J. R. Statist. Soc. B (2002)
64, *Part* 2, *pp.* 159–175

Dimension reduction for the conditional *k*th moment in regression

Xiangrong Yin

University of Georgia, Athens, USA

and R. Dennis Cook

University of Minnesota, St Paul, USA

[Received May 2000. Final revision September 2001]

Summary. The idea of dimension reduction without loss of information can be quite helpful for guiding the construction of summary plots in regression without requiring a prespecified model. Central subspaces are designed to capture all the information for the regression and to provide a population structure for dimension reduction. Here, we introduce the central *k*th-moment subspace to capture information from the mean, variance and so on up to the *k*th conditional moment of the regression. New methods are studied for estimating these subspaces. Connections with sliced inverse regression are established, and examples illustrating the theory are presented.

Keywords: Central subspaces; Dimension reduction subspaces; Permutation tests; Regression graphics; Sliced inverse regression

1. Introduction

In simple regression a two-dimensional plot of the response Y *versus* the predictor X displays all the sample information and can be quite helpful for gaining insights about the data and for guiding the choice of a first model. Such straightforward graphical displays of all the data are not generally possible with many predictors, but informative displays are still possible in situations where we can find low dimensional views, the only ones that are possible in practice, that provide 'sufficient' information about the regression. In regression graphics we seek to facilitate a visualization of the data by reducing the dimension of the $p \times 1$ predictor vector $\mathbf{X}$ without loss of information on the regression and without requiring a prespecified parametric model. We called this sufficient dimension reduction, borrowing terminology from classical statistics. Sufficient dimension reduction leads naturally to the idea of a *sufficient summary plot* that contains all of the information on the regression that is available from the sample.

As reviewed in Sections 2.1 and 2.2, *dimension reduction subspaces* (DRSs) and, in particular, *central subspaces* $\mathcal{S}_{Y|\mathbf{X}}$ are useful population foundations for pursuing sufficient dimension reduction in regression. A basic goal here is to estimate the fewest linear combinations $\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{X}, \ldots, \boldsymbol{\eta}_d^{\mathrm{T}}\mathbf{X}$, $d \leqslant p$, with the property that $Y|\mathbf{X}$ and $Y|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ have the same conditional distribution, where $\boldsymbol{\eta}$ is the $p \times d$ matrix $(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d)$. Thus, $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ contains all the information about the conditional distribution of $Y|\mathbf{X}$ that is available from $\mathbf{X}$ and, if $\boldsymbol{\eta}$ is known, a sufficient summary plot of Y *versus* $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ can be used to guide the analysis.

Address for correspondence: R. Dennis Cook, Department of Applied Statistics, University of Minnesota, 1994 Buford Avenue, St Paul, MN 55108, USA.

© 2002 Royal Statistical Society 1369–7412/02/64159

No model for $Y|\mathbf{X}$ is required. Methodology for estimating sufficient summary plots has proven to be particularly useful during model development and during model criticism where the response is replaced with a residual.

In Section 3 we introduce a new concept—*the central kth-moment subspace* (CKMS) $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$—and study its basic properties. In Section 4 we propose a method called cov_k to estimate the directions in $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$. The main goal of this new approach is to estimate the fewest linear combinations of $\mathbf{X}$ that contain all the information on just the first k moments of $Y|\mathbf{X}$. This shift in emphasis from the complete conditional distribution of $Y|\mathbf{X}$ to its first k moments distinguishes the new approach from previous work based on the central subspace.

There are at least three reasons that focusing attention on moments of $Y|\mathbf{X}$ might be useful in practice. First, the mean and variance functions are often of special interest. Our new approach allows investigators the opportunity to tailor a dimension reduction inquiry to those moments. Second, although we might be able to gain information on the first few moments of $Y|\mathbf{X}$ by using known methods for pursuing the distribution $Y|\mathbf{X}$, having the ability to focus on those moments should in principle produce better results. Finally, by contrasting results from known methods for studying $Y|\mathbf{X}$ with results on the first few moments, we may be able to tell whether the dependence of Y on $\mathbf{X}$ is confined substantially to those moments or whether higher order moments are involved.

A connection with sliced inverse regression (SIR) is presented in Section 5. In Section 6 we use the new method to study a real data set and present a few related simulation results. We put all the proofs in Appendix A.

We assume throughout that the scalar response Y and the $p \times 1$ vector of predictors $\mathbf{X}$ have a joint distribution, and that the data $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, are independent and identically distributed observations on $(Y, \mathbf{X})$ with finite moments. The goal of the regression is to infer, as far as possible with the data available, about the conditional distribution of $Y|\mathbf{X}$ with emphasis on its moments. The notation $\mathbf{U} \perp\!\!\!\perp \mathbf{V}|\mathbf{Z}$ means that the random vectors $\mathbf{U}$ and $\mathbf{V}$ are independent given any value for the random vector $\mathbf{Z}$. Subspaces will be denoted by $\mathcal{S}$, and $\mathcal{S}(\mathbf{B})$ means the subspace of $\mathcal{R}^t$ spanned by the columns of $t \times u$ matrix $\mathbf{B}$. $P_{\mathbf{B}}$ denotes the projection operator for $\mathcal{S}(\mathbf{B})$ with respect to the usual inner product and $Q_{\mathbf{B}} = I - P_{\mathbf{B}}$.

The data which are analysed in this paper can be obtained from

```
http://www.blackwellpublishers.co.uk/rss/
```

2. The central subspace

2.1. Definitions

Let $\mathbf{B}$ denote a fixed $p \times q$, $q \leqslant p$, matrix so that

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^{\mathrm{T}}\mathbf{X}. \tag{1}$$

This statement is equivalent to saying that the distribution of $Y|\mathbf{X}$ is the same as that of $Y|\mathbf{B}^{\mathrm{T}}\mathbf{X}$ for all possible values of $\mathbf{X}$. It implies that the $p \times 1$ predictor vector $\mathbf{X}$ can be replaced by the $q \times 1$ predictor vector $\mathbf{B}^{\mathrm{T}}\mathbf{X}$ without loss of regression information and thus represents a potentially useful reduction in the dimension of the predictor vector.

If condition (1) holds then it also holds when $\mathbf{B}$ is replaced with any matrix whose columns form a basis for $\mathcal{S}(\mathbf{B})$. Thus, condition (1) is appropriately viewed as a statement about $\mathcal{S}(\mathbf{B})$ which is called a DRS for the regression of Y on $\mathbf{X}$ (Li, 1991). Knowledge of the smallest DRS could be useful for parsimoniously characterizing how the distribution of $Y|\mathbf{X}$ changes with the value of $\mathbf{X}$.

Let $\mathcal{S}_{Y|\mathbf{X}}$ denote the intersection of all DRSs. Although $\mathcal{S}_{Y|\mathbf{X}}$ is always a subspace, it is not necessarily a DRS. Nevertheless, $\mathcal{S}_{Y|\mathbf{X}}$ is a DRS under various reasonable conditions (Cook, 1994a, 1996, 1998). In this paper, $\mathcal{S}_{Y|\mathbf{X}}$ is assumed to be a DRS and, following Cook (1994b), is called the *central* DRS, or simply the *central subspace*. The dimension $\dim(\mathcal{S}_{Y|\mathbf{X}})$ is the *structural dimension* of the regression; regressions are identified as having zero-dimensional, one-dimensional, ... structure.

The central subspace, which is taken as the inferential object for the regression, is the smallest DRS such that $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$, where the columns of the matrix $\boldsymbol{\eta}$ form a basis for the subspace. In effect, the central subspace is a metaparameter that can be used to characterize the regression of Y on $\mathbf{X}$. If an estimated basis $\hat{\boldsymbol{\eta}}$ of $\mathcal{S}_{Y|\mathbf{X}}$ is available then the summary plot of Y *versus* $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ can be used to guide subsequent analysis. For additional background on these ideas, see Cook (1998) and Cook and Weisberg (1999a).

Let $\Sigma_{\mathbf{X}} = \mathrm{var}(\mathbf{X})$, which is assumed to be non-singular, and let $\mathbf{Z} = \Sigma_{\mathbf{X}}^{-1/2}\{\mathbf{X} - E(\mathbf{X})\}$ be the standardized predictor. Then $\mathcal{S}_{Y|\mathbf{X}} = \Sigma_{\mathbf{X}}^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}$ (Cook (1998), proposition 6.1). The columns of the matrix $\boldsymbol{\gamma} = \Sigma_{\mathbf{X}}^{1/2}\boldsymbol{\eta}$ form a basis for $\mathcal{S}_{Y|\mathbf{Z}}$, the central subspace for the regression of Y on $\mathbf{Z}$.

