Comput Stat
DOI 10.1007/s00180-015-0611-9

ORIGINAL PAPER

Group-wise sufficient dimension reduction with principal fitted components

Kofi P. Adragni[1] · **Elias Al-Najjar**[1] · **Sean Martin**[1] · **Sai K. Popuri**[1] · **Andrew M. Raim**[2]

Received: 25 July 2014 / Accepted: 23 July 2015

Abstract Sufficient dimension reduction methodologies in regressions of Y on a p-variate X aim at obtaining a reduction $R(X) \in \mathbb{R}^d$, $d \leq p$, that retains all the regression information of Y in X. When the predictors fall naturally into a number of known groups or domains, it has been established that exploiting the grouping information often leads to more effective sufficient dimension reduction of the predictors. In this article, we consider group-wise sufficient dimension reduction based on principal fitted components, when the grouping information is unknown. Principal fitted components methodology is coupled with an agglomerative clustering procedure to identify a suitable grouping structure. Simulations and real data analysis demonstrate that the group-wise principal fitted components sufficient dimension reduction is superior to the standard principal fitted components and to general sufficient dimension reduction methods.

Keywords Inverse regression · Clustering · Prediction

1 Introduction

We consider a regression setting with a vector of p predictors X and a univariate response Y. We assume that the predictors are of a continuous type. When p is large, it is always worthwhile to reduce the dimensionality of X without losing any regression information. The reduction allows for a better visualization of the data, better modeling, mitigation of dimensionality issues in estimating the mean function $E(Y|X)$, and better

✉ Kofi P. Adragni
kofi@umbc.edu

[1] Department of Mathematics and Statistics, UMBC, Baltimore, MD 21250, USA

[2] United States Census Bureau, Washington, DC, USA

Published online: 07 August 2015

prediction of future observations. Replacing X by a lower dimensional function $R(X)$ is called dimension reduction. When $R(X)$ retains all the relevant information about Y, it is referred to as a *sufficient reduction*.

Formally, Cook (2007) defines a reduction $R : \mathbb{R}^p \to \mathbb{R}^d, d \leq p$, 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.

The reduction $R(X)$ captures all of the information about the response Y that is contained in X. 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. Thus, we can use an inverse regression to obtain a sufficient reduction of X and use this reduction in lieu of X in modeling $Y|X$.

Several methodologies have been developed to obtain a linear reduction $R(X) = \eta^T X$. The subspace $\mathcal{S}_\eta$ spanned by the columns of η is called a dimension reduction subspace. The reduction subspace with a minimal dimension d, called the central subspace (Cook 1998) and denoted by $\mathcal{S}_{Y|X}$, is often sought. Many of the methodologies developed to estimate the central subspace are nonparametric. These include sliced inverse regression (SIR; Li 1991), sliced average variance estimation (SAVE; Cook and Weisberg 1991), and direction regression (DR; Li and Wang 2007). Some recently developed methods, such as principal fitted components (PFC; Cook 2007) and likelihood acquired directions (LAD; Cook and Forzani 2009) are likelihood-based. Kernel-based methods such as the minimum average variance estimation (MAVE; Xia et al. 2002) and its variants have also been proposed.

In this paper, we develop methods for group-wise sufficient dimension reduction (SDR). We aim to partition the predictors into disjoint sets which are independent conditional on the response, while allowing predictors within a set to be conditionally dependent. These independent sets constitute the groups. Other authors have considered group-wise SDR; Li (2009) established a framework for grouped SDR, Li et al. (2010) developed a group-wise dimension reduction through the conditional mean function to obtain the group-wise central mean subspace, Guo et al. (2014) developed a group-wise dimension reduction using the so-called "direct sum envelope". These methodologies assume that the grouping information is known. In general, a known grouping structure can be imposed upon most SDR methods to obtain a group-wise SDR of the predictors given the response. However, to our knowledge no dimension reduction method has been devised to discover the group structure and estimate the sufficient reduction simultaneously.

The main contribution of this paper is a method for obtaining a sufficient dimension reduction of X when a grouping of the predictors is present or suspected, but the grouping information is unknown. We were unable to devise the methodology for any generic SDR method. However, we found that PFC models (Cook 2007) are well-suited to evaluate the grouping of the predictors. Consequently, this article is based upon utilizing a PFC model for a group-wise SDR.

The following is a brief description of a PFC model. Let X_y denote the p-vector random variable distributed as $X|(Y = y)$ and let $\bar{\mu} = E(X)$ and $\mu_y = E(X_y)$. The 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 $\mu_y - \bar{\mu}$ 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 $\Gamma^T\Gamma = \mathbf{I}_d$. We can then write $X_y \sim N(\bar{\mu} + \Gamma\boldsymbol{\nu}_y, \Delta)$ where $\boldsymbol{\nu}_y = \Gamma^T(\mu_y - \bar{\mu})$ is a function of y. Once the response values are observed, the unknown function $\boldsymbol{\nu}_y$ can be modeled as $\boldsymbol{\nu}_y = \beta(\mathrm{f}_y - E(\mathrm{f}_Y))$, where $\beta \in \mathbb{R}^{d\times r}$ is an unknown and unconstrained parameter of rank at most $d \leq \min\{p, r\}$, and $\mathrm{f}_y \in \mathbb{R}^r$ is a flexible set of basis function. The subsequent model, written as

$$X_y = \mu + \Gamma\beta\mathrm{f}_y + \Delta^{1/2}\boldsymbol{\varepsilon}, \tag{1}$$

where $\mu = \bar{\mu} - \Gamma\beta E(\mathrm{f}_Y)$ and $\varepsilon \sim N(0, I)$, is referred to as a PFC model. Typically, the function f_y 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 (2007) showed that a sufficient reduction of X is $\eta^T X$, where $\eta = \Delta^{-1}\Gamma$ is a function of the $p \times p$ covariance matrix Δ.

Under model (1), the grouping information is essentially embedded in Δ. For example, consider the gene expression data in the cardio data (Efron 2010). The data set is from a microarray experiment of $n_1 = 44$ healthy controls and $n_2 = 19$ cardiovascular patients. For each subject, measurements on $p = 20426$ genes were recorded. The initial data set was pruned via a t test to select genes that are differentially expressed. For illustration, the first 30 genes with the largest t statistics were selected. A PFC model was fitted. The 30 genes were treated as predictors and the response Y was categorical, representing either either a healthy control or a cardiovascular patient. The absolute correlation matrix obtained from the fitted covariance matrix is shown in Fig. 1a. Brighter yellow depicts a high correlation, and the blue depicts low or zero correlation. Rearranging the ordering of the genes (Fig. 1b) provides a visual guide for potential groupings; the yellow blocks indicate groups of genes which are conditionally dependent, while the blue stripes indicate conditional independence.

The information provided by these plots can be useful in understanding dependencies among these predictors, after conditioning on the response. Statistically, this

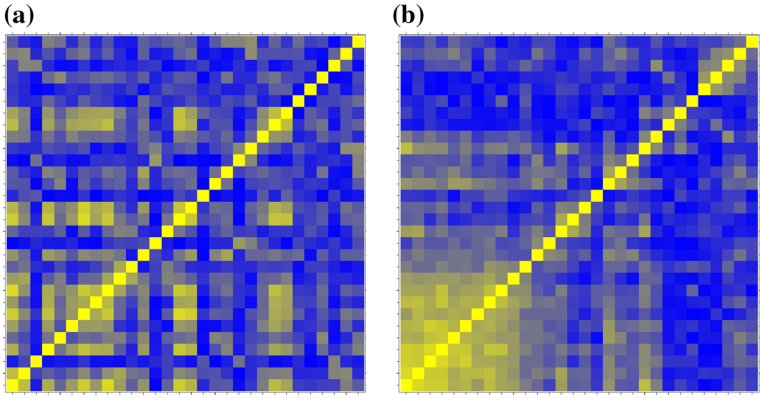

Fig. 1 Plot of the conditional correlation of the genes expression data

suggests that a structured Δ that allows for a conditional independence of some predictors should be used. A structured Δ removes any false dependency that could bias the sufficient dimension reduction, and will be more efficient than an unstructured Δ when the proposed group structure is true.

