Computational Statistics and Data Analysis 81 (2015) 107–120

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A sequential test for variable selection in high dimensional complex data

Kofi P. Adragni *, Moumita Karmakar

Department of Mathematics & Statistics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, United States

ARTICLE INFO

Article history:
Received 16 April 2013
Received in revised form 28 January 2014
Accepted 24 July 2014
Available online 2 August 2014

Keywords:
Inverse regression
Dimension reduction
Principal components
Variable selection
Variable screening
High dimensionality
Data analysis

ABSTRACT

Given a high dimensional p-vector of continuous predictors X and a univariate response Y, principal fitted components (PFC) provide a sufficient reduction of X that retains all regression information about Y in X while reducing the dimensionality. The reduction is a set of linear combinations of all the p predictors, where with the use of a flexible set of basis functions, predictors related to Y via complex, nonlinear relationship can be detected. In the presence of possibly large number of irrelevant predictors, the accuracy of the sufficient reduction is hindered. The proposed method adapts a sequential test to the PFC to obtain a "pruned" sufficient reduction that shed off the irrelevant predictors. The sequential test is based on the likelihood ratio which expression is derived under different covariance structures of $X|Y$. The resulting reduction has an improved accuracy and also allows the identification of the relevant variables.

1. Introduction

Consider the Big-Mac dataset (Enz, 1991), a simple dataset that gives average values in 1991 of several economic indicators for 45 world cities. It has nine continuous predictors and a continuous outcome variable. The outcome is the minimum labor to buy a Big Mac and fries in US dollars. A regression fitting to the raw data without any transformation of the response or predictors yields a multiple R^2, the square of the correlation between the observed and the fitted response to be 0.46. After a graphical exploration and the appropriate transformation of the variables, we obtained $R^2 = 0.87$. The reason of this drastic improvement is that the relationships between the response and the predictors, initially nonlinear (Fig. 1), were transformed into linear through a model fitting procedure guided by diagnostics (see Cook and Weisberg, 1982, Section 1.2). With nine predictors, this procedure is easily doable. However, when p is large, say 50 or more, a regression modeling using this iterative procedure is rather a daunting task, tedious and imponderable. Ubiquitously, a forward linear model is considered, and the relationship between individual predictors and the response is often unexplored because of the high dimensionality of the predictors. Diagnostic methods are seldom used for model checking. Using an ill-fitting model to solve a variable selection problem can result in reduced performance.

Most variable selection methods are constructed around forward linear regression models. Because the ordinary least squares estimation does not yield satisfactory results when p is large, it is often assumed that a large portion of these p predictors is irrelevant in explaining the response Y. The corresponding coefficients of these predictors in a linear regression model are shrunk or even set to zero. This brings the concept of sparsity into regression modeling with two induced consequences: parsimony of the model and accuracy in prediction. A flurry of research on algorithms and theory for variable

* Corresponding author.
E-mail addresses: kofi@umbc.edu, kofi.adragni@gmail.com (K.P. Adragni).

http://dx.doi.org/10.1016/j.csda.2014.07.016
0167-9473/

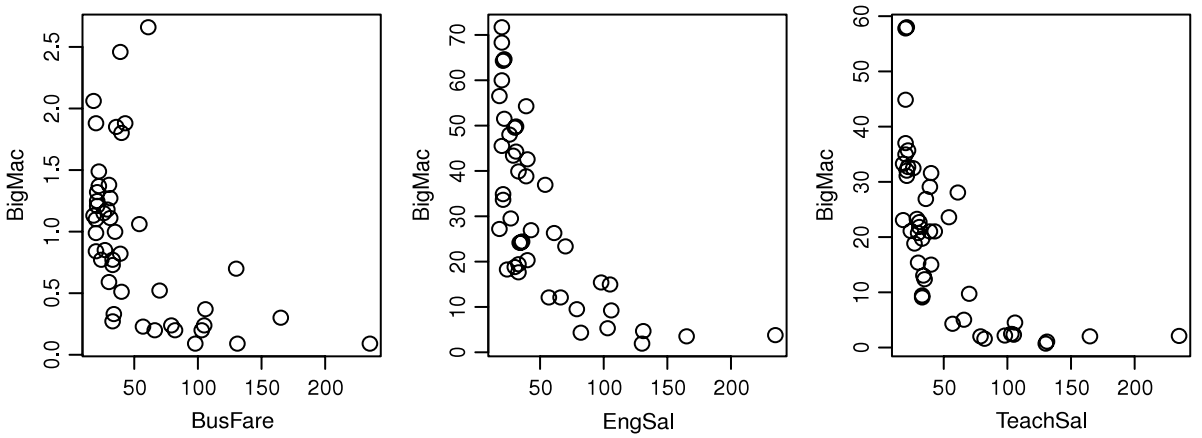


Fig. 1. BigMac partial scatter-matrix plot.

selection involving sparsity constraints have been observed in recent years. These methods include the soft thresholding (Donoho, 1995), the nonnegative garotte (Breiman, 1995), lasso (Tibshirani, 1996), the smoothly clipped absolute deviation penalty (SCAD; Fan and Li, 2001), elastic net (Zou and Hastie, 2005), and Dantzig selector (Candès and Tao, 2007) among many others. These methods work exceptionally well when the model is accurate. However they do not perform adequately when the predictors and the response have an arbitrary non-linear relationship.

A recent methodology proposed by Cook (2007) brings significant openings to address the shortcomings of linear models in capturing information about high dimensional predictors non-linearly related to the response. Cook (2007) proposed the concept of sufficient dimension reduction in regression and set up a new paradigm of dimension reduction through a likelihood-based approach called principal fitted components (PFC). A reduction $R : \mathbb{R}^p \to \mathbb{R}^d$, $d \leq p$, was defined to be sufficient if it satisfies one of the following three statements: (i) $Y|X \sim Y|R(X)$, (ii) $X|(Y, R(X)) \sim X|R(X)$, and (iii) $X \perp\!\!\!\perp Y|R(X)$. The symbol $\perp\!\!\!\perp$ stands for statistical independence, and $U \sim V$ stands for U and V having identical distribution. Statement (i) holds in a forward regression while statement (ii) holds in an inverse regression setup. Under a joint distribution of (Y, X) the three statements are equivalent.

Principal fitted components are a class of inverse regression models that yield a sufficient reduction of the predictors. Let X_y denote the random vector $X|(Y = y)$ and assume that there is a vector-valued function $\nu(Y) \in \mathbb{R}^d$, with $d \ll \min(n, p)$ and $\mathrm{E}[\nu(Y)] = 0$, so that X_y can be represented by the model $X_y \sim N(\mu + \Gamma\nu(y), \Delta)$. The term $\Gamma \in \mathbb{R}^{p\times d}$ is a semi-orthogonal matrix, and $\mu = \mathrm{E}(X)$. The covariance Δ is assumed to be independent of Y. Under this model the translated conditional means $\mathrm{E}(X_y) - \mu$ fall in the d-dimensional subspace $\mathrm{span}(\Gamma)$, and thus Γ captures the dependency between X and Y. Once the response is observed, the term $\nu(y)$ which is unobserved can be approximated using a flexible set of basis functions as $\nu(y) \approx \beta \mathrm{f}(y)$. The subsequent model

$$X_y = \mu + \Gamma\beta\mathrm{f}(y) + \Delta^{1/2}\varepsilon \tag{1}$$

is called a PFC model where ε is assumed to be normally distributed with mean 0 and variance I_p. Under this model, Cook (2007) showed that $\Gamma^T\Delta^{-1}X$ is a sufficient reduction of X. The choice of the basis function allows to capture predictors that are linearly and nonlinearly related to the response. The maximum likelihood estimators of the parameters in the model have been obtained (Cook, 2007; Cook and Forzani, 2008).

In high dimensional settings, irrelevant predictors, which often abound, can hinder the accuracy of the estimated sufficient reduction. Our goal is to obtain a "pruned" estimator of the sufficient reduction, which not only helps achieve accuracy, but also allows the identification of the relevant variables. By "pruning", we mean removing inactive predictors that do not contain any regression information about the response. This is often called a sparse estimator.

An estimation of the sparse reduction kernel $\Delta^{-1}\Gamma$ has been proposed by Li (2007) who established a framework to obtain the sparse sufficient reduction using a regression-type formulation with the lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005) penalties. Chen et al. (2010) proposed the coordinate independent sparse sufficient dimension reduction that shrinks row elements of $\Delta^{-1}\Gamma$ while preserving the orthogonality constraint of Γ. Both methodologies are apt when $n \gg p$. We herein construct a sequential likelihood ratio test that is reminiscent of the idea of testing predictor contribution in sufficient dimension reduction of Cook (2004). It helps obtain the sparse reduction under structures of Δ that allow $p > n$. We show the performance of the procedure through simulations.

2. A sequential test for sparse PFC

We assume that the p-vector predictor X can be partitioned as $(X_1^T, X_2^T)^T$, with $X_2 \in \mathbb{R}^{p_2}$, and let $(\Gamma_1^T, \Gamma_2^T)^T$, $\Delta = (\Delta_{ij})_{i,j=1,2}$ and $\Delta^{-1} = (\Delta^{ij})_{i,j=1,2}$ be the corresponding partitions of Γ, Δ and Δ^{-1} following the partition of X. Under

model (1), the sufficient reduction can be written as

$$\Gamma^T \Delta^{-1} X = (\Gamma_1^T \Delta^{11} + \Gamma_2^T \Delta^{21}) X_1 + (\Gamma_1^T \Delta^{12} + \Gamma_2^T \Delta^{22}) X_2. \tag{2}$$

Let us suppose that X_2 represents the set of predictors with no regression information about Y in the sense that $X_2|Y$ has the same distribution as X_2. Consequently, we have $\Gamma_2 = 0$. However, the sufficient reduction which becomes $\Gamma_1^T \Delta^{11} X_1 + \Gamma_1^T \Delta^{12} X_2$, still retains X_2. In order to obtain a sufficient reduction without the irrelevant X_2, we will assume that $\Delta^{12} = 0$. We summarize the result in the following proposition.

