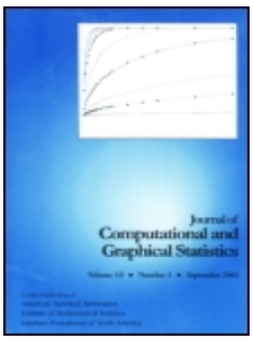

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Sufficient dimension reduction and variable selection for large-p-small-n data with highly correlated predictors

Haileab Hilafu & Xiangrong Yin

To cite this article: Haileab Hilafu & Xiangrong Yin (2016): Sufficient dimension reduction and variable selection for large-p-small-n data with highly correlated predictors, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2016.1164057

To link to this article: http://dx.doi.org/10.1080/10618600.2016.1164057

Sufficient dimension reduction and variable selection for large-*p*-small-*n* data with highly correlated predictors

Haileab Hilafu [*] and Xiangrong Yin [†]

Abstract

Sufficient Dimension Reduction (SDR) is a paradigm for reducing the dimension of the predictors without losing regression information. Most SDR methods require inverting the covariance matrix of the predictors. This hinders their use in the analysis of contemporary datasets where the number of predictors exceed the available sample size and the predictors are highly correlated. To this end, by utilizing the seeded SDR idea of Cook, Li and Chiaromonte (2007) and the sequential dimension reduction framework of Yin and Hilafu (2015), we propose a SDR method for high-dimensional data with correlated predictors. The performance of the proposed method is studied via extensive simulations. To demonstrate its use, an application to microarray gene expression data where the response is the production rate of riboflavin (vitamin B_2) is presented.

Key Words and Phrases: Central Subspace; High-dimensional data; Partial Inverse Regression; Partial Least Squares; Sufficient Dimension reduction

1 Introduction

Sufficient dimension reduction (SDR; Li, 1991; Cook, 1994, 1996) is a statistical methodology that is concerned with reducing the dimension of a predictor vector while preserving the regression

[*]Haileab Hilafu is Assistant Professor, Department of Business Analytics and Statistics, 233 Stokely Management Center, University of Tennessee, Knoxville, TN 37996. E-mail: hhilafu@utk.edu

[†]Xiangrong Yin is Professor, Department of Statistics, 319 Multidisciplinary Science Building, University of Kentucky, 725 Rose St., Lexington, KY 40536. E-mail:yinxiangrong@uky.edu.

relation with a response. Let X be a $p \times 1$ predictor vector, Y be a scalar response, $\mathcal{S}$ be a subspace of $\mathbb{R}^p$ and $P_{\mathcal{S}}$ be the orthogonal projection onto $\mathcal{S}$. If $\mathcal{S}$ is such that

$$Y \perp\!\!\!\perp X | P_{\mathcal{S}} X, \tag{1}$$

where $\perp\!\!\!\perp$ indicates statistical independence, then $\mathcal{S}$ is called *dimension reduction subspace* (Li 1991; Cook, 1998). The intersection of all such subspaces, if itself satisfies the above independence condition, is said to be the *central subspace* (CS), and is denoted by $\mathcal{S}_{Y|X}$. The dimension of $\mathcal{S}_{Y|X}$ is called the *structural dimension* of the regression, and is denoted by d, where d is often less than p ($d < p$). More precisely, suppose that β is a $p \times d$ basis matrix for the central subspace $\mathcal{S}_{Y|X}$. Then, we have $Y \perp\!\!\!\perp X | \beta^T X$. Thus, the regression of Y on X is equivalent to the regression of Y on the *sufficient predictor* $\beta^T X$ in the sense that $Y|X$ and $Y|\beta^T X$ have the same distribution. Therefore, a sufficient dimension reduction of the predictor vector X amounts to estimating a basis for the meta-parameter $\mathcal{S}_{Y|X}$, and its dimension d. The central subspace, $\mathcal{S}_{Y|X}$, is a well defined parameter under some mild conditions (Cook, 1996; Yin, Li, and Cook, 2008), and is assumed to exist hereafter.

An integral aspect of statistical analysis is variable selection. In the SDR framework, the sufficient predictors are linear (or non-linear) combinations of all the original predictors. Therefore, carrying out a simultaneous dimension reduction and variable selection is desirable. If Γ is a $p \times q$ matrix such that $Y \perp\!\!\!\perp X | \Gamma^T X$, where the columns of Γ consist unit vectors e_j with jth element being 1 and 0 otherwise, then the column space of Γ is called a *sufficient variable selection subspace* (SVS). The intersection of all such subspaces, if itself satisfies the above independence condition, is called the *central variable selection subspace* (Yin and Hilafu, 2015), denoted by $\mathcal{S}^V_{Y|X}$, with dimension s, $s \leq p$. Without loss of generality, we can take $\Gamma = (\mathbf{I}_s, \mathbf{0})^T$, where $\mathbf{I}_s$ is an $s \times s$ identify matrix (Cook, 2004). The two concepts, SDR and SVS, while closely related, are not the same. See Yin and Hilafu (2015) for a more detailed discussion on the similarities and differences between SDR and SVS.

The most widely used estimation methods for $\mathcal{S}_{Y|X}$ are based on inverse regression (Li, 1991; Duan and Li, 1991; Yin and Cook, 2003; Cook and Ni, 2005). These methods are easy to implement and have very fast algorithms. However, their implementation require inverting the sample covariance matrix of the predictors. When the number of predictors exceeds the available sample size, or the predictors are highly correlated, the inverse of the sample covariance matrix is not stable or does not exist. To this end, Cook, Li and Chiaromonte (2007) developed a SDR method that requires computing powers of the sample covariance matrix instead of its inverse. They utilized the ideas developed by Helland (1990) for partial least squares estimation, and established a connection between partial least squares and the SDR paradigm. However, even though their method avoids the need to invert the sample covariance matrix, it requires inverting a different but smaller matrix whose dimension depends on p (details in section 2.2). If p is large, say $p > 3n$, where n is the sample size, then this smaller matrix becomes singular and their method does not work (Cook, Li and Chiaromonte, 2007). To overcome this, Yin and Hilafu (2015) developed a dimension reduction framework for $p >> n$ through a sequential mechanism. Unlike Cook, Li and Chiaromonte (2007), Yin and Hilafu's (2015) framework requires inverting the sample covariance matrix – but only when the sample size exceeds the number of predictors, since they partition the predictor vector into sets of vectors of smaller dimension. Their novel sequential framework achieves SDR and SVS simultaneously for data with $p >> n$. However, their algorithms can not handle the case where predictors exhibit high correlations.

To the best of our knowledge, there is no sufficient dimension reduction method suited for high dimensional data ($p >> n$) with highly correlated predictors. In this paper, we fill this gap. We utilize the seeded sufficient dimension reduction idea in Cook, Li and Chiaromonte (2007) to deal with highly correlated predictors, and the sequential dimension reduction framework in Yin and Hilafu (2015) to deal with large p small n problem, simultaneously. In addition, we provide a bootstrap based sufficient variable selection.

The rest of the paper is organized as follows. In section 2 we review the work of Yin and Hilafu (2015), and that of Cook, Li and Chiaromonte (2007). We then present our proposed sequential partial inverse regression (SeqPIR) – including its estimation algorithms, thresholding based estimation of the structural dimension, a bootstrap-based sufficient variable selection approach, and a stable estimation procedure. In section 3 we present an extensive simulation study to demonstrate the effectiveness of the proposed method, and in section 4 we apply our method to a real data set. Section 5 contains a brief discussion.

The following notations will be used repeatedly in our exposition. For a matrix $A \in \mathbb{R}^{p\times p}$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^p$, $A\mathcal{S}$ stands for the set $\{AS : S \in \mathcal{S}\}$. For an arbitrary matrix B of p rows, span(B) denotes the subspace of $\mathbb{R}^p$ spanned by the columns of B. If $\mathbb{B} \in \mathbb{R}^{p\times q}$, then the projection onto span($\mathbb{B}$) relative to the Σ inner product has the matrix representation $\mathbb{B}(\mathbb{B}^T\Sigma\mathbb{B})^\dagger\mathbb{B}^T\Sigma$, where $\dagger$ indicates the Moore-Penrose inverse. A direct sum between two subspaces V_1 and V_2, denoted by $V_1 \bigoplus V_2$, is defined as: $V_1 \bigoplus V_2 = \{v_1 + v_2; v_1 \in V_1, v_2 \in V_2\}$.