2.2. Estimation methods

Many of the methods for estimating vectors in the central subspace are based on the following idea. Let $\hat{\mathbf{K}}$ be a consistent estimate of a population kernel matrix $\mathbf{K}$ with the property $\mathcal{S}(\mathbf{K}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. Then $\mathcal{S}(\mathbf{K})$ can be estimated as the span of the left singular vectors of $\hat{\mathbf{K}}$ whose singular values are inferred to be non-zero in the population.

Various ways of choosing $\mathbf{K}$ stem from the inverse regression of $\mathbf{Z}$ on Y and require the condition that $E(\mathbf{Z}|\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}) = P_{\boldsymbol{\gamma}}\mathbf{Z}$ where the columns of the matrix $\boldsymbol{\gamma}$ form a basis for $\mathcal{S}_{Y|\mathbf{Z}}$. This condition, which applies to the marginal distribution of the predictors, is equivalent to requiring that $E(\mathbf{Z}|\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})$ be linear, so we refer to it as the *linearity condition*. The linearity condition must hold only for the basis $\boldsymbol{\gamma}$ of the central subspace. Since $\boldsymbol{\gamma}$ is unknown, in practice we may require that the linearity condition holds for all possible $\boldsymbol{\gamma}$, which is equivalent to elliptical symmetry of the distribution of $\mathbf{Z}$ (Eaton, 1986). In particular, application is often facilitated when the predictors are normally distributed, as we shall see later in Section 4.2.2. Simultaneous power transformations $T(\mathbf{X}) = (t_j(X_j))$ of positive predictors $\mathbf{X}$ or weighting (Cook and Nachtsheim, 1994) are often effective for inducing near multivariate normality. Because the conditional distribution of $Y|\mathbf{X}$ is the same as that of $Y|T(\mathbf{X})$ and because no model is assumed for $Y|\mathbf{X}$, the role of a normalizing predictor transformation is to facilitate the analysis by changing the way that the conditional distributions are indexed by $\mathbf{X}$. Hall and Li (1993) showed that, as p increases, the linearity condition is satisfied for nearly every $p \times 1$ vector $\boldsymbol{\gamma}$. Placing a prior distribution on $\boldsymbol{\gamma}$, they then argued that there is a high probability that the linearity condition will be reasonable.

The *inverse mean subspace* is defined as

$$\mathcal{S}_{E(\mathbf{Z}|Y)} = \mathrm{span}\{E(\mathbf{Z}|Y), Y \in R(Y)\},$$

where $R(Y)$ is the support of Y. Under the linearity condition $\mathcal{S}_{E(\mathbf{Z}|Y)} \subseteq \mathcal{S}_{Y|\mathbf{Z}}$, and thus an estimate of a portion of the central subspace can be constructed by estimating $\mathcal{S}_{E(\mathbf{Z}|Y)}$. SIR (Li, 1991) is perhaps the most well-known method for estimating $\mathcal{S}_{E(\mathbf{Z}|Y)}$. It is based on the kernel matrix

$$\mathbf{K} = E\{E(\mathbf{Z}|\tilde{Y})\, E(\mathbf{Z}|\tilde{Y})^{\mathrm{T}}\}$$

where $\tilde{Y}$ is a discrete version of Y obtained by partitioning its range. This kernel matrix is the basis for estimating the inverse mean subspace $\mathcal{S}_{E(\mathbf{Z}|\tilde{Y})} \subseteq \mathcal{S}_{\tilde{Y}|\mathbf{Z}} \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. Although modifications of SIR have been proposed (see, for example, Schott (1994) and Velilla (1998)) they all share the estimation of $\mathcal{S}_{E(\mathbf{Z}|Y)}$ as a common goal. Recently, Bura and Cook (2001) studied parametric inverse regression methods for estimating $\mathcal{S}_{E(\mathbf{Z}|Y)}$; these methods do not require the linearity condition or slicing the response.

In the next section we begin our study of CKMSs and methods for estimating them.

3. *k*th-moment dimension reduction subspaces

The central subspace is designed to capture the entire conditional distribution of $Y|\mathbf{X}$ and thereby to give a full picture of the dependence of Y on $\mathbf{X}$. Cook and Li (2001) introduced a new class of subspaces, which they called *central mean subspaces*, and studied their properties. Following this direction, we now introduce *kth-moment DRSs* and *central kth-moment DRSs*, and study their properties. The construction of a CKMS is similar to that of the central subspace, but dimension reduction is aimed at reducing the mean function, variance function and up to the kth moment function, leaving the rest of $Y|\mathbf{X}$ as the 'nuisance parameter'.

When considering conditional moments, dimension reduction hinges on finding a $p \times q$ matrix $\boldsymbol{\eta}$, $q \leqslant p$, so that the random vector $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ contains all the information about Y that is available from $E(Y|\mathbf{X})$, $\mathrm{var}(Y|\mathbf{X}), \ldots, M^{(k)}(Y|\mathbf{X})$, where $M^{(k)}(Y|\mathbf{X}) = E[\{Y - E(Y|\mathbf{X})\}^k|\mathbf{X}]$ for $k \geqslant 2$. For notational convenience, $M^{(1)}(Y|\mathbf{X})$ stands for $E(Y|\mathbf{X})$.

Definition 1. If

$$Y \perp\!\!\!\perp \{M^{(1)}(Y|\mathbf{X}), \ldots, M^{(k)}(Y|\mathbf{X})\}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X},$$

then $\mathcal{S}(\boldsymbol{\eta})$ is a kth-moment DRS for the regression of Y on $\mathbf{X}$.

It follows from this definition that a DRS is necessarily a kth-moment DRS, which in turn is necessarily an ith-moment DRS for any $i \leqslant k$. Also, if $k = 1$, then $\mathcal{S}(\boldsymbol{\eta})$ is a mean DRS (Cook and Li, 2001), whereas, when $k \to \infty$, $\mathcal{S}(\boldsymbol{\eta})$ is a DRS when the moment-generating function of $Y|\mathbf{X}$ exists. In such a case, letting $k \to \infty$ is equivalent to requiring $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$.

The following proposition gives equivalent conditions for the conditional independence that was used in definition 1.

Proposition 1. The following statements are equivalent.

(a) $Y \perp\!\!\!\perp \{M^{(1)}(Y|\mathbf{X}), \ldots, M^{(k)}(Y|\mathbf{X})\}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$.
(b) $\mathrm{cov}\{Y^j, M^{(j)}(Y|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} = 0$, for $j = 1, \ldots, k$.
(c) $M^{(j)}(Y|\mathbf{X})$ is a function of $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$. Equivalently, $E(Y^j|\mathbf{X})$ is a function of $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ for $j = 1, \ldots, k$.
(d) $\mathrm{cov}\{Y^j, f(\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} = 0$ for $j = 1, \ldots, k$ and any function $f(\mathbf{X})$.

This proposition is a generalization of proposition 1 of Cook and Li (2001), but here part (d) is an additional equivalent condition.

Paralleling the development of central subspaces, we would like to have the smallest kth-moment DRS, as seen in the next definition.

Definition 2. Let $\mathcal{S}^{(k)}_{Y|\mathbf{X}} = \cap \mathcal{S}^{(k)}$ where the intersection is over all kth-moment DRSs $\mathcal{S}^{(k)}$. If $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ is itself a kth-moment DRS, it is called the central kth-moment DRS, or simply the CKMS.

The CKMS does not always exist, because the intersection of two kth-moment DRSs is not necessarily a kth-moment DRS. However, the existence of the CKMS can be guaranteed under various mild conditions that also guarantee the existence of the central subspace. For example, if the support of $\mathbf{X}$ is open and convex, then the CKMS and the central subspace exist. We do not view existence as a crucial practical issue and thus we assume throughout the rest of this paper that the CKMS exists.

The CKMS is always contained in the central subspace $\mathcal{S}_{Y|\mathbf{X}}^{(k)} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ because the former is the intersection of a larger collection of subspaces. Naturally, the following relationships hold:

$$\mathcal{S}_{Y|\mathbf{X}}^{(1)} \subseteq \ldots \subseteq \mathcal{S}_{Y|\mathbf{X}}^{(k)} \subseteq \ldots \subseteq \mathcal{S}_{Y|\mathbf{X}}$$

and

$$\mathcal{S}_{Y|\mathbf{X}} = \lim_{k\to\infty} (\mathcal{S}_{Y|\mathbf{X}}^{(k)}).$$

If $Y|\mathbf{X}$ depends only on moments up to the kth, then $\mathcal{S}_{Y|\mathbf{X}}^{(k)} = \mathcal{S}_{Y|\mathbf{X}}$. In fact, many common regression models for $Y|\mathbf{X}$ have $\mathcal{S}_{Y|\mathbf{X}}^{(2)} = \mathcal{S}_{Y|\mathbf{X}}$.