Proposition 2.1. *Consider PFC model (1) and the aforementioned partitioning of X, Γ and Δ. Then we have the following:*

(i) $X_2 \perp\!\!\!\perp Y \Leftrightarrow \Gamma_2 = 0$;
(ii) $X_1 \perp\!\!\!\perp X_2|Y \Leftrightarrow \Delta_{12} = 0$;
(iii) *If $X_2 \perp\!\!\!\perp Y$ and $X_1 \perp\!\!\!\perp X_2|Y$ then $\Gamma_1^T \Delta_{11}^{-1} X_1$ is a sufficient reduction of X.*

It is possible that the statement $X_2 \perp\!\!\!\perp Y$ holds and yet $\Delta_{12} \neq 0$. This implies that X_1 and X_2 are related through the error. One possibility of this occurrence is that both X_1 and X_2 are related to a latent variable that is unobserved. The assumption $\Delta_{12} = 0$ forces to ignore this latent dependency, and it may be reasonable in practice. However the assumption will be relaxed in Section 2.2 under an extended structure of Δ.

For now, we will assume that $\Delta_{12} = 0$. We aim to effectively separate the predictors into the active X_1 and the inactive X_2. Following Proposition 2.1, once a set X_2 is suspected to be unrelated to Y, we will test the hypothesis

$$H_0 : \Gamma_2 = 0 \text{ against } H_1 : \Gamma_2 \neq 0 \tag{3}$$

at a specified significance level α. Failing to reject H_0 implies that X_2 contains no relevant predictors.

We proceed with a likelihood ratio test. Let $\mathcal{L}_{H_0}$ and $\mathcal{L}_{H_a}$ be the log-likelihood evaluated at the maximum likelihood estimates of the parameters in the appropriate PFC model under the null and the alternative hypotheses respectively. Then under H_0, $\Lambda = 2(\mathcal{L}_{H_a} - \mathcal{L}_{H_0})$ follows approximately a chi-square distribution with dp_2 degrees of freedom. A number of structures of Δ have been discussed in the literature, including the *isotropic*, the *anisotropic*, and the *structured*. For predictors conditionally independent, we have $\Delta = \text{diag}(\delta_1^2, \ldots, \delta_p^2)$. This structure is referred to as *anisotropic* when $\delta_i^2 \neq \delta_j^2$ for some $i \neq j$, that is, the predictors are on different measurement scales. When $\delta_i^2 = \delta^2$ for all i, the model is referred to as *isotropic*: predictors are on same measurement scale. The structured model is rather adaptive and allows group conditional independence of the predictors.

The sequential LRT procedure relies essentially on the identification of candidates X_1 and X_2. We proceed using the strength of the marginal dependency between individual predictors and the response. Specifically, we start with a diagonal Δ. To show this, we rewrite Eq. (1) as

$$\Delta^{-1/2} X_y = \Delta^{-1/2} \mu + \Delta^{-1/2} \Gamma \beta \text{f}(y) + \varepsilon$$

where $\varepsilon \sim N(0, I)$. This transformation is to rescale the predictors to be on the same scale. Consequently, the rows of $\widetilde{\Gamma} = \Delta^{-1/2}\Gamma$ can be compared. Let $\tilde{\gamma}_j$ be the jth row vector of $\widetilde{\Gamma}$. The strength of the dependency between individual predictors and the response can be evaluated by $\|\tilde{\gamma}\|$. Larger values correspond to predictors with stronger relationships with the response. Predictors are then sorted in the decreasing order of the $\|\tilde{\gamma}_j\|$s. This has been proposed by Adragni and Cook (2008) for variable screening, and it was shown to subsume the sure independence screening of Fan and Lv (2008). The proposed algorithm is as follows.

Algorithm.

1. Fit an anisotropic PFC model, and estimate the dimension $\hat{d}$ of the reduction.
2. Form $\hat{\Delta}^{-1/2}\widehat{\Gamma} = (\hat{\zeta}_1^T, \hat{\zeta}_2^T, \ldots, \hat{\zeta}_p^T)^T$.
3. Sort the predictors in the decreasing order $x_{(1)}, \ldots, x_{(p)}$ according to their corresponding $\|\hat{\zeta}_i\|_2$.
4. Do sequentially from $i = \hat{d}$ until H_0 is not rejected
   a. Set $X_1 = (x_{(1)}, \ldots, x_{(i)})^T$ and $X_2 = (x_{(i+1)}, \ldots, x_{(p)})^T$.
   b. Test the hypothesis (3) at a level of significance α.
   c. If H_0 is rejected, then $i = i + 1$ and return to step a.; otherwise stop.

Fitting a PFC model requires user-provided basis function f(y). In general, a third degree polynomial or a piecewise constant are suggested. Once a basis function is chosen, the parameters can be estimated. The dimension d of Γ is estimated using either a likelihood ratio test or information criteria (AIC, BIC) as suggested by Cook and Forzani (2008).

To fix the idea, let us demonstrate on the Big-Mac dataset (Enz, 1991). It contains nine continuous predictors and a continuous outcome variable. The outcome is the minimum labor to buy a Big Mac and fries in US dollars. The nine predictor variables are X_1 the minimum labor to buy one kilogram of bread, X_2 the lowest cost of 10 km public transit, X_3 the electrical engineer's annual salary, X_4 the tax rate paid by electrical engineer, X_5 the annual cost of 19 services, X_6 the primary teacher's salary, X_7 the tax rate paid by primary teacher, X_8 the average days of vacation per year, and X_9 the average hours worked per year.

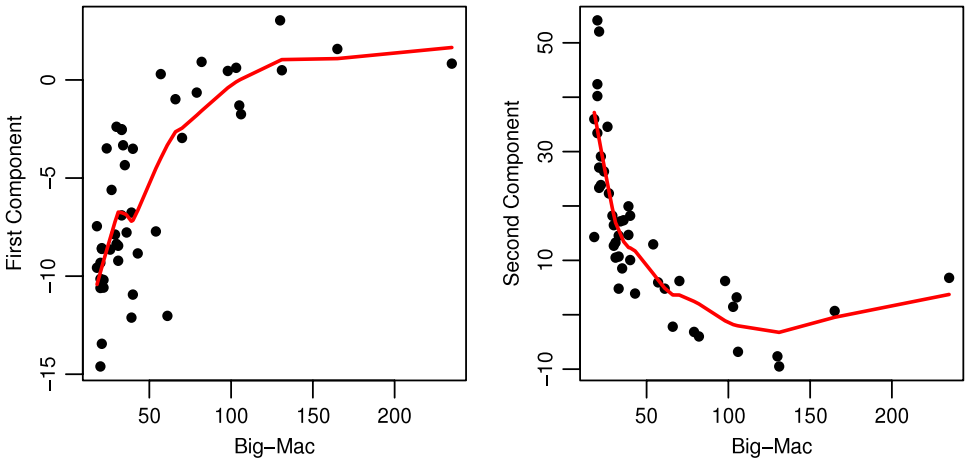


Fig. 2. Reductions of the predictors against the response.

We start with step (1) of the algorithm by fitting the anisotropic PFC model and estimating the dimension d. This first step was carried out using the R package `1dr` of Adragni and Raim (2014). The model was fitted using a piecewise constant basis with five slices. The guidance on the choice of the number of slices follows that of sliced inverse regression of Li (1991): the number of slices should be large enough to allow the estimation of d, and yield enough observations per slice to estimate the intra-slice parameters.

We proceeded to the estimation of the dimension d, using a likelihood ratio test. The hypothesis $H_0 : d = i$ against $H_a : d > i$ was tested sequentially for $i = 0, 1, 2, \ldots$. The null hypothesis was rejected for $d = 0$ and for $d = 1$, but was not rejected for $d = 2$ against $d > 2$. Thus, the estimated dimension of the reduction was $\hat{d} = 2$. Fig. 2 gives the plot of the two components of the reductions against the response. These two components retain all regression information about the response that is contained in the initial nine predictors. These two components can now be used to model the mean function $E(Y|X) = E(Y|\hat{\xi}^T X)$. The components were obtained with the two column-vectors of $\hat{\Delta}^{-1}\widehat{\Gamma}$ as (0.0132, −0.9896, −0.0527, −0.0280, −0.0072, −0.1180, −0.0543, 0.0070, 0.0013) and (−0.0236, −0.0916, 0.2842, −0.1352, 0.0026, 0.7450, −0.2296, −0.5331, 0.0063). With the estimated $\widehat{\Gamma}$ and $\hat{\Delta}$, steps (2) and (3) are carried out. The likelihood ratio test in step (3.b) depends on the structure to be used. In the following sections, we provide the expression of this test under the different structures of PFC model. We adopt the following notations throughout. Given n replications of $\{\mathbf{X}_i, y_i\}_{i=1}^n$, we denote by $\mathbb{X}$ the $n \times p$ data-matrix with its ith observation given by $(\mathbf{X}_i - \bar{\mathbf{X}})^T$, and $F \in \mathbb{R}^{n\times r}$ to be the data-matrix of the basis function, formed with its ith row being $\mathrm{f}(y_i)$. We let $\mathrm{P}_F = F(F^TF)^{-1}F^T$ be the projection operator that projects on the space spanned by the columns of F, where the basis function is chosen such that F^TF is invertible, and denote $\widehat{\Sigma}_{\text{fit}} = \mathbb{X}^T\mathrm{P}_F\mathbb{X}/n$, to be a $p \times p$ covariance matrix of rank at most $r < \min(p, n)$. Furthermore, we let $\widehat{\Sigma}_{\text{res}} = \mathbb{X}^T(\mathrm{I}_p - \mathrm{P}_F)\mathbb{X}/n$. Using the partition of X, we set $\mathbb{X}_j$ to be the data-matrix of X_j, and let $\widehat{\Sigma}_{jj} = \mathbb{X}_j^T\mathbb{X}_j/n$, $\widehat{\Sigma}_{jj,\text{fit}} = \mathbb{X}_j^T\mathrm{P}_F\mathbb{X}_j$, $\widehat{\Sigma}_{jj,\text{res}} = \mathbb{X}_j^T(\mathrm{I}_p - \mathrm{P}_F)\mathbb{X}_j$, $j = 1, 2$.

