Ann. Inst. Statist. Math.
Vol. 54, No. 4, 768–795 (2002)

SUFFICIENT DIMENSION REDUCTION AND GRAPHICS IN REGRESSION

FRANCESCA CHIAROMONTE[1] AND R. DENNIS COOK[2]

[1]*Department of Statistics, Pennsylvania State University, 411 Thomas Building, University Park, PA, 16802-2111, U.S.A.*
[2]*Department of Applied Statistics, School of Statistics, University of Minnesota, 352 Classroom-Office Building, 1994 Buford Avenue, St. Paul, MN 55108-6042, U.S.A.*

(Received September 1, 2000; revised July 24, 2001)

Abstract. In this article, we review, consolidate and extend a theory for *sufficient dimension reduction* in regression settings. This theory provides a powerful context for the construction, characterization and interpretation of low-dimensional displays of the data, and allows us to turn *graphics* into a consistent and theoretically motivated methodological body. In this spirit, we propose an iterative graphical procedure for estimating the meta-parameter which lies at the core of sufficient dimension reduction; namely, the central dimension-reduction subspace.

Key words and phrases: Sufficient dimension reduction, graphical displays, regression analysis.

1. Introduction

The overarching goal of a regression analysis is to understand how the conditional distribution of the univariate response Y given a vector X of p predictors depends on the value assumed by X. Attention is often restricted to the mean function $E(Y \mid X)$, and perhaps the variance function $\mathrm{Var}(Y \mid X)$. In full generality, though, the object of interest is the conditional distribution of $Y \mid X$, meant as a function of the value of X.

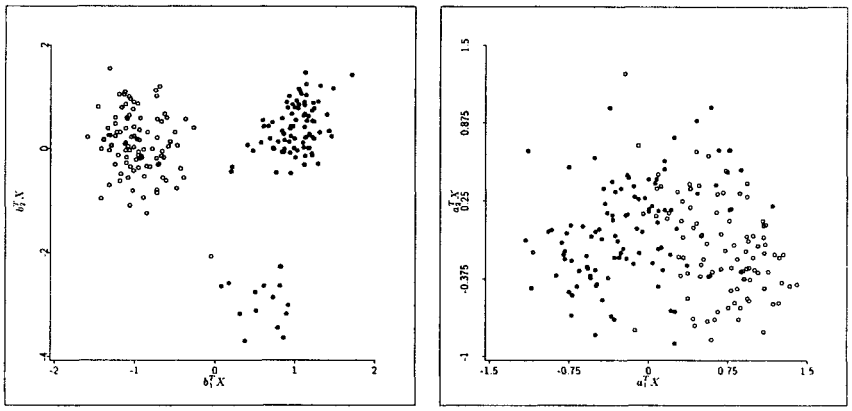
Graphical displays can be quite useful for investigating $Y \mid X$, especially when an adequate parsimoniously parameterized model is not available. Graphical displays can also be useful in the diagnostic phase of a model-based analysis, particularly when looking for patterns in the residuals. In the past decade, much literature has been devoted to using graphics in concert with dimension reduction. The latter is a leitmotif of statistics. For instance, if we are given a sample $z_1, \ldots, z_n$ from a normal distribution with mean μ and variance 1, we know that the sample mean $\bar{z}$ is sufficient for μ. Thus, we can replace the n-dimensional sample with the one-dimensional mean $\bar{z}$ without loss of information on μ. With an analogous rationale, one can attempt to reduce the dimension of X without losing information on $Y \mid X$, and without requiring a model for $Y \mid X$. Borrowing terminology from classical statistics, we call this *sufficient dimension reduction*. Sufficient dimension reduction has a two-fold connection with graphics: On the one hand, it leads to the pursuit of *sufficient summary plots*; that is, plots containing all of the regression information available from the sample. On the other, it provides a clear-cut setting for applying graphical methods in such a pursuit. As we will see, these methods employ so called supporting views.

1.1 *An introductory illustration*

Consider a regression involving counterfeit Swiss bank notes (Flury and Riedwyl (1988), p. 5). The binary response indicates a note's authenticity: $Y = 0$ for genuine notes and $Y = 1$ for counterfeit notes. There are $p = 6$ predictors in X, each giving a different aspect of the size of a note: length at the top, bottom, left and right edges, and along the diagonal and center. There are many ways to start an analysis of this regression. For example, we might inspect a scatter-plot matrix of the predictors, with the points marked to indicate the states of Y. Or we might begin with a logistic model, adding or deleting terms in the model as necessary in response to graphical or non-graphical diagnostics.

However, the idea of sufficient dimension reduction and the methods we describe later in the paper took us down a rather different data-analytic path. Without specifying a model for $Y \mid X$, we were able to assess that only two linear combinations of X, say $\beta_1' X$ and $\beta_2' X$, are needed to characterize $Y \mid X$ fully. Letting $\beta = (\beta_1, \beta_2)$, this is based on the inference that Y is independent of X given $\beta' X$, so that the conditional distribution of $Y \mid X$ and $Y \mid \beta' X$ are the same. In effect, we were able to reduce the dimension of the analysis, passing from the original six predictors to $(\beta_1' X, \beta_2' X)$, without any evidence in the data that this reduction would result in loss of information on $Y \mid X$. Since all the information about Y that is available from X is contained in the two linear combinations, a 3D plot of Y versus $(\beta_1' X, \beta_2' X)$ is a sufficient summary plot for the regression. With a binary Y, this is equivalent to a 2D binary response plot (Cook (1996*a*)) with $(\beta_1' X, \beta_2' X)$ on the axes, and points marked to indicate the states of Y.

In the previous description, β_1 and β_2 are unknown. Again using the methods we describe later on, we estimated these vectors and hence the linear combinations $b_1' X$ and $b_2' X$. This yielded the *estimated sufficient summary plot* for the regression shown in Fig. 1(a) (the values in b_1 and b_2 being unimportant for the present discussion). We thus concluded that all the information about Y that is available from X is contained in this one plot. The bimodal distribution within the counterfeit notes could indicate a change in the manufacturing process, or two different counterfeiting operations. Also,



a. Sufficient summary plot. b. An interesting view.

Fig. 1. Swiss bank note data. Open circles denote authentic notes; filled circles counterfeit notes.

there appears to be a outlying authentic note, which could be a mislabeled counterfeit note or an indication of a second low-frequency mode among the authentic notes. It seems unlikely that we would have found this summary plot without using the graphical methods discussed in this article. Fig. 1(b) will be discussed later.

There are a variety of approaches to the graphical exploration of regression data and the pursuit of interesting low-dimensional projections. The approach based on sufficient dimension reduction differs from others because it allows us to identify views that contain all the regression information. Thus, all subsequent analysis, including model building, can flow from the sufficient summary plot.

1.2 *Things to come*

The previous illustration makes use of a body of literature devoted to dimension reduction and graphics in regression. Although with different emphasis, vocabulary and levels of symbolic description, much of this literature draws upon a common core of population level concepts, results, and related inferential methods.

One aim of this article is to consolidate this core into a general and consistent framework. This is accomplished translating, integrating and generalizing existing concepts, results and methods at a new level of abstraction. Once in place, the framework allows us to clarify the geometry of dimension reduction, and to shed light on the conditions underlying effectiveness of various methods. Moreover, it allows us to design a new iterative procedure that combines graphical and non-graphical methodology.

When organized in the framework we describe here, sufficient dimension reduction provides a powerful context for the use and interpretation of low-dimensional graphical displays of the data, establishing fundamental connections between such displays and $Y \mid X$. In particular, we consider *projective regression views*. These are objects of the type $\{Y, P_S X\}$, where S is a linear subspace of $\mathbb{R}^p$ and P_S indicates the orthogonal projection operator on S with respect to the standard inner product. With a slight abuse of language, we often call $\{Y, P_S X\}$ a *marginal view*, or simply a *view* when no confusion seems likely. When discussing graphical methods, we also consider special varieties of marginal views that are obtained *conditioning* (i.e. restricting attention to subpopulations), and replacing Y and/or X with properly defined *residuals*.

In practice, the view $\{Y, P_S X\}$ is constructed by plotting Y against plotting coordinates which are just linear combinations of X determined by any basis for S. For instance, the view shown in Fig. 1(a) can be thought of as a *coordinate version* of the marginal view $\{Y, P_{S_b} X\}$, where S_b denotes the subspace spanned by $\{b_1, b_2\}$. The concept of marginal view, which refers to a subspace of $\mathbb{R}^p$, facilitates discussion of general results, while coordinate versions of a marginal view are necessary for implementation in practice.

Section 2 is devoted to the definition of a meta-parameter characterizing $Y \mid X$; the *central dimension-reduction subspace* $S_{Y|X} \subseteq \mathbb{R}^p$. The view shown in Fig. 1(a) is based on an estimate $\hat{S}_{Y|X}$ of the central subspace and can be described symbolically as $\{Y, P_{\hat{S}_{Y|X}} X\}$. In terms of our previous notation, $S_b = \hat{S}_{Y|X}$. In Section 3, we consolidate non-graphical methods for inference on $S_{Y|X}$, while Section 4 concerns graphical methods to investigate $S_{Y|X}$ through low-dimensional supporting views.

The new iterative procedure we propose is graphical, in the sense that it targets $S_{Y|X}$ by employing supporting views at each stage. At the same time, it takes advantage of non-graphical methods, which can be used to select these views, and to enhance the

overall performance. The procedure is first introduced in the examples of Section 4, and further discussed later in the article. Besides its direct application to inference on the central dimension-reduction subspace, this procedure can also be used in the more familiar model-building phases of a regression analysis. Subsection 4.2.5 shows how it can serve as a model-checking device when used on residuals. The bank note data are revisited at various stages of the article, while Section 5 contains a separate data example. We conclude with general discussion in Section 6.

Special cases and/or coordinate versions of some of the statements in this article have been discussed in Cook (1998*a*). However, here we present some crucial facts as formal propositions, and provide detailed proofs for them in a technical appendix. The goal is to give full insight into the conditional independence reasoning, and the mathematical machinery, that lie behind our framework.

2. Dimension reduction subspaces and sufficient views

Consider a regression with response $Y \in \mathbb{R}^1$ and random predictor vector $X \in \mathbb{R}^p$. We assume the data to consist of n iid observations from the joint distribution of (Y, X), and that first and second moments exist. When the dimension of the predictor vector p is larger than 2 or 3, only low-dimensional projective regression views $\{Y, P_S X\}$ can be visualized in practice, so we need to determine their relationship with $Y \mid X$ in order to understand the regression information they contain. Interesting projective views can often be found easily with modern visualization tools. For example, we encountered the view shown in Fig. 1(b) during a preliminary visual tour of the bank note data. This view seemed interesting, as it provides a fairly good separation between the authentic and counterfeit notes. However, marginal views $\{Y, P_S X\}$ can be misleading, unless we know how they relate to $Y \mid X$. To resolve this issue, we study the existence of low-dimensional projective views that provide sufficient information about the regression. For the bank note data, an estimated sufficient view is given in Fig. 1(a).

The approach we use permits reduction to occur in terms of linear combinations of X. In symbols, we investigate the existence of $k \leq p$ linearly independent vectors $\{\eta_1, \ldots, \eta_k\}$ in $\mathbb{R}^p$ such that

$$Y \perp\!\!\!\perp X \mid (\eta_1' X, \ldots, \eta_k' X) \tag{2.1}$$

where $\perp\!\!\!\perp$ indicates independence. The statement is thus that Y is independent of X given the k linear combinations $\eta_j' X$, $j = 1, \ldots, k$.

Let $S_\eta = \text{Span}(\eta_1, \ldots, \eta_k)$. (2.1) would equivalently hold for any other spanning system of the subspace. We therefore prefer to write

$$Y \perp\!\!\!\perp X \mid P_{S_\eta} X. \tag{2.2}$$

Passing from the basis notation in (2.1) to the subspace notation in (2.2) moves us away from interpretability in terms of the original predictor variables, but simplifies the discussion and facilitates geometric understanding. In particular, it shows us that the conditional independence we are after is a "coordinate-free" attribute of subspaces.

Let $Q_{S_\eta} = I - P_{S_\eta}$ be the projection on the orthogonal complement of our subspace (throughout the rest of the article, $Q_{(\cdot)}$ will always stay for $I - P_{(\cdot)}$). If (2.2) holds, then $Y \mid P_{S_\eta} X \sim Y \mid X$ and $Q_{S_\eta} X$ can be neglected in all further analyses, without loss of information on the regression. In other words, the view $\{Y, P_{S_\eta} X\}$ is equivalent to the full view $\{Y, X\}$.

We call any subspace $S \subseteq \mathbb{R}^p$ for which (2.2) holds a *dimension-reduction subspace* (DRS) for the regression of Y on X (Li (1991), Cook (1994*a*)). The corresponding view $\{Y, P_S X\}$ is called a *sufficient view*.