Let $\mathbf{W} = \mathbf{A}^{\mathrm{T}}\mathbf{X}$ for some invertible matrix $\mathbf{A}$. Then $\mathcal{S}_{Y|\mathbf{W}}^{(k)} = \mathbf{A}^{-1}\mathcal{S}_{Y|\mathbf{X}}^{(k)}$. Consequently, there is no loss of generality in standardizing $\mathbf{X}$ to have mean 0 and identity covariance matrix. In most of the subsequent developments then we work in terms of the standardized predictor $\mathbf{Z}$ defined in Section 2.1. Also, $\mathcal{S}_{Y|\mathbf{X}}^{(k)} = \Sigma_{\mathbf{X}}^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}^{(k)}$, which is the same as the relationship $\mathcal{S}_{Y|\mathbf{X}} = \Sigma_{\mathbf{X}}^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}$ between central subspaces that we mentioned previously.

4. Covariance methods

In this section we develop methods for finding directions in the CKMS.

4.1. Population structure

The following proposition indicates how to find vectors in $\mathcal{S}_{Y|\mathbf{Z}}^{(k)}$ from the covariance between $\mathbf{Z}$ and a polynomial in Y.

Proposition 2. Let γ be a basis for $\mathcal{S}_{Y|\mathbf{Z}}^{(k)}$, and assume that $E(\mathbf{Z}|\gamma^{\mathrm{T}}\mathbf{Z})$ is linear. If $f^{(k)}(Y)$ is any at most kth-degree polynomial of Y, then

$$E\{f^{(k)}(Y)\mathbf{Z}\} \in \mathcal{S}_{Y|\mathbf{Z}}^{(k)} \subseteq \mathcal{S}_{Y|\mathbf{Z}}.$$

In particular, we can take $f^{(k)}(Y) = Y^k$.

There are many ways of selecting $f^{(k)}(Y)$, but it seems that Y^k is a simple and convenient choice. The following two examples show the potential usefulness of this proposition.

4.1.1. Example 1

Let z_1, z_2, z_3 and ε be independent and identically distributed standard normal random variables, and let $Y = z_1 + z_1 z_2 + \varepsilon$. Then $E(Y\mathbf{Z}) = (1, 0, 0)^{\mathrm{T}}$, but

$$\begin{aligned} E(Y^2\mathbf{Z}) &= \mathrm{cov}\{\mathrm{var}(Y|\mathbf{Z}), \mathbf{Z}\} + \mathrm{cov}\{E(Y|\mathbf{Z})^2, \mathbf{Z}\} \\ &= (0, 0, 0)^{\mathrm{T}} + (0, 2, 0)^{\mathrm{T}}. \end{aligned}$$

Here $\mathcal{S}_{Y|\mathbf{Z}}^{(2)} = \mathcal{S}_{Y|\mathbf{Z}}$, which is two dimensional. While the population ordinary least squares (OLS) coefficient vector $E(Y\mathbf{Z})$ finds one direction, the method of proposition 2 finds the other.

4.1.2. Example 2

Let z_1, z_2, z_3 and ε be independent and identically distributed standard normal random variables, and let $Y = 1 + z_1 + \exp(z_2)\varepsilon$. Thus $E(Y\mathbf{Z}) = (1, 0, 0)^{\mathrm{T}}$, but

$$\begin{aligned} E(Y^2\mathbf{Z}) &= \operatorname{cov}\{\operatorname{var}(Y|\mathbf{Z}), \mathbf{Z}\} + \operatorname{cov}\{E(Y|\mathbf{Z})^2, \mathbf{Z}\} \\ &= (0, 2\mathrm{e}^2, 0)^{\mathrm{T}} + (2, 0, 0)^{\mathrm{T}} \\ &= (2, 2\mathrm{e}^2, 0)^{\mathrm{T}}. \end{aligned}$$

Here $\mathcal{S}^{(1)}_{Y|\mathbf{Z}} \subset \mathcal{S}^{(2)}_{Y|\mathbf{Z}} = \mathcal{S}_{Y|\mathbf{Z}}$. While OLS finds the direction in $\mathcal{S}^{(1)}_{Y|\mathbf{Z}}$, the method of proposition 2 finds the other direction, which is in $\mathcal{S}^{(2)}_{Y|\mathbf{Z}}$ but not in $\mathcal{S}^{(1)}_{Y|\mathbf{Z}}$.

Proposition 2 can be used to develop methodology for estimating a CKMS. Define the population kernel matrix $\mathbf{K} = (E(Y\mathbf{Z}), \ldots, E(Y^k\mathbf{Z}))$ and the corresponding *covariance subspace*

$$\mathcal{S}^{(k)}_{\mathrm{cov}} = \mathcal{S}(\mathbf{K}). \tag{2}$$

Then, under proposition 2, $\mathcal{S}^{(k)}_{\mathrm{cov}} \subseteq \mathcal{S}^{(k)}_{Y|\mathbf{Z}}$. Thus the subspace spanned by the left singular vectors of $\mathbf{K}$ corresponding to its non-zero singular values is a subspace of $\mathcal{S}^{(k)}_{Y|\mathbf{Z}}$. For application in practice, let $\hat{\mathbf{K}}$ be the straightforward moment estimate of $\mathbf{K}$. Then, if $d = \dim(\mathcal{S}^{(k)}_{\mathrm{cov}})$ is known, the subspace spanned by the left singular vectors of $\hat{\mathbf{K}}$ corresponding to the d largest singular values is a consistent estimate of $\mathcal{S}^{(k)}_{\mathrm{cov}}$. Usually we also require a method for inference about d. Such methods are addressed in the next section.

The kernel matrix $\mathbf{K}$ is one of many that could be used. For instance, we usually use the corresponding kernel matrix with Y centred and scaled:

$$\mathbf{K}_{\mathrm{c}} = (E(W\mathbf{Z}), \ldots, E(W^k\mathbf{Z})),$$

where $W = \{Y - E(Y)\}/\sigma(Y)$ and $\sigma(Y) = \surd\operatorname{var}(Y)$. Here the response is standardized in $\mathbf{K}_{\mathrm{c}}$ perhaps to induce some numerical stability and so all calculations will be invariant to unit changes in Y. Because $\mathcal{S}(\mathbf{K}_{\mathrm{c}}) = \mathcal{S}(\mathbf{K}) = \mathcal{S}^{(k)}_{\mathrm{cov}}$, methods based on $\mathbf{K}_{\mathrm{c}}$ should give results that are equivalent to those based on $\mathbf{K}$. This conclusion follows from the relationship $\mathbf{K}_{\mathrm{c}} = \mathbf{K}\mathbf{U}$, where $\mathbf{U}$ is a $k \times k$ non-singular, upper triangular matrix.

4.2. Methodology

In general, we are interested in the subspace

$$\mathcal{S}[E\{f^{(k)}_i(Y)\mathbf{Z}\}, i = 1, \ldots, h] \tag{3}$$

where $f^{(k)}_1(Y), \ldots, f^{(k)}_h(Y)$ are $h \leqslant \min(p, k)$ linearly independent known polynomials having up to kth degree. Usually we take $\min(p, k)$ as the default value of h. Particular interest is placed on the special case,

$$\mathcal{S}(\mathbf{K}_{\mathrm{c}}) = \mathcal{S}^{(k)}_{\mathrm{cov}}. \tag{4}$$

Our goal is to estimate the dimensions of these subspaces by using the singular values of corresponding sample kernel matrices, as mentioned in the previous section. Although expression (4) is the subspace that we normally pursue in practice, the results of the next section are presented in terms of expression (3) to allow for generality.

