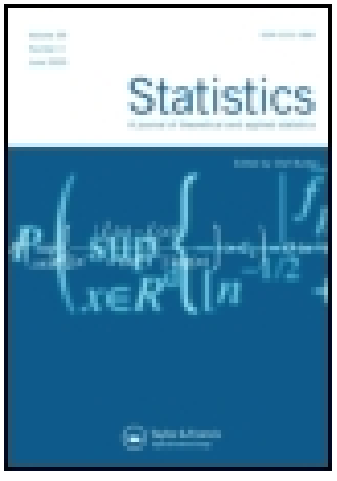

Statistics: A Journal of Theoretical and Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gsta20

Pruning a sufficient dimension reduction with a p-value guided hard-thresholding

Kofi P. Adragni[a] & Mingyu Xi[a]
[a] Department of Mathematics and Statistics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA
Published online: 11 Jun 2015.

To cite this article: Kofi P. Adragni & Mingyu Xi (2015): Pruning a sufficient dimension reduction with a p-value guided hard-thresholding, Statistics: A Journal of Theoretical and Applied Statistics, DOI: 10.1080/02331888.2015.1050019

To link to this article: http://dx.doi.org/10.1080/02331888.2015.1050019

Statistics, 2015
http://dx.doi.org/10.1080/02331888.2015.1050019

Pruning a sufficient dimension reduction with a *p*-value guided hard-thresholding

Kofi P. Adragni* and Mingyu Xi

Department of Mathematics and Statistics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA

(*Received 30 May 2013; accepted 4 May 2015*)

Principal fitted component (PFC) models are a class of likelihood-based inverse regression methods that yield a so-called sufficient reduction of the random p-vector of predictors X given the response Y. Assuming that a large number of the predictors has no information about Y, we aimed to obtain an estimate of the sufficient reduction that 'purges' these irrelevant predictors, and thus, select the most useful ones. We devised a procedure using observed significance values from the univariate fittings to yield a sparse PFC, a purged estimate of the sufficient reduction. The performance of the method is compared to that of penalized forward linear regression models for variable selection in high-dimensional settings.

Keywords: dimension reduction; principal fitted component; sparsity; variable selection

1. Introduction

Scientists now routinely collect large amount of data and formulate regressions for which the number of predictors p is often too large to allow a thorough graphical exploration of the data. It is observed that regression data are collected jointly on (Y, X) where $\mathrm{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is a random p-vector and Y is a univariate response. In a forward model of $Y \mid \mathrm{X}$, the predictors are usually treated as fixed, non-random; however, there is no compelling reason not to use the information of X given Y. A model of X given Y falls into the framework of inverse regression. Numerous inverse regression methodologies exist in the statistics literature, including sliced inverse regression ([1], SIR) sliced average variance estimation ([2], SAVE) and directional regression ([3], DR). In general, inverse methods are arguably the best suited to deal with the high dimensionality in regression.

Cook [4] proposed a likelihood-based inverse regression method called principal fitted components (PFC) models with the aim to estimate a sufficient reduction of X. Following is a formal definition of *sufficient reduction*,[4] as defined below.

DEFINITION 1.1 *A reduction $R : \mathbb{R}^p \to \mathbb{R}^d, d \leq p$, is sufficient if it satisfies one of the following three statements*: (i) $Y \mid \mathrm{X} \sim Y \mid R(\mathrm{X})$, (ii) $\mathrm{X} \mid (Y, R(\mathrm{X})) \sim \mathrm{X} \mid R(\mathrm{X})$, *and* (iii) $\mathrm{X} \perp\!\!\!\perp Y \mid R(\mathrm{X})$. *The symbol $\perp\!\!\!\perp$ stands for statistical independence, and $U \sim V$ stands for U and V having identical distribution.*

*Corresponding author. Email: kofi@umbc.edu

The reduction $R(\mathrm{X})$ captures all the information about the response Y that is contained in X. Statement (i) holds in a forward regression and requires the conditional distribution of $Y \mid \mathrm{X}$. For example, in prospective study designs such as clinical trials, X is fixed by design and $Y \mid \mathrm{X}$ is observed. Statement (ii) holds in an inverse regression and requires the conditional distribution of $\mathrm{X}|Y$. In retrospective studies, for example, case control designs, X is observed given the response or status Y. Statement (iii) requires the joint distribution of (Y, X). Under a joint distribution of (Y, X), the three statements are equivalent. Consequently, a model of $\mathrm{X} \mid Y$ can be used to obtain a sufficient reduction of X and use this reduction in lieu of X in modelling $Y \mid \mathrm{X}$.

A principal fitted components model is obtained as follows: Let X_y denote the p-vector random variable distributed as $\mathrm{X}|(Y = y)$ and let $\mu = E(\mathrm{X})$ and $\mu_y = E(\mathrm{X}_y)$. This model is based on the assumption that X_y has a multivariate normal distribution, and is therefore only appropriate for many-valued, quantitative, continuous or nearly-continuous predictors. It is assumed that $\mu_y - \mu$ falls in a subspace $\mathcal{S}$ of dimension d in $\mathbb{R}^p$ as y varies in its sample space. Let $\Gamma \in \mathbb{R}^{p\times d}$ denote a semi-orthogonal basis matrix of $\mathcal{S}$, such that $\Gamma^T\Gamma = \mathrm{I}_d$. We can then write $\mathrm{X}_y \sim N(\mu + \Gamma\nu_y, \Delta)$ where $\nu_y = \Gamma^T(\mu_y - \mu)$ is a function of y. The conditional variance Δ is assumed to be independent of Y. Once the response values are observed, the unknown function ν_y can be modelled as $\nu_y = \lambda(\mathrm{f}_y - E(\mathrm{f}_Y))$, where $\lambda \in \mathbb{R}^{d\times r}$ is an unknown and unconstrained parameter of rank at most $d \leq \min\{p, r\}$, and f_y is a flexible basis function. The subsequent model, written as

$$\mathrm{X}_y = \mu + \Gamma\lambda[\mathrm{f}_y - E(\mathrm{f}_Y)] + \Delta^{1/2}\varepsilon, \tag{1}$$

is referred to as a PFC model. Typically, the function $\mathrm{f}_y \in \mathbb{R}^r$ is a user-selected function. It helps capture the dependency of X on Y. Clearly, the dependence of the predictors on Y is captured through the row elements of Γ. Cook [4] showed that a sufficient reduction of X is $\Gamma^T\Delta^{-1}\mathrm{X}$ which is a set of d linear combinations of the p predictors. A predictor X_i is thus relevant in the reduction if its corresponding row of $\Delta^{-1}\Gamma$ is nonzero. We herein demonstrate that inverse regression can be an alternative to forward model for variable selection.

We propose a hard-thresholding procedure called 'p-value guided hard-thresholding' which is based on the magnitude of the scaled row elements of $\hat{\Gamma}$, an estimate of Γ. Hard-thresholding has been used in the literature [5,6] in wavelet estimation context for image-denoising. If any row element of $\hat{\Gamma}$ has an absolute value less than a specified threshold, then this procedure will set it to zero. Thus, the procedure serves to *prune* the sufficient reduction of irrelevant predictors.

2. Preliminaries

We start with the most restrictive covariance structure of Δ by assuming that, conditional to the response, the predictors are independent and on the same measurement scale, that is $\Delta = \delta^2\mathrm{I}$. The subsequent model is referred to as an *isotropic* PFC. The sufficient reduction of X is $\Gamma^T\mathrm{X}$.[4]

To describe the estimator of the sufficient reduction, we assume that n data points (x_i, y_i), $i = 1, \ldots, n$ from (X, Y) are observed. We denote by $\mathbb{X} \in \mathbb{R}^{n\times p}$, the centred data matrix with the ith row $(\mathrm{x}_i - \bar{\mathrm{x}})^T$ with $\bar{\mathrm{x}} = \sum_{i=1}^n \mathrm{x}_i/n$, and $\mathbb{F} \in \mathbb{R}^{n\times r}$ is the data matrix of the basis function with its ith row $(\mathrm{f}_{y_i} - \bar{\mathrm{f}})^T$ where $\bar{\mathrm{f}} = \sum_{i=1}^n \mathrm{f}_{y_i}/n$. We let $\mathrm{P}_{\mathbb{F}} = \mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}\mathbb{F}^T$ denote the projection operator on the space spanned by the columns of $\mathbb{F}$. The maximum likelihood estimator of Γ can be obtained as any orthonormal basis of the subspace spanned by the d eigenvectors of $\hat{\Sigma}_{\text{fit}} = \mathbb{X}^T\mathrm{P}_{\mathbb{F}}\mathbb{X}/n$ corresponding to the largest d eigenvalues. The estimated sufficient reduction $\hat{\Gamma}^T\mathrm{X}$ is the principal component of $\hat{\Sigma}_{\text{fit}}$, and is thus referred to as the PFC.