Note that this set-up admits as a special case *sufficient variable selection*, which occurs when some of the coordinates of X can be used as the linear combinations guaranteeing conditional independence in (2.1)—the corresponding coordinate space can then be used in (2.2). In symbols, partitioning X as $X' = (X_1', X_2')$, sufficiency of X_1 (redundancy of X_2) is expressed by $Y \perp\!\!\!\perp X \mid X_1$, or equivalently $Y \perp\!\!\!\perp X_2 \mid X_1$. Interestingly, some variable selection methods proposed for regressions with multivariate responses employ very similar notions (see McKay (1977), Fujikoshi (1982), and references therein). If response and predictor are jointly normal, and the response is univariate, the null hypothesis these authors consider is exactly equivalent to $Y \perp\!\!\!\perp X_2 \mid X_1$.

Back to the general set-up expressed by (2.2), it is straightforward to show that the origin $\{0\}$ is a DRS for the regression of Y on X if and only if response and predictor vector are independent ($Y \perp\!\!\!\perp X$ unconditionally). At the other extreme, any regression admits at least one obvious DRS; namely the whole $\mathbb{R}^p$. The conditional independence in (2.2) becomes interesting when it holds for a non-obvious S which has $\dim(S) < p$. Most regressions admit several DRS's and therefore several sufficient views, because any subspace containing a DRS is itself a DRS. Naturally, we are interested in reducing the dimension as much as possible. A subspace $S_m \subseteq \mathbb{R}^p$ is called a *minimum* DRS for the regression of Y on X if it is a DRS of minimal dimension; that is, if $\dim(S_m) \leq \dim(S_{drs})$ for any other DRS S_{drs} (Cook (1994*a*)). The corresponding view $\{Y, P_{S_m} X\}$ is called a *minimal sufficient view*. Since all regressions admit at least one DRS, they also admit at least one minimum DRS.

Unfortunately, requiring minimal dimension is not always enough to single out a unique subspace. Cook avoided this problem first restricting attention to regressions with a unique minimum DRS (1994*a*), and then introducing a new type of space into the picture (1994*b*, 1996*a*): Consider the subspace obtained by intersecting all the DRS's for the regression of Y on X. If this subspace is itself a DRS, we call it the *central dimension-reduction subspace* for the regression, and indicate it with $S_{Y|X}$. The corresponding view $\{Y, P_{S_{Y|X}} X\}$ is called the *central view*, and the dimension $d_{Y|X}$ of $S_{Y|X}$ is called the *structural dimension*; we will refer to regressions as having 0D, 1D, ..., pD structure. The view in Fig. 1(a) is exactly the estimated central view for the Swiss bank note data, for which we inferred 2D structure; $\hat{d}_{Y|X} = 2$.

Although the central DRS does not always exist, it does exist for a wide class of regressions (see Subsection 2.2). When this space exists, it is unique by construction, and thus constitutes a well-defined object of inference. It is important to notice that uniqueness is achieved replacing the minimal dimension requirement $\dim(S_m) \leq \dim(S_{drs})$ with the stronger inclusion requirement $S_{Y|X} \subseteq S_{drs}$. On a technical note, if the central DRS exists, it clearly is also the unique minimum DRS. However, one can construct examples in which there is a unique minimum DRS, but the central DRS does not exist (see Cook (1998*a*), p. 106). Cook and Weisberg (1999) recently gave an introductory account of regression graphics based on central subspaces.

2.1 *Conceptual importance of the central subspace*

So far we have identified three types of subspaces, an arbitrary DRS, a minimum DRS, and the central DRS. In his development of sliced inverse regression, Li (1991)

used an *effective* DRS, which was represented as the span of $(\beta_1, \ldots, \beta_k)$ in the "model"

$$Y = f(\beta_1' X, \ldots, \beta_k' X, \varepsilon) \tag{2.3}$$

where f is an unknown function and $\varepsilon \perp\!\!\!\perp X$. Effective DRS's are similar in spirit to those under discussion, but were not defined explicitly. Additionally, as pointed out by Cook ((1998*b*), rejoinder), the "model" representation in (2.3) complicates matters when the response is binary (see Carrol and Li (1995)), as for example in the Swiss bank note data—the problem being how to conceptualize the error ε when dealing with a binary response (Cox and Snell (1968)). Nevertheless, Li's approach is a clear signpost in the evolution of regression graphics.

Since $Y \perp\!\!\!\perp X \mid P_{S_{Y|X}} X$, knowledge of the central DRS allows us to reduce the whole regression analysis to Y on $P_{S_{Y|X}} X$. Regardless of how Y depends on X, such dependence will be entirely embodied by $P_{S_{Y|X}} X$. The issue can be turned around: the central DRS is not affected by the way Y depends on X, as long as the dependence is exhaustively (and minimally in terms of algebraic dimension) conveyed by the projection $P_{S_{Y|X}} X$. In particular, the definition of a central DRS does not rely on a model for $Y \mid X$. Correspondingly, neither do the methods to investigate DRS's that we present later in this article. The definitions and methods are also independent of some traits of the response's nature, as for example being discrete as opposed to continuous. In this sense, we can unify the treatment of all regressions for which $S_{Y|X}$ exists.

Our next task is to gain an understanding of this class of regressions. This is important from a practical point of view, because inferential methods for $S_{Y|X}$ can be quite elusive when the central DRS does not exist in the first place. For example, suppose a regression has two minimum DRS's, represented by two distinct lines (1-dimensional subspaces) of $\mathbb{R}^p$. Then a method such as sliced inverse regression (Li (1991)) will be prone to take their intersection $\{0\}$ as sufficient, and hence lead us to the erroneous conclusion that $Y \perp\!\!\!\perp X$. Additionally, the ability to distinguish between an arbitrary DRS and the central DRS is critically important in both theory and application.

2.2 *Existence of the central subspace*

Since we intend to restrict ourselves to regressions for which the central DRS exists, we need to determine whether this class of regressions is large enough to be relevant in practice. The following propositions characterize the class by means of sufficient conditions (detailed proofs can be found in the technical appendix). The law of X is denoted by $\mathcal{L}_X$, and $Supp_X$ is its closed support (the intersection of all closed sets having probability 1 under the law).

PROPOSITION 2.1. *Assume that $Supp_X$ contains an open set Ω with $\mathcal{L}_X(\Omega) = 1$. Moreover, suppose that $Y \perp\!\!\!\perp X \mid E(Y \mid X)$, and that $E(Y \mid X)$ can be expressed as an analytic function of X, X-a.s. Then, the central DRS $S_{Y|X}$ for the regression of Y on X exists.*

The first condition concerns the distribution of the predictor, and is usually guaranteed when $\mathcal{L}_X$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$ (some absolutely continuous predictor distributions make exception to this rule, but they are so peculiar as not to represent a concern in practical applications).

Second and third conditions concern the conditional distribution $Y \mid X$. With $Y \perp\!\!\!\perp X \mid E(Y \mid X)$ we restrict ourselves to *location regressions*, in which Y depends on

X only through the conditional mean $E(Y \mid X)$. Moreover, we require the conditonal mean to be an analytic function of the predictor.

This set-up is met by virtually all standard regression models, including generalized linear models like logistic and Poisson regression. For instance, consider the additive-error model $Y = g(X) + \varepsilon$, with $\varepsilon \perp\!\!\!\perp X$, $E(\varepsilon) = 0$. In this type of model all dependence is conveyed by $E(Y \mid X) = g(X)$, and $g(X)$ is usually a relatively simple parametric function of the X coordinates (e.g. a polynomial). As a matter of fact, the requirements of Proposition 2.1 are looser in that they allow, for instance, heteroschedasticity as a function of the mean: if $Y = g(X) + \sigma(g(X))\varepsilon$, with $\varepsilon \perp\!\!\!\perp X$, $E(\varepsilon) = 0$, one still has $E(Y \mid X) = g(X)$ and $Y \perp\!\!\!\perp X \mid g(X)$.

However, we wish to provide conditions for the existence of the central DRS that do not constrain $Y \mid X$ in any fashion. This is achieved in the next proposition.

PROPOSITION 2.2. *Assume that $Supp_X$ contains an open and convex set Ω with $\mathcal{L}_X(\Omega) = 1$. Then the central DRS $S_{Y|X}$ exists for the regression of any response Y on X.*

Eliminating constraints on $Y \mid X$ is important because of the asymmetric roles played by the distribution of $Y \mid X$ and X in regression analysis. In fact, while $Y \mid X$ is the object of study, the distribution of X may be at least partially known, and may be controllable in some studies. The condition on X we pose in Proposition 2.2 always holds when $\mathcal{L}_X$ is absolutely continuous and $Supp_X$ is convex. So it holds, for example, for any predictor with an everywhere positive density on $\mathbb{R}^p$.

The above propositions generalize results first introduced by Cook ((1994*a*), (1996*a*), Lemmas 1 and 2). These results employed similar but tighter assumptions, referred to location regressions only, and were proved in coordinate-based and non measure theoretical terms. Moreover, although we use convexity in the statement of Proposition 2.2 because of its intuitive appeal, the result we prove in the Appendix is more general, as it relies on linked sections—a condition *weaker than convexity*. The possibility of guaranteeing existence through requirements other than convexity has been hinted at elsewhere in the literature (see for example Carrol and Li (1995)). Relaxations beyond linked sections may be possible, but no complete and rigorous argument for them has been developed at this time.

We consider the class of regressions identified by Propositions 2.1, 2.2, and the generalization of the latter to linked sections, to be wide enough to recover a large share of practical applications. Also, results similar to Propositions 2.1 and 2.2 can be given for discrete predictors. From now on we always assume the central DRS to exist.

2.3 *Some properties of the central subspace*

Our next step is to introduce some general properties of the central DRS, which are employed in developing the dimension reduction methods we describe in following sections. Coordinate versions of these properties are proved in Cook (1996*a*, 1998*a*).

First, it is easy to show that *full rank affine transformations of the predictor vector* do not affect sufficient dimension reduction. If $a \in \mathbb{R}^p$, and $A : \mathbb{R}^p \to \mathbb{R}^p$ is a full rank linear operator, then the central DRS obeys the transformation law

$$S_{Y|a+AX} = (A')^{-1} S_{Y|X} \tag{2.4}$$

where $(A')^{-1} S_{Y|X} = \{(A')^{-1}x, x \in S_{Y|X}\}$. In particular, the structural dimension does not change ($d_{Y|a+AX} = d_{Y|X}$), and the central views relative to X and $a + AX$ are

equivalent: A set of plotting coordinates for $P_{S_{Y|X}}X$ can be obtained from a set of plotting coordinates for $P_{S_{Y|a+AX}}(a+AX)$ by a translation and a linear transformation. As an instance, consider standardization of the predictor: If $\Sigma_X = \mathrm{Var}(X)$ is positive definite, taking $Z = \Sigma_X^{-1/2}(X - E(X))$ we have $S_{Y|Z} = \Sigma_X^{1/2} S_{Y|X}$.

Second, we consider *transformations of the response*. Transforming the response does not lead us outside the central DRS, and whenever the information picture is not altered (that is, for bijections), the central DRS is left unchanged. In symbols

$$S_{\varphi(Y)|X} \subseteq S_{Y|X}, \qquad \varphi \text{ bijection} \Rightarrow S_{\varphi(Y)|X} = S_{Y|X}. \tag{2.5}$$

An important corollary of this fact is that strictly monotone transformations of the response normally employed to improve the appearance of plots do not affect $S_{Y|X}$.

As another instance of useful response transforms, consider a binary version of the response obtained by setting $\tilde{Y} = 1$ if $Y > c$ and $\tilde{Y} = 0$ otherwise, for some constant c. (2.5) tells us that $S_{\tilde{Y}|X} \subseteq S_{Y|X}$, and thus that we may be able to reconstruct a portion of $S_{Y|X}$ by investigating $S_{\tilde{Y}|X}$. The advantage is that $S_{\tilde{Y}|X}$ can be investigated graphically when $p = 3$ using 3D binary response plots (Cook (1996*a*); (1998*a*), Chapter 5). Extensions of this idea are immediate: We can partition the range of Y into *slices*, L_s, $s = 1, \ldots, K$ and define the *sliced response*: $\tilde{Y} = s$ if $Y \in L_s$. Again, $S_{\tilde{Y}|X} \subseteq S_{Y|X}$. Sliced responses are used in non-graphical methods for inference on the central DRS, as described in Section 3.

Third, we consider *projections of the predictor vector*. It is easy to show that projecting X onto any DRS does not affect the central subspace. In symbols

$$Y \perp\!\!\!\perp X \mid P_S X \Rightarrow S_{Y|P_S X} = S_{Y|X}. \tag{2.6}$$

Passing from the original view $\{Y, X\}$ to the sufficient view $\{Y, P_S X\}$ does not affect the regression, and hence, a fortiori, does not affect the central DRS. This intuitive fact has far-reaching consequences, as it constitutes the basis for any rigorous formulation of *sequential dimension reduction procedures*. We will see an instance of its use in Subsection 4.2.4, which describes a novel iterative procedure for the estimation of $S_{Y|X}$.

