Elastic-Net, OSCAR and FLASH methods from the family of shrinkage regression methods. From the SDR family, we focus on the SDR method proposed by Cook *et al.* [12] and the three main weight matrices developed there. More details on the relative advantages and disadvantages of these methods are provided based on their accuracy, variable selection and algorithm speed. We then apply all the selected methods from both families to two sets of real data and a number of simulated sets of data. Such analyses allow us to compare these two families of methods and the main methods within each family for the first time, and recommend conditions under which each family is more appropriate to be applied to high-dimensional data.

This paper is organized as follows. In Section 2, we review the basics of SDR approach and discuss the methods employed from this family in this paper. In Section 3 we describe shrinkage regression methods and their corresponding penalty functions. The advantages and disadvantages of these methods are also compared based on their penalty functions. Application and performance of the selected methods on the simulated and real data are conducted in Section 4. We conclude with a discussion of our findings in Section 5. In the rest of paper, we use the abbreviation form to simplify the notation.

2. Sufficient dimension reduction

There has been a great interest in SDR methods after 1990. The basic idea of dimension reduction approach is to map the random vector of predictors $X \in \mathbb{R}^p$ into another vector of lower dimension k ($k < p$) in a way that relevant information in the regression of the real response y on X is preserved. The goal is to find a sufficient reduction estimator function $\mathfrak{R} : \mathbb{R}^p \rightarrow \mathbb{R}^k$ that achieves such a mapping. In more detail, sufficient reduction estimator is defined as follows [9]:

Definition 2.1: *The function \mathfrak{R} is a sufficient reduction estimator for the regression of y on X if y is independent of $X \mid \mathfrak{R}(X)$. The image of \mathfrak{R} forms a d -dimensional subspace onto which the predictor X is projected which is called a dimension reduction subspace (S_{DRS}).*

Based on the above definition, the regression can be restricted to $\mathfrak{R}(X)$ once the reduction estimator \mathfrak{R} is specified, which has a lower dimension compared to the original regression problem. A more restrictive subspace is considered as follows [8]:

Definition 2.2: *Let $S_{y|X} = \cap S_{\text{DRS}}$ for any arbitrary S_{DRS} . If $S_{y|X}$ is an S_{DRS} itself, we call it a central dimension reduction subspace (S_{CDRS}).*

Remark 1: Following Cook *et al.* [12], we denote the dimension of S_{CDRS} by ‘ d ’ in the rest of paper.

It would correspond to a sufficient reduction, meaning that it preserves all the relevant information about response variable and if $\Psi(X)$ is any arbitrary sufficient reduction then $\mathfrak{R}(X)$ is a function of $\Psi(X)$. In this paper, we focus on minimal sufficient linear reductions for simplicity and refer to them as ‘sufficient reductions’:

Definition 2.3: *Minimal sufficient linear reduction is a reduction of the form $\mathfrak{R}(X) = \eta^T X$ where η is basis for S_{CDRS} .*

In the rest of this section, we review necessary tools for recognition and development of dimension reduction techniques, focusing more on the integrated dimension reduction methods developed in [12] that are further used in this paper.

One of the most widely methods in finding S_{CDRS} and sufficient reduction is SIR technique proposed in [26]. So if $(X^1, y_1), \dots, (X^n, y_n)$ represent the n realizations of the

random vector (X, y) , this regression results in the following relation:

$$X_i = \mu_X + \Gamma\beta(f(y_i) - \mu_f) + \varepsilon_i, \quad i = 1, 2, \dots, n, \tag{1}$$

where $X^i \in \mathbb{R}^p$, $\mu_X = E(X)$, the ε_i are independent realizations of a random vector $\varepsilon \in \mathbb{R}^p$ with mean 0 and variance Δ , Γ is a $p \times d$ -matrix ($d \leq p$) with rank d that is the basis matrix for $\Delta S_{y|X}$, β is an unknown $d \times r$ -matrix of coefficients ($d \leq r$) with rank d , and $f : \mathbb{R} \rightarrow \mathbb{R}^r$ is a known user-selected vector-valued function with $E(f(y)) = \mu_f$. Often, in high-dimensional data, we consider f to be a reasonably flexible set of basis functions, such as a vector-valued polynomial function of y . This regression model implies that the vector of $\beta(f(y_i) - \mu_f) \in \mathbb{R}^d$ gives the coordinates of $E(X|y) - E(X)$ in terms of the basis matrix Γ and is independent of ε . Based on this representation, $S_{\text{CDRS}} = \Delta^{-1} \text{span}(\Gamma)$.

A re-parametrization of the SIR method is proposed in [12] that avoids computation of the inverse of the covariance matrix, and hence eliminates some of the above shortcomings. Without loss of generality, the authors require that $\Gamma^T W \Gamma$ to be a diagonal matrix where W is a symmetric $p \times p$ positive-definite population weight matrix. Then they derive the SDR estimator function \mathfrak{R} as the coordinates of the projection of $X - \mu_X$ onto $\text{span}(\Gamma)$ in the Δ^{-1} inner product:

$$\mathfrak{R}(X) = \eta^T X := (\Gamma^T \Delta^{-1} \Gamma)^{-1} \Gamma^T \Delta^{-1} (X - \mu_X). \tag{2}$$

If we denote the estimation of \mathfrak{R} by $\hat{\mathfrak{R}}$, so finding $\hat{\mathfrak{R}}$ is based on using estimators of μ_X , Γ and β (referred to as \bar{X} , $\hat{\Gamma}$, and $\hat{\beta}$, respectively) when an estimate of W is specified. For a given sample weight matrix \hat{W} , estimating population weight matrix W , the function $\mathfrak{R}(X)$ is approximated as follows [12]:

$$\hat{\mathfrak{R}}(X) = (\hat{\Gamma}^T \hat{W} \hat{\Gamma})^{-1} \hat{\Gamma}^T \hat{W} (X - \bar{X}). \tag{3}$$

Different specifications of the positive-definite sample weight matrix \hat{W} results in different SDRs with different properties. In this paper, we apply the following three-dimension reduction estimators which were also employed in [12]. In the specifications below, let $\hat{\Delta}$ correspond to the residual sample covariance matrix for regression of X on $f(y)$.

- *Diagonal dimension reduction estimator $\hat{\mathfrak{R}}_{\text{diag}}$* : This estimator corresponds to the sample weight matrix of $\hat{W} = \text{diag}^{-1} \hat{\Delta}$, which ignores the correlations between residuals and only adjusts their variances. This estimator can be computed very quickly even though it may not be a good choice for highly correlated data [12].
- *Dimension reduction estimator $\hat{\mathfrak{R}}_{\hat{\Delta}}$* : This estimator requires the straight forward use of $\hat{W} = \hat{\Delta}^{-1}$ as the sample weight matrix. This estimator unifies the re-parametrization of Cook *et al.* [12] with earlier dimension reduction methods requiring direct calculation of the inverse of the covariance matrix. It is shown that this sample weight matrix is quite reasonable to use when $n > p+r+4$. However, the applicability of $\hat{\mathfrak{R}}_{\hat{\Delta}}$ is extended to cases when $n < p$ by using Moore–Penrose generalized inverse of $\hat{\Delta}$.
- *SPICE dimension reduction estimator $\hat{\mathfrak{R}}_{\text{spice}}$* : This estimator uses the SPICE method, developed in [32], to construct a sparse estimate for the inverse covariance matrix Δ^{-1} in high-dimensional data through an iterative process. The SPICE estimator is formed through L_1 -penalized likelihood optimization when a tuning parameter λ is specified

to control for how sparse of an estimate is required. When p is large, using a relatively large value of λ can lead to a substantial reduction in variability of the reduction estimator, while a smaller choice of λ creates less-sparse estimate of inverse error covariance matrix that slows down the computation time. In the analysis part, the optimal level of tuning parameter λ is employed.