Obtaining a sparse reduction from model (1) amounts to estimating a sparse principal component of $\hat{\Sigma}_{\text{fit}}$. The nonzero rows of $\hat{\Gamma}$ correspond to the relevant predictors to be identified. Several methods have been proposed to help estimate sparse principal components.[7–10] Zou

et al. [11] formulated principal component analysis as a regression-type optimization problem and proposed a sparse principal components analysis (SPCA) algorithm by imposing the lasso [12] constraint on the regression coefficients. The idea of SPCA was elaborated as a generic formulation for sparse sufficient dimension reduction by Li.[13] We adapted and implemented the algorithm of Li.[13] It is identical to the algorithm used by Chun and Keleş [14] to find the first partial least-squares direction vector. We studied its performance and observed that it was not sensitive to the inclusion of inactive predictors. To circumvent this, we propose an alternative method to estimate the sparse principal fitted component using a hard-thresholding method.

The proposed method includes predictors in the estimated reduction based on the significance of their relationship with the response. It is reminiscent of the method of Cook [15] on testing predictors contribution to the sufficient reduction in the context of SIR where large sample tests were used. With PFC, the normality assumption on $\mathrm{X} \mid Y$ provides an exact distribution to the tests of these predictors. It is worth noting that SIR is a moment-based sufficient dimension reduction method. It has been shown that when the response Y is categorical, SIR of Li [1] and PFC estimate the same minimal sufficient reduction subspace. When Y is continuous, SIR discretizes the response through a slicing procedure, which can inadvertently leave intra-slice information behind. On the other hand, the use of flexible basis functions in PFC can potentially help avoid such loss of information.[4,16]

In the remainder of the paper, we present the sparse estimation with the focus on single-index principal fitted components models where $d = 1$ in Section 3. In Section 4, we provide details for the sparse estimation for multiple-index PFC models where $d > 1$. Simulation studies comparing the performance of the sparse single-index PFC to some existing variable selection methods are in Section 5 where applications to two data sets are presented. We provide a theoretical justification of our simulation results in Section 6, followed by some discussions in Section 7.

3. A *P*-value guided hard-thresholding on single-index PFC

We continue with an isotropic single-index PFC model with $\Delta = \delta^2 \mathrm{I}$ and $\Gamma \in \mathbb{R}^{p\times 1}$. The sufficient reduction of X is $\Gamma^T \mathrm{X}$.[4] Let us partition the predictors into two groups, $\mathrm{X}_{(1)} \in \mathbb{R}^{p_1}$ and $\mathrm{X}_{(2)} \in \mathbb{R}^{p_2}$, and partition Γ accordingly. We assume that $\mathrm{X}_{(1)}$ is the set of active or relevant predictors that are linearly or nonlinearly related to the response, and that $\mathrm{X}_{(2)}$ is inactive and has no information about Y. The following proposition helps in setting up the variable pruning.

PROPOSITION 3.1 *Consider PFC model* (1) *where* $\Delta = \delta^2 \mathrm{I}$ *is independent of Y, and assume the aforementioned partitioning of* X *and* Γ. *Then, we have*

$$\mathrm{X}_{(2)} \perp\!\!\!\perp Y \Leftrightarrow \Gamma_2 = 0.$$

The statement in this proposition can be proved by observing that $E(\mathrm{X}_{(2)}|Y) - E(\mathrm{X}_{(2)}) = \Gamma_2 \nu_Y$, with $\nu_Y \neq 0$. This implies that $\Gamma_2 = 0$ if and only if $E(\mathrm{X}_{(2)}|Y) = E(\mathrm{X}_{(2)})$. With $\Gamma_2 = 0$, the sufficient reduction becomes $\Gamma_1^T \mathrm{X}_{(1)}$. The focus now is to properly identify both Γ_2 and Γ_1. Let $\Gamma\lambda = \Phi$. Model (1) can then be expressed as p independent simple linear regressions

$$X_i = \mu_i + \phi_i \mathrm{f}_y + \delta_i \varepsilon_i, \quad i = 1, \ldots, p, \tag{2}$$

where X_i and ϕ_i are respectively the ith predictor in X and the ith row of Φ. Thus, p independent models (2) that are individually a linear regression model can be fitted. 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 $\hat{\phi}_i$ be the ordinary least-squares estimator of ϕ_i. We consider the usual test statistic

$$F_i = \frac{n-r-1}{r} \cdot \frac{\sum_{j=1}^{n}[(\mathrm{x}_{ij} - \bar{\mathrm{x}}_{i\cdot})^2 - (\mathrm{x}_{ij} - \bar{\mathrm{x}}_{i\cdot} - \hat{\phi}_i \mathrm{f}_{y_j})^2]}{\sum_{j=1}^{n}(\mathrm{x}_{ij} - \bar{\mathrm{x}}_{i\cdot} - \hat{\phi}_i \mathrm{f}_{y_j})^2}$$

for testing $H_0 : \phi_i = 0$ against $H_a : \phi_i \neq 0$, where $\bar{\mathrm{x}}_{i\cdot} = \sum_{j=1}^{n} \mathrm{x}_{ij}/n$. The statistic F_i follows an F distribution with $(r, n-r-1)$ degrees of freedom. Let $\pi_i = P(F \geq F_i \,|\, \phi_i = 0)$ be the p-value resulting from the test, and let $\pi = (\pi_1, \ldots, \pi_p)^T$, and let 1_p be the p-vector of ones. The *crude* p-value guided thresholding estimator $\hat{\Gamma}_\alpha$ is obtained as

$$\hat{\Gamma}_\alpha = J(\pi \leq \alpha 1_p) \circ \hat{\Gamma}, \tag{3}$$

where $\hat{\Gamma}$ is an estimator of Γ. The inequality is element-wise, and $J(\cdot)$ represents the indicator function. The operator $\circ$ stands for the Hadamard product of matrices. With this procedure, $\pi_i \leq \alpha$ is an evidence against H_0 and X_i is assumed active; the corresponding row in $\hat{\Gamma}_\alpha$ 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 irrelevant predictors.

This hard-thresholding procedure is akin to multiple hypotheses testing where the concern is often to control the family-wise error rate. Testing the p hypotheses while controlling for the family-wise error rate ensures the simultaneous correctness of inferences about the p predictors. It thus guarantees a correct overall decision in the selection of the active predictors. The goal of this procedure, however, is not to control for the family-wise error rate; thus each of the p tests is carried out at level α, regardless of p.

The univariate model fittings (2) have been shown to subsume the sure independence screening of Fan and Lv [17] for variable screening.[18] Using the simplest basis function $\mathrm{f}_y = (y - \bar{y})$ in the univariate models turns ϕ_is into scalars. Larger values of the estimated $|\tilde{\phi}_i|$ correspond to predictors with stronger dependence on the response. It can be shown that $|\tilde{\phi}_i|$ is proportional to the sample correlation between X_i and y. Predictors with large values of $|\tilde{\phi}_i|$ are retained in the pruned sufficient reduction.

3.1. *Conditionally dependent predictors*

An unstructured and nonsingular Δ is now considered. We partition Δ and Δ^{-1} as $(\Delta_{ij})_{i,j=1,2}$ and $(\Delta^{(ij)})_{i,j=1,2}$ according to the partition of X. The sufficient reduction of X can then be rewritten as

$$\Gamma^T \Delta^{-1} \mathrm{X} = [\Gamma_1^T \Delta^{(11)} + \Gamma_2^T \Delta^{(21)}]\mathrm{X}_{(1)} + [\Gamma_1^T \Delta^{(12)} + \Gamma_2^T \Delta^{(22)}]\mathrm{X}_{(2)}.$$

With $\Gamma_2 = 0$, the sufficient reduction $\Gamma^T \Delta^{-1} \mathrm{X} = \Gamma_1^T \Delta^{(11)} \mathrm{X}_{(1)} + \Gamma_1^T \Delta^{(12)} \mathrm{X}_{(2)}$ still retains the inactive predictors $\mathrm{X}_{(2)}$ as long as $\Delta^{(12)} \neq 0$. If $\mathrm{X}_{(2)}$ is inactive and yet $\Delta_{12} \neq 0$, it may mean that $\mathrm{X}_{(1)}$ and $\mathrm{X}_{(2)}$ are related through an unobserved latent variable. We chose to ignore this conditional dependency between X_1 and X_2 and henceforth assume that $\Delta_{12} = 0$, which implies that $\Delta^{(12)} = 0$. The following proposition summarizes the assumptions and the result.

PROPOSITION 3.2 *PFC model* (1) *with the aforementioned partitioning of* X, Γ *and* Δ *bears the following*:

- *If (a)* $\mathrm{X}_{(2)} \perp\!\!\!\perp Y$ *and (b)* $\mathrm{X}_{(1)} \perp\!\!\!\perp \mathrm{X}_{(2)} | Y$ *then* $\Gamma_1^T \Delta_{11}^{-1} \mathrm{X}_{(1)}$ *is a sufficient reduction of* X.

