Sufficient Dimension Reduction and Prediction in Regression

By Kofi P. Adragni and R. Dennis Cook

University of Minnesota, 313 Ford Hall, 224 Church Street S.E. Minneapolis, MN 55455, USA

Dimension reduction for regression is a prominent issue today because technological advances now allow scientists to routinely formulate regressions in which the number of predictors is considerably larger than in the past. While several methods have been proposed to deal with such regressions, principal components still seem to be the most widely used across the applied sciences. We give a broad overview of ideas underlying a particular class of methods for dimension reduction that includes principal components, along with an introduction to the corresponding methodology. New methods are proposed for prediction in regressions with many predictors.

Keywords: Lasso, Partial least squares, Principal components, Principal component regression, Principal fitted components

1. Introduction

Consider the frequently encountered goal of determining a rule $m(\mathbf{x})$ for predicting a future observation of a univariate response variable Y at the given value $\mathbf{x}$ of a $p\times 1$ vector $\mathbf{X}$ of continuous predictors. Assuming that Y is quantitative, continuous or discrete, the mean squared error $\mathrm{E}(Y-m(\mathbf{x}))^2$ is minimized by choosing $m(\mathbf{x})$ to be the mean $\mathrm{E}(Y|\mathbf{X}=\mathbf{x})$ of the conditional distribution of $Y|(\mathbf{X}=\mathbf{x})$. Consequently, the prediction goal is often specialized immediately to the task of estimating the conditional mean function $\mathrm{E}(Y|\mathbf{X})$ from the regression of Y on $\mathbf{X}$. When the response is categorical with sample space S_Y consisting of h categories $S_Y=\{C_1,\ldots,C_h\}$, the mean function is no longer a relevant quantity for prediction. Instead, given an observation $\mathbf{x}$ on $\mathbf{X}$, the predicted category C_* is usually taken to be the one with the largest conditional probability $C_*=\arg\max \Pr(C_k|\mathbf{X}=\mathbf{x})$, where the maximization is over S_Y. When pursuing estimation of $\mathrm{E}(Y|\mathbf{X})$ or $\Pr(C_k|\mathbf{X})$ it is nearly always worthwhile to consider predictions based on a function $R(\mathbf{X})$ of dimension less than p, provided that it captures all of the information that $\mathbf{X}$ contains about Y so that $\mathrm{E}(Y|\mathbf{X})=\mathrm{E}(Y|R(\mathbf{X}))$. We can think of $R(\mathbf{X})$ as a function that concentrates the relevant information in $\mathbf{X}$. The action of replacing $\mathbf{X}$ with a lower dimensional function $R(\mathbf{X})$ is called *dimension reduction*; it is called *sufficient dimension reduction* when $R(\mathbf{X})$ retains all the relevant information about Y. A potential advantage of sufficient dimension reduction is that predictions based on an estimated R may be substantially less variable than those based on $\mathbf{X}$, without introducing worrisome bias. This advantage is not confined to predictions, but may accrue in other phases of a regression analysis as well.

One goal of this article is to give a broad overview of ideas underlying sufficient dimension reduction for regression, along with an introduction to the corresponding methodology. Sections 1a, 1b, 2 and 3 are devoted largely to this review. Sufficient dimension reduction methods are designed to estimate a population parameter called the central subspace, which is defined in §1b. Another goal of this article is to describe a new method of predicting quantitative responses following sufficient dimension reduction; categorical responses will be discussed only for contrast. The focus of this article shifts to prediction in §4 where we discuss four inverse regression models, describe the prediction methodology that stems from them, and give simulation results to illustrate their behaviour. Practical implementation issues are discussed in §5, along with additional simulation results.

(*a*) *Dimension reduction*

There are many methods available for estimating $\mathrm{E}(Y|\mathbf{X})$ based on a random sample $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, from the joint distribution of Y and $\mathbf{X}$. If p is sufficiently small and n is sufficiently large, it may be possible to estimate $\mathrm{E}(Y|\mathbf{X})$ adequately by using nonparametric smoothing (see, for example, Wand & Jones 1995). Otherwise, nearly all techniques for estimating $\mathrm{E}(Y|\mathbf{X})$ employ some type of dimension reduction for $\mathbf{X}$, either estimated or imposed as an intrinsic part of the model or method.

Broadly viewed, dimension reduction has always been a central statistical concept. In the second half of the nineteenth century 'reduction of observations' was widely recognized as a core goal of statistical methodology, and principal components was emerging as a general method for the reduction of multivariate observations (Adcock 1878). Principal components was established as a first reductive method for regression by the mid 1900s.

Dimension reduction for regression is a prominent issue today because technological advances now allow scientists to routinely formulate regressions in which p is considerably larger than in the past. This has complicated the development and fitting of regression models. Experience has shown that the standard iterative paradigm for model development guided by diagnostics (Cook & Weisberg 1982, p. 7) can be imponderable when applied with too many predictors. An added complication arises when p is larger than the number of observations n, leading to the so called '$n < p$' problem. Standard methods of fitting and corresponding inference procedures may no longer be applicable in such regressions. These and related issues have caused a shift in the applied sciences toward a different regression genre with the goal of reducing the dimensionality of the predictor vector as a first step in the analysis. Although large-p regressions are perhaps mainly responsible for renewed interest, dimension reduction methodology can be useful regardless of the size of p. For instance, it is often helpful to have an informative low-dimensional graphical summary of the regression to facilitate model building and gain insights. For this goal p may be regarded as large when it exceeds 2 or 3 since these bounds represent the limits of our ability to view a data set in full using computer graphics. Subsequent references to 'large p' in this article do not necessarily imply that $n < p$.

Reduction by principal components is ubiquitous in the applied sciences, particularly in bioinformatics applications where principal components have been called 'eigen-genes' (Alter *et al.* 2000) in microarray data analyses and 'meta-$k$mers' in analyses involving DNA motifs. The 2006 *Ad Hoc Committee Report on the 'Hockey Stick' Global Climate Reconstruction*, authored by E. Wegman, D. Scott and Y. Said and commissioned by the U.S. House Energy Committee, reiterates and makes clear that past influential analyses of data on global warming are flawed because of an inappropriate use of principal component methodology.

While principal components seem to be the dominant method of dimension reduction across the applied sciences, there are many other established and recent statistical methods that might be used to address large p regressions, including factor analysis, inverse regression estimation (Cook & Ni 2005), partial least squares, projection pursuit, seeded reductions (Cook *et al.* 2007), kernel methods (Fukumizu *et al.* 2009) and sparse methods like the lasso (Tibshirani 1996) that are based on penalization.

(*b*) *Sufficient Dimension Reduction*

Dimension reduction is a rather amorphous concept in statistics, changing its character and goals depending on context. Formulated specifically for regression, the following definition (Cook 2007) of a *sufficient reduction* will help in our pursuit of methods for reducing the dimension of $\mathbf{X}$ while en route to estimating $\mathrm{E}(Y|\mathbf{X})$:

Definition 1.1. *A reduction $R : \mathbb{R}^p \to \mathbb{R}^q$, $q \leq p$, is sufficient if it satisfies one of the following three statements:*

(i) inverse reduction, $\mathbf{X}|(Y, R(\mathbf{X})) \sim \mathbf{X}|R(\mathbf{X})$,

(ii) forward reduction, $Y|\mathbf{X} \sim Y|R(\mathbf{X})$,

(iii) joint reduction, $\mathbf{X} \perp\!\!\!\perp Y|R(\mathbf{X})$,

where $\perp\!\!\!\perp$ indicates independence, $\sim$ means identically distributed and $\mathbf{A}|\mathbf{B}$ refers to the random vector $\mathbf{A}$ given the vector $\mathbf{B}$.

Each of the three conditions in this definition conveys the idea that the reduction $R(\mathbf{X})$ carries all the information that $\mathbf{X}$ has about Y, and consequently all the information available to estimate $\mathrm{E}(Y|\mathbf{X})$. They are equivalent when $(Y, \mathbf{X})$ has a joint distribution. In that case we are free to determine a reduction inversely or jointly and then pass it to the conditional mean without additional structure: $\mathrm{E}(Y|\mathbf{X}) = \mathrm{E}(Y|R(\mathbf{X}))$. In some cases there may be a direct connection between $R(\mathbf{X})$ and $\mathrm{E}(Y|\mathbf{X})$. For instance, if $(Y, \mathbf{X})$ follows a nonsingular multivariate normal distribution then $R(\mathbf{X}) = \mathrm{E}(Y|\mathbf{X})$ is a sufficient reduction, $\mathrm{E}(Y|\mathbf{X}) = \mathrm{E}\{Y|\mathrm{E}(Y|\mathbf{X})\}$. This reduction is also minimal sufficient: if $T(\mathbf{X})$ is any sufficient reduction then R is a function of T. Further, because of the nature of the multivariate normal distribution, it can be expressed as a linear combination of the elements of $\mathbf{X}$: $R = \boldsymbol{\beta}^T\mathbf{X}$ is minimal sufficient for some vector $\boldsymbol{\beta}$.

Inverse reduction by itself does not require the response Y to be random, and it is perhaps the only reasonable reductive route when Y is fixed by design. For instance, in discriminant analysis $\mathbf{X}|Y$ is a random vector of features observed in one of a number of subpopulations indicated by the categorical response Y, and no discriminatory information will be lost if classifiers are restricted to R.

If we consider a generic statistical problem and reinterpret $\mathbf{X}$ as the total data D and Y as the parameter θ, then the condition for inverse reduction becomes $D|(\theta, R) \sim D|R$ so that R is a sufficient statistic. In this way, the definition of a sufficient reduction encompasses Fisher's (1922) classical definition of sufficiency.

One difference is that sufficient statistics are observable, while a sufficient reduction may contain unknown parameters and thus needs to be estimated. For example, if $(\mathbf{X}, Y)$ follows a nonsingular multivariate normal distribution then $R(\mathbf{X}) = \boldsymbol{\beta}^T\mathbf{X}$ and it is necessary to estimate $\boldsymbol{\beta}$.