The remainder of this article is organized as follows. We begin Sect. 2 with a discussion of the case when the grouping structure is known by adapting the work of Li (2009) to PFC. Section 3 presents methods that could be used to identify the grouping structure for dimension reduction, and Sect. 4 provides a set of simulations evaluating the performance of our proposed approach. Several real data applications are presented in Sect. 5, followed by an extensive simulation study comparing the new procedure to existing SDR methods in Sect. 6, and some discussions in Sect. 7.

2 Sufficient dimension reduction with known grouping

The grouping information about the predictors may be known in various conditions. Li (2009) provided a general framework for sufficient dimension reduction when this grouping information is known, where two scenarios of group-wise sufficient dimension reduction were identified. Predictors may be partitioned based on conditions inherent to the data collection. Among these are experimental, biological, and clinical conditions. In these conditions, a *joint sufficient dimension reduction* can be sought. The partitioning can also be based upon the statistical conditional independence of the predictors. This is referred to as *grouped sufficient dimension reduction*. Theoretical results of Li (2009) apply to sufficient dimension reduction via PFC when the partitioning is known. We assume for now that this partitioning information is known and provide the sufficient dimension reduction of the predictors under the two scenarios.

2.1 Joint sufficient dimension reduction

We continue with the PFC model (1) where the sufficient dimension reduction is $\Gamma^T\Delta^{-1}X$. We assume that the set of p predictors X can be written as $X = (X^{(1)T}, \ldots, X^{(g)T})^T$, where $X^{(k)} = (X_1^{(k)}, \ldots, X_{p_k}^{(k)})^T$, $k = 1, \ldots, g$ represent vectors of p_k predictors, with $p = \sum_{k=1}^{g} p_k$. Let Γ, and Δ^{-1} be partitioned as

$$\Gamma = \begin{pmatrix} \Gamma_1 \\ \vdots \\ \Gamma_g \end{pmatrix}, \quad \Delta^{-1} = \begin{pmatrix} \Delta^{(11)} & \ldots & \Delta^{(1g)} \\ \vdots & \ddots & \vdots \\ \Delta^{(g1)} & \ldots & \Delta^{(gg)} \end{pmatrix} \tag{2}$$

The following proposition gives the joint sufficient dimension reduction under PFC model (1). It is analogous to Proposition 1 of Li (2009).

Proposition 1 *Let* $(X^{(1)T}, \ldots, X^{(g)T})^T$ *be a partitioning of the* p *predictors* X*, and assume that* Γ *and* Δ^{-1} *are partitioned accordingly as in expressions* (2)*. Let* $\eta_i = \Delta^{(1i)}\Gamma_1 + \cdots + \Delta^{(gi)}\Gamma_g, i = 1, \ldots, g$*. Then* $Y \perp\!\!\!\perp X | \Gamma^T\Delta^{-1}X \Rightarrow Y \perp\!\!\!\perp (X^{(1)}, \ldots, X^{(g)}) | (\eta_1^T X^{(1)}, \ldots, \eta_g^T X^{(g)})$.

That is, $(\eta_1^T X^{(1)}, \ldots, \eta_g^T X^{(g)})$ is a joint sufficient dimension reduction of $(X^{(1)T}, \ldots, X^{(g)T})^T$. In this joint sufficient dimension reduction, the components η_i depend on $\Delta^{(ij)}, i, j = 1, \ldots, g$. Thus, the estimation of the full conditional covariance Δ is necessary in estimating η_is. One may wonder whether η_i can be obtained marginally using $X^{(i)}$, by ignoring the conditional co-dependency with the other sets of predictors. Simulation experiments by Li (2009) and our own unreported simulations suggest a decent performance under weak co-dependency, but the performance degrades as the co-dependency gets stronger.

Nevertheless, in some cases the natural grouping of the predictors may not adhere to the above joint sufficient dimension reduction. The cognitive impairment study (Cornish et al. 2008) described by Li (2009) is an example. A sufficient dimension reduction of the predictors per domain is sought for each set of predictors $X^{(i)}, i = 1, \ldots, g$. A PFC model (1) can be fitted as

$$X_y^{(k)} = \mu_k + \zeta_k \beta_k \mathrm{f}_y + \Omega_k^{1/2} \epsilon, \quad k = 1, \ldots, g, \tag{3}$$

where $\zeta_k \in \mathbb{R}^{p_k \times d_k}$ is semi-orthogonal, and Ω_k is a covariance matrix. The dimensions $d_k, k = 1, \ldots, g$ may differ across the g models while the basis function f_y may be assumed to be the same. The structure of Ω_k may not be the same across the g models. The dimension p_k of $X^{(k)}$ is assumed to be greater than one, otherwise the predictor will be a singleton and no reduction is needed. For each set of predictors, a minimal sufficient reduction is obtained as $\zeta_k^T \Omega_k^{-1} X^{(k)}$. Letting $\eta_k = \Omega_k^{-1} \zeta_k$, a joint sufficient dimension reduction of $(X^{(1)T}, \ldots, X^{(g)T})^T$ is $(\eta_1^T X^{(1)}, \ldots, \eta_g^T X^{(g)})$.

Turning now to the estimation of the parameters, results from Cook and Forzani (2008) provide all the maximum likelihood estimators of ζ_k and Ω_k. We provide these estimators herein for completeness. We adopt the following notation. Generic Γ and Δ are used in lieu of ζ_k and Ω_k, respectively and the index k is dropped. For a given matrix A, $P_A = A(A^T A)^{-1} A^T$ is the orthogonal projection operator onto the subspace spanned by the columns of A. We consider the maximum likelihood estimation of the parameters in model (1) with an unstructured Δ. We assume that there are n observations on (Y, X), denoted by $(y_i, \mathbf{x}_i), i = 1, \ldots, n$. We denote by $\mathbb{F}$ the $n \times r$ data-matrix with ith row $(\mathrm{f}_{y_i} - \bar{\mathrm{f}})^T$. Denote by $\mathbb{X}$ the $n \times p$ predictor data-matrix with ith row $(\mathbf{x}_i - \bar{\mathbf{x}})^T$, where $\bar{\mathrm{f}} = \sum_{i=1}^n \mathrm{f}_{y_i}/n$ and $\bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$. Let $\widehat{\Sigma} = \mathbb{X}^T \mathbb{X}/n$ be the sample covariance matrix of X, and let $\widehat{\Sigma}_{\mathrm{fit}} = \mathbb{X}^T P_{\mathbb{F}} \mathbb{X}/n$ be the covariance matrix of the fitted values from a multivariate regression of X on f_y. We also let $\widehat{\Sigma}_{\mathrm{res}} = \widehat{\Sigma} - \widehat{\Sigma}_{\mathrm{fit}}$. Let $\widehat{V}$ and $\widehat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ be the matrices of the ordered eigenvectors and eigenvalues of $\widehat{\Sigma}_{\mathrm{res}}^{-1/2} \widehat{\Sigma}_{\mathrm{fit}} \widehat{\Sigma}_{\mathrm{res}}^{-1/2}$, and assume that the nonzero $\hat{\lambda}_i$'s are distinct.