4.2.1. General known polynomial $f^{(k)}(Y)$

Let $\hat{\Sigma}_{\mathbf{X}}$ denote the usual estimate of $\Sigma_{\mathbf{X}}$, and define the standardized observations

$$\hat{\mathbf{Z}}_i = \hat{\Sigma}_{\mathbf{X}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}}), \qquad i = 1, \ldots, n,$$

where $\bar{\mathbf{X}}$ is the sample mean of the predictor vector. Define

$$\begin{aligned}
\mathbf{f}(Y) &= (f_1^{(k)}(Y), \ldots, f_h^{(k)}(Y))^{\mathrm{T}},\\
\mathbf{M}_{\mathbf{f}} &= E\{\mathbf{f}(Y)\},\\
M_i &= E\{f_i^{(k)}(Y)\},\\
\mathbf{K}_h &= (E[\{f_1^{(k)}(Y) - M_1\}\mathbf{Z}], \ldots, E[\{f_h^{(k)}(Y) - M_h\}\mathbf{Z}])\\
&= (E\{f_1^{(k)}(Y)\mathbf{Z}\}, \ldots, E\{f_h^{(k)}(Y)\mathbf{Z}\}).
\end{aligned}$$

The sample kernel matrix corresponding to $\mathbf{K}_h$ is

$$\hat{\mathbf{K}}_h = \left(\frac{1}{n}\sum_{i=1}^{n} f_1^{(k)}(Y_i)\hat{\mathbf{Z}}_i, \ldots, \frac{1}{n}\sum_{i=1}^{n} f_h^{(k)}(Y_i)\hat{\mathbf{Z}}_i\right).$$

Assume that $d = \dim\{\mathcal{S}(\mathbf{K}_h)\}$ is known. Let $\hat{s}_1 \geqslant \hat{s}_2 \geqslant \ldots \geqslant \hat{s}_h$ be the singular values of $\hat{\mathbf{K}}_h$ and let $\hat{l}_1, \ldots, \hat{l}_h$ be the corresponding left singular vectors. Then $\hat{\mathcal{S}}(\mathbf{K}_h) = \mathcal{S}(\hat{l}_1, \ldots, \hat{l}_d)$. Otherwise, inference about d is still required for use in practice. The linear combination $\hat{\boldsymbol{\eta}}_j^{\mathrm{T}}\mathbf{X}$, where $\hat{\boldsymbol{\eta}}_j = \hat{\Sigma}_{\mathbf{X}}^{-1/2}\hat{l}_j$, will be called the jth cov$_k$ predictor.

We base our estimate of d on the idea suggested by Li (1991). The statistic

$$\hat{\Lambda}_m = n \sum_{j=m+1}^{h} \hat{s}_j^2$$

is used as follows to estimate d. Beginning with $m = 0$, compare $\hat{\Lambda}_m$ with the percentage points of its distribution under the hypothesis $d = m$ and determine the p-value $\hat{p}_m$, which is the probability of exceeding the observed value of $\hat{\Lambda}_m$. If $\hat{p}_m$ is larger than a selected cut-off there is not sufficient information to contradict the hypothesis. If it is smaller, conclude that $d > m$, increment m by 1 and repeat the procedure. The estimate $\hat{d} = m$ follows when $\hat{p}_{m-1}$ is relatively small, implying that $d > m - 1$, whereas $\hat{p}_m$ is relatively large, so there is no information to contradict the hypothesis. The estimate of $\mathcal{S}(\mathbf{K}_h)$ is then given by $\mathcal{S}\{\hat{l}_1, \ldots, \hat{l}_{\hat{d}}\}$. These vectors can be back transformed to $\hat{\boldsymbol{\eta}}_j$ for consideration in the original scale.

Either the asymptotic distribution of $\hat{\Lambda}_d$ or a nonparametric alternative is required to implement this procedure in practice. Here we describe a permutation test that can be used to infer about d. The idea was suggested by Cook and Weisberg (1991) and studied further by Cook and Yin (2001). Background on permutation tests was given by Davison and Hinkley (1997) and Efron and Tibshirani (1993).

4.2.2. *Permutation test*

Assume without loss of generality that the kernel matrix $\mathbf{K}$ is a $p \times p$ positive semidefinite symmetric matrix. Starting with a non-symmetric kernel $\mathbf{A}$, for example, we can set $\mathbf{K} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. The test statistic is then

$$\hat{\Lambda}_m = n \sum_{j=m+1}^{p} \hat{\lambda}_j,$$

where the $\hat{\lambda}_j$ are the eigenvalues of $\hat{\mathbf{K}}$.

Let $\mathbf{U} = (\mathbf{u}_j)$ denote the $p \times p$ matrix of eigenvectors $\mathbf{u}_j$ of the population kernel matrix $\mathbf{K}$, let $d = \dim\{\mathcal{S}(\mathbf{K})\}$ and assume that $\mathcal{S}(\mathbf{K}) = \mathcal{S}_{Y|\mathbf{Z}}$. Consider testing the hypothesis that $d \leqslant m$

versus $d > m$. Partition $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ where $\mathbf{U}_1$ is $p \times m$ so that under the null hypothesis $\mathcal{S}_{Y|\mathbf{Z}} \subseteq \mathcal{S}(\mathbf{U}_1)$. The following proposition provides a basis for constructing permutation tests for inference on d. The proof follows directly from proposition 4.6 of Cook (1998) and is thus omitted.

Proposition 3. Let $\mathbf{U}$ be constructed as indicated previously and assume that $\mathbf{U}_1^{\mathrm{T}}\mathbf{Z} \perp\!\!\!\perp \mathbf{U}_2^{\mathrm{T}}\mathbf{Z}$. Then $\mathcal{S}(\mathbf{U}_1)$ is a DRS for the regression of Y on $\mathbf{Z}$ if and only if $(Y, \mathbf{U}_1^{\mathrm{T}}\mathbf{Z}) \perp\!\!\!\perp \mathbf{U}_2^{\mathrm{T}}\mathbf{Z}$.

According to this proposition we can gain information on d by testing the null hypothesis $(Y, \mathbf{U}_1^{\mathrm{T}}\mathbf{Z}) \perp\!\!\!\perp \mathbf{U}_2^{\mathrm{T}}\mathbf{Z}$. We propose to test this statement by comparing the observed test statistic $\hat{\Lambda}_m$ with its permutation distribution under the null hypothesis. This involves recomputing $\hat{\Lambda}_m$ for each of a selected number of random permutations of the elements of the sample version of $(Y, \mathbf{U}_1^{\mathrm{T}}\mathbf{Z})$, and then comparing the observed value with its permutation distribution to obtain the p-values. These p-values can be used to infer about d by using the general procedure described at the end of Section 4.2.1.

Li's (1991) test for dimension with SIR is based on the assumption of normal predictors. The permutation test requires that $\mathbf{U}_1^{\mathrm{T}}\mathbf{Z}$ and $\mathbf{U}_2^{\mathrm{T}}\mathbf{Z}$ be marginally independent. This condition is satisfied when $\mathbf{Z}$ is normally distributed, although normality is not a necessary condition for the permutation test.

5. Connection between $\mathcal{S}_{\mathrm{cov}}^{(k)}$ and $\mathcal{S}_{E(\mathbf{Z}|Y)}$

Having developed the CKMS, we establish in the next proposition a general connection between the inverse mean subspace $\mathcal{S}_{E(\mathbf{Z}|Y)}$ and the covariance subspace $\mathcal{S}_{\mathrm{cov}}^{(k)}$ defined at expression (2). This provides connections between SIR and cov_k.

Proposition 4.

(a) If Y has finite support $R(Y) = \{a_0, a_1, \ldots, a_k\}$, then

$$\mathcal{S}_{E(\mathbf{Z}|Y)} = \mathrm{span}\{E(Y^i\mathbf{Z}), i = 1, \ldots, k\} = \mathcal{S}_{\mathrm{cov}}^{(k)}.$$

(b) If Y is continuous and $\boldsymbol{\mu}_Y = E(\mathbf{Z}|Y)$ is continuous on $R(Y)$, then

$$\mathcal{S}_{E(\mathbf{Z}|Y)} = \mathrm{span}\{E(Y^i\mathbf{Z}), i = 1, 2, \ldots\} = \lim_{k\to\infty} (\mathcal{S}_{\mathrm{cov}}^{(k)}).$$

According to proposition 4, we can select a k depending on the support of Y so that the inverse mean subspace $\mathcal{S}_{E(\mathbf{Z}|Y)}$ and the covariance subspace $\mathcal{S}_{\mathrm{cov}}^{(k)}$ are the same. Thus methods like SIR (Li, 1991) and parametric inverse regression (Bura and Cook, 2001) for estimating the inverse mean subspace can be regarded as methods for estimating a covariance subspace $\mathcal{S}_{\mathrm{cov}}^{(k)}$ for some value of k. However, for any value of k, $\mathcal{S}_{\mathrm{cov}}^{(k)} \subseteq \mathcal{S}_{E(\mathbf{Z}|Y)}$ and thus methods for estimating $\mathcal{S}_{\mathrm{cov}}^{(k)}$ provide lower bounds on $\mathcal{S}_{E(\mathbf{Z}|Y)}$.