In some regressions $R(\mathbf{X})$ may be a nonlinear function of $\mathbf{X}$, and in extreme cases no reduction may be possible, so all sufficient reductions are one-to-one functions of $\mathbf{X}$ and thus equivalent to $R(\mathbf{X}) = \mathbf{X}$. Most often we encounter multi-dimensional reductions consisting of several linear combinations $R(\mathbf{X}) = \boldsymbol{\eta}^T\mathbf{X}$, where $\boldsymbol{\eta}$ is an unknown $p \times q$ matrix, $q \leq p$, that must be estimated from the data. Linear reductions may be imposed to facilitate progress, as in the moment-based approach reviewed in §3a. They can also arise as a natural consequence of modelling restrictions, as we will see in §3b. If $\boldsymbol{\eta}^T\mathbf{X}$ is a sufficient linear reduction then so is $(\boldsymbol{\eta}\mathbf{A})^T\mathbf{X}$ for any $q \times q$ full rank matrix $\mathbf{A}$. Consequently, only the subspace $\mathrm{span}(\boldsymbol{\eta})$ spanned by the columns of $\boldsymbol{\eta}$ can be identified – $\mathrm{span}(\boldsymbol{\eta})$ is called a *dimension reduction subspace.* If $\mathrm{span}(\boldsymbol{\eta})$ is a dimension reduction subspace then so is $\mathrm{span}(\boldsymbol{\eta}, \boldsymbol{\eta}_1)$ for any matrix $p \times q_1$ matrix $\boldsymbol{\eta}_1$. If $\mathrm{span}(\boldsymbol{\eta}_1)$ and $\mathrm{span}(\boldsymbol{\eta}_2)$ are both dimension reduction subspaces, then under mild conditions so is their intersection $\mathrm{span}(\boldsymbol{\eta}_1) \cap \mathrm{span}(\boldsymbol{\eta}_2)$ (Cook 1996, 1998). Consequently, the inferential target in sufficient dimension reduction is often taken to be the central subspace $\mathcal{S}_{Y|X}$, defined as the intersection of all dimension reduction subspaces (Cook 1994, 1996, 1998). A *minimal sufficient linear reduction* is then of the form $R(\mathbf{X}) = \boldsymbol{\eta}^T\mathbf{X}$, where the columns of $\boldsymbol{\eta}$ now form a basis for $\mathcal{S}_{Y|X}$. We assume that the central subspace exists throughout this article, and use $d = \dim(\mathcal{S}_{Y|X})$ to denote its dimension.

The ideas of a sufficient reduction and the central subspace can be used to further our understanding of existing methodology and to guide the development of new methodology. In Sections 2 and 3 we consider how sufficient reductions arise in three contexts: forward linear regression, inverse moment-based reduction and inverse model-based reduction.

2. Reduction in Forward Linear Regression

The standard linear regression model $Y = \beta_0 + \boldsymbol{\beta}^T\mathbf{X} + \epsilon$, with $\epsilon \perp\!\!\!\perp \mathbf{X}$ and $\mathrm{E}(\epsilon) = 0$, implies that $\mathcal{S}_{Y|X} = \mathrm{span}(\boldsymbol{\beta})$ and thus that $R(\mathbf{X}) = \boldsymbol{\beta}^T\mathbf{X}$ is minimal sufficient. The assumption of a linear regression then automatically focuses our interest on $\boldsymbol{\beta}$, which can be estimated straightforwardly using ordinary least squares (OLS) when n is sufficiently large, and it may appear that there is little to be gained from dimension reduction. However, dimension reduction has been used in linear regression to improve on the OLS estimator of $\boldsymbol{\beta}$ and to deal with $n < p$ regressions. One approach consists of regressing Y on $\mathbf{X}$ in two steps. The first is the reduction step: reduce $\mathbf{X}$ linearly to $\mathbf{G}^T\mathbf{X}$ using some methodology that produces $\mathbf{G} \in \mathbb{R}^{p\times q}$, $q \leq p$. The second step consists of using ordinary least squares to estimate the mean function $\mathrm{E}(Y|\mathbf{G}^T\mathbf{X})$ for the reduced predictors. To describe the resulting estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{G}}$ of $\boldsymbol{\beta}$ and establish notation for later sections, let $\mathbb{Y}$ be the $n \times 1$ vector of centred responses, let $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}/n$ denote the sample mean vector, let $\mathbb{X}$ be the $n \times p$ matrix with rows $(\mathbf{X}_i - \bar{\mathbf{X}})^T$, $i = 1, \ldots, n$, let $\widehat{\boldsymbol{\Sigma}} = \mathbb{X}^T\mathbb{X}/n$ denote the usual estimator of $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X})$, let $\widehat{\mathbf{C}} = \mathbb{X}^T\mathbb{Y}/n$, which is the usual estimator of $\mathbf{C} = \mathrm{cov}(\mathbf{X}, Y)$, and let $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}} = \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{C}}$ be the vector of coefficients from the OLS fit of Y on $\mathbf{X}$. Then (Cook & Forzani 2009*b*)

$$\widehat{\boldsymbol{\beta}}_{\mathbf{G}} = \mathbf{P}_{\mathbf{G}(\widehat{\boldsymbol{\Sigma}})}\widehat{\boldsymbol{\beta}}_{\text{ols}} = \mathbf{G}(\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\mathbf{G})^{-1}\mathbf{G}^T\widehat{\mathbf{C}}. \tag{2.1}$$

This estimator, which is the projection $\mathbf{P}_{\mathbf{G}(\widehat{\boldsymbol{\Sigma}})}$ of $\widehat{\boldsymbol{\beta}}_{\text{ols}}$ onto span$(\mathbf{G})$ in the $\widehat{\boldsymbol{\Sigma}}$ inner product, does not require computation of $\widehat{\boldsymbol{\Sigma}}^{-1}$ if $q < p$ and thus could be useful when $n < p$, depending on the size of q. In any case the estimator of $\mathrm{E}(Y|\mathbf{X})$ is

$$\widehat{\mathrm{E}}(Y|\mathbf{X}) = \bar{\mathbf{Y}} + \widehat{\boldsymbol{\beta}}_{\mathbf{G}}^T(\mathbf{X} - \bar{\mathbf{X}}). \tag{2.2}$$

If $\mathbf{G} = \mathbf{I}_p$ then $\widehat{\boldsymbol{\beta}}_{\mathbf{G}} = \widehat{\boldsymbol{\beta}}_{\text{ols}}$, which achieves nothing beyond $\widehat{\boldsymbol{\beta}}_{\text{ols}}$. If we choose the columns of $\mathbf{G}$ to be the first q eigenvectors of $\widehat{\boldsymbol{\Sigma}}$ then $\mathbf{G}^T\mathbf{X}$ consists of the first q principal components and $\widehat{\boldsymbol{\beta}}_{\mathbf{G}}$ is the standard principal component regression (PCR) estimator. Setting $\mathbf{G} = (\widehat{\mathbf{C}}, \widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{C}}, \ldots, \widehat{\boldsymbol{\Sigma}}^{q-1}\widehat{\mathbf{C}})$ yields the partial least squares (PLS) estimator with q factors (Helland 1990). Eliminating predictors by using an information criterion like AIC or BIC (see §5a) can result in a $\mathbf{G}$ with rows selected from the identity matrix $\mathbf{I}_p$, and again we obtain a reduction in $\mathbf{X}$ prior to estimation of $\boldsymbol{\beta}$. If span$(\mathbf{G})$ is a consistent estimator of a dimension reduction subspace $\mathcal{S}$ then $\widehat{\boldsymbol{\beta}}_{\mathbf{G}}$ may be a reasonable estimator of $\boldsymbol{\beta}$ since span$(\boldsymbol{\beta}) \subseteq \mathcal{S} \subseteq \mathbb{R}^p$, recalling that the intersection of any two dimension reduction subspaces is itself a dimension reduction subspace. However, while these estimators are well known, span$(\mathbf{G})$ may not be a consistent estimator of a dimension reduction subspace without additional structure, even if the linear model is accurate. The PCR estimator depends on $\mathbf{G}$ only through the marginal distribution of $\mathbf{X}$ and this alone cannot guarantee that span$(\mathbf{G})$ is consistent. The performance of the PLS estimator depends on the relationship between $\mathbf{C}$ and the eigenstructure of $\boldsymbol{\Sigma}$ (Naik & Tsai 2000).

A somewhat different approach is based on estimating $\boldsymbol{\beta}$ by using a penalized objective function like that for the lasso (Tibshirani, 1996). The lasso estimator is

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n}(Y_i - \bar{Y} - \boldsymbol{\beta}^T(\mathbf{X}_i - \bar{\mathbf{X}}))^2 + \lambda\sum_{j=1}^{p}|\beta_j| \right\},$$

where β_j is the j-th element of $\boldsymbol{\beta}$, $j = 1, \ldots, p$, and the tuning parameter λ is often chosen by cross validation. Several elements of $\widehat{\boldsymbol{\beta}}_{\text{lasso}}$ are typically zero, which corresponds to setting the rows of $\mathbf{G}$ to be the rows of the identity matrix $\mathbf{I}_p$ corresponding to the nonzero elements of $\widehat{\boldsymbol{\beta}}_{\text{lasso}}$. However, with this $\mathbf{G}$ we do not necessarily have $\widehat{\boldsymbol{\beta}}_{\text{lasso}} = \widehat{\boldsymbol{\beta}}_{\mathbf{G}}$, although the two estimators are often similar. Consequently, methodology based on penalization does not fit exactly the general form given in equation (2.1).

Pursuing dimension reduction based on linear regression may not produce useful results if the model is not accurate, particularly if the distribution of $Y|\mathbf{X}$ depends on more than one linear combination of the predictors. There are many diagnostic and remedial methods available to improve linear regression models when p is not too large. Otherwise, application of these methods can be quite burdensome.