2.1. Likelihood ratio test under structured Δ

We are assuming explicitly that, conditionally on the response, the active or relevant predictors X_1 are independent of the inactive ones. Thus, Δ is structured, being block-diagonal with $\Delta_{11} > 0$, $\Delta_{22} > 0$, and $\Delta_{12} = 0$. We adopted and implemented the maximum likelihood estimation procedure proposed by Cook and Forzani (2008). A linear structure of Δ was used with $\Delta = \sum_{i=1}^m \delta_i \mathrm{M}_i$, where $m \le p(p+1)/2$ and $\mathrm{M}_1, \ldots, \mathrm{M}_m$ are known real symmetric $p \times p$ linearly independent matrices and the elements of $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^T$ are functionally independent. The details of the algorithm for the maximum likelihood estimation is in Cook and Forzani (2008, Section 8). The following proposition gives the expression of the likelihood ratio test, with its proof in the Appendix.

Proposition 2.2. *Assuming that $\Delta_{12} = 0$, let $\hat{\Delta} = (\hat{\Delta}_{ii})_{i=1,2}$ be the MLE of Δ under H_a. Then the expression of LRT is given by*

$$\Lambda = n\left(p + \log|\widehat{\Sigma}_{22}| + \log|\widehat{\Sigma}_{11,\text{res}}|\right) + n\left[\sum_{i=d+1}^{p_1} \log[1 + \hat{\lambda}_i(\widehat{\Sigma}_{11,\text{res}}^{-1}\widehat{\Sigma}_{11,\text{fit}})] - \sum_{i=d+1}^{p} \lambda_i(\hat{\Delta}^{-1}\widehat{\Sigma}_{\text{fit}})\right]$$
$$- n\left(\log|\hat{\Delta}_{11}| + \log|\hat{\Delta}_{22}| + \mathrm{Tr}\{\hat{\Delta}_{11}^{-1}\widehat{\Sigma}_{11,\text{res}}\} + \mathrm{Tr}\{\hat{\Delta}_{22}^{-1}\widehat{\Sigma}_{22,\text{res}}\}\right). \quad (4)$$

It should be noted that the isotropic and the anisotropic structures are special cases of the structured Δ. The isotropic structure can be obtained by setting $m = 1$ and $\mathrm{M}_1 = \mathrm{I}_p$; the anisotropic structure can be obtained by setting $m = p$ and $\mathrm{M}_i =$

$\mathbf{e}_i\mathbf{e}_i^T$, $i = 1, \ldots, p$, where $\mathbf{e}_i \in \mathbb{R}^p$ with all 0 entries except 1 at the ith position. A simpler form of the LRT is obtained under the isotropic model, that is provided in the following proposition.

Proposition 2.3. *Let $\hat{\lambda}_i^{11,\text{fit}}$, $\hat{\lambda}_i^{\text{fit}}$, and $\hat{\lambda}_i$ be the ith largest eigenvalue of $\widehat{\Sigma}_{11,\text{fit}}$, $\widehat{\Sigma}_{\text{fit}}$, and $\widehat{\Sigma}$ respectively. The LRT statistic under the isotropic model is given by*

$$\Lambda = np \log \left[1 + \frac{\sum_{i=1}^{d} (\hat{\lambda}_i^{\text{fit}} - \hat{\lambda}_i^{11,\text{fit}})}{\sum_{i=1}^{p} \hat{\lambda}_i - \sum_{i=1}^{d} \hat{\lambda}_i^{\text{fit}}} \right]. \tag{5}$$

The proofs of these propositions are in the Appendix. It should be noted that there is no dimensionality issue with the isotropic and anisotropic models when $p \gg n$; the estimation of the parameters can be carried for any decent sample size. However, the estimation of the parameters requires $n \gg \max(p_1, p_2)$ when Δ_{11} and Δ_{22} are unstructured.

The choice of the structure of Δ could be determined by a test statistic in data-rich problem where $n \gg p$. For a fixed dimension d, a model with structured Δ can be compared to another with unstructured covariance following Cook and Forzani (2008, Section 8). When n is not large enough to estimate an unstructured Δ, an anisotropic model can be fitted.

Returning to the Big-Mac dataset, the sample size of 45 is not large enough to allow an accurate estimation of the 9×9 unstructured covariance Δ. Hence, we carried the algorithm using the anisotropic structure. The sequential test selected all the variables except X_9, the average hours worked per year. With the remaining variables, the two components of the estimated sufficient reduction are obtained and plotted against the response in Fig. 3. A reduction in the spread of the observations around the loess curve seems apparent on the plot with the second component compared to the original plot in Fig. 2.

The assumption of conditional independence may not always hold in general, and it might be of interest to evaluate the performance of the method when that assumption is violated. We will attempt an evaluation on a dataset in Section 3.1.

2.2. Extended Δ

So far, we have assumed that $X_1 \perp\!\!\!\perp X_2|Y$, which is equivalent to $\Delta_{12} = 0$. In the following, we will consider a structure that relaxes that assumption. Let Γ_0 be the orthogonal completion of Γ, so that (Γ, Γ_0) is a $p \times p$ orthogonal matrix. Let $P_\Gamma = \Gamma\Gamma^T$ be the projection operator on span(Γ), the subspace spanned by Γ, and let $Q_\Gamma = I - P_\Gamma = \Gamma_0\Gamma_0^T$ where I is a $p \times p$ identity matrix. Writing $\Delta = (P_\Gamma, Q_\Gamma)^T \Delta (P_\Gamma, Q_\Gamma)^T$ yields

$$\Delta = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T, \tag{6}$$

where $\Omega = \Gamma^T \Delta \Gamma \in \mathbb{R}^{d \times d}$ and $\Omega_0 = \Gamma_0^T \Delta \Gamma_0 \in \mathbb{R}^{(p-d)\times(p-d)}$. This is referred to as an extended structure of Δ (Cook, 2007). It assumes that $\mathcal{S}_\Gamma$, the subspace spanned by the columns of Γ, is a reducing subspace of Δ in the sense that $\Delta\mathcal{S}_\Gamma = \mathcal{S}_\Gamma$.

The mean function $E(X_y)$ in model (1) was obtained by approximating the true function $v(Y)$ by $\beta \mathrm{f}_Y$. This unknown function could be modeled as $v_y \sim N(\beta \mathrm{f}_y, \Psi)$, where $\Psi > 0$ is a $d \times d$ matrix. Consequently, using expression (6), model (1) can be rewritten as

$$X_y = \mu + \Gamma\beta \mathrm{f}_y + \Gamma \Phi^{1/2} \epsilon + \Gamma_0 \Omega_0^{1/2} \epsilon_0, \tag{7}$$

with $\Phi = \Omega + \Psi$ and $(\epsilon^T, \epsilon_0^T)^T \sim N(0, I)$. This model (7) also has an extended structure as described in Cook (2007). Under this structure, the sufficient reduction is $\Gamma^T X$. This structure allows some correlations among the conditional predictors but stops short of permitting an unstructured $\Delta > 0$. The following proposition provides the expression of likelihood ratio test under the extended structure. Its proof is in the Appendix.

Proposition 2.4. *Let $\widehat{\Gamma}$ and $\widehat{\Gamma}_1$ be semi-orthogonal representative bases of the MLEs of* span(Γ) *and* span(Γ_1) *under H_a and H_0 respectively, and let $\widehat{\Gamma}_0$ be the orthogonal completion of $\widehat{\Gamma}$. Let $\widehat{\Gamma}_{01}$ be the orthogonal completion of $(\widehat{\Gamma}_1^T, 0, \ldots, 0)^T \in \mathbb{R}^{p\times d}$. The expression of the likelihood ratio test is then*

$$\Lambda = n \left[\log |\widehat{\Gamma}_1^T \widehat{\Sigma}_{11,\text{res}} \widehat{\Gamma}_1| + \log |\widehat{\Gamma}_{01}^T \widehat{\Sigma} \widehat{\Gamma}_{01}| - \log |\widehat{\Gamma}^T \widehat{\Sigma}_{\text{res}} \widehat{\Gamma}| - \log |\widehat{\Gamma}_0^T \widehat{\Sigma} \widehat{\Gamma}_0| \right]. \tag{8}$$

The estimation of the parameters that led to expression (8) assumes that the columns of Γ_1 and Γ_{01} do not fall into the null eigenspace of $\widehat{\Sigma}_{11,\text{res}}$ and $\widehat{\Sigma}$, respectively, and similarly for $\widehat{\Gamma}$ and $\widehat{\Gamma}_0$ with respect to $\widehat{\Sigma}_{\text{res}}$ and $\widehat{\Sigma}$.

The estimation of Λ requires $n \gg p-d$ to allow the estimation of the $(p-d) \times (p-d)$ parameter Ω_0. Other substructures of the extended model could be considered. For example, Ω_0 can be assumed to have a diagonal structure, while Φ is unstructured. The performance and efficiency of the reduction under this extended model are yet to be addressed, as it requires more evolved parameter estimations over Grassmann manifolds.

