The Annals of Statistics
2002, Vol. 30, No. 2, 455–474

DIMENSION REDUCTION FOR CONDITIONAL MEAN IN REGRESSION

BY R. DENNIS COOK[1] AND BING LI[2]

University of Minnesota and Pennsylvania State University

In many situations regression analysis is mostly concerned with inferring about the conditional mean of the response given the predictors, and less concerned with the other aspects of the conditional distribution. In this paper we develop dimension reduction methods that incorporate this consideration. We introduce the notion of the Central Mean Subspace (CMS), a natural inferential object for dimension reduction when the mean function is of interest. We study properties of the CMS, and develop methods to estimate it. These methods include a new class of estimators which requires fewer conditions than pHd, and which displays a clear advantage when one of the conditions for pHd is violated. CMS also reveals a transparent distinction among the existing methods for dimension reduction: OLS, pHd, SIR and SAVE. We apply the new methods to a data set involving recumbent cows.

1. Introduction. Empirical evidence accumulated during the past few years indicates that recently developed dimension-reduction methods can be quite effective for constructing regression summary plots in a largely non-parametric context. In full generality, the goal of a regression is to infer about the conditional distribution of the univariate response variable Y given the $p \times 1$ vector of predictors $\mathbf{X}$: How does the conditional distribution of $Y|\mathbf{X}$ change with the value assumed by $\mathbf{X}$? The dimension-reduction methods approach this question through a population meta-parameter called the *central subspace* which is denoted by $\mathcal{S}_{Y|X}$. The central subspace is the smallest subspace of $\mathbb{R}^p$ such that Y is independent of $\mathbf{X}$ given $\boldsymbol{\eta}^T\mathbf{X}$, where the columns of the matrix $\boldsymbol{\eta}$ form any basis for the subspace. Knowledge of the central subspace is useful for parsimoniously characterizing how the distribution of $Y|\mathbf{X}$ changes with the value of $\mathbf{X}$. If $\mathcal{S}_{Y|X}$ is known, the *minimal sufficient summary plot* of Y versus $\boldsymbol{\eta}^T\mathbf{X}$ can be used to guide subsequent analysis. If an estimated basis $\widehat{\boldsymbol{\eta}}$ of $\mathcal{S}_{Y|X}$ is available then the summary plot of Y versus $\widehat{\boldsymbol{\eta}}^T\mathbf{X}$ can be used similarly.

Summary plots based on estimates of $\mathcal{S}_{Y|X}$ can be of significant value in many phases of a regression analysis, particularly during the initial phases when an adequate parsimoniously parameterized model is not yet available. Methods of estimating the central subspace or portions thereof include ordinary least squares

Received February 2000; revised August 2001.

[1]Supported in part by NSF Grants DMS-97-03777 and DMS-01-03983.

[2]Supported in part by NSF Grant DMS-96-26249.

AMS 2000 subject classifications. Primary 62G08; secondary 62-09, 62H05.

Key words and phrases. Central subspace, graphics, regression, pHd, SAVE, SIR, visualization.

(OLS), graphical regression [Cook (1994a)], principal Hessian directions [pHd; Li (1992)], sliced average variance estimation [SAVE; Cook and Weisberg (1991)], sliced inverse regression [SIR; Li (1991)] and parametric inverse regression [PIR; Bura and Cook (2001)]. Cook and Weisberg (1999) gave an introductory account of studying regressions via central subspaces. A comprehensive discussion is available in Cook (1998a).

While the central subspace is designed to give a complete picture of the dependence of Y on $\mathbf{X}$, certain characteristics of $Y|\mathbf{X}$ may often be of special interest. Indeed, regression is understood by some to imply a study of the mean function $\mathrm{E}(Y|\mathbf{X})$. Pursuing the mean function through the central subspace can be inefficient because the scope of the statistical inquiry may be much larger than necessary.

In Section 2 we introduce the *central mean subspace* (CMS) and study its properties. The construction of a CMS is similar in spirit to that for a central subspace, but dimension reduction is aimed at reducing the mean function alone, leaving the rest of $Y|\mathbf{X}$ as the "nuisance parameter."

In Section 3 we connect the central mean subspace to literature on the central subspace. In particular, we show that, in the population, many known techniques for constructing vectors in the central subspace in fact produce vectors in the central mean subspace. These results provide an important distinction between known methods – OLS, pHd, SIR and SAVE – for estimating vectors in the central subspace $\mathcal{S}_{Y|X}$. And they imply that in the first instance we can use two known methods – OLS and pHd – to estimate vectors in the central mean subspace.

In Section 4 we describe a new class of methods for estimating vectors in the CMS. This class may be useful because it requires fewer constraints than pHd on the distribution of the predictors. The new method is applied in Section 5. To avoid interrupting the development, we have placed most of the proofs in the Appendix.

2. Mean dimension-reduction subspaces and their properties. Consider a regression consisting of a univariate response variable Y and a $p \times 1$ vector of random predictors $\mathbf{X}$. We assume throughout this article that the data $\{y_i, \mathbf{x}_i\}$, $i = 1, \ldots, n$, are iid observations on $(Y, \mathbf{X})$ with finite moments. We use $\mathcal{S}(\mathbf{A})$ to denote a general subspace of $\mathbb{R}^p$, where $\mathbf{A}$ is a $p \times q$ matrix whose columns form a basis in $\mathcal{S}(\mathbf{A})$. We use $\mathcal{S}_{(\cdot)}$ to denote a dimension-reduction subspace, where the subscript indicates what response and predictors are involved, and whether the whole conditional distribution or only the conditional mean is of interest. In some instances we use $\mathcal{S}(\mathbf{b}, \mathbf{A})$, where $\mathbf{b}$ is a $p \times 1$ vector, to denote the subspace of $\mathbb{R}^p$ spanned by $\mathbf{b}$ and the columns of $\mathbf{A}$.

2.1. *Overview of central subspaces.* A *dimension-reduction subspace* [Li (1991)] for the regression of Y on $\mathbf{X}$ is any subspace $\mathcal{S}(\boldsymbol{\eta})$ such that

$$Y \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\eta}^T \mathbf{X}, \tag{2.1}$$

where $\perp\!\!\!\perp$ denotes independence and $\boldsymbol{\eta}$ denotes a $p \times q$ matrix with $q \leq p$. The statement is thus that Y is independent of $\mathbf{X}$ given $\boldsymbol{\eta}^T\mathbf{X}$. It is equivalent to saying that the distribution of $Y|\mathbf{X}$ is the same as that of $Y|\boldsymbol{\eta}^T\mathbf{X}$ for all 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 $\boldsymbol{\eta}^T\mathbf{X}$ without loss of regression information, and thus represents a potentially useful reduction in the dimension of the predictor vector. When the intersection of all dimension-reduction subspaces is itself a dimension-reduction subspace it is called the *central subspace* [Cook (1994b, 1998a)] and denoted by $\mathcal{S}_{Y|X}$. The central subspace is assumed to exist throughout this article.

As mentioned in the Introduction, the central subspace is designed to capture the complete conditional distribution of $Y|\mathbf{X}$ and thereby give a full picture of the dependence of Y on $\mathbf{X}$. On the other hand, when the conditional mean $\mathrm{E}(Y|\mathbf{X})$ is of special interest, it may be useful to adapt our inquiry to fit that more specific objective.

2.2. *Central mean subspace.* When focusing on the conditional mean, dimension reduction hinges on finding a $p \times k$ matrix $\boldsymbol{\alpha}$, $k \leq p$, so that the $k \times 1$ random vector $\boldsymbol{\alpha}^T\mathbf{X}$ contains all the information about Y that is available from $\mathrm{E}(Y|\mathbf{X})$. This is less restrictive than requiring that $\boldsymbol{\alpha}^T\mathbf{X}$ contain all the information about Y that is available from $\mathbf{X}$ as in the current literature associated with the central subspace. The following definition formalizes this idea.

DEFINITION 1. If $Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\boldsymbol{\alpha}^T\mathbf{X}$ then $\mathcal{S}(\boldsymbol{\alpha})$ is a mean dimension-reduction subspace for the regression of Y on $\mathbf{X}$.

It follows from this definition that a dimension-reduction subspace is necessarily a mean dimension-reduction subspace, because $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}$ implies $Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\boldsymbol{\alpha}^T\mathbf{X}$. The next proposition gives equivalent conditions for the conditional independence used in Definition 1.

PROPOSITION 1. *The following statements are equivalent*:
(i) $Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\boldsymbol{\alpha}^T\mathbf{X}$,
(ii) $\mathrm{Cov}[(Y, \mathrm{E}(Y|\mathbf{X}))|\boldsymbol{\alpha}^T\mathbf{X}] = 0$,
(iii) $\mathrm{E}(Y|\mathbf{X})$ *is a function of* $\boldsymbol{\alpha}^T\mathbf{X}$.

The first condition is the same as Definition 1. The second condition is that, given $\boldsymbol{\alpha}^T\mathbf{X}$, Y and $\mathrm{E}(Y|\mathbf{X})$ must be uncorrelated. The final condition is what might be suggested by intuition, $\mathrm{E}(Y|\mathbf{X}) = \mathrm{E}(Y|\boldsymbol{\alpha}^T\mathbf{X})$. Any of these three conditions could be taken as the definition of a mean dimension-reduction subspace.

Paralleling the development of central subspaces, we would like the smallest mean dimension-reduction subspace, as formalized in the next definition.

DEFINITION 2. Let $\mathcal{S}_{\mathrm{E}(Y|X)} = \bigcap \mathcal{S}_m$ where intersection is over all mean dimension-reduction subspaces $\mathcal{S}_m$. If $\mathcal{S}_{\mathrm{E}(Y|X)}$ is itself a mean dimension-reduction subspace, it is called the central mean dimension-reduction subspace, or simply the central mean subspace (CMS).

