The Annals of Statistics
2002, Vol. 30, No. 2, 475–497

SUFFICIENT DIMENSION REDUCTION IN REGRESSIONS WITH CATEGORICAL PREDICTORS

BY FRANCESCA CHIAROMONTE, R. DENNIS COOK[1] AND BING LI

Pennsylvania State University, University of Minnesota and Pennsylvania State University

In this article, we describe how the theory of sufficient dimension reduction, and a well-known inference method for it (sliced inverse regression), can be extended to regression analyses involving both quantitative and categorical predictor variables. As statistics faces an increasing need for effective analysis strategies for high-dimensional data, the results we present significantly widen the applicative scope of sufficient dimension reduction and open the way for a new class of theoretical and methodological developments.

1. Introduction. Typical regression analyses investigate the dependence of a response Y on a vector $\mathbf{X}$ of p predictors. Although the focus is often on the mean function $\mathrm{E}(Y|\mathbf{X})$, and perhaps the variance function $\mathrm{Var}(Y|\mathbf{X})$, the general object of interest is the conditional distribution of $Y|\mathbf{X}$, as a function of the value assumed by $\mathbf{X}$. For these settings, *sufficient dimension reduction* permits us to restrict attention to a projection $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ of the predictor vector $\mathbf{X}$ onto a subspace $\mathcal{S}$ of the predictor space, *without loss of information* on $Y|\mathbf{X}$. This reduction precedes the familiar model-building exercises which can therefore be limited to a number $d \leq p$ of new predictors, expressed as linear combinations of the original ones: $\mathbf{v}_1'\mathbf{X}, \ldots, \mathbf{v}_d'\mathbf{X}$, where $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is a basis of $\mathcal{S}$. In applications, the drop in dimension is often substantial even when starting with large p's; d's equal to 1 or 2 are common in practice, and they allow a fully informative and direct visualization of the original regression through a plot of Y versus the new predictors.

Methodological implementations such as sliced inverse regression [SIR; Li (1991)] and sliced average variance estimation [SAVE; Cook and Weisberg (1991)] of this general paradigm have been limited primarily to regressions with many-valued, possibly continuous, quantitative predictors because it is in such settings that dimension reduction may be particularly relevant. Straightforward application to regressions that include qualitative predictors such as species, sex or location may be inappropriate because then relevance of the linear combinations involving qualitative variables can be elusive.

In this article, we extend sufficient dimension reduction to regressions that include a qualitative predictor W in addition to a vector $\mathbf{X}$ of many-valued

Received January 2001; revised September 2001.
[1]Research for this article was supported in part by NSF Grant DMS-97-03777.
AMS 2000 subject classifications. Primary 62G08; secondary 62G09, 62H05.
Key words and phrases. Central subspace, graphics, SAVE, SIR, visualization.

predictors. The predictor W may represent a single qualitative variable, such as species, or a combination of such variables, such as species and location. It may also be a continuous predictor that is represented in too few values for linear combinations to be useful, or a categorical version of a continuous predictor. In short, we consider the regression of Y on $(\mathbf{X}, W)$, where the predictors in $\mathbf{X}$ are many-valued and W identifies a number of subpopulations, say $w = 1, \ldots, C$.

We approach dimension reduction in the regression of Y on $(\mathbf{X}, W)$ by seeking a projection $\mathbf{P}_{\mathcal{S}^{(W)}}\mathbf{X}$ of $\mathbf{X}$ that preserves information on $Y|(\mathbf{X}, W)$; that is, we "constrain" the reduction of $\mathbf{X}$ through the subpopulations established by W.

Letting $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ be a basis of $\mathcal{S}^{(W)}$, model-building is then aided by visualizing the original regression through a plot of Y versus the new predictors $\mathbf{v}_1'\mathbf{X}, \ldots, \mathbf{v}_d'\mathbf{X}$ with points marked to indicate the W subpopulations.

We refer to this as *partial* dimension reduction of $\mathbf{X}$, for the regression of Y on $(\mathbf{X}, W)$. We will see how this approach need not coincide with *marginal* dimension reduction for the regression of Y on $\mathbf{X}$, nor with *conditional* dimension reduction for the regression of Y on $\mathbf{X}$ within the subpopulations identified by W. But we will also see how partial, marginal and conditional dimension reduction are related to one another. These relationships are at the core of the inference methods we propose for partial dimension reduction.

In Section 2 we review some key concepts and briefly describe one inference method for sufficient dimension reduction: sliced inverse regression. In Section 3 we introduce partial sufficient dimension reduction and map its connections to marginal and conditional reduction. Section 4 is dedicated to the development of sliced inverse regression for partial reduction. This development includes large sample testing for the dimension of the subspace $\mathcal{S}^{(W)}$. Final remarks are given in Section 5. Proofs for most propositions in the paper are provided in a technical appendix.

2. Sufficient dimension reduction. Consider the regression of Y on $\mathbf{X} \in \mathbb{R}^p$. In sufficient dimension reduction, the main object of interest is the intersection of all subspaces $\mathcal{S} \subseteq \mathbb{R}^p$ such that

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{P}_{\mathcal{S}}\mathbf{X}, \tag{1}$$

where $\mathbf{P}_{(\cdot)}$ stands for the projection operator in the standard inner product and $\perp\!\!\!\perp$ indicates independence. The statement is thus that Y is independent of $\mathbf{X}$ given any value for $\mathbf{P}_{\mathcal{S}}\mathbf{X}$. When the intersection itself satisfies the above conditional independence condition, it is called the *central subspace* (CS) of the regression and is indicated with $\mathcal{S}_{Y|\mathbf{X}}$. Its dimension $d_{Y|\mathbf{X}} = \dim(\mathcal{S}_{Y|\mathbf{X}})$ is called the *structural dimension* of the regression. A plot of Y versus $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$, the latter being expressed through any basis of $\mathcal{S}_{Y|\mathbf{X}}$, is called the *central view* of the regression, with the understanding that it may be directly visualizable only when $d_{Y|\mathbf{X}}$ is small.

The CS does not exist for all regressions; some assumptions are required to guarantee that the intersection of all subspaces satisfying (1) does itself satisfy

the condition. Because these assumptions permit a broad range of practical applications, we do not view the existence issue as worrisome and thus we assume that central subspaces exist for all regressions considered in this article. For background on this issue and an introduction to the associated literature, see Cook (1998, Chapter 6), Cook and Weisberg (1999a) and Chiaromonte and Cook (2002).

The CS represents the minimal subspace that preserves the original information relative to the regression, in the sense that the conditional distribution of $Y|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$ is the same as the conditional distribution of $Y|\mathbf{X}$. Being defined as an intersection, this minimal subspace is unique and thus constitutes a well-defined object of inference.

$\mathcal{S}_{Y|\mathbf{X}}$ has several useful properties, among which is straightforward behavior under full rank affine transformations of the predictor vector: if $\mathbf{a} \in \mathbb{R}^p$, and $\mathbf{A}$: $\mathbb{R}^p \to \mathbb{R}^p$ is a full rank linear operator, then

$$\mathcal{S}_{Y|\mathbf{a}+\mathbf{A}\mathbf{X}} = (\mathbf{A}')^{-1}\mathcal{S}_{Y|\mathbf{X}}. \tag{2}$$

The structural dimension does not change ($d_{Y|\mathbf{a}+\mathbf{A}\mathbf{X}} = d_{Y|\mathbf{X}}$), and the two spaces can be obtained from one another through the $\mathbf{A}$ operator. Because of (2), indicating with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ the mean and covariance of $\mathbf{X}$ and assuming that $\boldsymbol{\Sigma}$ is invertible, we usually shift attention to the standardized predictor $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ and the corresponding space $\mathcal{S}_{Y|\mathbf{Z}}$. The CS on the original predictor scale is then given by $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}$.

Several graphical and numerical methods for the estimation of the CS are available. In this paper, we concentrate on sliced inverse regression [Li (1991)], although other methods can be adapted to partial dimension reduction along the lines developed here. The following review of SIR departs slightly from the literature, as it is given for later comparison with SIR for partial dimension reduction.

2.1. *Sliced inverse regression.* Sliced inverse regression is based on a fundamental result by Li (1991): if the interdependencies within the predictor vector are linear along the CS, that is,

$$\mathrm{E}\big(\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}\big) = \mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}, \tag{3}$$

then $\mathrm{E}(\mathbf{Z}|Y) \in \mathcal{S}_{Y|\mathbf{Z}}$. To allow relatively easy estimation of the inverse mean vector $\mathrm{E}(\mathbf{Z}|Y)$, the response is replaced with a discrete version $\tilde{Y}$, with finite support constructed by partitioning the range of Y into H slices. Under (3) one has then

$$\mathrm{E}(\mathbf{Z}|\tilde{Y}) \in \mathcal{S}_{\tilde{Y}|\mathbf{Z}} \subseteq \mathcal{S}_{Y|\mathbf{Z}}$$

so that the covariance of the inverse mean vector

$$\boldsymbol{\Theta} = \mathrm{Cov}\big(\mathrm{E}(\mathbf{Z}|\tilde{Y})\big) = \mathrm{E}\big(\mathrm{E}(\mathbf{Z}|\tilde{Y})\mathrm{E}(\mathbf{Z}|\tilde{Y})'\big) \tag{4}$$

has column span $\mathrm{Span}(\boldsymbol{\Theta}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. In addition to the linearity condition (3), it is also often assumed that $\boldsymbol{\Theta}$ allows us to recover the whole CS: letting $\mathbf{P}_\Theta$ be the projection on $\mathrm{Span}(\boldsymbol{\Theta})$, one postulates that

$$Y \perp\!\!\!\perp \mathbf{Z}|\mathbf{P}_\Theta\mathbf{Z}, \tag{5}$$

which is equivalent to $\text{Span}(\boldsymbol{\Theta}) \supseteq \mathcal{S}_{Y|\mathbf{Z}}$. Together with (3), this coverage condition guarantees $\text{Span}(\boldsymbol{\Theta}) = \mathcal{S}_{Y|\mathbf{Z}}$. In summary, the population basis for SIR is as follows:

PROPOSITION 2.1. *Under linearity* (3) *and coverage* (5), $\boldsymbol{\Theta}$ *as defined in* (4) *has* $\text{Span}(\boldsymbol{\Theta}) = \mathcal{S}_{Y|\mathbf{Z}}$.