For application in practice, SIR requires that a continuous response Y be replaced by a discrete version $\tilde{Y}$ with finite support constructed by slicing. Although Li (1991) indicated that the number of slices is not crucial, the results can be sensitive to the choice (Cook (1998), chapter 11). From proposition 4 we can think of the number of slices as controlling the maximum order k of the covariance $E(Y^k\mathbf{Z})$ used to summarize the distribution of $Y|\mathbf{Z}$. If interest is mostly in the lower order covariances there may be little reason to choose a large number of slices. Alternatively, we may wish to conduct analyses with different numbers of slices, focusing first on low order covariances (few slices) and eventually on high order covariances (many slices) to gain a more detailed understanding of $Y|\mathbf{Z}$. This perhaps partly answers one of the open questions posed by Kent (1991).

SIR and cov_k can also be connected in a related way. Consider a class of subspace estimation methods based on a population kernel matrix with columns $E\{g_j(Y)\mathbf{Z}\}, j = 1, \ldots, k$, where the g_j are known functions. The cov_k methods are based on choosing polynomials for the g_j, whereas SIR is based on choosing step functions (slicing) for the g_j.

Finally, proposition 4 can be used to understand the limitations of the inverse mean subspace, and why SIR fails to find directions in symmetric regressions like $Y = Z^2 + \varepsilon$ where Z is standard normal and independent of ε which has mean 0. In this case $E(Y^k Z) = 0$ for all $k = 1, 2, \ldots$, so $\mathcal{S}_{E(Z|Y)} = \mathcal{S}(0)$. Like SIR, cov_k will also fail in such symmetric regressions, and this accounts for many situations in which $\mathcal{S}^{(k)}_{\text{cov}}$ is a proper subset of $\mathcal{S}^{(k)}_{Y|\mathbf{Z}}$.

6. Soil evaporation

In this example we consider the regression of daily soil evaporation (Y) on four predictors: the integrated area and range of the daily air temperature and daily humidity curves. The data set consists of 46 observations and a few additional predictors that are not used here. It was presented by Freund (1979) and analysed in a related context by Cook (1994a). Our goal was to study the first two moments of $Y|\mathbf{Z}$ by using $\mathcal{S}^{(2)}_{\text{cov}}$. An inspection of various plots of the predictors did not reveal any notable non-linearities and thus we assumed the linearity condition.

We based our analysis on the sample kernel matrix

$$\hat{\mathbf{K}}_{\text{c}} = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{W}_i \hat{\mathbf{Z}}_i, \frac{1}{n} \sum_{i=1}^{n} \hat{W}_i^2 \hat{\mathbf{Z}}_i \right)$$

where $\hat{W}_i = (Y_i - \bar{Y})/\hat{\sigma}(Y)$. As an aside, the squared length of the first column of $\hat{\mathbf{K}}_{\text{c}}$ is n^{-1} times the score test statistic for the hypothesis $\boldsymbol{\beta} = 0$ in the homoscedastic single-index, normal linear model $Y = \beta_0 + g(\boldsymbol{\beta}^{\text{T}}\mathbf{X}) + \varepsilon$. The squared length of the second column of $\hat{\mathbf{K}}_{\text{c}}$ is $2n^{-1}$ times the score test statistic for the hypothesis $\boldsymbol{\alpha} = 0$ in the heteroscedastic linear model $Y = \beta_0 + \exp(\boldsymbol{\alpha}^{\text{T}}\mathbf{X})\varepsilon$, where ε is normal with mean 0 and variance σ^2. This kernel matrix thus gains its information on the second-moment subspace in a rather classical way. Further, although the usual score test statistic for $\boldsymbol{\alpha} = 0$ requires that the mean function be correct (Cook and Weisberg, 1983), it can also indicate an incorrect mean function, even if the variance function is constant. The second column of the kernel matrix $\hat{\mathbf{K}}_{\text{c}}$ can therefore contain information on both the mean function and the variance function. This was previously illustrated in the two examples of Section 4.1.

Returning to soil evaporation, the permutation p-values for cov_2 based on 1000 replications are 0 and 0.032; thus, we infer that $\dim(\mathcal{S}^{(2)}_{\text{cov}}) = 2$. The singular values of the sample kernel matrix are about 1.13 and 0.27, suggesting that most of the variation is in the first direction. Additionally, the permutation p-values for cov_3 are 0, 0.035 and 0.914, lending support to the possibility that the regression information is contained in the first two moments and that $\mathcal{S}^{(2)}_{\text{cov}} = \mathcal{S}^{(2)}_{Y|\mathbf{Z}} = \mathcal{S}_{Y|\mathbf{Z}}$.

The orthonormal cov_2 coefficient vectors $\hat{\boldsymbol{\eta}}_1$ and $\hat{\boldsymbol{\eta}}_2$ are shown in Table 1 for the original predictors and those predictors standardized by their sample standard deviations. A standardized coefficient is just the original coefficient times the sample standard deviation of the associated predictor. As in linear regression, it may be easier to assess the relative magnitudes of coefficients when the predictors all have the same sample standard deviation. The interpretation of the coefficients in Table 1 depends on the specific application context.

Table 1. Raw and marginally standardized estimated coefficient vectors $\hat{\eta}_j$ from the cov$_2$ analysis of the evaporation data

Predictor **X**	$\hat{\eta}_1$	*Standardized* $\hat{\eta}_1$	$\hat{\eta}_2$	*Standardized* $\hat{\eta}_2$
Air temperature area	0.096	0.114	0.242	0.531
Air temperature range	0.621	0.148	−0.797	−0.353
Humidity area	−0.562	−0.935	0.177	0.547
Humidity range	−0.538	−0.301	0.524	0.542

Nevertheless, we can see that the humidity area is the dominant predictor in the first standardized direction in Table 1, and all predictors contribute about equally to the second standardized direction. Since the coefficient vectors $\hat{\eta}_1$ and $\hat{\eta}_2$ only define a subspace, the interpretation might be enriched by considering linear combinations of them.

A three-dimensional summary plot of *Y versus* the two cov$_2$ predictors along with the OLS fit $\hat{Y}$ of a full quadratic model in those two predictors is represented in Fig. 1. Various residual plots and other standard diagnostics did not indicate deficiencies in the fit. The OLS fit $\hat{Y}_{\text{full}}$ of a full quadratic in the original four predictors did not offer any notable improvement. The sample correlation between $\hat{Y}$ and $\hat{Y}_{\text{full}}$ is 0.984. The quadratic surface shown in Fig. 1 appears to be twisted, suggesting the possibility of interactions. This possibility is supported by investigating the interaction terms in the OLS fit of a full quadratic model in the original four predictors.

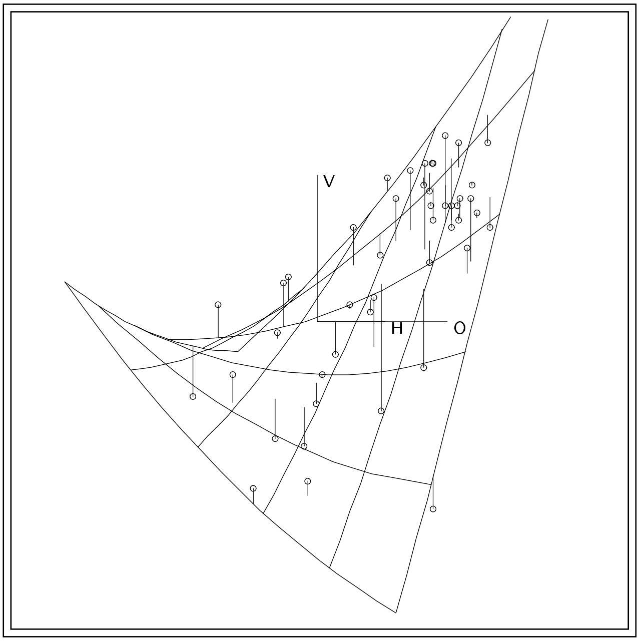

Fig. 1. Three-dimensional summary plot of the response *versus* the two cov$_2$ predictors: the surface is a full quadratic fitted to the plot by using OLS

For contrast we also applied SIR as implemented in `Arc` (Cook and Weisberg, 1999b). Not surprisingly, results from the application of SIR depend on the number of slices. Using the asymptotic test associated with SIR, there was a firm indication of one-dimensional structure. Depending on the number of slices, the SIR test hinted at a second dimension as well. For instance, using SIR with three slices (SIR$_3$), the SIR p-values are 0.000 and 0.967. But, using SIR with four slices (SIR$_4$), the p-values are 0.000, 0.075 and 1.00. Table 2 gives the R^2-values from four different OLS regressions distinguished only by their responses which are the first two SIR predictors with three and four slices. The predictors for each of the four regressions are the two cov$_2$ predictors $\hat{\eta}_1^{\mathrm{T}}\mathbf{X}$ and $\hat{\eta}_2^{\mathrm{T}}\mathbf{X}$. For example, $R^2 = 0.982$ for the regression of the first SIR predictor with three slices on the two cov$_2$ predictors. The results in Table 2 indicate that the first SIR predictors are essentially linear combinations of the two cov$_2$ predictors, but the second SIR predictors are not. Further investigation indicated that the first SIR predictors are very similar to the first cov$_2$ predictor, but the second SIR predictors are not similar to the second cov$_2$ predictor. Evidently, the methods differ in their second directions.