The CMS does not always exist, because the intersection of two mean dimension-reduction subspaces is not necessarily a mean dimension-reduction subspace. However, when it does exist, $\mathcal{S}_{\mathrm{E}(Y|X)} \subseteq \mathcal{S}_{Y|X}$ because the former is the intersection of a larger collection of subspaces. Hence it may be possible to reduce the dimension from that of $\mathcal{S}_{Y|X}$ if $\mathrm{E}(Y|\mathbf{X})$ alone is concerned. Under mild conditions, the existence and uniqueness of the CMS can be guaranteed in the same way that the existence of the central subspace is guaranteed. For instance, if the domain of $\mathbf{X}$ is open and convex, then the CMS exists and is unique. For location regressions where $Y \perp\!\!\!\perp \mathbf{X}|\mathrm{E}(Y|\mathbf{X})$, the central subspace and the CMS are the same, $\mathcal{S}_{\mathrm{E}(Y|X)} = \mathcal{S}_{Y|X}$. Additional existence results are available in such cases [Cook (1994a; 1998a), page 111]. We assume in the remainder of this article that the CMS always exists.

The CMS is intended to play the same role when considering the conditional mean as the central subspace does when inquiring about the full conditional distribution of $Y|\mathbf{X}$. If an estimated basis $\widehat{\boldsymbol{\alpha}}$ of $\mathcal{S}_{\mathrm{E}(Y|X)}$ is available then the summary plot of Y versus $\widehat{\boldsymbol{\alpha}}^T\mathbf{X}$ can provide a low-dimensional visualization of the mean function. Methods of estimating $\mathcal{S}_{\mathrm{E}(Y|X)}$ are discussed later in this article.

The central subspace is invariant under one-to-one transformations T of the response, $\mathcal{S}_{T(Y)|X} = \mathcal{S}_{Y|X}$. This property does not extend to the CMS because $\mathcal{S}_{\mathrm{E}(Y|X)}$ does not in general equal $\mathcal{S}_{\mathrm{E}(T(Y)|X)}$, although the central subspace is an invariant upper bound on the CMS, $\mathcal{S}_{\mathrm{E}(T(Y)|X)} \subseteq \mathcal{S}_{Y|X}$. Standard response transformation methodology in linear regression exploits this flexibility by attempting to find a T so that $\dim(\mathcal{S}_{\mathrm{E}(T(Y)|X)}) = 1$ and $\mathrm{E}(T(Y)|\mathbf{X})$ is linear.

It is sometimes helpful to transform $\mathbf{X}$ linearly so that the transformed predictors are uncorrelated, and then study the relation between Y and the transformed predictors. In general, if $\mathbf{Z} = \mathbf{A}^T\mathbf{X} + \mathbf{b}$ for some invertible matrix $\mathbf{A}$ and some vector $\mathbf{b}$, then $\mathcal{S}_{\mathrm{E}(Y|Z)} = \mathbf{A}^{-1}\mathcal{S}_{\mathrm{E}(Y|X)}$ is the CMS for the regression of Y on $\mathbf{Z}$.

In most of the subsequent developments, we work in terms of the standardized predictor

$$\mathbf{Z} = \boldsymbol{\Sigma}_{xx}^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X})),$$

where $\boldsymbol{\Sigma}_{xx} = \mathrm{Var}(\mathbf{X})$ is assumed to be positive definite. In terms of this standardized predictor, the CMS is $\mathcal{S}_{\mathrm{E}(Y|X)} = \boldsymbol{\Sigma}_{xx}^{-1/2}\mathcal{S}_{\mathrm{E}(Y|Z)}$. The corresponding sample version $\widehat{\mathbf{Z}}$ is obtained by replacing $\boldsymbol{\Sigma}_{xx}$ and $\mathrm{E}(\mathbf{X})$ with their usual moment estimates, $\widehat{\boldsymbol{\Sigma}}_{xx}$ and $\mathrm{E}_n(\mathbf{X})$.

3. Vectors in the central mean subspace. Having established some basic properties of the CMS, we now turn our attention to finding population vectors in that subspace. We will survey available methods for constructing vectors in the central subspace and demonstrate that some of them in fact produce vectors in CMS. By categorizing and assessing these methods in relation to CMS, we set the stage for a new estimation method introduced in Section 4.

3.1. *Objective functions and OLS.* In their pioneering article, Li and Duan (1989) demonstrated that a class of estimators, which includes OLS, correctly estimate the direction of the regression parameter regardless of the shape of the regression function, provided that the predictor satisfies a linearity condition. They considered an objective function of the form $R(a, \mathbf{b}) = \mathrm{E}[L(a + \mathbf{b}^T\mathbf{Z}, Y)]$ where $a \in \mathbb{R}^1$ and $\mathbf{b} \in \mathbb{R}^p$. Here, the expectation is with respect to the joint distribution of Y and $\mathbf{Z}$, and $L(K, Y)$ is strictly convex in K. This use of an objective function is not meant to imply that any associated model is true or even provides an adequate fit of the data. Nevertheless, there is a useful connection between $\mathcal{S}_{\mathrm{E}(Y|Z)}$ and the vectors derived from these objective functions.

Let

$$(\alpha, \boldsymbol{\beta}) = \arg\min_{a,\mathbf{b}} R(a, \mathbf{b}) \tag{3.1}$$

denote the population minimizers, and let $\boldsymbol{\eta}$ be a basis matrix for $\mathcal{S}_{Y|Z}$. Li and Duan (1989) showed, in effect, that if $\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})$ is linear in $\mathbf{Z}$ and $\dim(\mathcal{S}_{Y|Z}) = 1$, then $\boldsymbol{\beta} \in \mathcal{S}_{Y|Z}$. From this it is straightforward to relax the dimension restriction: if $\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})$ is linear in $\mathbf{Z}$ then $\boldsymbol{\beta} \in \mathcal{S}_{Y|Z}$ [Cook (1998a), pages 143–147].

Where $\boldsymbol{\beta}$ falls in relation to $\mathcal{S}_{\mathrm{E}(Y|Z)}$ depends on the choice of L: for some choices $\boldsymbol{\beta}$ belongs to $\mathcal{S}_{\mathrm{E}(Y|Z)}$ while for others it may belong to $\mathcal{S}_{Y|Z} \setminus \mathcal{S}_{\mathrm{E}(Y|Z)}$. However, if we restrict attention to objective functions

$$L(a + \mathbf{b}^T\mathbf{Z}, Y) = -Y(a + \mathbf{b}^T\mathbf{Z}) + \phi(a + \mathbf{b}^T\mathbf{Z}) \tag{3.2}$$

based on the natural exponential family for some strictly convex function ϕ, then $\boldsymbol{\beta}$ always belongs to $\mathcal{S}_{\mathrm{E}(Y|Z)}$:

THEOREM 1. *Let $\boldsymbol{\gamma}$ be a basis matrix for $\mathcal{S}_{\mathrm{E}(Y|Z)}$, assume that $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ is a linear function of $\mathbf{Z}$ and let $\boldsymbol{\beta}$ be as defined in* (3.1) *using an exponential family objective function* (3.2). *Then $\boldsymbol{\beta} \in \mathcal{S}_{\mathrm{E}(Y|Z)}$.*

The exponential family objective function (3.2) covers many estimation methods used in practice. In particular, OLS is obtained by setting $\phi(K) = K^2/2$. For future reference we denote the OLS coefficient vector $\mathrm{E}(Y\mathbf{Z})$ by $\boldsymbol{\beta}_{yz}$.

3.2. *SIR and SAVE.* It is easy to see that SIR [Li (1991)] and SAVE [Cook and Weisberg (1991)] can find vectors in $\mathcal{S}_{Y|Z} \setminus \mathcal{S}_{\mathrm{E}(Y|Z)}$: Suppose that $\mathcal{S}_{Y|Z}$ is spanned by the columns of the matrix $\boldsymbol{\eta}$ and $\mathbf{P}_\eta$ is the projection onto $\mathcal{S}_{Y|Z}$. Then, $\mathrm{E}(\mathbf{Z}|Y) = \mathrm{E}[\mathrm{E}(\mathbf{Z}|Y, \boldsymbol{\eta}^T\mathbf{Z})|Y] = \mathrm{E}[\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})|Y] = \mathbf{P}_\eta \mathrm{E}(\mathbf{Z}|Y)$, so that $\mathrm{E}(\mathbf{Z}|Y)$ belongs to $\mathcal{S}_{Y|Z}$. However, if we replace $\boldsymbol{\eta}$ by $\boldsymbol{\gamma}$, the basis for $\mathcal{S}_{\mathrm{E}(Y|Z)}$, then the second equality need not hold because we do not necessarily have the conditional independence $Y \perp\!\!\!\perp \mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}$. Hence, in general, $\mathrm{E}(\mathbf{Z}|Y)$ is in $\mathcal{S}_{Y|Z}$ but not necessarily in $\mathcal{S}_{\mathrm{E}(Y|Z)}$. The same is true of the vectors derived from SAVE.

3.3. *y-based pHd.* Li's (1992) y-based method of principal Hessian directions (pHd) is based on using the third moment matrix $\boldsymbol{\Sigma}_{yzz} = \mathrm{E}\{(Y - \mathrm{E}(Y))\mathbf{Z}\mathbf{Z}^T\}$ to infer about $\mathcal{S}_{Y|Z}$, assuming normally distributed predictors. However, as described in the next theorem, pHd in fact targets the CMS.

THEOREM 2. *Let $\boldsymbol{\gamma}$ be a basis matrix for $\mathcal{S}_{\mathrm{E}(Y|Z)}$. If $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ is a linear function of $\mathbf{Z}$ and if $\mathrm{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ is uncorrelated with Y, then $\mathcal{S}(\boldsymbol{\beta}_{yz}, \boldsymbol{\Sigma}_{yzz}) \subseteq \mathcal{S}_{\mathrm{E}(Y|Z)}$.*

The requirement that $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ be linear implies $\boldsymbol{\beta} \in \mathcal{S}_{\mathrm{E}(Y|Z)}$ as an application of Theorem 1. Both conditions stated in this theorem are used to obtain the conclusion that $\mathcal{S}(\boldsymbol{\Sigma}_{yzz}) \subseteq \mathcal{S}_{\mathrm{E}(Y|Z)}$.