The central subspace can then be reconstructed as

$$\mathcal{S}_{Y|\mathbf{Z}} = \text{Span}(\boldsymbol{\Theta}) = \text{Span}(\mathbf{t}_1, \ldots, \mathbf{t}_d),$$

where $d = \text{rank}(\boldsymbol{\Theta})$, and the $\mathbf{t}_j$'s are the eigenvectors corresponding to eigenvalues $\theta_j \neq 0$ in the spectral decomposition

$$\boldsymbol{\Theta} = \sum_{j=1}^{p} \theta_j \mathbf{t}_j \mathbf{t}_j'$$

(here and elsewhere, eigenvalues are listed in nonincreasing order). Thus, in practice, $\mathcal{S}_{Y|\mathbf{Z}}$ is estimated as the span of the eigenvectors of $\hat{\boldsymbol{\Theta}}$, a sample version of $\boldsymbol{\Theta}$, whose eigenvalues are inferred to correspond to nonzero population eigenvalues.

Assuming that an iid sample $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, from the joint distribution of $(\mathbf{X}, Y)$ is available, SIR is applied according to the following algorithm: Using moment estimates $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ for the mean and covariance of $\mathbf{X}$, the predictor observations are standardized to $\hat{\mathbf{Z}}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X}_i - \hat{\boldsymbol{\mu}})$, $i = 1, \ldots, n$. Then, creating a system of $s = 1, \ldots, H$ slices on the sample range of Y, one calculates intraslice mean vectors as

$$\bar{\mathbf{Z}}_s = \frac{1}{n_s} \sum_{i|s} \hat{\mathbf{Z}}_i, \qquad s = 1, \ldots, H,$$

where the sum is over the indices i of response observations Y_i that fall into slice s, and n_s is the number of observations in slice s. Since the mean vectors $\bar{\mathbf{Z}}_s$ average to 0 over the slices ($\frac{1}{n}\sum_{s=1}^{H} n_s \bar{\mathbf{Z}}_s = 0$), a sample version of $\boldsymbol{\Theta}$ is then constructed as

$$\hat{\boldsymbol{\Theta}} = \sum_{s=1}^{H} \frac{n_s}{n} \bar{\mathbf{Z}}_s \bar{\mathbf{Z}}_s'$$

and decomposed spectrally as

$$\hat{\boldsymbol{\Theta}} = \sum_{j=1}^{p} \hat{\theta}_j \hat{\mathbf{t}}_j \hat{\mathbf{t}}_j'.$$

The eigenvalues of $\hat{\boldsymbol{\Theta}}$ are used to produce an estimate $\hat{d}$ of $d = \text{rank}(\boldsymbol{\Theta})$ (see Section 2.2), while the eigenvectors are used to construct $\text{Span}(\hat{\mathbf{t}}_1, \ldots, \hat{\mathbf{t}}_{\hat{d}})$, which estimates a lower bound under (3), and the whole $\mathcal{S}_{Y|\mathbf{Z}}$ under (3) and (5).

Back to the $\mathbf{X}$-scale, one forms SIR *predictors* as $\hat{\mathbf{v}}_1'\mathbf{X}, \ldots, \hat{\mathbf{v}}_p'\mathbf{X}$, where

$$\hat{\mathbf{v}}_j = \hat{\mathbf{\Sigma}}^{-1/2}\hat{\mathbf{t}}_j, \qquad j = 1, \ldots, p,$$

and constructs the estimated $(1 + \hat{d})$-dimensional central view for the regression as the plot of Y versus the first $\hat{d}$ of such predictors.

2.2. *Large sample testing for the rank of* $\mathbf{\Theta}$. A summary plot of the response versus the first few SIR predictors, or a scatter-plot matrix of the response and all p SIR predictors, is generally informative regardless of d. Nevertheless, since d is usually unknown, inference about it is helpful to the practical effectiveness of SIR. Using the test statistic

$$T(m) = n \sum_{j=m+1}^{p} \hat{\theta}_j \tag{6}$$

Li (1991) proposed to estimate d by testing a series of hypotheses of the form H_o: $d = m$ versus H_a: $d > m$. Beginning with $m = 0$, compare $T(m)$ to the percentage points of its distribution under the null hypothesis $d = m$. If it is smaller, there is no significant evidence against the null hypothesis. If it is larger, we conclude that $d > m$, increment m by 1 and repeat the procedure. The estimate $\hat{d} = m$ follows when $T(m-1)$ is relatively large, implying that $d > m - 1$, while $T(m)$ is relatively small, so that $d = m$ cannot be rejected. Implementation of this procedure requires a null distribution for $T(d)$. Li showed that, when the predictors are normally distributed and the coverage assumption holds, the asymptotic distribution of $T(d)$ is χ^2 with $(H - d - 1)(p - d)$ degrees of freedom. Extensions of this distributional result were studied by Cook [(1998), Chapter 11], Schott (1994) and Velilla (1998). For comparison with SIR for partial dimension reduction, we state the following:

PROPOSITION 2.2. *Let* $d = \text{rank}(\mathbf{\Theta})$, *let* $\mathbf{P}_\Theta$ *be the projection on* $\text{Span}(\mathbf{\Theta})$ *and let* $\mathbf{Q}_\Theta = \mathbf{I}_p - \mathbf{P}_\Theta$. *If* (a) $Y \perp\!\!\!\perp \mathbf{Z}|\mathbf{P}_\Theta\mathbf{Z}$, (b) $E(\mathbf{Z}|\mathbf{P}_\Theta\mathbf{Z}) = \mathbf{P}_\Theta\mathbf{Z}$ *and* (c) $\text{Cov}(\mathbf{Z}|\mathbf{P}_\Theta\mathbf{Z}) = \mathbf{Q}_\Theta$, *then the asymptotic distribution of* $T(d)$ *defined in* (6) *is a* $\chi^2_{(H-d-1)(p-d)}$, *where* H *is the number of response slices used in the SIR algorithm.*

Note that (a) corresponds to coverage (5), and (b) to linearity (3) under coverage; (c) is a constant covariance condition required to obtain the asymptotic χ^2. Note also that linearity and constant covariance are satisfied when the predictor vector is normal [see, e.g., Cook (1998), Section 7.3.2].

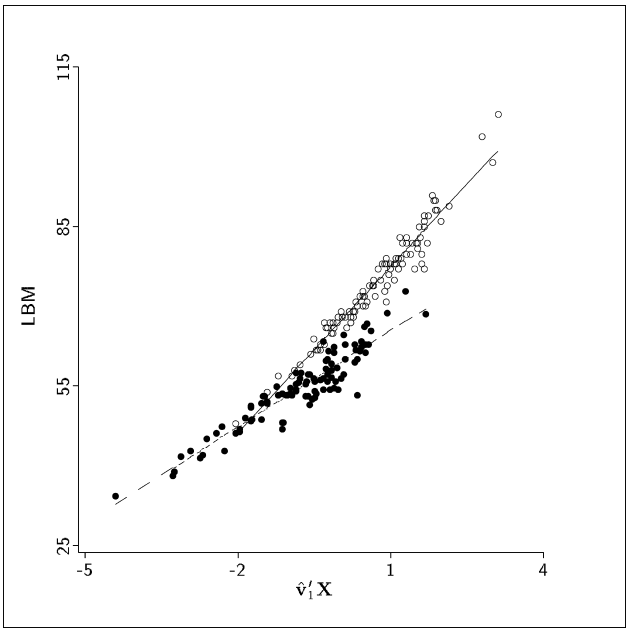
2.3. *Illustration: lean body mass regression.* Several studies have investigated the relationship between body fat and various predictors with the goal of identifying overweight individuals and understanding factors that may be associated with this condition. An introduction to the literature on this topic was given by Nevill and Holder (1995). In this illustration, which is continued later in the article, we consider a data set discussed by Cook and Weisberg (1994, 1999b). Lean body mass L is regressed on the logarithms of height, weight, red cell count, white cell count and hemoglobin, plus an indicator for gender, for a sample of 202 individuals training at the Australian Institute of Sport. Thus, we have five many-valued predictors that comprise $\mathbf{X}$ and one qualitative predictor, gender, which we denote as G with $G = 1$ for males and $G = 2$ for females.

When attempting to reduce the dimension of the predictors to facilitate visualization, we are faced with the issue of how to deal with G. There are several possibilities represented in the literature. A first possibility is to simply apply a dimension reduction method such as SIR to the regression of L on $(\mathbf{X}, G)$, but practical and theoretical issues raise some doubt about its usefulness. As mentioned previously, the relevance of linear combinations involving qualitative predictors may not be clear. In addition, the linearity condition (3) may become tenuous and special uniqueness and existence issues arise for the central subspace.