Having addressed the issue of existence of the central DRS, and described a few of its properties, we now pass to non-graphical and graphical methods for making inference on this space based on data from (Y, X). Throughout the rest of the article, we work mostly in terms of standardized predictor Z. This involves no loss of generality because of (2.4), and will facilitate presentation. For use in practice, Z is constructed with Σ_X and $E(X)$ replaced by the usual estimates.

3. Non-graphical methods for dimension reduction

This Section provides a unified account of four non-graphical methods for inference on $S_{Y|Z}$. As we will see shortly, these methods allow us to recover linear portions of the central DRS, i.e. *lower bounds* to $S_{Y|Z}$, estimating directions within it. From our perspective, they are useful as pre-processors and aids to the graphical tools we describe in Section 4.

Suppose we have a consistent estimate $\hat{M}$ of a $p \times m$ population-level *kernel matrix* M, whose column span $S_M = \mathrm{Span}(M)$ is contained in the space of interest: $S_M \subseteq S_{Y|Z}$. Then, at least a linear portion of $S_{Y|Z}$ can be estimated based on $\hat{M}$. Let $\hat{u}_1, \ldots, \hat{u}_q$ denote the left singular vectors of $\hat{M}$ ordered according to the magnitude of its singular

values from largest to smallest, where $q = \min(p, m)$. Assuming that $k = \dim(S_M)$ is known ($k \leq q$)

$$\hat{S}_M = \text{Span}(\hat{u}_1, \ldots, \hat{u}_k)$$

is a consistent estimate of S_M. For use in practice, k will typically need to be replaced with an estimate $\hat{k}$ equal to the number of singular values which are inferred to be non-zero in the population. Many existing non-graphical methods can be fit under this umbrella, for proper specifications of the kernel matrix M, and conditions to guarantee $S_M \subseteq S_{Y|Z}$.

A first instance is *Ordinary Least Squares* (OLS). Under the assumption that

$$E(Z \mid P_{S_{Y|Z}} Z) = P_{S_{Y|Z}} Z, \qquad Z\text{-a.s.} \tag{3.1}$$

the $p \times 1$ kernel $M = E(YE(Z \mid Y)) = \text{Cov}(Z, Y)$ belongs to $S_{Y|Z}$ (see Li and Duan (1989) and Cook (1998*a*), Proposition 8.1). Consequently, if we believe the structural dimension to be $d_{Y|Z} = 1$, we can take $\hat{S}_{Y|Z} = \text{Span}(\hat{M})$: The 2D plot of Y versus $\hat{M}'Z$ is an estimated central view. Because of (2.4), this plot is equivalent to the plot of Y versus the fitted values from the OLS linear regression of Y on the non-standardized X. If the structural dimension of the regression is larger than 1, the OLS vector will still estimate a direction within the central subspace.

A second instance is *Sliced Inverse Regression* (SIR), as introduced by Li (1991). Again under assumption (3.1), the column span of the $p \times p$ kernel matrix $E[E(Z \mid Y)E(Z \mid Y)']$ lies within $S_{Y|Z}$. SIR thus employs the "approximate" kernel $M = E[E(Z \mid \tilde{Y})E(Z \mid \tilde{Y})']$, where $\tilde{Y}$ is a sliced response as described previously. Inference methods for the dimension of S_M are discussed in Li's original paper, and extended by Cook ((1998*a*), Chapter 11).

A third instance is *Sliced Average Variance Estimation* (SAVE). Under (3.1), and the further second moment assumption

$$\text{Var}(Z \mid P_{S_{Y|Z}} Z) = Q_{S_{Y|Z}}, \qquad Z\text{-a.s.} \tag{3.2}$$

the column span of the $p \times p$ kernel matrix $E(I - \text{Var}(Z \mid Y))$ lies within $S_{Y|Z}$. Again, a sliced version is used: $M = E(I - \text{Var}(Z \mid \tilde{Y}))$. SAVE was proposed by Cook and Weisberg (1991), and developed further by Cook and Lee (1999) who discuss inference methods for $\dim(S_M)$.

Last, *Principal Hessian Directions* (pHd), proposed by Li (1992) and extended by Cook (1998*b*), refers to the $p \times p$ kernel matrix

$$\begin{aligned} M &= E[(Y - E(Y))ZZ'] \\ &= E[(Y - E(Y))[E(Z \mid Y)E(Z \mid Y)' - (I - \text{Var}(Z \mid Y))]]. \end{aligned}$$

Under (3.1) and (3.2), also this matrix has a column span within $S_{Y|Z}$.

Cook and Lee (1999) demonstrated the common nature of these methods proving that under (3.1) and (3.2), *any linear combination and/or (weighted) averaging of* $E(Z \mid Y)$, $E(Z \mid Y)E(Z \mid Y)'$ *and* $(I - \text{Var}(Z \mid Y))$ *is guaranteed to span a subspace of* $S_{Y|Z}$.

Condition (3.1) is equivalent to requiring that $E(Z \mid P_{S_{Y|Z}} Z)$ be linear (Cook (1998*a*), p. 57), while (3.2) is equivalent to requiring that $\text{Var}(Z \mid P_{S_{Y|Z}} Z)$ be constant. Both conditions involve Z and $Y \mid Z$ through $S_{Y|Z}$. However, they can be guaranteed

through stronger requirements on Z alone. For example, they hold when the predictor is normally distributed (for further discussion see Cook ((1998*a*), Subsection 8.3).

While it has been demonstrated that each of these methods can perform well in practice, they are all potentially fallible in the right situations, depending on the appropriateness of (3.1) and (3.2), the accuracy of the large sample methods for inference on $\dim(S_M)$, and their intrinsic operating characteristics. However, from our perspective, we are less concerned about the performance of these methods per se. In fact, we use them as pre-processors to *construct linear combinations of Z that are ordered based on their likely relation to the central* DRS. These new predictors are then used in concert with the graphical methods developed in the next section.

For future reference, we will denote by $\{a_1, \ldots, a_p\}$ the left singular vectors of the estimated SAVE kernel, and call $a_j'Z$, $j = 1, \ldots, p$, the *SAVE predictors*. Similarly, $\{h_1, \ldots, h_p\}$ will denote the left singular vectors of the estimated pHd kernel, and $h_j'Z$, $j = 1, \ldots, p$ will be called the *pHd predictors*.

4. Graphical methods for dimension reduction

This section is devoted to graphical methods for inference on $S_{Y|Z}$. At the core of these methods, is the possibility of using low-dimensional *supporting views* as tools to produce a DRS S; that is, a sufficient view $\{Y, P_S Z\}$ for the regression of Y on Z. Our aim is then to employ supporting views iteratively, generating a sequence of nested DRS's, This will allows us to "approach" the central DRS from above. Moreover, we will use non-graphical methods to aid the choice of supporting views at each stage.

We introduce results concerning three main types of supporting views; namely, *conditional*, *marginal* and *residual* views. We discuss conditions under which they produce DRS's (effectiveness), details of their iterative application, and conditions that can improve their performance (efficiency). The bank note data are used as running example.

In the following, the symbol $\oplus$ indicates the sum of subspaces, implemented as $T \oplus S = \{t + s, t \in T, s \in S\}$ (note the operation does not require $T \cap S = \{0\}$). A simple $+$ is used when adding matrices corresponding to linear operators (e.g. with orthogonal projections, $P_T + Q_S$).

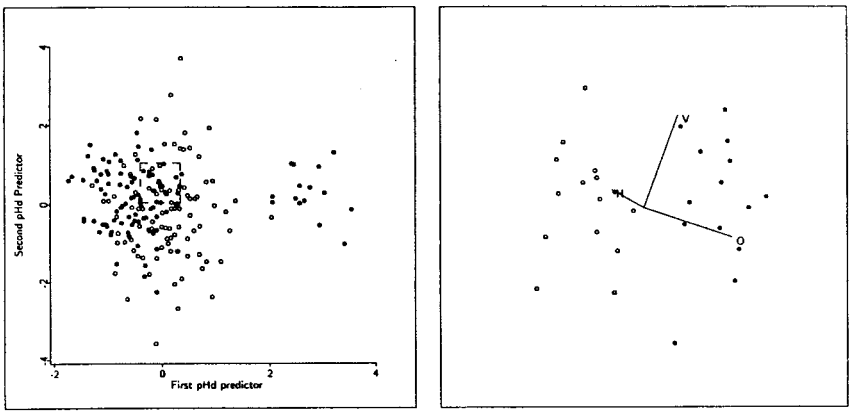
4.1 *Using conditional views to assess sufficiency of a view*

One approach to producing a DRS is to select a candidate subspace, and assess whether it is sufficient for the regression of Y on Z. Through the use of *conditional views*, we can turn this problem into assessing independence for a collection of regressions with a low-dimensional predictor vector. The theoretical basis is provided by the following:

PROPOSITION 4.1. *Let S and T be subspaces of $\mathbb{R}^p$ such that $S \oplus T = \mathbb{R}^p$. Then, S is a DRS for the regression of Y on Z if and only if $Y \perp\!\!\!\perp P_T Z \mid P_S Z$.*

Thus, S is a DRS if and only if the regressions of Y on $P_T Z$ defined for each conditioning value of $P_S Z$, all have 0D structure. These regressions are captured by the conditional views $\{Y, P_T Z \mid P_S Z\}$.

In practice, one considers a candidate with $\dim(S) = p - 1, p - 2$ or possibly $p - 3$, and takes T with $\dim(T) = 1, 2$ or 3 as to complement it with $\{0\}$ intersection. Conditioning is approximated by "slicing" on $P_S Z$, and considering a finite collection of "intra-slice" regressions of Y on the low-dimensional predictor $P_T Z$. Within each of



a. Y vs. $(h_1'Z, h_2'Z)$ b. Y vs. $(H, V, O) = (h_4'Z, h_5'Z, h_6'Z)$

Fig. 2. Two views from a pHd analysis of the Swiss bank note data. Plot (b) shows only those points in the selection rectangle of plot (a).

the low-dimensional "intra-slice" views, 0D structure can be easily assessed by visual inspection, as described in Cook (1994*a*, 1996*a*, 1998*a*) and Cook and Weisberg (1994).

Proposition 4.1 guarantees effectiveness of conditional views, posing no assumptions on either Z or $Y \mid Z$. The price for this is two-fold: First, one has to analyze a whole collection of low-dimensional views. Second, those low-dimensional views are constructed by "slicing" on a possibly high-dimensional S. If $\dim(S) = p - \dim(T)$ is larger than 3, serious practical complications due to sparseness of the data can result.

One possible choice for the complementary space is the orthogonal complement to S, $T = S^{\perp}$. The corresponding conditional views $\{Y, P_T Z \mid P_S Z\}$ are called the *uncorrelated views* for S, since $\mathrm{Cov}(P_T Z, P_S Z) = 0$. In this respect, Proposition 4.1 generalizes the notion of uncorrelated views used by Cook and Weisberg (1994).

Returning to the bank note data, a candidate S might be chosen by using any of the methods mentioned in Section 3, or from a visual tour using, for example, XGobi (Swayne *et al.* (1998)). We selected the candidate $S = \mathrm{Span}(h_1, h_2)$ based on pHd. The corresponding binary response plot is shown in Fig. 2(a). The issue now is whether there is information in the data to contradict the sufficiency of this view. According to Proposition 4.1, we can check this possibility by studying the conditional views $\{Y, P_T Z \mid P_S Z\}$ for a suitable choice of the subspace T. We selected $T = S^{\perp}$, which is the same as the 4-dimensional subspace corresponding to the remaining four pHd predictors, $\mathrm{Span}(h_3, h_4, h_5, h_6)$. Figure 2(b) shows a projection of the 4D uncorrelated view containing the points in the 2D slice shown in Fig. 2(a). Since Y is clearly dependent on $P_T Z$ in this view, we can conclude that $\{Y, P_S Z\}$ is *not* sufficient.

4.2 *Using a marginal view to identify a sufficient view*

When $p \leq 2$, (or 3 when the response is binary or perhaps trinary) the central subspace $S_{Y|Z}$ can be estimated from 2D or 3D plots along the lines described in Cook (1994*a*, 1996*a*, 1998*a*, Chapters. 4–5) and Cook and Weisberg ((1994), Chapters. 6–8). The methods for these low-dimensional settings are straightforward and free of assumptions, because of our ability to view $\{Y, Z\}$ fully. In the following, we use them

on low-dimensional marginal views to gain information on $S_{Y|Z}$ when $p > 2$.

The approach we describe here is an alternative to the use of conditional views. Through the use of a *marginal view*, we can turn the problem of identifying a DRS for the regression of Y on Z into that of identifying the central DRS for a regression with a low-dimensional predictor vector. The idea is to select a low-dimensional subspace T, which plays the role of a sort of "reduction window". Then, we find the central subspace $S_{Y|P_TZ}$ for the regression of Y on P_TZ, applying the methods referenced above to a coordinate version of the marginal view $\{Y, P_TZ\}$. Last, we piece together $S_{Y|P_TZ}$ and the complement of T. Under a crucial assumption we discuss below, this piecing together produces a DRS for Y on the whole Z.