Cook [(1998b), Theorem 1] showed that the inference procedure proposed by Li (1992) is not as straightforward as originally claimed. However, the inference procedure simplifies greatly if $\mathrm{Cov}(Y, \mathbf{Z}) = 0$. This condition will not normally hold, but is guaranteed to hold if we replace Y with population OLS residuals $r = Y - \mathrm{E}(Y) - \boldsymbol{\beta}_{yz}^T\mathbf{Z}$. In short, we can use the relatively straightforward inference procedure suggested by Cook (1998b) for inferring about the related subspace $\mathcal{S}(\boldsymbol{\Sigma}_{rzz})$, where $\boldsymbol{\Sigma}_{rzz}$ is constructed as $\boldsymbol{\Sigma}_{yzz}$, except that Y is replaced with the residual r:

$$\boldsymbol{\Sigma}_{rzz} = \mathrm{E}(r\mathbf{Z}\mathbf{Z}^T). \tag{3.3}$$

For this to be useful, we must understand the relationship between $\mathcal{S}_{\mathrm{E}(Y|Z)}$, the subspace we would like to know, and $\mathcal{S}_{\mathrm{E}(r|Z)}$, the CMS that we can infer about easily. This is the topic of the next proposition.

Recall that $\boldsymbol{\gamma}$ is the basis matrix for $\mathcal{S}_{\mathrm{E}(Y|Z)}$, and that $\boldsymbol{\beta}_{yz} = \mathrm{E}(Y\mathbf{Z})$. Define a residual

$$r = r(Y, \boldsymbol{\beta}_{yz}^T\mathbf{Z}) = a(\boldsymbol{\beta}_{yz}^T\mathbf{Z}) + b(\boldsymbol{\beta}_{yz}^T\mathbf{Z})Y \tag{3.4}$$

to be a function of Y and $\boldsymbol{\beta}_{yz}^T\mathbf{Z}$ that is linear in Y. For example, $r = Y - \mathrm{E}(Y) - \boldsymbol{\beta}_{yz}^T\mathbf{Z}$ satisfies these conditions as does the Pearson residual computed from a logistic regression model.

PROPOSITION 2. *Assume that* $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ *is linear. Then*

$$\mathcal{S}_{\mathrm{E}(Y|Z)} = \mathcal{S}_{\mathrm{E}(r|Z)} + \mathcal{S}(\boldsymbol{\beta}_{yz}), \tag{3.5}$$

where the summation on the right means the collection of vectors of the form $\boldsymbol{\beta}+\boldsymbol{\beta}'$ *with* $\boldsymbol{\beta}$ *in* $\mathcal{S}_{\mathrm{E}(Y|Z)}$ *and* $\boldsymbol{\beta}'$ *in* $\mathcal{S}(\boldsymbol{\beta}_{yz})$.

Proposition 2 is useful because it allows us to infer about $\mathcal{S}_{\mathrm{E}(Y|Z)}$ by inferring about $\mathcal{S}(\boldsymbol{\beta}_{yz})$ and $\mathcal{S}_{\mathrm{E}(r|Z)}$ using relatively straightforward procedures in the literature. In particular, it follows from Theorem 2 applied to the regression of r on $\mathbf{Z}$ that $\mathcal{S}(\boldsymbol{\Sigma}_{rzz}) \subseteq \mathcal{S}_{\mathrm{E}(r|Z)}$, and we can infer about $\mathcal{S}(\boldsymbol{\Sigma}_{rzz})$ by using the inference procedures discussed by Cook (1998b).

Finally, assuming that $\mathrm{E}(\mathbf{Z}|\boldsymbol{\beta}_{yz}^T\mathbf{Z})$ is linear and $\mathrm{Var}(\mathbf{Z}|\boldsymbol{\beta}_{yz}^T\mathbf{Z})$ is constant, it can be verified that

$$\boldsymbol{\Sigma}_{rzz} = \boldsymbol{\Sigma}_{yzz} - \mathbf{P}_{\beta_{yz}}\mathrm{E}(\boldsymbol{\beta}_{yz}^T\mathbf{Z})^3/\|\boldsymbol{\beta}_{yz}\|^2.$$

Thus, $\mathcal{S}(\boldsymbol{\beta}_{yz}, \boldsymbol{\Sigma}_{rzz}) = \mathcal{S}(\boldsymbol{\beta}_{yz}, \boldsymbol{\Sigma}_{yzz})$, which returns us to the regression of Y on $\mathbf{Z}$.

4. Vectors in the CMS that require only the linear conditional means.

4.1. *Population structure.* The vectors in the CMS described in the last section require two essential conditions:

C.1: the conditional mean $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ is linear in $\mathbf{Z}$,
C.2: the conditional variance $\mathrm{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ is uncorrelated with Y.

Predictor linearity conditions such as C.1 used in Theorem 1 are common in dimension reduction. In general, the requirement that $\mathrm{E}(\mathbf{Z}|\mathbf{A}^T\mathbf{Z})$ be linear in $\mathbf{Z}$ is equivalent to requiring that $\mathrm{E}(\mathbf{Z}|\mathbf{P}_A\mathbf{Z}) = \mathbf{P}_A\mathbf{Z}$ where $\mathbf{P}_A$ is the projection operator for $\mathcal{S}(\mathbf{A})$ with respect to the standard inner product [Cook (1998a), page 57]. If $\mathbf{Z}$ follows an elliptically contoured distribution then $\mathrm{E}(\mathbf{Z}|\mathbf{A}^T\mathbf{Z})$ is linear for all conforming $\mathbf{A}$'s. Hall and Li (1993) show that such linearity will hold to a reasonable approximation in many problems. The intuition here is that conditional expectations of the form $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ become more linear as p increases with $\dim(\mathcal{S}_{\mathrm{E}(Y|Z)})$ fixed. This is related to the work of Diaconis and Freedman (1984) who argue that almost all low-dimensional projections of high-dimensional data sets are nearly normal. In addition, these conditions might be induced by using predictor transformations and predictor weighting [Cook and Nachtsheim (1994)].

With condition C.1 we can use known methods based on exponential family objective functions to estimate a vector in $\mathcal{S}_{\mathrm{E}(Y|Z)}$ (Theorem 1). Such methods may be sufficient when $\dim(\mathcal{S}_{\mathrm{E}(Y|Z)}) = 1$, but other methods are needed when $\dim(\mathcal{S}_{\mathrm{E}(Y|Z)}) > 1$. To obtain multiple vectors in $\mathcal{S}_{\mathrm{E}(Y|Z)}$ using y-based pHd we had to require condition C.2 (Theorem 2) which is implied when $\mathrm{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ is

constant. Both conditions are implied when $\mathbf{Z}$ is normally distributed, although normality is not a necessary condition. In this section we introduce a new class of vectors in the CMS that requires only C.1. The next theorem lies at the core of this method.

THEOREM 3. *Suppose that U and V are measurable functions of $\boldsymbol{\gamma}^T\mathbf{Z}$ and that $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ is linear in $\mathbf{Z}$. Then $\mathrm{E}\{(UY+V)\mathbf{Z}\} \in \mathcal{S}_{\mathrm{E}(Y|Z)}$, provided that $(UY+V)\mathbf{Z}$ is integrable.*

This theorem provides a method of forming vectors in the CMS, which in turn provides a basis for constructing estimates of at least a portion of the CMS. We refer to these as COZY vectors since they come from COvariances between $\mathbf{Z}$ and and constructed responses $Y^* = UY + V$.

To see how Theorem 3 might be used to construct vectors in the CMS, suppose we know one vector $\boldsymbol{\delta}_0 \in \mathcal{S}_{\mathrm{E}(Y|Z)}$. We can find another vector $\boldsymbol{\delta}_1 \in \mathcal{S}_{\mathrm{E}(Y|Z)}$ by choosing appropriate functions $u : R \mapsto R$ and $v : R \mapsto R$, and then forming the covariance $\boldsymbol{\delta}_1 = \mathrm{E}(Y_1^*\mathbf{Z})$ between $Y_1^* = u(\boldsymbol{\delta}_0^T\mathbf{Z})Y + v(\boldsymbol{\delta}_0^T\mathbf{Z})$ (which is an instance of the constructed response Y^*) and the standardized predictor vector $\mathbf{Z}$. This process can then be iterated in the hope of finding additional vectors in the CMS:

$$\begin{aligned}\boldsymbol{\delta}_j &= \mathrm{E}\{[u(\boldsymbol{\delta}_{j-1}^T\mathbf{Z})Y + v(\boldsymbol{\delta}_{j-1}^T\mathbf{Z})]\mathbf{Z}\} \\ &= \mathrm{E}\{Y_{j-1}^*\mathbf{Z}\}, \qquad j = 1, 2, \ldots,\end{aligned} \tag{4.1}$$

where $Y_{j-1}^* = u(\boldsymbol{\delta}_{j-1}^T\mathbf{Z})Y + v(\boldsymbol{\delta}_{j-1}^T\mathbf{Z})$. Of course, we would need at most $k = \dim(\mathcal{S}_{\mathrm{E}(Y|Z)})$ vectors.

Two general questions remain regarding the construction of COZY vectors: How can we find the first vector $\boldsymbol{\delta}_0$ that is needed to prime the process? And how should we choose the functions u and v used in forming the constructed responses Y^*? We address these issues in the next section when developing a new method of estimating the CMS based on the COZY class.