We conducted a small simulation study to see whether this example reflects general operating characteristics of the methods or perhaps just isolated aspects of the data. Letting $\hat{Y}$ represent the fitted values from the quadratic fit shown in Fig. 1 using the first two cov$_2$ predictors $\hat{\eta}_1^{\mathrm{T}}\mathbf{X}$ and $\hat{\eta}_2^{\mathrm{T}}\mathbf{X}$, we generated simulated data sets as

$$Y_{\mathrm{sim},i} = \hat{Y}_i + 4\varepsilon_i, \qquad i = 1, \ldots, 46,$$

where ε is a standard normal pseudorandom variable and the standard deviation is roughly the same as that observed from an OLS fit of the full quadratic model in all four predictors. The simulated data have two-dimensional structure with the two coefficient vectors $(\hat{\eta}_1, \hat{\eta}_2)$ in Table 1 spanning $\mathcal{S}_{\mathrm{cov}}^{(2)} = \mathcal{S}_{Y_{\mathrm{sim}}|\mathbf{X}}$. We then applied cov$_2$ and SIR$_3$ to the regression of Y_{sim} on the original four predictors. To assess the accuracy of the methods, we calculated the R^2-values from the four OLS regressions of the first and second cov$_2$ and SIR$_3$ predictors from the simulated data on the two true predictors $\hat{\eta}_1^{\mathrm{T}}\mathbf{X}$ and $\hat{\eta}_2^{\mathrm{T}}\mathbf{X}$ from the original analysis. The simulation was then repeated 100 times.

The first four rows of Table 3 following the column headings give the 5, 50 and 95 percentage points of the empirical distribution of R^2 for the first and second cov$_2$ and SIR$_3$ predictors. For example, the first row gives the results from the OLS regression of the first cov$_2$ predictor from the simulated data on $\hat{\eta}_1^{\mathrm{T}}\mathbf{X}$ and $\hat{\eta}_2^{\mathrm{T}}\mathbf{X}$. The results indicate that the first cov$_2$ and SIR$_3$ predictors reliably captured a linear combination of the true predictors. Additionally, the second cov$_2$ predictor gave a good representation of a second, orthogonal,

Table 2. R^2-values from the OLS regressions of the first two SIR predictors with three and four slices on the two cov$_2$ predictors

Method	*R^2-values for the following SIR predictors:*	
	1st	*2nd*
SIR$_3$	0.982	0.018
SIR$_4$	0.999	0.510

Table 3. SIR and cov_2 applied to the simulated evaporation data

Method	*Predictor*	*5% R^2*	*Median R^2*	*95% R^2*
cov_2	1st	0.989	0.995	0.999
cov_2	2nd	0.890	0.979	0.998
SIR_3	1st	0.973	0.984	0.995
SIR_3	2nd	0.016	0.308	0.936
cov_2	1st	0.987	0.996	0.999
cov_2	2nd	0.878	0.971	0.998
SIR_4	1st	0.972	0.987	0.998
SIR_4	2nd	0.102	0.213	0.741
SIR_3, $Y_{\text{sim}} \perp\!\!\!\perp \mathbf{X}$	2nd	0.085	0.516	0.931

direction. The results for the second SIR_3 predictor are not nearly as good because the median R^2-value is lower and the variability is larger.

In an independent simulation, the next four rows of Table 3 were constructed in the same way as the first four rows, except that we used SIR_4 instead of SIR_3. The results are qualitatively similar to the previous results, but the performance of the second SIR_4 predictor was somewhat worse than that for the second SIR_3 predictor.

Finally, to provide an external basis for comparison, we ran the previous cases using $Y_{\text{sim},i} = 4\varepsilon_i$ so that $Y_{\text{sim}} \perp\!\!\!\perp \mathbf{X}$. In all cases the R^2-values seemed consistent with observations on a uniform [0, 1] random variable. The final row of Table 2 gives one set of results for contrast. Comparing that row with the previous results, it seems that the second SIR direction did not do as well as selecting a direction at random.

We suggested in Section 1 that, when the first few moments of $Y|\mathbf{X}$ are of special interest, methods that target those moments should in principle produce results that are better than those from methods that pursue the distribution of $Y|\mathbf{X}$, particularly when $\mathcal{S}^{(k)}_{Y|\mathbf{Z}} = \mathcal{S}_{Y|\mathbf{Z}}$. This example supports that idea since the results for cov_2 were found to be substantially better than those for SIR.

7. Discussion

In this section we discuss a few issues related to aspects of this paper.

7.1. Assumptions

We used three primary assumptions in the developments of this paper. The first is that Y and $\mathbf{X}$ have a joint distribution, the second is the linearity condition of proposition 2—$E(\mathbf{Z}|\gamma^{\mathrm{T}}\mathbf{Z})$ is linear or approximately so—and the third is the independence condition that was used in the permutation test of Section 4.2.2. The first assumption was used mainly to facilitate the exposition. It is not essential and the results here apply to designed experiments in which $\mathbf{X}$ is fixed. The linearity condition, although not a severe restriction, is important because extreme violations of this condition can produce misleading results. When used in the context of designed experiments, the linearity condition should be interpreted relative to the design distribution. Many classical designs, particularly central composite designs, satisfy the linearity condition to a good approximation (Ibrahimy and Cook, 1995).

As mentioned in Section 2.2, it may be possible to induce approximate normality in the predictors by using power transformations or weighting, which would ensure both the linearity condition and the independence condition for the permutation test. To see what can happen when the linearity condition fails, let $\mathcal{L}$ be the subspace spanned by $\{E(\mathbf{Z}|\gamma^{\mathrm{T}}\mathbf{Z} = \mathbf{v})\}$ as $\mathbf{v}$ varies in the sample space of $\gamma^{\mathrm{T}}\mathbf{Z}$, where γ is still a basis for $\mathcal{S}^{(k)}_{Y|\mathbf{Z}}$. Then it follows immediately from the proof of proposition 2 that $E(Y^k\mathbf{Z}) \in \mathcal{L}$ without requiring the linearity condition. In addition,

$$E\{E(\mathbf{Z}|\gamma^{\mathrm{T}}\mathbf{Z})\mathbf{Z}^{\mathrm{T}}\gamma\} = E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\gamma) = \gamma$$

and thus $\mathcal{S}^{(k)}_{Y|\mathbf{Z}} \subseteq \mathcal{L}$ with equality when the linearity condition holds. This shows that $\mathcal{L}$ is always a kth-moment DRS, although it is not necessarily central unless the linearity condition holds. A practical consequence of this result is that, when the linearity condition fails, cov_k may 'overestimate' $\mathcal{S}^{(k)}_{Y|\mathbf{Z}}$ by including directions that are in the kth-moment DRS $\mathcal{L}$ but not in the CKMS $\mathcal{S}^{(k)}_{Y|\mathbf{Z}}$.

The permutation test requires the independence condition $\mathbf{U}_1^{\mathrm{T}}\mathbf{Z} \perp\!\!\!\perp \mathbf{U}_2^{\mathrm{T}}\mathbf{Z}$. If this condition fails then, because $\mathrm{cov}(\mathbf{U}_1^{\mathrm{T}}\mathbf{Z}, \mathbf{U}_2^{\mathrm{T}}\mathbf{Z}) = 0$ by construction, one of two cases must hold: either

(a) $E(\mathbf{U}_2^{\mathrm{T}}\mathbf{Z}|\mathbf{U}_1^{\mathrm{T}}\mathbf{Z}) = 0$ and the dependence is in some higher conditional moments or
(b) $E(\mathbf{U}_2^{\mathrm{T}}\mathbf{Z}|\mathbf{U}_1^{\mathrm{T}}\mathbf{Z})$ is a non-linear function of $\mathbf{U}_1^{\mathrm{T}}\mathbf{Z}$.