Condition (a) is equivalent to $\Gamma_2 = 0$ which means that X_2 is inactive. Condition (b) is equivalent to $\Delta_{12} = 0$; it is a classical result of multivariate normal distribution.

The univariate test is still carried out to assess the dependency between individual predictors and the response. The following procedure is used to obtain the pruned single-index sufficient reduction of the predictors.

(1) Proceed with a thorough investigation for a PFC model fitting:
   (a) select the appropriate basis function for the data,
   (b) select the variance structure,
   (c) fit the PFC model to collect $\hat{\Delta}$ and $\hat{\Gamma}$.
(2) Fit the p univariate linear regressions (2) to collect the p-values.
(3) Form the 'pruned' estimated sufficient reduction $\hat{\Gamma}_\alpha^T \hat{\Delta}_\alpha^{-1} X_\alpha$ of $\Gamma_1^T \Delta_{11}^{-1} X_{(1)}$ at the level of significance α.

Choices of the basis function in step (a) include polynomial, piecewise constant, piecewise polynomial, and Fourier bases.[4,16,19] When nonlinear relationships are suspected, high degree polynomial bases are suggested. A number of variance structures have been discussed in the literature, including the isotropic ($\Delta = \delta^2 I$), the anisotropic [$\Delta = \text{diag}(\delta_1^2, \ldots, \delta_p^2)$], the structured, and the unstructured Δ. A likelihood ratio test can be used to determine the appropriate structure.[16] In step (3), the choice of the level of significance α can be arbitrary, e.g. 5% or 10%. It can also be determined by cross-validation if a prediction performance is used to guide the selection, as detailed in the following section.

3.2. *Prediction-based selection*

It is possible that not all predictors selected at level α are relevant in predicting the response. We propose selecting the most important predictors by cross-validation.

Let $\alpha_1 < \alpha_2 < \cdots < \alpha_m = \alpha$ be a sequence of levels of significance, where α is the maximum to be considered, that is, 0.1. For each α_i, the mean square prediction error can be calculated in the usual fashion.[20] We split the n observations of data set into K sets $D_1, \ldots, D_K$ of roughly equal sizes $n_1, \ldots, n_K$. The set $D_{(-k)}$, that is, the set D with D_k held out, is used as a training set to estimate Γ and Δ. The p-value guided thresholding estimator $\hat{\Gamma}$ of Γ that yields the best prediction of the response is

$$\hat{\Gamma} = \underset{i \in \{1,\ldots,m\}}{\arg\min} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{y_j \in D_k} [y_j - \hat{E}(Y \mid \hat{\Gamma}_{\alpha_i}^T \hat{\Delta}_{\alpha_i}^{-1} x^{(k)})]^2, \tag{4}$$

where $x^{(k)}$ are observations from the testing set D_k. We adopted the prediction method of Adragni and Cook [19] which does not assume a model for $Y \mid X$ to obtain $\hat{E}(Y \mid \hat{\Gamma}_\alpha^T \hat{\Delta}_\alpha^{-1} x)$. Let $\hat{g}$ denote the estimated density function of $\Gamma^T \Delta^{-1} X$. The predicted response is obtained as

$$\hat{E}\{Y \mid \hat{\Gamma}_\alpha^T \hat{\Delta}_\alpha^{-1} \mathbf{x}\} = \sum_{i=1}^{n} w_i(\mathbf{x}) y_i = \sum_{i=1}^{n} \frac{\hat{g}(\hat{\Gamma}_\alpha^T \hat{\Delta}_\alpha^{-1} \mathbf{x} \mid y_i)}{\sum_{j=1}^{n} \hat{g}(\hat{\Gamma}_\alpha^T \hat{\Delta}_\alpha^{-1} \mathbf{x} \mid y_j)} y_i,$$

The prediction-based method proposed herein is reminiscent of the variable selection for the single-index model of Kong and Xia [21] where candidate models were obtained with all nonempty subsets of the p predictors and a nonparametric model was assumed. Our approach is similar except for the following: (i) the predictors are sorted and included incrementally in the sufficient reduction based on their p-values, and (ii) there is no kernel bandwidth to be estimated by cross-validation which is commonly needed by nonparametric methods.

3.3. *A connection between PFC and forward linear model*

Under restrictive assumptions on $\mathrm{X} \mid Y$, we establish herein a connection between the inverse regression (1) and a forward linear regression model through the following theorem.

THEOREM 3.3 *Assume that* (Y, X) *are jointly observed, and that a forward linear model* $Y = \beta^T \mathrm{X} + \varepsilon$ *holds as well as the isotropic model* (1) *with* $\mathrm{Var}(\mathrm{X}_y) = \delta^2 \mathrm{I}$, *a restricted basis function* $\mathrm{f}_y = (y - \bar{y})$, *and* $\lambda \in \mathbb{R}$. *Let* $\mathrm{Var}(Y) = \sigma_Y^2$. *Then, we have*

$$\beta = \frac{\lambda \sigma_Y^2}{\delta^2 + \lambda^2 \sigma_Y^2} \Gamma. \tag{5}$$

This theorem suggests that the single-index isotropic PFC with a linear basis function is equivalent to forward linear model in selecting relevant predictors when the predictors, conditionally on the response, are independent. The maximum likelihood estimator of the parameters in Equation (1) with an isotropic structure are given in the following theorem.

THEOREM 3.4 *Assume model* (1), *and let* $\tilde{\mathrm{C}} = \mathbb{X}^T \mathbb{Y}/n$, $\tilde{\sigma}_y^2 = \mathbb{Y}^T \mathbb{Y}/n$, *and* $\hat{\Sigma} = \mathbb{X}^T \mathbb{X}/n$, *where* $\mathbb{Y} = (y_1, \ldots, y_n)^{\mathrm{T}}$. *The maximum likelihood estimators of the parameters are* $\hat{\mu} = \bar{\mathrm{X}}$, $\hat{\lambda} = \|\tilde{\mathrm{C}}\|/\tilde{\sigma}_y^2$, $\hat{\Gamma} = \tilde{\mathrm{C}}/\|\tilde{\mathrm{C}}\|$, $\hat{\delta}^2 = (\mathrm{Tr}\{\hat{\Sigma}\} - n\|\tilde{\mathrm{C}}\|^2/\tilde{\sigma}_y^2)/p$.

There is no dimensionality issue when estimating the parameters under model (1), and estimation of the parameters can be performed with a decent sample size.

4. Pruning Multiple-Index PFC

We consider a PFC model as in Equation (1), except that $\Gamma \in \mathbb{R}^{p \times d}$, where $d \leq \min(r, p)$, $\lambda \in \mathbb{R}^{d \times r}$, and we continue to assume that $\Delta_{12} = 0$ as in Proposition 3.2. We still have $\Phi = \Gamma\lambda$. Here again, if X_i and Y are dependent, then ϕ_i, the ith row vector of Φ should be nonzero. Let $\hat{\Gamma}$ be an estimator without thresholding of Γ. With $\pi = (\pi_1, \pi_2, \ldots, \pi_p)^{\mathrm{T}}$ be vector of the p-values from the p fitted models, the *crude* p-value guided thresholding estimator $\hat{\Gamma}_\alpha$ is obtained as

$$\hat{\Gamma}_\alpha = J((\pi \otimes 1_d) \leq \alpha 1_{pd}) \circ \hat{\Gamma},$$

The following procedure is used to obtain $\hat{\Gamma}_\alpha$.

(1) Proceed with a thorough PFC model fitting:
   (a) select the appropriate basis function for the data,
   (b) select the variance structure,
   (c) estimate the dimension d of the reduction,
   (d) fit the PFC model to collect $\hat{\Delta}$ and $\hat{\Gamma}$.
(2) Fit the p univariate linear regressions (2) to collect the p-values.
(3) Form the 'pruned' estimated sufficient reduction $\hat{\Gamma}_\alpha^T \hat{\Delta}_\alpha^{-1} \mathrm{X}_\alpha$ of $\Gamma_1^T \Delta_{11}^{-1} \mathrm{X}_{(1)}$ at the level of significance α.

This procedure is similar to the previous given in Section 3.1, except that the dimension d of the reduction subspace must be estimated now. The estimation of d can be carried out using either a sequential likelihood ratio test or information criteria.[16] We provide more details about this sequential likelihood ratio test in Appendix A.4. A cross-validation method can also be used to estimate d.[19] In step (3), the prediction-based selection of Section 3.2 can be used.