4.2. *Iterative Hessian transformation.* A first application of Theorem 3 is straightforward: Setting $U = 1$ and $V = 0$ with probability 1, $\mathrm{E}\{(UY+V)\mathbf{Z}\} = \boldsymbol{\beta}_{yz}$ which we know belongs to $\mathcal{S}_{\mathrm{E}(Y|Z)}$ under condition C.1 (Theorem 1). Since any function of $\boldsymbol{\beta}_{yz}^T\mathbf{Z}$ is measurable with respect to $\boldsymbol{\gamma}^T\mathbf{Z}$, we can use this first result to construct additional vectors in the CMS:

COROLLARY 1. *Let $u : R \mapsto R$ and $v : R \mapsto R$ be any functions such that $\{u(\boldsymbol{\beta}_{yz}^T\mathbf{Z})Y + v(\boldsymbol{\beta}_{yz}^T\mathbf{Z})\}\mathbf{Z}$ is integrable. Assume condition* C.1. *Then any vector of the form $\mathrm{E}[\{u(\boldsymbol{\beta}_{yz}^T\mathbf{Z})Y + v(\boldsymbol{\beta}_{yz}^T\mathbf{Z})\}\mathbf{Z}]$ belongs to $\mathcal{S}_{\mathrm{E}(Y|Z)}$.*

To apply Corollary 1 we must choose the functions u and v. The next corollary gives two specific COZY sets obtained by choosing (i) $u(t) = t$ and $v(t) = -tE(Y)$, and (ii) $u(t) = t$ and $v(t) = -tE(Y) - t^2$. Starting with $\boldsymbol{\delta}_0 = \boldsymbol{\beta}_{yz}$, we then use (4.1) to form subsequent COZY vectors.

COROLLARY 2. *Assume condition* C.1. *Then*:
(i) $\text{Span}\{\boldsymbol{\Sigma}_{yzz}^j \boldsymbol{\beta}_{yz} : \ j = 0, 1, \ldots\} \subseteq \mathcal{S}_{\text{E}(Y|Z)}$, *and*
(ii) $\text{Span}\{\boldsymbol{\Sigma}_{rzz}^j \boldsymbol{\beta}_{yz} : \ j = 0, 1, \ldots\} \subseteq \mathcal{S}_{\text{E}(Y|Z)}$, *where* $\boldsymbol{\Sigma}_{rzz}$ *is as defined in* (3.3).

We restrict most of the following discussion to case (i); similar comments hold for case (ii).

It follows from Corollary 2 that, unless $\boldsymbol{\beta}_{yz}$ is an eigenvector of $\boldsymbol{\Sigma}_{yzz}$, the sequence of COZY vectors $\boldsymbol{\beta}_{yz}$, $\boldsymbol{\Sigma}_{yzz}\boldsymbol{\beta}_{yz}$, $\boldsymbol{\Sigma}_{yzz}^2\boldsymbol{\beta}_{yz}, \ldots$ provides a set of different vectors in $\mathcal{S}_{\text{E}(Y|Z)}$. One question that remains is how large j must be in order for the first j vectors, $\boldsymbol{\beta}_{yz}, \ldots, \boldsymbol{\Sigma}_{yzz}^{j-1}\boldsymbol{\beta}_{yz}$, to exhaust all possible vectors in the sequence. This is important because in practice we can compute only a finite number of these vectors. This question is answered by the next proposition; its proof is straightforward and omitted.

PROPOSITION 3. *Let* $\mathbf{A}$ *be a* $p \times p$ *matrix and* $\boldsymbol{\beta}$ *be a* p*-dimensional vector. If* $\mathbf{A}^j\boldsymbol{\beta}$ *belongs to the subspace spanned by* $\boldsymbol{\beta}, \ldots, \mathbf{A}^{j-1}\boldsymbol{\beta}$, *then so does* $\mathbf{A}^s\boldsymbol{\beta}$ *for any* $s > j$.

Since $\boldsymbol{\Sigma}_{yzz}^j \boldsymbol{\beta}_{yz}$ belongs to $\mathcal{S}_{\text{E}(Y|Z)}$, which has dimension k, Proposition 3 implies that there is an integer $s \le k$ such that the first s vectors in the sequence, $\boldsymbol{\beta}_{yz}, \ldots, \boldsymbol{\Sigma}_{yzz}^{s-1}\boldsymbol{\beta}_{yz}$, are linearly independent, and all the subsequent vectors are linearly dependent on them. This suggests the following estimation scheme (at the population level). First, compute the p COZY vectors $\boldsymbol{\beta}_{yz}$, $\boldsymbol{\Sigma}_{yzz}\boldsymbol{\beta}_{yz}, \ldots, \boldsymbol{\Sigma}_{yzz}^{p-1}\boldsymbol{\beta}_{yz}$. Let $\mathbf{B}$ be the p by p matrix with these vectors as columns and form the matrix $\mathbf{B}\mathbf{B}^T$. If $\mathbf{v}$ is an eigenvector of $\mathbf{B}\mathbf{B}^T$, then by Corollary 2 it is also the eigenvector of $\mathbf{P}_\gamma\mathbf{B}\mathbf{B}^T\mathbf{P}_\gamma$, where $\mathbf{P}_\gamma$ is the projection matrix onto $\mathcal{S}_{\text{E}(Y|Z)}$, and hence $\mathbf{v} \in \mathcal{S}_{\text{E}(Y|Z)}$. Now let $\mathbf{v}_1, \ldots, \mathbf{v}_p$ be all the eigenvectors of $\mathbf{B}\mathbf{B}^T$, so ordered that their corresponding eigenvalues form a descending sequence $\lambda_1 \ge \cdots \ge \lambda_k \ge 0 = \cdots = 0$. Then $\mathbf{v}_1$ is the first eigenvector in $\mathcal{S}_{\text{E}(Y|Z)}$, $\mathbf{v}_2$ is the second eigenvector, and so on.

Because this method is based on the iterative transformation of $\boldsymbol{\beta}_{yz}$ by the Hessian matrix $\boldsymbol{\Sigma}_{yzz}$, we call it the Iterative Hessian Transformation method, or IHT. Note, however, that when $\mathbf{Z}$ is non-Gaussian $\boldsymbol{\Sigma}_{yzz}$ ceases to have the interpretation of the Hessian matrix of the regression function [Cook (1998b)].

The *IHT directions* $\mathbf{v}_j$ can be back-transformed to the original scale: $\mathbf{u}_j = \boldsymbol{\Sigma}_{xx}^{-1/2}\mathbf{v}_j$, $j = 1, \ldots, p$. *Sample IHT* directions can be constructed by replacing

$\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{yzz}$ and $\boldsymbol{\Sigma}_{rzz}$ by their sample estimates, and $\boldsymbol{\beta}_{yz}$ by the vector of least squares coefficients. The linear combinations $\mathbf{v}_j^T\mathbf{Z} = \mathbf{u}_j^T(\mathbf{X} - \mathrm{E}(\mathbf{X}))$ will be called the *IHT predictors*.

Let the columns of the matrix $\boldsymbol{\eta}$ form a basis for $\mathcal{S}_{Y|Z}$. Li (1992, Theorem 6.1) showed in effect that if $\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})$ is linear and $\boldsymbol{\beta} \in \mathcal{S}_{Y|Z}$ then $\mathcal{S}(\boldsymbol{\Sigma}_{yzz}\boldsymbol{\beta}) \in \mathcal{S}_{Y|Z}$. While Li's result is related to that of Corollary 2, there are two key differences. First, Li's result is in terms of the central subspace and not the more specific CMS as defined here. Second, Li used his result to address the robustness of pHd rather than to suggest the possibility of iterating to find additional vectors in the CMS.

4.3. *Illustrations.* The results of the previous section show that we may need all p COZYvectors to be assured of exhausting $\mathcal{S}_{\mathrm{E}(Y|Z)}$, but in practice we may not need more than the first $k = \dim(\mathcal{S}_{\mathrm{E}(Y|Z)})$ of them. For instance, if $k = 2$, $\boldsymbol{\beta}_{yz}$ and $\boldsymbol{\Sigma}_{yzz}\boldsymbol{\beta}_{yz}$ could be sufficient, as illustrated in the following example.

To provide some insight into the IHT method, consider the regression model

$$Y = \boldsymbol{\alpha}_1^T\mathbf{Z} + f(\boldsymbol{\alpha}_2^T\mathbf{Z}) + \varepsilon,$$

where $\varepsilon \perp\!\!\!\perp \mathbf{Z}$ and without loss of generality $\mathrm{E}(Y) = 0$. We assume that $\mathrm{E}(\mathbf{Z}|\boldsymbol{\alpha}_1^T\mathbf{Z})$, $\mathrm{E}(\mathbf{Z}|\boldsymbol{\alpha}_2^T\mathbf{Z})$ and $\mathrm{E}(\mathbf{Z}|\boldsymbol{\alpha}_1^T\mathbf{Z}, \boldsymbol{\alpha}_2^T\mathbf{Z})$ are linear. Also, without loss of generality, we constrain $\boldsymbol{\alpha}_2^T\mathbf{Z}$ and $f(\boldsymbol{\alpha}_2^T\mathbf{Z})$ to be uncorrelated.

The CMS is spanned by $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$, and

$$\begin{aligned}\boldsymbol{\beta}_{yz} &= \mathrm{Cov}(Y, \mathbf{Z}) \\ &= \boldsymbol{\alpha}_1 + \frac{\boldsymbol{\alpha}_2^T\mathrm{E}[\boldsymbol{\alpha}_2^T\mathbf{Z} f(\boldsymbol{\alpha}_2^T\mathbf{Z})]}{\|\boldsymbol{\alpha}_2\|^2} \\ &= \boldsymbol{\alpha}_1.\end{aligned}$$