The MLEs of parameters μ, β, $\mathcal{S}_\Gamma$ the subspace spanned by the columns of Γ, and Δ are respectively $\hat{\mu} = \bar{\mathbf{x}}$, $\hat{\beta} = (\widehat{\Gamma}^T \widehat{\Delta}^{-1} \widehat{\Gamma})^{-1} \widehat{\Gamma}^T \widehat{\Delta}^{-1} B$, $\mathcal{S}_{\widehat{\Gamma}} = \widehat{\Delta}^{1/2} \mathrm{span}\{\widehat{V}_d\}$, and $\widehat{\Delta} = \widehat{\Sigma}_{\mathrm{res}} + \widehat{\Sigma}_{\mathrm{res}}^{1/2} \widehat{V} \widehat{K} \widehat{V}^T \widehat{\Sigma}_{\mathrm{res}}^{1/2}$, where $\widehat{K} = \mathrm{diag}(0, \ldots, 0, \hat{\lambda}_{d+1}, \ldots, \hat{\lambda}_p)$, and $B = \mathbb{X}^T \mathbb{F} (\mathbb{F}^T \mathbb{F})^{-1}$. The columns of $\widehat{V}_d$ are the d eigenvectors corresponding to the largest eigenvalues of $\widehat{\Delta}^{-1/2} \widehat{\Sigma}_{\mathrm{fit}} \widehat{\Delta}^{-1/2}$. The dimension d of Γ is obtained by a likelihood ratio test (LRT), Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

2.2 A structured PFC model for a grouped SDR

For now, we assume that $\{X^{(i)}\}_{i=1}^{g}$ represent *groups* of predictors, where $X^{(i)} = (X_1^{(i)}, X_2^{(i)}, \ldots, X_{p_i}^{(i)})^T$ and each $X_m^{(i)} \in \mathbb{R}$. Our definition of *group* is provided by the following two conditions

(i) $X^{(i)}$ and $X^{(j)}$ are made of disjoint sets of predictors, and $X^{(i)} \perp\!\!\!\perp X^{(j)}|Y$, for $i \neq j$.
(ii) There are no sub-groups $X^{(i_1)}$ and $X^{(i_2)}$ of $X^{(i)}$ such that $X^{(i)} = (X^{(i_1)T}, X^{(i_2)T})^T$ and $X^{(i_1)} \perp\!\!\!\perp X^{(i_2)}|Y$.

Condition (i) states that the groups are disjoint sets of predictors and are statistically independent, given the response. Condition (ii) assures that any predictor within a group is assumed to be statistically dependent on at least one other predictor in the group, and that the group cannot be further decomposed into independent sets.

Because of the conditional independence of the groups given the response, the grouping of the predictors induces a structure in the covariance Δ of model (1). Two possible PFC model fitting approaches may be considered. The first is fitting g separate PFC models, one for each of the g groups of predictors $X^{(i)}$. The grouped sufficient dimension can be obtained as in Sect. 2.1. The second approach is a single PFC model with a structured Δ that is group-wise constrained. The following structured Δ is used.

$$\Delta = \begin{bmatrix} \Lambda_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Lambda_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Lambda_g \end{bmatrix} \tag{4}$$

The covariance Δ is block diagonal, and each block Λ is assumed unstructured. Under this model, the sufficient reduction is still $\Gamma^T \Delta^{-1} X$ with the central subspace given by $\mathcal{S}_{Y|X} = \Delta^{-1}\mathcal{S}_\Gamma$. This approach may be helpful when individual group reductions are not necessarily needed. The reduction could be passed to a forward model of the form $Y = f(\eta^T X, \epsilon)$ for model building and prediction.

We now turn to maximum likelihood estimation of the parameters in model (1) with a structured Δ. The maximum likelihood estimators of the parameters μ, β and $\mathcal{S}_\Gamma$ are as in Sect. 2.1. The maximum likelihood estimator of Δ maximizes the function

$$\mathcal{L}_d(\Delta) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Delta|) - \frac{n}{2}\mathrm{Tr}\left[\Delta^{-1}\widetilde{\Sigma}_{\mathrm{res}}\right] - \frac{n}{2}\sum_{i=d+1}^{p}\lambda_i\left(\Delta^{-1}\widehat{\Sigma}_{\mathrm{fit}}\right). \tag{5}$$

There is no closed-form solution to the estimators of Δ. In their development of the PFC model, (Cook and Forzani 2008, Appendix B) proposed an algorithm to help estimate the structured Δ. The algorithm assumes that Δ has a linear structure $\Delta = \sum_{i=1}^{m}\delta_i G_i$ with $m \leq p(p+1)/2$ where $G_1, \ldots, G_m$ are known real symmetric $p \times p$ linearly independent matrices, and that the elements of $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^T$ are functionally independent. Let $\widetilde{G} = \{\mathrm{vec}(G_1), \mathrm{vec}(G_2), \ldots, \mathrm{vec}(G_m)\}$. The following algorithm

solves $\partial \mathcal{L}_d(\Delta(\boldsymbol{\delta}))/\partial \boldsymbol{\delta} = 0$ iteratively. The starting point is the value that maximizes $\mathcal{L}_{\tilde{d}}$ when $d = r$, which can be found explicitly.

1. Set $\boldsymbol{\delta}_0 = (\delta_1^0, \ldots, \delta_m^0)^T = (\widetilde{G}^T \widetilde{G})^{-1} \widetilde{G}^T \mathrm{vec}(\widetilde{\Sigma}_{\mathrm{res}})$
2. Compute $\Delta_0 = \sum_{i=1}^m \delta_i^0 G_i$.
3. Compute until convergence, $k = 1, 2, \ldots,$

$$\mathrm{vec}(\Delta_k) = (\widetilde{G}^T \widetilde{G})^{-1} \widetilde{G}^T \Big[\mathrm{vec}(\widetilde{\Sigma}_{\mathrm{res}}) + \sum_{i=d+1}^p \lambda_i^{\Delta_{k-1}^{-1}} \mathrm{vec}\{\Delta_{k-1}^{1/2} \bar{u}_i^{\Delta_{k-1}^{-1}} (\bar{u}_i^{\Delta_{k-1}^{-1}})^T \Delta_{k-1}^{1/2}\} \Big], \tag{6}$$

with $\lambda_i^{\Delta_{k-1}^{-1}}$ and $\bar{u}_i^{\Delta_{k-1}^{-1}}$ being respectively an eigenvalue and eigenvector of $\Delta_{k-1}^{-1/2} \widehat{\Sigma}_{\mathrm{fit}} \Delta_{k-1}^{-1/2}$, and $\lambda_i^{\Delta_{k-1}^{-1}}$ are in the decreasing order for $i = d+1, ..., p$.

To model an anisotropic $\Delta = \mathrm{diag}(\delta_1^2, \ldots, \delta_p^2)$ we set $G_i = e_i e_i^T$, where $e_i \in \mathbb{R}^p$ contains a 1 in the ith position and zeros elsewhere, for $i = 1, \ldots, p$. For Δ with two blocks $\Delta_1 \in \mathbb{R}^{p_1 \times p_1}$ and $\Delta_2 \in \mathbb{R}^{p_2 \times p_2}$ for example, there are $0.5p_1(p_1+1) + 0.5p_2(p_2+1)$ matrices G_i. A matrix G corresponding to any entry δ_k^2 of Δ_1 can be obtained as $G = e_l e_m^T + J(l \neq m) e_m e_l^T$ where $J(.)$ stands for the indicator function.

3 Dimension reduction with unknown grouping

The knowledge of the grouping among predictors is often times unknown, and yet a grouping structure might be suspected or hypothesized. In the following, we seek to reveal the grouping structure of the predictors by using a clustering technique. The true Δ provides the main information about the grouping structure of the predictors, given the response Y. We consider using the crude maximum likelihood estimator of Δ from fitting an unstructured PFC model to help determine the suitable structure. We consider an agglomerative clustering method on the correlation matrix derived from $\widehat{\Delta}$.