3. Inverse Reduction

Inverse regression $\mathbf{X}|Y$ provides an alternative approach to estimating a sufficient reduction. It can deal straightforwardly with regressions that depend on more that one linear combination of the predictors, and does not necessarily suffer from the modelling problems that plague forward regression when p is large. There are two general paradigms for determining a sufficient reduction inversely. The first is by specifying a parametric model for the inverse regression of $\mathbf{X}$ on Y, as discussed in Sections 3b and 4. In this model-based approach, minimal sufficient reductions can in principle be determined from the model itself. For example, we saw previously that $\mathrm{E}(Y|\mathbf{X})$ is a sufficient reduction when $(Y, \mathbf{X})$ is normally distributed. The second, which is discussed §3a, is the moment-based approach in which derived moment relations are used to estimate a sufficient reduction by way of the central subspace. Model accuracy is nearly always an issue in the model-based approach, while efficiency is worrisome in the moment-based approach. Some of the best moment-based methods have turned out to be quite inefficient in relatively simple settings. (See Cook & Forzani 2009*a*, for an instance of this inefficiency.)

(*a*) *Moment-based inverse reduction*

In contrast to model-based reduction, there is no law to guide the choice of $R(\mathbf{X})$ in moment-based reduction. However progress is still possible by restricting consideration to multi-dimensional linear reduction and pursuing estimation of the central subspace $\mathcal{S}_{Y|X}$, as discussed in §1b. Recall that $d = \dim(\mathcal{S}_{Y|X})$ and that the columns of the $p \times d$ matrix $\boldsymbol{\eta}$ are a basis for $\mathcal{S}_{Y|X}$.

Sliced inverse regression (SIR, Li 1991) and sliced average variance estimation (SAVE, Cook & Weisberg 1991) were the first moment-based methods proposed for dimension reduction. Although the concept of the central subspace was not developed until a few years after SIR and SAVE were introduced, it is now known that these methods in fact estimate $\mathcal{S}_{Y|X}$ under two key conditions: (a) $\mathrm{E}(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ is a linear function of $\mathbf{X}$ (*linearity condition*) and (b) $\mathrm{var}(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ is a nonrandom matrix (*constant covariance condition*). We forgo discussion of these conditions, which involve only the marginal distribution of $\mathbf{X}$, since they are well known and widely regarded as mild. A good recent discussion of them was given by Li & Wang (2007). Under the linearity condition $\mathrm{E}(\mathbf{X}|Y) - \mathrm{E}(\mathbf{X}) \in \boldsymbol{\Sigma}\mathcal{S}_{Y|X}$, which is the population foundation for SIR. Under the linearity and constant covariance conditions $\mathrm{span}(\boldsymbol{\Sigma} - \mathrm{var}(\mathbf{X}|Y)) \in \boldsymbol{\Sigma}\mathcal{S}_{Y|X}$, which is population basis for SAVE. When the response is categorical, $\mathrm{E}(\mathbf{X}|C_k)$ can be estimated straightforwardly as the average predictor vector $\bar{\mathbf{X}}_k$ in category C_k. The SIR estimator of $\mathcal{S}_{Y|X}$, which requires $n > p$, is then the span of the first d eigenvectors of $\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{M}\mathbf{M}^T$, where $\mathbf{M}$ is the $p \times h$ matrix with columns $\bar{\mathbf{X}}_k - \bar{\mathbf{X}}$. Continuous responses are treated by slicing the observed range of Y into h categories C_k and then applying the method for a categorical response. The SAVE estimator uses a similar construction. Routines for computing SIR, SAVE and other moment-based estimates of $\mathcal{S}_{Y|X}$ are available in the R package ‘dr’ (http://cran.us.r-project.org/web/packages/dr/index.html) and in the Arc software (www.stat.umn.edu/arc/software.html).

SIR and SAVE both provide $\sqrt{n}$ consistent estimators of $\mathcal{S}_{Y|X}$ under standard conditions, but by itself consistency does not guarantee good performance in practice. It is known that SIR has difficulty finding directions that are associated with certain types of nonlinear trends in $\mathrm{E}(Y|\mathbf{X})$. SAVE was developed in response to this limitation but its ability to find linear trends is generally inferior to SIR's. Several moment-based methods have been developed in an effort to improve on the estimates of $\mathcal{S}_{Y|X}$ provided by SIR and SAVE. Using the same population foundations as SIR, Cook & Ni (2005) developed an asymptotically optimal method of estimating $\mathcal{S}_{Y|X}$ called IRE (inverse regression estimation). Ye & Weiss (2003) and Zhu *et al.* (2005) attempted to combine the advantages of SIR and SAVE by using linear combinations of them. Cook & Forzani (2009*a*) used a likelihood-based objective function to develop a method called LAD (likelihood acquired directions) that apparently dominates all dimension reduction methods based on the same population foundations as SIR and SAVE. These methods have been developed and studied mostly in regressions where $p \ll n$, although there are some results for other settings (Li 2007; Li & Yin 2008). SIR, SAVE, IRE and LAD come with a range of inference capabilities, including methods for estimating d and tests of conditional independence hypotheses such as Y is independent of $\mathbf{X}_1$ given $\mathbf{X}_2$, where we have partitioned $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$.

Moment-based sufficient dimension reduction methods provide estimates of the minimal sufficient linear reduction, but they are not designed specifically for prediction and do not produce predictive methods *per se*. Instead, once $\widehat{R} = \widehat{\boldsymbol{\eta}}^T\mathbf{X}$ has been determined, it is considered fixed and standard model development methods are typically used to estimate $\mathrm{E}(Y|\widehat{\boldsymbol{\eta}}^T\mathbf{X})$ and thereby obtain predictions. Model-based sufficient reduction methods allow a more direct route to estimation of $\mathrm{E}(Y|\mathbf{X})$.

(*b*) *Model-based inverse reduction*

Model-based sufficient dimension reduction is relatively new (Cook 2007). There are several useful characteristics of this approach that may become clear in the sections that follow. One is that a model for $\mathbf{X}|Y$ can itself be inverted to provide a method for estimating the forward mean function $\mathrm{E}(Y|\mathbf{X})$ without specifying a model for the full joint distribution of $(\mathbf{X}, Y)$. For convenience we denote the densities of $\mathbf{X}$ and $\mathbf{X}|Y$ by $g(\mathbf{X})$ and $g(\mathbf{X}|Y)$, and so on, keeping in mind that the symbol g indicates a different density in each case. Because densities will always appear together with their arguments this should cause no ambiguity. We assume that $R(\mathbf{X})$ has a density as well. With these understandings we have

$$\begin{aligned} \mathrm{E}\{Y|\mathbf{X}=\mathbf{x}\} &= \mathrm{E}\{Yg(\mathbf{x}|Y)\}/\mathrm{E}\{g(\mathbf{x}|Y)\} && (3.1)\\ &= \mathrm{E}\{Y|R(\mathbf{x})\} && (3.2)\\ &= \mathrm{E}\{Yg(R(\mathbf{x})|Y)\}/\mathrm{E}\{g(R(\mathbf{x})|Y)\}, && (3.3) \end{aligned}$$

where all right-hand side expectations are with respect to the marginal distribution of Y. Equation (3.1) provides a relationship between the mean function $\mathrm{E}\{Y|\mathbf{X}\}$ and the conditional density of $\mathbf{X}|Y$, while (3.2) is a restatement of the equality of the mean functions for the regressions of Y on $\mathbf{X}$ and $R(\mathbf{X})$. The final equality (3.3) establishes a relationship between these forward mean functions and the conditional density of $R|Y$, and provides a method to estimate $\mathrm{E}(Y|\mathbf{X})$:

$$\begin{aligned} \widehat{\mathrm{E}}\{Y|\mathbf{X}=\mathbf{x}\} &= \sum_{i=1}^{n} w_i(\mathbf{x})Y_i \\ w_i(\mathbf{x}) &= \frac{\widehat{g}(\widehat{R}(\mathbf{x})|Y_i)}{\sum_{i=1}^{n}\widehat{g}(\widehat{R}(\mathbf{x})|Y_i)}, \end{aligned} \tag{3.4}$$

where $\widehat{g}$ denotes an estimated density and $\widehat{R}$ is the estimated reduction. This estimator is reminiscent of a nonparametric kernel estimator (Simonoff 1996, ch. 4), but there are important differences. The weights in a kernel estimator do not depend on the response, while the weights w_i here do. Kernel weights typically depend on the full vector of predictors $\mathbf{X}$, while the weights here depend on $\mathbf{X}$ only through the estimated reduction $\widehat{R}(\mathbf{x})$. If d is small, the estimator (3.4) may avoid the curse of dimensionality. Multivariate kernels are usually taken to be the product of univariate kernels, corresponding here to constraining the components of R to be independent. Finally, there is no explicit bandwidth in our weights since they are determined entirely from $\widehat{g}$, which eliminates the need for bandwidth estimation by, for example, cross validation.

The success of this approach depends on obtaining good estimators of the reduction and of its conditional density. In the next section we address these issues by using normal models for the conditional distribution of $\mathbf{X}|Y$. The models in §4a, §4b and §4d were introduced by Cook (2007). The model in §4c is from Cook & Forzani (2009*b*). Our discussion includes a review of the results that are needed to use (3.4).