2 Proposed Method

We first review the sequential dimension reduction framework of Yin and Hilafu (2015) in section 2.1, and the seeded dimension reduction method of Cook, Li and Chiaromonte (2007) in section 2.2. Then, we present our proposed *sequential partial inverse regression* (SeqPIR) method, along with its estimation algorithms and bootstrap variable selection in section 2.3.

2.1 Sequential Sufficient Dimension Reduction

Partition the predictor vector $X \in \mathbb{R}^p$ as $X^T = (X_1^T, X_2^T)$, where X_1 is $p_1 \times 1$ and X_2 is $(p - p_1) \times 1$, and let Y be the scalar response variable that could be either categorical or quantitative. Suppose the matrix $\beta_1 \in \mathbb{R}^{p_1\times d_1}$, $d_1 < p_1$, is such that

$$Y \perp\!\!\!\perp X_1 | (\beta_1^T X_1, X_2). \tag{2}$$

Then,

$$P(Y|X) = P(Y|X_1, X_2) = P(Y|\beta_1^T X_1, X_1, X_2) = P(Y|\beta_1^T X_1, X_2), \quad (3)$$

where the last equality holds because of condition (2). Since the dimension of $\beta_1^T X_1$ is less than that of X_1, equation (3) achieves sufficient dimension reduction of X_1. Therefore, we can replace the original predictor vector $X = (X_1, X_2)$ by the lower dimensional predictor vector $X^* = (\beta_1^T X_1, X_2)$ without loss of regression information on $Y|X$. Yin and Hilafu (2015) provided conditions that produce such a β_1, which we reproduce below for completeness.

Proposition 1 *(*Proposition 1: Yin and Hilafu, 2015*): Let X_1, X_2 and Y be as described above. Statement (i) or (ii) below implies statement (iii).*

$(i) : (Y, X_2) \perp\!\!\!\perp X_1|\beta_1^T X_1,$

$(ii) : X_2 \perp\!\!\!\perp X_1|(\beta_1^T X_1, Y)$ *and* $Y \perp\!\!\!\perp X_1|\beta_1^T X_1,$

$(iii) : Y \perp\!\!\!\perp X_1|(\beta_1^T X_1, X_2).$

Proposition 1 provides two ways to obtain a β_1 that satisfies the desired condition (2). Statement (i) can be viewed as a multivariate regression problem where $Y^* = (Y, X_2)$ serves as the multivariate response and X_1 is the predictor vector. Similarly, first part of statement (ii) can be viewed as a multivariate regression problem with a combination of categorical and quantitative predictors, where X_2 serves as a response, X_1 is the set of quantitative predictors and Y serves as the categorical predictor; the second part of statement (ii) is the usual univariate categorical response regression problem. Once an estimate for β_1 is obtained using an existing multivariate regression method, the process is repeated on the reduced set of predictors until no further reduction is possible. See Yin and Hilafu (2015) for more details.

2.2 Seeded Dimension Reduction

Inverse regression based sufficient dimension reduction methods, such as sliced inverse regression (SIR; Li, 1991) and sliced average variance estimation (SAVE; Cook and Weisberg, 1991), have attracted great attention due to their easy and fast computation. However, their implementation requires inverting the sample covariance matrix of the predictors. When the predictors exhibit high correlations, the sample covariance matrix is unstable or even singular – which hinders the use of these methods. To overcome this issue, Cook, Li and Chiaromonte (2007) proposed the seeded dimension reduction method that utilizes ideas from partial least squares (Helland, 1990) and avoids the need to invert the sample covariance matrix. We briefly describe their method below (also, see Li, Cook and Tsai, 2007).

Define a population seed as any matrix ν such that $\text{span}(\nu) \subseteq \Sigma\mathcal{S}_{Y|X}$ and possibly $\text{span}(\nu) = \Sigma\mathcal{S}_{Y|X}$, where Σ is the covariance matrix of the predictors. Then, if Σ is invertible, the columns of $\Sigma^{-1}\nu$ span $\mathcal{S}_{Y|X}$. For example, the seed vector for OLS is the $p \times 1$ covariance vector $\nu = \text{cov}(X, Y)$, and $\mathcal{S}_{Y|X}$ can be obtained by $\Sigma^{-1}\text{cov}(X, Y)$, if $d = 1$. Some methods for estimating $\mathcal{S}_{Y|X}$, including the least squares, require the linearity condition (A_1): *For any vector $b \in \mathbb{R}^p$, $E(b^T X|\beta^T X)$ is a linear function of $\beta^T X$*, where β spans the central subspace of interest. Hall and Li (1993) showed that, as p increases with d fixed, this linearity condition holds to a good approximation in many problems; see also Diaconis and Freedman (1984). Additional discussion of this condition, which is typically regarded as mild, was given by Cook and Ni (2005).

Suppose there is a subspace $\mathcal{M} \in \mathbb{R}^p$ that contains $\mathcal{S}_{Y|X}$. Then, projecting $\Sigma^{-1}\nu$ onto $\mathcal{M}$ with respect to the Σ inner product will result in the column space of $\Sigma^{-1}\nu$ itself, which spans the central subspace of interest $\mathcal{S}_{Y|X}$. More specifically, if the columns of a matrix R form a basis for $\mathcal{M}$, then $\Sigma^{-1}\nu = R(R^T\Sigma R)^{-1}R\nu$. Consequently, $R(R^T\Sigma R)^{-1}R\nu$ forms a basis for $\mathcal{S}_{Y|X}$ and Σ^{-1} is not required. Thus, estimating $\mathcal{S}_{Y|X}$ amounts to identifying $\mathcal{M}$ and ν. We seek to identify a space $\mathcal{M}$ large enough to cover $\mathcal{S}_{Y|X}$, but small enough to allow reasonable estimation based on the available data without

involving the inversion of Σ. Now define R_u as the following Krylov space,

$$R_u \equiv (\nu, \Sigma\nu, \cdots, \Sigma^{u-1}\nu), u = 1, 2, \cdots, \tag{4}$$

and note that as u increases, the $p \times ud$ matrices in (4) form a nondecreasing sequence of nested subspaces, where d is the dimension of ν. Also define

$$\mathcal{M}_{Y|X} = \bigoplus_{j=1}^{m} P_j \mathcal{S}_{Y|X} = \bigoplus_{j \in K} P_j \mathcal{S}_{Y|X} \tag{5}$$

where $\bigoplus$ denotes a direct sum between subspaces, P_j, $j = 1, \cdots, m$, is the orthogonal projection operator to the jth eigenspace relative to the Σ inner product, m is the number of non-zero eigenvalues of Σ, and $K \subseteq \{1, \cdots, m\}$ is the set of indices of the eigenspaces of Σ that are not orthogonal to $\mathcal{S}_{Y|X}$. Let k be the cardinality of K, and note that $\mathcal{S}_{Y|X} \subseteq \mathcal{M}_{Y|X}$. Theorem 1 of Cook, Li and Chiaromonte (2007) states that, if Σ is positive definite with q distinct non-zero eigenvalues k of which correspond to eigenspaces not orthogonal to $\mathcal{S}_{Y|X}$, there exists an integer $\tilde{u}, 1 \leq \tilde{u} \leq k$, such that span($R_u$) is strictly increasing until $u = \tilde{u}$, and settles upon $\mathcal{M}_{Y|X}$ thereafter. Since $\mathcal{S}_{Y|X} \subseteq \mathcal{M}_{Y|X}$, the subspaces span($R_u$), $u = 1, \cdots$, will grow to contain $\mathcal{S}_{Y|X}$ after some u^* steps, $u^* \leq \tilde{u} \leq k$. Therefore, we have $\Sigma^{-1}\nu = R_{u^*}(R_{u^*}^T \Sigma R_{u^*})^{-1} R_{u^*}^T \nu$. Since the matrix R_{u^*} only require calculations of powers of Σ, inverting Σ is not required. Thus, if we have an estimate of u^*, we can obtain an estimate of $\mathcal{S}_{Y|X}$ using the matrix R_{u^*} without inverting Σ. Cook, Li and Ciaromonte (2007) proposed to use the d eigenvectors of the SIR kernel matrix, Cov$\{\mathbb{E}(X|Y)\}$, corresponding to the d non-zero eigenvalues as the seed matrix ν. It is also worth noting that if ν = Cov(X,Y) the method is equivalent to the partial least squares (Helland, 1990).