For dimension reduction within a low-dimensional marginal view $\{Y, P_TZ\}$ to be useful in the identification of a DRS for Y on Z, and ultimately in the pursuit of the overall central subspace, we must clarify the relation between $S_{Y|P_TZ}$, the subspace that we can estimate straightforwardly, and $S_{Y|Z}$ itself. The next proposition is our first step in this direction.

PROPOSITION 4.2. *For any subspace* $T \subseteq \mathbb{R}^p$, $S_{Y|P_TZ} \subseteq P_TS_{Y|Z} \oplus S_{P_VZ|P_TZ}$ *where, for notational convenience,* V *is the projection of* $S_{Y|Z}$ *onto* $T^\perp$; *i.e.* $V = Q_TS_{Y|Z}$.

This proposition expresses a fully general property of marginal views. The inclusion means that $P_TS_{Y|Z} \oplus S_{P_VZ|P_TZ}$ is a DRS for the regression of Y on P_TZ. The *marginal central* DRS $S_{Y|P_TZ}$ can depend on $Y \mid Z$ via the *coordinate subspace* $P_TS_{Y|Z}$, and on Z via $S_{P_VZ|P_TZ}$, the central DRS for the regression of P_VZ on P_TZ, which we call the *predictor subspace*. What we would like, is to use the marginal central DRS $S_{Y|P_TZ}$ to infer about the coordinate subspace $P_TS_{Y|Z}$, which would then tell us something useful about the overall central subspace $S_{Y|Z}$.

Before proceeding, we illustrate the meaning of Proposition 4.2 through a small example: Consider the typical linear regression $Y \sim \alpha_0 + \alpha_1'Z + \varepsilon$, $\varepsilon \perp\!\!\!\perp Z$, $E(\varepsilon) = 0$. Following (2.1), the central DRS is $S_{Y|Z} = \text{Span}(\alpha_1)$. Within the marginal view $\{Y, P_TZ\}$ the mean function is

$$E(Y \mid P_TZ) = \alpha_0 + (P_T\alpha_1)'Z + E[(Q_T\alpha_1)'Z \mid P_TZ]. \tag{4.1}$$

We see clearly how the coordinate subspace $P_TS_{Y|Z} = P_T\text{Span}(\alpha_1)$ captures the term $(P_T\alpha_1)'Z$. The term $E[(Q_T\alpha_1)'Z \mid P_TZ]$, on the other hand, is captured by the predictor subspace $S_{P_VZ|P_TZ}$. In fact, here $V = Q_TS_{Y|Z} = Q_T\text{Span}(\alpha_1)$.

4.2.1 *Marginal consistency*

Proposition 4.2 allows for the possibility that the marginal central DRS $S_{Y|P_TZ}$ is a proper subset of $P_TS_{Y|Z} \oplus S_{P_VZ|P_TZ}$. In particular, it allows for the possibility that the marginal central DRS does not contain the coordinate subspace: $P_TS_{Y|Z} \not\subseteq S_{Y|P_TZ}$. In this case, an investigation of the marginal central DRS in $\{Y, P_TZ\}$ will miss part of the coordinate subspace. On the other hand, if one has equality in Proposition 4.2 $S_{Y|P_TZ} = P_TS_{Y|Z} \oplus S_{P_VZ|P_TZ}$ or more generally if

$$S_{Y|P_TZ} \supseteq P_TS_{Y|Z} \tag{4.2}$$

the marginal view $\{Y, P_TZ\}$ contains *all* the information that is relevant to the coordinate subspace. We call this the *marginal consistency assumption*.

Marginal consistency is the key condition for effectiveness of a marginal view. In fact, under (4.2), the space obtained piecing together the marginal DRS and the complement of T contains the overall central subspace

$$S = S_{Y|P_T Z} \oplus T^{\perp} \supseteq P_T S_{Y|Z} \oplus T^{\perp} \supseteq S_{Y|Z}$$

and is therefore a DRS for the regression of Y on Z. As a consequence, we can use the marginal view $\{Y, P_T Z\}$ to identify the central DRS for Y on $P_T Z$, and construct a sufficient view for Y on Z as

$$\{Y, (P_{S_{Y|P_T Z}} + Q_T)Z\}. \tag{4.3}$$

Marginal consistency is a fairly weak condition, likely to hold in most cases of practical interest. For example, in the typical linear regression example introduced above, $P_T S_{Y|Z} = P_T \text{Span}(\alpha_1) \not\subseteq S_{Y|P_T Z}$ cannot occur, because of the term $(P_T \alpha_1)' Z = (P_T \alpha_1)'(P_T Z)$ in the mean function (4.1). Whenever this term is not 0, $P_T \text{Span}(\alpha_1) \subseteq S_{Y|P_T Z}$. Moreover, the term will be 0 if and only if the subspace T is orthogonal to $S_{Y|Z}$, i.e. $P_T \alpha_1 = 0$, in which case $P_T \text{Span}(\alpha_1) = \{0\} \subseteq S_{Y|P_T Z}$ holds trivially—note that by using a standardized predictor vector we are assuming its non-singularity, but if Z were singular and T orthogonal to its linear support, i.e. $P_T Z = 0$, $Z \in T^{\perp}$ would imply $S_{Y|Z} \subseteq T^{\perp}$, and therefore again $P_T \alpha_1 = 0$.

In general, $P_T S_{Y|Z} = P_T \text{Span}(\alpha_1) \not\subseteq S_{Y|P_T Z}$ would imply existence of a direction η that is relevant to the regression of Y on Z, is not annihilated by the marginalization ($P_T \eta \neq 0$), and yet becomes irrelevant when considering the regression of Y on $P_T Z$. Loss of relevant directions by marginalization is not impossible, but requires peculiar combinations of regression structure and predictor distribution. The regression of Y on $P_{S_{Y|Z}} Z$ is irreducible by construction, in the sense that (see (2.6)) $S_{Y|P_{S_{Y|Z}} Z} = S_{Y|Z}$. Instances in which (4.2) fails can be constructed if this irreducible, and often very low-dimensional, "core" of the original regression admits some reducible marginals. For example, suppose there were a subspace $W \subseteq S_{Y|Z}$ such that $Y \perp\!\!\!\perp P_W Z$. Because of irreducibility, Y and $P_W Z$ cannot be *conditionally independent* given $P_{W^{\perp_{S_{Y|Z}}}} Z$; otherwise $Y \perp\!\!\!\perp P_{S_{Y|Z}} Z \mid P_{W^{\perp_{S_{Y|Z}}}} Z$ and $W^{\perp_{S_{Y|Z}}} \subseteq S_{Y|Z}$ would be a dimension reduction subspace for the regression because of (2.6). But at the same time, Y and $P_W Z$ are *marginally independent*. For any subspace contained in this "special" marginal independence region within the central space of the regression, $T \subseteq W \subseteq S_{Y|Z}$, we would then have

$$Y \perp\!\!\!\perp P_W Z \Leftrightarrow S_{Y|P_T Z} = \{0\}, \qquad P_T S_{Y|Z} = T.$$

This type of situation is seldom in applications, especially since it ought to pertain not to the original regression, with all its potential redundancies, but to its irreducible "core". Moreover, coexistence of marginal independence and conditional dependence, when it occurs, is often conveyed by dependencies among the predictors themselves. Thus, working in terms of the standardized Z, which eliminates linear dependencies among predictors, reduces the chances of (4.2) failing—advantages of standardization will be discussed again relative to efficiency of marginal views in Subsection 4.2.6.

Assuming (4.2), we now illustrate how the above results can be exploited in practice through an iterative strategy. Non-graphical methods are employed to aid the choice of marginal view at each stage.

4.2.2 *Bank note data using iterated marginal views based on SAVE*

Our estimated central view $\{Y, (b_1'X, b_2'X)\}$ in Fig. 1(a) contains the first two SAVE predictors: $b_1'X = a_1'Z$ and $b_2'X = a_2'Z$, apart from additive constants that can be neglected because of the affine invariance in (2.4). We next describe how we reached this as conclusion, assuming marginal consistency (4.2).

We used the last three SAVE predictors to form $T = \text{Span}(a_4, a_5, a_6)$, our first "reduction window", and thus $\{Y, P_T Z\}$, our first marginal view. A corresponding coordinate version is given by the 3D binary response plot of Y versus $(a_4'Z, a_5'Z, a_6'Z)$. We found no visual evidence of dependence, because the relative density of authentic notes appeared uniform throughout the plot (see Cook (1996*a*) for more details on the interpretation of binary response plots). Thus, we inferred that $Y \perp\!\!\!\perp P_T Z$, i.e. that $\hat{S}_{Y|P_T Z} = \{0\}$. Using (4.2), this allowed us to take

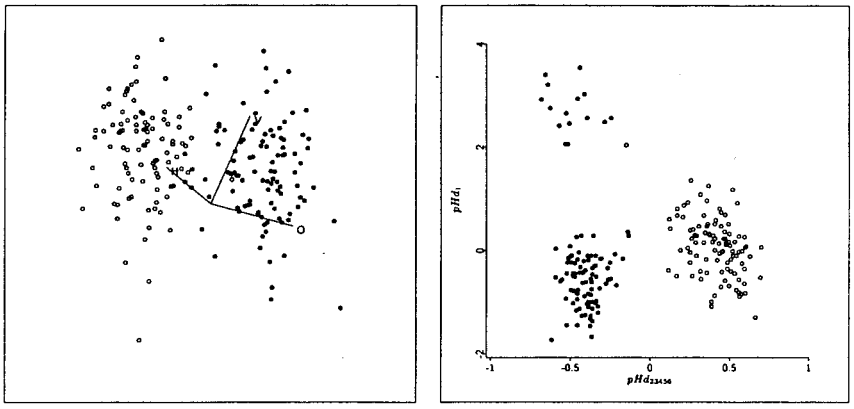
$$S_1 = \hat{S}_{Y|P_T Z} \oplus T^{\perp} = \{0\} \oplus T^{\perp} = \text{Span}(a_1, a_2, a_3)$$

as a DRS for the regression of Y on Z.

Because of (2.6), we know that the search for the central DRS can be restricted to any DRS. Thus, we passed to the regression of Y on $P_{S_1} Z$. This can be visualized directly in the 3D binary response plot of Y versus the first three SAVE predictors $(a_1'Z, a_2'Z, a_3'Z)$; we inferred 2D structure with $S_2 = \hat{S}_{Y|P_{S_1} Z} = \text{Span}(a_1, a_2)$ which constitutes a second DRS for the overall regression, $S_2 \subseteq S_1$. The regression of Y on $P_{S_2} Z$ could not be reduced any further, so we set $\hat{S}_{Y|Z} = \hat{S}_{Y|P_{S_2} Z} = \text{Span}(a_1, a_2)$ itself. This resulted in the estimated central view in Fig. 1(a).

4.2.3 *Bank note data using iterated marginal views based on pHd*

We now turn to a similar analysis based on pHd. Taking the last three pHd predictors to form the first "reduction window" $T = \text{Span}(h_4, h_5, h_6)$, we found clear evidence of marginal dependence in the corresponding 3D binary response plot of Y on $(h_4'Z, h_5'Z, h_6'Z)$. Thus we inferred that $\text{Span}(h_1, h_2, h_3)$ is not sufficient. Here, assuming that (4.2) holds in Z, we were able to go a step further by estimating $S_{Y|P_T Z}$ and hence constructing a DRS through (4.3). Using again methods described in Cook (1996*a*), we



a. Y vs. $(H, V, O) = (h_4'Z, h_5'Z, h_6'Z)$ b. Y vs. $(h_{23456}'Z, h_1'Z)$

Fig. 3. Two views from a pHd analysis of the Swiss bank note data under condition (4.2); $h_{456}'Z$ is on the horizontal screen axis of (a).

inferred that $\dim(S_{Y|P_T Z}) = 1$, and replaced (h_4, h_5, h_6) with a single linear combination of them, h_{456}: $\hat{S}_{Y|P_T Z} = \text{Span}(h_{456})$. Figure 3(a) shows a projection of the 3D binary response plot of Y versus $(h_4'Z, h_5'Z, h_6'Z)$ with $h_{456}'Z$ on the screen horizontal axis. This allowed us to take

$$S_1 = \hat{S}_{Y|P_T Z} \oplus T^{\perp} = \text{Span}(h_{456}, h_1, h_2, h_3)$$

as a DRS for the regression of Y on Z.

Because of (2.6), we next considered the regression of Y on $P_{S_1} Z$. Since this can not yet be visualized directly, we iterated the above reasoning *within* S_1. We formed our second "reduction window" as $T = \text{Span}(h_{456}, h_2, h_3)$. The corresponding 3D binary response plot of Y on $(h_4'Z, h_5'Z, h_6'Z)$ presented again 1D structure, with $\hat{S}_{Y|P_T Z} = \text{Span}(h_{23456})$. Hence, we took

$$S_2 = \hat{S}_{Y|P_T Z} \oplus T^{\perp S_1} = \text{Span}(h_1, h_{23456})$$

as a new DRS for Y on Z, contained in the one identified at the first stage; $S_2 \subseteq S_1$. The resulting regression, Y on $P_{S_2} Z$, can be visualized directly in the 2D binary response plot of Y versus $(h_1'Z, h_{23456}'Z)$. Since this could not be reduced any further, we set

$$\hat{S}_{Y|Z} = \hat{S}_{Y|P_{S_2} Z} = \text{Span}(h_1, h_{23456})$$

itself. This resulted in the estimated central view in Fig. 3(b) which, aside from orientation, is nearly identical to SAVE's estimated central view in Fig. 1(a).