4. Normal Inverse Models

Let $\mathbf{X}_y$ denote a random vector distributed as $\mathbf{X}|(Y=y)$, and assume that $\mathbf{X}_y$ is normally distributed with mean $\boldsymbol{\mu}_y$ and constant variance matrix $\boldsymbol{\Delta} > 0$, where the inequality means that $\boldsymbol{\Delta}$ is positive definite. Let $\bar{\boldsymbol{\mu}} = \mathrm{E}(\mathbf{X})$ and let $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$ denote a basis matrix whose columns form a basis for the d-dimensional subspace $\mathcal{S}_{\boldsymbol{\Gamma}} = \mathrm{span}\{\boldsymbol{\mu}_y - \bar{\boldsymbol{\mu}}|y \in S_Y\}$, where S_Y denotes the sample space of Y. Then we can write (Cook 2007)

$$\mathbf{X}_y = \bar{\boldsymbol{\mu}} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \boldsymbol{\varepsilon}, \tag{4.1}$$

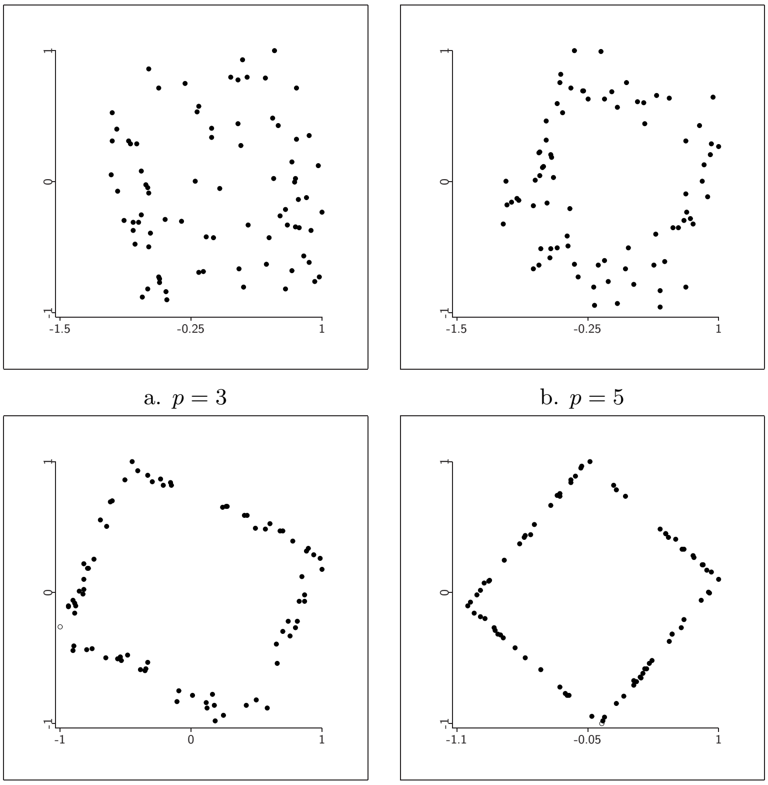
where $\boldsymbol{\varepsilon}$ is independent of Y and normally distributed with mean 0 and covariance matrix $\boldsymbol{\Delta}$, and $\boldsymbol{\nu}_y = (\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T(\boldsymbol{\mu}_y - \bar{\boldsymbol{\mu}}) \in \mathbb{R}^d$; we assume that $\mathrm{var}(\boldsymbol{\nu}_Y) > 0$. The basis matrix $\boldsymbol{\Gamma}$ is not identifiable in this model since for any full rank $d \times d$ matrix $\mathbf{A}$ we can always obtain an equivalent parameterization as $\boldsymbol{\Gamma}\boldsymbol{\nu}_y = (\boldsymbol{\Gamma}\mathbf{A}^{-1})(\mathbf{A}\boldsymbol{\nu}_y)$. However, $\mathrm{span}(\boldsymbol{\Gamma})$ is identifiable and estimable, and for this reason we assume without loss of generality that $\boldsymbol{\Gamma}$ is a semi-orthogonal matrix, $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_d$.

Model (4.1) represents the fact that the translated conditional means $\boldsymbol{\mu}_y - \bar{\boldsymbol{\mu}}$ fall in the d-dimensional subspace $\mathcal{S}_{\boldsymbol{\Gamma}}$. Under model (4.1) $R(\mathbf{X}) = \boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$ is minimal sufficient (Cook & Forzani 2009*b*), and the goal is thus to estimate $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}} = \{\boldsymbol{\Delta}^{-1}\mathbf{z} : \mathbf{z} \in \mathcal{S}_{\boldsymbol{\Gamma}}\}$. Since the minimal sufficient reduction is linear it follows that $\mathcal{S}_{Y|X} = \boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$. In other words, in the class of models represented by equation (4.1), moment-based and model-based dimension reduction coincide in the population.

As a convenient notation for describing estimators of $\mathbf{\Delta}^{-1}\mathcal{S}_{\mathbf{\Gamma}}$, let $\mathcal{S}_d(\mathbf{A},\mathbf{B})$ denote the span of $\mathbf{A}^{-1/2}$ times the first d eigenvectors of $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$, where $\mathbf{A}$ and $\mathbf{B}$ are symmetric matrices and $\mathbf{A}$ is nonsingular. The subspace $\mathcal{S}_d(\mathbf{A},\mathbf{B})$ can also be described as the span of $\mathbf{A}^{-1}$ times the first d eigenvectors of $\mathbf{B}$. We refer to errors having covariance matrix $\mathbf{\Delta} = \sigma^2\mathbf{I}_p$ as *isotonic*. Isotonic models are models with isotonic errors. For notational simplicity we will use $\mathbf{X}_i$ when referring to observations rather than the more awkward notation $\mathbf{X}_{y_i}$.

(a) *The PC model*

The isotonic version of model (4.1) is called the *PC model* since the maximum likelihood estimator (MLE) of $\mathbf{\Delta}^{-1}\mathcal{S}_{\mathbf{\Gamma}} = \mathcal{S}_{\mathbf{\Gamma}}$ is $\mathcal{S}_d(\mathbf{I}_p,\widehat{\mathbf{\Sigma}})$ and thus the d components of $\widehat{R}(\mathbf{X})$ are simply the first d principal components. This relatively simple result is due to the nature of $\mathbf{\Delta}$. Since the errors are isotonic, the contours of $\mathbf{\Delta}$ are circular. When the signal $\mathbf{\Gamma}\boldsymbol{\nu}_y$ is added the contours of $\mathbf{\Sigma} = \mathbf{\Gamma}\mathrm{var}(\boldsymbol{\nu}_Y)\mathbf{\Gamma}^T + \sigma^2\mathbf{I}_p$ become p-dimensional ellipses with their longest d axes spanning $\mathcal{S}_{\mathbf{\Gamma}}$. The MLE $\widehat{\sigma}^2$ of σ^2 is $\widehat{\sigma}^2 = \sum_{j=d+1}^{p}\widehat{\lambda}_j/p$, where $\widehat{\lambda}_1 > \ldots > \widehat{\lambda}_{d+1} \geq \ldots \geq \widehat{\lambda}_p$ are the eigenvalues of $\widehat{\mathbf{\Sigma}}$, and the MLE of $\bar{\boldsymbol{\mu}}$ is simply $\bar{\mathbf{X}}$.



a. $p = 3$ b. $p = 5$

c. $p = 25$ d. $p = 500$

Figure 1. Plots of the estimated sufficient reduction from the PC model with $n = 80$ observations and varying number of predictors p.

We performed a small simulation to provide some intuition. Observations on $\mathbf{X}$ were generated as $\mathbf{X}_y = \boldsymbol{\Gamma}^* \boldsymbol{\nu}_y^* + \boldsymbol{\varepsilon}$, where $\boldsymbol{\nu}_y^*$ was sampled uniformly from the boundary of the square $[-1, 1]^2$, the elements of the $p \times 2$ matrix $\boldsymbol{\Gamma}^*$ were sampled independently from a standard normal distribution, and the error vector $\boldsymbol{\varepsilon}$ was sampled from a normal distribution with mean 0 and variance matrix $\mathbf{I}_p$. In terms of model (4.1), $\bar{\boldsymbol{\mu}} = 0$, $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^*(\boldsymbol{\Gamma}^{*T}\boldsymbol{\Gamma}^*)^{-1/2}$ and $\boldsymbol{\nu}_y = (\boldsymbol{\Gamma}^{*T}\boldsymbol{\Gamma}^*)^{1/2}\boldsymbol{\nu}_y^*$. This sampling process, which does not require an explicit choice of Y, was repeated $n = 80$ times for various values of p. Figure a shows plots of the first two principal components for four values of p. We see that for small p the square is not recognizable, but for larger values of p the square is quite clear. In figure ad, there are $p = 500$ predictors, while the number of observations is still $n = 80$. The sides of the estimated square in figure ad do not align with the coordinate axes because the method is designed to estimate only the subspace $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$, which is equal to $\mathcal{S}_{\boldsymbol{\Gamma}}$ with isotonic errors.

Turning to prediction, let $\widehat{\boldsymbol{\Gamma}}$ denote the $p \times d$ matrix with columns consisting of the first d eigenvectors of $\widehat{\boldsymbol{\Sigma}}$. Then the weights (3.4) can be written as

$$\begin{aligned} w_i(\mathbf{x}) &\propto \exp\{-(2\widehat{\sigma}^2)^{-1}\|\widehat{\boldsymbol{\Gamma}}^T(\mathbf{x} - \mathbf{X}_i)\|^2\} \\ &= \exp\{-(2\widehat{\sigma}^2)^{-1}\|\widehat{R}(\mathbf{x}) - \widehat{R}(\mathbf{X}_i)\|^2\}, \end{aligned} \tag{4.2}$$

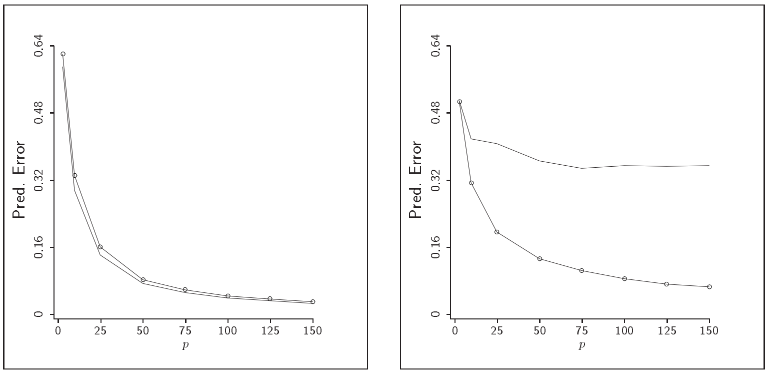
where $\widehat{R}(\mathbf{x}) = \widehat{\boldsymbol{\Gamma}}^T\mathbf{x}$ is the estimated reduction. In this case $\widehat{\mathrm{E}}(Y|\mathbf{X} = \mathbf{x})$ is in the form of a normal product kernel density estimator with bandwidth $\widehat{\sigma}$. Here and in other weights we assumed that d is known. Methods for choosing d in the context of prediction are discussed in §5.