In Section 4, we compare the performance of reduction estimators $\hat{\mathfrak{N}}_{\text{diag}}$, $\hat{\mathfrak{N}}_{\hat{\Delta}}$, and $\hat{\mathfrak{N}}_{\text{spice}}$ on data alongside of the selected methods from shrinkage regression family of methods outlined in the next section.

3. Shrinkage regression

Shrinkage regression is one of the most important methods in dimension reduction and variable selection for datasets with large number of predictor variables. These methods are specially useful when multicollinearity is present between predictor variables. The general approach is to impose a penalty term $P_{\lambda}(\beta)$ on the regression coefficients with the goal of reducing the magnitude and/or number of non-zero regression coefficients. The coefficients are then estimated such that the penalized squared error is minimized:

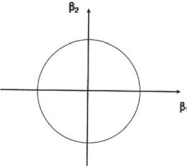
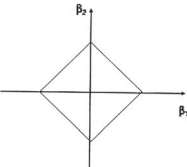
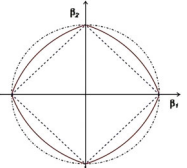
$$\hat{\beta} = \arg_{\beta} \min(\|y - X\beta\|^2 + P_{\lambda}(\beta)). \quad (4)$$

The penalty function can be adjusted by a tuning parameter λ to achieve a desired level of coefficient shrinkage. However, the selection of λ should be done with care as inappropriately high or low values of λ may result in underestimation or overestimation of the regression coefficients. The penalty function $P_{\lambda}(\beta)$ is typically expressed as one norm or a combination of different norms of the regression coefficients such as L_1 -norm, L_2 -norm and pairwise L_{∞} -norm. Thus the way the coefficients are shrunk in a shrinkage regression method reflects the properties of the corresponding norm(s) employed in its penalty function. For example a penalty function constructed by L_1 -norm can lead to a sparse regression model because of non-differentiability of the norm at 0, but a penalty function based on the L_2 -norm keeps all of the predictors in the model but forces highly correlated predictors to be averaged. This is also in contrast with a penalty based on pairwise L_{∞} -norm which encourages equality of the coefficient.

Some of the most important and widely used methods further tested in this paper are Ridge, LASSO, Elastic-Net, OSCAR, SCAD, and FLASH regression methods. Table 1 outlines the penalty functions used in these methods, and illustrates the shape of their level curves. The table also summarizes the advantages and disadvantages of each of these methods. The choice of each of these methods for a given dataset can then be made based on the nature of the data and aim of the analysis.

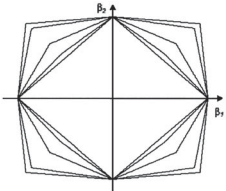
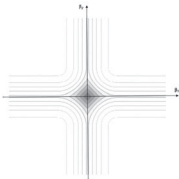
Some properties of the shrinkage regression methods are evident from Table 1. Ridge regression uses an L_2 -norm penalty which does not have the ability to select variables. LASSO method corrects for this by using an L_1 -norm penalty which makes the fitted model sparse and more interpretable. However, LASSO exhibits poor performance when $p > n$ or strong multicollinearity exists between predictor variables. Elastic-Net method is introduced as a compromise between these two techniques, which has a penalty that is the weighted sum of the these two methods [40]. The OSCAR method is another generalization of the LASSO method whose penalty function is a weighted sum of L_1 and L_{∞} norms of the

Table 1. Some of the most important shrinkage regression methods and their properties.

Method	Disadvantages	Advantages	Penalty Function ($P_\lambda(\beta)$)	Level curve(s) of $P_\lambda(\beta)$ based on only two of the regression coefficients β_1 and β_2
Ridge	Unable to select variables and results in biased estimators [35]	Better accuracy of prediction compared to the OLS approach [35]	$\lambda \sum_{j=1}^p \beta_j^2$	
LASSO	Has a poor performance when $p > n$ or when strong collinearity exists between predictor variables, results in biased estimators, and fails to do grouped selection [17,35,40]	Variable selection, consistency, and reduced over-fitting compared to other models [35,40]	$\lambda \sum_{j=1}^p \beta_j $	
Elastic-Net	Places positively correlated variables in the same group, and does not have an exact grouping property [40]	Selects variables, has a better accuracy of prediction compared to LASSO, can select the true model even when the LASSO fails, and is a good option for high-dimensional data [40]	$\lambda_1 \sum_{j=1}^p \beta_j + \lambda_2 \sum_{j=1}^p \beta_j^2$	

(continued).

Table 1. Continued.

Method	Disadvantages	Advantages	Penalty Function ($P_\lambda(\beta)$)	Level curve(s) of $P_\lambda(\beta)$ based on only two of the regression coefficients β_1 and β_2
OSCAR	Cannot handle high-dimensional data [1]	Selects variables, encourages grouping effect, and is a good option for highly correlated data [1]	$\lambda(\sum_{j=1}^p \beta_j + c \sum_{k < j} \max\{ \beta_j , \beta_k \})$	
SCAD	Non-differentiable at the origin [19]	Selects variables, and results in a sparse model with approximately unbiased coefficients for large coefficients [19]	$\sum_{j=1}^p (\lambda \beta_j I_{(0 \leq \beta_j < \lambda)} + \frac{(a^2 - 1)\lambda^2 - (\beta_j - a\lambda)^2}{2(a - 1)} I_{(\lambda \leq \beta_j < a\lambda)} + \frac{(a + 1)\lambda^2}{2} I_{(\beta_j \geq a\lambda)})$	
FLASH	Poor performance for $p < n$ [31]	Selects variables, adjusts the level of shrinkage to optimize the selection of the next variable, and is a good option for high-dimensional data [31]	The method has no specific penalty; it is a combination of LASSO and forward selection regression with adaptive tuning of parameter λ .	This method has no specific penalty.

regression coefficients. This method encourages sparsity of predictor variables in the fitted model as well as equality of coefficients for correlated predictors that have similar relationships with the response variable [1]. The SCAD method also reduces to LASSO when $a \rightarrow \infty$. Unlike the LASSO procedure, the SCAD penalty for coefficients that are larger than $a\lambda$ is a constant, making the estimation of these coefficients unbiased. The SCAD method uses two parameters to specify the penalty function (similar to Elastic-Net and OSCAR), but it is shown that SCAD performs best when $a = 3.7$ [19].

As mentioned in Table 1, the FLASH method is an adoptive procedure based on LASSO and forward selection. First it considers the model without any variables and then iteratively adds variables that are the most highly correlated with the current residual vector. In each iteration, the residuals are recomputed using the OLS solution based on the currently selected variables, and the level of shrinkage (choice of λ) is updated so as to optimize the selection of the next variable. This procedure is repeated until all variables have been added to the model [31]. The adoptive nature of this method maintains some of the desirable properties of LASSO such as variable selection, but circumvents some of its limitations so that it can be well applied to high-dimensional data.

4. Simulated and real data analysis

In this section, we apply the selected methods in Sections 2 and 3 to simulated and real data and explore their performance. We use three criteria to compare the performance of the selected methods in both families of shrinkage regression and dimension reduction: residual mean square (RMS), the computation time, and the variable selection capability. RMS is calculated as

$$\text{RMS} = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n},$$

where \hat{y}_i in shrinkage regression methods represents the predicted value for observation i when the appropriate penalty function is applied. In dimension reduction methods, \hat{y}_i is determined by

$$\hat{y}_i = \hat{E}(y | X = X_i) = \sum_{i=1}^n \left[\frac{\hat{f}(\hat{\mathfrak{R}}(X) | y_i)}{\sum_{k=1}^n \hat{f}(\hat{\mathfrak{R}}(X) | y_k)} \right] y_i,$$

where \hat{f} is the estimate of the conditional density function f of $\mathfrak{R} | y$. For more details on functions f and \hat{f} , see, for example, Cook *et al.* [12].