3.1 Hierarchical agglomerative clustering

There is long list of clustering methods and algorithms in the literature (Hastie et al. 2001). Some of these methods are parametric, such as k-means, k-medoids, and Gaussian mixture. Others, like hierarchical agglomerative and divisive clustering, are nonparametric. Many algorithms exist within the agglomerative clustering methods, including single linkage, average linkage, and complete linkage variations. Agglomerative clustering methods require the specification of a distance metric to evaluate the closeness of data points. In this study, we arbitrarily consider the complete linkage agglomerative clustering algorithm. We use correlation coefficients to evaluate the closeness of predictors. As PFC models assume normality of X_y, the conditional correlation between two predictors determines whether or not they are conditionally independent.

An agglomerative clustering methodology is a bottom-up approach. It starts at the bottom level where each predictor constitutes its own group. This corresponds to an anisotropic structure with $\Delta = \text{diag}(\delta_1^2, \ldots, \delta_p^2)$. As the algorithm proceeds to higher levels, it builds groups or clusters using the correlation-based closeness criterion to group predictors together. At the final level, all of the predictors are placed in one single group as for an unstructured Δ.

Given Δ, the correlation matrix can be obtained as $R = D^{-1/2}\Delta D^{-1/2}$, where $D \in \mathbb{R}^{p\times p}$ is the matrix of the diagonal elements of Δ. The distance between two predictors X_i and X_j is $d_{ij} = 1 - |R_{ij}|$, where $R_{ij} = \rho(X_i, X_j|Y)$ is the correlation coefficient of the predictors X_i and X_j, given Y. In complete linkage clustering two closest clusters, say C_k and C_l, are merged using the following metric

$$d(C_k, C_l) = \max_{X_i\in C_k, X_j\in C_l} d_{ij}.$$

Following is a brief description of the complete linkage clustering.

1. Start with each variable in its own singleton cluster.
2. Merge the closest two clusters.
3. Repeat (2) until there is a single cluster.

As an illustration, consider Fig. 2 obtained with $p = 8$ predictors. When the cluster number is four, the clusters $\{X_1\}$, $\{X_2, X_3, X_4\}$, $\{X_5\}$, and $\{X_6, X_7, X_8\}$ are assumed to be independent, conditional on Y. For each possibility for the count of clusters $1, \ldots, p$, a PFC model can be fitted with a structure of Δ induced by the conditional independence of the clusters of predictors. With p predictors there will be p different partitions of the predictors, ranging from p singleton clusters corresponding to an anisotropic Δ, to a single cluster corresponding to an unstructured Δ. We assume that the true structure of Δ is identified by one of the sets of clusters; our goal is to identify it.

To proceed, we notice that for each structure, a PFC model can be fitted with the appropriate structured Δ. The p models can be compared with respect to a likelihood ratio test, an information criterion, or a prediction performance.

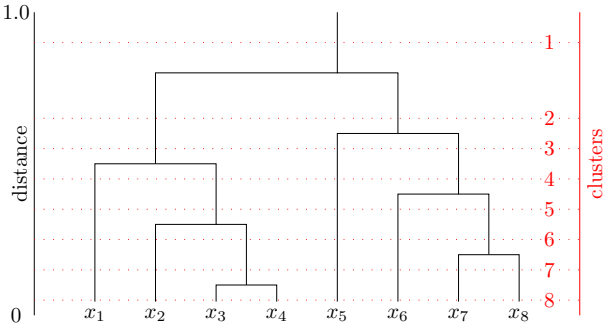

Fig. 2 Dendrogam with *eight* predictors

3.2 Model selection

With the clustering algorithm, p different PFC models are fitted, each with a specified structure of Δ. We can then proceed with the selection of the most suitable model. At least four methods can be used for the model selection: Akaike information criterion (AIC), Bayesian information criterion (BIC), likelihood ratio test (LRT), or prediction performance based on a cross-validation (CV).

Let $\mathcal{L}_M$ be the log-likelihood under the model M, where M is one of the p models with different Δ structure.

We will denote by M_k the model with k clusters, so that M_1 is the model with unstructured Δ and M_p is the model with anisotropic Δ. With AIC and BIC the model is obtained by minimizing $-2\mathcal{L}_M + h(n)g(M)$ over $M \in \{M_1, \ldots, M_p\}$, where $h(n)$ is equal to 2 for AIC and $\log n$ for BIC, and $g(M)$ is the number of parameters to be estimated in model M.

For a likelihood ratio test, we first note that predictors are added to clusters one at a time. Second, the models are fitted with the same basis function and identical dimension d of Γ. With these facts, the fitted PFC models are nested sequentially from bottom to top, that is

$$M_p \subset M_{p-1} \subset \cdots \subset M_2 \subset M_1.$$

Model M_k, $1 < k \leq p$, has a structured covariance Δ_k. It can be compared to model M_1 with an unstructured Δ_1. We will refer to the unstructured model as a *full* model, and to the structured model as a *reduced* model. We test the null hypothesis

$$H_0 : \Delta_k = \Delta_1 \text{ against } H_a : \Delta_k \neq \Delta_1. \tag{7}$$

The test statistic is $T^2 = 2(\mathcal{L}_{M_1} - \mathcal{L}_{M_k})$, where $\mathcal{L}_{M_1}$ is the log-likelihood from the PFC fit assuming an unstructured Δ_1 and $\mathcal{L}_{M_k}$ is the log-likelihood from the PFC fit assuming the structure Δ_k suggested by the clustering algorithm. Then under H_0, T^2 follows $\chi^2_{g(M_1)-g(M_k)}$.

We propose a sequential likelihood ratio test to select the appropriate model among the p candidates.

1. Fit an unstructured PFC model to obtain $\widehat{\Delta}$ and obtain the associated correlation matrix.
2. Use a clustering method on the correlation matrix to obtain the list of cluster assignments $C_1, \ldots, C_p$.
3. Sequentially test the hypothesis (7) from $k = p, \ldots, 2$ until the first time with $k = k_0$ when H_0 is not rejected.
4. If a rejection of H_0 occurred for $k_0 \geq 2$, then fit a structured PFC model with the grouping induced by C_{k_0} and obtain the sufficient reduction; otherwise, select the unstructured PFC model.

Lastly, a cross-validation method can be used to select the best predictive model. We used the prediction method devised by Adragni and Cook (2009). Specifically, normal inverse regression models for $X|Y$ have been inverted to provide a method for

estimating the forward mean function $E(Y|X)$ without specifying a model for the full joint distribution of (X, Y). With the observed response $(Y_1, \ldots, Y_n)^T$, the predicted value $\hat{Y}$ for a given observation $X = x$ can be obtained as

$$\widehat{\mathrm{E}}\,[Y|X = x] = \sum_{i=1}^{n} Y_i \frac{\hat{m}(x|Y_i)}{\sum_{i=1}^{n} \hat{m}(x|Y_i)}$$

where $\hat{m}(x|Y_i)$ is the estimated density of $\eta^T X|Y_i$. The best predictive model is the one with the smallest mean squared prediction error obtained by a cross-validation.

4 Numerical studies

We have performed several sets of simulations in order to compare and evaluate the relative effectiveness of AIC, BIC, and LRT in selecting the suitable grouping structure of the predictors. We refrained from using a cross-validation in these simulation studies as the predictive method was computationally expensive. However, it is used for comparison purposes in Sect. 5. For all data sets generated, we have set a pre-specified number of groups and a number of predictors in each group.