The method of estimating $\mathrm{E}(Y|\mathbf{X})$ characterized by (4.2) is distinct from the standard PCR estimator given in equation (2.2) with the columns of $\mathbf{G}$ being the first d eigenvectors of $\widehat{\boldsymbol{\Sigma}}$. Predictions from (4.2) and (2.2) both use principal components, but the way in which they are used is quite different. Model (4.1), which leads to principal components as the MLE of the sufficient reduction, has a very general mean function, but its error structure is restrictive. Nevertheless, this error structure has recently been used in studies of gene expression ($\mathbf{X}$) that are complicated by stratification and heterogeneity (Leek & Storey 2007). On the other hand, the usual linear model has a restrictive mean function, and under that model alone there seems to be no clear rationale for reduction by principal components (Cox 1968).

We performed a small simulation study to illustrate the relative behaviour of predictions using (2.2) and (4.2). The basic simulation scenario $\mathbf{X}_y = \boldsymbol{\Gamma}^*\boldsymbol{\nu}_y^* + \boldsymbol{\varepsilon}$ was the same as that leading to figure a, but in this case we set $n = 200$, $d = 1$, $\boldsymbol{\Gamma}^* = (1, \ldots, 1)^T$, $\mathrm{var}(\boldsymbol{\varepsilon}) = 5^2\mathbf{I}_p$ and $\boldsymbol{\nu}_Y^* = 2.5Y$, where Y is a standard normal random variable. In this setup, $(\mathbf{X}, Y)$ has a multivariate normal distribution so both (2.2) and (4.2) are appropriate, but the predictions from PCR (2.2) use the fact that the mean function $\mathrm{E}(Y|\mathbf{X})$ is linear, while predictions using (4.2) do not. For each of 100 datasets generated in this way the mean squared prediction error PE_k for the estimated mean function $\widehat{\mathrm{E}}(Y|\mathbf{X})$ was determined by sampling 200 new observations $(\mathbf{X}^*, Y^*)$:

$$\mathrm{PE}_k = \sum_{i=1}^{200}(Y_i^* - \widehat{\mathrm{E}}(Y|\mathbf{X} = \mathbf{X}_i^*))^2/200, \quad k = 1, \ldots, 100. \tag{4.3}$$

The final prediction error was then determined by averaging, $\sum_{k=1}^{100} \mathrm{PE}_k/100$. The prediction errors are shown in figure aa for $3 \leq p \leq 150$. While the PCR predictions (2.2) perform a bit better at all values of p, the loss when using predictions from the inverse model is negligible. Figure ab was constructed in the same way, except that we set $\boldsymbol{\nu}_Y^* = Y + Y^2 + Y^3$. In this case the predictions (4.2) from the inverse model are much better than the PCR predictions because they automatically adapt to $\mathrm{E}(Y|\mathbf{X})$ regardless of the nature of $\boldsymbol{\nu}_Y$.



a. $\boldsymbol{\nu}_Y^* = 2.5Y$ b. $\boldsymbol{\nu}_Y^* = Y + Y^2 + Y^3$

Figure 2. Comparison of prediction errors: The line passing through the plotted points corresponds to predictions from the inverse model using (3.4) with weights (4.2). The other line is for PCR predictions (2.2).

(*b*) *The Isotonic PFC model*

It will often be possible and useful to model the coordinate vectors as $\boldsymbol{\nu}_y = \boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})$, where $\mathbf{f}_y \in \mathbb{R}^r$ is a known vector-valued function of y with linearly independent elements and $\boldsymbol{\beta} \in \mathbb{R}^{d \times r}$, $d \leq \min(r, p)$, is an unrestricted rank d matrix. Under this model for $\boldsymbol{\nu}_y$ each coordinate X_{yj}, $j = 1, \ldots, p$, of $\mathbf{X}_y$ follows a linear model with predictor vector $\mathbf{f}_y$. Consequently, when Y is quantitative we are able to use inverse response plots (Cook 1998, ch. 10) of X_{yj} versus y, $j = 1, \ldots, p$, to gain information about suitable choices for $\mathbf{f}_y$, which is an ability that is not generally available in the forward regression of Y on $\mathbf{X}$. When p is too large for such graphical inspection, the PC model can be used as a tool to aid in selecting $\mathbf{f}_y$. Under the PC model the MLE of $\boldsymbol{\nu}_y$ for the i-th observed case is $\widehat{\boldsymbol{\nu}}_i = (\widehat{\boldsymbol{\Gamma}}^T \widehat{\boldsymbol{\Gamma}})^{-1} \widehat{\boldsymbol{\Gamma}}^T (\mathbf{X}_i - \bar{\mathbf{X}}) = \widehat{\boldsymbol{\Gamma}}^T (\mathbf{X}_i - \bar{\mathbf{X}})$; the second equality follows since $\widehat{\boldsymbol{\Gamma}}$ is a semi-orthogonal matrix. These vectors $\widehat{\boldsymbol{\nu}}_i$ can then be plotted against Y_i, $i = 1, \ldots, n$, and used in the same way as the inverse response plots. In cases like figure ad, brushing a histogram of the response while observing movement of the corresponding points around the square may be useful. We can also consider $\mathbf{f}_y$s that contain a reasonably flexible set of basis functions, like polynomial terms in Y, which may also be useful when it is impractical to apply graphical methods to all of the predictors. Piecewise polynomials could also be used.

In some regressions there may be a natural choice for $\mathbf{f}_y$. Suppose for instance that Y is categorical, taking values in one of h categories C_k, $k = 1, \ldots, h$. We can then set $r = h - 1$ and specify the k-th element of $\mathbf{f}_y$ to be $J(y \in C_k)$, where J is the indicator function. Another option consists of 'slicing' the observed values of a continuous Y into h bins (categories) C_k, $k = 1, \ldots, h$, and then specifying the k-th coordinate of $\mathbf{f}_y$ as for the case of a categorical Y. This has the effect of approximating each conditional mean $\mathrm{E}(X_{yj})$ as a step function of y with h steps,

$$\mathrm{E}(X_{yj}) \approx \bar{\mu}_j + \sum_{k=1}^{h-1} \boldsymbol{\gamma}_j^T \mathbf{b}_k \{J(y \in C_k) - \Pr(Y \in C_k)\},$$

where $\boldsymbol{\gamma}_j^T$ is the j-th row of $\boldsymbol{\Gamma}$ and $\mathbf{b}_k$ is the k-th column of $\boldsymbol{\beta}$.

Models with $\boldsymbol{\nu}_y = \boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})$ are called *principal fitted component (PFC) models.* To describe the MLE of $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ when the errors are isotonic, let $\mathbb{F}$ denote the $n \times r$ matrix with rows $(\mathbf{f}_i - \bar{\mathbf{f}})^T$. Then the $n \times p$ matrix of centred fitted vectors from the linear regression of $\mathbf{X}$ on $\mathbf{f}$ is $\widehat{\mathbb{X}} = \mathbf{P}_{\mathbb{F}}\mathbb{X}$ and sample covariance matrix of these fitted vectors is $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}} = \mathbb{X}^T\mathbf{P}_{\mathbb{F}}\mathbb{X}/n$, where $\mathbf{P}_{\mathbb{F}}$ denotes the linear operator that projects onto the subspace spanned by the columns of $\mathbb{F}$. Under this *isotonic PFC model* the MLE of $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ is $\mathcal{S}_d(\mathbf{I}_p, \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}})$. Johnson (2008) gave sufficient conditions for this estimator to converge at the usual $\sqrt{n}$ rate.

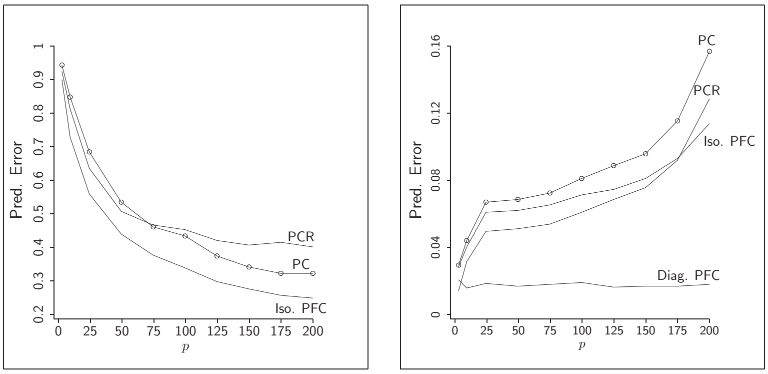
Under the isotonic PFC model the sufficient reduction is estimated as $\widehat{R}(\mathbf{x}) = \widehat{\boldsymbol{\Gamma}}^T\mathbf{x}$, where the columns of $\widehat{\boldsymbol{\Gamma}}$ are the first d eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}$. The elements of this $\widehat{R}$ are called *principal fitted components.* Additionally, the MLE of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Gamma}}^T\mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}$ and the MLE of σ^2 is $\widehat{\sigma}^2 = (\sum_{j=1}^p \widehat{\lambda}_j - \sum_{j=1}^d \widehat{\lambda}_j^{\mathrm{fit}})/p$, where the $\widehat{\lambda}_j^{\mathrm{fit}}$s are the ordered eigenvalues of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}$.

To describe the weights under this model, let $\mathbf{B}_{\mathrm{ols}} = \mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}$ be the coefficient matrix from the multivariate OLS fit of $\mathbf{X}$ on $\mathbf{f}$, and let the fitted vectors be denoted by $\widehat{\mathbf{X}}_i = \bar{\mathbf{X}} + \mathbf{B}_{\mathrm{ols}}(\mathbf{f}_i - \bar{\mathbf{f}})$. The weights from (3.4) are then

$$w_i(\mathbf{x}) \propto \exp\{-(2\widehat{\sigma}^2)^{-1}\|\widehat{R}(\mathbf{x}) - \widehat{R}(\widehat{\mathbf{X}}_i)\|^2\}. \tag{4.4}$$

These weights are of the same general form as those in equation (4.2) from the PC model, but there are two important differences. First, the isotonic PFC reduction uses the eigenvectors from $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}$, while the PC reduction uses the eigenvectors from $\widehat{\boldsymbol{\Sigma}}$. Second, the reduction is applied to the observed predictor vectors $\mathbf{X}_i$ in the PC weights (4.2), while the reduction is applied to the fitted predictor vectors $\widehat{\mathbf{X}}_i$ in the isotonic PFC weights (4.4).