Of course, a method with smaller RMS and shorter computation time is preferred, but specially for datasets with large number of predictor variables, the ability of a method to select variables is highly desired as well in order to reduce the complexity of the problem. Therefore, we report the number of selected variables in each of the models that correspond to the ones with non-zero coefficients. Note that for the dimension reduction methods, we linearly map the original predictors into the central subspace with a much lower dimension. In our experiments, we found that when the results of estimation from the central subspace are mapped back to the original space of the predictors, many of the predictors contribute to the response variable with a small coefficient. Therefore, in each of our experiments and

for each of the SDR methods, we also report the dimension of the central subspace (computed according to the permutation test developed by Cook and Yin [15]), along with the number of predictors that contribute with a coefficient outside of the range $(-0.05, 0.05)$ (denoted by P^*).

All methods were implemented in *R* software package. For shrinkage regression methods, optimal tuning parameters were also computed to minimize the penalized squared error, making the results more reliable.

4.1. Simulation study

In this section, we use a simulation study to assess the performance of shrinkage regression and SDR families with different correlation structure between predictor variables. We consider the relationship between the predictor and response variables to be of the form $y = X\beta + \varepsilon$ in which we set all elements of β except five of them to be zero. Therefore, we know that the actual number of effective variables is 5. We assume that ε follows a normal distribution with mean zero and standard deviation 1 and is independent of X . But the predictors X are assumed to follow a multivariate normal distribution, $N_p(0, \Sigma)$, in which the covariance matrix $\Sigma = [\sigma_{ij}]$ is set such that $\sigma_{ij} = \exp(-\phi/20)|i-j|$, $\phi \in (0, 1]$, for $i \neq j$ and 1 otherwise, and the parameter ϕ is allowed to vary in the range $(0, 1]$. In our simulation study, we test three values of $\phi = 0.1, 0.5$ and 1 to assess the performance of our selected methods when the correlation structure of the data changes. Figure 1 depicts the correlation structure of the predictor variables in simulated data for these three values of ϕ . Based on this figure, changing the values of ϕ changes the correlation structure of the predictors so that for $\phi = 1$ or 0.5, we have positive and negative correlations between the predictors, while for $\phi = 0.1$, the correlation between the predictor variables is all positive. Furthermore, $\phi = 0.1$ makes the correlation between the predictor variables stronger than $\phi = 1$ or 0.5. Hence, smaller values of ϕ makes the predictors more highly correlated.

We generate 60 observations of X and ε , i.e. $n = 60$, and consider two levels of the number of predictors $p = 400$ and $p = 56$. Note that the level of difficulty of the problem for these two levels of p are different as in the former we have $p > n$ but in the latter we have $p < n$. For each level of p , we test our selected methods when ϕ is set to either 0.1, 0.5, or 1.

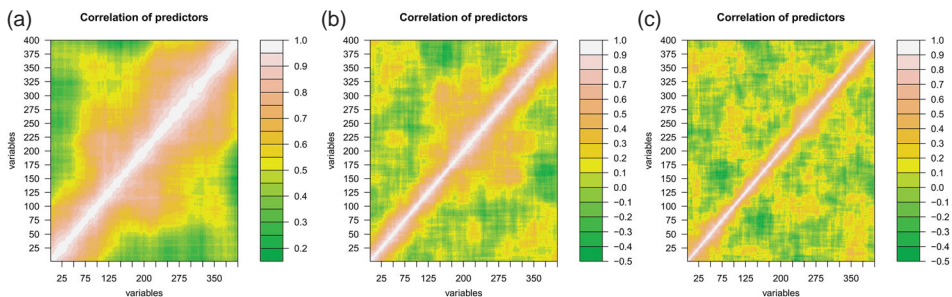


Figure 1. Diagram of the correlation between the predictor variables in the simulated data. For simplicity, we show predictor X_i by just ' i ' on both axes. The bright and dark colors indicate weak and strong correlation between predictors, respectively. (a) Correlation between the predictors for $\phi = 0.1$. (b) Correlation between the predictors for $\phi = 0.5$. (c) Correlation between the predictors for $\phi = 1$.

Now we apply the selected methods in Sections 2 and 3 and assess the performance of these two methods. Note that when $p = 400$, the OSCAR method cannot be applied (as $p > n$), and also when $p = 56$, the FLASH method cannot be applied (as $p < n$). The results of applying the remaining methods are summarized in Tables 2 and 3.

When $p > n$, we can see in Table 2 that $\hat{\mathfrak{N}}_{\text{spice}}$ has the best levels of RMS among dimension reduction methods. In terms of computational complexity, the $\hat{\mathfrak{N}}_{\text{diag}}$ method is more time efficient than $\hat{\mathfrak{N}}_{\hat{\Delta}}$ and $\hat{\mathfrak{N}}_{\text{spice}}$, but the computation time of $\hat{\mathfrak{N}}_{\text{spice}}$ is reasonable as well. However, when $p < n$, Table 3 shows that among dimension reduction methods $\hat{\mathfrak{N}}_{\hat{\Delta}}$ and $\hat{\mathfrak{N}}_{\text{spice}}$ have the best levels of RMS, while the best method of choice among the two switches to $\hat{\mathfrak{N}}_{\hat{\Delta}}$. The computation time of these two methods are also very similar. Therefore, we can conclude that $\hat{\mathfrak{N}}_{\hat{\Delta}}$ and $\hat{\mathfrak{N}}_{\text{spice}}$ are the preferred methods for the dimension reduction family as they better capture the correlation structure between the predictor variables than $\hat{\mathfrak{N}}_{\text{diag}}$ depending on the number of parameters in relation to the data size.

In the choice of dimension reduction methods, it is also important to know how much of the variations in the data are explained by the model. This can be detected visually by comparing plots of response variables versus fitted values for each of the methods. The plots for the three selected dimension reduction methods for the two levels of p are depicted in Figures 2 and 3, which illustrate that $\hat{\mathfrak{N}}_{\text{spice}}$ is more successful in describing the relationship between predictors and the response variable, with $\hat{\mathfrak{N}}_{\hat{\Delta}}$ providing second best level of fit when $p < n$.

In all three-dimension reduction methods, we found that the dimension of the central subspaces to vary between 1 and 4 across the different simulated datasets. All central subspaces were found to have high p -values according to the permutation test of Cook and

Table 2. Performance comparison of selected dimension reduction and shrinkage regression methods for simulated data.