Next,

$$\begin{aligned}\boldsymbol{\Sigma}_{yzz}\boldsymbol{\beta}_{yz} &= \mathrm{E}(Y\mathbf{Z}\mathbf{Z}^T)\boldsymbol{\alpha}_1 \\ &= \mathrm{E}[(\boldsymbol{\alpha}_1^T\mathbf{Z})^2\mathbf{Z}] + \mathrm{E}[f(\boldsymbol{\alpha}_2^T\mathbf{Z})\mathbf{Z}\mathbf{Z}^T\boldsymbol{\alpha}_1] \\ &= \frac{\boldsymbol{\alpha}_1}{\|\boldsymbol{\alpha}_1\|^2}\mathrm{E}[(\boldsymbol{\alpha}_1^T\mathbf{Z})^3] + (\mathbf{P}_{\alpha_1} + \mathbf{P}_{Q_{\alpha_1}\alpha_2})\mathrm{E}[f(\boldsymbol{\alpha}_2^T\mathbf{Z})\mathbf{Z}\mathbf{Z}^T\boldsymbol{\alpha}_1] \\ &= \frac{\boldsymbol{\alpha}_1}{\|\boldsymbol{\alpha}_1\|^2}\mathrm{E}[(\boldsymbol{\alpha}_1^T\mathbf{Z})^3] + \frac{\boldsymbol{\alpha}_1}{\|\boldsymbol{\alpha}_1\|^2}\mathrm{E}[(\boldsymbol{\alpha}_1^T\mathbf{Z})^2 f(\boldsymbol{\alpha}_2^T\mathbf{Z})] \\ &\quad + \frac{\mathbf{Q}_{\alpha_1}\boldsymbol{\alpha}_2}{\|\mathbf{Q}_{\alpha_1}\boldsymbol{\alpha}_2\|^2}\mathrm{E}[f(\boldsymbol{\alpha}_2^T\mathbf{Z})(\boldsymbol{\alpha}_2^T\mathbf{Q}_{\alpha_1}\mathbf{Z})(\boldsymbol{\alpha}_1^T\mathbf{Z})].\end{aligned} \tag{4.2}$$

The first two terms on the right-hand side of (4.2) are in $\mathcal{S}(\boldsymbol{\beta}_{yz})$. The first term will be 0 if $\boldsymbol{\alpha}_1^T\mathbf{Z}$ is symmetric. The second term will generally be nonzero. The third term is in the subspace spanned by the part of $\boldsymbol{\alpha}_2$ that is orthogonal to $\boldsymbol{\alpha}_1$ and

that is what we want. It generally depends on higher order moments and we would expect it to be nonzero, although it can be zero in carefully constructed cases. For example, it equals zero if $\boldsymbol{\alpha}_1$ is orthogonal to $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_1^T\mathbf{Z} \perp\!\!\!\perp \boldsymbol{\alpha}_2^T\mathbf{Z}$.

Using $\boldsymbol{\Sigma}_{rzz}$ we obtain similar results,

$$\begin{aligned}\boldsymbol{\Sigma}_{rzz}\boldsymbol{\beta}_{yz} &= \mathrm{E}[f(\boldsymbol{\alpha}_2^T\mathbf{Z})\mathbf{Z}\mathbf{Z}^T\boldsymbol{\alpha}_1] \\ &= \frac{\boldsymbol{\alpha}_1}{\|\boldsymbol{\alpha}_1\|^2}\mathrm{E}[(\boldsymbol{\alpha}_1^T\mathbf{Z})^2 f(\boldsymbol{\alpha}_2^T\mathbf{Z})] \\ &\quad + \frac{\mathbf{Q}_{\alpha_1}\boldsymbol{\alpha}_2}{\|\mathbf{Q}_{\alpha_1}\boldsymbol{\alpha}_2\|^2}\mathrm{E}[f(\boldsymbol{\alpha}_2^T\mathbf{Z})(\boldsymbol{\alpha}_2^T\mathbf{Q}_{\alpha_1}\mathbf{Z})(\boldsymbol{\alpha}_1^T\mathbf{Z})].\end{aligned} \tag{4.3}$$

This result differs from (4.2) by the term $\boldsymbol{\alpha}_1\mathrm{E}[(\boldsymbol{\alpha}_1^T\mathbf{Z})^3]/\|\boldsymbol{\alpha}_1\|^2$ which belongs to $\mathcal{S}(\boldsymbol{\beta}_{yz})$, but otherwise the results are essentially the same. In particular,

$$\mathcal{S}(\boldsymbol{\beta}_{yz}, \boldsymbol{\Sigma}_{yzz}\boldsymbol{\beta}_{yz}) = \mathcal{S}(\boldsymbol{\beta}_{yz}, \boldsymbol{\Sigma}_{rzz}\boldsymbol{\beta}_{yz}) = \mathcal{S}_{\mathrm{E}(Y|Z)}$$

provided none of the key terms vanishes.

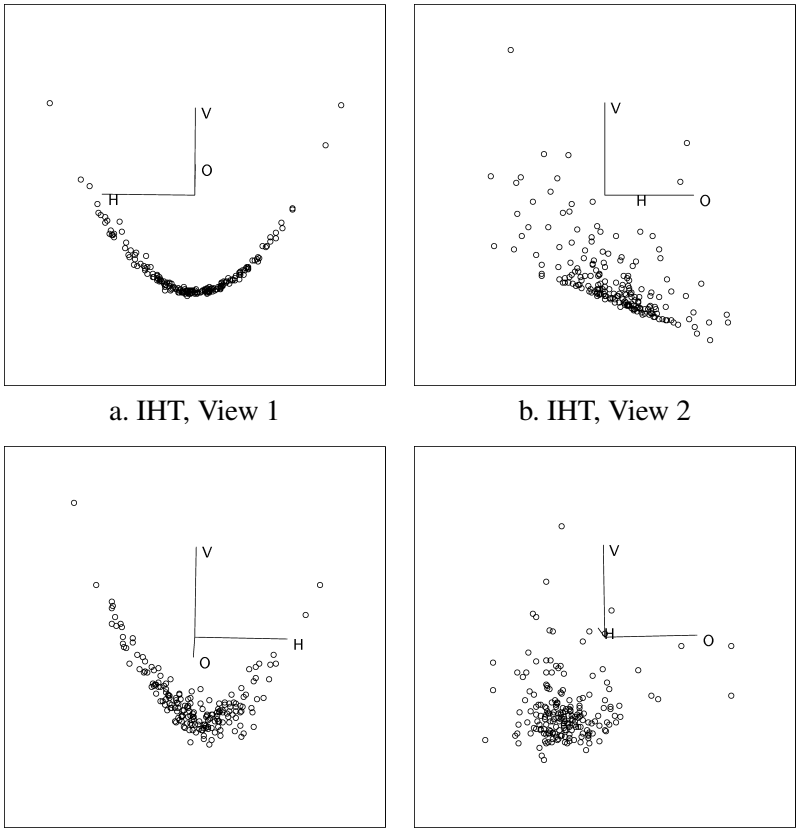
For a numerical illustration, we generated 200 observations on 5 predictors and a response as follows:

$$\begin{aligned} X_1 &= \varepsilon_1, \\ X_2|X_1 &= X_1 + \varepsilon_2, \\ X_3 &= \varepsilon_3, \\ X_4|X_2 &= (1 + X_2/2)\varepsilon_4, \\ X_5 &= \varepsilon_5, \\ Y &= X_1 + X_2^2/2. \end{aligned}$$

All errors ε_k are independent standard normal random variables. The response Y was generated without error to emphasize the qualitative nature of the results. The CMS is spanned by $(1, 0, 0, 0, 0)^T$ and $(0, 1, 0, 0, 0)^T$. Condition C.1 holds, but C.2 does not hold because $\mathrm{Var}(X_4|X_2) = (1 + X_2/2)^2$ which is correlated with Y. Table 1 gives the first two pHd directions, $\widehat{\mathbf{h}}_1$ and $\widehat{\mathbf{h}}_2$, and the first two sample IHT

TABLE 1
Sample pHd directions $\widehat{\mathbf{h}}_1$ and $\widehat{\mathbf{h}}_2$ and IHT directions $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$ from the simulated data

	$\widehat{\mathbf{h}}_1$	$\widehat{\mathbf{h}}_2$	$\widehat{\mathbf{u}}_1$	$\widehat{\mathbf{u}}_2$
X_1	−0.098	0.166	0.022	−0.996
X_2	−0.984	−0.011	0.999	0.024
X_3	0.050	0.151	0.001	0.062
X_4	−0.142	−0.974	0.032	−0.016
X_5	−0.017	0.035	0.016	0.017



a. IHT, View 1 b. IHT, View 2

c. pHd, View 1 d. pHd, View 2

FIG. 1. *Summary views from IHT and pHd for the simulated data. View* 1 *shows the quadratic component while View* 2 *is provided for contrast. For IHT,* $V = Y$ *and* $(H, O) = (\widehat{\mathbf{v}}_1^T\mathbf{Z}, \widehat{\mathbf{v}}_2^T\mathbf{Z})$. *For pHd,* $V = Y$ *and* $(H, O) = (\widehat{\mathbf{h}}_1^T\mathbf{Z}, \widehat{\mathbf{h}}_2^T\mathbf{Z})$.

directions, $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$. pHd found X_2 and X_4 to be the important predictors, while IHT correctly picked X_1 and X_2. In effect, pHd missed the linear component in favor of the quadratic component and X_4. Figure 1 gives a visual representation of these results. The response surface from IHT gives a very good representation of the true surface, while the surface for pHd shows only a relatively rough quadratic.

The fact that pHd missed the linear trend in this example may not be surprising because it is not very effective at finding linear trends in the first place [Cook (1998b)]. In addition, as illustrated by this example, failure of condition C.2 can influence the behavior of pHd. In contrast, IHT can find linear trends and does not require C.2 to be effective.

In a variety of similar simulations, IHT was found to be superior to pHd whenever there is a clear linear trend in the data, even if condition C.2 holds. The linear trend is necessary to provide a reasonable estimate of $\boldsymbol{\beta}_{yz}$, which is needed to prime the procedure. If there is not a linear trend, IHT failed as expected. Basing

IHT on the residuals r (i.e., using $\boldsymbol{\Sigma}_{rzz}^{j}\boldsymbol{\beta}_{yz}$) instead of the response worked well in many situations.

5. Asymptotic behavior. In the previous sections we focused on the characterization and estimation of the CMS *at the population level*. A full-fledged asymptotic analysis of the sample properties exceeds the scope of this paper, and will be developed in the authors' future work. In this section we provide a few basic ideas and results. Three types of asymptotic problems will concern us: the $\sqrt{n}$-consistency of the estimators, the asymptotic effects of standardization, and the test statistic for determining the dimension of the CMS. We will now outline the asymptotic development for IHT.