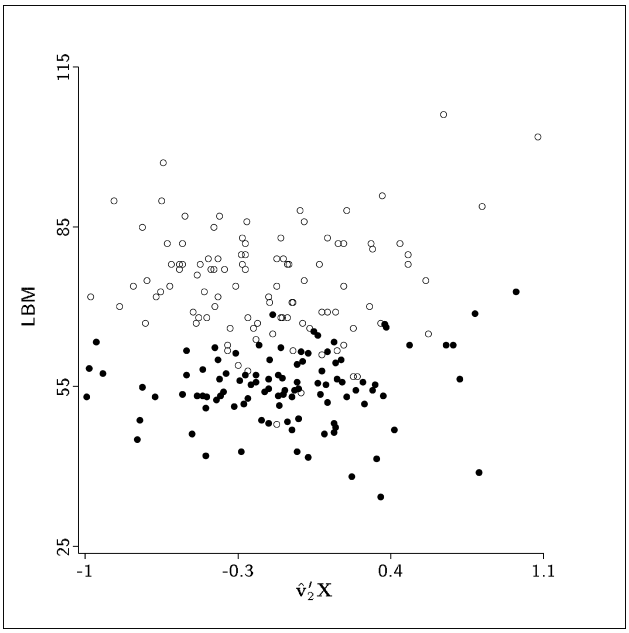
A second possibility is to apply a dimension reduction method to the marginal regression of L on $\mathbf{X}$ and attempt to incorporate G when interpreting the results. A third possibility is to proceed conditionally, applying a dimension reduction method within each gender and then comparing the results. These approaches are combined by Cook and Weisberg [(1994), Section 8.2], who studied a restricted regression involving only three of the five many-valued predictors mentioned above. Following their strategy, we start by applying SIR to the marginal regression of L on $\mathbf{X}$. This produces two relevant SIR predictors, $\hat{\mathbf{v}}_1'\mathbf{X}$ and $\hat{\mathbf{v}}_2'\mathbf{X}$, because the large sample tests described in Section 2.2 lead to an estimate of $\hat{d} = 2$ for the structural dimension. Figure 1 shows plots of L versus $\hat{\mathbf{v}}_1'\mathbf{X}$ and $\hat{\mathbf{v}}_2'\mathbf{X}$. Given the obvious dependence between $\mathbf{X}$ and G, the first open question concerns the relevance of these marginal results to the investigation of $L|(\mathbf{X}, G)$.

Next, we applied SIR to males and females separately, identifying only one relevant predictor in each case, $\hat{\mathbf{v}}_{1m}'\mathbf{X}$ and $\hat{\mathbf{v}}_{1f}'\mathbf{X}$. The sample correlation between $\hat{\mathbf{v}}_{1m}'\mathbf{X}$ and $\hat{\mathbf{v}}_1'\mathbf{X}$ is 0.977, and that between $\hat{\mathbf{v}}_{1f}'\mathbf{X}$ and $\hat{\mathbf{v}}_1'\mathbf{X}$ is 0.986, suggesting that the relevant linear combinations for males and females are the same and coincide with the first SIR predictor identified by the marginal analysis. Two more questions arise: Can this informal finding be placed on a better foundation? Is it in contradiction with our previous conclusion that the marginal central subspace $\mathcal{S}_{L|\mathbf{X}}$ is two-dimensional?

It must also be noticed that informal comparison of the outcomes of conditional dimension reduction becomes unwieldy when the qualitative predictor W has several levels, especially if dependencies between $\mathbf{X}$ and W increase the likelihood



(a)



(b)

FIG. 1. *Summary plots from the regression of L on* $\mathbf{X}$: (a) *L versus the first SIR predictor* $\hat{\mathbf{v}}_1'\mathbf{X}$ (*lines represent OLS fitted values within gender*); (b) *L versus the second SIR predictor* $\hat{\mathbf{v}}_2'\mathbf{X}$; ∘ *males,* • *females.*

that the linear combinations relevant for each subpopulation regression differ [Cook and Critchley (2000)].

Postulating that subpopulation regressions share relevant linear combinations, Carroll and Li (1995) investigated methods for estimating the unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}^1$ in models of the form

$$Y = g\big(\alpha \times \mathrm{Ind}(W = 2) + \boldsymbol{\beta}'\mathbf{X}; \varepsilon\big), \tag{7}$$

where g is an unknown function, the error ε is independent of $(\mathbf{X}, W)$, W has only two levels and $\mathrm{Ind}(W = 2)$ is an indicator variable for the second. Thus, they confined attention to a two-subpopulation case, in which *one* linear combination suffices for each subpopulation, and *the same* linear combination $\boldsymbol{\beta}'\mathbf{X}$ serves for both. In addition, (7) limits the effect of W to an additive shift of $\boldsymbol{\beta}'\mathbf{X}$ in the first argument of g. Carroll and Li suggested a two-stage estimation procedure when $\mathbf{X}$ and W are independent: first use SIR on the marginal regression of $Y|\mathbf{X}$ to estimate $\mathrm{Span}(\boldsymbol{\beta})$ and then, given that estimate, use a different nonparametric method to estimate α. When $\mathbf{X}$ and W are dependent, they suggest proceeding conditionally for the first stage: use SIR within each subpopulation to produce two independent estimates of $\mathrm{Span}(\boldsymbol{\beta})$ and then combine these estimates using a weighted average. Back to our data, assuming temporarily that indeed $\mathbf{v}_1 = \mathbf{v}_{1m} = \mathbf{v}_{1f}$, Figure 1a does not support a model of the form given in (7).

In the next sections, we introduce a framework that allows us to connect marginal and conditional dimension reduction to the pursuit of a projection of $\mathbf{X}$ that preserves information on $Y|(\mathbf{X}, W)$. We then present methodology for combining conditional dimension reduction outcomes in what we call partial dimension reduction. This methodology overcomes complications due to the number of levels in W, and dependencies between $\mathbf{X}$ and W. Moreover, it does not refer to models such as (7) in which dimension reduction and the nature of the effect of W are postulated at the outset. For example, our approach would allow consideration of models of the form

$$Y = g\left(\sum_{w=1}^{C}\{\alpha_w \times \mathrm{Ind}(W = w)\} + \sum_{w=1}^{C}\{\boldsymbol{\beta}_w' \times \mathrm{Ind}(W = w)\}\mathbf{X}; \varepsilon\right)$$

where $\mathrm{Ind}(W = w)$, $w = 1, \ldots, C$, are indicator variables for the levels of W.

3. Partial dimension reduction. Recall that $\mathbf{X} \in \mathbb{R}^p$ is the predictor vector with respect to which one wishes to perform dimension reduction, while W is an additional predictor that is not to be included in the reduction—in our interpretation, we think of W as representing one or more categorical variables that identify $w = 1, \ldots, C$ subpopulations.

We consider the intersection of all subspaces $\mathcal{S} \subseteq \mathbb{R}^p$ such that

$$Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W). \tag{8}$$

Under the assumption that such intersection itself satisfies (8), we call it the *partial central subspace* relative to $\mathbf{X}$, for the regression of Y on $(\mathbf{X}, W)$, and indicate it with $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$. Correspondingly, $d_{Y|\mathbf{X}}^{(W)} = \dim(\mathcal{S}_{Y|\mathbf{X}}^{(W)})$ is the *partial structural dimension*. A plot of Y versus $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$, expressed through any basis of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ and with points marked to indicate the W subpopulation, is a *partial central view*. Existence conditions and properties of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ can be derived immediately from those of generic central subspaces.

The idea of partial dimension reduction is similar to Dawid's (1979) notion of sufficient covariates in experimental design. Dawid said that a set of covariates $\mathbf{U}$ is sufficient if the individual experimental units contain no further information about the response, given the treatment and $\mathbf{U}$. Thinking of $\mathbf{X}$ as characterizing the experimental unit, W as the treatment and $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ as the sufficient covariates, this corresponds exactly to the partial dimension reduction condition in (8); we require the response to be independent of the experimental unit, given the treatment and the sufficient covariates. In particular, we look for the minimal sufficient covariate vector, which is given by $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$.

Marginally (i.e., integrating out W), we have a distribution for $(\mathbf{X}, Y)$ and a CS, $\mathcal{S}_{Y|\mathbf{X}}$, for the marginal regression of Y on $\mathbf{X}$. Conditionally (i.e., within the subpopulations identified by W), we have distributions for $(\mathbf{X}, Y)|(W = w)$ and CS's, $\mathcal{S}_{Y|(\mathbf{X}, W=w)}$, $w = 1, \ldots, C$, for the regression of Y on $\mathbf{X}$ within each subpopulation. For notational simplicity, we will use $(\mathbf{X}_w, Y_w)$ to indicate a generic pair distributed like $(\mathbf{X}, Y)|(W = w)$. Correspondingly, $\mathcal{S}_{Y|(\mathbf{X}, W=w)} = \mathcal{S}_{Y_w|\mathbf{X}_w}$. Since we know how to produce inferences for $\mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}_{Y_w|\mathbf{X}_w}$, $w = 1, \ldots, C$, our next step is to understand the relationships between such spaces and the partial CS, $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$.

Finally, we note for completeness that the notation $\mathcal{S}_{Y|(\mathbf{X}, W)}$ is not defined here because, as mentioned in the Introduction, it is not clear how to perform dimension reduction on $\mathbf{X}$ and W simultaneously.

3.1. *Partial and marginal dimension reduction.* As suggested by comparing (1) and (8), $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ need not coincide with $\mathcal{S}_{Y|\mathbf{X}}$. Although untouched by the dimension reduction exercise, W participates in the informational structure of the regression and thus shapes the conditional independence relation through which the reduction of $\mathbf{X}$ is performed. Depending on the features of the overall joint distribution of $(\mathbf{X}, W, Y)$, $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ and $\mathcal{S}_{Y|\mathbf{X}}$ might overlap in any fashion. We are particularly interested in identifying conditions under which

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} \supseteq \mathcal{S}_{Y|\mathbf{X}} \quad \text{and/or} \quad \mathcal{S}_{Y|\mathbf{X}}^{(W)} \subseteq \mathcal{S}_{Y|\mathbf{X}}.$$

We have the following two results:

PROPOSITION 3.1. *If* $W \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$ *or* $W \perp\!\!\!\perp Y|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$, *then* $\mathcal{S}_{Y|\mathbf{X}}^{(W)} \supseteq \mathcal{S}_{Y|\mathbf{X}}$.

PROPOSITION 3.2. *If* $W \perp\!\!\!\perp Y|\mathbf{X}$, *then* $\mathcal{S}_{Y|\mathbf{X}}^{(W)} \subseteq \mathcal{S}_{Y|\mathbf{X}}$.