		$n = 60, p = 400$								
Method		FLASH	LASSO	Elastic-Net	Ridge	SCAD	OSCAR	$\hat{\mathfrak{N}}_{\text{diag}}$	$\hat{\mathfrak{N}}_{\hat{\Delta}}$	$\hat{\mathfrak{N}}_{\text{spice}}$
$\phi = 1$	RMS	0.021	0.120	0.111	2.760	0.043	–	3.333	2.631	0.535e–3
	Number of selected variables	24	40	47	400	400	–	400	400	400
	Dimension of central subspace (P^*)	–	–	–	–	–	–	1(6)	1(30)	1(5)
	Computation time (seconds)	19.126	2.590	1.903	7.816	2950.382	–	0.702	471.448	4922.220
$\phi = 0.5$	RMS	0.052	0.144	0.166	2.035	0.064	–	1.312	1.080	0.297e–3
	Number of selected variables	16	21	27	400	400	–	400	400	400
	Dimension of central subspace (P^*)	–	–	–	–	–	–	1(7)	1(18)	1(6)
	Computation time (seconds)	18.207	0.624	0.437	1.966	542.157	–	0.330	130.936	5150.437
$\phi = 0.1$	RMS	0.017	0.029	0.033	1.582	0.041	–	4.099	0.232	0.326e–3
	Number of selected variables	14	30	39	400	400	–	400	400	400
	Dimension of central subspace (P^*)	–	–	–	–	–	–	1(1)	1(38)	1(6)
	Computation time (seconds)	3.908	0.478	0.480	1.745	953.501	–	0.337	140.303	4689.515

P^* : Number of variables with coefficients outside of the range $(-0.05, 0.05)$.

Table 3. Performance comparison of selected dimension reduction and shrinkage regression methods for simulated data.

		$n = 60, p = 56$								
Method		FLASH	LASSO	Elastic-Net	Ridge	SCAD	OSCAR	$\hat{\mathfrak{N}}_{diag}$	$\hat{\mathfrak{N}}_{\Delta}$	$\hat{\mathfrak{N}}_{spice}$
$\phi = 1$	RMS	–	0.091	0.147	3.732	0.388	0.798e–4	1.901	0.843e–4	0.002
	Number of selected variables	–	17	19	56	56	56	56	56	56
	Dimension of central subspace (P^*)	–	–	–	–	–	–	2(39)	2(56)	2(55)
	Computation time (seconds)	–	0.158	0.219	0.297	6.562	5825.186	0.141	14.484	16.922
$\phi = 0.5$	RMS	–	0.076	0.110	2.452	0.042	0.123e–3	2.510	0.621e–6	0.002
	Number of selected variables	–	16	14	56	56	56	56	56	56
	Dimension of central subspace (P^*)	–	–	–	–	–	–	1(9)	1(53)	1(8)
	Computation time (seconds)	–	2.246	0.780	1.170	9.032	6656.775	1.031	16.228	13.432
$\phi = 0.1$	RMS	–	0.049	0.055	1.086	0.011	0.387e–3	0.650	0.340e–18	0.395e–3
	Number of selected variables	–	7	11	56	56	56	56	56	56
	Dimension of central subspace (P^*)	–	–	–	–	–	–	4(54)	4(56)	4(56)
	Computation time (seconds)	–	0.359	0.281	0.374	16.850	6338.784	0.452	27.832	12.639

P^* : Number of variables with coefficients outside of the range $(-0.05, 0.05)$.

Yin [15], meaning that the central subspaces can be described by one or a few linear combinations of the predictor variables. Looking at the variables selected in the specification of the central subspace, we found that all predictors contributed in explaining the response variable in all three methods. Therefore, all dimension reduction methods perform poorly in selecting significant variables. However, if we ignore predictors whose coefficients are close enough to zero and lie in the range $(-0.05, 0.05)$, then $\hat{\mathfrak{N}}_{spice}$ is the closest method in estimating the correct number of effective predictor variables (i.e. the chosen number of 5) when $p > n$ but its performance in this respect is not as satisfactory when $p < n$.

Among the family of shrinkage regression methods, when $p > n$, Table 2 illustrates that the FLASH method outperforms the others with the smallest RMS and a reasonable computation time. In terms of variable selection ability of the methods, the FLASH method chooses the least number of variables, so this method is the best choice in the category of shrinkage regression methods. SCAD, Elastic-Net and LASSO have somewhat small RMS, and are the second best compared to FLASH when $p > n$. However, the SCAD method is much slower specially when optimizing the tuning parameter and is unable to select variables.

When $p < n$, the FLASH method is no longer applicable, and Table 3 illustrates that the OSCAR method outperforms the others with the smallest RMS. However, this method is not able to properly select effective variables. LASSO and Elastic-Net provide second best levels of RMS while having superior variable selection capability compared to OSCAR.

Comparing the two families of shrinkage regression and dimension reduction methods, we can conclude that overall $\hat{\mathfrak{N}}_{spice}$ and OSCAR methods provide the best accuracy when $p > n$ and $p < n$, respectively, specially when the data exhibit strong multicollinearity. However, these methods may not be the best choice if variable selection and/or computational

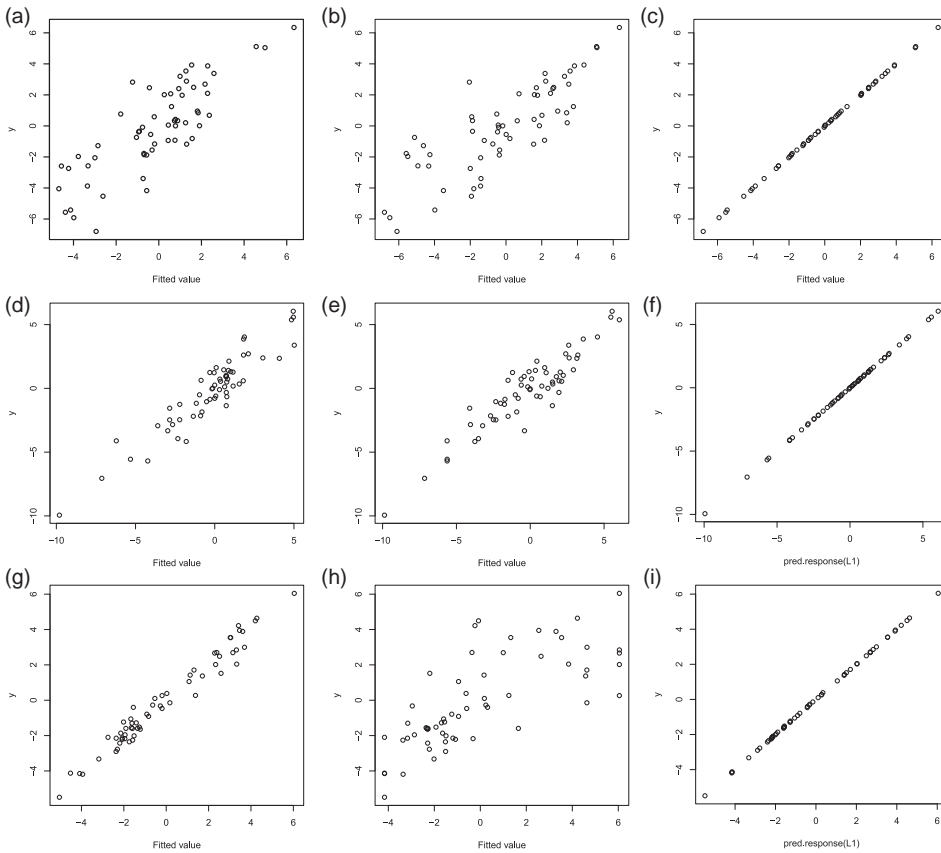


Figure 2. Comparison of explanatory power of dimension reduction methods for simulated data ($n = 60, p = 400$). (a) Response variables vs. fitted values for $\hat{\eta}_{\Delta}(\phi = 1)$. (b) Response variables vs. fitted values for $\hat{\eta}_{\text{diag}}(\phi = 1)$. (c) Response variables vs. fitted values for $\hat{\eta}_{\text{SPICE}}(\phi = 1)$. (d) Response variables vs. fitted values for $\hat{\eta}_{\Delta}(\phi = 0.5)$. (e) Response variables vs. fitted values for $\hat{\eta}_{\text{diag}}(\phi = 0.5)$. (f) Response variables vs. fitted values for $\hat{\eta}_{\text{SPICE}}(\phi = 0.5)$. (g) Response variables vs. fitted values for $\hat{\eta}_{\Delta}(\phi = 0.1)$. (h) Response variables vs. fitted values for $\hat{\eta}_{\text{diag}}(\phi = 0.1)$. (i) Response variables vs. fitted values for $\hat{\eta}_{\text{SPICE}}(\phi = 0.1)$.

speed are most important. In this case, other methods from shrinkage regression family are the best: when $p > n$, the FLASH method is the best method of choice, and when $p < n$, LASSO and Elastic-Net have the best performance in this respect. We also observe that the relative ranking of these methods, as explained above, remains more or less similar as the level of multicollinearity in the data (i.e. ϕ) changes.