5. Numerical studies

We applied our methodology to two data sets: the diabetes and the Big-Mac data sets. The diabetes data set was described in [22] and can be found in the R package `lars`.[23] It has 64 predictors including 10 main predictors and certain of their interactions with 442 observations. The Big-Mac data set [24] has 9 predictors with 45 observations.

We also conducted simulations to study the performance of our procedure in selecting the relevant predictors. The maximum significance level was set to 0.1. The selection was guided by prediction performance, and all tuning parameters were determined by leave-one-out cross-validation for real data sets and by 10-fold cross-validation in simulations.

5.1. *Data analysis*

Diabetes Data I: We consider the data set with the nine continuous predictors without the interactions. We removed the categorical covariate *sex* because our methodology is designed for continuous variables. The response variable is a measure of disease progression one year after baseline. It was centred and scaled to have unit variance.

We fit the models with $\mathrm{f}_y = y - \bar{y}$ and an isotropic covariance structure. The dimension d of the reduction is consequently one and was not estimated. We obtained the selected predictors along with the prediction error. Out of the following variables, *(age, bmi, map, tc, ldl, hdl, tch, ltg, glu)*, the hard-thresholding estimator of the PFC kernel matrix $\hat{\Gamma}$ was $(0.00, -0.49, -0.37, -0.20, -0.17, 0.33, -0.36, -0.47, -0.31)^T$. So all variables, except *age* were actively related to the response, and thus contributed to the reduction. With this sufficient reduction, the mean-squared prediction error was $0.51(se = 0.026)$.

This example follows Section 3.3 where the forward linear model and the single PFC yield equivalent results, ensuing in a fair comparison between our method and forward linear models. We considered `glmnet` of [25], a R package of the recently proposed coordinate descent algorithm implementation of the lasso for variable selection. The results yielded $\hat{\beta} = (0.00, 6.32, 2.11, 0.00, 0.00, -1.09, 0.00, 5.49, 0.00)^{\mathrm{T}}$ and a mean-squared prediction error of $0.52(se = 0.030)$.

The two methodologies disagreed on the variables *(tc, ldl, tch, glu)* although they are individually related to the response (e.g. p-values in the univariate models are less than the significance level). Understandably, with PFC, all predictors with a significant information about the response are included in the estimated reduction, while for forward linear models, predictors are included while controlling for terms previously included. The two methodologies should agree with each other when the predictors are statistically independent, conditionally on the response as in Theorem 3.3. Otherwise, sparse PFC would potentially select more active variables than the lasso.

Diabetes Data II: We continue with the diabetes data set with the larger set of predictors. We removed the categorical predictor and its interactions with other predictors. The remaining subset had 54 predictors.

We fit the model with a piecewise constant basis function with five slices. The choice of the number of slices was arbitrary. The guidance on the choice of the number of slices follows that of sliced inverse regression of Li [1]: 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. A likelihood ratio test of an isotropic structure against an unstructured Δ rejects the null hypothesis. Similarly, an anisotropic structured was rejected against an unstructured Δ. A significance level of 5% was used for these tests. With a model with an unstructured Δ, a likelihood ratio test was used to estimate the dimension d. A sequential test was used, following Cook and

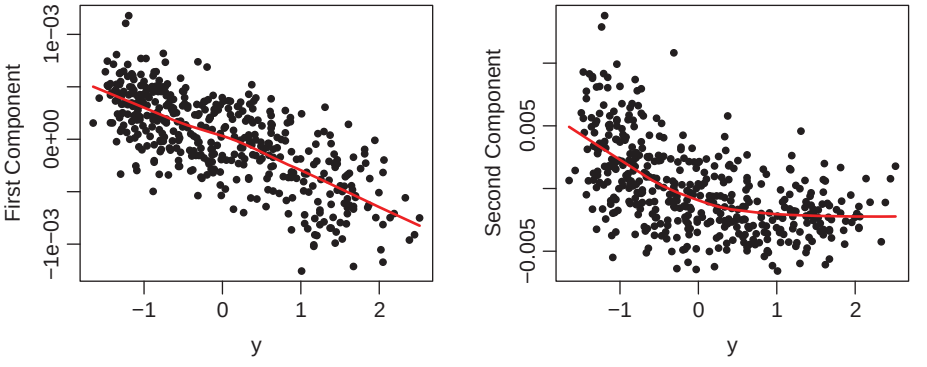

Figure 1. Plots of the sufficient reduction of the predictors against the response, using all 54 predictors.

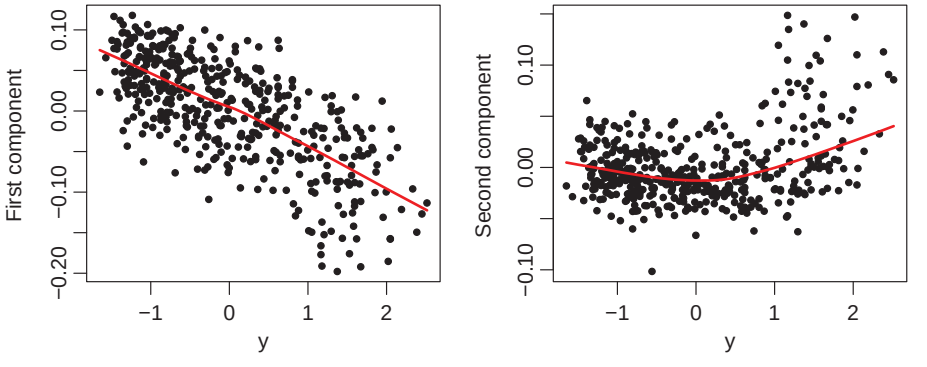

Figure 2. Plots of the sufficient reduction of the predictors against the response, using the ten selected predictors.

Forzani.[16] The hypotheses $H_0 : d = i$ against $H_1 : d > i$ were tested for $i = 0, 1, 2, \ldots$, until the first time the null hypothesis is not rejected. For this data set, we failed to reject the null hypothesis that $d = 2$ against $d > 2$, and thus the estimated dimension used was $\hat{d} = 2$. These steps were carried out using the R package `ldr` of [26]. The two plots in Figure 1 show the two components of the sufficient reduction against the response. These two components of the sufficient reduction are obtained as the two columns vectors of the data-matrix $\mathbb{X}\Delta^{-1}\Gamma$ where $\mathbb{X}$ represents the 442×54 data matrix of the predictors. The pruned sufficient reduction of the 54 predictors was made with two linear combinations of the following predictors and interaction terms ($bmi, map, tc, hdl, tch, ltg, glu, bmi^2, bmi : map, bmi : ltg$). The two components of the pruned sufficient reduction are plotted against the response on Figure 2. The two components of the pruned sufficient reduction can be used to effectively replace the 54 predictors X in a regression of Y on X. While the first component shows a straight line relationship with the response, the second component suggests a quadratic relationship. These two components can now be used in the modelling of the response.

We have observed a 10-fold cross-validation prediction error of 0.52 (se = 0.023). For comparison, the `glmnet` implementation of the lasso selected ($bmi, map, hdl, ltg, glu^2, bmi : map$) with a prediction error of 0.51 (se = 0.030).

Big-Mac Data: The Big-Mac data set [24] contains a continuous response variable that is the minimum labour to buy a Big Mac and fries in US Dollars, and nine predictors with 45 observations. The response Y was centred and scaled to have unit variance. These predictors

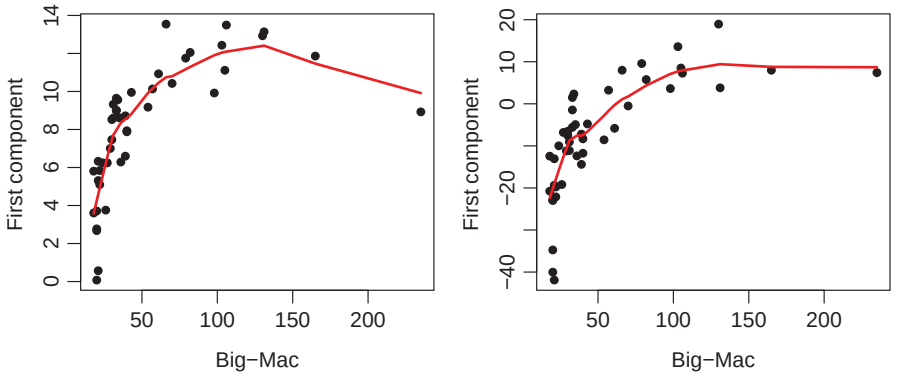

Figure 3. Sufficient reduction against the response of Big-Mac data set: all variables (left); after pruning (right).

are *Bread* the minimum labour to buy one kilogram of bread, *BusFare* the lowest cost of 10 kilometres public transit, *EngSal* the electrical engineer annual salary, *EngTax* the tax rate paid by electrical engineer, *Services* the annual cost of 19 services, *TeachSal* the primary teacher salary, *TeachTax* the tax rate paid by primary teacher, *V acDays* the average days of vacation per year, and *WorkHrs* the average hours worked per year.