In terms of interpretation, if $W \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$ or $W \perp\!\!\!\perp Y|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$ as required by Proposition 3.1, partial dimension reduction does not miss any of the directions that are relevant to the marginal regression of Y on $\mathbf{X}$. Consideration of W while attempting to reduce the $\mathbf{X}$ vector does not make any of these directions superfluous and might lead us to retain some additional ones. On the other hand, if $W \perp\!\!\!\perp Y|\mathbf{X}$ as required by Proposition 3.2, partial dimension reduction does not add to the directions that are relevant to the marginal regression of Y on $\mathbf{X}$. Consideration of W while attempting to reduce the $\mathbf{X}$ vector might make some of these directions superfluous and does not lead us to retain any additional one. Finally, if $W \perp\!\!\!\perp Y|\mathbf{X}$ and $W \perp\!\!\!\perp Y|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$, or if $W \perp\!\!\!\perp Y|\mathbf{X}$ and $W \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$, partial and marginal dimension reduction identify exactly the same space.

It is interesting to notice that $\mathcal{S}_{Y|\mathbf{X}}^{(W)} \supseteq \mathcal{S}_{Y|\mathbf{X}}$ in designed experiments, where W represents a treatment randomly assigned to experimental units characterized by the value of $\mathbf{X}$. In these situations, $\mathbf{X} \perp\!\!\!\perp W$ by design, and therefore $W \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$, so the first of the two sufficient conditions in Proposition 3.1 holds.

3.2. *Partial and conditional dimension reduction.* The following proposition connects the partial CS, $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$, to the CS's $\mathcal{S}_{Y_w|\mathbf{X}_w}$ for Y_w on $\mathbf{X}_w$, $w = 1, \ldots, C$; $\oplus$ indicates the direct sum between two subspaces ($V_1 \oplus V_2 = \{v_1 + v_2;\ v_1 \in V_1,\ v_2 \in V_2\}$).

PROPOSITION 3.3. $\mathcal{S}_{Y|\mathbf{X}}^{(W)} = \bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{X}_w}$.

Although the within-subpopulation spaces $\mathcal{S}_{Y_w|\mathbf{X}_w}$, $w = 1, \ldots, C$, can overlap in any fashion, the partial CS always coincides with their direct sum.

Cook and Critchley [(2000), Proposition 1] established a general relation between the marginal regression of Y on $\mathbf{X}$ and the conditional subpopulation regressions of Y_w on $\mathbf{X}_w$, $w = 1, \ldots, C$:

$$\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{W|\mathbf{X}} \oplus \left\{ \bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{X}_w} \right\}, \tag{9}$$

where $\mathcal{S}_{W|\mathbf{X}}$ is the central subspace for the regression of the qualitative predictor W on $\mathbf{X}$. In effect, $\mathcal{S}_{W|\mathbf{X}}$ carries the "joining information" to connect the individual subpopulation regressions. They also argue that, while it is theoretically possible to have proper containment in (9), in practice we may normally expect equality. Combining (9) with Proposition 3.3 we have

$$\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{W|\mathbf{X}} \oplus \mathcal{S}_{Y|\mathbf{X}}^{(W)}. \tag{10}$$

If $\mathbf{X} \perp\!\!\!\perp W$, then we are led back to the conclusion of Proposition 3.1.

Equation (10) is useful because it suggests that, when equality holds, we can expect the marginal CS $\mathcal{S}_{Y|\mathbf{X}}$ to be larger than the partial CS $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ and that the "difference" is due to the joining regression of W on $\mathbf{X}$. This may then account for the finding of $\hat{d} = 2$ in the regression of Section 2.3: the dimension of $\mathcal{S}_{L|\mathbf{X}}$ may consist of one dimension from $\mathcal{S}^{(G)}_{L|\mathbf{X}} = \bigoplus_{g=1}^{2} \mathcal{S}_{L_g|\mathbf{X}_g}$ and one dimension from $\mathcal{S}_{G|\mathbf{X}}$. If that were the case, then the partial central subspace would be one-dimensional, and a single two-dimensional plot would summarize the regression of L on $(\mathbf{X}, G)$. But we still need a sound method to estimate $\mathcal{S}^{(G)}_{L|\mathbf{X}}$.

Proposition 3.3 suggests that $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ can be estimated by combining dimension reduction within subpopulations. Focusing on a generic subpopulation, indicate with $\boldsymbol{\mu}_w$ and $\boldsymbol{\Sigma}_w$ the mean and covariance of $\mathbf{X}_w$, assume that $\boldsymbol{\Sigma}_w$ is invertible and standardize $\mathbf{X}_w$ to $\mathbf{Z}_w = \boldsymbol{\Sigma}_w^{-1/2}(\mathbf{X}_w - \boldsymbol{\mu}_w)$. From (2), we have $\mathcal{S}_{Y_w|\mathbf{X}_w} = \boldsymbol{\Sigma}_w^{-1/2}\mathcal{S}_{Y_w|\mathbf{Z}_w}$ and therefore

$$\mathcal{S}^{(W)}_{Y|\mathbf{X}} = \bigoplus_{w=1}^{C} \boldsymbol{\Sigma}_w^{-1/2}\mathcal{S}_{Y_w|\mathbf{Z}_w}.$$

This relationship is the starting point to adapt SIR for inference on $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$. We call this method *partial sliced inverse regression*.

4. Partial sliced inverse regression. Here, we introduce the simplifying assumption that the predictors covariance structure is the same across subpopulations:

$$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{\text{pool}}, \qquad w = 1, \ldots, C. \tag{11}$$

Departures from this assumption do not alter the overall logic of partial SIR, but introduce scaling issues affecting spectral decompositions and large sample testing for rank. Also, this assumption may be appropriate in an important subclass of regressions, including designed experiments as discussed previously. If (11) holds, we can rewrite $\mathbf{Z}_w = \boldsymbol{\Sigma}_{\text{pool}}^{-1/2}(\mathbf{X}_w - \boldsymbol{\mu}_w)$ for each subpopulation, and

$$\mathcal{S}^{(W)}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}_{\text{pool}}^{-1/2}\left(\bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{Z}_w}\right).$$

In ordinary SIR we were able to concentrate on $\mathcal{S}_{Y|\mathbf{Z}}$, eventually back-transforming to $\mathcal{S}_{Y|\mathbf{X}}$ through $\boldsymbol{\Sigma}^{-1/2}$. Likewise, we can now concentrate on $\bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{Z}_w}$, eventually back-transforming to $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ through $\boldsymbol{\Sigma}_{\text{pool}}^{-1/2}$. Our next step is to combine subpopulation SIR analyses to recover $\bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{Z}_w}$.

Let $\tilde{Y}_w$ denote a discrete version of Y_w, the response in subpopulation w, and consider $\mathrm{E}(\mathbf{Z}_w|\tilde{Y}_w)$ with its covariance marix

$$\boldsymbol{\Theta}_w = \mathrm{Cov}\big(\mathrm{E}(\mathbf{Z}_w|\tilde{Y}_w)\big) = \mathrm{E}\big(\mathrm{E}(\mathbf{Z}_w|\tilde{Y}_w)\mathrm{E}(\mathbf{Z}_w|\tilde{Y}_w)'\big).$$

Averaging these matrices over subpopulations, we obtain

$$\boldsymbol{\Theta}^{(W)} = \sum_{w=1}^{C} \Pr(W = w)\boldsymbol{\Theta}_w = \mathrm{E}(\boldsymbol{\Theta}_W) = \mathrm{Cov}\big(\mathrm{E}(\mathbf{Z}_W|\tilde{Y}_W)\big), \tag{12}$$

which, in expanded notation for clarity, corresponds to $\mathrm{Cov}(\mathrm{E}(\mathbf{Z}|\tilde{Y}, W))$. The last equality in (12) can be derived as follows:

$$\mathrm{Cov}\big(\mathrm{E}(\mathbf{Z}|\tilde{Y}, W)\big) = \mathrm{Cov}\big\{\mathrm{E}\big(\mathrm{E}(\mathbf{Z}|\tilde{Y}, W)|W\big)\big\} + \mathrm{E}\big\{\mathrm{Cov}\big(\mathrm{E}(\mathbf{Z}|\tilde{Y}, W)\big)|W\big\}.$$

Because of the intraslice centering, $\mathrm{E}(\mathrm{E}(\mathbf{Z}|\tilde{Y}, W)|W) = \mathrm{E}(\mathbf{Z}|W) = 0$ and thus we are left with

$$\mathrm{E}\big\{\mathrm{Cov}\big(\mathrm{E}(\mathbf{Z}|\tilde{Y}, W)\big)|W\big\} = \mathrm{E}\big\{\mathrm{E}\big[\mathrm{E}(\mathbf{Z}|\tilde{Y}, W)\mathrm{E}(\mathbf{Z}|\tilde{Y}, W)'|W\big]\big\} = \mathrm{E}(\boldsymbol{\Theta}_W) = \boldsymbol{\Theta}^{(W)}.$$

Now, if linearity (3) holds within each subpopulation, that is,

$$\mathrm{E}\big(\mathbf{Z}_w|\mathbf{P}_{\mathcal{S}_{Y_w|\mathbf{Z}_w}}\mathbf{Z}_w\big) = \mathbf{P}_{\mathcal{S}_{Y_w|\mathbf{Z}_w}}\mathbf{Z}_w, \qquad w = 1, \ldots, C, \tag{13}$$

then we have $\mathrm{E}(\mathbf{Z}_w|\tilde{Y}_w) \in \mathcal{S}_{\tilde{Y}_w|\mathbf{Z}} \subseteq \mathcal{S}_{Y_w|\mathbf{Z}_w}$, or equivalently $\mathrm{Span}(\boldsymbol{\Theta}_w) \subseteq \mathcal{S}_{Y_w|\mathbf{Z}_w}$ for each w, and therefore

$$\mathrm{Span}\big(\boldsymbol{\Theta}^{(W)}\big) = \bigoplus_{w=1}^{C} \mathrm{Span}(\boldsymbol{\Theta}_w) \subseteq \bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{Z}_w}.$$

Assuming that the coverage condition (5) holds within each subpopulation, which we call *subpopulation coverage*, gives $\bigoplus_{w=1}^{C} \mathrm{Span}(\boldsymbol{\Theta}_w) = \bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{Z}_w}$.