Shown in figure ba are results from a simulation study to illustrate the potential advantages of using fitted components. The data were generated as for figure ab. We used $\mathbf{f}_y = (y, y^2, y^3)^T$ for the fitted model and, to broaden the scope, we increased the overall variability by reducing the sample size to $n = 50$ and increasing the conditional variance matrix to $\mathrm{var}(\boldsymbol{\varepsilon}) = 15^2\mathbf{I}_p$. This is a highly variable regression, but all three methods respond quickly as the number of predictors increases. The PCR predictions do better than the inverse predictions from the PC model for relatively small p, but the situation is reversed for larger values of p. Most importantly, the prediction errors from isotonic PFC are the smallest at all values of p. Figure bb is discussed in the next section.



a. Isotonic PFC, $\mathbf{f}_y = (y, y^2, y^3)^T$ b. Diagonal PFC, $f_y = y$

Figure 3. Comparison of prediction errors from principal component regression (PCR; eq. (2.2)), inverse principal components (PC; eq. (4.2)), isotonic principal fitted components (Iso. PFC; eq. (4.4)) and diagonal principal fitted components (Diag. PFC; eq. (4.5)).

(*c*) *The diagonal PFC model*

The isotonic PFC model requires that, given the response, the predictors must be independent and have the same variance. While this model will be useful in some applications, the requirement of equal variances is restrictive relative to the range of applications in which reduction may be desirable. In this section we expand the scope of application by permitting a diagonal error covariance matrix $\boldsymbol{\Delta} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. This allows for different measurement scales of the predictors, but still requires that they be conditionally independent. The estimated sufficient reduction for this model is $\widehat{R}(\mathbf{x}) = \widehat{\boldsymbol{\Gamma}}^T \widehat{\boldsymbol{\Delta}}^{-1}\mathbf{x}$, where $\widehat{\boldsymbol{\Delta}}$ is the MLE of $\boldsymbol{\Delta}$ and $\widehat{\boldsymbol{\Gamma}}$ is any basis for the MLE of span$(\boldsymbol{\Gamma})$. With this $\widehat{R}$ the weights have the same form as the weights for the isotonic PFC model:

$$w_i(\mathbf{x}) \propto \exp\{-(1/2)\|\widehat{R}(\mathbf{x}) - \widehat{R}(\widehat{\mathbf{X}}_i)\|^2\}, \tag{4.5}$$

where $\widehat{\mathbf{X}}$ is as defined for equation (4.4).

It remains to determine $\widehat{\boldsymbol{\Delta}}$ and $\widehat{\boldsymbol{\Gamma}}$. For this model we were unable to find a closed-form expression for these estimators. However, a straightforward alternating algorithm can be developed based on the following reasoning. If the inverse mean function is specified then the variances σ_j^2 can be estimated by using the sample variances of the centred variables $\mathbf{X}_i - \bar{\boldsymbol{\mu}} - \boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f}_i - \bar{\mathbf{f}})$. If $\boldsymbol{\Delta}$ is specified then we can standardize the predictor vector to obtain an isotonic PFC model in $\mathbf{Z} = \boldsymbol{\Delta}^{-1/2}\mathbf{X}$:

$$\mathbf{Z} = \boldsymbol{\Delta}^{-1/2}\bar{\boldsymbol{\mu}} + \boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f} - \bar{\mathbf{f}}) + \boldsymbol{\varepsilon}, \tag{4.6}$$

where $\boldsymbol{\varepsilon}$ is normal with mean $\mathbf{0}$ and variance matrix $\mathbf{I}_p$. Consequently, we can estimate span$(\boldsymbol{\Gamma})$ as $\boldsymbol{\Delta}^{1/2}$ times the estimate $\widetilde{\boldsymbol{\Gamma}}$ of $\boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma}$ from the isotonic model (4.6). Alternating between these two steps leads to the following algorithm:

1. Fit the isotonic PFC model to the original data, getting initial estimates $\widehat{\boldsymbol{\Gamma}}_{(1)}$ and $\widehat{\boldsymbol{\beta}}_{(1)}$. The MLE of the intercept $\bar{\boldsymbol{\mu}}$ is always $\bar{\mathbf{X}}$ so there is no need to include this parameter explicitly in the algorithm.

2. For some small $\epsilon > 0$, repeat for $j = 1, 2 \ldots$ until $\mathrm{tr}\{(\widehat{\mathbf{\Delta}}_{(j)} - \widehat{\mathbf{\Delta}}_{(j+1)})^2\} < \epsilon$
   (a) Calculate $\widehat{\mathbf{\Delta}}_{(j)} = \mathrm{diag}\{(\mathbb{X} - \mathbb{F}\widehat{\boldsymbol{\beta}}^T_{(j)}\widehat{\mathbf{\Gamma}}^T_{(j)})^T(\mathbb{X} - \mathbb{F}\widehat{\boldsymbol{\beta}}^T_{(j)}\widehat{\mathbf{\Gamma}}^T_{(j)})\}$,
   (b) Transform $\mathbf{Z} = \widehat{\mathbf{\Delta}}^{-1/2}_{(j)}\mathbf{X}$,
   (c) Fit the isotonic PFC model to $\mathbf{Z}$, yielding estimates $\widetilde{\mathbf{\Gamma}}$, and $\widetilde{\boldsymbol{\beta}}$,
   (d) Backtransform the estimates to the original scale $\widehat{\mathbf{\Gamma}}_{(j+1)} = \widehat{\mathbf{\Delta}}^{1/2}_{(j)}\widetilde{\mathbf{\Gamma}}$, $\widehat{\boldsymbol{\beta}}_{(j+1)} = \widetilde{\boldsymbol{\beta}}$.

Cook & Forzani (2009*b*) proposed a different algorithm for fitting the diagonal PFC model. The algorithm here is more reliable when $n < p$.

We again report partial results of a simulation study to illustrate the potential benefits of the diagonal PFC model. The scenario for generating the data was the same as that for figure aa with $3 \leq p \leq 200$ and $n = 50$, modified to induce different conditional variances. The conditional variances $(\sigma_1^2, \ldots, \sigma_{200}^2)$ were generated once as the order statistics for a sample of size 200 from 60 times a chi-squared random variable with 2 degrees of freedom. The smallest order statistic was $\sigma_1^2 = 0.07$ and the largest was $\sigma_{200}^2 = 499$. We then used $(\sigma_1^2, \ldots, \sigma_p^2)$ for a regression with p predictors. In this way the conditional variances increased as we added predictors. While it is unlikely that predictors would be so ordered in practice, this arrangement seems useful to highlight the relative behaviour of the methods. In each case we fitted the isotonic and diagonal PFC model using $\mathbf{f}_y = y$. Predictions were again assessed using 200 new simulated observations, and the entire setup was again replicated 100 times to obtain the average prediction errors shown in figure bb. With $p = 3$ predictors all methods perform well because the first three conditional variances are similarly small. The prediction error for the diagonal PFC model changed very little as we added predictors, while the prediction errors for all other methods increased substantially. This happened because the MLEs from the diagonal PFC model can weight each predictor according to its conditional variance, while the predictors are treated equally by the other methods. Predictions from the PC model were consistently the worst. This is perhaps to be expected since the other three methods use mean functions that are appropriate for this simulation.

Predictions from the diagonal PFC model should perform well also when $\mathbf{\Delta} > 0$ is not a diagonal matrix, provided that the conditional predictor correlations are small to moderate. This is in agreement with Prentice & Zhao (1991, p. 830) who recommended in a related context that independence models generally should be adequate for a broad range of applications in which the dependencies are not strong. The methodology discussed in the next section may be useful in the presence of strong conditional correlations.

(*d*) *The PFC model*

In this section we describe the estimator (3.4) for general $\mathbf{\Delta} > 0$, allowing the predictors to be conditionally dependent with different variances. The methods deriving from the isotonic and diagonal PFC models do not require $p < n$. In contrast, the methods of this section work best in data rich regressions where $p \ll n$. The analysis relies on the MLEs given by Cook (2007, §7.2) and studied in more detail by Cook & Forzani (2009*b*).

The following notation will help describe the MLEs under the PFC model. Let $\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}} = \widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}} > 0$ be the sample covariance matrix of the residuals vectors $\mathbf{X}_i - \widehat{\mathbf{X}}_i$ from the OLS fit of $\mathbf{X}$ on $\mathbf{f}$. Let $\widehat{\omega}_1 > \ldots \geq \widehat{\omega}_p$ and $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_p)$ be the eigenvalues and corresponding matrix of eigenvectors $\widehat{\mathbf{v}}_j$ of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}$. Finally, let $\widehat{\mathbf{K}}$ 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 $\widehat{\omega}_{d+1}, \ldots, \widehat{\omega}_p$. Then the MLE of $\boldsymbol{\Delta}$ is

$$\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1/2}\widehat{\mathbf{V}}(\mathbf{I}_p + \widehat{\mathbf{K}})\widehat{\mathbf{V}}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1/2}.$$

If $d = r$ then $\widehat{\mathbf{K}} = 0$ and this estimator reduces to $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}$. Otherwise $\widehat{\boldsymbol{\Delta}}$ recovers information on variation due to overfitting ($r > d$). The MLE of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Gamma}}^T\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Delta}}^{-1})}\mathbf{B}_{\mathrm{ols}}$, where $\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Delta}}^{-1})}\mathbf{B}_{\mathrm{ols}}$ is the projection of $\mathbf{B}_{\mathrm{ols}}$ onto the span of $\widehat{\boldsymbol{\Gamma}}$ in the $\widehat{\boldsymbol{\Delta}}^{-1}$ inner product, and the MLE of $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ is $\mathcal{S}_d(\widehat{\boldsymbol{\Delta}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}})$, which is equal to the SIR estimator of $\mathcal{S}_{Y|X}$ when the response is categorical (Cook & Forzani, 2009*b*).