2.3 Sequential Partial Inverse Regression

Our development of the proposed *sequential partial inverse regression* (SeqPIR) is through statement (*i*) [(*ii*)] when the response variable is quantitative [categorical]. Suppose that the response

variable Y is quantitative. Our goal is to characterize the conditional distribution of Y given X. In section 2.1 we have shown that if there exists a $p_1 \times d_1$ matrix β_1 such that $Y \perp\!\!\!\perp X_1 | (\beta_1^T X_1, X_2)$, then $P(Y|X) = P(Y|X_1, X_2) = P(Y|\beta_1^T X_1, X_2)$. Statement (i) of proposition 1 established that: $(Y, X_2) \perp\!\!\!\perp X_1 | \beta_1^T X_1$ implies $Y \perp\!\!\!\perp X_1 | (\beta_1^T X_1, X_2)$. Observe that $(Y, X_2) \perp\!\!\!\perp X_1 | \beta_1^T X_1$ is a usual dimension reduction problem with a multivariate response, (Y, X_2). Therefore, we could carry out a dimension reduction with multivariate response to obtain an estimate for β_1. This is the approach we take, and we describe it below.

For ease of exposition, denote the multivariate response (Y, X_2) by Y^* with dimension q, and let t be a random vector generated from the unit sphere in $\mathbb{R}^q$. Also, let $\mathcal{S}_{Y^*|X_1}$ denote the central subspace for the regression of Y^* on X_1. Therefore, we could take β_1 to be the column space that spans $\mathcal{S}_{Y^*|X_1}$. To obtain a basis for $\mathcal{S}_{Y^*|X_1}$ we proceed as follows. Using Theorem 3.1 of Li, Wen and Zhu (2008), if the linearity condition (A_1) on page 6 holds, we have

$$\mathbf{M} \equiv \mathbb{E}_t\{\mathrm{Cov}\{\mathbb{E}(X_1|t^T Y^*)\}\} \subseteq \Sigma_1 \mathcal{S}_{Y^*|X_1}, \tag{6}$$

where Σ_1 is the covariance matrix of the predictor vector X_1. The matrix $\mathrm{Cov}\{\mathbb{E}(X_1|t^T Y^*)\}$ is the sliced inverse regression kernel matrix for the regression of $t^T Y^*$ on X_1 (Li, 1991). Equation (6) suggests that to estimate the central subspace $\mathcal{S}_{Y^*|X_1}$, it suffices to obtain estimates for the central subspaces $\mathcal{S}_{t^T Y^*|X_1}$, for $t \in \mathbb{R}^q$, and pool these subspaces. Of course, it is not practical to obtain $\mathrm{Cov}\{\mathbb{E}(X_1|t^T Y^*)\}$ for every $t \in \mathbb{R}^q$. Thus, we will do so for a sufficiently large (relative to the sample size) collection of t. More specifically, let $\mathbf{M}_t$ be a sample estimate for $\mathrm{Cov}\{\mathbb{E}(X_1|t^T Y^*)\}$. We pool these estimates to obtain an estimate $\mathbf{M}_{n,m}$ of $\mathbf{M}$ as: $\mathbf{M}_{n,m} = \frac{1}{m}\sum_{i=1}^m \mathbf{M}_{t_i}$, where m is the number of projection vectors t that are considered and n is the sample size. Following, we use the d_1 eigenvectors corresponding to the d_1 largest eigenvalues of $\mathbf{M}_{n,m}$ as the seed matrix to obtain an estimate for β_1 using the seeded dimension reduction method described in section 2.2. Detailed algorithms for the estimation procedure of our method are given below.

2.3.1 Estimation Procedure for SeqPIR

Estimation for the proposed sequential partial inverse regression is achieved through algorithm 1. Steps 1, 2 and 3 are repeated until the dimension of the predictor vector no longer exceeds the sample size.

Algorithm 1 : SeqPIR

1. Decompose $X \in \mathbb{R}^p$ into $X = (X_1, X_2)$, where X_1 is a $p_1 \times 1$ vector, X_2 is $(p - p_1) \times 1$.
2. Consider the dimension reduction problem $(Y, X_2) \perp\!\!\!\perp X_1 | \beta_1^T X_1$. Use **Algorithm 2** to obtain an estimate for β_1, with (Y, X_2) as the response and X_1 as the predictor.
3. Replace predictor X by $(\beta_1^T X_1, X_2)$ and go back to step 1.

In theory, one could partition the predictor vector X arbitrarily into two sets of vectors X_1 and X_2. However, if one has a prior information on the partitioning of the predictors, say, sets of predictors that are functionally related, then one can use such a partitioning scheme. In our numerical studies, we first order the predictors based on their marginal correlation coefficient with the response variable – where predictors with smaller correlation are put in X_1. We report simulation results for $p_1 = n/2, n, 2n$.

Step 2 in algorithm 1 uses algorithm 2 to perform *multivariate response partial inverse regression*. For a given q-dimensional multivariate response Y, p-dimensional predictor vector X, let (X_i, Y_i), $i = 1, \cdots, n$, be an *iid* realizations from the joint distribution of (X, Y). Then, for given values of d and u (estimation of which will be discussed in section 2.3.2), algorithm 2 presents the details for multivariate response partial inverse regression. In addition to the values of d and u, algorithm 2 also requires choosing a monte carlo sample size m, the number of random projections of the multivariate response. This monte carlo sample size should be chosen such that it goes to ∞ faster than n (Li, Wen and Zhu, 2008). This is required for the estimate $\mathbf{M}_{n,m}$ to be a consistent estimate of the population kernel $\mathbf{M}$ (Li, Wen and Zhu, 2008). In addition, in order for the projections to maintain the integrity of the information that relates the response and the predictor vectors,

m should be chosen to be larger than q. On the other hand, theoretically, Hilafu and Yin (2013) showed that if the dimension of the central subspace is d, there exists at most $m = d$ projections that completely recover the central subspace, though they could not tell how to choose these projections. In all our simulations and application, we use $m = n^{3/2}$, which is what Li, Wen and Zhu, (2008) suggested, and it seems to work very well.

Algorithm 2 : MultiPIR

1. Generate an *iid* sample $t_1, \cdots, t_m$ from the uniform distribution on the unit sphere $\mathbb{S}^q = \{t \in \mathbb{R}^q : ||t|| = 1\}$. E.g. take $t_j = G_j/||G_j||$, where $G_1, \cdots, G_m$ are *iid* $N(0, I_q)$.
2. Compute moment estimates $\widehat{\mu}$ and $\widehat{\Sigma}$ for the mean and covariance matrix of X.
3. For each t_j, $j = 1, \cdots, m$, compute the $p \times H$ SIR kernel matrix from $(X_1, t_j^T Y_1), \cdots, (X_n, t_j^T Y_n)$, where H is the number of slices, as follows:
   (a) Divide range of $t_j^T Y = (t_j^T Y_1, ..., t_j^T Y_n)$ into H slices, $I_1, ..., I_H$; let the proportion of the $t_j^T Y_i$'s that fall in slice h be $\widehat{p}_h, h = 1, ..., H$.
   (b) Within each slice, compute the sample mean of X, $\widehat{\mu}_h(t_j) = (1/n\widehat{p}_h) \sum_{\{t_j^T Y_i \in I_h\}} X_i$.
   (c) Compute the $p \times H$ kernel matrix: $\mathbf{M}_n(t_j) = \sum_{h=1}^{H} \widehat{p}_h \left(\widehat{\mu}_h(t_j) - \widehat{\mu}\right)\left(\widehat{\mu}_h(t_j) - \widehat{\mu}\right)^T$.
4. Compute the $p \times H$ aggregate kernel matrix: $\mathbf{M}_{n,m} = \frac{1}{m} \sum_{j=1}^{m} \mathbf{M}_n(t_j)$.
5. Let $\widehat{v}_1, ..., \widehat{v}_d$ be the d eigenvectors of $\mathbf{M}_{n,m}$ corresponding to its d largest eigenvalues. Set the seed matrix $v = (\widehat{v}_1, ..., \widehat{v}_d)$.
6. Return $\beta = R_u(R_u^T \widehat{\Sigma} R_u)^{-1} R_u^T v$ as the basis for the target subspace $\mathcal{S}_{Y|X}$, where $R_u = (v, \widehat{\Sigma}v, \cdots, \widehat{\Sigma}^{u-1}v)$.