A graphical exploration of the data showed that the relationships between the predictors and the response were nonlinear. And thus, we fit a PFC model with a piecewise discontinuous linear polynomial basis function with three slices and an unstructured Δ. A sequential likelihood ratio test was used to test $H_0 : d = i$ against $H_a : d > i$ for $i = 0, 1, \ldots$. The test failed to reject $d = 1$ against $d > 1$ (p-value $= 0.19$), and consequently a single-index PFC model was fitted. The p-value guided thresholding selected the following predictors *Bread, BusFare, EngSal, Service*, and *TeachSal*. The selection was guided by prediction performance via cross-validation. Figure 3 shows the inverse response plots [27] of the sufficient reductions against the response using all the variable (left plot) and after pruning the variables (right plot). Both plots show a clear nonlinear relationship between the sufficient reduction and the response. Further analysis showed that all the predictors selected by PFC were nonlinearly related to the response. The nonlinear relationships between the response and the predictors were captured through the use of the basis function. We should point out that several forward regression methods could capture these nonlinear relationships, including kernel regressions, nonlinear regressions, or by expanding the predictors using splines [20]; however, these methods are often easily hindered by high dimensionality of the predictors.

5.2. *Simulations*

We studied the variable selection behaviour of single index sparse isotropic PFC. Our interest was in calculating the true positive rate (TPR) and the false discovery rate (FDR). To describe TPR and FDR, we suppose that p hypotheses are tested with the following summary.

	Declared non-significant	Declared significant	Total
True H_0	U	V	$p - p_0$
True H_a	T	S	p_0
	$p - W$	W	p

The false discovery rate was defined as $E(V/W)$, the expected value of the proportion of the rejected null hypotheses which are erroneously rejected, while TPR is $E(T)/p_0$, the expected proportion of correctly rejected null hypotheses among the true alternative hypotheses. A seminal work of Benjamini and Hochberg (HB) [28] provides a procedure to control FDR. We compared the variable selection performance of our method to that of BH procedure, and to three implemented variants of the lasso in R: (i) `lars`, the least angle regression,[22] (ii) `elasticnet`,[29] and (iii) `glmnet`, the coordinate descent implementation.[25] These simulations were conducted for different ratios of the signal strength to the noise.

We considered scenarios where the relationship between the predictors and the response is linear as in Theorem 3.3. This is the simplest setup for a fair comparison with penalized forward linear regression methods.

We used three approaches to generate the data with n observations and p predictors: the *forward generation*, the *inverse generation*, and the *latent generation*. With forward generation, the observations of the predictors $\mathbb{X}$ were first generated independently of the response from the normal distribution with mean 0 and variance $\delta^2 \mathrm{I}_p$. Then for a given parameter β, we obtained $\mathbb{Y} = \mathbb{X}\beta + \sigma_{Y|\mathrm{X}}\mathbf{e}$, where the elements of $\mathbf{e}$ were obtained from the standard normal distribution.

With the inverse generation, the response observations $\mathbb{Y}$ were first generated from a given distribution, and centred to have sample mean zero. The observations on the predictors were then generated as $\mathbb{X} = \mathrm{G}\mathbb{Y} + \delta\mathbb{E}$ with specified G. Here, the elements of $\mathbb{E}$ were generated from the multivariate standard normal. The data-matrix $\mathbb{X}$ is made with two parts: the signal $\mathrm{G}\mathbb{Y}$ and the noise $\delta\mathbb{E}$.

The latent generation follows the inverse pattern. We first generated observations $\mathbb{U}$ from a latent variable with a specified distribution. The response observations $\mathbb{Y}$ were obtained as $\mathbb{Y} = \mathbb{U} + \sigma\mathbf{e}$ and centred to have sample mean zero. The predictors were obtained as $\mathbb{X} = \mathrm{G}\mathbb{U} + \delta\mathbb{E}$. This setup is similar to a single component factor analysis model [30] and has been used by [31] in the context of supervised principal components.

In each of these cases, the first p_0 predictors were linearly related to the response and the remaining $p - p_0$ predictors were not related to the response. Letting 1_k and 0_k represent k-vectors of 1s and 0s, respectively, the data sets were generated using the following setup.

Forward: $\beta = \tau \cdot (1^T_{p_0}, 0^T_{p-p_0})^{\mathrm{T}}$, $\sigma_{Y|\mathrm{X}} = 1$, $p = 50$, $p_0 = 4$, $n = 100$.

Inverse: $\mathrm{G} = \tau \cdot (1^T_{p_0}, 0^T_{p-p_0})^{\mathrm{T}}$, $\delta = 1$ and $Y \sim N(0, 9)$, $p = 100$, $p_0 = 50$, $n = 200$.

Latent: $\mathrm{G} = \tau \cdot (1^T_{p_0}, 0^T_{p-p_0})^{\mathrm{T}}$, $\delta = 1$, $U \sim N(0, 3)$, and $\sigma^2 = 0.05$, $p = 100$, $p_0 = 50$, $n = 200$.

The value of τ determines the strength of the signal in $\mathbb{X}$. We increased the signal to noise ratio by fixing the noise strength and increasing the signal intensity with $\tau = i/10, i = 1, \ldots, 12$). For each value of τ, a total of 1000 data sets were used to estimate TPR and FDR. The selection under our method was guided by prediction performance.

The results in Figure 4 suggest that the performance of the methods depends on the type of data generation. The p-value guided approach works fairly well in terms of both TPR and FDR. While `elasticnet` and `lars` mostly dominated sparse PFC for TPR, they also had much larger rates for FDR. It is observed that all the methods reached a plateau less than 90% under inverse and less than 80% under latent data-generation scenarios as the signal to noise increases. Overall, sparse PFC has a favourable performance compared to the variants of the lasso in terms of both TPR and FDR. Of the three linear model-based methods, `glmnet` seems to have the worst performance in TPR, and the best performance in FDR. Overall, the BH procedure performs better than the other methods in terms of FDR.

These plots illustrate the delicate balance often sought when dealing with variable selection: a high TPR is desired with a small FDR. While HB consistently yields the best FDR, it also gives the worst TPR. On the other hand, `elasticnet` and `lars` have the highest TPR, yet they yield the worst FDR. Our method seems to be cautiously in the middle.