In summary, the population basis for partial SIR is as follows:

PROPOSITION 4.1. *Assume the common covariance condition* (11) *holds. Under linearity* (13) *and subpopulation coverage*, $\boldsymbol{\Theta}^{(W)}$ *as defined in* (12) *has* $\mathrm{Span}(\boldsymbol{\Theta}^{(W)}) = \bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{Z}_w}$.

As in ordinary SIR, we can then reconstruct our space as

$$\mathrm{Span}\big(\boldsymbol{\Theta}^{(W)}\big) = \mathrm{Span}(\mathbf{t}_1, \ldots, \mathbf{t}_d),$$

where $d = \mathrm{rank}(\boldsymbol{\Theta}^{(W)})$, and the $\mathbf{t}_j$'s are eigenvectors corresponding to eigenvalues $\theta_j \neq 0$ in the spectral decomposition

$$\boldsymbol{\Theta}^{(W)} = \sum_{j=1}^{p} \theta_j \mathbf{t}_j \mathbf{t}_j'.$$

Assuming that an iid sample $(\mathbf{X}_i, W_i, Y_i)$, $i = 1, \ldots, n$, from the joint distribution of $(\mathbf{X}, W, Y)$ is available, partial SIR is applied according to the following

algorithm: Form moment estimates $\hat{\boldsymbol{\mu}}_w$, $\hat{\boldsymbol{\Sigma}}_w$ for the mean and covariance of $\mathbf{X}_w$ in each subpopulation, pool the latter in an estimate of the common covariance:

$$\hat{\boldsymbol{\Sigma}}_{\text{pool}} = \sum_{w=1}^{C} \frac{n_w}{n} \hat{\boldsymbol{\Sigma}}_w, \tag{14}$$

where n_w is the number of observations from w. Standardize the predictor to $\hat{\mathbf{Z}}_{iw} = \hat{\boldsymbol{\Sigma}}_{\text{pool}}^{-1/2}(\mathbf{X}_{iw} - \hat{\boldsymbol{\mu}}_w)$, $i = 1, \ldots, n_w$, $w = 1, \ldots, C$. Next, following the usual recommendations for SIR [Li (1991), Cook (1998)], create a system of $s = 1, \ldots, H_w$ slices on the sample range of Y_w within subpopulation w, and calculate the intraslice mean vectors as

$$\bar{\mathbf{Z}}_{sw} = \frac{1}{n_{sw}} \sum_{i|s} \hat{\mathbf{Z}}_{iw}, \qquad s = 1, \ldots, H_w, \ w = 1, \ldots, C,$$

where the sum is over indexes i of response observations Y_{iw} that fall into slice s, and n_{sw} is the number of observations in slice s, for subpopulation w. Since these mean vectors average to 0 over the slices within each subpopulation ($\frac{1}{n_w} \times \sum_{s=1}^{H_w} n_{sw} \bar{\mathbf{Z}}_{sw} = 0$), sample versions of $\boldsymbol{\Theta}_w$ for the various subpopulations are given by

$$\hat{\boldsymbol{\Theta}}_w = \sum_{s=1}^{H_w} \frac{n_{sw}}{n_w} \bar{\mathbf{Z}}_{sw} \bar{\mathbf{Z}}'_{sw}, \qquad w = 1, \ldots, C.$$

Now, a sample version of $\boldsymbol{\Theta}^{(W)}$ can be constructed as

$$\hat{\boldsymbol{\Theta}}^{(W)} = \sum_{w=1}^{C} \frac{n_w}{n} \hat{\boldsymbol{\Theta}}_w$$

and spectrally decomposed as

$$\hat{\boldsymbol{\Theta}}^{(W)} = \sum_{j=1}^{p} \hat{\theta}_j \hat{\mathbf{t}}_j \hat{\mathbf{t}}'_j.$$

The eigenvalues of $\hat{\boldsymbol{\Theta}}^{(W)}$ are used to produce an estimate of the rank $\hat{d}$ (see Section 4.1), while the eigenvectors are used to construct $\text{Span}(\hat{\mathbf{t}}_1, \ldots, \hat{\mathbf{t}}_{\hat{d}})$, which estimates a lower bound under (13), and the whole $\bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{Z}_w}$ under (13) and subpopulation coverage.

Returning to the $\mathbf{X}$-scale, one forms *partial SIR predictors* as $\hat{\mathbf{v}}'_1\mathbf{X}, \ldots, \hat{\mathbf{v}}'_p\mathbf{X}$, where

$$\hat{\mathbf{v}}_j = \hat{\boldsymbol{\Sigma}}_{\text{pool}}^{-1/2} \hat{\mathbf{t}}_j, \qquad j = 1, \ldots, p,$$

and constructs the estimated $(1 + \hat{d})$-dimensional *partial* central view for the regression as a plot of Y versus the first $\hat{d}$ such predictors, with points marked according to subpopulation.

4.1. *Large sample testing for rank of* $\mathbf{\Theta}^{(W)}$. A summary plot of the response versus the first few partial SIR predictors, or a scatter-plot matrix of the response and all p partial SIR predictors, is generally informative. Nevertheless, as with SIR, inference about d is important for full effectiveness in practice. A test statistic of the form

$$T(m) = n \sum_{j=m+1}^{p} \hat{\theta}_j \tag{15}$$

again can be used in an iterative fashion to estimate the rank. The next proposition concerns the asymptotic distribution of $T(d)$ and is proved in detail in the Appendix.

PROPOSITION 4.2. *Assume the common covariance condition* (11) *holds. Let* $d = \text{rank}(\mathbf{\Theta}^{(W)})$, *let* $\mathbf{P}_{\Theta^{(W)}}$ *be the projection on* $\text{Span}(\mathbf{\Theta}^{(W)})$ *and let* $\mathbf{Q}_{\Theta^{(W)}} = \mathbf{I}_p - \mathbf{P}_{\Theta^{(W)}}$. *If, within each subpopulation* $w = 1, \ldots, C$:

(a) $Y_w \perp\!\!\!\perp \mathbf{Z}_w | \mathbf{P}_{\Theta^{(W)}} \mathbf{Z}_w$,
(b) $E(\mathbf{Z}_w | \mathbf{P}_{\Theta^{(W)}} \mathbf{Z}_w) = \mathbf{P}_{\Theta^{(W)}} \mathbf{Z}_w$ *and*
(c) $\text{Cov}(\mathbf{Z}_w | \mathbf{P}_{\Theta^{(W)}} \mathbf{Z}_w) = \mathbf{Q}_{\Theta^{(W)}}$,

then the asymptotic distribution of $T(d)$ *defined in* (15) *is* $\chi^2_{(H-d-C)(p-d)}$, *where* $H = \sum_{w=1}^{C} H_w$ *is the sum of the number of response slices for each subpopulation used in the partial SIR algorithm.*

Again, (a) corresponds to coverage, (b) corresponds to linearity under coverage and (c) is a constant covariance condition required to obtain the asymptotic χ^2. Note also that (b) and (c) are satisfied if $\mathbf{X}_w$ is normal, and therefore $\mathbf{Z}_w \sim N(0, \mathbf{I}_p)$, within subpopulations.

4.2. *More on the lean body mass regression.* Returning to the lean body mass regression introduced in Section 2.3, we used the likelihood methods implemented in the computer program *Arc* [Cook and Weisberg (1999b)] to investigate simultaneous power transformations of the five many-valued predictors so that the conditional distribution of the transformed predictors $\mathbf{X}|(G = g)$ is approximately normal with common covariance matrix for $g = 1, 2$. This led to the log transformation for each of the predictors. This procedure is usually effective for insuring that common covariance, as well as subpopulation linearity and constant covariance, conditions are met to a reasonable approximation.

Next, using five slices within each subpopulation, we applied partial SIR to the regression of L on $(\mathbf{X}, G)$. The first two large sample SIR p-values were 0.000 and 0.326, indicating that the partial central subspace is one-dimensional, as hinted by our previous analysis. The summary plot of L versus the first partial SIR predictor shown in Figure 2 is quite similar to that shown in Figure 1a. The correlation between the first ordinary SIR predictor of Figure 1a and the first partial SIR

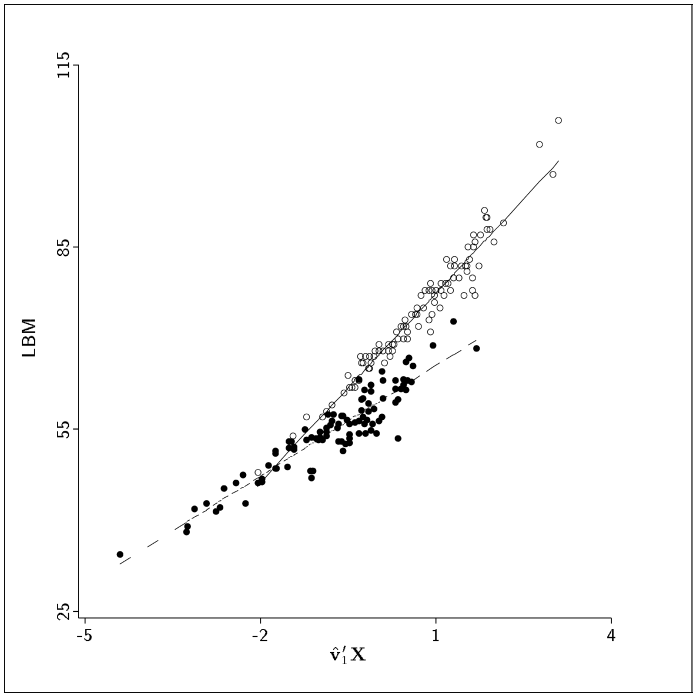


FIG. 2. *Lean body mass L versus the first partial SIR predictor* $\hat{\mathbf{v}}_1'\mathbf{X}$ *from regression of L on* $(\mathbf{X}, G)$; ∘ *males,* • *females.*

predictor of Figure 2 is 0.985, indicating that they are likely estimating the same linear combination of $\mathbf{X}$.