Algorithm 2 is designed for when the response variable is quantitative. We turn the problem into a multivariate response dimension reduction problem through statement (i) of proposition 1. However, if the response variable is categorical, we use statement (ii) of proposition 1 to turn the problem into a multivariate response dimension reduction problem with a combination of categorical and quantitative predictors: $X_2 \perp\!\!\!\perp X_1 | (\beta_1^T X_1, Y)$. Suppose that Y has K categories, and n_k is the number of observations that fall into category $k, k = 1, ..., K$. To obtain the estimate $\mathbf{M}_{n,m}$, we repeat steps 1, 2, 3 and 4 on the sub-populations created by the categories of Y. More specifically,

let $\mathbf{M}_{n_k,m} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{M}_{n_k}(t_j)$ be the kernel matrix in step 4 for the k^{th} sub-population corresponding to $Y = k$. Then, we compute the aggregate kernel matrix $\mathbf{M}_{n,m}$ as: $\mathbf{M}_{n,m} = \frac{1}{n}\sum_{k=1}^{K} n_k\, \mathbf{M}_{n_k,m}$, where $n = n_1 + ... + n_K$.

2.3.2 Estimation of d and u

Our estimation algorithms in section 2.3.1 assume both d and u are known. However, in practice, d and u also need to be estimated from the data. We discuss a simple and intuitive approach to estimate both d and u. To proceed, note that d is the number of non-zero eigenvalues of the $p \times H$ population kernel matrix, $\mathbf{M}$, whose estimate is given by $\mathbf{M}_{n,m}$ in algorithm 2. The last $(p - d)$ eigenvalues of the kernel matrix $\mathbf{M}$ are equal to zero. However, their sample versions will be small but not necessarily exactly zero. Therefore, we use the following thresholding approach to estimate d.

$$\widehat{d} = \sum_{j=1}^{p-1} I(r_j > \alpha), \text{with } \; r_j = \lambda_j/\lambda_{j+1},\, j = 1, \cdots, p-1 \tag{7}$$

where $I(.)$ is the indicator function, α is a pre-specified thresholding value, and $\lambda_1, \cdots, \lambda_p$ are the eigenvalues of $\mathbf{M}_{n,m}$ in descending order. Li, Cook and Tsai (2007) suggested the use of $\alpha = 1.5$ and our empirical sensitivity analysis confirmed that this choice of α works well (see Figure 1). Therefore, we used α to be 1.5 in all our numerical studies.

Note that, the seeded dimension reduction discussion in section 2.2 required that $\nu \subseteq \text{span}(\gamma_1, \cdots, \gamma_m)$, where $\gamma_1, ..., \gamma_m$ are the eigenvectors of Σ that are not orthogonal to $\mathcal{S}_{Y|X}$. Therefore, u should be chosen no less than m. Note also that the sequence of subspaces $\{\Sigma^j\nu,\, j = 0, 1, \cdots\}$ provide a basis for $\text{span}(\gamma_1, \cdots, \gamma_m)$. Therefore, u should be chosen such that $du \leq p$ to avoid a singular R_u. That is, u should be chosen such that $m < u < p/d$. An optimal choice of u is the rank of the matrix $R_p = (\nu, \Sigma\nu, \cdots, \Sigma^{p-1}\nu)$ or $R_pR_p^T$. In the case where p is very large, we use the estimated rank of the matrix $R_{Kn}R_{Kn}^T$ for a $K_n \to \infty$ as $n \to \infty$ as an estimate for u. For example, we

can choose $K_n = O(\log(n)^{3/4})$) (Zhu and Zhu, 2009). We follow the threshoding approach given in equation (7) to obtain an estimate $\widehat{u}$ of the rank of $R_{Kn}R_{Kn}^T$. Our experience with this thresholding method for obtaining both $\widehat{d}$ and $\widehat{u}$ is that values of α between 1 and 2 generally tend to work well. However, one could try different values of α and choose $\widehat{d}$ and $\widehat{u}$ that are selected most frequently.

2.3.3 Sufficient Variable Selection

Sufficient variable selection is a very important aspect of statistical analysis. The sufficient predictors that a sufficient dimension reduction method yields often are linear (or non-linear) combinations of all the original predictors. However, in many applications, it is often believed that only a subset of the predictors are truly important in predicting the response variable. Therefore, identifying these important variables is critical both for interpretability and enhancing the predictive ability of the subsequent model. In the SDR context, determining whether a predictor is truly informative is equivalent to testing whether the corresponding row in the $p \times d$ basis matrix β is truly different from zero (Cook, 2004). Li and Yin (2008) induced a combination of l_1 and l_2 norms to the least squares formulation of the sliced inverse regression proposed by Cook (2004) to achieve variable selection for high-dimensional data. However, their method is restricted to the sliced inverse regression (Li, 1991). Li (2007) proposed sparse SDR by casting the inverse regression SDR methods as a generalized eigenvalue decomposition problem, and imposing a column-wise l_1 norm penalty to achieve variable selection. However, his method requires inverting the sample covariance matrix and, therefore, it is not applicable to the high-dimensional setting. Yin and Hilafu (2015) incorporated sparse SDR (Li, 2007) in their sequential dimension reduction framework to achieve sufficient dimension reduction and sufficient variable selection simultaneously for high-dimensional data. While their approach is attractive in principle, it is computationally intensive as the tuning parameters for both the l_1 and l_2 norm penalties need to be chosen at each step of the sequential reduction framework. For computational simplicity, in this paper, we resort to a simple bootstrap based variable selection that does not require selecting multiple tuning parameters.

The bootstrap variable selection procedure is as follows. First, we generate B bootstrap samples as follows: for $m = 1, \cdots, B$, we draw with replacement n pairs $(y_i^{(m)}, \mathbf{x}_i^{(m)})$ from the original sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ at random. Then, we apply sequential partial inverse regression to each bootstrap sample to obtain estimates $\widehat{\beta}^{(m)}$ of β. Let $\widehat{\Sigma}_j$ be the bootstrap covariance matrix of the estimates corresponding to x_j from the B bootstrap samples, $j = 1, \cdots, p$, where $\widehat{\Sigma}_j$ is a $d \times d$ matrix, if β is $p \times d$. Let the $\widehat{\beta}_j$ be the $d \times 1$ coefficient vector of the jth element in the estimate $\widehat{\beta}$ based on the original data. We compute the squared Mahalanobis distance between $\widehat{\beta}_j$ and the origin as as follows

$$\Gamma_j = \widehat{\beta}_j^T \widehat{\Sigma}_j^{-1} \widehat{\beta}_j.$$

We then use the χ_d^2 as the approximate reference distribution for Γ_j. If $\Gamma_j > \chi_d^2(0.05)$, we declare x_j as an important variable. A similar bootstrap based variable selection approach has been used for the sliced inverse regression by Zhong et al. (2005).