Violations under case (b) are potentially more serious in practice and are linked to violations of the linearity condition. If the linearity holds under the hypothesis $Y \perp\!\!\!\perp \mathbf{U}_2^{\mathrm{T}}\mathbf{Z}|\mathbf{U}_1^{\mathrm{T}}\mathbf{Z}$ then we must be in case (a) and modest violations of the independence condition should not have much effect on the permutation test. If the linearity condition fails then case (b) holds and both estimation and testing will be affected. The outcome here is more difficult to predict, although results from various simulations indicate that overestimation is likely.

7.2. Residual analysis

The methods of this paper are applicable as model diagnostics by replacing the response Y with a residual r. For example, suppose that we wish to check the homoscedastic linear model $Y = \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{U} + \varepsilon$ where $\mathbf{U}$ denotes the vector of predictors and ε is normal with mean 0 and variance σ^2. Letting r_i denote the ith ordinary residual standardized by the maximum likelihood estimate of σ, a kernel matrix for cov_2 is just

$$\hat{\mathbf{K}}_{\mathrm{c}} = (\hat{\mathbf{k}}_1, \hat{\mathbf{k}}_2) = \left(\frac{1}{n}\sum_{i=1}^{n} r_i\hat{\mathbf{Z}}_i, \frac{1}{n}\sum_{i=1}^{n} r_i^2\hat{\mathbf{Z}}_i\right). \tag{5}$$

If the predictors $\mathbf{U}$ used in the model are the same as the predictors used in the cov_2 analysis then the first column of the kernel matrix is 0.

7.3. Score test for heteroscedasticity

Cook and Weisberg (1983) developed a diagnostic score test for $\boldsymbol{\alpha} = 0$ in the linear model

$$Y = \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{U} + \exp(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Z})\varepsilon,$$

where ε is normal with mean 0 and variance σ^2. For ease of exposition and without loss of generality, we have written the variance function in terms of the standardized predictor vector $\mathbf{Z}$ which may contain predictors in $\mathbf{U}$. The score test statistic for $\boldsymbol{\alpha} = 0$ can now be expressed as $n/2$ times the squared length of the second column $\hat{\mathbf{k}}_2$ of the kernel matrix (5).

Cook and Weisberg (1983) also proposed an associated diagnostic plot of r_i^2 *versus* the fitted values $\mathbf{a}^{\mathrm{T}}\hat{\mathbf{Z}}_i$ from the OLS regression of r_i^2 on $\hat{\mathbf{Z}}_i$, referring to the estimated coefficient vector $\mathbf{a}$ as a 'quick estimate' of $\boldsymbol{\alpha}$ without any apparent justification. In the context of this paper, $\mathbf{a} = \hat{\mathbf{k}}_2$ and, under the condition that $E(\mathbf{Z}|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Z})$ is linear, it can be shown that $\mathbf{a}$ is a consistent estimate of $c\boldsymbol{\alpha}$ for some positive constant c, providing support for the diagnostic plot proposed by Cook and Weisberg (1983).

7.4. Regression mixtures

Cook and Critchley (2000) studied foundations of dimension reduction in the presence of regression mixtures and outliers, focusing primarily on the central subspace. They concluded that methodology for estimating the central subspace enables

> '. . . the analyst to construct low dimensional summary plots that may show the main regression structure as well as outliers and mixtures without the need to pre-specify a parametric model'.

One of their illustrations of this conclusion involved applying SIR to the regression of an athlete's lean body mass Y on height, weight and red blood cell count. They found that SIR indicates at least two-dimensional structure; the possibility of three-dimensional structure was difficult to assess because the results depend on the number of slices. In the end, they found that SIR identifies three distinct subpopulations.

To check further on the abilities of the cov_k methodology, we applied cov_3 to the athlete's regression study of Cook and Critchley (2000). The three permutation p-values based on 1000 replications are 0.000, 0.000 and 0.008, indicating that no dimension reduction is possible in this regression. From this point it was straightforward to reproduce the graphical displays and conclusions reported by Cook and Critchley using the three cov_3 predictors. For example, a plot of the response *versus* the third cov_3 predictor is very similar to Cook and Critchley's Fig. 2. Thus, cov_3 found the same structure as SIR, illustrating that cov_k methodology can perform well in complicated regressions.

7.5. Availability

The permutation test discussed in Section 4.2.2 is available as an extension to the dimension reduction methods in `Arc` (Cook and Weisberg, 1999b). It can be obtained at the Internet site `www.stat.umn.edu/arc` by following the link to 'Text extensions'. cov_k methods are also available as extensions to dimension reduction methods in `Arc`.

Acknowledgements

We thank the referees for helpful comments on earlier versions of this paper. The work of Yin was supported in part by the University of Georgia Research Foundation and the work of Cook was supported in part by National Science Foundation grants DMS-9703777 and DMS-0103983.

Appendix A: Justifications

A.1. Proposition 1

That part (a) of proposition 1 implies part (b), part (c) implies part (a) and part (d) implies part (b) are immediate. We now show that part (b) implies part (c).

$E(Y|\mathbf{X})$ is a function of $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ by proposition 1 of Cook and Li (2001) for $Y \perp\!\!\!\perp E(Y|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$.

Let us suppose that, for $i = 1, \ldots, k-1$, $M^{(i)}(Y|\mathbf{X})$ are functions of $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$. Thus by expanding $M^{(k)}(Y|\mathbf{X}) = E[\{Y - E(Y|\mathbf{X})\}^k|\mathbf{X}]$, we have

$$M^{(k)}(Y|\mathbf{X}) = E(Y^k|\mathbf{X}) + g(\mathbf{X}),$$

where $g(\mathbf{X})$ is a function of $M^{(i)}(Y|\mathbf{X})$ for $i = 1, \ldots, k-1$; hence $g(\mathbf{X})$ is a function of $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$, and we next show that $M^{(k)}(Y|\mathbf{X})$ is a function of $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$. This will imply that $E(Y^k|\mathbf{X})$ is a function of $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$.

The proof proceeds by first using part (b),

$$\operatorname{cov}\{Y^k, M^{(k)}(Y|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} = 0,$$

to show that

$$E\{M^{(k)}(Y|\mathbf{X})\, E(Y^k|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} = E\{E(Y^k|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\}\, E\{M^{(k)}(Y|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\}.$$

Next, by assumption, $g(\mathbf{X})$ is a function of $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$. We then have

$$E\{M^{(k)}(Y|\mathbf{X})\, g(\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} = E\{g(\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\}\, E\{M^{(k)}(Y|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\}.$$

Adding the last two equations gives

$$E\{M^{(k)}(Y|\mathbf{X})\, M^{(k)}(Y|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} = E\{M^{(k)}(Y|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\}\, E\{M^{(k)}(Y|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\}.$$

Therefore

$$E[\{M^{(k)}(Y|\mathbf{X})\}^2|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}] = [E\{M^{(k)}(Y|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\}]^2.$$

This implies that $\operatorname{var}\{M^{(k)}(Y|\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} = 0$. Thus, given $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$, $M^{(k)}(Y|\mathbf{X})$ is a constant, and it follows that part (b) implies part (c).

We now show that part (c) implies part (d), as follows:

$$\begin{aligned}
\operatorname{cov}\{Y^k, f(\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} &= E\{Y^k f(\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} - E(Y^k|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X})\, E\{f(\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} \\
&= E[E\{Y^k f(\mathbf{X})|\mathbf{X}\}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}] - E(Y^k|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X})\, E\{f(\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} \\
&= E(Y^k|\mathbf{X})\, E\{f(\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} - E(Y^k|\mathbf{X})\, E\{f(\mathbf{X})|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}\} \\
&= 0.
\end{aligned}$$

The third equality follows from condition (c).

A.2. Proposition 2

Using proposition 1, the linearity condition and $f^{(k)}(Y) = Y^k$,

$$\begin{aligned}
E(Y^k\mathbf{Z}) &= E\{E(Y^k\mathbf{Z}|\mathbf{Z})\} \\
&= E\{E(Y^k|\gamma^{\mathrm{T}}\mathbf{Z})\mathbf{Z}\} \\
&= E\{E(Y^k|\gamma^{\mathrm{T}}\mathbf{Z})\, E(\mathbf{Z}|\gamma^{\mathrm{T}}\mathbf{Z})\} \qquad (6) \\
&= P_\gamma\, E(Y^k\mathbf{Z}).
\end{aligned}$$

The proof for general $f^{(k)}(Y)$ is essentially the same.