The $\sqrt{n}$-consistency is easy to demonstrate. Let $\widehat{\mathbf{B}}$ be the sample version of $\mathbf{B}$; that is,

$$\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\beta}}_{yz}, \widehat{\boldsymbol{\Sigma}}_{yzz}\widehat{\boldsymbol{\beta}}_{yz}, \ldots, \widehat{\boldsymbol{\Sigma}}_{yzz}^{p-1}\widehat{\boldsymbol{\beta}}_{yz}),$$

where $\widehat{\boldsymbol{\beta}}_{yz} = \mathrm{E}_n(Y\mathbf{Z})$, $\widehat{\boldsymbol{\Sigma}}_{yzz} = \mathrm{E}_n(Y\mathbf{Z}\mathbf{Z}^T)$. By the central limit theorem $\widehat{\boldsymbol{\beta}}_{yz}$ and $\widehat{\boldsymbol{\Sigma}}_{yzz}$ are $\sqrt{n}$-consistent. Hence $\widehat{\mathbf{B}}\widehat{\mathbf{B}}^T$, being a (matrix-valued) smooth function of $\widehat{\boldsymbol{\beta}}_{yz}$ and $\widehat{\boldsymbol{\Sigma}}_{yzz}$, is a $\sqrt{n}$-consistent estimator of $\mathbf{B}\mathbf{B}^T$. Consequently the eigenvectors of the former are $\sqrt{n}$-consistent estimators of the eigenvectors of the latter.

We have described the estimators for CMS in terms of the standardized explanatory vectors $\mathbf{Z}$ to simplify the presentation. In practice, we first transform the observed $\mathbf{X}$ into $\widehat{\mathbf{Z}}$ by $\widehat{\mathbf{Z}} = \widehat{\boldsymbol{\Sigma}}_{xx}^{-1/2}(\mathbf{X} - \mathrm{E}_n(\mathbf{X}))$, and then apply IHT to $\widehat{\mathbf{Z}}$ to obtain the estimated vectors in $\mathcal{S}_{\mathrm{E}(Z|Y)}$, say $\widehat{\boldsymbol{\nu}}_1, \ldots, \widehat{\boldsymbol{\nu}}_k$. They are then multiplied by $\widehat{\boldsymbol{\Sigma}}_{xx}^{-1/2}$ to become the estimated vectors in $\mathcal{S}_{\mathrm{E}(Y|X)}$, say $\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_k$. Since both transformations involve only $\sqrt{n}$-consistent estimators such as $\widehat{\boldsymbol{\Sigma}}_{xx}$ and $\mathrm{E}_n(\mathbf{X})$, they do not affect the $\sqrt{n}$-consistency of the estimators of the CMS. In other words $\widehat{\boldsymbol{\nu}}_1, \ldots, \widehat{\boldsymbol{\nu}}_k$ are $\sqrt{n}$-consistent estimators of the vectors in $\mathcal{S}_{\mathrm{E}(Y|Z)}$, and $\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_k$ are $\sqrt{n}$-consistent estimators of the vectors in $\mathcal{S}_{\mathrm{E}(Y|X)}$.

An asymptotic test for dimension can be developed along the following lines. If the rank of $\mathbf{B}$ is k, then the smallest $p - k$ eigenvalues $\mathbf{B}\mathbf{B}^T$ are 0. Hence the corresponding eigenvalues of $\widehat{\mathbf{B}}\widehat{\mathbf{B}}^T$ behave like noise and their sum should converge to a definite distribution. This sum, together with its asymptotic distribution, can then be used to test the hypothesis $H_0 : \mathrm{rank}(\mathbf{B}\mathbf{B}^T) = k$. The largest k for which this hypothesis is rejected is an estimator of the dimension of the CMS. To see how this asymptotic distribution can be derived, let vec be the transformation that maps a matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_p)$ to the column vector $(\mathbf{a}_1^T, \ldots, \mathbf{a}_p^T)^T$. For any p-dimensional vector $\boldsymbol{\beta}$ and $p \times p$ matrix $\boldsymbol{\Sigma}$, let

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \left(\boldsymbol{\beta}^T, \mathrm{vec}(\boldsymbol{\Sigma})^T\right)^T \quad \text{and} \quad \mathbf{b}(\boldsymbol{\theta}) = \mathrm{vec}\left(\boldsymbol{\beta}, \ldots, \boldsymbol{\Sigma}^{p-1}\boldsymbol{\beta}\right),$$

and denote $\boldsymbol{\theta}(\widehat{\boldsymbol{\beta}}_{yz}, \widehat{\boldsymbol{\Sigma}}_{yzz})$ by $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}(\boldsymbol{\beta}_{yz}, \boldsymbol{\Sigma}_{yzz})$ by $\boldsymbol{\theta}_0$. Because the components of $\widehat{\boldsymbol{\theta}}$ are moment estimators, by the central limit theorem $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in

distribution to a normal random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}(\boldsymbol{\theta}_0)$. By construction, $\mathbf{b}(\boldsymbol{\theta})$ is a vector-valued smooth function of $\boldsymbol{\theta}$ and $\mathbf{b}(\boldsymbol{\theta}_0) = \operatorname{vec}(\mathbf{B})$. Hence, applying the delta method, we have the convergence

$$\sqrt{n}\big(\operatorname{vec}(\widehat{\mathbf{B}}) - \operatorname{vec}(\mathbf{B})\big) \xrightarrow{\mathcal{L}} N\left(\mathbf{0},\ \frac{\partial \mathbf{b}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T}\ \mathbf{V}(\boldsymbol{\theta}_0)\ \frac{\partial \mathbf{b}^T(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right).$$

From here we can apply the Eaton–Tyler factorization [Eaton and Tyler (1994)] to derive the desired asymptotic distribution.

6. Recumbent cows. For unknown reasons, pregnant dairy cows can become recumbent—they lie down—either shortly before or after calving. This condition can be serious, and frequently leads to death of the cow. Clark, Henderson, Hoggard, Ellison and Young (1987) analyze data collected at the Ruakura (N.Z.) Animal Health Laboratory on a sample of recumbent cows. We use 254 cases with complete records to illustrate the dimension-reduction methods described in this article. The response is binary, $Y = 1$ for surviving cows, $Y = 0$ otherwise.

The $p = 3$ predictors are log(AST), logarithm of serum asparate amino transferase (U/l at 30C); log(CK), logarithm of serum creatine phosphokinase (U/l at 30C); and log(UREA), logarithm of serum urea (mmol/l). These predictors were used in one of the candidate logistic models studied by Clark et al. We deleted two anomalous cases at the outset, as did Clark et al. (1987).

The transformations of the blood measurements to logarithms effectively induced the linear conditional predictor expectations of the kind discussed in this article (condition C.1), although this rationale was not stated in the report by Clark et al. (1987). However, an analysis of the predictors indicated that condition C.2 might not be reasonable. For example, the score test for heteroscedasticity [Cook and Weisberg (1983)] in the linear regression of log(UREA) on (log(AST), log(CK)) has a p-value of 0.001. The lack of constant variances $\operatorname{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ may cause problems for pHd because this allows $\operatorname{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ and Y to be correlated. To avoid this issue we would like to use a dimension reduction method that requires only condition C.1. SIR is one possibility, but because the response is binary it can find at most one direction in $\mathcal{S}_{\mathrm{E}(Y|Z)}$, a limitation that is shared by all methods covered by Theorem 1. In short, IHT introduced in Section 4.2 seems to be the only known method that can find multiple directions in $\mathcal{S}_{\mathrm{E}(Y|Z)}$ and requires only condition C.1.

Table 2 shows the sample IHT directions from the regression of the binary response on the three predictors using a sample version of the population calculations described in Section 4.2. The directions were computed in the scale of the original predictors, except the predictors have been standardized marginally to have sample standard deviation equal to one. All three sample IHT vectors $\widehat{\boldsymbol{\Sigma}}^s_{yzz}\widehat{\boldsymbol{\beta}}_{yz}$ were used to form the sample version of $\mathbf{B}$.

Applying the graphical regression methods proposed by Cook (1996, 1998a) to the regression of Y on the three IHT predictors, we inferred that a good

TABLE 2
Sample IHT directions $\widehat{\mathbf{u}}_j = \widehat{\boldsymbol{\Sigma}}_{xx}^{-1/2}\widehat{\mathbf{v}}_j$ from the recumbent cow data. Predictors have been standardized marginally to have sample standard deviation equal to one

	$\widehat{\mathbf{u}}_1$	$\widehat{\mathbf{u}}_2$	$\widehat{\mathbf{u}}_3$
log(AST)	0.761	0.041	0.716
log(CK)	0.387	−0.595	−0.695
log(UREA)	0.521	0.803	−0.061

summary is provided by the binary response plot for the first two IHT predictors shown in Figure 2. The correlation between the first IHT predictor and the linear combination of the predictors from a logistic fit is 0.999, so this discussion covers what might be a first step for many. Our interpretation of the plot in Figure 2 is based on visual comparisons of the empirical conditional distributions of the IHT predictors given the response. We see from the figure that the conditional distributions for the first two IHT predictors differ primarily in location and scale;