2.3.4 Stable Estimation

Algorithm 2 employs the well known sliced inverse regression method (SIR; Li, 1991). When the response variable is quantitative, estimating the kernel matrix using the sliced inverse regression requires partitioning the response variable into H non-overlapping slices. Even though the number of slices may affect the estimation accuracy, there are no criteria or guidelines on how to choose the number of slices. To stabilize the estimate against the choice of the number of slices, we propose to use an ensemble of estimates obtained using different number of slices. More specifically, let the set $\mathcal{H}$ contain a collection of the potential number of slices. Also, let $\widehat{\beta}_H$ be the estimate using $H \in \mathcal{H}$. We aggregate these estimates as follows. Set

$$\mathbb{M} = \frac{1}{|\mathcal{H}|} \sum_{H \in \mathcal{H}} \widehat{\beta}_H \widehat{\beta}_H^T, \tag{8}$$

where $|.|$ is the cardinality of a set. Then, we take the final estimate to be the d eigenvectors of $\mathbb{M}$ corresponding to its d largest eigenvalues. An alternative way to aggregate the estimates is to pool

the kernel matrices instead of the estimated directions. That is, define the pooled kernel matrix as follows:

$$\mathbf{M}_{n,m} = \sum_{H \in \mathcal{H}} \mathbf{M}^H_{n,m} (\mathbf{M}^H_{n,m})^T = \sum_{H \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{j=1}^{m} \mathbf{M}_n(t_j) \mathbf{M}_n(t_j)^T \right\}, \tag{9}$$

where $\mathbf{M}^H_{n,m}$ is the kernel matrix obtained from steps 1, 2, 3 and 4 of algorithm 2 with H slices. The ensemble approach through (9) has recently been used in Wu and Yin (2015) and Cook and Zhang (2014). However, our simulation studies revealed that ensemble estimates through (8) are slightly better than through (9), specially for variable selection accuracy. We suspect that this could be because aggregation through (9) contains noise directions while (8) only aggregates the informative directions. Therefore, we report simulation results for ensemble estimates only through (8).

3 Simulations

3.1 Estimation and Variable Selection Accuracy

We assess the performance of the proposed method through extensive simulations. Our simulations are characterized by the structure of the covariance matrix of the predictors Σ, and the size of p_1. We consider three values for p_1 ($p_1 = n/2, n, 2n$), and three structures for Σ determined by ρ, where the diagonal entries of Σ are 1's and the off-diagonal entries are all equal to ρ ($\rho = 0, 0.5, 0.9$). We simulate the predictor vector X from a Gaussian distribution with 0 mean and covariance matrix Σ. We set the dimension of the vector of predictors to $p = 500$ and the sample size to $n = 100$. The error term ε is generated from a Gaussian distribution with mean 0 and variance 1, and the noise level $\sigma = 0.5$. The existing approaches, such as Cook, Li and Chiaromonte (2007), would not work for the models here as $p >> 3n$. We set the number of slices for the SIR estimation to 5. In section 3.3, we offer additional simulation results to study the effect of the choice of number of slices. Data was generated using the following four models.

Model 1 : $y = e^{(2-X\beta)} + \sigma\varepsilon$, where $\beta = (-0.5, 1, 0.5, 1, -1, -0.8, 0.8, 1, 0.5, 0.75, 0, ..., 0)$.

Model 2 : $y = e^{0.75(X\beta)}\sigma\varepsilon$, where $\beta = (-0.5, 1, 0.5, 1, -1, -0.8, 0.8, 1, 0.5, 0.75, 0, ..., 0)$.

Model 3 : $X|y = 0 \sim N(0, \Sigma), X|y = 1 \sim N(\beta, \Sigma)$,
where $\beta = (-0.5, 1, 0.5, 1, -1, -0.8, 0.8, 1, 0.5, 0.75, 0, ..., 0)$.

Model 4 : $X_y = \mu + \Gamma f_y + E$, where $y \sim U(0, 1)$, $f_y = (y, y^2), \mu = 0$, $E \sim 0.5 * N(0,\Sigma)$. We set $\Gamma = (\beta_1, \beta_2)$ where $\beta_1 = (0.5, -0.75, 0, ..., 0)$ and $\beta_2 = (0, 0, 0.75, 0.5, 0, ..., 0)$

The forms of models 1 and 2 are adopted from the *partial inverse regression* paper by Li, Cook and Tsai (2007), but we modified the non-zero coefficients to have different magnitude and sign. Model 1 is homoscedastic, while model 2 is heteroscedastic. Model 3 is an inverse model, where the distribution of X depends on Y only through its mean. Model 4 is a 2-dimensional inverse homoscedastic model which we adopted from Cook, Forzani and Rothman (2012). To assess estimation accuracy, for the 1-dimensional models we compute $|\text{corr}(\widehat{\beta}^T X, \beta^T X)|$, the absolute correlation between the estimated sufficient predictor and the true sufficient predictor as commonly used by Li, Cook and Tsai (2007), and Cook, Li and Chiaromonte (2007), among others. For model 4 we compute the squared trace correlation coefficient. For a pair of generic random vectors U and V, the squared trace correlation coefficient is defined as $r^2 = \text{tr(A)/dim(A)}$, where $\text{A} = \Sigma_V^{-1/2}\Sigma_{VU}\Sigma_U^{-1}\Sigma_{UV}\Sigma_V^{-1/2}$, Σ_V, Σ_U are the variance matrices of U and V, respectively, and Σ_{UV} is the covariance matrix between U and V. We report $|r| = \sqrt{r^2(U, V)}$ with $U = \widehat{\beta}^T X$ and $V = \beta^T X$. The reported values are mean and standard deviation of the performance metric from 100 simulated datasets. The closer the absolute correlation (and the squared trace correlation) to one the better estimation of the central subspace. To assess variable selection performance, we report two metrics: the true positive rate (TPR) defined as the proportion of truly important variables that are identified as important by the method, and the false positive rate (FPR) defined as the proportion of truly unimportant variables that are identified as important by the method. TPR is also referred to as sensitivity, and FPR as 1-specificity. We also compare *sequential partial inverse regression* (SeqPIR) with the *sequential sufficient dimension reduction* (SeqSDR; Yin and Hilafu, 2015).

Table 1 presents the simulation results for all the models. Concerning estimation accuracy, for uncorrelated predictors ($\rho = 0.0$), our method is comparable to the sequential sufficient dimension reduction method of Yin and Hilafu (2015), denoted by SeqSDR in Table 1. For SeqSDR, we report results only with $p_1 = n/2 = 50$ since the estimation requires inverting the sample covariance matrix of the predictors. We observe that our proposed SeqPIR outperforms SeqSDR when the predictors are correlated ($\rho = 0.5, 0.9$). In addition, the performance of the estimates using our SeqPIR method do not seem to heavily depend on the choice of p_1. Nevertheless, the results for $p_1 = 50$ and $p_1 = 100$ are slightly better than the results for $p_1 = 200$, albeit not significantly. Simulation results not reported in this manuscript has revealed that the performance of SeqPIR estimates deteriorate if p_1 becomes too large relative to the sample size (say, $p_1 > 2n$) – which confirms the findings of Cook, Li and Chiaromonte (2007). Our experience from the extensive simulation is that optimum results are achieved when $p_1 \approx n$. Our simulations also show that SeqPIR results for correlated predictors are generally superior to the results for uncorrelated predictors.

The bootstrap based variable selection also works very well in identifying the important variables – as evidenced by the high TPR for all models and all model settings. The method also discards most of the unimportant variables as shown by the relatively small FPRs. The variable selection by incorporating the sparse SDR idea of Li (2007) into SeqSDR also works well. The two methods are comparable with regard to variable selection although SeqSDR tends to have higher FPR when the predictors are correlated. Overall, our simulation results show that our proposed method is a viable dimension reduction method for high-dimensional data with correlated predictors.

To give the reader an idea of the computation time involved, we have measured the time it takes to obtain an estimate of the subspace of interest given a dataset. We run this for model 2 by setting $p = 500, n = 100, m = n^{3/2} = 1000$, and $p_1 = n = 100$. We run the model on a personal laptop with the following specifications. Processor: Intel(R) Core(TM) i5-4200U CPU @ 2.30 GHz RAM: 6

GB.

Model	Method	Setting	Time(s)
2	Seeded DR	$n = 100, p = 100, h = 5$	0.19
	SeqPIR	$n = 100, p = 500, h = 5, p_1 = 100, m = n^{3/2} = 1000$	10.52

The table above summarizes the comparison of our SeqPIR method with the seeded dimension reduction (Cook, Li and Chiaromonte, 2007). The SeqPIR would need 5 steps of the seeded dimension reduction since $p = 500$ and $p_1 = 100$. In each step, we use $m = 1000$ projections of the response variable. Since the algorithm for the sliced inverse regression is very fast, our method does not take long to run.