4.2. Real data analysis

In this subsection, we apply the selected methods in Sections 2 and 3 to two sets of data: Cookie dough data used in [30] and Prostate Cancer data employed in [24,34]. In the cookie dough data, the number of predictors (400) is much larger than the number of

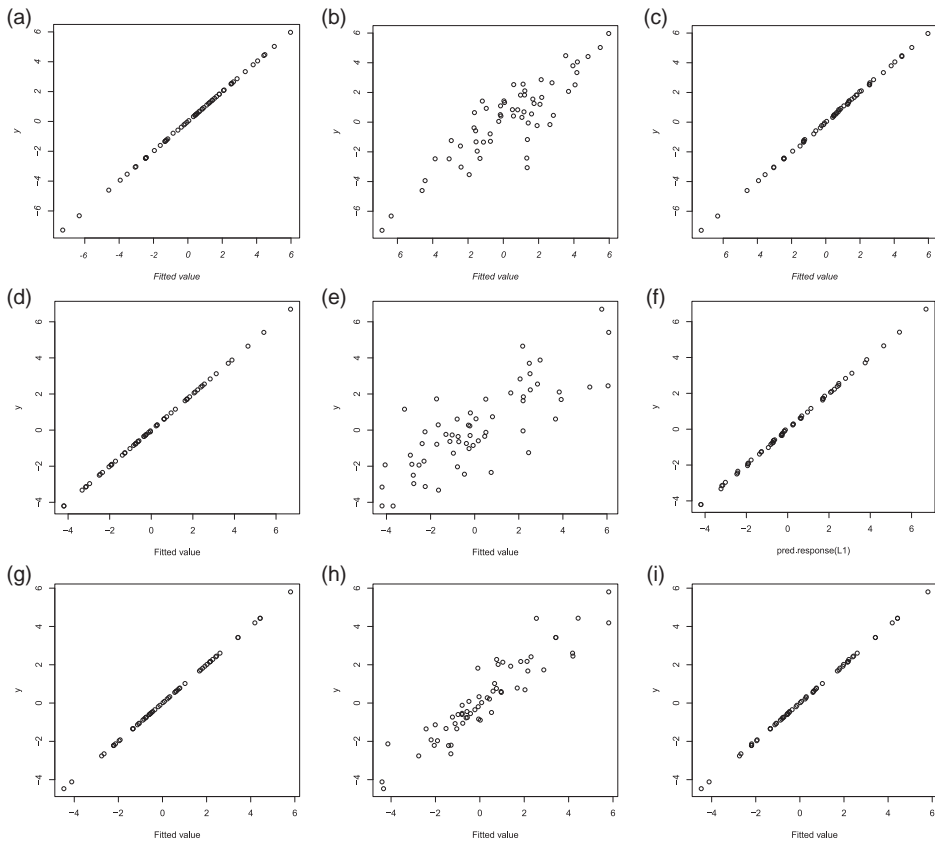


Figure 3. Comparison of explanatory power of dimension reduction methods for simulated data ($n = 60, p = 56$). (a) Response variables vs. fitted values for $\hat{\eta}_{\hat{\Delta}}(\phi = 1)$. (b) Response variables vs. fitted values for $\hat{\eta}_{\text{diag}}(\phi = 1)$. (c) Response variables vs. fitted values for $\hat{\eta}_{\text{SPICE}}(\phi = 1)$. (d) Response variables vs. fitted values for $\hat{\eta}_{\hat{\Delta}}(\phi = 0.5)$. (e) Response variables vs. fitted values for $\hat{\eta}_{\text{diag}}(\phi = 0.5)$. (f) Response variables vs. fitted values for $\hat{\eta}_{\text{SPICE}}(\phi = 0.5)$. (g) Response variables vs. fitted values for $\hat{\eta}_{\hat{\Delta}}(\phi = 0.1)$. (h) Response variables vs. fitted values for $\hat{\eta}_{\text{diag}}(\phi = 0.1)$. (i) Response variables vs. fitted values for $\hat{\eta}_{\text{SPICE}}(\phi = 0.1)$.

observations, a feature of a big dataset that complicates the use of traditional statistical methods. In comparison, the Prostate Cancer data represent a more standard type of dataset in the sense that the number of observation (97) is larger than the number of the predictor variables (8). In addition, as can be seen in Figure 4 and Table 4, each pair of the predictor variables in both datasets are highly correlated (with more correlation existing in the Cookie dough data). This multicollinearity adds another layer of challenge to data analysis and variable selection.

4.2.1. Analysis of cookie dough data.

Cookie dough data, originally used in [30], arose from an experiment to non-destructively control the sucrose content of cookies with the help of near-Infrared (NIR) spectroscopy, (see also [3]). The predictor variables are measurements of the NIR spectroscopy

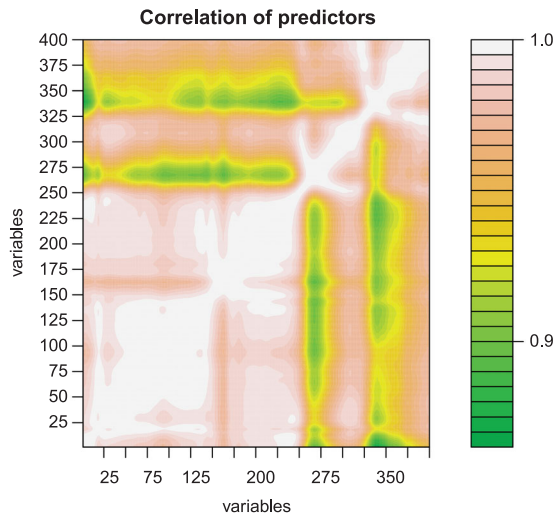


Figure 4. Diagram of the correlation between the predictor variables in the Cookie Dough data. For simplicity, we show predictor X_i by just ‘ i ’ on both axis. The bright and dark colors indicate weak and strong correlation between predictors, respectively.

Table 4. Correlation between the predictor variables in the Prostate Cancer data.

Predictors	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
lcavol	1.000	0.281	0.225	0.027	0.539	0.675	0.432	0.435
lweight	0.281	1.000	0.348	0.442	0.155	0.165	0.057	0.107
age	0.225	0.348	1.000	0.350	0.118	0.128	0.269	0.276
lbph	0.027	0.442	0.350	1.000	-0.086	-0.007	0.078	0.078
svi	0.539	0.155	0.118	-0.086	1.000	0.673	0.320	0.458
lcp	0.675	0.165	0.128	-0.007	0.673	1.000	0.515	0.633
gleason	0.432	0.057	0.269	0.078	0.320	0.515	1.000	0.752
pgg45	0.435	0.107	0.276	0.078	0.458	0.633	0.752	1.000

Table 5. Performance comparison of selected dimension reduction and shrinkage regression methods for cookie dough data.