4.2.4 *Iterating*

In Subsections 4.2.2 and 4.2.3 we saw how estimation of the central DRS can be approached iteratively by estimating a sequence of nested DRS's for the regression on Y on Z. At the population level, this is justified by (2.6), which indicates that the search for $S_{Y|Z}$ can be performed within any sufficient view $\{Y, P_S Z\}$. From a practical standpoint, it is interesting because, even when operating in large dimension, DRS's can be identified using low-dimensional marginal views under the sole marginal consistency assumption (4.2).

We produce a first DRS S_1 for Y on Z chosing a subspace $T \subseteq \mathbb{R}^p$, and using the view $\{Y, P_T Z\}$. We then restrict ourselves to the regression of Y on $P_{S_1} Z$, and produce a DRS S_2 for the latter chosing a subspace $T \subseteq S_1$, and using the view $\{Y, P_T Z\}$. Iterating the procedure we generate a sequence of nested DRS's $S_1 \supseteq S_2 \cdots \supseteq S_t$ descending towards the central DRS. Iteration is continued until an irreducible regression is reached. In fact, if Y on $P_{S_t} Z$ does not admit a DRS strictly contained in S_t, then $S_{Y|Z} = S_t$ by construction.

At any iteration stage, the subspace T corresponds to the "reduction window", i.e. the linear region of the current DRS within which we attempt a further reduction. Although the T's could in principle be taken in any fashion, e.g. at random, we customarily use the ranked directions produced by a non-graphical method, as SAVE or pHd. In other words, we attempt reduction in linear regions spanned by low-ranking directions, as those are the ones less likely to be relevant to the regression (i.e. to contain a portion of $S_{Y|Z}$). The ways in which directions produced by non-graphical methods can be used to improve the performance of an iterative graphical investigation will be further discussed in Subsection 4.4.

In practice, if $\dim(S_t) \leq 2$ (or 3 when the response is binary) we can positively asses irreducibility because we can visualize $\{Y, P_{S_t} Z\}$ directly. This is what happened

in the bank note example: we were able to proceed all the way down to a 2-dimensional subspace, at which point irreducibility could be checked by direct visualization.

On the other hand, suppose at stage t of the iteration, with $\dim(S_t)$ still too large for direct visualization, we found $S_{Y|P_T Z} = T$. In this situation, we are stuck with $T \oplus T^{\perp S_t} = S_t$ itself. We do know that our current space, being a DRS, contains the central DRS; $S_{Y|Z} \subseteq S_t$. However, $S_{Y|P_T Z} = T$ is not enough to conclude that $S_{Y|P_{S_t} Z}$, and therefore that $S_{Y|Z}$ *coincides* with S_t. Whatever approach we used to select T (even if we used the ranked directions of a non-graphical method), we might just have gotten the wrong "reduction window"; that is, a viewpoint from which we could not see that the central DRS is smaller than the current S_t. In such a situation, if indeed $S_{Y|Z} \subset S_t$, attempting reduction with several T's (i.e. considering several viewpoints) could lead us to by-pass the obstacle. Conversely, finding $S_{Y|P_T Z} = T$ for several T's in S_t enforces our confidence that $S_{Y|Z} = S_t$, but does not yet guarantee equality.

Another important issue here is the dimensionality of our "reduction window(s)". Ceteris paribus, the larger $\dim(T)$, the lower the chances that a strict containment $S_{Y|Z} \subset S_t$ will not show when we view the data from T. This motivates taking $\dim(T)$ as large as practically possible with the available software. A second rationale is that the variability along the response axis in $\{Y, P_T Z\}$ tends to be larger the smaller $\dim(T)$. As an illustration, consider two nested subspaces $\tilde{T} \subseteq T$. Reasoning in the familiar terms of variability partitioning

$$\begin{aligned} E[\mathrm{Var}(Y \mid P_T Z)] &= \mathrm{Var}(Y) - \mathrm{Var}[E(Y \mid P_T Z)] \\ &\leq \mathrm{Var}(Y) - \mathrm{Var}[E(Y \mid P_{\tilde{T}} Z)] = E[\mathrm{Var}(Y \mid P_{\tilde{T}} Z)]. \end{aligned}$$

Hence, $\dim(T)$ should be large to maximize visual accuracy in detecting dependencies.

In his development of pHd, Li ((1992), Section 6) suggested that under certain conditions it may be desirable to inspect all possible $p(p-1)/2$ coordinate views of the form $\{Y, (h_i' Z, h_j' Z)\}$ in an effort to identify the linear combinations driving the regression. The procedure proposed here is quite different, requiring fewer assumptions and fewer plots. As the bank note example illustrates, we may be able to reach the estimated central view by inspecting only $p/2$ plots, which is considerably less than the number suggested by Li. Moreover, the possibility of premature termination can be eliminated in many circumstances. We will come back to this issue several times in the reminder of the paper.

4.2.5 *Assessing independence between response and predictors: an application of the iterative procedure to residual analysis*

In the procedure described in Subsection 4.2.4, iteration proceeds all the way down to $S_{Y|Z}$ when its dimension is $d_{Y|Z} = 0$. In other words, our iterative procedure always "recognizes" situations in which $Y \perp\!\!\!\perp Z$. Formally, this means that reaching $S_t = \{0\}$ is a necessary and sufficient condition for declaring $S_{Y|Z} = \{0\}$. Sufficiency is straightforward: $S_t = \{0\}$ implies $S_{Y|Z} = \{0\}$ because we traveling along DRS's guarantees $S_t \supseteq S_{Y|Z}$. Necessity is the crucial point here, as it means that we will indeed travel all the way down to $\{0\}$ without incurring in premature termination. This is guaranteed by the fact that if $Y \perp\!\!\!\perp Z$, then $Y \perp\!\!\!\perp P_T Z$ for any subspace T; that is, if $S_{Y|Z} = \{0\}$, we will find $S_{Y|P_T Z} = \{0\}$ for any choice of "reduction window" T along the way, and thus necessarily keep reducing our current DRS all the way down to $S_t = \{0\}$.

A lengthier argument can be given here, that emphasises the roles of coordinate and predictor spaces introduced near Proposition 4.2 (this type of reasoning will be

used when discussing premature termination in a later section). Assume again that $S_{Y|Z} = \{0\}$. Then, for any choice of T, the coordinate space is $P_T S_{Y|Z} = \{0\}$. Moreover $V = Q_T S_{Y|Z} = \{0\}$, and therefore the predictor space is $S_{P_{\{0\}}Z|P_T Z} = S_{0|P_T Z} = \{0\}$. Now $S_{Y|P_T Z} = \{0\}$ follows immediately from Proposition 4.2. In other words, when $S_{Y|Z} = \{0\}$, regardless of the selected "reduction window" T, we have that (*a*) *the coordinate subspace* $P_T S_{Y|Z}$ *cannot be misleading on the dimension of* $S_{Y|Z}$, *and* (*b*) *the predictor subspace cannot affect* $\{Y, P_T Z\}$ *by "inflating"* $S_{Y|P_T Z}$ *with respect to* $P_T S_{Y|Z}$.

This capability of reliably recognizing independence allows us to use the iterative procedure in residual analysis. Suppose we have a model for Y on Z, with the usual property that the model is correct if and only if its population residual r is $r \perp\!\!\!\perp Z$. Diagnostics for model adequacy is often performed in practice by plotting sample residuals versus individual predictor variables or fitted values, in effect checking whether $r \perp\!\!\!\perp b'Z$ (i.e. $S_{r|b'Z} = \{0\}$) for a few selected values of the vector b, which is *not* enough to conclude $r \perp\!\!\!\perp Z$ (i.e. $S_{r|Z} = \{0\}$).

On the other hand, assuming marginal consistency ((4.2) with r in place of Y) for the regression of r on Z, the iterative procedure provides a necessary and sufficient condition for the model to be correct: $S_{r|Z} = \{0\}$ if and only if we reach $S_t = \{0\}$. The possibility of developing a graphical procedure for exhaustive model checking was discussed by Cook and Wetzel (1993) and Cook (1994*a*).

4.2.6 *Refinements: efficient marginal views and improved iteration*

In this section we consider various situations that can improve the performance of a graphical analysis, and we explain one reason why working in the Z scale is desirable.

Whenever marginal consistency (4.2) holds, a marginal view $\{Y, P_T Z\}$ is effective, because it contains *all* the information relative to the coordinate subspace $P_T S_{Y|Z}$. This is what allows us to take $S = S_{Y|P_T Z} \oplus T^{\perp}$ as a DRS for Y on Z. However, the marginal view can be affected by the distribution of Z, as well as by $Y \mid Z$. Following Proposition 4.2, even under (4.2), the marginal central DRS $S_{Y|P_T Z}$ can be anywhere between $P_T S_{Y|Z}$ and $P_T S_{Y|Z} \oplus S_{P_V Z|P_T Z}$. A large predictor space allows $S_{Y|P_T Z}$ to exceed $P_T S_{Y|Z}$ substantially, which in turn makes the marginal view unefficient, in the sense that the DRS it produces might be larger than it needs to be. At the extreme, we can have $S_{Y|P_T Z} = T$ even if $P_T S_{Y|Z} = \{0\}$. It is therefore desirable to limit the potential predictor contribution to the marginal view, by reducing the dimension of $S_{P_V Z|P_T Z}$.

Since there are no linear dependencies within the standardized predictor Z, we would expect the dimension of a predictor subspace in Z to be less than the dimension of the corresponding subspace in X. Thus, working in terms of Z will usually facilitate the analysis without loss of generality or introduction of constraints.

Ideally, we would like $S_{P_V Z|P_T Z} \subseteq P_T S_{Y|Z}$. In this case $S_{Y|P_T Z} \subseteq P_T S_{Y|X}$; that is, the marginal view $\{Y, P_T Z\}$ contains *only* information that is relevant to the coordinate subspace $P_T S_{Y|X}$. This, combined with marginal consistency (4.2), gives an *efficient* marginal view defined as one in which $S_{Y|P_T Z} = P_T S_{Y|X}$. When using an efficient marginal view, one achieves the largest dimension reduction allowed by the chosen "reduction window" T. Generalizing Lemma 4.1 in Cook (1994*a*), the desired inclusion is achieved when $C = P_T S_{Y|Z}$ is a DRS for the regression of $P_V Z$ on $P_T Z$. In symbols

$$P_V Z \perp\!\!\!\perp P_T Z \mid P_C Z \Rightarrow S_{Y|P_T Z} \subseteq P_T S_{Y|Z}.$$

This condition involves both Z and the conditional distribution of $Y \mid Z$ via the central subspace $S_{Y|Z}$. We can eliminate the involvement of $Y \mid Z$ by adding to the constraints

on Z:

$$Q_T Z \perp\!\!\!\perp P_T Z \Rightarrow S_{Y|P_T Z} \subseteq P_T S_{Y|Z}. \tag{4.4}$$

Here $Q_T Z \perp\!\!\!\perp P_T Z$ ensures that $S_{P_V Z|P_T Z} = \{0\}$, regardless of $S_{Y|Z}$ and $V = Q_T S_{Y|Z}$. In particular, if Z is normal, $Q_T Z \perp\!\!\!\perp P_T Z$ holds for any choice of subspace T. It is important to stress that predictor normality is not related to effectiveness of marginal views, which relies on (4.2) alone. The bank note predictors are clearly not normal, for example. Normality guarantees efficiency, and is therefore a desirable but not a necessary condition for our graphical analysis.

In terms of the iterative procedure, *efficiency accelerates the descent towards* $S_{Y|Z}$. Moreover, it allows us to make progress on premature termination: *If the marginal views employed at each stage are efficient and based on "reduction windows" that exceed the dimension of the central subspace, then iteration is guaranteed to proceed all the way down to* $S_{Y|Z}$. The argument for this is as follows: (a) If $\dim(T) > d_{Y|Z}$ the coordinate subspace cannot be misleading about the dimension of $S_{Y|Z}$, because $P_T S_{Y|Z} \subset T$. (b) Efficiency removes the predictor space from the picture, ensuring $S_{Y|P_T Z} = P_T S_{Y|Z}$. Thus, $S_{Y|P_T Z} \subset T$. For example, employing marginal views with $\dim(T) = 2$ we will reach $S_{Y|Z}$ when $d_{Y|Z} = 0$ or 1. And when the response is binary, with $\dim(T) = 3$ we will reach the central DRS when its dimension is $0, 1$ or even 2.

This discussion has another important consequence. Suppose we got stuck at S_t with $\dim(S_t)$ still too large for direct visualization. While it could still be that $S_{Y|Z} \subset S_t$ strictly, efficient views would allow us to positively conclude that $d_{Y|Z} \geq \dim(T)$.