3.2 Estimating Structural Dimension

This section reports simulation results to assess the accuracy of the method in section 2.3.2 in estimating the structural dimension (d). We consider the four models above. As in the simulations above, we take $n = 100$ and $p = 500$, but we report results only for $\rho = 0.9$. Figure 1 presents the results for a range of α values between 1 and 5, with increments of 0.25. The reported accuracy refers to the proportion of times the method identifies the structural dimension correctly. We repeat the process for 1000 simulated data sets. For models 1 and 4 the method identifies d correctly with 100% accuracy for values of α between 1 and 2. For model 2, the method identifies d correctly with at least 90% accuracy for values of α between 1 and 2. For the 1-dimensional inverse model with binary response (model 3), we estimate the kernel matrix using the sliced average variance estimation (Cook and Weisberg, 1991) as SIR is known to yield only one direction when the response is binary. The plot for this model shows that the method identifies d correctly with 100% accuracy for all the α values we consider in our simulation, though, of course one would expect a deterioration of the results as we continue to increase α. The results for this model are not surprising as the inverse regression methods work best when the model is inverse. Similar

behavior is observed in the simulation results because the estimation accuracy is generally better for this model.

3.3 Stable Estimation

To study the effect of the choice of the number of slices, we use the ensemble approach. Our simulation results in Table 1 were obtained by setting the number of slices $h = 5$ when estimating the SIR kernel matrix $\text{Cov}\{\mathbb{E}(X|Y)\}$, for quantitative Y. Here, we report simulation results using the ensemble estimation presented in section 2.3.4. Table 2 presents these simulation results for model 2. The left half of the table presents results using $h = 5$ and right half presents results from the ensemble estimation using $\mathcal{H} = \{4, ..., 20\}$. The results show that the ensemble estimates are very similar to the estimates using just $h = 5$, with slightly improved false positive rates (FPR).

4 Application

We apply our method to a data set about riboflavin (vitamin B_2) production with *B. subtilis*. The data is adopted from Bühlmann, Kalisch and Meier (2014), and can be found at http://www.annualreviews.org/doi/suppl/10.1146/annurev-statistics-022513-115545 (file name `riboflavin`). There is a single real-valued response variable, which is the logarithm of the riboflavin production rate, and $p = 4{,}088$ (co)variables that measure the logarithm of the expression level of 4,088 genes; these gene expression profiles were normalized using the default in the `R package affy` (Gautier et al. 2004). The data consists of $n = 71$ samples that were hybridized repeatedly during a fed-batch fermentation process in which different engineered strains and strains grown under different fermentation conditions were analyzed (fore more details, also see Lee et al., 2001 , and Zamboni et al., 2005).

First, we randomly split the data into a training set of 50 samples and a test set of 21 samples. Following, using the training set, we employed variable screening using a marginal simple linear

regression of the logarithm of the production rate on the logarithm of the expression profiles of each of the genes. This type of marginal variable screening is commonly used in microarray gene expression analysis as a pre-processing step. We selected the most significant 500 genes (with the smallest *p*-vlaues) for further analysis. Panel (a) in Figure 2 presents a histogram of the pair-wise absolute correlation coefficients for these selected 500 genes. We see that about 40% of the pair-wise absolute correlation coefficients exceed 0.5, with about 25% exceeding 0.6. It is clear that many of the genes are highly-correlated.

Following, we applied sequential partial inverse regression to the training data with the 500 genes. The thresholding method in section 2.3.2 estimated the structural dimension to be $\widehat{d} = 1$, with $\alpha = 1.5$. We also used the bootstrap based variable selection method, with 1000 bootstrap samples, to determine the important genes. This procedure yielded 25 informative genes (gene names given below). Figure 2 panel (b) depicts the sufficient summary plot, the scatter plot of the sufficient predictor and the log-riboflavin-production rate, with an embedded loess (locally weighted scatterplot smoothing) fit. We used this loess fit to compute predicted values both for the training and the test samples. Figure 2 panel (c) and panel (d) display the scatterplot of the actual and predicted values for the 50 training samples and 21 test samples, respectively. The plots show that the non-parametric model using the resulting sufficient predictor performs well in terms of prediction. As expected, the prediction performance is better for the training samples than the test samples.

The names of the 25 genes that were selected by our method are as follows: COMP_at, YVRK_at, YTNP_at, YMDA_at, YUKJ_at, YHEM_at, YQJG_at, YTLP_at, YSFE_at, AHPC_at, YUKD_at, MUTM_at, YDFO_at, YRKH_at, YSIA_at, RPSE_at, RPLF_at, YTIA_at, YTGC_at, YTGD_at, YTGA_at, YCEA_at, YTGB_at, YDAR_at, YCKC_at.

5 Discussion

In this paper we presented sequential partial inverse regression, a sufficient dimension reduction method that takes advantage of ideas from partial least squares estimation and sequential reduction framework of Yin and Hilafu (2015). This method is shown to work well for large-p-small-n problems with correlated predictors. Extensive empirical studies demonstrate the efficacy of the method in a variety of settings. We have also applied the method to a microarray gene expression data which exhibits high correlation among expression profiles. Our bootstrap based variable selection yielded 25 genes whose profile expressions could be used to predict the riboflavin production rate. Since most high-dimensional data sets exhibit high collinearity among the predictors, the proposed method broadens the scope of sufficient dimension reduction.

Acknowledgments

We thank the editor, an associate editor, and two referees for helpful comments and suggestions which greatly improved the manuscript.

References

[1] Cook, R. D. (1994), On the interpretation of regression plots. *Journal of the American Statistical Association*, **89**, 177–190.

[2] Cook, R. D. (1996), Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983–992.

[3] Cook, R. D. (1998), Regression Graphics: Ideas for studying regressions through graphics. Wiley: New York.

[4] Cook, R. D. (2004), Testing predictor contributions in sufficient dimension reduction. *Annals of Statistics*, **32**, 1061-1092.

[5] Cook, R.D, Forzani, L. and Rothman, A. J. (2012), Estimating Sufficient Reductions reductions of the Predictors in Abundant High-dimensional Regressions. *Annals of Stat.*, **40**, 353-384.

[6] Cook, R.D, Li, B. and Chiaromonte, F. (2007), Dimension reduction in regression without matrix inversion. *Biometrika*, **94**, 569-584.

[7] Cook, R.D, and Ni, L. (2005), Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of American Stat. Assoc.* , **100**, 410-428.

[8] Cook, R. D and Weisberg, S. (1991), Discussion of Li 1991. *Jour. of Amer. Stat. Assoc.*, **86**, 328-332.

[9] Cook, R. D and Zhang, X. (2014), Fused Estimators of Central Subspaces in Sufficient Dimension Reduction. *Jour. of Amer. Stat. Assoc.*, **109**, 815-827.

[10] Diaconis, P. and Freedman D. (1984), Asymptotics of graphical projection pursuit. *Ann. Statist.*, **12**, 793815.

[11] Duan, N. and Li, K.C. (1991), Slicing Regression: a link-free regression method. *Ann. Statist.*, **19**, 505–530.

[12] Gautier, L., Cope, L., Bolstad, B. and Irizarry, R. (2004), Affyanalysis of Affymetrix-GeneChip data at the probe level. *Bioinformatics*, **20**, 307–15.

[13] Hall, P. and Li, K. C. (1993), On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.*, **21**, 867–889.

[14] Hilafu, H. and Yin, X. (2013), Sufficient Dimension Reduction for Multivariate Regressions with Categorical Predictors. *Comp. Stat. and Data Anal.*, **63**, 139-147.

[15] Lee, J-M., Zhang, S., Saha, S., Anna, SS., Jiang, C. and Perkins, J. (2001), RNA expression analysis using an antisense Bacillus subtilis genome array. *J. Bacteriol.*, **183**, 737180.

[16] Li, B., Wen, S and Zhu, L. (2008), On a projective resampling method for dimension reduction with multivariate responses. *Jour. of the Amer. Stat. Assoc.*, **103**, 1177-1186.