Method	FLASH	LASSO	Elastic-Net	Ridge	SCAD	$\hat{\eta}_{diag}$	$\hat{\eta}_{\Delta}$	$\hat{\eta}_{spice}$
RMS	0.267	0.649	0.731	6.474	0.710	30.717	0.108	0.448
Number of selected variables	16	24	101	400	400	400	400	400
Dimension of central subspace (P^*)	-	-	-	-	-	1(378)	1(400)	1(373)
Computation time (seconds)	12.981	8.354	17.556	6.029	572.386	1.545	524.187	4191.847

P^* : Number of variables with coefficients outside of the range $(-0.05, 0.05)$.

reflectance spectrum, with 400 wavelengths measured from 1400 to 2198 nm in steps of 2 nanometers, while the dependent variable is the sucrose content of a piece of cookie dough. Similar to [23], 23rd and 61st observations (out of 70) are removed as outliers. The performance of selected methods in both dimension reduction and shrinkage regression families applied to this dataset is compared in Table 5. Note that as $p > n$ in these data, the OSCAR method cannot be applied.

According to Table 5, in the family of dimension reduction methods $\hat{\eta}_{diag}$ performs the worst. As mentioned in Section 2, the difference between these methods is due to the choice

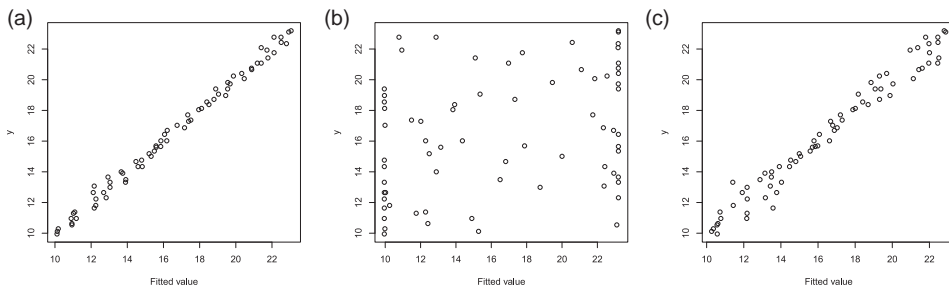


Figure 5. Comparison of explanatory power of dimension reduction methods for cookie dough data. (a) Response variables vs. fitted values for $\hat{\eta}_{\hat{\Delta}}$. (b) Response variables vs. fitted values for $\hat{\eta}_{\text{diag}}$. (c) Response variables vs. fitted values for $\hat{\eta}_{\text{SPICE}}$.

of the sample weight matrix and its ability to capture the correlation structure between the predictor variables with a reasonable accuracy and in a reasonable time. The poor performance of $\hat{\eta}_{\text{diag}}$ can be expected since the predictor variables are highly correlated and a diagonal weight matrix cannot support this property. However, $\hat{\eta}_{\hat{\Delta}}$ performs the best among the dimension reduction methods, with the SPICE method coming to second, both in terms of RMS and computation time. In the choice of dimension reduction methods, it is also important to know how much of the variations in the data are explained by the model. This can be detected visually by comparing plots of response variables versus fitted values for each of the methods. The plots for the three selected dimension reduction methods are depicted in Figure 5, which illustrates that $\hat{\eta}_{\hat{\Delta}}$ is more successful in describing the relationship between predictors and the response variable.

In all three-dimension reduction methods, the central subspaces are one-dimensional according to the permutation test of Cook and Yin [15] with high p -values, meaning that the central subspace can be described by a single linear combination of the predictor variables. Looking at the variables selected in the specification of the central subspace (the ones with non-zero coefficient in the linear combination), we found that all predictors contributed in explaining the response variable in all three methods. Therefore, all dimension reduction methods perform poorly in selecting significant variables. However, if we ignore predictors whose coefficients are close enough to zero and lie in the range $(-0.05, 0.05)$, then 400, 378, 373 variables contribute to constructing a sufficient reduction in $\hat{\eta}_{\hat{\Delta}}$, $\hat{\eta}_{\text{diag}}$ and $\hat{\eta}_{\text{spice}}$, respectively. Therefore, in terms of variable selection, $\hat{\eta}_{\text{spice}}$ results in a less complex sufficient reduction in a sense and is desired over the other two methods. However, none of the methods are able to create a sparse model.

Among the family of shrinkage regression methods, Table 5 illustrates that the FLASH method outperforms the others with the smallest RMS and a reasonable computation time. In terms of variable selection ability of the methods, the FLASH method chooses the least number of variables while providing the best performance, so this method remains the best choice in the category of shrinkage regression methods. LASSO, SCAD and Elastic-Net have somewhat similar RMS, and are the second best compared to FLASH, however, the SCAD method is much slower specially when optimizing the tuning parameter and is unable to select variables.

Comparing the two families of shrinkage regression and dimension reduction methods, according to Table 5 we can conclude that overall $\hat{\eta}_{\hat{\Delta}}$ method (associated to the weight matrix $\hat{W} = \hat{\Delta}^{-1}$) provides the best accuracy specially for the cookie dough data that has more predictors than observations and exhibits strong multicollinearity. However, this method may not be the best choice if variable selection and/or computational speed is most important. In this case, the FLASH method from shrinkage regression family is the best alternative which is superior in terms of variable selection and time efficiency while remaining relatively accurate.

4.2.2. Analysis of prostate cancer data.

The Prostate Cancer data, first studied in [34], examine the correlation between the level of prostate-specific antigen (PSA) and a number of clinical measures in 97 men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (lpsa) from measurements of log of cancer volume (lcavol), log of prostate weight (lweight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), gleason score (gleason), and percent of gleason scores 4 or 5 (pgg45). Similar to the cookie dough data, the predictor variables are highly correlated (see Table 4), but here the number of predictors is less than the number of observations (i.e. $p < n$). As a result, the FLASH method cannot be applied to this dataset. The results of applying the remaining methods in both dimension reduction and shrinkage regression families to this data are summarized in Table 6.

As we see in Table 6, among dimension reduction methods $\hat{\eta}_{\text{SPICE}}$ and $\hat{\eta}_{\hat{\Delta}}$ have the best and similar levels of RMS. Therefore, these methods are better equipped to capture the correlation structure between the predictor variables than $\hat{\eta}_{\text{diag}}$. However, RMS for $\hat{\eta}_{\text{diag}}$ method is not too far away from the other two methods as observed for the cookie dough data, which is expectable due to the relatively weaker correlation between pairs of predictors in this dataset. In terms of computational complexity, unlike the previous data, the $\hat{\eta}_{\text{SPICE}}$ method is more time efficient than $\hat{\eta}_{\hat{\Delta}}$, but the computation time of $\hat{\eta}_{\hat{\Delta}}$ is reasonable as well. Consequently, for these data with $p < n$, the $\hat{\eta}_{\text{SPICE}}$ method is preferred even though $\hat{\eta}_{\hat{\Delta}}$ is also the best alternative. Comparing the plots of response variables versus fitted values for each of the dimension reduction methods depicted in Figure 6, we see a relatively good and similar explanatory capability for the three-dimension reduction methods as the data points in these plots are similarly scattered around the 45° line, even though the $\hat{\eta}_{\text{diag}}$ method seems to explain somewhat less than the other two methods.

In all three-dimension reduction methods, the central subspaces are one-dimensional according to the permutation test of Cook and Yin [15] with high p -values. This means that

Table 6. Performance comparison of selected dimension reduction and shrinkage regression methods for prostate cancer data.