We close the discussion with a note on a PFC model with a general unstructured covariance when n is too small for the estimation of Δ. A sparse estimation of both Δ and Γ can be sought. The positive-definite l_1-penalized estimation of large covariance matrices of Xue et al. (2012) could be considered in a general alternating algorithm to estimate Δ and Γ. This should be studied by its own right and will be worth pursuing.

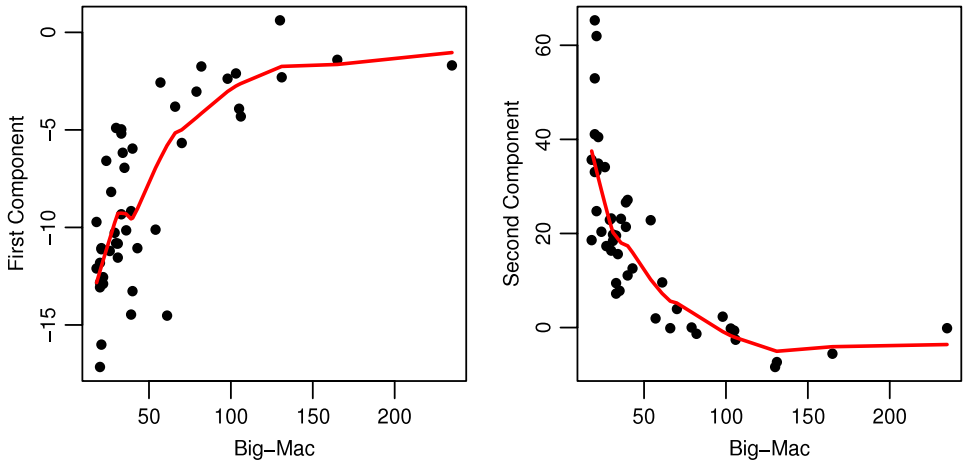


Fig. 3. Plots of the reductions against the response of the Big-Mac data after the sequential LRT.

2.3. Other sparse sufficient dimension reduction methods

A number of sufficient dimension reduction methods have been proposed in the statistics literature since the seminal paper on sliced inverse regression (SIR; Li, 1991). Among these methods are the principal Hessian directions (PHD; Li, 1992), the inverse regression estimation (IRE; Cook and Ni, 2005), and the directional regression (DR; Li and Wang, 2007). Some of these methods are distribution-free, while others like PFC are likelihood-based. Nearly all sufficient dimension reduction methods seek a set of linear combinations $\zeta^T X$ of the predictors that retains all the regression information of Y on X. Sparse sufficient dimension reduction of Li (2007) is a generic formulation that can be adapted to any of the aforementioned methodologies.

Li's formulation uses a lasso or an elastic-net type of tuning parameters to shrink coefficients of inactive predictors to zero. An information criterion helps determine the optimal values of the tuning parameters. The information criterion is a function of the inverse of the covariance matrix of the predictors Σ^{-1}, and its computation brings a numerical challenge when p is large compared to n. To avoid this issue in our adaptation of Li's formulation to PFC, we used a cross-validation approach with the prediction approach of Adragni and Cook (2009). We will later refer to this method as the PFC-enet.

Sliced inverse regression (Li, 1991) has been well studied, and there is a number of proposed methods for the estimation of its sparse sufficient reduction (Zhong et al., 2005; Li and Yin, 2008; Wang and Zhu, 2013, among others). A connection of SIR with PFC was established by Cook (2007, Section 6.3) and further elaborated in Cook and Forzani (2008, Section 4). Cook (2007) showed that when the response Y is categorical, then SIR and PFC estimate the same minimal sufficient reduction subspace. When Y is continuous, SIR discretizes the response through a slicing procedure and can leave intra slice information behind. On the other hand, the flexible basis functions of PFC potentially help avoid such loss of information.

We propose another method to estimate the sparse PFC. It is a p-value guided thresholding that sets to zero, rows of Γ corresponding to predictors that do not have a marginal relationship with Y through f_y. To describe it, let $\boldsymbol{\varphi} = \Gamma\beta$ and assume a diagonal structure for Δ. Then model (1) can be expressed as p independent simple linear regressions

$$X_i = \mu_i + \varphi_i \mathrm{f}(y) + \delta_i \varepsilon_i, \quad i = 1, \ldots, p, \tag{9}$$

where φ_i is the ith row vector of $\boldsymbol{\varphi}$. If X_i and Y are dependent, then φ_i should be nonzero. The parameter φ_i can be tested for equality to zero at a specified level of significance α. Given the data, let F_i be the test statistic for test $H_0 : \varphi_i = 0$ against $H_1 : \varphi_i \neq 0$. The statistic F_i follows an F distribution with $(r, n-r-1)$ degrees of freedom. Let $\pi_i = P(F \geq F_i | \varphi_i = 0)$ be the p-value resulting from the test, let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)^T$, and let $\mathbf{1}_p$ be the p-vector of ones. The p-value guided thresholding estimator $\widehat{\Gamma}_\alpha$ is obtained as

$$\widehat{\Gamma}_\alpha = J(\boldsymbol{\pi} \leq \alpha \mathbf{1}_p)\widehat{\Gamma}. \tag{10}$$

The inequality is element-wise, and $J(\cdot)$ represents the indicator function. With this procedure, $\pi_i \leq \alpha$ is an evidence against H_0 and X_i is assumed active; the corresponding row in $\widehat{\Gamma}_\alpha$ is nonzero. These rows form the estimate of Γ_1. One advantage of this procedure is that all predictors with corresponding p-values $\pi_i > \alpha$ are discarded, and this may significantly prune the predictors of the irrelevant ones. This method will be referred to as PFC-pv. We will later compare the proposed sequential LRT method to PFC-enet and to PFC-pv through numerical simulations.

3. Numerical studies

We illustrate the performance of the sequential likelihood ratio test for sparse sufficient reduction estimation and variable selection with PFC on two datasets and also through a simulation study. With the first dataset, the performance of

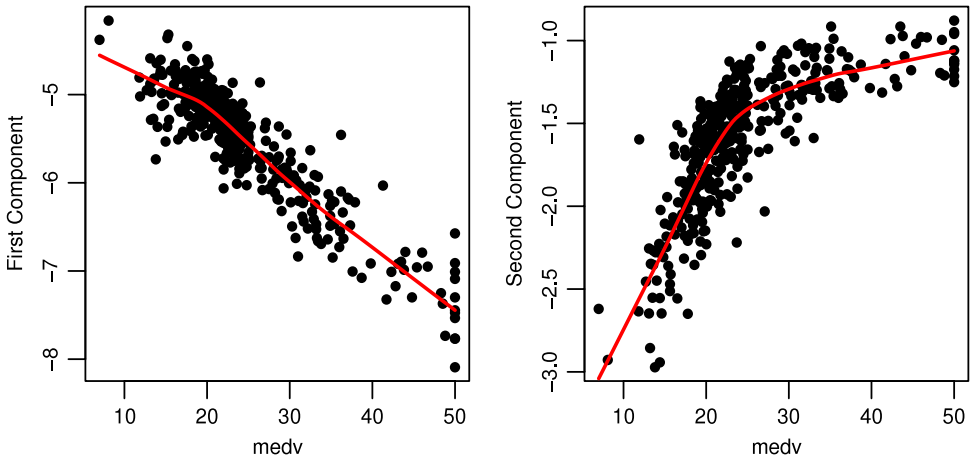


Fig. 4. Plots of the reduction of the predictors against the response of the Boston data assuming conditional dependence of the predictors.

the method is evaluated when the assumption of the conditional independence is violated. The second dataset is a case where the sufficient reduction methodology leads to fitting a linear regression model and its related shrinkage methodologies for variable selection.

3.1. The Boston housing dataset

The Boston housing dataset has been widely used in the literature and is available on the web site http://lib.stat.cmu.edu/datasets/boston_corrected.txt. It has 506 observations with 13 predictors. These predictors are the following: per capita crime rate by town (*crim*), proportion of residential land zoned for lots over 25,000 sq.ft (zn), proportion of non-retail business acres per town (*indus*), Charles River dummy variable (*chas*), nitric oxides concentration (*nox*), average number of rooms per dwelling (*rm*), proportion of owner-occupied units built prior to 1940 (*age*), weighted distances to five Boston employment centers (*dis*), index of accessibility to radial highways (*rad*), full-value property-tax rate (*tax*), pupil–teacher ratio by town (*ptratio*), proportion of blacks by town (black), and percentage of lower status of the population (*lstat*). The response (*medv*) is the median value of owner-occupied homes in each of the 506 census tracts in the Boston Standard Metropolitan Statistical Areas in $1000s. We removed *chas* and *rad* as they are categorical with two and nine levels respectively. Previous studies of the data (Li, 1991; Chen et al., 2010) suggested removing observations corresponding to crime rate greater than 3.2, as a few predictors remain constant except for 3 observations in this case. We thus used 374 observations.

In an initial investigation, we sought to determine the appropriate structure of Δ. We fit PFC models to the data assuming first that the dimension d was five, using a piecewise constant basis function with eight slices. The choice of the dimension $d = 5$ was arbitrary since the true dimension is yet to be estimated. We believed that the true dimension would be less than five, thus a dimension larger than expected helps avoid losing relevant information. A hypothesis test of an isotropic against an unstructured model rejected the null hypothesis. Similarly, an anisotropic structure was rejected against an unstructured model. The data contradicted the assumption of conditional independence of the predictors. The correlation among the predictors, conditionally on the response, ranged from 0.02 to 0.96. The dimension of the sufficient reduction was then estimated using a sequential likelihood ratio test that rejected $d = 0$ against $d \geq 1$, and also rejected $d = 1$ against $d \geq 2$, but failed to reject $d = 2$ against $d \geq 3$. All the tests were carried at a 5% significance level. The final model was obtained and the reduction was plotted as shown in Fig. 4. The solid line is a nonparametric loess curve.