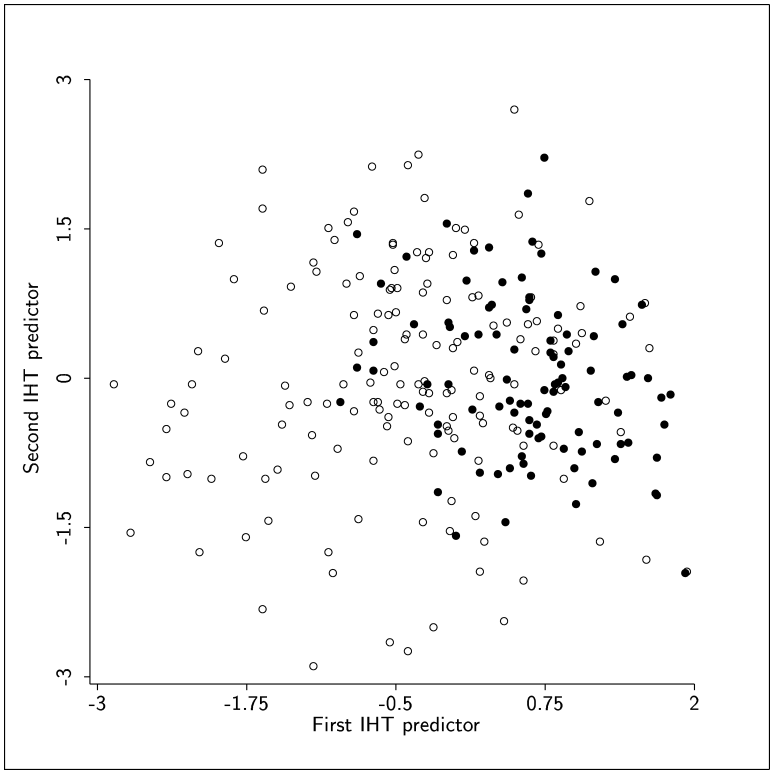


FIG. 2. *Binary response plot for the first two IHT predictors in the recumbent cow data. Filled circles correspond to survivors.*

difference in orientation seems less important. Additionally, both conditional distributions seem consistent with samples from bivariate normal distributions. Using this as a guide, we concluded that a first logistic model should include (a) linear and quadratic terms in the first IHT predictor since the conditional distributions for that predictor differ in both location and scale, and (b) a quadratic term in the second IHT predictor since the conditional distributions for that predictor differ primarily in scale. The fit of a logistic model with the three IHT predictors, their squares, and a cross product between the first two predictors sustains these conclusions, as the ratios of the estimated coefficients to their standard errors are less than one for all terms except the linear and quadratic terms in the first predictor and the quadratic in the second. And the absolute coefficients for these terms are all larger than about 2 standard errors.

The goal of this example is to illustrate that IHT summary plots can be informative and can capture the primary dependence in the data. Subsequent use of the summary plot must depend strongly on the applications context.

7. Discussion.

7.1. *Overview.* We addressed the problem of how to do dimension reduction in regression when the mean function alone, rather than the whole conditional distribution, is of interest. From this consideration the Central Mean Subspace arises naturally as the inferential object. The CMS is that part of the central subspace that captures all the information about the mean function available in the predictors, and shares many parallel properties with the latter. Moreover, the CMS categorizes transparently the previous methods into two kinds: those, such as OLS and pHd, that estimate vectors in the CMS and those, such as SIR and SAVE, that can pick up vectors outside the CMS. This gives us a clear idea of what these estimators do, and thereby provides guidance on their use.

The new IHT method, as well as other methods that might be developed from the class of COZY vectors, have two potential advantages: they allow the investigator to focus on the mean function and they do not require constraints on the conditional variances variances $\mathrm{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$, although linear conditional expectations $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ are still required. Our simulations show that IHT performs better than pHd when the conditional variance $\mathrm{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ is not a constant. We make no claim that IHT method is more "powerful" than pHd when both conditions C.1 and C.2 hold. The IHT predictors, like the pHd predictors, are partially ordered on their likely relation to the CMS. Plots of the response versus the first few IHT predictors and, in particular, graphical regression can be used to aid in determining the number of "significant" directions just as they are used with other dimension reduction methods in regression [see Cook (1998a) for an overview]. However, formal testing methods are required for full effectiveness. Such methods are under consideration, as outlined in Section 5.

7.2. *Comparing $\mathcal{S}_{Y|Z}$ and $\mathcal{S}_{\mathrm{E}(Y|Z)}$.* A comparison of $\mathcal{S}_{Y|Z}$ and $\mathcal{S}_{\mathrm{E}(Y|Z)}$ may be of general interest in all regressions and of particular interest in some regressions. Because $\mathcal{S}_{\mathrm{E}(Y|Z)} \subseteq \mathcal{S}_{Y|Z}$, we can use their dimensions as a rough indicator of the difference between the two subspaces. If $\dim(\mathcal{S}_{\mathrm{E}(Y|Z)}) = \dim(\mathcal{S}_{Y|Z})$ then all of the regression structure is captured by the mean, while there may be much more to the regression structure otherwise.

One method for inferring about aspects of $\mathcal{D} = \mathcal{S}_{Y|Z} \setminus \mathcal{S}_{\mathrm{E}(Y|Z)}$ can be developed from recent results on *partial dimension reduction* by Chiaromonte, Cook and Li (2002). Let $\boldsymbol{\gamma}$ be an orthonormal basis for $\mathcal{S}_{\mathrm{E}(Y|Z)}$ and let $(\boldsymbol{\gamma}, \boldsymbol{\gamma}_0)$ be an orthonormal basis for $\mathbb{R}^p$. Define $\mathbf{W} = \boldsymbol{\gamma}^T\mathbf{Z}$ and $\mathbf{V} = \boldsymbol{\gamma}_0^T\mathbf{Z}$. Then, assuming it exists, the *central partial subspace* $\mathcal{S}_{Y|\mathbf{V}}^{(\mathbf{W})}$ for $Y|(\mathbf{V}, \mathbf{W})$ is the intersection of all subspaces $\mathcal{S}(\mathbf{A})$ such that

$$Y \perp\!\!\!\perp \mathbf{V} | (\mathbf{A}^T\mathbf{V}, \mathbf{W}). \tag{7.1}$$

It can be shown that $\mathcal{S}_{Y|Z} = \mathcal{S}_{\mathrm{E}(Y|Z)} + \boldsymbol{\gamma}_0 \mathcal{S}_{Y|\mathbf{V}}^{(\mathbf{W})}$ and consequently we can infer about $\mathcal{D}$ via $\mathcal{S}_{Y|\mathbf{V}}^{(\mathbf{W})}$.

For example, suppose $\mathbf{Z}$ is normally distributed so that $\mathbf{W} \perp\!\!\!\perp \mathbf{V}$. Then (7.1) holds if and only if $(Y, \mathbf{W}) \perp\!\!\!\perp \mathbf{V} | \mathbf{A}^T\mathbf{V}$. Thus, we can infer about $\mathcal{S}_{Y|\mathbf{V}}^{(\mathbf{W})}$ by applying SIR/SAVE to the multivariate regression of $(Y, \mathbf{W})$ on $\mathbf{V}$. Because $\boldsymbol{\gamma}$ will usually be estimated rather than known, new asymptotic tests need to be developed for careful application.

APPENDIX: PROOFS

PROOF OF PROPOSITION 1. That (i) implies (ii) is immediate. That (iii) implies (i) is also immediate, because, if $\mathrm{E}(Y|\mathbf{X})$ is a function of $\boldsymbol{\alpha}^T\mathbf{X}$, then, given $\boldsymbol{\alpha}^T\mathbf{X}$, $\mathrm{E}(Y|\mathbf{X})$ is a constant, and hence independent of any other random variable. Now let's prove that (ii) implies (iii). By (ii),

$$\mathrm{E}\{Y\mathrm{E}(Y|\mathbf{X})|\boldsymbol{\alpha}^T\mathbf{X}\} = \mathrm{E}(Y|\boldsymbol{\alpha}^T\mathbf{X})\mathrm{E}\{\mathrm{E}(Y|\mathbf{X})|\boldsymbol{\alpha}^T\mathbf{X}\}.$$

The left hand side is

$$\mathrm{E}[\mathrm{E}\{Y\mathrm{E}(Y|\mathbf{X})|\mathbf{X}\}|\boldsymbol{\alpha}^T\mathbf{X}] = \mathrm{E}\{[\mathrm{E}(Y|\mathbf{X})]^2|\boldsymbol{\alpha}^T\mathbf{X}\},$$

and the right hand side is $\{\mathrm{E}[\mathrm{E}(Y|\mathbf{X})|\alpha^T\mathbf{X}]\}^2$. Therefore $\mathrm{Var}[\mathrm{E}(Y|\mathbf{X})|\boldsymbol{\alpha}^T\mathbf{X}] = 0$. Thus, given $\boldsymbol{\alpha}^T\mathbf{X}$, $\mathrm{E}(Y|\mathbf{X})$ is a constant. □

PROOF OF THEOREM 1. We first rewrite $R(a, \mathbf{b})$ making use of the fact that $\boldsymbol{\gamma}$ is a basis for the central mean subspace:

$$\begin{aligned}
R(a, \mathbf{b}) &= \mathrm{E}[-Y(a + \mathbf{b}^T\mathbf{Z}) + \phi(a + \mathbf{b}^T\mathbf{Z})] \\
&= \mathrm{E}[-\mathrm{E}(Y|\boldsymbol{\gamma}^T\mathbf{Z})(a + \mathbf{b}^T\mathbf{Z}) + \phi(a + \mathbf{b}^T\mathbf{Z})] \\
&\geq \mathrm{E}[-\mathrm{E}(Y|\boldsymbol{\gamma}^T\mathbf{Z})(a + \mathbf{b}^T\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})) + \phi(a + \mathbf{b}^T\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}))] \\
&= \mathrm{E}[-Y(a + \mathbf{b}^T\mathbf{P}_\gamma\mathbf{Z}) + \phi(a + \mathbf{b}^T\mathbf{P}_\gamma\mathbf{Z})].
\end{aligned}$$

The second equality follows because $\boldsymbol{\gamma}$ is a basis for $\mathcal{S}_{\mathrm{E}(Y|Z)}$. The inequality follows because of convexity. The next equality stems from the linearity of $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ which is equivalent to requiring that $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}) = \mathbf{P}_\gamma\mathbf{Z}$, where $\mathbf{P}_\gamma$ is the projection onto $\mathcal{S}_{\mathrm{E}(Y|Z)}$ with respect to the usual inner product. Thus,

$$R(a, \mathbf{b}) \geq R(a, \mathbf{P}_\gamma\mathbf{b})$$

and the conclusion now follows because $\boldsymbol{\beta}$ is unique. □

PROOF OF THEOREM 2. It is sufficient to show that if $\mathrm{E}(Y) = 0$, then $\mathcal{S}\{\mathrm{E}(Y\mathbf{Z}\mathbf{Z}^T)\} \subseteq \mathcal{S}_{\mathrm{E}(Y|Z)}$. Assume then $\mathrm{E}(Y) = 0$. Now

$$\begin{aligned}\mathrm{E}(Y\mathbf{Z}\mathbf{Z}^T) &= \mathrm{E}\{Y\mathrm{E}(\mathbf{Z}\mathbf{Z}^T|\boldsymbol{\gamma}^T\mathbf{Z})\}\\ &= \mathrm{E}\{Y(\mathbf{P}_\gamma\mathbf{Z}\mathbf{Z}^T\mathbf{P}_\gamma + \mathrm{Var}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z}))\}\\ &= \mathbf{P}_\gamma\mathrm{E}(Y\mathbf{Z}\mathbf{Z}^T)\mathbf{P}_\gamma.\end{aligned}$$