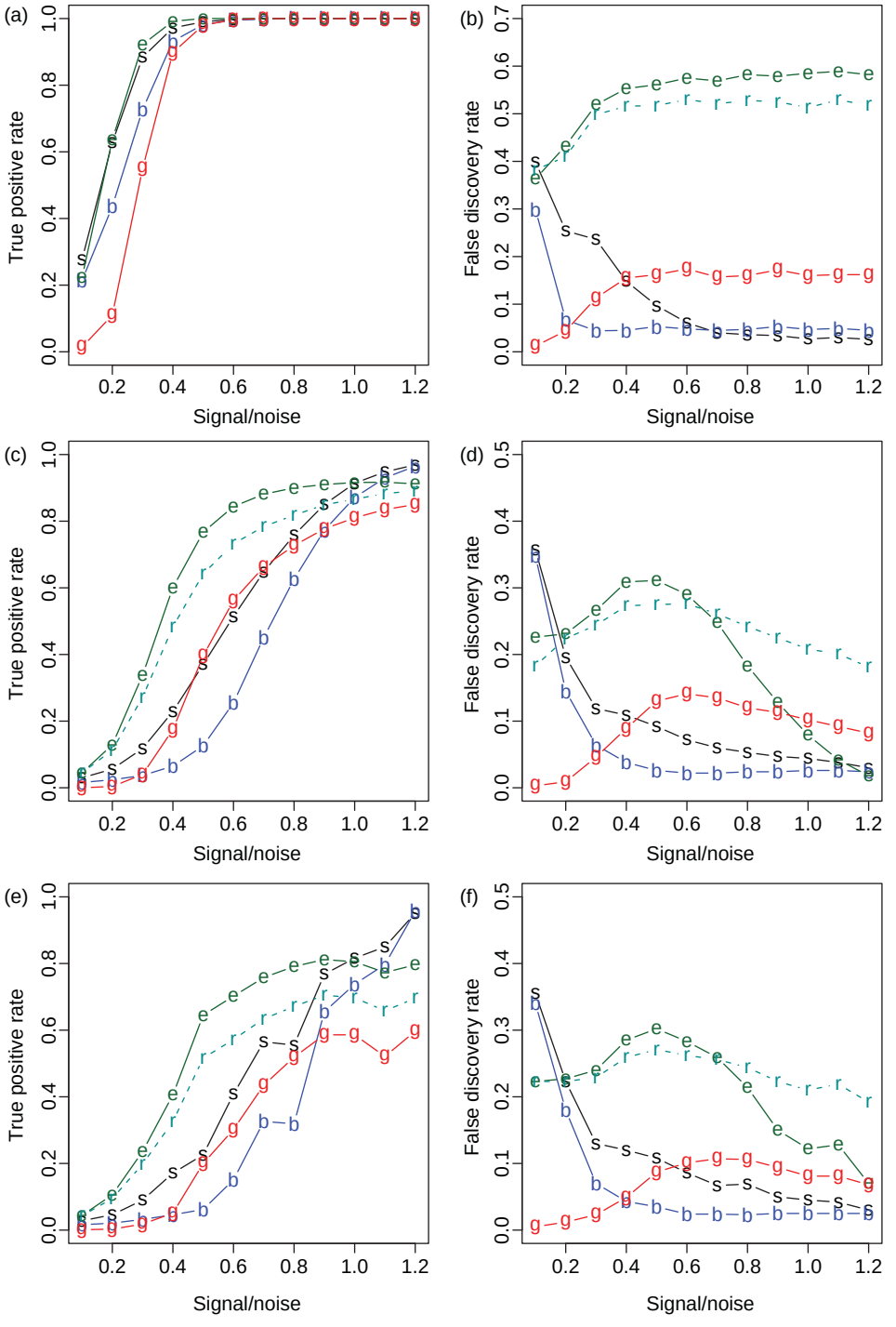

Figure 4. TPR and FDR against signal to noise ratio for sparse PFC ('s'), `glmnet` ('g'), `lars` ('r'), `elasticnet` ('e'), and Benjamini-Hochberg procedure ('b') with forward data generation (a. and b.), the inverse generation (c. and d.), and latent generation (e. and f.). Note that 'r' and 'e' overlap on plot a.

6. On coordinate descent algorithm for lasso

Sparse PFC is an inverse regression method that explicitly uses the randomness of the predictors. Forward regression methods, such as the lasso and its variants, use the randomness of the response and condition on the predictors. The forward linear model

$$Y = \beta^T \mathrm{X} + \varepsilon \tag{6}$$

is assumed to be true, and a sparse β is estimated. It is understandable that the lasso and its variants would perform best on forward generated data, and that sparse PFC would perform best under an inverse-generated data scheme. Simulation results on Figure 4(e). showed an under-performance of `glmnet` compared to the other methods in terms of TPR. With inverse and latent generated data schemes, `glmnet` reached a plateau much lower than the other methods considered. To further elucidate the behaviour of `glmnet` under the inverse and latent data-generation schemes, we looked into the coordinate descent algorithm [25] for the lasso, where the goal is to obtain

$$\hat{\beta} = \min_{\beta \in \mathbb{R}^p} \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \zeta \sum_{i=1}^{p} |\beta_j| \right], \tag{7}$$

where β_j is the jth element of β and ζ is a tuning parameter. The `glmnet` algorithm uses a cyclical coordinate descent approach. Let $\tilde{y}_i^{(j)}$ be the fitted value excluding the contribution of x_{ij}, the ith observation of X_j, and let $z = \sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)})/n$. The coordinate-wise update has the form

$$\tilde{\beta}_j \leftarrow \begin{cases} z - \zeta & \text{if } z > 0 \text{ and } \zeta < |z| \\ z + \zeta & \text{if } z < 0 \text{ and } \zeta < |z| \\ 0 & \text{if } \zeta \geq |z|. \end{cases} \tag{8}$$

We consider the population version of z. Let $\tilde{Y}^{(j)} = \sum_{i \neq j} \tilde{\beta}_i X_i$, and assume that the predictors $X_i, i = 1, \ldots, p$ and the response Y are scaled to have unit variance. The following proposition gives the expression of $\mathrm{Cov}(X_j, Y - \tilde{Y}^{(j)})$, the population version of z, under the forward and the inverse models.

PROPOSITION 6.1 *Assume that the forward linear model* (6) *holds. Then*

$$\mathrm{Cov}(X_j, Y - \tilde{Y}^{(j)}) = \beta_j + \sum_{i \neq j} E(X_i X_j)[\beta_i - \tilde{\beta}_i]. \tag{9}$$

Assume that both model (1) *holds with* $\Gamma \in \mathbb{R}^p$, $\lambda \in \mathbb{R}$, $\mathrm{f}_y = y - \bar{y}$, *and that Equation* (6) *also holds. Then*

$$\mathrm{Cov}(X_j, Y - \tilde{Y}^{(j)}) = \beta_j(\delta^2 + \lambda^2)[1 - \sum_{i \neq j} \tilde{\beta}_i \beta_i (\delta^2 + \lambda^2)]. \tag{10}$$

The coordinate-wise update (8) is based on result (9), where the predictors can be assumed orthogonal to have $E(X_i X_j) = 0$ for $i \neq j$. However, expression (10) suggests that $\mathrm{Cov}(X_j, Y - \tilde{Y}^{(j)})$ could be zero even if $\beta_j \neq 0$. This seems to explain the under-performance of `glmnet` with inverse and latent data-generation simulations, compared to sparse PFC. Note that our intent is not to criticize the `glmnet` algorithm, but rather to explain what we observed in the simulations.

7. Discussions

We presented the *p*-value guided thresholding for pruning a sufficient reduction of the predictors using PFC models. This methodology uses univariate PFC models to select predictors that are statistically related to the response based on a test statistic. Predictors related to the response are admitted in the sufficient reduction of the predictors. The methodology allows to prune away irrelevant predictors.

A single-index PFC model is restrictive because of the dimension of the reduction is assumed to be one; however, we showed a connection between a single-index PFC and a forward linear model for variable selection. We observed through simulations that `glmnet`, the coordinate descent implementation of the lasso in R, loses its edge in selecting the important predictors when these predictors are random. A theoretical justification of this latter fact was provided. We also described the methodology when the dimension of the reduction is greater than one.

The simulation results showed that the FDR performance of our methodology seems comparable to that of HB for strong signals. It is of our interest to study whether our method controls FDR as does HB procedure as suggested by Figure 4(b)–4(d).

Pruning a sufficient reduction can be reliably used to identify all active predictors for a use in modelling forward regression. The sufficient reduction can also be used in forward regressions of the form $Y = g\{\eta_1^T \mathrm{X}, \eta_2^T \mathrm{X}, \ldots, \eta_d^T \mathrm{X}; \epsilon\}$ is Y is continuous, or in logistic regressions of the form $\mathrm{logit}(Y) = g\{\eta_1^T \mathrm{X}, \eta_2^T \mathrm{X}, \ldots, \eta_d^T \mathrm{X}\}$.

A referee had requested a simulation study comparing our method to existing sparse sufficient dimension reduction methods: the algorithm of Li,[13] the methodology of Cook,[15] and the algorithm of Chen et al.[32] We have implemented all three methodologies and studied their behaviour in terms of variable selection. The choice of the tuning parameters of the algorithms of Li [13] and of Chen et al. [32] was based on a criterion similar to the Akaike information criterion. We found their behaviour very unsatisfactory when p is relatively large.

With respect to,[15] we found its comparison to this new method somewhat unfair: first, Cook's method was developed around SIR and was based on an asymptotic test where p is assumed to be relative small. Second, SIR behaves best when the data is close to a normal distribution, which is the setup for PFC. And third, using a piecewise constant basis function for PFC, the result for SIR and PFC are equivalent.[4] With our simulation setups where p was relatively large, Cook's method was unfairly overwhelmed and not worth reporting. Perhaps a comparative study of Li,[13] Cook,[15] and Chen et al. [32] could be interesting, since they all require $n > p$.

Acknowledgements

The authors thank the two referees and the associate editor for their constructive comments and suggestions that helped substantially improve this paper. The authors are grateful to Professor Anindya Roy for his earlier comments and suggestions, and to Dr. Heather L. White for proofreading the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Li KC. Slice inverse regression for dimension reduction. J Am Statist Assoc. 1991;86:316–327.
[2] Cook RD, Weisberg S. Discussion of Li (1991). J Am Statist Assoc. 1991;86:328–332.
[3] Li B, Wang S. On directional regression for dimension reduction. J Am Statist Assoc. 2007;102(479):997–1008.
[4] Cook RD. Fisher lecture - dimension reduction in regression (with discussion). Statist Sci. 2007;22:1–26.
[5] Donoho DL, Johnstone IM. Adapting to unknown smoothness via wavelet shrinkage. J Am Statist Assoc. 1995;90:1200–1224.