Although Δ should be unstructured with the Boston data, we proceeded nevertheless fitting a PFC with an anisotropic structure. The dimension of the reduction was again obtained as $\hat{d} = 2$ using LRT. The plots of the two components of the reduction plotted against the response are in Fig. 5. This result, and other unreported simulations suggest that we can expect a decent performance of the methodology when conditional independence is used while the true Δ is unstructured. All this analysis was carried using the R package ldr of Adragni and Raim (2014) that is available on CRAN at http://cran.r-project.org/web/packages/ldr/. We applied the sequential likelihood ratio test with $\hat{d} = 2$ and an unstructured Δ. The test did not identify any set of inactive predictors. Similar result was obtained using an anisotropic structure.

3.2. The Los Alamos National Lab dataset

The dataset was from a large simulation developed at Los Alamos National Laboratory (LANL) to aid in a study of an environmental contaminant introduced into an ecosystem. It was extracted from the statistical software ARC (http://www.stat.umn.edu/arc/software.html). A description of the data can be found in Cook (1998) where a brief statistical analysis is presented. The dataset has $p = 84$ predictors with a continuous outcome and $n = 500$ observations. The initial response

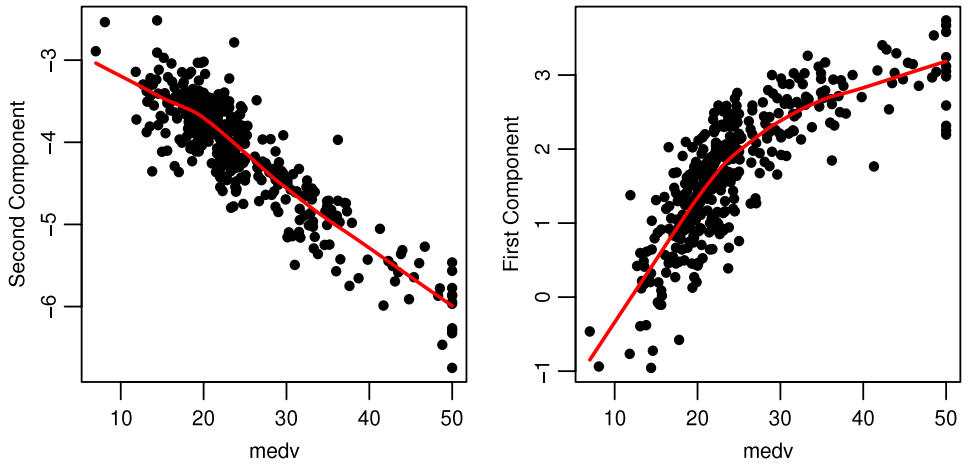


Fig. 5. Plots of the reduction of the predictors against the response of the Boston data assuming conditional independence of predictors.

variable Y is highly skewed toward larger observations and was replaced by its logarithm transformed $\log(Y)$. No transformation was made on the predictors.

We were not able to fit a PFC with an unstructured Δ as n is too small to estimate all the $p(p+1)/2 = 3528$ parameters in Δ. We thus used an anisotropic structure. With a cubic polynomial basis function, the dimension d was estimated as $\hat{d} = 1$. The plot of the reduction gave no evidence of nonlinear dependency between the predictors and the response. As such, this dataset can be analyzed using a linear model. Highly efficient shrinkage methods such as the lasso (Tibshirani, 1996), the smoothly clipped absolute deviation penalty (SCAD; Fan and Li, 2001), elastic net (Zou and Hastie, 2005), or Dantzig selector (Candès and Tao, 2007) provide the estimators of the unidirectional sparse reduction. They can be used in selecting the important predictors among the 84. We used the recently proposed coordinate descent method (glmnet; Friedman et al., 2008) that selected 14 important predictors. The use of our sequential LRT included 11 predictors in its unidirectional reduction; nine of these 11 predictors were selected by glmnet.

3.3. Sparse sufficient reduction simulation

In this simulation example, the goal is to show that more precise reduction of the data can be obtained using the sparse estimation of the sufficient reduction. To show this, we generated a Swiss roll type dataset with a number of inactive variables.

The dataset was generated as follows: We used $n = 200$ observations with $p = 600$ predictors and $d = 2$. We generated $\mathbb{Y}$ from the uniform $(0, 1)$, and $\mathbb{F} = (2\mathbb{Y}\cos(6\pi\mathbb{Y}), 2\mathbb{Y}\sin(6\pi\mathbb{Y}))^T$. With $\mathrm{G} = (\mathrm{G}_1^T, \mathrm{G}_2^T)^T \in \mathbb{R}^{p\times 2}$ where $\mathrm{G}_1 \in \mathbb{R}^{p_1\times 2}$ and $\mathrm{G}_2 \in \mathbb{R}^{p_2\times 2}$, we set $\mathrm{G}_2 = 0$ and the elements of G_1 were generated from the uniform $(1.5, 2.5)$ and used $p_1 = 100$ and $p_2 = 500$. The ith observation of $\mathbb{X}_i \in \mathbb{R}^p$ was generated as $\mathbb{X}_i = 6\mathrm{G}\mathbb{F}_i + \mathbb{E}_i$. The error term $\mathbb{E}_i$ was generated from the multivariate normal with mean 0 and diagonal covariance with its first p_1 entries being 0.5 and the remaining 10.

We applied PFC to obtain a sufficient reduction of the data. We plot the first two directions from the fitted PFC with and without the sequential LRT procedure. The fit was performed using a piecewise constant basis with an anisotropic structure. The results in Fig. 6 show a refined improvement of the sufficient reduction (b) compared to the raw reduction (a).

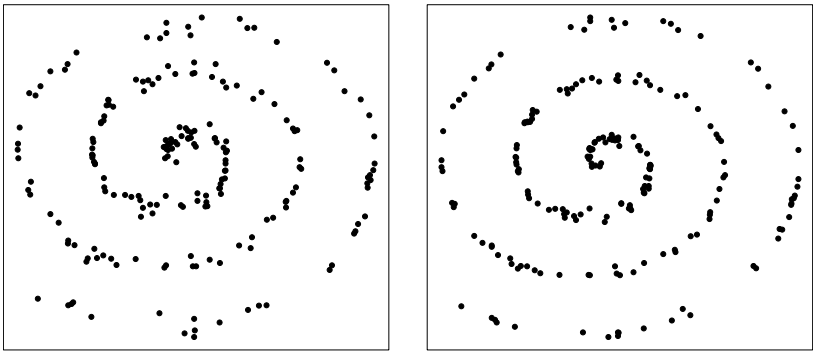
3.4. Variable selection simulation study

The performance of the method is studied for variable selection in the presence of complex relationships between the predictors and the response. We simulated datasets where the relevant and irrelevant predictors were known. We used each of the methods to estimate the true positive rate (TPR) and the false discovery rate (FDR). We define TPR and FDR as follows where ideally, TPR should be to 1 and FDR to be 0.

$$TPR = \frac{\text{number of true relevant selected}}{p_1}.$$

$$FDR = \frac{\text{number of false relevant selected}}{\text{total number selected}}.$$

We present two sets of simulations covering scenarios where the relationship between the response and the relevant predictors is nonlinear. We used two different setups to generate the data. In the first setup, we first generated a so-called latent variable and obtained the response and relevant predictors as noisy functions of the latent observations. In the second setup, the predictors were generated and the response was then obtained using a combination of the relevant predictors. The



(a) Crude reduction. (b) Sparse reduction.

Fig. 6. Plot of the first two directions of the sufficient reduction (a) without the sequential LRT, and (b) with it.

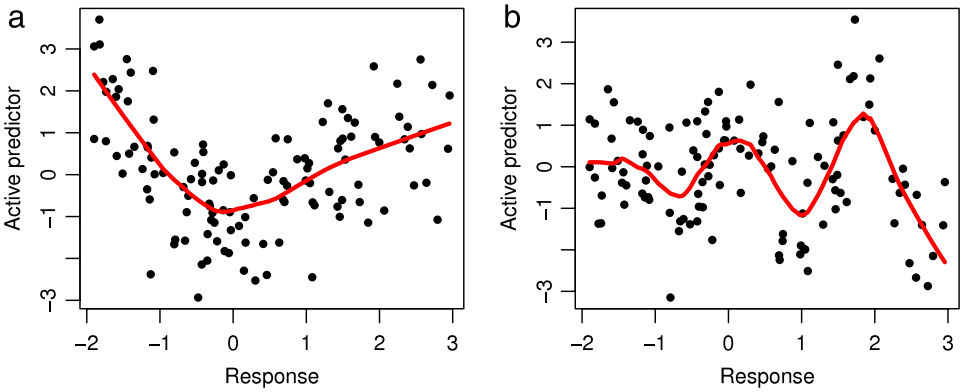


Fig. 7. Typical relationships between the predictors and the response in simulation #1 (plot a) and simulation #2 (plot b).

strength of the relationship between the relevant predictors and the response were increased from weak to strong. We used 100 data replications and obtained the estimate of TPR and FDR. A significance level of 0.05 was used for the sequential LRT. In all cases, $\mathbb{E}$ and $\mathbf{e}$ represent n observations generated respectively from the multivariate standard normal, and univariate standard normal distribution; $\mathbf{1}_n$ represents an n-vector of 1's. Fig. 7(a) and (b) shows typical relationships between the active predictors and the response in the first and second simulation sets respectively. The solid line on these plots is a nonparametric loess curve. We compare the proposed sequential likelihood ratio test (PFC-lrt) to PFC-enet, PFC-pv, and also to sparse PLS of Chun and Keles (2010). For sparse PLS, we used the R (R Development Core Team, 2013) software package *spls* of Chung et al. (2012). PFC-enet and sparse PLS are both prediction-based method. We used a ten-fold cross-validation in our simulations.