The conclusion follows since $\mathcal{S}\{\mathbf{P}_\gamma\mathrm{E}(Y\mathbf{Z}\mathbf{Z}^T)\mathbf{P}_\gamma\} \subseteq \mathcal{S}_{\mathrm{E}(Y|Z)}$. □

PROOF OF PROPOSITION 2. By definition of $\mathcal{S}_{\mathrm{E}(Y|Z)}$, $Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\boldsymbol{\gamma}^T\mathbf{X}$. Because of the linearity of $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$, $\boldsymbol{\beta}_{yz} \in \mathcal{S}_{\mathrm{E}(Y|Z)}$ and hence

$$r(Y, \boldsymbol{\beta}_{yz}^T\mathbf{Z}) \perp\!\!\!\perp r(\mathrm{E}(Y|\mathbf{Z}), \boldsymbol{\beta}_{yz}^T\mathbf{Z})|\boldsymbol{\gamma}^T\mathbf{Z}.$$

Because r is linear in Y,

$$\mathrm{E}[r(Y, \boldsymbol{\beta}_{yz}^T\mathbf{Z})|\mathbf{Z}] = r(\mathrm{E}(Y|\mathbf{Z}), \boldsymbol{\beta}_{yz}^T\mathbf{Z}).$$

Thus $r \perp\!\!\!\perp \mathrm{E}(r|\mathbf{Z})|\boldsymbol{\gamma}^T\mathbf{Z}$, so that $\mathcal{S}_{\mathrm{E}(Y|Z)}$ is a mean dimension-reduction subspace for the regression of r on $\mathbf{Z}$, and

$$\mathcal{S}_{\mathrm{E}(r|Z)} \subseteq \mathcal{S}_{\mathrm{E}(Y|Z)}. \tag{A.1}$$

Next, let $\boldsymbol{\rho}$ be the matrix whose columns form a basis of $\mathcal{S}_{\mathrm{E}(r|Z)}$; that is, $\mathcal{S}_{\mathrm{E}(r|Z)} = \mathcal{S}(\boldsymbol{\rho})$. Then $r \perp\!\!\!\perp \mathrm{E}(r|\mathbf{Z})|\boldsymbol{\rho}^T\mathbf{Z}$. But by Proposition 1, given $\boldsymbol{\rho}^T\mathbf{Z}$, $\mathrm{E}(r|\mathbf{Z})$ is constant. Hence,

$$r \perp\!\!\!\perp \mathrm{E}(r|\mathbf{Z})|(\boldsymbol{\rho}^T\mathbf{Z}, \boldsymbol{\beta}_{yz}^T\mathbf{Z})$$

and therefore

$$(r, \boldsymbol{\beta}_{yz}^T\mathbf{Z}) \perp\!\!\!\perp (\mathrm{E}(r|\mathbf{Z}), \boldsymbol{\beta}_{yz}^T\mathbf{Z})|(\boldsymbol{\rho}^T\mathbf{Z}, \boldsymbol{\beta}_{yz}^T\mathbf{Z}).$$

Since $Y = (r - a(\boldsymbol{\beta}_{yz}^T\mathbf{Z}))/b(\boldsymbol{\beta}_{yz}^T\mathbf{Z})$, it follows that $Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{Z})|(\boldsymbol{\rho}^T\mathbf{Z}, \boldsymbol{\beta}_{yz}^T\mathbf{Z})$ and thus $\mathcal{S}_{\mathrm{E}(r|Z)} + \mathcal{S}(\boldsymbol{\beta}_{yz})$ is a mean dimension-reduction subspace for the regression of Y on $\mathbf{Z}$. Combining this with (A.1) gives the desired conclusion

$$\mathcal{S}_{\mathrm{E}(Y|Z)} \subseteq \mathcal{S}_{\mathrm{E}(r|Z)} + \mathcal{S}(\boldsymbol{\beta}_{yz}) \subseteq \mathcal{S}_{\mathrm{E}(Y|Z)} + \mathcal{S}(\boldsymbol{\beta}_{yz}) = \mathcal{S}_{\mathrm{E}(Y|Z)},$$

where the last equality follows because the linearity of $\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})$ implies $\boldsymbol{\beta}_{yz} \in \mathcal{S}_{\mathrm{E}(Y|Z)}$. $\square$

PROOF OF THEOREM 3. Let $W = U\mathrm{E}(Y|\mathbf{Z}) + V$ and, as before, let $\mathbf{P}_\gamma$ be the projection matrix onto $\mathcal{S}(\boldsymbol{\gamma})$. Then

$$\mathrm{E}\{(UY + V)\mathbf{Z}\} = \mathrm{E}\{(UE(Y|\mathbf{Z}) + V)\mathbf{Z}\} = \mathrm{E}(W\mathbf{Z}).$$

Since U and V are measurable with respect to $\boldsymbol{\gamma}^T\mathbf{Z}$, and since $\mathrm{E}(Y|\mathbf{Z}) = \mathrm{E}(Y|\boldsymbol{\gamma}^T\mathbf{Z})$ by Proposition 1, the random variable W is measurable with respect to $\boldsymbol{\gamma}^T\mathbf{Z}$. Hence

$$\mathrm{E}(W\mathbf{Z}) = \mathrm{E}\left\{W\mathrm{E}(\mathbf{Z}|\boldsymbol{\gamma}^T\mathbf{Z})\right\} = \mathbf{P}_\gamma \mathrm{E}(W\mathbf{Z}),$$

where the right hand side belong to $\mathcal{S}(\boldsymbol{\gamma})$. $\square$

PROOF OF COROLLARY 2. We only prove part (i); part (ii) can be proved similarly. In Corollary 1, take $u(t) = t$ and $v(t) = -t\mathrm{E}(Y)$. Then

$$\mathrm{E}\left[\left\{u\left(\boldsymbol{\beta}_{yz}^T\mathbf{Z}\right)Y + v\left(\boldsymbol{\beta}_{yz}^T\mathbf{Z}\right)\right\}\mathbf{Z}\right] = \boldsymbol{\Sigma}_{yzz}\,\boldsymbol{\beta}_{yz}$$

belongs to $\mathcal{S}(\boldsymbol{\gamma})$. Now suppose that $\boldsymbol{\delta} = \boldsymbol{\Sigma}_{yzz}^{r-1}\,\boldsymbol{\beta}_{yz}$ belongs to $\mathcal{S}(\boldsymbol{\gamma})$. Then $\boldsymbol{\delta}^T\mathbf{Z}$ is measurable with respect to $\boldsymbol{\gamma}^T\mathbf{Z}$. By Theorem 3, $\mathrm{E}\{(\boldsymbol{\delta}^T\mathbf{Z})Y\mathbf{Z}\} = \boldsymbol{\Sigma}_{yzz}\,\boldsymbol{\delta} = \boldsymbol{\Sigma}_{yzz}^{r}\,\boldsymbol{\beta}_{yz}$ also belongs to $\mathcal{S}(\boldsymbol{\gamma})$. Hence, by induction, all vectors of the form $\boldsymbol{\Sigma}_{yzz}^{r}\,\boldsymbol{\beta}_{yz}$ belong to $\mathcal{S}(\boldsymbol{\gamma})$. $\square$

Acknowledgments. We would like to thank two referees and an Associate Editor for their suggestions that led to significant improvement of this article.

REFERENCES

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

CHIAROMONTE, F., COOK, R. D. and LI, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Ann. Statist.* **30** 475–497.

CLARK, R. G., HENDERSON, H. V., HOGGARD, G. K. ELLISON, R. S. and YOUNG, B. J. (1987). The ability of biochemical and haematological tests to predict recovery in periparturient recumbent cows. *New Zealand Veterinary Journal* **35** 126–133.

COOK, R. D. (1992). Regression plotting based on quadratic predictors. In *L1-Statistical Analysis and Related Methods* (Y. Dodge, ed.) 115–127. North-Holland, Amsterdam.

COOK, R. D.(1994a). On the interpretation of regression plots. *J. Amer. Statist. Assoc.* **89** 177–189.

COOK, R. D. (1994b). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the Section on Physical and Engineering Sciences* 18–25. Amer. Statistic. Assoc., Alexandria, VA.

COOK, R. D. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.* **91** 983–992.

COOK, R. D.(1998a). *Regression Graphics*. Wiley, New York.

COOK, R. D. (1998b). Principal Hessian directions revisited. *J. Amer. Statist. Assoc.* **93** 84–100.

COOK, R. D. and NACHTSHEIM, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *J. Amer. Statist. Assoc.* **89** 592–599.

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

COOK, R. D. and WEISBERG, S. (1991). Discussion of "Sliced inverse regression for dimension reduction." *J. Amer. Statist. Assoc.* **86** 328–332.

COOK, R. D. and WEISBERG, S. (1999). Graphics in statistical analysis: Is the medium the message? *The American Statistician* **53** 29–37.

DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** 793–815.

EATON, M. L. and TYLER, D. (1994). The asymptotic distribution of singular values with application to canonical correlations and correspondence analysis. *J. Multivariate Anal.* **50** 238–264.

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

LI, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–342.

LI, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87** 1025–1039.

LI, K.-C. and DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1952.

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
1994 BUFORD AVENUE
ST. PAUL, MINNESOTA 55108
E-MAIL: dennis@stat.umn.edu

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
326 THOMAS BUILDING
UNIVERSITY PARK, PENNSYLVANIA 16802