Our partial SIR analysis allows us to conclude that a single linear combination of the predictors is sufficient to describe the regression of L on $\mathbf{X}$ for both males and females. Moreover, it provides us with an estimate of this linear combination, the first partial SIR predictor $\hat{\mathbf{v}}_1'\mathbf{X}$. Using the estimated partial central view in Figure 2 as a guide, we can now develop linear models for the intragender regressions. The view suggests that location may be well captured by a mean function of the form

$$\mathrm{E}(L|\mathbf{X}, G) = \mathrm{E}(L|\mathbf{v}'\mathbf{X}, G) = \big(\alpha_1 + \alpha_2 \operatorname{Ind}(G = 2)\big) + \big(\gamma_1 + \gamma_2 \operatorname{Ind}(G = 2)\big)\mathbf{v}'\mathbf{X},$$

while the variance function $\mathrm{Var}(L|\mathbf{X}, G) = \mathrm{Var}(L|\mathbf{v}'\mathbf{X}, G)$ may depend nontrivially on $\mathbf{v}'\mathbf{X}$, especially for females.

5. Discussion. In this article, we described how the theory of sufficient dimension reduction, and a well-known inference method for it (SIR), can be extended to regression analyses involving both quantitative and categorical predictors. This extension significantly widens the applicative scope of dimension reduction, as a very large number of actual data sets do contain variables of both

types. All assumptions employed in our extension are similar to the ones used in traditional dimension reduction and play similar roles, except for the common covariance assumption (11) used in partial SIR.

The results we presented open the way for a whole new class of theoretical and methodological developments, which are currently under investigation by the authors. First, the common covariance assumption can be abandoned without altering the logic of partial SIR. This requires taking scaling into account when selecting relevant directions and deriving large sample testing for rank, an issue that is not relevant in ordinary SIR, but should be considered when faced with several subpopulations with very different covariance structures for the predictors.

Second, other well-known inference procedures for dimension reduction such as ordinary least squares [Li and Duan (1989)], principal Hessian directions [Li (1992)] and sliced average variance estimation [Cook and Weisberg (1991)] could be adapted to partial dimension reduction, in a fashion similar to the one we detailed for SIR.

Third, partial dimension reduction can be viewed as a form of constrained exercise, in which one considers the informational role of all predictors, but limits reduction to a subset of them. With some modifications, our approach can be employed in cases in which W is not categorical, but is itself a vector of many-valued or continuous variables. This would be of particular interest in applications in which some predictors play a particular role, and must therefore be "shielded" from the reduction process.

Last, the relationships we drew among partial, marginal and conditional central spaces have very interesting connections with (i) the nature of the interaction between $\mathbf{X}$ and W and (ii) theoretical and methodological aspects of dimension reduction for regressions with multiple responses.

Progress on these fronts is likely to generate a significant leap in a branch of statistical theory and methodology (sufficient dimension reduction) whose scope is widening due to the increasing need for effective analysis strategies for high-dimensional data.

The partial SIR method has been implemented as an add-on to *Arc* [Cook and Weisberg (1999b)]. The add-on can be obtained from an Internet address that is available from the authors.

APPENDIX

LEMMA A.1. *For generic random variables V_1, V_2, V_3, V_4, the following equivalences hold*:

$$\begin{aligned} & V_1 \perp\!\!\!\perp V_2 | (V_3, V_4) \quad \textit{and} \quad V_1 \perp\!\!\!\perp V_4 | V_3 \\ \iff \quad & V_1 \perp\!\!\!\perp V_4 | (V_2, V_3) \quad \textit{and} \quad V_1 \perp\!\!\!\perp V_2 | V_3 \\ \iff \quad & V_1 \perp\!\!\!\perp (V_2, V_4) | V_3. \end{aligned}$$

A discussion of this result can be found in Cook [(1998), Chapter 6].

PROOF OF PROPOSITION 3.1. From Lemma A.1 we have that, for a generic subspace $\mathcal{S}$,

$$W \perp\!\!\!\perp \mathbf{X}|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, Y) \quad \text{and} \quad Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}$$
$$\iff \quad Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W) \quad \text{and} \quad W \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}.$$

Thus, under the assumption that $W \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$,

$$Y \perp\!\!\!\perp \mathbf{X}|\big(\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}, W\big) \quad \Longrightarrow \quad Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$$

and therefore $\mathcal{S}_{Y|\mathbf{X}}^{(W)} \supseteq \mathcal{S}_{Y|\mathbf{X}}$. Again using Lemma A.1,

$$Y \perp\!\!\!\perp W|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, \mathbf{X}) \quad \text{and} \quad Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}$$
$$\iff \quad Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W) \quad \text{and} \quad Y \perp\!\!\!\perp W|\mathbf{P}_{\mathcal{S}}\mathbf{X}.$$

Thus, under the assumption that $W \perp\!\!\!\perp Y|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$,

$$Y \perp\!\!\!\perp \mathbf{X}|\big(\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}, W\big) \quad \Longrightarrow \quad Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$$

and therefore $\mathcal{S}_{Y|\mathbf{X}}^{(W)} \supseteq \mathcal{S}_{Y|\mathbf{X}}$. □

PROOF OF PROPOSITION 3.2. Again we use

$$Y \perp\!\!\!\perp W|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, \mathbf{X}) \quad \text{and} \quad Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}$$
$$\iff \quad Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W) \quad \text{and} \quad Y \perp\!\!\!\perp W|\mathbf{P}_{\mathcal{S}}\mathbf{X}$$

as in the Proof of Proposition 3.1, noting that $Y \perp\!\!\!\perp W|(\mathbf{P}_S\mathbf{X}, \mathbf{X})$ is obviously guaranteed by $Y \perp\!\!\!\perp W|\mathbf{X}$. Under such an assumption

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X} \quad \Longrightarrow \quad Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}, W)$$

and therefore $\mathcal{S}_{Y|\mathbf{X}} \supseteq \mathcal{S}_{Y|\mathbf{X}}^{(W)}$. □

PROOF OF PROPOSITION 3.3. It is immediate to see that, for a generic subspace $\mathcal{S}$,

$$(16) \quad Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W) \quad \iff \quad Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W = w) \qquad \forall w = 1, \ldots, C.$$

Since $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ satisfies the left-hand side of (16), it also satisfies

$$Y \perp\!\!\!\perp \mathbf{X}|\big(\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}, W = w\big) \qquad \forall w = 1, \ldots, C.$$

This in turn implies $\mathcal{S}_{Y|\mathbf{X}}^{(W)} \supseteq \mathcal{S}_{Y_w|\mathbf{X}_w}$, $w = 1, \ldots, C$, and therefore

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} \supseteq \bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{X}_w}.$$

Since $\bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{X}_w} \supseteq \mathcal{S}_{Y_w|\mathbf{X}_w}$, $w = 1, \ldots, C$, the sum space satisfies the right-hand side of (16). Hence

$$Y \perp\!\!\!\perp \mathbf{X} | \big(\mathbf{P}_{\bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{X}_w}} \mathbf{X}, W\big),$$

which implies

$$\bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{X}_w} \supseteq \mathcal{S}_{Y|\mathbf{X}}^{(W)}.$$

We can conclude that $\mathcal{S}_{Y|\mathbf{X}}^{(W)} = \bigoplus_{w=1}^{C} \mathcal{S}_{Y_w|\mathbf{X}_w}$. $\square$

PROOF OF PROPOSITION 4.2.

Definitions and outline. Define $\hat{\alpha}_w = n_w/n$, $\hat{a}_w = \sqrt{\hat{\alpha}_w}$, $\hat{\phi}_{sw} = n_{sw}/n_w$ and $\hat{f}_{sw} = \sqrt{\hat{\phi}_{sw}}$. Also, let

$$\bar{\mathbf{Z}}_{\cdot w} = (\hat{f}_{1w}\bar{\mathbf{Z}}_{1w}, \ldots, \hat{f}_{H_w w}\bar{\mathbf{Z}}_{H_w w})$$

be the $p \times H_w$ matrix of weighted slice means for subpopulation $w = 1, \ldots, C$, and let

$$\bar{\mathbf{Z}}_{\cdot\cdot} = (\hat{a}_1\bar{\mathbf{Z}}_{\cdot 1}, \ldots, \hat{a}_C\bar{\mathbf{Z}}_{\cdot C})$$

be the $p \times H$ matrix of weighted subpopulation matrices, where $H = \sum_{w=1}^{C} H_w$. Then $\hat{\mathbf{\Theta}}^{(W)}$ can be expressed as

$$\hat{\mathbf{\Theta}}^{(W)} = \bar{\mathbf{Z}}_{\cdot\cdot}\bar{\mathbf{Z}}_{\cdot\cdot}'.$$