3.4.1. Simulation #1

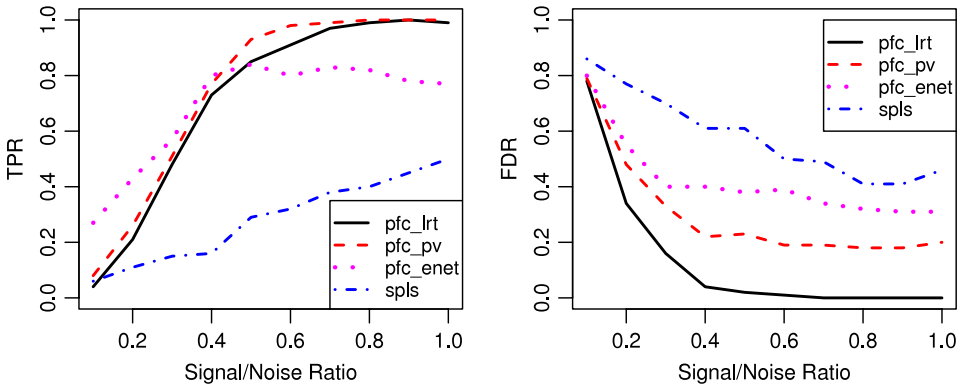
We generated the response and predictors from observations $\mathbb{U}$ of a latent variable. We used $n = 200$ observations with $p = 500$ predictors including $p_1 = 10$ active ones. The latent observations $\mathbb{U}$ were generated from the uniform $(-2, 3)$. We obtained $\mathbb{Y} = \mathbb{U} + 0.05\mathbf{e}$ and $\mathbb{X} = \lambda G\mathbb{F} + \mathbb{E}$ where $G = (\mathbf{1}_{p_1}^T, \mathbf{0}_{p-p_1}^T)^T$, and $\mathbb{F} = 0.4(\mathbb{U} + 2\mathbf{1}_n)\sin[1.2\pi(\mathbb{U} + 2\mathbf{1}_n)]$. As such, the active predictors were nonlinearly related to the response. The term $\lambda = i/10$, $i = 1, \ldots, 10$ is a constant that controls the strength of the signal input compared to the noise.

The results are plotted in Fig. 8: TPR increased and FDR decreased from weak to strong signal. PFC-pv dominated uniformly PFC-lrt for TPR, while it gave larger false discovery rates. PFC-enet flattened out around 80% TPR, which seems to indicate the inconsistency of prediction-based variable selection method (Leng et al., 2006), and sparse PLS compared unfavorably to the PFC-based methods on these data.

3.4.2. Simulations #2

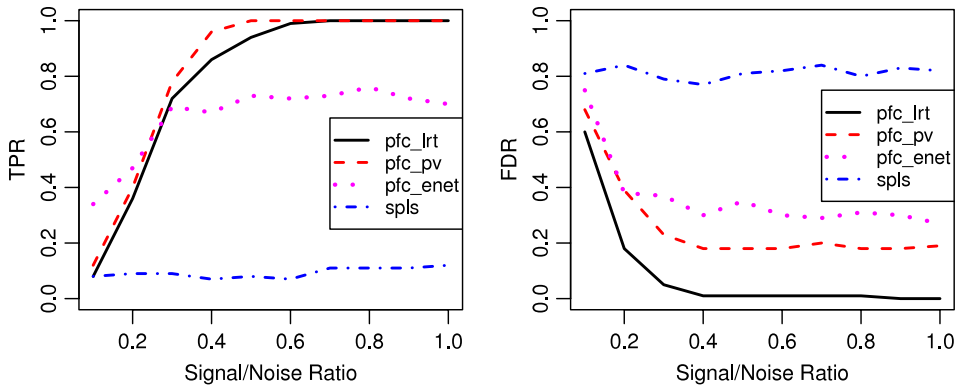
With $n = 200$ observations, we used $p = 1000$ predictors including $p_1 = 10$ active ones. We proceeded similarly as in the previous simulation setup, except that $\mathbb{F} = (\mathbb{Y}^2 - \mathbf{1}_n)I(\mathbb{Y} \le \mathbf{1}_n) + \log[\mathbb{Y}I(\mathbb{Y} \ge \mathbf{1}_n)]$.

The results in Fig. 9 are similar to those obtained in the previous simulations. The true positive rate for sparse PLS is around the level of a random selection while the false discovery is constantly at 80% regardless of the signal to noise ratio.



(a) True positive rate. (b) False discovery rate.

Fig. 8. TPR and FDR under inverse nonlinear model — Simulation #1.



(a) True positive rate. (b) False discovery rate.

Fig. 9. TPR and FDR under inverse nonlinear model — Simulation #2.

In these two simulation examples, the predictors were obtained to be nonlinearly related to the response where the correlation was essentially zero. With the use of basis functions, all active predictors are likely to be selected using the proposed sequential likelihood ratio test to form the PFC-based sparse sufficient reduction of the predictors.

4. Discussions

We have presented a sequential likelihood ratio test to obtain a sparse estimate of the sufficient reduction of the data with PFC in high dimensional setup when the relationship between the active predictors and the response is nonlinear. The sparse sufficient reduction also yields the active or important predictors relevant in explaining the response.

The sparse sufficient reduction can be readily carried into a forward model for prediction or classification. With the reduction of the dimensionality from large p to d often less than five, a graphical exploration could be carried for a proper forward modeling.

The simulation studies suggest a favorable behavior and a possible consistency of the selection of active predictors as n increases. Theoretical study of the consistency of the selection is yet to be established. Specifically, if S is the set of truly active predictors, and S^*_α is the selected set at level of significance α, what are theoretical conditions to have $S \subseteq S^*_\alpha$ with a large probability?

Appendix

Proof of Proposition 2.1. Statement (i) is obtained by noting that $\Gamma_2 = \text{span}\{E(X_2|Y = y) - E(X_2)\}$ when y varies in the sample space. Statement (ii) is a classical result of multivariate normal distribution. For statement (iii), we start with the expression (2) of the sufficient reduction and note that $\Delta^{12} = -(\Delta_{11} - \Delta_{12}\Delta_{22}^{-1}\Delta_{21})^{-1}\Delta_{12}\Delta_{22}^{-1}$. Assuming $\Delta_{12} = 0$ and $\Gamma_2 = 0$ the result follows.

Proof of the remaining propositions

The LRT is obtained as $-2(\mathcal{L}_{H_0} - \mathcal{L}_{H_a})$ where $\mathcal{L}_{H_0}$ and $\mathcal{L}_{H_a}$ are the maximized log-likelihood under $H_0 : \Gamma_2 = 0$ and $H_a : \Gamma_2 \neq 0$ respectively. These maximized log-likelihood are obtained below. In all cases, we assume that $\{X_i, y_i\}_{i=1}^n$ is observed. Furthermore, X, Γ are assumed to be partitioned as $X^T = (X_1^T, X_2^T)$, and $\Gamma^T = (\Gamma_1^T, \Gamma_2^T)$. The log-likelihood function is obtained as

$$\mathcal{L} = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Delta|) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu - \Gamma\beta \mathrm{f}_{y_i})^T\Delta^{-1}(X_i - \mu - \Gamma\beta \mathrm{f}_{y_i}). \tag{11}$$

We are assuming that $\sum_{i=1}^n \mathrm{f}_{y_i} = 0$. Maximizing $\mathcal{L}$ over μ yields $\hat{\mu} = \bar{X}$. Using the data matrices $\mathbb{X}$ and F, the partially maximized log-likelihood can be rewritten as

$$\mathcal{L} = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Delta|) - \frac{1}{2}\mathrm{Tr}\{(\mathbb{X} - F\beta^T\Gamma^T)\Delta^{-1}(\mathbb{X} - F\beta^T\Gamma^T)^T\}. \tag{12}$$

Proof of Proposition 2.2. There is no closed-form solution to the MLE of the structured Δ under H_a. Cook and Forzani (2008) provided an algorithm for such estimation, where a linear structure of both Δ and Δ^{-1} was assumed. Once Δ is estimated, under H_a, the maximized log-likelihood is

$$\begin{aligned}\mathcal{L}_{H_a} = &-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\hat{\Delta}_{11}|) - \frac{n}{2}\log(|\hat{\Delta}_{22}|)\\ &- \frac{n}{2}\mathrm{Tr}\{\hat{\Delta}_{11}^{-1}\widehat{\Sigma}_{11,\mathrm{res}}\} - \frac{n}{2}\mathrm{Tr}\{\hat{\Delta}_{22}^{-1}\widehat{\Sigma}_{22,\mathrm{res}}\} - \frac{n}{2}\sum_{i=d+1}^{p}\lambda_i(\hat{\Delta}^{-1}\widehat{\Sigma}_{\mathrm{fit}}).\end{aligned} \tag{13}$$

We now obtain the log-likelihood under H_0. Let

$$P_{\Gamma(\Delta^{-1})} = \Gamma(\Gamma^T\Delta^{-1}\Gamma)^{-1}\Gamma^T\Delta^{-1}.$$

Holding Δ and Γ fixed, the log-likelihood (12) is maximized by

$$\tilde{\beta} = \Gamma_1^T P_{\Gamma_1(\Delta^{11})}\mathbb{X}_1^T F(F^TF)^{-1}.$$