[17] Li, K-C. (1991), Sliced inverse regression for dimension reduction (with discussion). *Jour. of the Amer. Stat. Assoc.*, **86** 316-342.

[18] Li, L., Cook, R. D and Tsai, C. L. (2007), Partial Inverse Regression. *Biometrika*, **94**, 615–625.

[19] Li, L. and Yin, X. (2008), Sliced inverse regression with regulations. *Biometrics*, **64**, 124–131.

[20] Wu, W. and Yin, X. (2015), Stable Estimation in Dimension Reduction. *Jour. Of Comp. and Graphical Stats.*, **24**, 104–120.

[21] Yin, X. and Cook,R.D. (2003), Estimating central subspaces via inverse third moments. *Biometrika*, **90**, 113–125.

[22] Yin, X., Li, B. and Cook, R. D. (2008), Successive direction extraction for estimating the central subspace in a Multiple-index regression. *Journal of Multivariate Analysis*, **99**, 1733–1757.

[23] Yin, X. and Hilafu, H. (2015), Sequential Sufficient Dimension Reduction for Large *p* Small *n* Problems. *J. of the Royal Stat. Soc. - B*, **77**, 879–892.

[24] Zamboni, N., Fischer, E., Muffler, A., Wyss, M., Hohmann ,H-P. and Sauer, U. (2005), Transient expression and flux changes during a shift from high to low riboflavin production in continuous cultures of *Bacillus subtilis*. *Biotechnol. Bioeng.*, **89**, 219–232.

[25] Zhong, W., Zeng, P, Ma, P., Liu, J.S., and Zhu, Y. (2005), RSIR: Regularized sliced inverse regression for motif discovery. *Bioinformatics*, **21**, 4169–4175.

[26] Zhu, L-P., and Zhu, L. (2009), On distribution-weighted partial least squares with diverging number of highly correlated predictors. *Jour. of the Royal Stat. Soc.: B*, **71**, 525–548.

Table 1: Mean and standard deviation (in []) of the absolute correlation (trace correlation) for models 1, 2 and 3 (model 4); The true positive rate (TPR) and false positive rate (FPR). All based on 100 replications.

		SeqPIR			SeqSDR		
Model	Setting	$\|r\|$	TPR	FPR	$\|r\|$	TPR	FPR
1	$\rho = 0.0, p_1 = 50$	$0.682_{[0.073]}$	0.700	0.055	$0.541_{[0.043]}$	0.965	0.0647
	$\rho = 0.0, p_1 = 100$	$0.694_{[0.077]}$	0.864	0.061			
	$\rho = 0.0, p_1 = 200$	$0.653_{[0.072]}$	1.000	0.057			
	$\rho = 0.5, p_1 = 50$	$0.745_{[0.079]}$	0.800	0.051	$0.423_{[0.077]}$	0.832	0.094
	$\rho = 0.5, p_1 = 100$	$0.856_{[0.069]}$	0.900	0.045			
	$\rho = 0.5, p_1 = 200$	$0.801_{[0.023]}$	1.000	0.055			
	$\rho = 0.9, p_1 = 50$	$0.850_{[0.140]}$	1.000	0.055	$0.436_{[0.112]}$	0.921	0.197
	$\rho = 0.9, p_1 = 100$	$0.938_{[0.143]}$	1.000	0.071			
	$\rho = 0.9, p_1 = 200$	$0.911_{[0.111]}$	1.000	0.088			
2	$\rho = 0.0, p_1 = 50$	$0.593_{[0.161]}$	0.735	0.047	$0.745_{[0.128]}$	0.928	0.097
	$\rho = 0.0, p_1 = 100$	$0.594_{[0.153]}$	0.930	0.071			
	$\rho = 0.0, p_1 = 200$	$0.578_{[0.159]}$	1.000	0.065			
	$\rho = 0.5, p_1 = 50$	$0.702_{[0.191]}$	0.900	0.045	$0.421_{[0.122]}$	0.947	0.086
	$\rho = 0.5, p_1 = 100$	$0.800_{[0.150]}$	0.850	0.031			
	$\rho = 0.5, p_1 = 200$	$0.712_{[0.163]}$	1.000	0.065			
	$\rho = 0.9, p_1 = 50$	$0.812_{[0.196]}$	1.000	0.073	$0.319_{[0.165]}$	0.941	0.359
	$\rho = 0.9, p_1 = 100$	$0.859_{[0.207]}$	1.000	0.071			
	$\rho = 0.9, p_1 = 200$	$0.806_{[0.196]}$	1.000	0.096			
3	$\rho = 0.0, p_1 = 50$	$0.867_{[0.032]}$	0.983	0.042	$0.851_{[0.109]}$	0.961	0.091
	$\rho = 0.0, p_1 = 100$	$0.905_{[0.038]}$	0.997	0.097			
	$\rho = 0.0, p_1 = 200$	$0.891_{[0.032]}$	1.000	0.123			
	$\rho = 0.5, p_1 = 50$	$0.907_{[0.038]}$	0.980	0.078	$0.869_{[0.099]}$	0.969	0.093
	$\rho = 0.5, p_1 = 100$	$0.934_{[0.028]}$	1.000	0.096			
	$\rho = 0.5, p_1 = 200$	$0.918_{[0.020]}$	1.000	0.104			
	$\rho = 0.9, p_1 = 50$	$0.916_{[0.045]}$	0.850	0.079	$0.718_{[0.094]}$	0.950	0.271
	$\rho = 0.9, p_1 = 100$	$0.939_{[0.033]}$	0.974	0.093			
	$\rho = 0.9, p_1 = 200$	$0.872_{[0.042]}$	1.000	0.105			
4	$\rho = 0.0, p_1 = 50$	$0.666_{[0.103]}$	0.750	0.078	$0.625_{[0.106]}$	0.942	0.174
	$\rho = 0.0, p_1 = 100$	$0.649_{[0.157]}$	1.000	0.091			
	$\rho = 0.0, p_1 = 200$	$0.621_{[0.155]}$	1.000	0.106			
	$\rho = 0.5, p_1 = 50$	$0.617_{[0.112]}$	1.000	0.103	$0.608_{[0.113]}$	0.981	0.102
	$\rho = 0.5, p_1 = 100$	$0.694_{[0.156]}$	1.000	0.109			
	$\rho = 0.5, p_1 = 200$	$0.672_{[0.177]}$	1.000	0.083			
	$\rho = 0.9, p_1 = 50$	$0.709_{[0.133]}$	1.000	0.089	$0.430_{[0.097]}$	0.977	0.204
	$\rho = 0.9, p_1 = 100$	$0.762_{[0.143]}$	1.000	0.029			
	$\rho = 0.9, p_1 = 200$	$0.702_{[0.099]}$	0.750	0.060			

Table 2: Mean and standard deviation (in []) of the absolute correlation ($|r|$), the true positive rate (TPR) and false positive rate (FPR), over 100 replications.

		SeqPIR			Ensemble		
Model	Setting	$\|r\|$	TPR	FPR	$\|r\|$	TPR	FPR
2	$\rho = 0.0, p_1 = 50$	$0.593_{[0.161]}$	0.735	0.047	$0.603_{[0.092]}$	0.796	0.097
	$\rho = 0.0, p_1 = 100$	$0.594_{[0.153]}$	0.930	0.071	$0.613_{[0.094]}$	0.973	0.056
	$\rho = 0.0, p_1 = 200$	$0.578_{[0.159]}$	1.000	0.065	$0.567_{[0.059]}$	1.000	0.023
	$\rho = 0.5, p_1 = 50$	$0.702_{[0.191]}$	0.900	0.045	$0.697_{[0.078]}$	0.959	0.035
	$\rho = 0.5, p_1 = 100$	$0.800_{[0.150]}$	0.850	0.031	$0.801_{[0.149]}$	0.968	0.079
	$\rho = 0.5, p_1 = 200$	$0.712_{[0.163]}$	1.000	0.065	$0.745_{[0.097]}$	1.000	0.049
	$\rho = 0.9, p_1 = 50$	$0.812_{[0.196]}$	1.000	0.073	$0.812_{[0.043]}$	1.000	0.023
	$\rho = 0.9, p_1 = 100$	$0.859_{[0.207]}$	1.000	0.071	$0.862_{[0.095]}$	1.000	0.098
	$\rho = 0.9, p_1 = 200$	$0.806_{[0.196]}$	1.000	0.096	$0.821_{[0.099]}$	1.000	0.131