Turning to prediction, we substituted these MLEs into the multivariate normal density for $\mathbf{X}|Y$ and simplified the resulting expression by in part ignoring proportionality constants not depending on the observation i to obtain the weights (3.4)

$$w_i(\mathbf{x}) \propto \exp\{-(1/2)(\mathbf{x} - \widehat{\mathbf{X}}_i)^T[\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Delta}}^{-1}](\mathbf{x} - \widehat{\mathbf{X}}_i)\}, \tag{4.7}$$

where $\widehat{\mathbf{X}}_i$ is as defined for (4.4). We next discuss how these weights can be simplified to a more intuitive and computationally efficient form.

Let $\widetilde{\mathbf{V}} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\widehat{\mathbf{V}}(\mathbf{I}_p + \widehat{\mathbf{K}})^{-1/2}$. Then it is easy to show that $\widehat{\boldsymbol{\Delta}}^{-1} = \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^T$, and that the columns of $\widehat{\boldsymbol{\Delta}}^{1/2}\widetilde{\mathbf{V}}$ are the normalized eigenvectors of $\widehat{\boldsymbol{\Delta}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}\widehat{\boldsymbol{\Delta}}^{-1/2}$. Let $\widetilde{\mathbf{V}}_d$ and $\widehat{\mathbf{V}}_d$ denote the $p \times d$ matrices consisting of the first d columns of $\widetilde{\mathbf{V}}$ and $\widehat{\mathbf{V}}$. Then, since the MLE of $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ is $\mathcal{S}_d(\widehat{\boldsymbol{\Delta}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}})$, it follows that $\mathcal{S}_d(\widehat{\boldsymbol{\Delta}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}) = \mathrm{span}(\boldsymbol{\Delta}^{-1/2}\boldsymbol{\Delta}^{1/2}\widetilde{\mathbf{V}}) = \mathrm{span}(\widetilde{\mathbf{V}}_d) = \mathrm{span}(\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\widehat{\mathbf{V}}_d)$. Consequently, we may take $\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\boldsymbol{\Gamma}} = \widetilde{\mathbf{V}}_d = \widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\widehat{\mathbf{V}}_d$, which implies the reduction

$$\widehat{R}(\mathbf{x}) = \widehat{\mathbf{V}}_d^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\mathbf{x} = (\widehat{\mathbf{v}}_1^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\mathbf{x}, \ldots, \widehat{\mathbf{v}}_d^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\mathbf{x})^T.$$

Using these results in (4.7) we have $\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\boldsymbol{\Gamma}} = (\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Delta}}^{-1})\widehat{\boldsymbol{\Delta}}(\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\boldsymbol{\Gamma}}) = \widetilde{\mathbf{V}}_d^T\widehat{\boldsymbol{\Delta}}\widetilde{\mathbf{V}}_d = \mathbf{I}_d$ and thus

$$\begin{aligned} w_i(\mathbf{x}) &\propto \exp\{-(1/2)(\mathbf{x} - \widehat{\mathbf{X}}_i)^T[\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Delta}}^{-1}](\mathbf{x} - \widehat{\mathbf{X}}_i)\} \\ &= \exp\{-(1/2)(\mathbf{x} - \widehat{\mathbf{X}}_i)^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\widehat{\mathbf{V}}_d\widehat{\mathbf{V}}_d^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}(\mathbf{x} - \widehat{\mathbf{X}}_i)\} \\ &= \exp\{-(1/2)\|\widehat{R}(\mathbf{x}) - \widehat{R}(\widehat{\mathbf{X}}_i)\|^2\}. \end{aligned} \tag{4.8}$$

The reduction in this case is applied to $\mathbf{x}$ and $\widehat{\mathbf{X}}$ in the same way that it was for the weights from the isotonic and diagonal PFC models, but the reduction itself differs. A perhaps clearer connection between the PFC reduction and the isotonic PFC reduction can be seen by noting that $\mathrm{span}(\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\widehat{\mathbf{V}}_d)$ is equal to the span of the first d eigenvectors of $\widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}$. Under the isotonic PFC model the reduction is computed by using the first d eigenvectors of $\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}$, while under the PFC model the reduction is computed using the first d eigenvectors of $\widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}$. Additionally, since $\mathcal{S}_d(\widehat{\mathbf{\Sigma}}_{\mathrm{res}}, \widehat{\mathbf{\Sigma}}_{\mathrm{fit}}) = \mathcal{S}_d(\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{\Sigma}}_{\mathrm{res}})$ the reduction could be computed also as the first d eigenvectors of $\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}$.

5. Implementation

In this section we present additional simulation results and provide some suggestions for issues that must be addressed prior to implementation. We consider only the isotonic, diagonal and general PFC models and address the implementation issues first.

(*a*) *Choice of d*

Two general methods for choosing $d \in \{0, 1, ..., \min(r, p)\}$ have been studied in literature. One is by using an information criterion like AIC or BIC. Let $L(d_0)$ denote the value of the maximized log likelihood for any of the three inverse models under consideration fitted with a value d_0 for d. Then the dimension is selected that minimizes the information criterion $-2L(d_0) + h(n)g(d_0)$, where $h(n)$ is equal to $\log(n)$ for BIC and 2 for AIC, and $g(d_0) = p + d_0(p - d_0) + dr + D(\mathbf{\Delta})$ is the number of parameters to be estimated as a function of d_0. The first addend in this count corresponds to $\boldsymbol{\mu}$, the second to $\mathcal{S}_{\mathbf{\Gamma}}$, and the third to $\boldsymbol{\beta}$. The final term $D(\mathbf{\Delta})$ is the number of parameters required for $\mathbf{\Delta}$, which is equal to 1, p and $p(p+1)/2$ for the isotonic, diagonal and PFC models. The second method is based on using the likelihood ratio statistic $2(L(\min(r, p)) - L(d_0))$ in a sequential scheme to choose d: Using a common test level and starting with $d_0 = 0$, choose the estimate of d as the hypothesized value that is not rejected. See Cook (2007) and Cook & Forzani (2009*a*) for further discussion of these methods.

These methods for choosing d have been shown to work well in data-rich regressions where $p \ll n$, but can be unreliable otherwise. Consequently, in the context of this article, we will determine d using the cross-validation method proposed by Hastie *et al.* (2001, ch. 7).

(*b*) *Screening*

When dealing with large p regressions there may be a possibility that a substantial subset $\mathbf{X}_2$ of the predictors is inactive; that is, $\mathbf{X}_2$ is independent of the response given the remaining predictors $\mathbf{X}_1$. In terms of PFC models $\mathbf{X}_2$ is inactive if and only if the corresponding rows of $\mathbf{\Delta}^{-1}\mathbf{\Gamma}$ are all equal to zero. Addressing this possibility in terms of the general PFC models requires a data-rich regression, and is outside the scope of this article (see Cook & Forzani 2009*b* for further discussion). For the isotonic and diagonal PFC models the predictors $\mathbf{X}_2$ are inactive if and only if the corresponding rows of $\mathbf{\Gamma}$ are all equal to 0. In these models we can test if the k-th row of $\mathbf{\Gamma}$ is equal to 0 straightforwardly by using an F statistic to test if $\boldsymbol{\alpha} = 0$ in the univariate regression model of the k-th predictor X_k on $\mathbf{f}$, $X_k = \mu + \boldsymbol{\alpha}^T\mathbf{f} + \varepsilon$. This method with a common test level of 0.1 was used for screening in the simulations that follow. When $d = 1$ and $\mathbf{f} = y$ this method is equivalent to sure independence screening proposed by Fan & Lv (2008).

(c) *Simulation results*

To allow for some flexibility, we used the following general setup as the basis for our simulations. Let $\mathbf{J}_k$ and $\mathbf{O}_k$ denote vectors of length k having all elements equal to 1 and 0. A response was generated from a uniform $(0, 3)$ distribution. Then with $\mathbf{f}_y = (y, \mathrm{e}^{2y})^T$ a predictor vector of length p was generated using an integer $p_0 \leq p/2$ as $\mathbf{X}_y = \boldsymbol{\Gamma}^* \boldsymbol{\beta}^* \mathbf{f}_y + \boldsymbol{\varepsilon}$, where $\boldsymbol{\Gamma}^* = (\boldsymbol{\Gamma}_1^*, \boldsymbol{\Gamma}_2^*)$ with $\boldsymbol{\Gamma}_1^* = (\mathbf{J}_{p_0}^T, \mathbf{O}_{p-p_0}^T)^T$ and $\boldsymbol{\Gamma}_2^* = (\mathbf{O}_{p_0}^T, \mathbf{J}_{p_0}^T, \mathbf{O}_{p-2p_0})^T$, $\boldsymbol{\beta}^* = \mathrm{diag}(2/\sqrt{20}, 1/\sqrt{20})$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$ with $\boldsymbol{\Delta} = \mathrm{diag}(\sigma_0^2 \mathbf{J}_{p_0}^T, \sigma_1^2 \mathbf{J}_{p_0}^T, \sigma_0^2 \mathbf{J}_{p-2p_0}^T)$, $\sigma_0^2 = 2$ and $\sigma_1^2 = 20$. Datasets of n independent observations generated in this way have the following properties: (1) one set of p_0 predictors is linearly related to the response, another set of p_0 is nonlinearly related and the remaining $p - 2p_0$ predictors are independent of the response; (2) the conditional covariance of $\mathbf{X}|Y$ has a diagonal structure; (3) $\boldsymbol{\Gamma}$ is a $p \times 2$ matrix and thus the sufficient reduction is composed of $d = 2$ linear combinations of the predictors.

We present results from three cases of this simulation model. In each case we fitted the diagonal PFC model with $\mathbf{f}_y = (y^1, \ldots, y^4)^T$, so the fitted $\mathbf{f}_y$ was not the same as that used to generate the data. The prediction errors were determined using the method of equation (4.3).