The log-likelihood becomes

$$\begin{aligned}\mathcal{L} = &-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Delta_{11}|) - \frac{n}{2}\log(|\Delta_{22}|) - \frac{1}{2}\mathrm{Tr}\{\mathbb{X}\Delta^{-1}\mathbb{X}^T\}\\ &- \frac{1}{2}\mathrm{Tr}\{-\mathbb{X}\Delta^{-1}\Gamma\tilde{\beta}F^T - F\tilde{\beta}^T\Gamma^T\Delta^{-1}\mathbb{X}^T + F\tilde{\beta}^T\Gamma^T\Delta^{-1}\Gamma\tilde{\beta}F^T\}.\end{aligned} \tag{14}$$

Now, we have the following:

$$\mathrm{Tr}\{\mathbb{X}\Delta^{-1}\mathbb{X}^T\} = n\mathrm{Tr}\{\widehat{\Sigma}_{11}\Delta_{11}^{-1}\} + n\mathrm{Tr}\{\widehat{\Sigma}_{22}\Delta^{22}\}. \tag{15}$$

$$\mathrm{Tr}\{\mathbb{X}\Delta^{-1}\Gamma\tilde{\beta}F^T\} = n\mathrm{Tr}\{\widehat{\Sigma}_{\mathbf{1}1,\mathrm{fit}}\Delta_{11}^{-1}\Gamma_1\Gamma_1^T P_{\Gamma_1(\Delta_{11}^{-1})}\} \tag{16}$$

$$\mathrm{Tr}\{F\tilde{\beta}^T\Gamma^T\Delta^{-1}\mathbb{X}^T\} = n\mathrm{Tr}\{\widehat{\Sigma}_{11,\mathrm{fit}}P^T_{\Gamma_1(\Delta_{11}^{-1})}\Gamma_1\Gamma_1^T\Delta_{11}^{-1}\} \tag{17}$$

$$\mathrm{Tr}\{F\tilde{\beta}^T\Gamma^T\Delta^{-1}\Gamma\tilde{\beta}F^T\} = n\mathrm{Tr}\{\widehat{\Sigma}_{11,\mathrm{fit}}\Delta_{11}^{-1}\Gamma_1\Gamma_1^T P_{\Gamma_1(\Delta_{11}^{-1})}\}. \tag{18}$$

And the partially maximized log-likelihood simplifies to

$$\begin{aligned}\mathcal{L} = &-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Delta_{11}|) - \frac{n}{2}\log(|\Delta_{22}|) - \frac{n}{2}\mathrm{Tr}\{\widehat{\Sigma}_{11}\Delta_{11}^{-1}\}\\ &- \frac{n}{2}\mathrm{Tr}\{\widehat{\Sigma}_{22}\Delta_{22}^{-1}\} + \frac{n}{2}\mathrm{Tr}\{\widehat{\Sigma}_{11,\mathrm{fit}}P^T_{\Gamma_1(\Delta_{11}^{-1})}\Gamma_1\Gamma_1^T\Delta_{11}^{-1}\}.\end{aligned} \tag{19}$$

Noting that $P^T_{\Gamma_1(\Delta_{11}^{-1})}\Gamma_1\Gamma_1^T\Delta_{11}^{-1} = P^T_{\Gamma_1(\Delta_{11}^{-1})}\Gamma_1\Gamma_1^T\Delta_{11}^{-1} = \Delta_{11}^{-1/2}P_{\Gamma_1(\Delta_{11}^{-1/2})}\Delta_{11}^{-1/2}$, we then obtain

$$\begin{aligned}\mathcal{L} = &-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Delta_{11}|) - \frac{n}{2}\log|\Delta_{22}| - \frac{n}{2}\mathrm{Tr}\{\widehat{\Sigma}_{11}\Delta_{11}^{-1}\}\\ &- \frac{n}{2}\mathrm{Tr}\{\widehat{\Sigma}_{22}\Delta_{22}^{-1}\} + \frac{n}{2}\mathrm{Tr}(\Delta_{11}^{-1/2}\widehat{\Sigma}_{11,\mathrm{fit}}\Delta_{11}^{-1/2}P_{\Gamma_1(\Delta_{11}^{-1/2})}).\end{aligned} \tag{20}$$

Keeping Δ_{11} fixed, $\mathcal{L}(\Gamma_1)$ is maximized by the eigenvectors corresponding to the first d largest eigenvalues of $\Delta_{11}^{-1}\widehat{\Sigma}_{11,\text{fit}}$. This leads to

$$\mathcal{L} = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Delta_{22}|) - \frac{n}{2}\text{Tr}\{\widehat{\Sigma}_{22}\Delta^{22}\} - \frac{n}{2}\log(|\Delta_{11}|) - \frac{n}{2}\text{Tr}\{\widehat{\Sigma}_{11}\Delta^{11}\} + \frac{n}{2}\sum_{i=1}^{d}\hat{\lambda}_i(\Delta_{11}^{-1}\widehat{\Sigma}_{11,\text{fit}}), \tag{21}$$

where $\hat{\lambda}_i(A)$ is the ith largest eigenvalues of A. As a function of Δ_{22}, $\mathcal{L}$ is maximized by $\hat{\Delta}_{22} = \widehat{\Sigma}_{22}$. The partially maximized log-likelihood becomes

$$\mathcal{L} = -\frac{np}{2}\log(2\pi) - \frac{np_2}{2} - \frac{n}{2}\log(|\widehat{\Sigma}_{22}|) - \frac{n}{2}\log(|\Delta_{11}|) - \frac{n}{2}\text{Tr}\{\widehat{\Sigma}_{11,\text{res}}\Delta_{11}^{-1}\} - \frac{n}{2}\sum_{i=d+1}^{p_1}\hat{\lambda}_i(\Delta_{11}^{-1}\widehat{\Sigma}_{11,\text{fit}}). \tag{22}$$

This last expression, as a function of Δ_{11} is similar to Eq. (4) in Cook and Forzani (2008) where Theorem 3.1 provides the MLE and the final maximized log-likelihood under H_0 is

$$\mathcal{L}_{H_0} = -\frac{np}{2}\log(2\pi) - \frac{np}{2} - \frac{n}{2}\log(|\widehat{\Sigma}_{22}|) - \frac{n}{2}\log|\widehat{\Sigma}_{11,\text{res}}| - \frac{n}{2}\sum_{i=d+1}^{p_1}\log[1+\hat{\lambda}_i(\widehat{\Sigma}_{11,\text{res}}^{-1}\widehat{\Sigma}_{11,\text{fit}})]. \tag{23}$$

Proof of Proposition 2.3. The partially maximized log-likelihood function under H_a with $\Delta = \sigma^2 I$, is

$$\mathcal{L} = -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|\text{Vec}(\mathbb{X}) - (\Gamma\otimes F)\text{Vec}(\beta^T)\|^2. \tag{24}$$

As a function of β, holding Γ and σ^2 fixed, the log-likelihood is maximized with $\tilde{\beta} = \Gamma^T\mathbb{X}^T F(F^TF)^{-1}$. Substituting $\tilde{\beta}$ in the log-likelihood yields

$$\mathcal{L}(\sigma^2, \Gamma) = -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\text{Tr}\{\widehat{\Sigma} - \widehat{\Sigma}_{\text{fit}}P_\Gamma\}. \tag{25}$$

As a function of Γ alone, $\mathcal{L}(\Gamma)$ is maximized by the subspace spanned by the first d largest eigenvalues of $\widehat{\Sigma}_{\text{fit}}$. Let $\hat{\lambda}_1^{\text{fit}} \geq \cdots \geq \hat{\lambda}_p^{\text{fit}}$ be the ordered eigenvalues of $\widehat{\Sigma}_{\text{fit}}$. The MLE of σ^2 is obtained by maximizing the partially maximized log-likelihood

$$\mathcal{L}(\sigma^2) = -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\text{Tr}\{\widehat{\Sigma} - \widehat{\Sigma}_{\text{fit}}P_{\widehat{\Gamma}}\}$$

which yields $\hat{\sigma}^2 = [\sum_{i=1}^{p}\hat{\lambda}_i - \sum_{j=1}^{d}\hat{\lambda}_j^{\text{fit}}]/p$. Under H_a, the maximized log-likelihood for a fixed d becomes

$$\mathcal{L}_{H_a} = -\frac{np}{2}\log(2\pi) - \frac{np}{2} - \frac{np}{2}\log\left[\left(\sum_{i=1}^{p}\hat{\lambda}_i - \sum_{j=1}^{d}\hat{\lambda}_j^{\text{fit}}\right)\Big/ p\right]. \tag{26}$$

Under H_0 $\tilde{\beta} = \Gamma_1^T\mathbb{X}^T F(F^TF)^{-1}$. Replacing $\tilde{\beta}$ in the log-likelihood (24), the MLE of Γ_1 is obtained as the eigenvectors corresponding to the first d largest eigenvalues of $\widehat{\Sigma}_{11,\text{fit}}$. The MLE of σ^2 is then $\hat{\sigma}^2 = [\sum_{i=1}^{p}\hat{\lambda}_i - \sum_{i=1}^{d}\hat{\lambda}_i^{11,\text{fit}}]/p$, where $\hat{\lambda}_i^{11,\text{fit}}$ is the ith largest eigenvalue of $\widehat{\Sigma}_{\text{fit}}^{11}$. The maximized log-likelihood is

$$\mathcal{L}_{H_0} = -\frac{np}{2}\log(2\pi) - \frac{np}{2} - \frac{np}{2}\log\left[\frac{\sum_{i=1}^{p}\hat{\lambda}_i - \sum_{i=1}^{d}\hat{\lambda}_i^{11,\text{fit}}}{p}\right]. \tag{27}$$

Proof of Proposition 2.4. The partially maximized log-likelihood function under H_a, with $\Delta = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0$ is,

$$\begin{aligned}\mathcal{L}(\Omega, \Omega_0, \Gamma, \beta) = &-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Omega|) - \frac{n}{2}\log(|\Omega_0|)\\ &-\frac{1}{2}\text{Tr}\{(\mathbb{X} - F\beta^T\Gamma^T)(\Gamma\Omega^{-1}\Gamma + \Gamma_0\Omega_0^{-1}\Gamma_0)(\mathbb{X} - F\beta^T\Gamma^T)^T\}.\end{aligned}$$