A.3. Proposition 4

A.3.1. Part (a)

Define $\boldsymbol{\mu}_i = E(\mathbf{Z}|Y = a_i)$ and $f_i = \Pr(Y = a_i)$ for $i = 0, \ldots, k$. Therefore we have $\Sigma_{i=0}^k f_i = 1$ and $\Sigma_{i=0}^k f_i\boldsymbol{\mu}_i = 0$.

$$\mathcal{S}\{E(Y\mathbf{Z}), \ldots, E(Y^k\mathbf{Z})\} = \mathcal{S}\left\{ (\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_k) \begin{pmatrix} f_0 & 0 & \ldots & 0 \\ 0 & f_1 & \ldots & 0 \\ . & . & \ldots & . \\ 0 & 0 & \ldots & f_k \end{pmatrix} \begin{pmatrix} 1 & a_0 & a_0^2 & \ldots & a_0^k \\ 1 & a_1 & a_1^2 & \ldots & a_1^k \\ . & . & . & \ldots & . \\ 1 & a_k & a_k^2 & \ldots & a_k^k \end{pmatrix} \right\}$$

$$= \mathcal{S}(\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_k).$$

The last two matrices in the first equation are non-singular, the first because it is a diagonal matrix with positive f_i whereas the determinant of the second is $\Pi_{k \geqslant i > j \geqslant 0}(a_i - a_j)$ which is not 0. The result follows since $\mathcal{S}_{E(\mathbf{Z}|Y)} = \mathcal{S}(\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_k)$ by definition.

A.3.2. Part (b)

First note that $E[Y^i\{E(\mathbf{Z}|Y)\}] \in \mathcal{S}_{E(\mathbf{Z}|Y)}$ for $i = 1, 2, \ldots$. Hence

$$\mathcal{S}\{E(Y^i\mathbf{Z}), i = 1, 2, \ldots\} \subseteq \mathcal{S}_{E(\mathbf{Z}|Y)}.$$

Now we need to prove

$$\mathcal{S}_{E(\mathbf{Z}|Y)} \subseteq \mathcal{S}\{E(Y^i\mathbf{Z}), i = 1, 2 \ldots\}.$$

It follows from Cook (1998), proposition 11.1, that

$$\mathcal{S}_{E(\mathbf{Z}|Y)} = \mathcal{S}[\mathrm{var}\{E(\mathbf{Z}|Y)\}] = \mathcal{S}\{E(\boldsymbol{\mu}_Y \boldsymbol{\mu}_Y^{\mathrm{T}})\},$$

and, if we define the vector $\mathbf{Y}_q = (1, Y, Y^2, \ldots, Y^q)^{\mathrm{T}}$ for any non-negative integer q, then

$$\mathcal{S}\{E(\boldsymbol{\mu}_Y \mathbf{Y}_q^{\mathrm{T}})\} \subseteq \mathcal{S}\{E(Y^i\mathbf{Z}), i = 1, 2, \ldots\}.$$

Let us first assume that Y has a compact support, say, $R(Y) = [a, b]$. On the basis of the Weierstrass theorem (the theorem is stated in Rudin (1987)), $\forall \varepsilon > 0$, $\exists$ non-negative integer q and $p \times q$ matrix $\mathbf{C}_p$ such that

$$\boldsymbol{\mu}_Y = \mathbf{C}_p \mathbf{Y}_q + \boldsymbol{\varepsilon}_p \qquad \text{for all } Y \in [a, b],$$

where $\boldsymbol{\varepsilon}_p$ is a $p \times 1$ vector and

$$\boldsymbol{\varepsilon}_p^{\mathrm{T}} \boldsymbol{\varepsilon}_p \leqslant \varepsilon \qquad \text{for all } Y \in [a, b].$$

Therefore, for any vector $\mathbf{B} \in \mathcal{S}_{E(\mathbf{Z}|Y)}$, say $\mathbf{B} = E(\boldsymbol{\mu}_Y \boldsymbol{\mu}_Y^{\mathrm{T}})\mathbf{B}_0$, we can have

$$\mathbf{B} = E(\boldsymbol{\mu}_Y \mathbf{Y}_q^{\mathrm{T}}) \mathbf{C}_P^{\mathrm{T}} \mathbf{B}_0 + E(\boldsymbol{\mu}_Y \boldsymbol{\varepsilon}_p^{\mathrm{T}}) \mathbf{B}_0.$$

So the first term on the right-hand side is a vector belonging to $\mathcal{S}\{E(Y^i\mathbf{Z}), i = 1, 2, \ldots\}$ whereas the second term tends to 0 uniformly as $\varepsilon \to 0$. Thus any vector in $\mathcal{S}_{E(\mathbf{Z}|Y)}$ is a limit of the vectors in $\mathcal{S}\{E(Y^i\mathbf{Z}), i = 1, 2, \ldots\}$, which means that this vector is in $\mathcal{S}\{E(Y^i\mathbf{Z}), i = 1, 2, \ldots\}$, since $\mathcal{S}_{E(\mathbf{Z}|Y)}$ is a compact vector space. So

$$\mathcal{S}_{E(\mathbf{Z}|Y)} \subseteq \mathcal{S}\{E(Y^i\mathbf{Z}), i = 1, 2, \ldots\}.$$

In the case that Y does not have a compact support, since no element of $\mathrm{var}(\boldsymbol{\mu}_Y)$ is infinite, as usual there are a and b such that we can ignore the Ys on $R(Y) - [a, b]$ and thus apply the above result to $[a, b]$. We then have $\mathcal{S}_{E(\mathbf{Z}|Y)} \subseteq \mathcal{S}\{E(Y^i\mathbf{Z}), i = 1, 2, \ldots\}$. Hence the conclusion holds.

References

Bura, E. and Cook, R. D. (2001) Estimating the structural dimension of regressions via parametric inverse regression. *J. R. Statist. Soc.* B, **63**, 393–410.

Cook, R. D. (1994a) On the interpretation of regression plots. *J. Am. Statist. Ass.*, **89**, 177–190.

——(1994b) Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proc. Phys. Engng Sci. Sect. Am. Statist. Ass.*, 18–25.

———(1996) Graphics for regressions with a binary response. *J. Am. Statist. Ass.*, **91**, 983–992.

———(1998) *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

Cook, R. D. and Critchley, F. (2000) Detecting regression outliers and mixtures graphically. *J. Am. Statist. Ass.*, **95**, 781–794.

Cook, R. D. and Li, B. (2001) Dimension reduction for the conditional mean in regression. *Ann. Statistn*, to be published.

Cook, R. D. and Nachtsheim, C. J. (1994) Re-weighting to achieve elliptically contoured covariates in regression. *J. Am. Statist. Ass.*, **89**, 592–599.

Cook, R. D. and Weisberg, S. (1983) Diagnostics for heteroscedasticity in regression. *Biometrika*, **70**, 1–10.

———(1991) Discussion of Li (1991). *J. Am. Statist. Ass.*, **86**, 328–332.

———(1999a) Graphs in statistical analyses: is the medium the message? *Am. Statistn*, **53**, 29–37.

———(1999b) *Applied Regression Including Computing and Graphics*. New York: Wiley.

Cook, R. D. and Yin, X. (2001) Dimension reduction and visualization in discriminant analysis (with discussion). *Aust. New Z. J. Statist.*, **43**, 147–199.

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Eaton, M. L. (1986) A characterization of spherical distributions. *J. Multiv. Anal.*, **20**, 272–276.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Freund, R. J. (1979) *Proc. Statist. Comput. Sect. Am. Statist. Ass.*, 111–112.

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

Ibrahimy, A. and Cook, R. D. (1995) Regression design for one-dimensional subspaces. In *Model Oriented Data Analysis* (eds C. P. Kitsos and W. G. Müller), pp. 125–134. New York: Springer.

Kent, J. T. (1991) Discussion of Li (1991). *J. Am. Statist. Ass.*, **86**, 336–337.

Li, K. C. (1991) Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Ass.*, **86**, 316–342.

Rudin, W. (1987) *Real and Complex Analysis*. New York: McGraw-Hill.

Schott, J. R. (1994) Determining the dimensionality in sliced inverse regression. *J. Am. Statist. Ass.*, **89**, 141–148.

Velilla, S. (1998) Assessing the number of linear components in a general regression problem. *J. Am. Statist. Ass.*, **93**, 1088–1098.