Case 1: $p \ll n$**.** In this regression we set $n = 400$, $p = 40$ and $p_0 = 20$, so all predictors are relevant and no screening was used. The results are shown in the third column of table 1. The first row shows the prediction errors when d is determined by the cross-validation method of Hastie *et al.* (2001, ch. 7). In this case the estimated d can vary across the 100 replicate data sets. The next four rows show the prediction errors when d is fixed at the indicated value across all datasets. The final two rows show the prediction errors for PLS and the lasso applied straightforwardly to the linear regression of Y on $\mathbf{X}$ using the R packages 'pls' and 'lars'. We see that the PFC prediction errors are fairly stable across the choices for d and that they are roughly one third of the prediction error for the forward methods.

Case 2: $n < p$**.** In this simulation we set $n = 100$, $p_0 = 60$ and $p = 120$, so again all predictors are relevant and no screening was used. The results, which are shown in the fourth column of table 1, are qualitatively similar to those for Case 1, except the prediction error for the lasso is notably less than that for PLS.

Case 3: $n < p$**.** For this case we set $n = 100$, $p_0 = 20$ and $p = 120$. Now there are only 40 relevant predictors out of 120 total predictors and, as discussed in §5b, screening was used prior to fitting the diagonal PFC model. The results shown in the final column of table 1 are qualitatively similar to the other two cases. Here we were a bit surprised that the lasso did not perform better since it is designed specifically for sparse regressions. Evidently, the adaptability of the diagonal PFC model through the choices of $\mathbf{f}$ and d can be important for good predictions.

(d) *Illustration*

We conclude this section with an illustrative analysis of an economic regression using data from Enz (1991). The data and additional discussion were given by

Table 1. *Prediction errors (4.3) for three simulation scenarios, with standard errors given in parentheses*

		case 1	case 2	case 3
PFC	$\widehat{d}$ by c.v.	0.063 (0.0031)	0.07 (0.006)	0.090 (0.0047)
	$d = 1$	0.073 (0.0036)	0.08 (0.006)	0.092 (0.0053)
	$d = 2$ (true)	0.063 (0.0031)	0.07 (0.005)	0.087 (0.0049)
	$d = 3$	0.065 (0.0031)	0.07 (0.005)	0.089 (0.0046)
	$d = 4$	0.066 (0.0030)	0.08 (0.005)	0.093 (0.0045)
forward methods	lasso	0.18 (0.002)	0.19 (0.003)	0.26 (0.005)
	PLS	0.18 (0.002)	0.27 (0.004)	0.27 (0.004)

Table 2. *BigMac dataset*

methods	prediction error
PFC - cubic polynomial	933
PFC - piecewise constant	771
elastic net	1198
ridge regression	1211
lasso	1412
PLS	1426
OLS	2268

Cook & Weisberg (1999, p. 140). The response, which was measured in 1991, is the average minutes of labour required to buy a BigMac hamburger and French fries in $n = 45$ world cities. There are nine continuous predictors. We fitted the diagonal PFC model as discussed in §4c, and estimated the prediction error using leave-one-out cross-validation. We set $d = 1$ to allow an evenhanded comparison with forward methods.

The results are shown in table 2. Two entries are given for PFC. The first corresponds to fitting PFC with the cubic polynomial basis $\mathbf{f} = (y, y^2, y^3)^T$ and the second uses a more elaborate piecewise constant polynomial with ten slices for $\mathbf{f}$. PFC with the piecewise constant polynomial basis gives a better prediction error than PFC with the cubic polynomial basis, but both methods outperform the predictions by five forward methods, the elastic net (Zou & Hastie 2005), the lasso, ridge regression, partial least squares (PLS) and ordinary least squares (OLS). The relative behaviour of PFC illustrated here is typical of our experiences, regardless of the number of predictors.

6. Discussion

The methodology introduced in this article applies across a useful range of applications with continuous predictors that, given the response, are mildly correlated. We view the ability to choose $\mathbf{f}_y$ as an important feature, since it allows a level of adaptability that is not really available when directly modelling the mean function $\mathrm{E}(Y|\mathbf{X})$ in large-p regressions. The assumption of normal errors is not crucial, but the predictions may not perform well under extreme deviations from normality. For instance, we expect that it is possible to develop substantially better methodology for regressions in which all the predictors are binary, perhaps employing a quadratic exponential family for $\mathbf{X}|Y$ in place of the multivariate normal. Work along these lines is in progress.

It also seems possible to develop models 'between' the diagonal and general PFC models that allow for some of the conditional predictor correlations to be substantial but stop short of permitting a general $\boldsymbol{\Delta} > 0$. One route is to model the conditional covariance matrix as $\boldsymbol{\Delta} = \boldsymbol{\Gamma}\mathbf{M}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\mathbf{M}_0\boldsymbol{\Gamma}_0^T$, where $\boldsymbol{\Gamma}$ is the same as that in model (4.1), $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ is an orthogonal $p \times p$ matrix, $\mathbf{M} > 0$ and $\mathbf{M}_0 > 0$ (Cook 2007). The minimal sufficient reduction in this case is $R(\mathbf{X}) = \boldsymbol{\Gamma}^T\mathbf{X}$, which does not involve the $\mathbf{M}$s. This model for $\boldsymbol{\Delta}$ allows arbitrary correlations among the elements of R and among the elements of $\boldsymbol{\Gamma}_0^T\mathbf{X}$, but requires that R and $\boldsymbol{\Gamma}_0^T\mathbf{X}$ be conditionally independent. Predictions under this model in a data rich regression can be obtained as an extension of results by Cook (2007). The behaviour of this model in $n < p$ regressions is under study.

Research for this article was supported in part by the U.S. National Science Foundation, Microsoft Research and the Newton Institute for Mathematical Sciences, Cambridge, U.K.

References

Adcock, R. J. 1878 A problem in least squares. *The Analyst* **5**, 53–54.

Alter, O., Brown, P. & Botstein, D. 2000 Singular value decomposition for gene-wide expression data processing and modelling. *Proc. Nat. Acad. of Sciences* **97**, 10101–10106.

Cook, R. D. 1994 Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proc. Sect. Phys. Eng. Sc.*, pp. 18-25. Alexandria, VA: American Statistical Association.

Cook, R. D. 1996 Graphics for regressions with a binary response. *J. Amer. Statist. Ass.* **91**, 983–992.

Cook, R. D. 1998 *Regression Graphics.* New York: Wiley.

Cook, R. D. 2007 Fisher Lecture: Dimension Reduction in Regression (with discussion). *Statistical Science* **22**, 1–26.

Cook, R. D. & Forzani, L. 2009*a* Likelihood-based sufficient dimension reduction. *J. Amer. Statist. Ass.*, to appear.

Cook, R. D. & Forzani, L. 2009*b* Principal fitted components for dimension reduction in regression. *Statistical Science*, to appear.

Cook. R. D., Li. B. & Chiaromonte, F. 2007 Dimension reduction without matrix inversion. *Biometrika* **94**, 569–584.

Cook, R. D. & Ni, L. 2005 Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Ass.* **100**, 410–428.

Cook, R. D. & Weisberg, S. 1982 *Residuals and Influence in Regression.* London: Chapman and Hall. (Available on line at http://www.stat.umn.edu/rir)

Cook, R. D. & Weisberg, S. 1991 Discussion of 'Sliced inverse regression for dimension reduction' by K. C. Li. *J. Amer. Statist. Ass.* **86**, 328–332.

Cook, R. D. & Weisberg, S. 1999 *Applied Regression Including Computing and Graphics.* New York: Wiley.

Cox, D. R. 1968 Notes on some aspects of regression analysis. *J. R. Statist. Soc., A* **131**, 265–279.

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

Fan, J. & Lv, J. 2008 Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc., B* **70**, 849–911.

Fisher, R. J. 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transact. R. Statist. Soc. A* **222**, 309–368.

Fukumizu, K., Bach, F. R. & Jordan, M. I. 2009 Kernel dimension reduction in regression. *Ann. Statist.*, to appear.

Hastie, T., Tibshirani, R. & Friedman, J. 2001 *The Elements of Statistical Learning - Data Mining, Inference, and Prediction.* New York: Springer.

Helland, I. S. 1990 Partial least squares regression and statistical models. *Scand. J. Statist.* **17**, 97–114.

Johnson, O. 2008 Theoretical properties of Cook's PFC dimension reduction algorithm for linear regression. *Electronic J. Statist.* **2**, 807–827.

Leek, J. T. & Storey, J. D. 2007 Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLos Genetics* **3**, 1724–1735.

Li, K-C. 1991 Sliced inverse regression for dimension reduction. *J. Amer. Statist. Ass.* **86**, 316–342.

Li, L. 2007 Sparse sufficient dimension reduction. *Biometrika* **94**, 603–613.

Li, L. & Yin, X. 2008 Sliced inverse regression with regularizations. *Biometrics* **64**, 124–131.

Li, B. & Wang, S. 2007 On directional regression for dimension reduction. *J. Amer. Statist. Ass.* **102**, 997–1008.

Naik, P. & Tsai, C-L. 2000 Partial least squares estimator for single-index models. *J. R. Statist. Soc., B* **62**, 763–771.

Prentice, R. L. & Zhao, L. P. 1991 Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825–838.

Simonoff, J. S. (1996): *Smoothing Methods in Statistics.* New York: Springer.

Tibshirani, R. 1996 Regression shrinkage and selection via the lasso. *J. R. Statist. Soc., B* **58**, 267-288.

Wand, M. P. & Jones, M. C. 1995 *Kernel Smoothing.* London: Chapman and Hall.

Ye, Z. & Weiss, R. 2003 Using the Bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Ass.* **98**, 968–978.

Zou, H. & Hastie, T. 2005 Regularization and variable selection via the elastic net. *J. R. Statist. Soc., B* **68**, 301–320.

Zhu, L., Ohtaki, M. & Li, Y. 2005 On hybrid methods of inverse regression-based algorithms. *Computational Statistics and Data Analysis* **51**, 2621–2635.