Let $\mathcal{L}_1$ be the summand function of Ω_0 and Γ_0 only.

$$\mathcal{L}_1(\Omega_0, \Gamma_0) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Omega_0| - \frac{1}{2}\text{Tr}\{\mathbb{X}\Gamma_0\Omega_0^{-1}\Gamma_0^T\mathbb{X}^T\}. \tag{28}$$

Holding Γ_0 fixed, $\mathcal{L}_1(\Omega_0)$ is maximized by $\widetilde{\Omega}_0 = \Gamma_0\widehat{\Sigma}\Gamma_0$ to become

$$\mathcal{L}_1(\Gamma_0) = -\frac{np}{2}\log(2\pi) - \frac{n(p-d)}{2} - \frac{n}{2}\log|\Gamma_0\widehat{\Sigma}\Gamma_0|. \tag{29}$$

The remaining summand of $\mathcal{L}$ is

$$\begin{aligned}\mathcal{L}_2(\Omega, \beta, \Gamma) &= \frac{n}{2}\log|\Omega| - \frac{1}{2}\mathrm{Tr}\{(\mathbb{X} - F\beta^T\Gamma^T)\Gamma\Omega^{-1}\Gamma^T(\mathbb{X} - F\beta^T\Gamma^T)^T\}\\ &= \frac{n}{2}\log|\Omega| - \frac{1}{2}\|\mathrm{Vec}(\mathbb{X}\Gamma\Omega^{-1/2} - F\beta^T\Omega^{-1/2})\|^2\\ &= \frac{n}{2}\log|\Omega| - \frac{1}{2}\|\mathrm{Vec}(\mathbb{X}\Gamma\Omega^{-1/2}) - (\Omega^{-1/2}\otimes F)\mathrm{Vec}(\beta^T)\|^2.\end{aligned}$$

As a function of β, holding Γ and Ω fixed, $\mathcal{L}_2$ is maximized with $\tilde{\beta} = \Gamma^T\mathbb{X}^TF(F^TF)^{-1}$. This leads to the partially maximized $\mathcal{L}_2$

$$\mathcal{L}_2(\Omega, \Gamma) = -\frac{n}{2}\log|\Omega| - \frac{1}{2}\mathrm{Tr}\{\Gamma^T(\widehat{\Sigma} - \widehat{\Sigma}_{\mathrm{fit}})\Gamma\Omega\}. \tag{30}$$

As a function of Ω alone, while holding Γ fixed, $\mathcal{L}_2(\Omega)$ is maximized by $\widetilde{\Omega} = \Gamma^T\widehat{\Sigma}_{\mathrm{res}}\Gamma$. The partially maximized log-likelihood becomes

$$\mathcal{L}(\Gamma) = -\frac{np}{2}\log(2\pi) - \frac{np}{2} - \frac{n}{2}\log|\Gamma^T\widehat{\Sigma}_{\mathrm{res}}\Gamma| - \frac{n}{2}\log|\Gamma_0^T\widehat{\Sigma}\Gamma_0|.$$

This latter objection function is maximized over the Grassmann manifold of dimension $d(p-d)$ to obtain the estimate of the subspace spanned by the columns of Γ. Let $\widehat{\Gamma}$ be a basis representative of this estimated subspace, and $\widehat{\Gamma}_0$ be the orthogonal completion of $\widehat{\Gamma}$. Under H_a and for a fixed d, the maximized log-likelihood becomes

$$\mathcal{L}_{H_a} = -\frac{np}{2}\log(2\pi) - \frac{np}{2} - \frac{n}{2}\log|\widehat{\Gamma}^T\widehat{\Sigma}_{\mathrm{res}}\widehat{\Gamma}| - \frac{n}{2}\log|\widehat{\Gamma}_0^T\widehat{\Sigma}\widehat{\Gamma}_0|. \tag{31}$$

Under H_0, $\tilde{\beta} = \Gamma_1^T\mathbb{X}_1^TF(F^TF)^{-1}$. The summand $\mathcal{L}_2$ in (30) becomes

$$\mathcal{L}_2(\Omega, \Gamma) = -\frac{n}{2}\log|\Omega| - \frac{1}{2}\mathrm{Tr}\{\Gamma_1^T\widehat{\Sigma}_{11,\mathrm{res}}\Gamma_1\Omega\} \tag{32}$$

which is maximized over Ω, holding Γ fixed, by $\widetilde{\Omega} = \Gamma_1^T\widehat{\Sigma}_{11,\mathrm{res}}\Gamma_1$. The parameter Γ_1 is estimated by maximizing the following partially maximized log-likelihood function

$$\mathcal{L}(\Gamma_1) = -\frac{np}{2}\log(2\pi) - \frac{np}{2} - \frac{n}{2}\log|\Gamma_1^T\widehat{\Sigma}_{11,\mathrm{res}}\Gamma_1| - \frac{n}{2}\log|\Gamma_0^T\widehat{\Sigma}\Gamma_0|$$

over the Grassmann manifold of dimension $d(p_1 - d)$, where we assumed $p_1 > d$. Let $\widehat{\Gamma}_{01}$ be the orthogonal completion of $(\widehat{\Gamma}_1^T, \mathbf{0}^T)^T$ where $\mathbf{0} \in \mathbb{R}^{p_2\times d}$. The maximized log-likelihood can be written as

$$\mathcal{L}_{H_0} = -\frac{np}{2}\log(2\pi) - \frac{np}{2} - \frac{n}{2}\log|\widehat{\Gamma}_1^T\widehat{\Sigma}_{11,\mathrm{res}}\widehat{\Gamma}_1| - \frac{n}{2}\log|\widehat{\Gamma}_{01}^T\widehat{\Sigma}\widehat{\Gamma}_{01}|. \tag{33}$$

References

Adragni, K.P., Cook, R.D., 2008. Discussion on the sure independence screening for ultrahigh dimensional feature space of Jianqing Fan and Jinchi Lv (2007). J. R. Stat. Soc. Ser. B 70, 1–35.

Adragni, K.P., Cook, R.D., 2009. Sufficient dimension reduction and prediction in regression. Phil. Trans. R. Soc. A 367, 1906.

Adragni, K.P., Raim, A., 2014. ldr: Methods for likelihood-based dimension reduction in regression. R package version 1.3, http://CRAN.R-project.org/package=ldr.

Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37 (4), 373–384.

Candès,, Tao,, 2007. The Dantzig selector: statistical estimation when p is much larger than n. Ann. Statist. 35 (6), 2313–2351.

Chen, X., Zou, C., Cook, R.D., 2010. Coordinate-independent sparse sufficient dimension reduction and variable selection. Ann. Statist. 38 (6), 3696–3723.

Chun, H., Keles, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (1), 3–25.

Chung, D., Chun, H., Keles, S., 2012. spls: Sparse Partial Least Squares (SPLS) Regression and Classification. R package version 2.1-2.

Cook, R.D., 1998. Regression Graphics: Ideas for Studying Regression Through Graphics. Wiley, New York.

Cook, R.D., 2004. Testing predictor contributions in sufficient dimension reduction. Ann. Statist. 32 (3), 1062–1092.

Cook, R.D., 2007. Fisher lecture. Statist. Sci. 22 (1), 1–26.

Cook, R.D., Forzani, L., 2008. Principal fitted components for dimension reduction in regression. Statist. Sci. 23, 485–501.

Cook, R.D., Ni, L., 2005. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. J. Amer. Statist. Assoc. 100, 927–1010.

Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman & Hall, London & New York.

Donoho, D., 1995. De-noising by soft-thresholding. IEEE Trans. Inform. Theory 41 (3), 613–627.

Enz, R., 1991. Prices and Earnings Around the Globe. Union Bank of Switzerland.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456).

Fan, J., Lv, J., 2008. Sure independence screening for ultra-high dimensional feature space. J. R. Stat. Soc. Ser. B 70, 849–911.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), http://www.stanford.edu/~hastie/Papers/glmnet.pdf.

Leng, C., Lin, Y., Wahba, G., 2006. A note on the lasso and related procedures in model selection. Statist. Sinica 16, 1273–1284.

Li, K.C., 1991. Sliced inverse regression for dimension reduction (with discussion). J. Amer. Statist. Assoc. 86, 316–342.
Li, K.C., 1992. On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. J. Amer. Statist. Assoc. 87, 1025–1039.
Li, L., 2007. Sparse sufficient dimension reduction. Biometrika 94 (3), 603–613.
Li, B., Wang, S., 2007. On directional regression for dimension reduction. J. Amer. Statist. Assoc. 102, 997–1008.
Li, L., Yin, X., 2008. Sliced inverse regression with regularizations. Biometrics 64 (1), 124–131.
R Development Core Team 2013. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.
Tibshirani, R., 1996. Regression shrinkage and and selection via the lasso. J. R. Stat. Soc. Ser. B 58, 267–288.
Wang, Tao, Zhu, Lixing, 2013. Sparse sufficient dimension reduction using optimal scoring. Comput. Statist. Data Anal. 57, 223–232.
Xue, L., Ma, S., Zou, H., 2012. Positive-definite l_1-penalized estimation of large Covariance Matrices. J. Amer. Statist. Assoc. 107 (500), 1480–1491.
Zhong, W., Zeng, P., Ma, P., Liu, J.S., Zhu, Y., 2005. RSIR: regularized sliced inverse regression for motif discovery. Bioinformatics 21 (22), 4169–4175.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the Elastic Net. J. R. Stat. Soc. Ser. B 67, 301–320. Part 2.