[6] Fryzlewicz P. Bivariate hard thresholding in wavelet function estimation. Statist Sin. 2007;17:1457–1481.
[7] McCabe GP. Principal variables. Technometrics. 1984;26:137–144.
[8] Cadima J, Jolliffe IT. Loading and correlations in the interpretation of principal components. J Appl Stat. 1995;22:203–214.
[9] Jolliffe I. Principal component analysis. New York: Springer; 1995.
[10] Jolliffe IT, Trendafilov NT, Uddin M. A modified principal component technique based on the lasso. J Comput Graph Stat. 2003;12(3):531–547.
[11] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. J Comput Graph Stat. 2006;15:265–286.
[12] Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc. Ser B (Methodol). 1996;58(1):267–288.
[13] Li L. Sparse sufficient dimension reduction. Biometrika. 2007;94:603–613.
[14] Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc: Ser B (Statist Methodol). 2010;72(1):3–25.
[15] Cook RD. Testing predictor contributions in sufficient dimension reduction. Ann Stat. 2004;32(3):1062–1092.
[16] Cook RD, Forzani L. Principal fitted components for dimension reduction in regression. Statist Sci. 2008;23(4):485–501.
[17] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Statist Soc: Ser B (Statist Methodol). 2008;70:849–911.
[18] Adragni KP. Independent screening in high-dimensional exponential family predictors' space. J Appl Stat. 2015;42(2):347–359.
[19] Adragni KP, Cook RD. Sufficient dimension reduction and prediction in regression. Philos Trans R Soc, Ser A. 2009;367(1906):4385–4405.
[20] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning - data mining, inference, and prediction. New York: Springer; 2001.
[21] Kong E, Xia Y. Variable selection for the single-index model. Biometrika. 2007;94(1):217–229.
[22] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat. 2004;32(2):407–499.
[23] R Development Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2013. Available from: http://www.R-project.org/.
[24] Enz R. Prices and Earnings Around the Globe. Zurich: Union Bank of Switzerland; 1991.
[25] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Statist Softw. 2010;33(1):1–22.
[26] Adragni KP, Raim A. An R software package for likelihood-based sufficient dimension reduction. J Statist Softw. 2014;61(3). Available from: http://www.jstatsoft.org/v61/i03.
[27] Cook RD. Regression graphics. New York: Wiley; 1998.
[28] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc, Ser B. 1995;57(1):289–300.
[29] Zou H, Hastie T. Regularization and variable selection via the Elastic Net. J R Statist Soc: Ser B (Statist Methodol). 2005;67(Part 2):301–320.
[30] Anderson TW. An introduction to multivariate statistical analysis. 3rd ed. Hoboken (NJ): Wiley; 2003.
[31] Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. J Am Statist Assoc. 2006;101(473):119–137.
[32] Chen X, Zou C, Cook RD. Coordinate-independent sparse sufficient dimension reduction and variable selection. Ann Stat. 2010;38(6):3696–3723.

Appendix

The following notations are adopted throughout this appendix. For an orthogonal matrix $\mathbf{A}$, we define $\mathrm{P}_{\mathbf{A}}$ to be the projection operator that projects onto span($\mathbf{A}$). With a semi-orthogonal Γ, we define Γ_0 to be its orthogonal completion such that $(\Gamma, \Gamma_0)^T(\Gamma, \Gamma_0) = \mathrm{I}_p$.

A.1. *Proof Theorem 3.3*

We assume that X and Y are centred to have mean 0, and consider the forward and the inverse models

$$Y \mid \mathrm{X} = \beta^T \mathrm{X} + \sigma_{Y \mid \mathrm{X}} \epsilon, \tag{A1}$$

$$\mathrm{X} \mid Y = \Gamma \lambda Y + \delta \boldsymbol{\varepsilon}, \tag{A2}$$

where $\epsilon \sim N(0,1)$ and $\varepsilon \sim N_p(0, \mathrm{I}_p)$. At the population level, let $\Sigma = \mathrm{Var}(\mathrm{X})$ and $\mathrm{C} = \mathrm{Cov}(\mathrm{X}, Y)$ and write from Equation (A1).

$$\beta = \Sigma^{-1}\mathrm{C}. \tag{A3}$$

Using the fact that $\Sigma = \mathrm{Var}(\mathrm{E}(\mathrm{X} \mid Y)) + \mathrm{E}(\mathrm{Var}(\mathrm{X} \mid Y))$, from Equation (A2), we have

$$\Sigma = (\lambda^2\sigma_Y^2 + \delta^2)\Gamma\Gamma^T + \delta^2\Gamma_0\Gamma_0^T,$$

$$\Sigma^{-1} = \frac{1}{\lambda^2\sigma_Y^2 + \delta^2}\Gamma\Gamma^T + \frac{1}{\delta^2}\Gamma_0\Gamma_0^T.$$

With $\mathrm{C} = \mathrm{E}[\mathrm{X}Y] = \mathrm{E}[\mathrm{E}(\mathrm{X}Y \mid Y)] = \mathrm{E}[\Gamma\lambda Y^2] = \Gamma\lambda\sigma_Y^2$, the expression of β in Equation (A3) becomes

$$\beta = \frac{\lambda\sigma_Y^2}{\lambda^2\sigma_Y^2 + \delta^2}\Gamma. \tag{A4}$$

A.2. *Proof Theorem 3.4*

The log-likelihood under model (1) using (X_i, y_i), $i = 1, \ldots, n$, and assuming $\sum_i^n y_i = 0$ is

$$\mathcal{L}(\mu, \Gamma, \lambda, \sigma) = -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log\delta^2 - \frac{1}{2\delta^2}\sum_{i=1}^{n}\|\mathrm{X}_i - \mu - \Gamma\lambda y_i\|^2.$$

The estimate of μ is obtained holding all parameters fixed as $\hat{\mu} = \bar{\mathrm{X}}$. With Γ fixed, the parameter λ is identifiable. Using $\mathrm{I}_p = \Gamma\Gamma^T + \Gamma_0\Gamma_0^T$, the partially maximized log-likelihood becomes

$$\mathcal{L} = -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log\delta^2 - \frac{1}{2\delta^2}\sum_{i=1}^{n}(\mathrm{X}_i - \bar{\mathrm{X}} - \Gamma\lambda y_i)^T(\Gamma\Gamma^T + \Gamma_0\Gamma_0^T)(\mathrm{X}_i - \bar{\mathrm{X}} - \Gamma\lambda y_i).$$

The estimation of λ is obtained using the following summand of the log-likelihood:

$$\mathcal{L}(\lambda) = -\frac{1}{2\delta^2}\sum_{i=1}^{n}(\mathrm{X}_i - \bar{\mathrm{X}} - \Gamma\lambda y_i)^T\Gamma\Gamma^T(\mathrm{X}_i - \bar{\mathrm{X}} - \Gamma\lambda y_i) = -\frac{1}{2\delta^2}\|\Gamma^T\mathbb{X}^T - \lambda\mathbb{Y}^T\|^2.$$

This function is partially maximized, holding Γ fixed by $\tilde{\lambda} = \Gamma^T\mathbb{X}^T\mathbb{Y}(\mathbb{Y}^T\mathbb{Y})^{-1}$. With $\hat{\Sigma} = \mathbb{X}^T\mathbb{X}/n$, $\hat{\Sigma}_{\mathrm{fit}} = \mathbb{X}^T\mathrm{P}_{\mathbb{Y}}\mathbb{X}/n$ and $\tilde{\sigma}_y^2 = \mathbb{Y}^T\mathbb{Y}/n$, the partially maximized log-likelihood is

$$\begin{aligned}\mathcal{L}(\Gamma, \hat{\mu}, \tilde{\lambda}, \delta^2) &= -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log\delta^2 - \frac{1}{2\delta^2}\mathrm{Tr}\{(\mathbb{X}^T - \Gamma\tilde{\lambda}\mathbb{Y}^T)^T(\mathbb{X}^T - \Gamma\tilde{\lambda}\mathbb{Y}^T)\} \\ &= -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log\delta^2 - \frac{n}{2\delta^2}\mathrm{Tr}\{\hat{\Sigma}\} + \frac{n}{2\delta^2}\mathrm{Tr}\{\hat{\Sigma}_{\mathrm{fit}}\mathrm{P}_\Gamma\}. \end{aligned} \tag{A5}$$

Holding δ^2 fixed, expression (A5) is maximized in Γ by the eigenvector of $\hat{\Sigma}_{\mathrm{fit}}$ corresponding to its largest eigenvalue. Since $\hat{\Sigma}_{\mathrm{fit}}$ is of rank one, we obtain $\hat{\Gamma} = \mathbb{X}^T\mathbb{Y}/\|\mathbb{X}^T\mathbb{Y}\|$ corresponding to the largest eigenvalue $\|\mathbb{X}^T\mathbb{Y}\|^2/(n\tilde{\sigma}_y^2)$. The MLE of δ^2 and λ are then

$$\hat{\delta}^2 = \frac{\mathrm{Tr}\{\hat{\Sigma}\} - \|\mathbb{X}^T\mathbb{Y}\|^2/(n\tilde{\sigma}_y^2)}{p} \quad \text{and} \quad \hat{\lambda} = \frac{\|\mathbb{X}^T\mathbb{Y}\|}{n\tilde{\sigma}_y^2}. \tag{A6}$$