The data sets were generated as follows: the n response observations $\mathbb{Y} \in \mathbb{R}^n$ were generated from the normal distribution with mean 0 and variance 4. The matrix of basis functions $\mathbb{F} = (\mathbb{Y}, |\mathbb{Y}|) \in \mathbb{R}^{n\times 2}$ is column-wise centered to have sample mean 0. We obtained Γ as two eigenvectors of a $p \times p$ generated positive definite matrix. The covariance Δ was set to be block diagonal, with blocks

$$\Lambda_{(i)} = \sigma^2 \left(\rho_i \mathbf{1}_{p_i} \mathbf{1}_{p_i}^T + (1 - \rho_i) I_{p_i}\right),$$

where I_{p_i} is an identity $p_i \times p_i$ matrix, and $\mathbf{1}_{p_i} \in \mathbb{R}^{p_i}$ represents a column-vector of ones. The error terms $\mathbb{E} \in \mathbb{R}^{n\times p}$ were generated from a multivariate normal with mean 0 and variance Δ with $\sigma^2 = 2$. The data matrix of the predictors $\mathbb{X} \in \mathbb{R}^{n\times p}$ was obtained as $\mathbb{X} = \mathbb{F}\Gamma^T + \mathbb{E}$.

We ran three sets of simulations. In each of them, the number of observations n was increased from 50 to 400. The first simulation set involves two groups, the second involves three groups, and the last set involves seven groups of predictors. Given n, p, and ρ, a data set is generated. In all cases, we fit an unstructured PFC model with a cubic polynomial basis function. We assumed that the dimension d of the reduction is known to be two, and we obtained the unstructured estimate $\widehat{\Delta}$. The clustering procedure was used to find the different sets of clusters. We then used LRT, AIC and BIC to select the appropriate model. We determined the proportion of times the "true structure" was selected out of 1000 replications. By "true structure", we mean the trio made of the exact true and the two closest to the true. For example, on Fig. 2, if the true cluster corresponds to 3, then we use 2 and 4 to be the two closest. In some cases, the exact true structure is not found by the clustering algorithm. In those cases, the structure with the correct number of groups is selected together with its two closest.

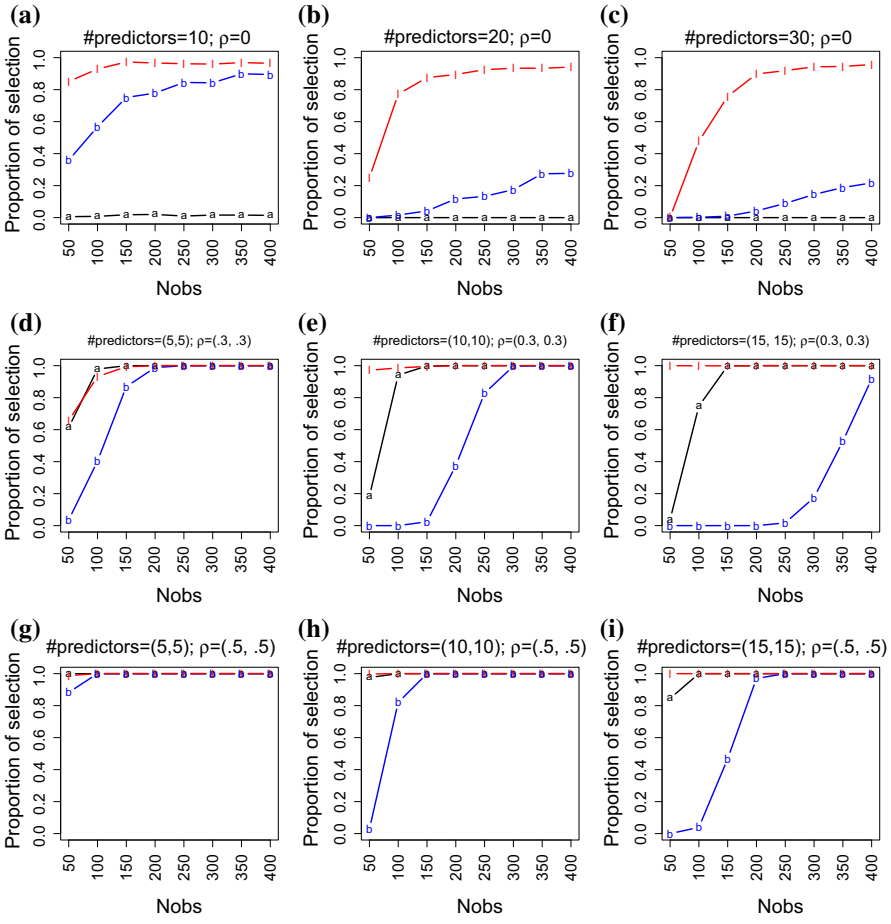

Fig. 3 Proportion of times the true model was selected using BIC (__b) , AIC (__a), and LRT (__l) with two groups. The p_i predictors in (p_1, p_2) have a conditional correlation ρ_i in $\boldsymbol{\rho} = (\rho_1, \rho_2)$

Simulation #1: Three different values of p were used: 10, 20, and 30. The p predictors are clustered into two groups of $p/2$ predictors. For each value of p, three correlation values were used: $\rho = 0, 0.3$ and 0.5. The results are on Fig. 3. Figure 3a–c show the anisotropic model, where the predictors are conditionally independent and each predictor constitutes its own cluster. The likelihood ratio test seems to be performing better than AIC and BIC, while BIC does somewhat fine when p is small. Of the three methods, AIC had the worst performance.

Figure 3d through 3i show that all three methods perform relatively well when a strong correlation is used and the sample size is large. Contrary to the first three plots, AIC showed a greater performance compared to BIC, while LRT seems better than the two.

Simulation #2: The p predictors are now clustered into three groups with $p = (3, 3, 4)$ for Fig. 4a, d, $p = (6, 6, 8)$ for Fig. 4b, e and $p = (10, 10, 10)$ for Fig. 4c, f. For

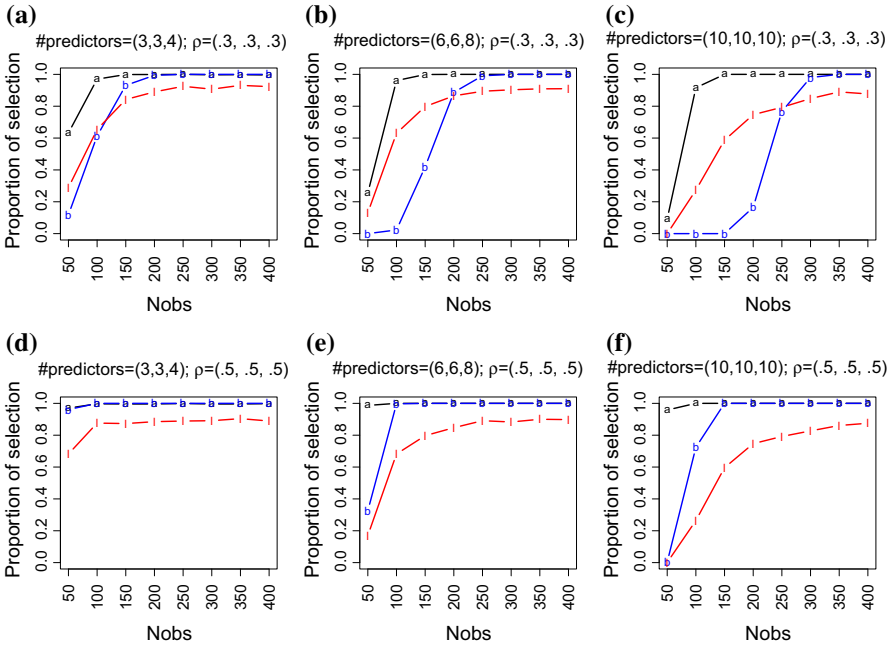

Fig. 4 Proportion of times the true model was selected using BIC (__b) , AIC (__a), and LRT (__l) with three groups. The conditional correlations within the groups are $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$