4.3 *Using residual views instead of marginal views*

In this section we introduce the use of various residuals as another means of refining our graphical analysis. Consider again a low-dimensional subspace T, define the predictor residual

$$r_{1|2} = r_{P_T Z|Q_T Z} = P_T Z - E(P_T Z \mid Q_T Z) \in T$$

and let $\omega(Y, Q_T Z)$ be a function of Y and $Q_T Z$, which is to play the role of a "working response". This could be a second residual from a fit of Y on $Q_T Z$. We now introduce the analogue of marginal consistency in this setting. Suppose we had (*residual consistency assumption*)

$$S_{\omega|r_{1|2}} \supseteq P_T S_{Y|Z} \tag{4.5}$$

then $S = S_{\omega|r_{1|2}} \oplus T^{\perp}$ would provide a DRS for the regression of Y on Z in the same way that $S_{Y|P_T Z} \oplus T^{\perp}$ provided a DRS following (4.2). Consequently, we could use the *residual view* $\{\omega(Y, Q_T Z), r_{1|2}\}$ to identify the central DRS for ω on $r_{1|2}$, and then construct a sufficient view for Y on Z as

$$\{Y, (P_{S_{\omega|r_{1|2}}} + Q_T)Z\}.$$

We have again turned the problem into identifying the central DRS for a regression with a low-dimensional predictor vector. The effectiveness of a residual view is guaranteed by condition (4.5), which is similar to condition (4.2), and can be interpreted along the same lines. The difference is that in (4.2) we define the marginal central DRS through

the regression of Y on $P_T Z$, while in (4.5) we define it through the regression of ω on the residual $r_{1|2}$. Under (4.5), residual views can be used in the same way as marginal views were used in Subsections 4.2.2–4.2.5.

Like marginal views, residual views can be inefficient, albeit effective, under their consistency assumption. But there are situations in which we can guarantee the inclusion $S_{\omega|r_{1|2}} \subseteq P_T S_{Y|Z}$, and therefore equality under (4.5):

PROPOSITION 4.3. *Let $T \subseteq \mathbb{R}^p$. If $P_T S_{Y|Z}$ is a DRS for the regression of $Q_T Z$ on $r_{1|2}$, then $S_{\omega|r_{1|2}} \subseteq P_T S_{Y|Z}$.*

As for marginal views, we can add to the constraints on Z to avoid involving $Y \mid Z$ through $S_{Y|Z}$. For example

$$Q_T Z \perp\!\!\!\perp r_{1|2} \Rightarrow S_{\omega|r_{1|2}} \subseteq P_T S_{Y|Z}. \tag{4.6}$$

Residual views are generally preferable to marginal views because they remain efficient under a wider variety of dependencies among the predictor variables. Regardless of the choice of ω, the condition $Q_T Z \perp\!\!\!\perp P_T Z$ in (4.4) is more restrictive than the corresponding condition $Q_T Z \perp\!\!\!\perp r_{1|2}$ in (4.6). As a special case, if Z is normal the two conditions are equivalent since $r_{1|2} = P_T Z$, and hold for any T.

Concerning the choice of ω, one possibility is to stay with the original response, $\omega(Y, Q_T Z) = Y$, and hence consider $\{Y, r_{1|2}\}$. However, other specifications are possible. In particular, if $\omega(\cdot, Q_T Z)$ is a bijection Z-a.s., the marginal central DRS's coincide due to (2.5); $S_{\omega|r_{1|2}} = S_{Y|r_{1|2}}$. Thus, there is no loss of information on the coordinate subspace $P_T S_{Y|Z}$ when passing from $\{Y, r_{1|2}\}$ to $\{\omega, r_{1|2}\}$, and the same residual consistency assumption works for both cases; $S_{\omega|r_{1|2}} = S_{Y|r_{1|2}} \supseteq P_T S_{Y|Z}$.

The above discussion has an interesting practical consequence, which further increases the appeal of residual views. When we try to identify the central DRS in a low-dimensional view, the spread of the data along the response axis affects the resolution with which we are able to discern dependency patterns, and thus the accuracy of our visual investigation. If the view in question is a residual view, and if we can find a bijection such that $\mathrm{Var}(\omega(Y, Q_T Z)) < \mathrm{Var}(Y)$, switching from $\{Y, r_{1|2}\}$ to $\{\omega, r_{1|2}\}$ will improve resolution at no cost because all the information on $P_T S_{Y|Z}$ is retained.

Consider $\omega(Y, Q_T Z) = e_{Y|2}$, the residual from the population OLS linear regression of Y on $Q_T Z$. This transformation reduces (or leaves unchanged) the variation. If Z is normal, $e_{Y|2}$ is a bijection in Y. Moreover, the residual from the OLS linear regression of $P_T Z$ on $Q_T Z$ is $e_{1|2} = r_{1|2} = P_T Z$. Applying (4.6) and imposing consistency we have then

$$S_{e_{Y|2}|P_T Z} = P_T S_{Y|Z} \tag{4.7}$$

which allows us to use $\{e_{Y|2}, P_T Z\}$ as residual view: $\{Y, (P_{S_{e_{Y|2}|P_T Z}} + Q_T)Z\}$ will be a sufficient view for the regression of Y on Z. This represents a *novel application of added-variable plots* (AVP's). In the standard use, an AVP serves to explore the dependence of Y on $P_T Z$ *after* $Q_T X$ (See Cook (1996*b*) for an introduction to the literature). Notice that using $\{e_{Y|2}, P_T Z\}$ as residual view does *not* require modeling assumptions on the conditional distribution of $Y \mid Q_T X$ (the OLS linear regression used to generate $e_{Y|2}$ need not be the correct model).

4.4 *Where not to look: (non-graphical) lower bounds and further improvements on iteration*

In this section, we go back to the issue of premature termination of the iterative procedure presented in Subsections 4.2.2–4.2.5, and show how the situation can be improved if we know a lower bound to the central DRS, say $S_{(L)} \subseteq S_{Y|Z}$. This knowledge tells us where we should positively not be looking when attempting reduction, and thus allows us to modify the procedure by limiting ourselves to the complementary linear region.

At stage t, let S_t indicate the current DRS, and $R = S_{(L)}^{\perp S_t}$ the orthogonal complement of the lower bound within S_t. Suppose S_t is still too large for us to look at $\{Y, P_{S_t}Z\}$, but the portion of the space complementary to the lower bound is actually small, so that we can visualize $\{Y, P_R Z\}$ directly. In other words, suppose that the difference $\dim(S_t) - \dim(S_{(L)})$ is ≤ 2 (or 3 if the response is binary). Then we can positively assess irreducibility inspecting the second view. In fact, since $S_{(L)} \subseteq S_{Y|Z}$, $S_{Y|P_{S_t}Z} = S_t$ if and only if $S_{Y|P_R Z} = R$.

A similar logic allows us to extend the discussion at the end of Subsection 4.2.6 to cases in which an $S_{(L)}$ is available. If we select a "reduction window" that is orthogonal to the lower bound within the current DRS, $T \subseteq R$, then the coordinate subspace $P_T S_{Y|Z}$ cannot be misleading about the dimension of $S_{Y|Z}$, as long as $\dim(T) > d_{Y|Z} - \dim(S_{(L)})$. Adding efficiency to remove potential predictor space effects, we have the following: *If the marginal views employed at each stage are efficient, orthogonal to the lower bound, and based on "reduction windows" that exceed the dimension of the unknown portion* $S_{(L)}^{\perp S_{Y|Z}}$, *then iteration is guaranteed to proceed all the way down to* $S_{Y|Z}$.

In Section 3 we saw how non-graphical methods can be used to estimate directions in the central DRS, and thus lower bounds for it. For example, under the linearity condition (3.1), the OLS direction is in the central DRS. Suppose the distribution of Z is such that linearity and efficiency of marginal views hold. Then, taking $S_{(L)} = \text{Span}(\text{Cov}(Z, Y))$ and employing 2-dimensional "reduction windows" selected in its orthogonal complement at each iteration stage, we will reach $S_{Y|Z}$ when $d_{Y|Z} = 0, 1$ or 2. And if the response is binary, so that we can increase the dimension of our "reduction windows" to 3, we will reach the central DRS when $d_{Y|Z} = 0, 1, 2$ or even 3.

Potentially stricter lower bounds $S_M \subseteq S_{Y|Z}$ could be obtained through the other non-graphical methods we mentioned, but besides the stronger requirements on Z, their reliability would depend crucially on the accuracy of the inference concerning $\dim(S_M)$. The example in the next section employs the iterative procedure with residual views in place of marginal views, and the OLS lower bound.

5. The reaction yield data

Box and Draper ((1987), p. 368) reported 32 observations on the percentage yield Y from a two-stage chemical process characterized by temperatures T_1, T_2 (in degrees Celsius), log reaction times $\log(t_1)$, $\log(t_2)$ at the two stages, and percent concentration C of one of the reactants. Previous studies using steepest ascent indicated that the maximum yield was likely to be found in the region of the factor space covered by the experiment. Accordingly, Box and Draper fitted a full second-order response surface in the five predictors $X_1 = T_1$, $X_2 = \log(t_1)$, $X_3 = C$, $X_4 = T_2$ and $X_5 = \log(t_2)$.

Cook (1998*a*,1998*b*), and Cheng and Li (1995) analyzed these data with different versions of pHd. Using tests on the singular values of pHd's $\hat{M}$, Cook concluded that the

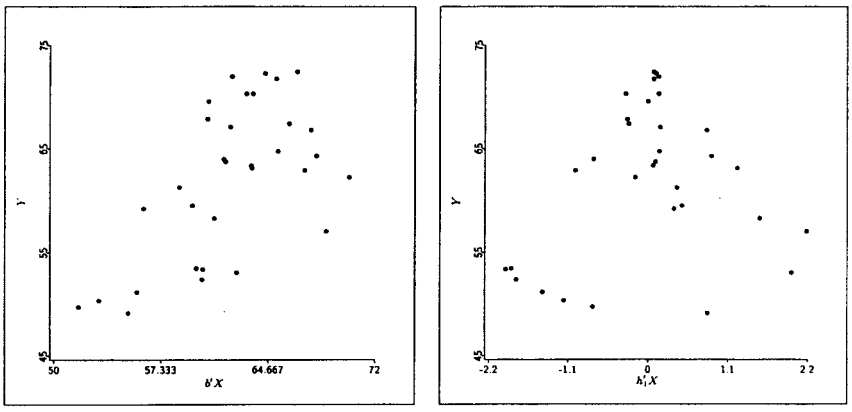
regression of Y on X has 2D structure. Cheng and Li effectively based their analysis on an assumption of at most 2D structure at the outset. Working in the Z-scale, we revisit these data, arguing that our graphical approach identifies a third relevant dimension.

We assume (4.2) and (4.5) to hold whenever needed. The scatter-plot matrix of the predictors showed no notable nonlinearities, and hence no evidence against the linearity condition (3.1) and efficiency. See Cook (1998*b*) for further discussion of the behavior of the predictors in these data.

We started the analysis by obtaining the sample coefficient vector b_0 from the OLS linear regression of Y on Z; under the linearity condition, this estimates the direction in $S_{Y|Z}$ that corresponds to the linear trend, as discussed in Section 3. Next, following Cook (1998*b*), we applied pHd to the regression of the residuals $Y - b_0'Z$ on Z and obtained the vectors h_j, $j = 1, \ldots, 5$. These have a reasonable chance of spanning directions that correspond to curvature in the data, although whether they do so or not will not affect the overall effectiveness of our graphical analysis. Finally, we sequentially orthogonalized the first four pHd directions against b_0 to obtain an orthonormal basis of $\mathbb{R}^5$ consisting of the five vectors b_j, $j = 0, \ldots, 4$. Here, b_1 is the projection of h_1 onto $\text{Span}(b_0)^\perp$, b_2 is the projection of h_2 onto $\text{Span}(b_0, b_1)^\perp$ and so on. The 3D plot of Y against $(b_0'Z, b_1'Z)$ is quite similar to the summary plots found by Cook (1998*a*,1998*b*) and Cheng and Li (1995). Figure 4 contains the two marginal views $\{Y, b_0'Z\}$ and $\{Y, b_1'Z\}$.