A.3. *Proof Proposition 6.1*

We assume that model (A1) holds. Let $\tilde{Y}^{(i)} = \sum_{j\neq i}\tilde{\beta}_j X_j$, where $\tilde{\beta}$ is obtained while withholding variable X_i from X. Under forward model (A1),

$$\begin{aligned}\mathrm{Cov}(X_i, Y - \hat{Y}^{(i)}) &= \mathrm{E}(X_i Y) - \mathrm{E}(X_i\tilde{Y}^{(i)}) \\ &= \mathrm{E}[\mathrm{E}(X_i Y \mid X_i)] - \sum_{j\neq i}\mathrm{E}(X_i\tilde{\beta}_j X_j) \\ &= \mathrm{E}\left[X_i\sum_j \beta_j X_j\right] - \sum_{j\neq i}\mathrm{E}(X_i\tilde{\beta}_j X_j) \\ &= \beta_i\mathrm{E}(X_i^2) + \sum_{j\neq i}\beta_j\mathrm{E}(X_i X_j) - \sum_{j\neq i}\mathrm{E}(X_i\tilde{\beta}_j X_j) \\ &= \beta_i\mathrm{E}(X_i^2) + \sum_{j\neq i}(\beta_j - \tilde{\beta}_j)\mathrm{E}(X_i X_j), \end{aligned} \tag{A7}$$

and result (9) follows assuming that $\mathrm{Var}(X_i) = 1,\ i = 1, \ldots, p$.

We now assume that model (1) holds. Let $\Gamma = (\gamma_1, \ldots, \gamma_p)^T$. Then

$$\begin{aligned}
\mathrm{Cov}(X_i, Y - \tilde{Y}^{(i)}) &= \mathrm{E}(X_i Y) - \mathrm{E}\left(X_i \sum_{j\neq i} \tilde{\beta}_j X_j\right) \\
&= \mathrm{E}(\mathrm{E}(X_i Y \mid Y)) - \mathrm{E}\left(X_i \sum_{j\neq i} \tilde{\beta}_j X_j\right) \\
&= \mathrm{E}(Y\mathrm{E}(X_i \mid Y)) - \sum_{j\neq i} \mathrm{E}(X_i \tilde{\beta}_j X_j) \\
&= \lambda\gamma_i \mathrm{E}(Y^2) - \sum_{j\neq i} \tilde{\beta}_j \mathrm{E}(X_i X_j) \\
&= \lambda\gamma_i \sigma_Y^2 - \sum_{j\neq i} \tilde{\beta}_j \mathrm{E}[\mathrm{E}(X_i X_j \mid Y)] \\
&= \lambda\gamma_i \sigma_Y^2 - \sum_{j\neq i} \tilde{\beta}_j \mathrm{E}[\lambda^2 \gamma_i \gamma_j Y^2] \\
&= \lambda\gamma_i \sigma_Y^2 - \sum_{j\neq i} \tilde{\beta}_j \lambda^2 \gamma_i \gamma_j \sigma_Y^2. \qquad \text{(A8)}
\end{aligned}$$

With $\|\mathrm{C}\|^2 = \lambda^2\sigma_Y^4$, and using the result (3.3), we have

$$\begin{aligned}
\mathrm{Cov}(X_i, Y - \tilde{Y}^{(i)}) &= (\delta^2 + \lambda^2\sigma_Y^2)\beta_i - \sum_{j\neq i} \tilde{\beta}_j \lambda^2 \left(\frac{\delta^2 + \lambda^2\sigma_Y^2}{\lambda\sigma_Y^2}\right)^2 \beta_i\beta_j \\
&= (\delta^2 + \lambda^2\sigma_Y^2)\beta_i - \sum_{j\neq i} \tilde{\beta}_j \left(\frac{(\delta^2 + \lambda^2\sigma_Y^2)^2}{\sigma_Y^2}\right) \beta_i\beta_j \\
&= (\delta^2 + \lambda^2\sigma_Y^2)\beta_i \left[1 - \sum_{j\neq i} \left(\frac{\delta^2}{\sigma_Y^2} + \lambda^2\right) \tilde{\beta}_j\beta_j\right]. \qquad \text{(A9)}
\end{aligned}$$

Result (10) follows assuming that $\sigma_Y^2 = 1$.

A.4. *On the Choice of the Dimension d*

The dimension d of the dimension reduction subspace $\mathcal{S}$ is to be estimated. The following details are from Cook and Forzani.[16] At least two methods were suggested to estimate d: information criterion – Akaike (AIC) and Bayesian (BIC), and a sequential likelihood ratio test.

Let $\mathcal{L}_{d_0}$ be the log-likelihood value estimated for dimension $d = d_0$. With information criterion, the dimension d is selected to minimize over d_0 the function

$$\mathrm{IC}(d_0) = -2\mathcal{L}_{d_0} + h(n)g(d_0),$$

where $g(d_0)$ is the number of parameters to be estimated, and $h(n)$ is equal to $\log(n)$ for the Bayesian criterion and 2 for Akaike's. The term n is the number of observations.

The likelihood ratio test $\Lambda(d_0) = 2(\mathcal{L}_p - \mathcal{L}_{d_0})$ is used in a sequential manner for the hypothesis $H_0 : d = d_0$ against $H_a : d > d_0$ to choose d, starting with $d = 0$. The first hypothesized value of d that is not rejected is the estimated dimension d. The term $\mathcal{L}_p$ denotes the value of the maximized log-likelihood for the full model with $d_0 = p$ and $\mathcal{L}_{d_0}$ is the maximum value of the log-likelihood function when the dimension of the reduction is d_0. Under the null hypothesis $\Lambda(d_0)$ is distributed asymptotically as a chi-squared random variable with degrees of freedom $(r - d_0)(p - d_0)$ where r is the dimension of f_y and p is the number of predictors.

The expression of the log-likelihood depends on the structure of $\Delta = \mathrm{Cov}(X_y)$. There is no closed-form expression of the log-likelihood under an anisotropic model. We provide below the expressions for the isotropic and the unstructured models.

Let $\hat{\lambda}_i^{\mathrm{fit}}, i = 1, \ldots, d_0$ be the eigenvalues of $\hat{\Sigma}_{\mathrm{fit}} = \mathbb{X}^T \mathrm{P}_{\mathbb{Y}} \mathbb{X}/n$, as defined in Section 2, such that $\hat{\lambda}_1^{\mathrm{fit}} > \cdots > \hat{\lambda}_{d_0}^{\mathrm{fit}}$, and let $\hat{\lambda}_i, i = 1, \ldots, p$ be the eigenvalues of $\hat{\Sigma}$, the sample covariance of X. The estimate of σ^2 is $\hat{\sigma}^2 = (\sum_{i=1}^{p} \hat{\lambda}_i -$

$\sum_{i=1}^{d_0} \hat{\lambda}_i^{\text{fit}})/p$. The maximized log-likelihood under the isotropic model is

$$\mathcal{L}_{d_0} = -\frac{np}{2}(1 + \log(2\pi)) - \frac{np}{2}\log(\hat{\sigma}^2).$$

For the unstructured PFC, let $\hat{\Sigma}_{\text{res}} = \hat{\Sigma} - \hat{\Sigma}_{\text{fit}}$, and let $\hat{\omega}_1 > \cdots \geq \hat{\omega}_p$ and $\hat{V} = (\hat{v}_1, \ldots, \hat{v}_p)$ be the eigenvalues and the corresponding eigenvectors of $\hat{\Sigma}_{\text{res}}^{-1/2}\hat{\Sigma}_{\text{fit}}\hat{\Sigma}_{\text{res}}^{-1/2}$. Finally, let $\hat{K}$ be a $p \times p$ diagonal matrix, with the first d diagonal elements equal to zero and the last $p - d$ diagonal elements equal to $\hat{\omega}_{d+1}, \ldots, \hat{\omega}_p$. Then, the MLE of Δ is $\hat{\Delta} = \hat{\Sigma}_{\text{res}}^{1/2}\hat{V}(I + \hat{K})\hat{V}^\top\hat{\Sigma}_{\text{res}}^{1/2}$. [16, Theorem 3.1] The MLE of $\mathcal{S}_{Y|\mathbf{X}}$ is then obtained as the span of $\hat{\Delta}^{-1}$ times the first d eigenvectors of $\hat{\Sigma}_{\text{fit}}$. The maximized value of the log-likelihood is given by

$$L_{d_0} = \frac{np}{2}(1 + \log(2\pi)) - \frac{n}{2}\log|\hat{\Sigma}_{\text{res}}| - \frac{n}{2}\sum_{i=d_0+1}^{p}\log(1 + \hat{\omega}_i).$$