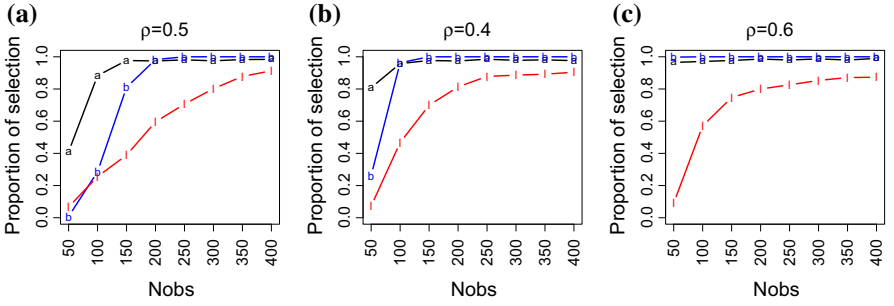

Fig. 5 Proportion of times the true model was selected using BIC (__b) , AIC (__a), and LRT (__l) with seven groups of three predictors. The conditional correlation between the predictors within the groups is ρ

each value of p, two correlation values were used: $\rho = 0.3$ for first row, and $\rho = 0.5$ for second row. The results on these figures showed that for large n and a relatively strong correlation, all three methods perform adequately well in selecting the model with the correct structure. However, with weaker correlations, AIC shows an overall better performance compared to LRT and BIC, especially for $n \geq 150$.

Simulation #3: We increased the number of groups to seven, each with three predictors. Three correlation values were used: $\rho = 0.3, 0.4$ and 0.6. The results are shown in Fig. 5a–c. Both AIC and BIC performed well in selecting the correct model; they both

improved as the correlation increased. Likelihood ratio test uniformly had the worst performance even when the correlation increased.

None of the three methods AIC, BIC, or LRT performed uniformly better compared than the others in detecting the true model structure. Overall, with large sample sizes, BIC performs adequately well when p is small; LRT gives the most stable performance although it is outperformed in some scenarios; AIC has its worst performance when the structure is anisotropic; otherwise, it dominates LRT and BIC when p is small.

5 Applications

We applied the methodology to a number of data sets. Unlike the simulation studies, we used the mean squared prediction error obtained by leave-one-out cross-validation to select the structure of the conditional covariances and evaluate the performance of the procedure. For the data sets considered, our goal was to obtain the sufficient reduction of the predictors under the PFC model by obtaining, if relevant, a structured Δ that best describes the conditional independence of the predictors. The best model was obtained as the one that yields the smallest mean squared prediction error. The following data sets were used.

Baseball data: The set contains salary information for 337 Major League Baseball players who are not pitchers and played at least one game during both the 1991 and 1992 seasons. The purpose of the study is to determine whether a baseball player's salary is a reflection of his offensive performance. For each player, the salary from the 1992 season along with 12 offensive statistics from the 1991 season were collected. In addition to these variables, there are four indicator variables which identify free agency and eligibility for arbitration. The data set was retrieved from Boos (2014).

Big-Mac data (Enz 1991): The data set contains a continuous response variable that is the minimum labor to buy a Big Mac and fries in US dollars, and nine predictors with 45 observations. These predictors are $Bread$ the minimum labor to buy one kilogram of bread, $BusFare$ the lowest cost of 10 km 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, $VacDays$ the average days vacation per year, and $WorkHrs$ the average hours worked per year.

Cardio data: The data set is from the Cardio Study (Efron 2010), a microarray experiment comparing $n_1 = 44$ healthy controls to $n_2 = 19$ cardiovascular patients, each measured on $p = 20426$ genes. The predictor matrix X is the doubly standardized expression data; that is, the columns are standardized by individually subtracting the mean and dividing by the standard deviation of each column, and the rows are similarly standardized. The response is categorical, taking values "healthy" and "patient".

Diabetes data: (Efron et al. 2004) The set contains blood and other measurements in diabetics patients. Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of $n = 442$

diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.

Pollution data: (McDonald and Schwing 1973) This data set has 15 independent variables and a measure of mortality on 60 US metropolitan areas in 1959–1961. It was obtained from Boos (2014).

Pyrimidine data: The data set contains structural information on 74 2,4-diamino-5-substituted benzyl pyrimidines used as inhibitors of dihydrofolate reductase in *E. coli*. There are three positions where chemical activity occurs and nine attributes per position leading to 27 total predictors. The response is the quantitative structure-activity relationships (QSAR) of the inhibition of dihydrofolate reductase (DHFR) by pyrimidines. The set was obtained from Boos (2014).

The response Y, when continuous, was centered and scaled to have unit variance. A piecewise constant basis function f_y with five slices is used for all of the data sets considered, except for the cardio data. For the cardio data, let C_1 and C_2 denote "healthy" and "patient" groups. We set $\mathbf{f}_y$ to be $J(y \in C_1) - n_1/n$, where $J(.)$ is the indicator function, $n_1 = 44$ and $n = 63$. For details on the choice of basis functions, please see *basis functions* in the R package ldr (Adragni and Raim 2014b).

In all six cases, the dimension d of the reduction was estimated by a likelihood ratio test using all of the data. A complete linkage agglomerative clustering in the R package cluster (Maechler et al. 2015) was used. The mean squared prediction error was obtained by leave-one-out cross-validation.

Of the six data sets, the cardio data is notable because of its size. Obviously, the dimension p of the predictors is very large, compared to the sample size. We verified and observed that a large portion of these predictors retain no regression information about the response. To remove these irrelevant variables, we proceeded with a screening procedure based a test statistic following Adragni (2015). We fit p separate linear models of $X_i = \beta_0 + \beta_{1i}\mathbf{f}_y + \epsilon$ for $i = 1, \ldots, p$, and tested the hypotheses $H_{0i} : \beta_{1i} = 0$ against the alternative $H_{ai} : \beta_{1i} \neq 0$. We collected all of the p values of these tests and performed the multiple testing procedure of Benjamini and Hochberg (1995) to control the false positive rate in multiple comparisons. This yielded a set of $p_0 = 143$ selected predictors having the strongest dependence with the response. Since the number of selected predictors was still far greater than the sample size, a direct unstructured PFC model could not be fitted. We proceeded by obtaining a sparse estimate of the unstructured Δ using the graphical lasso of Friedman et al. (2007) before applying the group-wise dimension reduction procedure.

We summarize the results of the data analysis in Fig. 6 and Table 1. Fig. 6a–f provide the mean squared prediction error for the p different models induced by the clustering procedure. Except for the diabetes data set in Fig. 6d, the mean squared error is minimal for a structured model obtained by cross-validation. The other three methodologies, AIC, BIC and LRT, were also used to select the best model. Their results are added to Table 1. Under the columns corresponding to AIC, BIC and LRT are the mean squared prediction errors corresponding to their respective selected model; the number of clusters corresponding to the selected model is in parentheses.