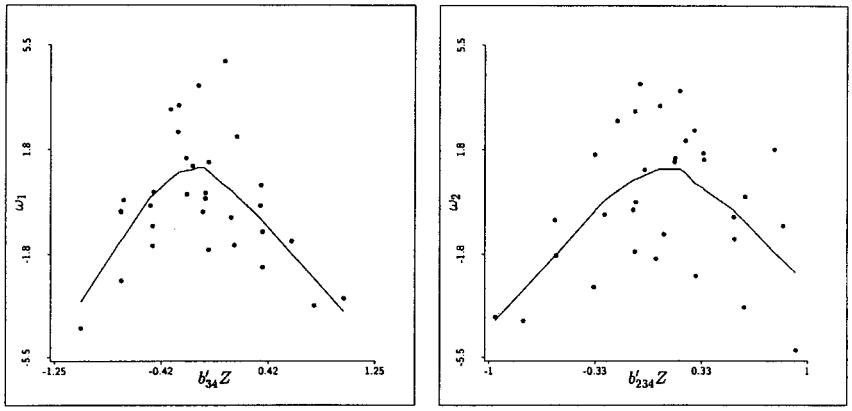
For the first iteration, we took $T = \text{Span}(b_3, b_4) \subset \text{Span}(b_0)^\perp$, and constructed a coordinate version of the corresponding view as the 3D plot of Y versus $(b_3'Z, b_4'Z)$. The plot was difficult to interpret, so to improve resolution we replaced Y with the residuals ω_1 from fitting the full second-order model in $b_0'Z$, $b_1'Z$ and $b_2'Z$. Using the methods discussed by Cook ((1998*a*), Chapter 4), analysis of the 3D plot of ω_1 versus $(b_3'Z, b_4'Z)$ resulted in an inference of 1D structure, with $\hat{S}_{\omega_1|P_T Z} = \text{Span}(b_{34}'Z)$. The corresponding central view $\{\omega_1, b_{34}'Z\}$ is shown in Fig. 5(a). Thus, we took $S_1 = \hat{S}_{\omega_1|P_T Z} \oplus T^\perp = \text{Span}(b_0, b_1, b_2, b_{34})$ as a DRS for Y on Z, and restricted ourselves to a regression with four rather than five predictors.

For the second iteration, we took $T = \text{Span}(b_2, b_{34}) \subset \text{Span}(b_0)^{\perp S_1}$. Repeating



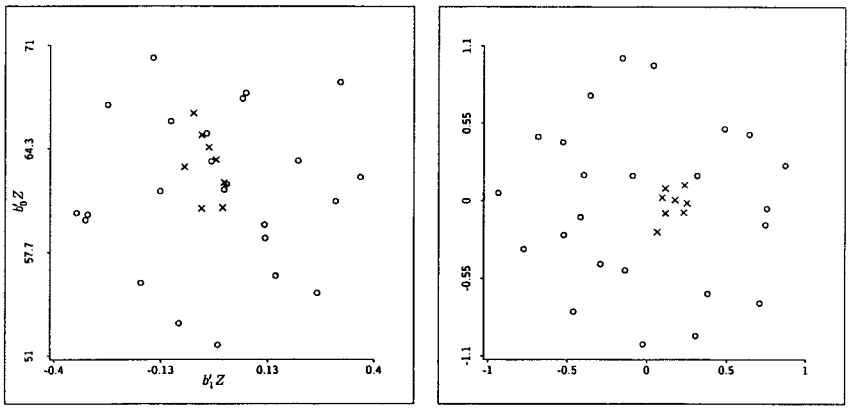
a. Y vs. $b_0'Z$ b. Y vs. $b_1'Z$

Fig. 4. Two marginal views from the reaction yield data.



a. first iteration b. second iteration

Fig. 5. Central views extracted from the 3D residual views (lowess smooths superimposed). Reaction yield data.



a. b'_0Z vs. b'_1Z b. Isolated maximum

Fig. 6. Two dichotomized-response views from the central subspace $S_{Y|Z}$ ($\times$ = high-response points). Reaction yield data.

the analysis of the first iteration, we once again inferred 1D structure with $\hat{S}_{\omega_2|P_T Z} = \text{Span}(b'_{234}Z)$. The corresponding central view is shown in Fig. 5(b). Consequently, we took $S_2 = \text{Span}(b_0, b_1, b_{234})$ as a DRS. At this point, although we were not yet in the position of visualizing Y on $P_{S_2}Z$, we could visualize directly what was left between the lower bound $\text{Span}(b_0)$ and S_2; that is, $\text{Span}(b_1, b_{234}) = \text{Span}(b_0)^{\perp_{S_2}}$. Analyzing the 3D plot of Y versus $(b'_1Z, b'_{234}Z)$, we inferred 2D structure. This allowed us to positively conclude that Y on S_2 is irreducible, and therefore to take $\hat{S}_{Y|Z} = \text{Span}(b_0, b_1, b_{234})$. To illustrate the need for a third dimension after b_0 and b_1, we passed to a binary version $\tilde{Y}$ of the response that isolates the eight observations with highest Y. We then constructed the 3D binary response plot of $\tilde{Y}$ versus $(b'_0Z, b'_1Z, b'_{234}Z)$. Figure 6 shows two projections

of this plot. In Fig. 6(a), which corresponds to the solutions of Cook (1998*b*) and Cheng and Li (1995), high-response points are intertwined with low-response points, and no clear spatial separation seems possible. On the other hand, if we rotate using the third direction b_{234}, we get to the view shown in Fig. 6(b) which gives a perfect spatial mapping of the maximum yield: The eight high-response points are concentrated in a small region, and the low-response ones are scattered around it.

6. Discussion

As we mentioned in the Introduction, there are a variety of approaches to the graphical exploration of regression data, and the pursuit of interesting low-dimensional views. Our basic premise in this article is that the usefulness of modern graphics might be greatly increased if one could recognize situations in which low-dimensional views provide exhaustive information on the regression. The ability to recognize such situations requires a context for connecting the graphics with the statistics.

The one we proposed relies on the idea of sufficiency, and has at its core a meta-parameter named the central dimension-reduction subspace. We devoted most of our attention to graphical inference methods for it; that is, to procedures that pursue the central view by means of other low-dimensional views. But we addressed non-graphical methods, and showed how they can be used to enhance the graphical analysis.

The context we described is quite broad; it requires few scope-limiting conditions, most of which can be guaranteed through the predictor distribution. While the concepts and inference procedures in this article can be effectively applied to model-checking, they do not rely upon the introduction of a model, or more generally on substantial constraints on the form of $Y \mid X$. Because of this, reduction to $S_{Y|X}$ can be interpreted as the "phase-0" of a regression study, after which the modeling effort can be guided by the estimated central view. Without any loss of information, the usual practice of model construction and selection will begin where dimension reduction ends.

Throughout our discussion, the term "low-dimensional" referred to a view small enough to be visualized directly. In practice, with one axis devoted to the response, this means dimension ≤ 2 for the predictor, or possibly ≤ 3 when the response is binary or trinary. Computer-aided tools allowing quasi-direct visualization in 4D would extend this domain. They would allow us to employ 3-dimensional "reduction windows" in our iterations, (4-dimensional, when the response is binary or trinary), and thus further decrease the chances of premature termination: With such windows, and using the 1-dimensional OLS lower bound, we would be guaranteed to reach the central subspace of any regression with structural dimension ≤ 3 (4 when the response is binary).

It is important to remark that the concepts and results we introduced do not rely on Y being 1-dimensional. But visualization, and thus the implementation of graphical methods, clearly do. With two axes devoted to two continuous responses, we are left with only one axis available for the predictor; if one of the responses is binary or trinary, an additional axes can be freed for the predictor representing the discrete response through different plotting symbols, etc. But in general, a bivariate response forces us to employ smaller "reduction windows" in our iterations, possibly increasing the chances of premature termination. Of course lower bounds obtained from non-graphical methods can be used to balance out the risks involved in using smaller reduction windows. In particular, Cook (1998*a*) implemented and demonstrated the use of *bivariate sliced inverse regression*. Previously, Li *et al.* (1995) proposed a variant of SIR for multivari-

ate response regressions, involving Hotelling's notion of most predictable variates (these developments can be related to those in Fujikoshi (1982)).

Also, existence of first and second order moments of (Y, X) clearly plays a role in the non-graphical methods summarized in Section 3, but is not crucial to most of the graphical developments. What we presented in terms of the standardized Z can be rephrased in terms of the original X, and moments really come into the picture only in Proposition 2.1 and in Subsection 4.3.

Interesting issues concern the application of dimension reduction methods to data (Y_i, X_i), $i = 1, \ldots, n$ that do not represent an iid sample from the joint distribution of (Y, X), or in which the predictors are not random, as for example in designed experiments. In principle, nothing prevents the extensions of dimension reduction methodology to the case of non-iid observations; this is an extremely promising and challenging research venue, for which no literature is yet available. In the case of designed experiments, while the meaning of the conditional distribution $Y \mid X$ at the available X values is unchanged, that of the joint distribution (Y, X) becomes elusive. However, designed predictors may actually help in avoiding "geometrically pathological" data clouds, or failure of some basic assumptions. Regarding estimation of relevant directions through non-graphical methods such as OLS, SIR, SAVE or pHd, one can simply take expectations involving predictors with respect to their empirical (design) distribution. The biases thus introduced can be controlled through the choice of experimental design itself. So, for example, assumptions such as (3.1) or (3.2) could be "enforced by design". The situation is more delicate regarding some tests commonly employed by non-graphical methods for dimensional inferences, but as we remark at the end of Section 3, this is of little concern for the approach we propose in this paper. We use non-graphical methods only as pre-processors generating a set of directions to guide the graphical analysis, and we never employ any of the tests associated with them. More on the issue of designed predictors can be found in Cook ((1998*a*, Comment by Li and Rejoinder), Ibrahimy and Cook (1995) and Cheng and Li (1995).

Finally, issues related to dimension reduction and graphical exploration are of paramount importance also in multivariate settings in which no variable is elected to the role of response. In these cases, one would like to recognize situations in which low-dimensional projections provide exhaustive information on some target defined to embody structural traits of a multivariate distribution. Furthermore, one would like to do so assuming as little as possible on the nature of the latter. Developments along these lines can be found in Chiaromonte (1997, 1998, 2001).

Acknowledgements

This work was supported in part by National Science Foundation grant DMS 0103983. We would like to thank John R. Baxter and the referees of a previous version of this manuscript for their helpful comments.

Appendix

Existence of $S_{Y|X}$ results

We will make use of a well known lemma from real analysis:

LEMMA A.1. *Let $\Omega \subseteq \mathbb{R}^p$ be an open set, and $\alpha : \mathbb{R}^p \to \mathbb{R}^1$ an analytic function. Then given any two subspaces $S_1, S_2 \subseteq \mathbb{R}^p$, $\alpha(x) = \alpha(P_{S_1}x) = \alpha(P_{S_2}x)$, $\forall x \in \Omega$ implies $\alpha(x) = \alpha(P_{S_1 \cap S_2}x)$, $\forall x \in \Omega$.*

For a set A and two subspaces V, W consider $T = V \cap W$ and the sections

$$A(T^{\perp V} \oplus T^{\perp W}; t) = \{P_{T^{\perp V} \oplus T^{\perp W}} x, x \in A : P_T x = t\}, \qquad t \in P_T A.$$

We say that A has *linked sections* if for any choice of subspaces, any $t \in P_T A$ and any $y_1, y_2 \in A(T^{\perp V} \oplus T^{\perp W}; t)$, there exists a sequence $l_n \in A(T^{\perp V} \oplus T^{\perp W}; t)$, $n = 1, \ldots, N$ such that: (i) $l_1 = y_1$ and $l_N = y_2$, (ii) for all $n = 2, \ldots, N$ either $P_{T^{\perp V}} l_n = P_{T^{\perp V}} l_{n-1}$ or $P_{T^{\perp W}} l_n = P_{T^{\perp W}} l_{n-1}$. Notice that an open and convex set has linked sections.

LEMMA A.2. *Let $\Omega \subseteq \mathbb{R}^p$ be a set with linked sections, and $\alpha : \mathbb{R}^p \to \mathbb{R}^1$. Then given any two subspaces $S_1, S_2 \subseteq \mathbb{R}^p$, $\alpha(x) = \alpha(P_{S_1}x) = \alpha(P_{S_2}x)$, $\forall x \in \Omega$ implies $\alpha(x) = \alpha(P_{S_1 \cap S_2}x)$, $\forall x \in \Omega$.*

PROOF. Let $T = S_1 \cap S_2$. Since Ω has linked sections, for any $t \in P_T\Omega$ and any $y_1, y_2 \in \Omega(T^{\perp S_1} \oplus T^{\perp S_2}; t)$, there exists a linking sequence within the section as defined above. Thus, for any y_1, y_2 in the section one can move along such sequence keeping the value of $\alpha(\cdot)$ constant. In fact, $\alpha(y_1) = \alpha(l_1) = \alpha(P_{S_1} l_1) = \alpha(P_{T^{\perp S_1}} l_1 + t)$ and also $= \alpha(P_{T^{\perp S_2}} l_1 + t)$. Suppose $P_{T^{\perp S_1}} l_2 = P_{T^{\perp S_1}} l_1$, then $\alpha(y_1) = \alpha(P_{T^{\perp S_1}} l_2 + t)$ simply by coincidence of the argument points. The same reasoning applies with the third point of the sequence, etc. At the last step, one will have for example $\alpha(y_1) = \alpha(P_{T^{\perp S_2}} l_N + t)$ $= \alpha(P_{S_2} l_N) = \alpha(l_N) = \alpha(y_2)$. The statement follows.