Method	LASSO	Elastic-Net	Ridge	SCAD	OSCAR	$\hat{\eta}_{\text{diag}}$	$\hat{\eta}_{\hat{\Delta}}$	$\hat{\eta}_{\text{spice}}$
RMS	0.547	0.548	0.552	0.464	0.449	0.611	0.440	0.439
Number of selected variables	3	4	8	8	8	8	8	8
Dimension of central subspace (P^*)	–	–	–	–	–	1(5)	1(5)	1(5)
Computation time (second)	1.118	0.870	0.853	384.049	1517.931	0.187	24.138	0.257

P^* : Number of variables with coefficients outside of the range $(-0.05, 0.05)$.

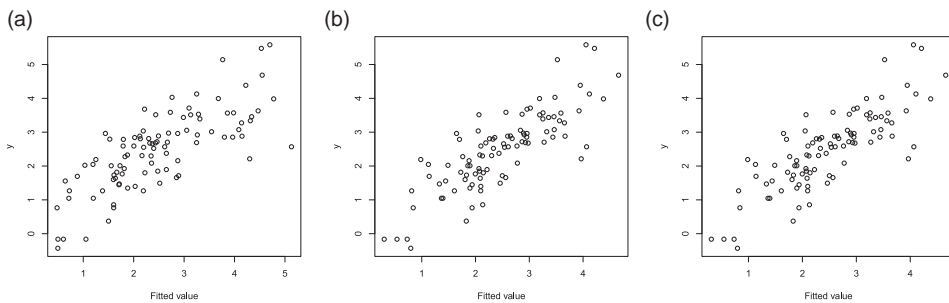


Figure 6. Comparison of explanatory power of dimension reduction methods for prostate cancer data. (a) Response variables versus fitted values for $\hat{\eta}_{diag}$. (b) Response variables versus fitted values for $\hat{\eta}_{\hat{\Delta}}$. (c) Response variables versus fitted values for $\hat{\eta}_{SPICE}$.

the sufficient reductions can be described by a single linear combination of the eight original predictor variables. Looking at the variables selected in the specification of the central subspace (the ones with non-zero coefficient in the linear combination), we found that all predictors contributed in explaining the response variable in all three methods. Therefore, all dimension reduction methods perform poorly in selecting significant variables. However, if we ignore predictors whose coefficients are close enough to zero and lie in the range $(-0.05, 0.05)$, then five of the original predictor variables contribute to constructing a sufficient reduction in each of the three methods. Therefore, in terms of variable selection, all three methods obtain a similar level of model complexity, but none of them are able to create a sparse model. In addition, for constructing sufficient predictors for $\hat{\eta}_{\hat{\Delta}}$ and $\hat{\eta}_{spice}$, the variables ‘age’, ‘gleason’ and ‘pgg45’ are not very effective, and the values of coefficients of the other five variables are very close to one another in these two methods. In contrast, the variables ‘age’, ‘lbph’ and ‘pgg45’ are not effective in constructing a sufficient reduction for $\hat{\eta}_{diag}$ method, and all of the coefficients that this method proposes are negative. Therefore, the behavior of $\hat{\eta}_{\hat{\Delta}}$ and $\hat{\eta}_{spice}$ is similar but different from that of $\hat{\eta}_{diag}$.

Among the family of shrinkage regression methods, Table 6 illustrates that SCAD and OSCAR methods have the least RMS values, even though the RMS of the other shrinkage regression methods is also close. However, SCAD and OSCAR are unable to reduce the number of selected variables and have relatively large computation time. Alternatively, LASSO and Elastic-Net methods are able to significantly reduce the number of effective variables down to 3 and 4, respectively. Running the significance test for LASSO proposed by [4,5,18,29,36] and we found that LASSO enters one less variable than Elastic-Net at a significance level of .05. Also these two methods run very time efficient (less than 2 seconds) without much sacrifice on the accuracy (RMS). Therefore, if variable selection capability is desired, LASSO or Elastic-Net methods can be a method of choice. Comparing the regression coefficients produced by these methods, we observed that all coefficients in LASSO and Elastic-Net methods were non-negative, while the coefficients of ‘age’ variable in Ridge and OSCAR methods were negative which seems to be against the intuition. As a result, the shrinkage regression methods provide different interpretation of the influence of predictors on the response variables and the choice of one method over another should be done with more care.

Comparing the two families of shrinkage regression and dimension reduction methods, according to Table 6 we can conclude that overall dimension reduction method of $\hat{\mathfrak{N}}_{\hat{\Delta}}$ and $\hat{\mathfrak{N}}_{\text{spice}}$ provides the best accuracy over other methods for the prostate cancer data which is characterized by less predictors than observations and exhibits strong multicollinearity. This also agrees with the comparison of these two families of methods for the cookie dough data. But it should be noted that the accuracy improvement obtained from using these two methods over shrinkage regression methods is much larger in the cookie dough data with $p > n$, even though they took longer to run. However, if the aim of the analysis is to select effective variables, dimension reduction methods are not a good choice. LASSO and Elastic-Net methods from shrinkage regression family are the best alternative which are superior in terms of variable selection while remaining relatively accurate and time efficient.

5. Conclusion

In this paper, for the first time, we analytically and empirically compare commonly used methods in SDR and shrinkage regression families of methods for analyzing standard and high-dimensional data. We select Ridge, LASSO, SCAD, Elastic-Net, OSCAR and FLASH methods from the family of shrinkage regression methods. From the dimension reduction family, we focus on the method proposed by Cook *et al.* [12] that integrates many of the existing dimension reduction methods by the choice of a weight matrix, and experiment with three main weight matrices developed there. We review the fundamentals of the two families of methods and summarize their relative analytical advantages and disadvantages.

We then apply all the selected methods from both families to simulated data as well as to two commonly used sets of real data. The first dataset was the cookie dough data in which the number of predictors were more than number of observations, which is a candidate for high-dimensional data. In contrast, the second dataset was the prostate cancer data exemplifying a more standard type of data in the sense that the number of observations is larger than the number of the predictors. Both datasets as well as our simulated data exhibit strong multicollinearity that often complicates the use of standard predictive methods.

We compare the selected methods based on three measures: accuracy, computation time, and variable selection capability. Each of the SDR and shrinkage regression methods that we studied can be considered usefully in the right situation. But according to our experimental results on real and simulated data, encompassing both high-dimensional and standard real data that exhibit significant multicollinearity, we make the following conclusions:

- *For standard data ($p < n$):* From the family of shrinkage regression methods, OSCAR is the most accurate one. Also, among the SDR family, we found that using weight matrices corresponding to the SPICE and $\hat{\Delta}^{-1}$ provide the best accuracy. These two weight matrices provide a comparable accuracy with the OSCAR method, where the relative rank between them may change based on specifics of each dataset such as its multicollinearity structure. Also, OSCAR is computationally more demanding than the other SDR methods, while the computation time remains at an acceptable level. However, these methods cannot provide proper variable selection among the parameters. If variable selection is highly desired, FLASH and Elastic-Net methods from the shrinkage

regression family are the best alternatives while they remain considerably accurate as well.

- For *high-dimensional data* ($\mathbf{p} > \mathbf{n}$): In the shrinkage regression family, note that the most accurate method for standard data, OSCAR, cannot be applied for high-dimensional data. In this family, we found the FLASH method to be the most accurate one. However, among the SDR methods, we found that using weight matrices corresponding to the SPICE and $\hat{\Delta}^{-1}$ provide the best accuracy. Comparing the two families of methods, we found that the SDR method with the former two weights are superior in terms of accuracy, where the relative rank between using these two weights may vary based on specifics of each dataset (e.g. in our simulation study, we found that the SPICE weight to be the best, while in the cookie dough data $\hat{\Delta}^{-1}$ was the best). However, this level of accuracy comes at the cost of variable selection capability. Based on our experiments, FLASH and LASSO methods can effectively select variables while they remain relatively accurate.