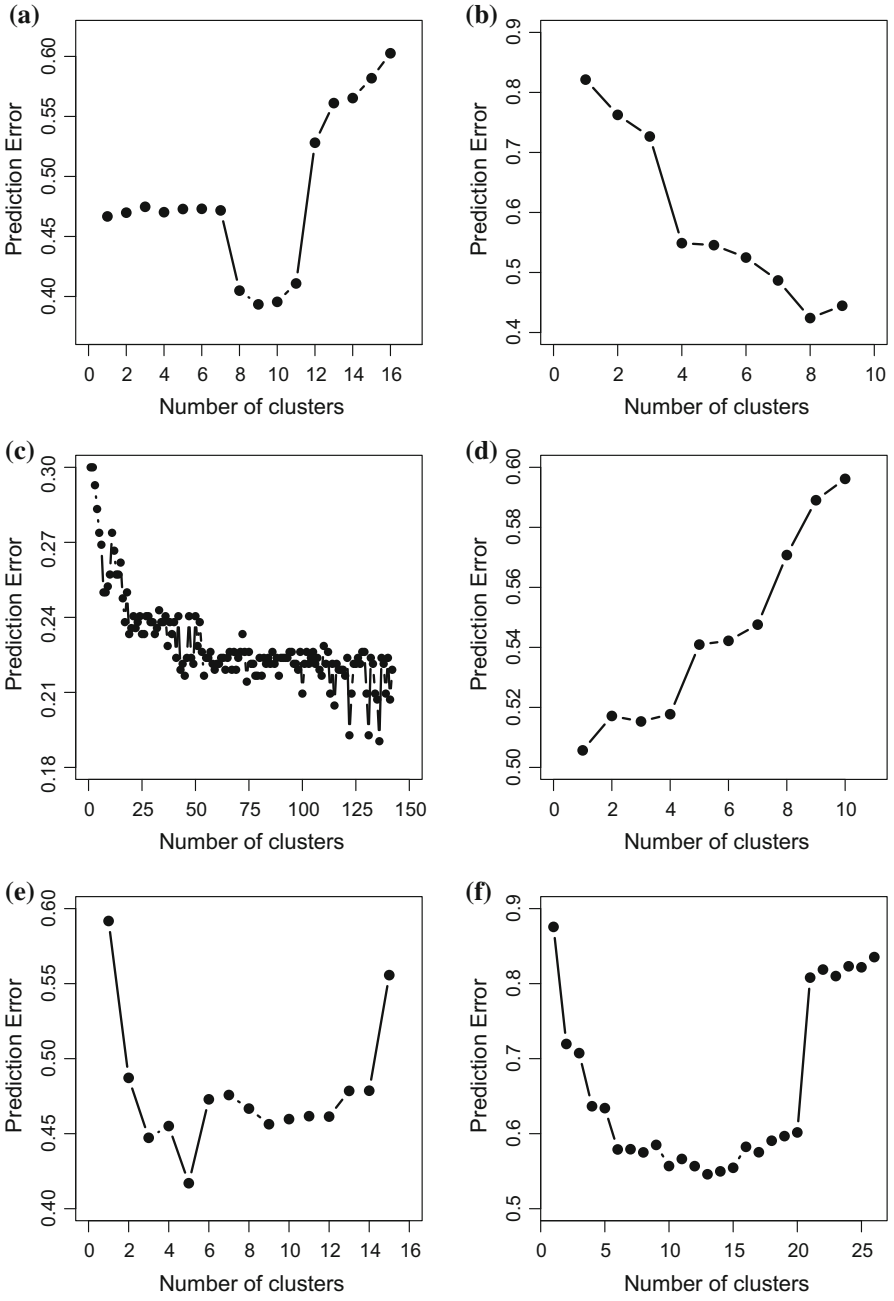

Fig. 6 Mean squared prediction error against the number of conditionally independent clusters of predictors. **a** Baseball. **b** Big-Mac. **c** Cardio. **d** Diabetes. **e** Pollution. **f** Pyrimidine

Table 1 Mean squared prediction error for unstructured PFC model (Unstr.), model selected by cross-validation (CV), and for the model selected by AIC, BIC and LRT

Data sets $[n, p]$	Unstr.	CV	AIC	BIC	LRT
Baseball [337, 16]	0.47	0.39 (9)	0.47 (1)	0.47 (1)	0.47 (1)
Big-Mac [45, 9]	0.82	0.42 (8)	0.76 (2)	0.72 (3)	0.76 (2)
Cardio [63 ,143]	0.30	0.19 (136)	0.30 (1)	0.30 (1)	0.26 (12)
Diabetes [442, 10]	0.51	0.51 (1)	0.51 (1)	0.51 (1)	0.51 (1)
Pollution [60, 15]	0.70	0.46 (5)	0.70 (1)	0.70 (1)	0.70 (1)
Pyrimidine [74, 26]	0.88	0.55 (13)	0.88 (1)	0.72 (2)	0.88 (1)

The number of clusters for the selected model is in parentheses

The diabetes data set has 442 observations for ten predictors, and all four methods agreed on the structure of Δ. For the baseball data set, there was a slight gain in prediction using nine clusters instead of one cluster suggested by AIC, BIC and LRT. For the other four data sets, the sample sizes are too small for AIC, BIC and LRT to be trustworthy, and cross-validation would be relied upon.

6 Group-wise SDR with other methods

We have considered a few existing SDR methods for group-wise sufficient dimension reduction. If the grouping information of the predictors is known, Li's method (Li 2009) applies to most sufficient dimension reduction methods directly. However, when this grouping information is unavailable, PFC seems best suited to help evaluate the conditional independence of predictors, and also to evaluate the subsequent models via likelihood-based procedures.

The method devised in this article to obtain the group-wise SDR in the absence of grouping information is a two-stage procedure. The first stage is to obtain the sets of clusters of conditionally independent predictors based on $\Delta = \text{Cov}(X|Y = y)$. The second stage is to evaluate the models under the different structures of Δ induced by the clustering. In our investigation, we have considered developing a similar procedure for the following methods: SIR of Li (1991), SAVE of Cook and Weisberg (1991), LAD of Cook and Forzani (2009), and MAVE of Xia et al. (2002). However, major issues arose. For SIR and SAVE, normality is not assumed. Thus, zero correlation among the predictors does not imply independence. Even if we assume normality, it is not obvious to us how to incorporate the structured conditional covariance into the estimation of the central subspace. MAVE is built upon a local regression of $Y|X$. It does not make any assumption on X, which is essentially treated as fixed. LAD is a likelihood-based method that relies on the normality assumption of $X|Y$. It estimates the central subspace using the conditional mean $E(X|Y)$ and conditional variance function $\text{Var}(X|Y)$ that are both assumed to be dependent on Y. For LAD, the conditional independence could change as Y varies in its sample space, which carries an extra layer of complication. We were unable to devise modifications of the four aforementioned SDR methods that would incorporate the conditional independence

of predictors into the estimation of the central subspace. Nevertheless, we conducted a simulation study to compare the performance of group-wise PFC to these four methods.

Three simulations are herein reported. For each of the three simulations, 100 data replications were generated. For each data set, the total number of predictors is p, and the number of observations was n. The predictors are grouped into three sets with p_1, p_2, and p_3 predictors where $p_1 + p_2 + p_3 = p$. Group-wise PFC, LAD, MAVE, SAVE and SIR were used to estimate the central subspace, which is compared to the true population central subspace.

The data sets were obtained as follows: we first generated the n response observations $\mathbb{Y}$ from the normal distribution with mean 0 and standard deviation σ. The predictors, given the response, were obtained as $\mathbb{X} = f(\mathbb{Y})G^T + \delta\mathbb{E}$, where $f(\mathbb{Y}) \in \mathbb{R}^{n\times r}$ are provided, $G \in \mathbb{R}^{p\times r}$ is a fixed semi-orthogonal matrix, and $\mathbb{E} \in \mathbb{R}^{n\times p}$ are obtained from a normal with mean zero and a structured covariance Δ, and $\delta \in \mathbb{R}$. The covariance Δ was obtained as a block diagonal matrix with three blocks Δ_{11} for the first five predictors, Δ_{22} for the next five, and Δ_{33} for the last five. The blocks were obtained as

$$\Delta_{ii} = \sigma^2\left(\rho_i \mathbf{1}_{p_i}\mathbf{1}_{p_i}^T + (1-\rho_i)I_{p_i}\right).$$

In all cases, we used $n = 300$, $p_i = 5, i = 1, 2, 3$, $(\rho_1, \rho_2, \rho_3) = (0, 0.4, 0.8)$ Following are the specificities for each simulation. For *simulation 1:* $\sigma = 3$, $\delta = 1$, $f(\mathbb{Y}) = 2(\mathbb{Y} - \overline{\mathbb{Y}})$; for *simulation 2:* $\sigma = 2$, $\delta = 2$, $f(\mathbb{Y}) = 2(\mathbb{Y}^3 - \overline{\mathbb{Y}^3})$; and for *simulation 3:* $\sigma = 2$, $\delta = 2$, $f(\mathbb{Y}) = 2(\mathbb{Y}^2 - \overline{\mathbb{Y}^2})$. In all three cases $r = 1$ and $f(\mathbb{Y}) \in \mathbb{R}^n$.