Let $\mathcal{S}_{Y|X}$ indicate the class of all DRS's for the regression of Y on $X \in \mathbb{R}^p$. By construction, the central DRS exists if and only if $\bigcap_{\mathcal{S}_{Y|X}} S \in \mathcal{S}_{Y|X}$, or equivalently, if $\mathcal{S}_{Y|X}$ is closed under intersection. We are now in a position to prove Propositions 2.1 and 2.2 restated as follows. Recall we indicate with $\mathcal{L}_X$ and $Supp_X$ the law and closed support of X. Existence the first moment for Y is required in 2.1.

PROPOSITION A.1. *Assume that $Supp_X$ contains an open set Ω with $\mathcal{L}_X(\Omega) = 1$. Suppose furthermore that Y admits finite first order moments, $Y \perp\!\!\!\perp X \mid E(Y \mid X)$, and $E(Y \mid X)$ can be expressed as an analytic function of X, X-a.s. Then $S_1 \cap S_2 \in \mathcal{S}_{Y|X}$ for all $S_1, S_2 \in \mathcal{S}_{Y|X}$.*

PROOF. By definition, $S \in \mathcal{S}_{Y|X}$ if and only if $Y \perp\!\!\!\perp X \mid P_S X$. Since $Y \perp\!\!\!\perp X \mid E(Y \mid X)$, the conditional independence statement is equivalent to $E(Y \mid X) = E(Y \mid P_S X)$, X-a.s. Using an appropriate measurable function, this can be rewritten as $\alpha(X) = \alpha(P_S X)$, X-a.s. Taking $S_1, S_2 \in \mathcal{S}_{Y|X}$ we have then $\alpha(x) = \alpha(P_{S_1}x) = \alpha(P_{S_2}x)$, $\forall x \in \Omega$ $\subseteq Supp_X$. By Lemma A.1 this implies $\alpha(x) = \alpha(P_{S_1 \cap S_2}x)$, $\forall x \in \Omega$, which in turn implies $S_1 \cap S_2 \in \mathcal{S}_{Y|X}$ as $\mathcal{L}_X(\Omega) = 1$.

PROPOSITION A.2. *Assume that $Supp_X$ contains an open and convex set Ω with $\mathcal{L}_X(\Omega) = 1$. Then $S_1 \cap S_2 \in \mathcal{S}_{Y|X}$ for all S_1, $S_2 \in \mathcal{S}_{Y|X}$*

PROOF. The conditional independence statement $Y \perp\!\!\!\perp X \mid P_S X$ can be equivalently expressed as $\mathcal{L}_{Y|X}(B) = \mathcal{L}_{Y|P_S X}(B)$, X-a.s., for any measurable $B \subseteq \mathbb{R}^1$. Using appropriate measurable functions, this can be rewritten as $\alpha_B(X) = \alpha_B(P_S X)$, X-a.s., $\forall B$.

Taking $S_1, S_2 \in \mathcal{S}_{Y|X}$ we have then $\alpha_B(x) = \alpha_B(P_{S_1}x) = \alpha_B(P_{S_2}x)$, $\forall x \in \Omega \subseteq Supp_X$, $\forall B$. Furthermore, being open and convex, Ω has linked sections. By Lemma A.2 we have $\alpha(x) = \alpha(P_{S_1 \cap S_2}x)$, $\forall x \in \Omega$, $\forall B$, which in turn implies $S_1 \cap S_2 \in \mathcal{S}_{Y|X}$ as $\mathcal{L}_X(\Omega) = 1$.

Clearly the same proof works whenever $Supp_X$ contains a set having linked sections and probability 1. Relaxing convexity is very important, given the crucial role played by existence of the central DRS in the sufficiency-based theory of dimension reduction.

Results for conditional, marginal and residual views. Results are established in terms of X, and we introduce moments only when referring to predictor residuals. We make use of two lemmas that easily follow from Dawid (1979, 1980); see also Basu and Pereira (1983). Z here indicates a generic random variable, and not the standardized predictor.

LEMMA A.3. *Let U, W and Z be three random variables defined on a common probability space, and let $\gamma(\cdot, \cdot)$ be any measurable function defined on the domain space of (W, Z). Then $U \perp\!\!\!\perp W \mid Z$ implies $U \perp\!\!\!\perp \gamma(W, Z) \mid Z$. If furthermore $\gamma(\cdot, z)$ is a bijection for every z, $U \perp\!\!\!\perp W \mid Z$ is equivalent to $U \perp\!\!\!\perp \gamma(W, Z) \mid Z$.*

LEMMA A.4. *Let U, W and Z be as above, and let $\delta(\cdot)$ be any measurable function defined on the domain space of Z. Then $U \perp\!\!\!\perp W \mid Z$ and $U \perp\!\!\!\perp Z \mid \delta(Z)$, implies $U \perp\!\!\!\perp W \mid \delta(Z)$.*

We thus can prove.

LEMMA A.5. *Let $S \in \mathcal{S}_{Y|X}$ and $\rho, \delta : \mathbb{R}^p \to \mathbb{R}^1$ be measurable functions such that $\rho(X) \perp\!\!\!\perp P_S X \mid \delta(P_S X)$. Then $\eta(Y, P_S X) \perp\!\!\!\perp \rho(X) \mid \delta(P_S X)$ for any measurable $\eta : \mathbb{R}^1 \times \mathbb{R}^p \to \mathbb{R}^1$.*

PROOF. Since $S \in \mathcal{S}_{Y|X}$, $Y \perp\!\!\!\perp X \mid P_S X$. Applying Lemma A.3 twice we have $\eta(Y, P_S X) \perp\!\!\!\perp \rho(X) \mid P_S X$. At the same time we have $\rho(X) \perp\!\!\!\perp P_S X \mid \delta(P_S X)$. So $\eta(Y, P_S X) \perp\!\!\!\perp \rho(X) \mid \delta(P_S X)$ by Lemma A.4.

Proposition 4.1 is an immediate consequence of Lemma A.3: $Y \perp\!\!\!\perp X \mid P_S X$ is equivalent to $Y \perp\!\!\!\perp P_T X \mid P_S X$, as $T \oplus S = \mathbb{R}^p$ guarantees that X is a bijection in $P_T X$ once $P_S X$ is given. Recalling that $V = Q_T S_{Y|X}$, $r_{1|2} = P_T X - E(P_T X \mid Q_T X)$ and $\omega : \mathbb{R}^1 \times \mathbb{R}^p \to \mathbb{R}^1$ is a measurable transformation, Propositions 4.2 and 4.3 can be restated and proved as follows:

PROPOSITION A.3. *For any subspace $T \subseteq \mathbb{R}^p$, $P_T S_{Y|X} \oplus S_{P_V X|P_T X}$ is a DRS for the regression of Y on $P_T X$.*

PROOF. For notational simplicity, let $C = P_T S_{Y|X}$, $R = S_{P_V X|P_T X}$, and set $S = C \oplus V \oplus R$. S is clearly a DRS for Y on X, as $S_{Y|X} \subseteq C \oplus V \subseteq S$. Also, set $\rho(X) = P_T X$ and $\delta(P_{C \oplus V \oplus R} X) = P_{C \oplus R} X$. $C \oplus R$ is clearly a DRS for $P_V X$ on $P_T X$, as it contains the central DRS R. So we have $P_V X \perp\!\!\!\perp P_T X \mid P_{C \oplus R} X$ ($P_{C \oplus R}(P_T X) = P_{C \oplus R} X$). But since $P_{C \oplus V \oplus R} X$ is a bijection in $P_V X$ once $P_{C \oplus R} X$ is given, this is equivalent to

$P_T X \perp\!\!\!\perp P_{C\oplus V\oplus R} X \mid P_{C\oplus R} X$. Hence, by Lemma A.5, $\eta(Y, P_{C\oplus V\oplus R} X) \perp\!\!\!\perp P_T X \mid P_{C\oplus R} X$ for any choice of $\eta(\cdot,\cdot)$. In particular $Y \perp\!\!\!\perp P_T X \mid P_{C\oplus R} X$.

PROPOSITION A.4. *Let $T \subseteq \mathbb{R}^p$. Assume that X has finite first order moments and that $P_T S_{Y|X}$ is a DRS for the regression of $Q_T X$ on $r_{1|2}$. Then $P_T S_{Y|X}$ is a DRS for the regression of $\omega(Y, Q_T X)$ on $r_{1|2}$.*

PROOF. Again, let $C = P_T S_{Y|X}$, and set $S = C \oplus T^{\perp}$. S is clearly a DRS for Y on X, as $S_{Y|X} \subseteq C \oplus T^{\perp}$. Also, set $\rho(X) = r_{1|2}$ and $\delta(P_{C\oplus T^{\perp}} X) = P_C r_{1|2}$. C is a DRS for $Q_T X$ on $r_{1|2}$ by assumption: we have $Q_T X \perp\!\!\!\perp r_{1|2} \mid P_C r_{1|2}$, which implies $Q_T X + P_C E(P_T X \mid Q_T X) \perp\!\!\!\perp r_{1|2} \mid P_C r_{1|2}$. But since $P_{C\oplus T^{\perp}} X$ is a bijection in $Q_T X + P_C E(P_T X \mid Q_T X)$ once $P_C r_{1|2}$ is given, this is equivalent to $r_{1|2} \perp\!\!\!\perp P_{C\oplus T^{\perp}} X \mid P_C r_{1|2}$. Hence, by Lemma A.5, $\eta(Y, P_{C\oplus T^{\perp}} X) \perp\!\!\!\perp r_{1|2} \mid P_C r_{1|2}$ for any choice of $\eta(\cdot,\cdot)$. In particular $\omega(Y, Q_T X) \perp\!\!\!\perp r_{1|2} \mid P_C r_{1|2}$.

REFERENCES

Basu, D. and Pereira, C. A. B. (1983). Conditional independence in statistics, *Sankhyā Ser. A*, **45**, 324–337.

Box, G. E. P. and Draper, N. (1987). *Empirical Model-Building and Response Surfaces*, Wiley, New York.

Carrol, R. J. and Li, K. C. (1995). Binary regressors in dimension reduction models: A new look at treatment comparison, *Statist. Sinica*, **5**, 667–688.

Cheng, C. S. and Li, K. C. (1995). A study of the method of principal Hessian direction for analysis of designed experiments, *Statist. Sinica*, **5**, 617–640.

Chiaromonte, F. (1997). *A reduction paradigm for multivariate laws, L1 Statistical Procedures and Related Topics*, (ed. Y. Dodge), *lecture notes-monograph series, IMS Lecture Notes Monogr. Ser.*, 229–240, Hayward, California.

Chiaromonte, F. (1998). On multivariate structures and exhaustive reductions, *Computing Science and Statistics*, (ed. S. Weisberg), **30**, 204–213, The Interface Foundation of North America, Fairfax Station, Virginia.

Chiaromonte, F. (2001). Structures and exhaustive reductions: A general framework for the simplification of multivariate data (submitted).

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

Cook, R. D. (1994*b*). Using dimension-reduction subspaces to identify important inputs in models of physical systems, *Proceedings of the Section on Physical and Engineering Sciences*, 18–25, American Statistical Association, Alexandria, Virginia.

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

Cook, R. D. (1996*b*). Added-variable plots and curvature in linear regression, *Technometrics*, **38**, 275–278.

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

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

Cook, R. D. and Lee, H. (1999). Dimension reduction in binary response regression, *J. Amer. Statist. Assoc.*, **97**, 1187–1200.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction", *J. Amer. Statist. Assoc.*, **86**, 316–342.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*, Wiley, New York.

Cook, R. D. and Weisberg, S. (1999). Graphics in statistical analysis: Is the medium the message?, *Amer. Statist.*, **53**, 29–37.

Cook, R. D. and Wetzel, N. (1993). Exploring regression structure with graphics (with discussion), *Test*, **2**, 33–100.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals, *J. Roy. Statist. Soc. Ser. B*, **30**, 248–275.

Dawid, A. P. (1979). Conditional independence in statistical theory, *J. Roy. Statist. Soc. Ser. B*, **41**, 1–31.

Dawid, A. P. (1980). Conditional independence for statistical operations, *Ann. Statist.*, **8**, 598–617.

Flury, B. and Riedwyl, H. (1998). *Multivariate Statistics: A Practical Approach*, Chapman and Hall, London.

Fujikoshi, Y. (1982). A test for additional information in canonical correlation analysis, *Ann. Inst. Statist. Math.*, **34**, 523–530.

Ibrahimy, A. and Cook, R. D. (1995). Regression design for one-dimensional subspaces, *MODA4—Advances in Model-Oriented Data Analysis* (eds. C. P. Kitsas and W. G. Muller), **86**, 125–132, Physica, Heidelberg.

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.

Li, K. C., Aragon Y. and Thomos-Agan C. (1995). Analyzing multivariate outcome data: SIR and a non linear theory of Hotelling's most predictable variates, (to appear in J. Amer. Statist. Assoc.).

McKay, R. J. (1977). Variable selection in multivariate regression: An application of simultaneous test procedures, *J. Roy. Statist. Soc. Ser. B.*, **39**, 371–380.

Swayne, D. F., Cook, D. and Buja, A. (1998). XGobi: Interactive dynamic data visualization in the X window system, *J. Comput. Graph. Statist.*, **7**, 113–130.