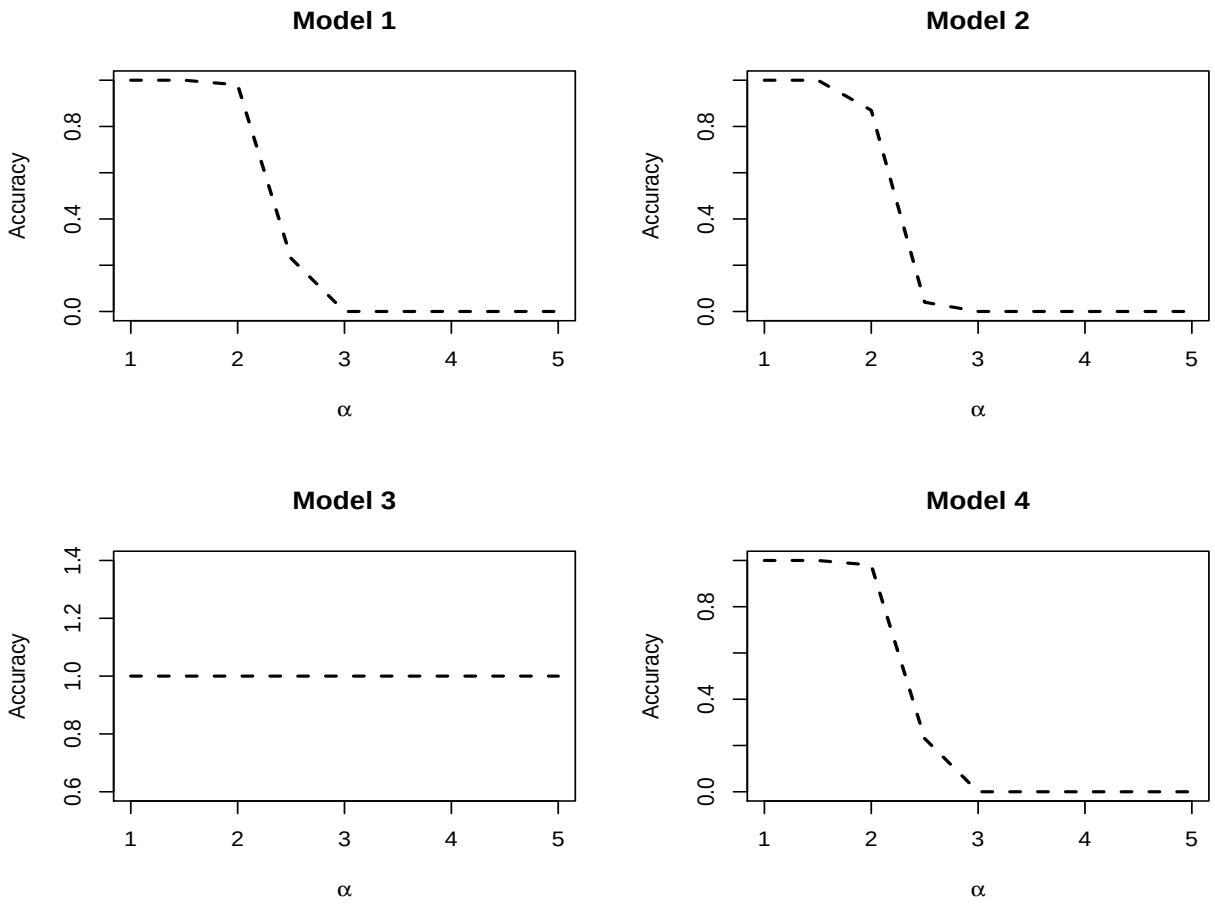

Figure 1: Estimation accuracy of the structural dimension (d) estimation method for a range of α values.

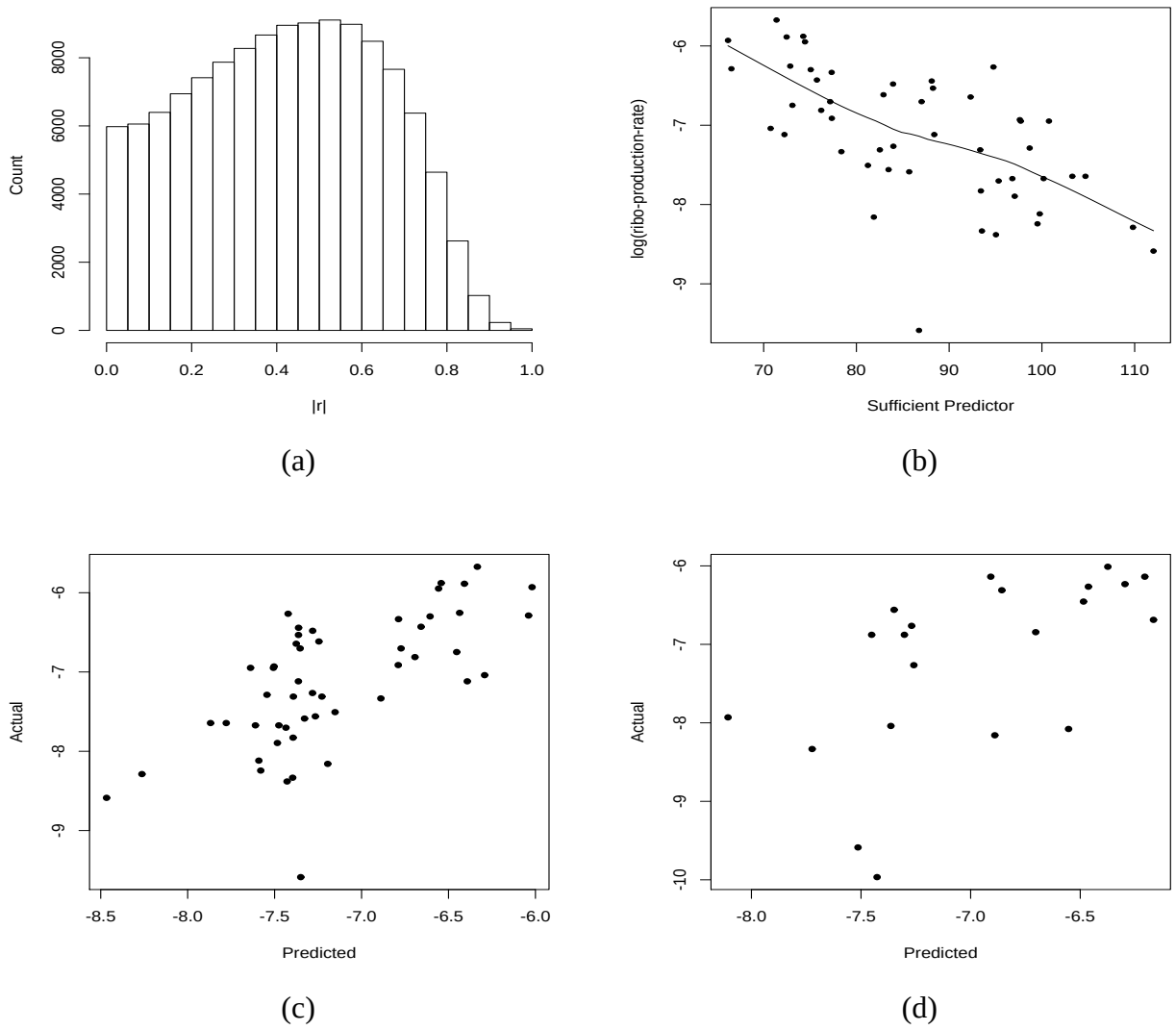

Figure 2: Panel (a) is a histogram of the pair-wise absolute correlations for the selected 500 genes; Panel (b) is the sufficient summary plot; Panel (c) is a scatterplot of the actual and predicted values for the 50 training samples; Panel (d) is a scatterplot of the actual and predicted values for the 21 test samples (both in log scale)