The true central subspace is spanned by the column of $\Delta^{-1}G$. All of the simulations were carried out in the statistical software R (R Core Team 2015). We obtained the estimates of the central subspace via SIR and SAVE using the R package `dr` of Weisberg (2002). The central subspace was also estimated by LAD using the package `ldr` of Adragni and Raim (2014a). We implemented the algorithm of MAVE following Xia et al. (2002). Figure 7 shows the angle in degrees between the true and the estimated central subspaces, both of dimension one in $\mathbb{R}^{15}$. As a reference, the average angle between two randomly generated directions in $\mathbb{R}^{15}$ is 77°.

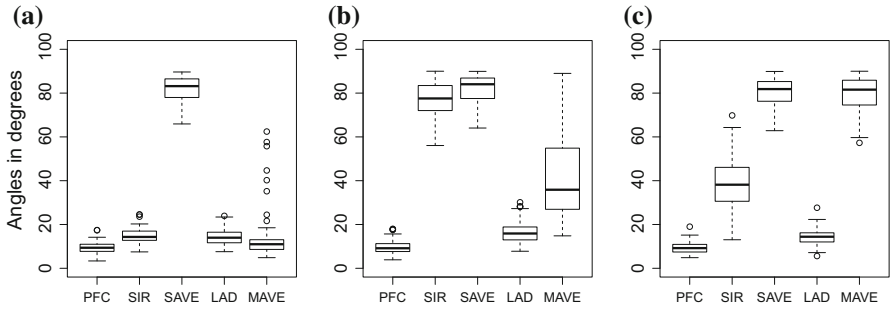

Fig. 7 Angle in degrees between the true and the estimated central subspaces. **a** Simulation 1. **b** Simulation 2. **c** Simulation 3

Overall, the simulation results show the superior performance of group-wise PFC compared to the other four listed SDR methods. Its performance is followed by that of LAD in all three simulations. SIR performed relatively well in simulation 1 shown in Fig. 7a, whereas SAVE gave the worse performance. The first reduction direction of both SIR and SAVE also failed to capture the true direction in simulation 2 shown on Fig. 7b and in simulation 3 shown on Fig. 7c. It is known that SIR fails to find the reduction directions when $Y = f(B^T X) + \epsilon$, where f is a symmetric function of $B^T X$, and SAVE was designed to handle this case. These simulations provide another look at cases where both SIR and SAVE seem to be failing. Finally, MAVE performed well as expected when X is linearly related to Y, but its performance worsens when X is related to Y^2 or to Y^3. This behavior of MAVE was not a surprise since the data were generated under an inverse regression setting. Intrinsically, the data do not follow the model upon which MAVE directions were obtained.

7 Concluding remarks

We have developed a procedure to obtain a group-wise sufficient dimension reduction of the predictors when the grouping information is unknown. The methodology uses a clustering algorithm based on the correlation structure of the conditional predictors X_y. Information criteria and a likelihood ratio test were investigated for use in finding the proper structure.

Our simulation studies showed that none of the three methods AIC, BIC, and LRT, uniformly performed better than others in selecting the true structure. However, at least one of the three methods would perform adequately well when the sample size is large enough. In all cases, LRT exhibits greater stability compared to AIC and BIC across the different structures of Δ and dimensionality p. Smaller sample sizes tend to induce spurious high correlations between the predictors; consequently all three methods tend to select an unstructured Δ. In these cases of smaller sample sizes, a leave-one-out cross-validation is recommended.

The application of the methodology showed great advantages for prediction of the response; the prediction errors were substantially reduced for several data sets. Finally, we note that a cross-validation method can always be used regardless of the dimension p and the sample size n. Its main drawback is that it can be computationally expensive when n is large.

References

Adragni KP (2015) Independent screening in high-dimensional exponential family predictors' space. J Appl Stat 42(2):347–359

Adragni KP, Cook RD (2009) Sufficient dimension reduction and prediction in regression. Philos Trans R Soc Ser A 367:4385–4405

Adragni KP, Raim A (2014a) An R software package for likelihood-based sufficient dimension reduction. J Stat Softw, 61(3). http://www.jstatsoft.org/v61/i03

Adragni KP, Raim A (2014b) ldr: methods for likelihood-based dimension reduction in regression. http://cran.r-project.org/web/packages/ldr/index.html. R package version 1.3

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57(1):289–300

Boos D (2014) Boos-stefanski variable selection home. http://www4.stat.ncsu.edu/~boos/var.select/
Cook RD (1998) Regression graphics. Wiley, Hoboken
Cook RD (2007) Fisher lecture—dimension reduction in regression (with discussion). Stat Sci 22(1):1–26
Cook RD, Forzani L (2008) Principal fitted components for dimension reduction in regression. Stat Sci 23(4):486–501
Cook RD, Forzani L (2009) Likelihood-based sufficient dimension reduction. J Am Stat Assoc 104(485):197–208
Cook RD, Weisberg S (1991) Discussion of sliced inverse regression by KC Li. J Am Stat Assoc 86:328–332
Cornish KM, Li L, Kogan CS, Jacquemont S, Turk J, Dalton A, Hagerman RJ, Hagerman PJ (2008) Age-dependent cognitive changes in carriers of the fragile x syndrome. Cortex 44:628–636
Efron B (2010) Large-scale inference empirical bayes methods for estimation, testing, and prediction. Cambridge University Press, Cambridge
Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32(2):407–499
Enz R (1991) Prices and earnings around the globe: a comparison of purchasing power in 48 cities. Union Bank of Switzerland, Economic Research Dept., Zurich
Friedman J, Hastie T, Tibshirani R (2007) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9:432–441
Guo Z, Li L, Lu W, Li B (2014) Groupwise dimension reduction via envelope method. J Am Stat Assoc. http://dx.doi.org/10.1080/01621459.2014.970687
Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: data mining, inference and prediction. Springer, New York
Li B, Wang S (2007) On directional regression for dimension reduction. J Am Stat Assoc 102:997–1008
Li KC (1991) Sliced inverse regression for dimension reduction (with discussion). J Am Stat Assoc 86:316–342
Li L (2009) Exploiting predictor domain information in sufficient dimension reduction. Comput Stat Data Anal 53(7):2665–2672
Li L, Li B, Zhu L (2010) Groupwise dimension reduction. J Am Stat Assoc 105(491):1188–1201
Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2015) Cluster: cluster analysis basics and extensions, R package version 2.0.3
McDonald G, Schwing R (1973) Instabilities of regression estimates relating air pollution to mortality. Technometrics 15:463–481
R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
Weisberg S (2002) Dimension reduction regression in R. J Stat Softw 7(1):1–22. http://www.jstatsoft.org/v07/i01
Xia Y, Tong H, Li WK, Zhu L (2002) An adaptive estimation of dimension reduction space. J R Stat Soc Ser B 64(3):363–410