In addition, our analysis indicates that different levels of multicollinearity between predictor variables has little impact on the relative ranking of the methods that we studied. This finding essentially simplifies the task of method selection when dealing with a new dataset. Of course, our analysis is limited to the commonly used form of collinearity structure that we considered in our simulation study and observed in the two real datasets, as well as the sizes of the predictor variables and sample size. More examination of the relative performance of the selected methods would be needed for datasets with p and n much larger than the ones considered here. Also, development of variable selection capability in dimension reduction methods can prove to be worthwhile specially in better analysis of high-dimensional data with multicollinearity.

Acknowledgments

We would like to thank referees whose helpful comments led to substantial improvements in this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- [1] H.D. Bondell and B.J. Reich, *Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar*, *Biometrics* 64 (2008), pp. 115–123.
- [2] L. Breiman, *Heuristics of instability and stabilization in model selection*, *Ann. Stat.* 24 (1996), pp. 2350–2383.
- [3] P.J. Brown, T. Fearn, and M. Vannucci, *Bayesian wavelet regression on curves with application to a spectroscopic calibration problem*, *J. Amer. Statist. Assoc.* 96 (2001), pp. 398–408.
- [4] P. Bühlmann, L. Meier, and S. Van De Geer, *Discussion: “A significance test for the lasso”*, *Ann. Stat.* 42 (2014), pp. 469–477.
- [5] T.T. Cai and M. Yuan, *Comments on “a significance test for the lasso”*, *Ann. Stat.* 42 (2014), pp. 478–482.
- [6] F. Chiaromonte and J. Martinelli, *Dimension reduction strategies for analyzing global gene expression data with a response*, *Math. Biosci.* 176 (2002), pp. 123–144.

- [7] F. Chiaromonte, R.D. Cook, and B. Li, *Sufficient dimensions reduction in regressions with categorical predictors*, Ann. Stat. 30 (2002), pp. 475–497.
- [8] R.D. Cook, *On the interpretation of regression plots*, J. Amer. Statist. Assoc. 89 (1994), pp. 177–189.
- [9] R.D. Cook, *Regression graphics: Ideas for studying regressions through graphics*, Wiley Series in Probability and Statistics, New York, 1998.
- [10] R.D. Cook and F. Chiaromonte, *Sufficient dimension reduction and graphics in regression*, Ann. Instit. Statist. Math. 54 (2002), pp. 768–795.
- [11] R.D. Cook and L. Forzani, *Likelihood-based sufficient dimension reduction*, J. Amer. Statist. Assoc. 104 (2009), pp. 197–208.
- [12] R.D. Cook, L. Forzani, and A.J. Rothman, *Estimating sufficient reductions of the predictors in abundant high-dimensional regressions*, Ann. Stat. 40 (2012), pp. 353–384.
- [13] R.D. Cook and L. Ni, *Sufficient dimension reduction via inverse regression: a minimum discrepancy approach*, J. Amer. Statist. Assoc. 100 (2005), pp. 410–428.
- [14] R.D. Cook and S. Weisberg, *Discussion of sliced inverse regression for dimension reduction*, by C.-K. Li, J. Amer. Statist. Assoc. 86 (1991), pp. 328–332.
- [15] R.D. Cook and X. Yin, *Dimension reduction and visualization in discriminant analysis*, Aust. N.Z. J. Stat. 43 (2001), pp. 147–199.
- [16] Y. Dong and B. Li, *Dimension reduction for non-elliptically distributed predictors: second-order methods*, Biometrika 97 (2010), pp. 279–294.
- [17] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression (with discussion)*, Ann. Stat. 32 (2004), pp. 407–499.
- [18] J. Fan and Z.T. Ke, *Discussion on “a significance test for the lasso”*, Ann. Stat. 42 (2014), pp. 483–492.
- [19] J. Fan and R. Li, *Variable selection via nonconvex penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc. 96 (2001), pp. 1348–1360.
- [20] J. Fan and R. Li, *Statistical challenges with high dimensionality: Feature selection in knowledge discovery*, in *Proceedings of the International Congress of Mathematicians*, Vol. III, M. Sanzsole, J. Soria, J.L. Varona, J. Verdera, eds., European Mathematical Society, Zürich, 2006, pp. 595–622.
- [21] K. Fukumizu, F.R. Bach, and M.I. Jordan, *Kernel dimension reduction in regression*, Ann. Stat. 37 (2009), pp. 1871–1905.
- [22] Y. Guan and H. Wang, *Sufficient dimension reduction for spatial point processes directed by gaussian random fields*, J. R. Statist. Soc. Ser. B (Statist. Methodol.) 72 (2010), pp. 367–387.
- [23] C. Hans, *Elastic net regression modeling with the orthant normal prior*, J. Amer. Statist. Assoc. 106 (2011), pp. 1383–1393.
- [24] T. Hastie, R. Tibshirani, and J.H. Friedman, *The elements of statistical learning: Data mining inference and prediction*, Springer, New York, 2009.
- [25] A.E. Hoerl and R.W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics 12 (1970), pp. 55–67.
- [26] K.-C. Li, *Sliced inverse regression for dimension reduction*, J. Amer. Statist. Assoc. 86 (1991), pp. 316–327.
- [27] B. Li and Y. Dong, *Dimension reduction for nonelliptically distributed predictors*, Ann. Stat. 37 (2009), pp. 1272–1298.
- [28] L. Li and X. Yin, *Sliced inverse regression with regularizations*, Biometrics 64 (2008), pp. 124–131.
- [29] J. Lv and Z. Zheng, *Discussion: a significance test for the lasso*, Ann. Stat. 42 (2014), pp. 493–500.
- [30] B.G. Osbourne, T. Fearn, A.R. Miller, and S. Douglas, *Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs*, J. Sci. Food Agric. 35 (1984), pp. 99–105.
- [31] P. Radchenko and G.M. James, *Improved variable selection with Forward-Lasso adaptive shrinkage*, Ann. Appl. Stat. 5 (2011), pp. 427–448.
- [32] A. Rothman, P. Bickel, E. Levina, and J. Zhu, *Sparse permutation invariant covariance estimation*, Ann. Stat. 2 (2008), pp. 494–515.

- [33] N.J. Salkind, *Encyclopedia of measurement and statistics, Vol. II*, Sage, Thousand Oaks, 2007.
- [34] T. Stamey, J. Kabalin, J. McNeal, I. Johnstone, F. Freiha, E. Redwine, and N. Yang, *Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients*, *J. Urology* 16 (1989), pp. 1076–1083.
- [35] R. Tibshirani, *Regression shrinkage and selection via the lasso*, *J. R. Statist. Soc: Ser B (Methodol.)* 58 (1996), pp. 267–288.
- [36] R. Tibshirani, J. Taylor, R. Lockhart, and R.J. Tibshirani, *A significance test for the lasso*, *Ann. Stat.* 42 (2014), pp. 413–468.
- [37] Y. Wu and L. Li, *Asymptotic properties of sufficient dimension reduction with a diverging number of predictors*, *Statist. Sin.* 21 (2011), pp. 707–730.
- [38] Y. Xia, H. Tong, W.K. Li, and L.X. Zhu, *An adaptive estimation of optimal regression subspace*, *J. R. Statist. Soc: Ser. B* 64 (2002), pp. 363–410.
- [39] Y. Xia, D. Zhang, and J. Xu, *Dimension reduction and semiparametric estimation of survival models*, *J. Amer. Statist. Assoc.* 105 (2010), pp. 278–290.
- [40] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, *J. R. Statist. Soc: Ser. B* 67 (2005), pp. 301–320.