We first investigate the joint asymptotic distribution of the smallest $\min(p - d, H - d)$ singular values of $\bar{\mathbf{Z}}_{\cdot\cdot}$ using the general approach developed by Eaton and Tyler (1994). We then find the asymptotic distribution of $T(d)$ using the fact that the nonzero eigenvalues of $\hat{\mathbf{\Theta}}^{(W)}$ are the squares of the singular values of $\bar{\mathbf{Z}}_{\cdot\cdot}$. The matrix $\bar{\mathbf{Z}}_{\cdot w}$ converges in probability to

$$\mathbf{B}_{\cdot w} = \big(f_{1w}\mathrm{E}(\mathbf{Z}|\tilde{Y} = 1, W = w), \ldots, f_{H_w w}\mathrm{E}(\mathbf{Z}|\tilde{Y} = H_w, W = w)\big),$$

where $\hat{f}_{sw} \xrightarrow{p} f_{sw} = [\Pr(\tilde{Y}_w = s)]^{1/2}$, and $\mathbf{\Theta}_w = \mathbf{B}_{\cdot w}\mathbf{B}_{\cdot w}'$. The matrix $\bar{\mathbf{Z}}_{\cdot\cdot}$ converges in probability to

$$\mathbf{B}_{\cdot\cdot} = (a_1\mathbf{B}_{\cdot 1}, \ldots, a_C\mathbf{B}_{\cdot C}),$$

where $\hat{a}_w \xrightarrow{p} a_w = [\Pr(W = w)]^{1/2}$, and $\boldsymbol{\Theta}^{(W)} = \mathbf{B}_{\cdot\cdot}\mathbf{B}_{\cdot\cdot}'$. Now construct the singular value decomposition of $\mathbf{B}_{\cdot\cdot}$:

$$\mathbf{B}_{\cdot\cdot} = \boldsymbol{\Gamma}' \begin{pmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{pmatrix} \boldsymbol{\Psi},$$

where $\boldsymbol{\Gamma}'$ and $\boldsymbol{\Psi}$ are orthonormal matrices with dimensions $p \times p$ and $H \times H$, and $\mathbf{D}$ is a $d \times d$ diagonal matrix of singular values. Next partition $\boldsymbol{\Gamma}' = (\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_0)$ and $\boldsymbol{\Psi}' = (\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_0)$, where $\boldsymbol{\Gamma}_0$ is $p \times (p-d)$ and $\boldsymbol{\Psi}_0$ is $H \times (H-d)$. Then it follows from Eaton and Tyler (1994) that the asymptotic distribution of the smallest $\min(p-d, H-d)$ singular values of $\sqrt{n}\bar{\mathbf{Z}}_{\cdot\cdot}$ is the same as the asymptotic distribution of the singular values of the $(p-d) \times (H-d)$ matrix

$$\sqrt{n}\mathbf{U} = \sqrt{n}\boldsymbol{\Gamma}_0'(\bar{\mathbf{Z}}_{\cdot\cdot} - \mathbf{B}_{\cdot\cdot})\boldsymbol{\Psi}_0, = \sqrt{n}\boldsymbol{\Gamma}_0'\bar{\mathbf{Z}}_{\cdot\cdot}\boldsymbol{\Psi}_0,$$

where $\mathbf{U}$ is defined implicitly. Thus, the asymptotic distribution of $T(d)$ is the same as that of

$$T = n \times \operatorname{trace}[\boldsymbol{\Gamma}_0'\bar{\mathbf{Z}}_{\cdot\cdot}\boldsymbol{\Psi}_0(\boldsymbol{\Gamma}_0'\bar{\mathbf{Z}}_{\cdot\cdot}\boldsymbol{\Psi}_0)'],$$

which is the sum of the squared singular values of $\sqrt{n}\mathbf{U}$. Also

$$T = n \operatorname{vec}(\mathbf{U})' \operatorname{vec}(\mathbf{U}),$$

where $\operatorname{vec}(\mathbf{U})$ is the $(p-d)(H-d) \times 1$ vector constructed by stacking the columns of $\mathbf{U}$.

Partition $\boldsymbol{\Psi}_0 = (\boldsymbol{\Psi}_{01}', \ldots, \boldsymbol{\Psi}_{0C}')'$, where $\boldsymbol{\Psi}_{0w}$ has dimension $H_w \times (H-d)$. Then, because $\boldsymbol{\Gamma}_0'\mathbf{B}_{\cdot w} = 0$ for all w, we have

$$\begin{aligned} \sqrt{n}\mathbf{U} &= \sqrt{n}\boldsymbol{\Gamma}_0' \sum_{w=1}^{C} (\hat{a}_w \bar{\mathbf{Z}}_{\cdot w} - a_w \mathbf{B}_{\cdot w})\boldsymbol{\Psi}_{0w} \\ &= \sum_{w=1}^{C} \sqrt{n_w}\boldsymbol{\Gamma}_0'(\bar{\mathbf{Z}}_{\cdot w} - \mathbf{B}_{\cdot w})\boldsymbol{\Psi}_{0w} \equiv \sum_{w=1}^{C} \sqrt{n_w}\mathbf{U}_w, \end{aligned} \tag{17}$$

where $\mathbf{U}_w$ is defined implicitly. Because the $\mathbf{U}_w$ are mutually independent, we can investigate the limiting distribution of a typical term $\sqrt{n_w}\,\mathbf{U}_w$ and then add the results.

The rest of the justification is organized as follows: First, using (11), we show that the limiting distribution of $\sqrt{n_w}\operatorname{vec}(\mathbf{U}_w)$ is multivariate normal with mean 0 and covariance matrix $\boldsymbol{\Omega}_w$, and thus that the limiting distribution of $\sqrt{n}\operatorname{vec}(\mathbf{U})$ is normal with mean 0 and covariance matrix $\boldsymbol{\Omega} = \sum_w \boldsymbol{\Omega}_w$. Next, we use conditions (a), (b) and (c) to simplify each $\boldsymbol{\Omega}_w$. We then show that $\boldsymbol{\Omega}$ is an idempotent matrix with rank $(p-d)(H-d-C)$. Consequently, T and thus $T_{\boldsymbol{\Theta}^{(W)}}(d)$ are distributed asymptotically as χ^2 random variables with $(p-d)(H-d-C)$ degrees of freedom.

Limiting distribution of $\sqrt{n_w}\,\mathrm{vec}(\mathbf{U}_w)$. The next step is to express $\sqrt{n_w}\,\mathbf{U}_w$ in terms of the original independent predictors. Working now in terms of $\mathbf{X}$, let

$$\bar{\mathbf{X}}_{\cdot w} = (\bar{\mathbf{X}}_{1w}, \ldots, \bar{\mathbf{X}}_{H_w w})$$

be the $p \times H_w$ matrix of sliced means subpopulation $w = 1, \ldots, C$, let $\mathbf{f}_w$ be the $H_w \times 1$ vector with elements f_{sw}, $s = 1, \ldots, H_w$, and let $\hat{\mathbf{f}}_w$ be the corresponding construction in terms of $\hat{f}_{sw}$. Finally, for any vector $\mathbf{u}$ let $\mathbf{D}_\mathbf{u}$ denote a diagonal matrix with elements $\mathbf{u}$, let $\mathbf{P}_\mathbf{u} = \frac{\mathbf{u}\mathbf{u}'}{\mathbf{u}'\mathbf{u}}$ be the projection on the span of $\mathbf{u}$, and let $\mathbf{Q}_\mathbf{u} = \mathbf{I} - \mathbf{P}_\mathbf{u}$. With these definitions we can write

$$\bar{\mathbf{Z}}_{\cdot w} = \hat{\boldsymbol{\Sigma}}_{\mathrm{pool}}^{-1/2} \bar{\mathbf{X}}_{\cdot w} \mathbf{D}_{\hat{\mathbf{f}}_w} \mathbf{Q}_{\hat{\mathbf{f}}_w},$$

$$\mathbf{B}_{\cdot w} = \boldsymbol{\Sigma}_{\mathrm{pool}}^{-1/2} \mathrm{E}(\bar{\mathbf{X}}_{\cdot w}) \mathbf{D}_{\mathbf{f}_w} \mathbf{Q}_{\mathbf{f}_w}$$

and

$$\begin{aligned}\sqrt{n_w}\,\mathbf{U}_w = \sqrt{n_w}\,\boldsymbol{\Gamma}_0' \big(&\hat{\boldsymbol{\Sigma}}_{\mathrm{pool}}^{-1/2} - \boldsymbol{\Sigma}_{\mathrm{pool}}^{-1/2} + \boldsymbol{\Sigma}_{\mathrm{pool}}^{-1/2}\big) \\ &\times \big(\bar{\mathbf{X}}_{\cdot w} - \mathrm{E}(\bar{\mathbf{X}}_{\cdot w}) + \mathrm{E}(\bar{\mathbf{X}}_{\cdot w})\big) \\ &\times (\mathbf{D}_{\hat{\mathbf{f}}_w} \mathbf{Q}_{\hat{\mathbf{f}}_w} - \mathbf{D}_{\mathbf{f}_w} \mathbf{Q}_{\mathbf{f}_w} + \mathbf{D}_{\mathbf{f}_w} \mathbf{Q}_{\mathbf{f}_w}) \boldsymbol{\Psi}_{0w}.\end{aligned}$$

Expanding this expression and collecting the four $o_p(1)$ terms leaves

$$\begin{aligned}\sqrt{n_w}\,\mathbf{U}_w = {}& \sqrt{n_w}\,\boldsymbol{\Gamma}_0'(\mathbf{A} - \mathbf{I}) \mathbf{B}_{\cdot w} \boldsymbol{\Psi}_{0w} \\ &+ \sqrt{n_w}\,\boldsymbol{\Gamma}_0' \boldsymbol{\Sigma}_{\mathrm{pool}}^{-1/2} \big(\bar{\mathbf{X}}_{\cdot w} - \mathrm{E}(\bar{\mathbf{X}}_{\cdot w})\big) \mathbf{D}_{\mathbf{f}_w} \mathbf{Q}_{\mathbf{f}_w} \boldsymbol{\Psi}_{0w} \\ &+ \sqrt{n_w}\,\boldsymbol{\Gamma}_0' \boldsymbol{\Sigma}_{\mathrm{pool}}^{-1/2} \mathrm{E}(\bar{\mathbf{X}}_{\cdot w}) (\mathbf{D}_{\hat{\mathbf{f}}_w} \mathbf{Q}_{\hat{\mathbf{f}}_w} - \mathbf{D}_{\mathbf{f}_w} \mathbf{Q}_{\mathbf{f}_w}) \boldsymbol{\Psi}_{0w} \\ &+ \sqrt{n_w}\,\boldsymbol{\Gamma}_0' \mathbf{B}_{\cdot w} \boldsymbol{\Psi}_{0w} \\ &+ o_p(1),\end{aligned}$$

where $\mathbf{A} = \hat{\boldsymbol{\Sigma}}_{\mathrm{pool}}^{-1/2} \boldsymbol{\Sigma}_{\mathrm{pool}}^{1/2}$. The first term in this expansion contributes to the distribution of $\sqrt{n_w}\,\mathbf{U}_w$ but not to the distribution of $\sum_{w=1}^{C} \sqrt{n_w}\,\mathbf{U}_w$:

$$\begin{aligned}\boldsymbol{\Gamma}_0'(\mathbf{A} - \mathbf{I}) \sum_{w=1}^{C} \sqrt{n_w}\,\mathbf{B}_{\cdot w} \boldsymbol{\Psi}_{0w} &= \boldsymbol{\Gamma}_0'(\mathbf{A} - \mathbf{I}) \sqrt{n} \sum_{w} a_w \mathbf{B}_{\cdot w} \boldsymbol{\Psi}_{0w} + o_p(1) \\ &= \boldsymbol{\Gamma}_0'(\mathbf{A} - \mathbf{I}) \sqrt{n}\,\mathbf{B}_{\cdot\cdot} \boldsymbol{\Psi}_0 + o_p(1) = o_p(1)\end{aligned}$$

because $\mathbf{B}_{\cdot\cdot} \boldsymbol{\Psi}_0 = 0$ by construction. Next, the third term equals 0: Since $\mathbf{D}_{\mathbf{f}_w} \mathbf{Q}_{\mathbf{f}_w} = (\mathbf{I} - \mathbf{D}_{\mathbf{f}_w} \mathbf{f}_w \mathbf{1}_w') \mathbf{D}_{\mathbf{f}_w}$ we have

$$\boldsymbol{\Gamma}_0' \mathbf{B}_{\cdot w} = \boldsymbol{\Gamma}_0' \boldsymbol{\Sigma}_{\mathrm{pool}}^{-1/2} \mathrm{E}(\bar{\mathbf{X}}_{\cdot w}) (\mathbf{I} - \mathbf{D}_{\mathbf{f}_w} \mathbf{f}_w \mathbf{1}_w') \mathbf{D}_{\mathbf{f}_w} = 0,$$

where $\mathbf{1}_w$ is an $H_w \times 1$ vector of 1's. Thus, multiplying both sides from the right by the nonsingular diagonal matrix $\mathbf{D}_{\mathbf{f}_w}^{-1}$ and rearranging terms yields

$$\boldsymbol{\Gamma}_0' \boldsymbol{\Sigma}_{\text{pool}}^{-1/2} \mathrm{E}(\bar{\mathbf{X}}_{\cdot w}) = \boldsymbol{\Gamma}_0' \boldsymbol{\Sigma}_{\text{pool}}^{-1/2} \mathrm{E}(\bar{\mathbf{X}}_{\cdot w}) \mathbf{D}_{\mathbf{f}_w} \mathbf{f}_w \mathbf{1}_w'.$$

Substituting this into the third term gives the conclusion. Finally, the fourth term is also 0 and thus

$$\sqrt{n_w}\, \mathbf{U}_w = \sqrt{n_w}\, \boldsymbol{\Gamma}_0' \boldsymbol{\Sigma}_{\text{pool}}^{-1/2} \big(\bar{\mathbf{X}}_{\cdot w} - \mathrm{E}(\bar{\mathbf{X}}_{\cdot w})\big) \mathbf{D}_{\mathbf{f}_w} \mathbf{Q}_{\mathbf{f}_w} \boldsymbol{\Psi}_{0w} + o_p(1). \tag{18}$$

Writing a typical column of $\sqrt{n_w}(\bar{\mathbf{X}}_{\cdot w} - \mathrm{E}(\bar{\mathbf{X}}_{\cdot w}))$ as $\hat{f}_{sw}^{-1} \sqrt{n_{sw}} (\bar{\mathbf{X}}_{sw} - \mathrm{E}(\bar{\mathbf{X}}_{sw}))$, and then applying the central limit theorem and the multivariate version of Slutsky's theorem, it follows that, as all $n_{sw} \to \infty$, $\sqrt{n_w}\, \mathrm{vec}(\bar{\mathbf{X}}_{\cdot w} - \mathrm{E}(\bar{\mathbf{X}}_{\cdot w}))$ converges in distribution to a normal random vector with mean 0 and $pH_w \times pH_w$ covariance matrix

$$(\mathbf{D}_{\mathbf{f}_w}^{-1} \otimes \mathbf{I}_p) \mathbf{V}_w^* (\mathbf{D}_{\mathbf{f}_w}^{-1} \otimes \mathbf{I}_p),$$

where $\mathbf{V}_w^*$ is a $pH_w \times pH_w$ block diagonal matrix with diagonal blocks $\mathrm{Cov}(\mathbf{X}_w | \tilde{Y}_w = s)$, $s = 1, \ldots, H_w$. It then follows that $\sqrt{n_w}\, \mathrm{vec}(\mathbf{U}_w)$ converges in distribution to a normal random vector with mean 0 and $(p-d)(H_w - d) \times (p-d)(H_w - d)$ covariance matrix

$$\boldsymbol{\Omega}_w = (\boldsymbol{\Psi}_{0w}' \mathbf{Q}_{\mathbf{f}_w} \otimes \mathbf{I}_{(p-d)}) \mathbf{V}_w (\mathbf{Q}_{\mathbf{f}_w} \boldsymbol{\Psi}_{0w} \otimes \mathbf{I}_{(p-d)}), \tag{19}$$

where $\mathbf{V}_w$ is a $(p-d)H_w \times (p-d)H_w$ block diagonal matrix with diagonal blocks $\boldsymbol{\Gamma}_0' \mathrm{Cov}(\mathbf{Z}_w | \tilde{Y}_w = s) \boldsymbol{\Gamma}_0$, $s = 1, \ldots, H_w$.

It follows immediately from (17) and (19) that, as $n \to \infty$, $\sqrt{n}\mathbf{U}$ converges to a normal random vector with mean 0 and covariance matrix $\boldsymbol{\Omega} = \sum_w \boldsymbol{\Omega}_w$. Before considering $\boldsymbol{\Omega}$, we simplify $\boldsymbol{\Omega}_w$ using conditions in the proposition.

Simplifying $\boldsymbol{\Omega}_w$. The behavior of $\boldsymbol{\Omega}_w$ hinges on the block covariance matrices in $\mathbf{V}_w$:

$$\begin{aligned}
\boldsymbol{\Gamma}_0' \mathrm{Cov}(\mathbf{Z}_w | \tilde{Y}_w) \boldsymbol{\Gamma}_0 &= \mathrm{E}\big[\mathrm{Cov}(\boldsymbol{\Gamma}_0' \mathbf{Z}_w | \boldsymbol{\Gamma}_1' \mathbf{Z}_w, \tilde{Y}_w) | \tilde{Y}_w\big] \\
&\quad + \mathrm{Cov}\big[\mathrm{E}(\boldsymbol{\Gamma}_0' \mathbf{Z}_w | \boldsymbol{\Gamma}_1' \mathbf{Z}_w, \tilde{Y}_w) | \tilde{Y}_w\big] \\
&= \mathrm{E}\big[\mathrm{Cov}(\boldsymbol{\Gamma}_0' \mathbf{Z}_w | \boldsymbol{\Gamma}_1' \mathbf{Z}_w) | \tilde{Y}_w\big] + \mathrm{Cov}\big[\mathrm{E}(\boldsymbol{\Gamma}_0' \mathbf{Z}_w | \boldsymbol{\Gamma}_1' \mathbf{Z}_w) | \tilde{Y}_w\big] \\
&= \mathrm{E}\big[\mathrm{Cov}(\boldsymbol{\Gamma}_0' \mathbf{Z}_w | \boldsymbol{\Gamma}_1' \mathbf{Z}_w) | \tilde{Y}_w\big] \\
&= \mathbf{I}_{(p-d)}.
\end{aligned}$$

The second equality follows from condition (a) in the proposition. The third equality follows from condition (b), noting that $\mathrm{Span}(\boldsymbol{\Gamma}_1) = \mathrm{Span}(\boldsymbol{\Theta}^{(W)})$. The fourth equality follows from condition (c). Thus, $\mathbf{V}_w = \mathbf{I}$ and

$$\boldsymbol{\Omega}_w = \boldsymbol{\Psi}_{0w}' \mathbf{Q}_{\mathbf{f}_w} \boldsymbol{\Psi}_{0w} \otimes \mathbf{I}_{(p-d)}.$$

Calculating $\mathbf{\Omega}$. We have

$$\begin{aligned}\mathbf{\Omega} &= \sum_w \mathbf{\Omega}_w = \sum_{w=1}^{C} \{\mathbf{\Psi}'_{0w}\mathbf{Q}_{\mathbf{f}_w}\mathbf{\Psi}_{0w}\} \otimes \mathbf{I}_{(p-d)} \\ &= \mathbf{I}_{((H-d)(p-d))} - \left\{\sum_{w=1}^{C} \mathbf{\Psi}'_{0w}\mathbf{f}_w\mathbf{f}'_w\mathbf{\Psi}_{0w}\right\} \otimes \mathbf{I}_{(p-d)} \\ &= \mathbf{I}_{((H-d)(p-d))} - \left\{\mathbf{\Psi}'_0 \sum_{w=1}^{C} \mathbf{F}_w\mathbf{F}'_w\mathbf{\Psi}_0\right\} \otimes \mathbf{I}_{(p-d)},\end{aligned}$$

where

$$\mathbf{F}'_w = (\mathbf{0}'_{H_1}, \ldots, \mathbf{0}'_{H_{(w-1)}}, \mathbf{f}'_w, \mathbf{0}'_{H_{(w+1)}}, \ldots, \mathbf{0}'_{H_C})'$$

is an $H \times 1$ vector with $\mathbf{0}_m$ indicating an $m \times 1$ vector of 0's. Because $\mathbf{B}_{\cdot\cdot}\mathbf{F}_w = 0$, $\mathbf{F}_w \in \text{Span}(\mathbf{\Psi}_0)$. Thus, noting that $\|\mathbf{F}_w\| = 1$, $\mathbf{F}_w\mathbf{F}'_w$ is a projection onto a subspace of $\text{Span}(\mathbf{\Psi}_0)$. Because the $\mathbf{F}_w$'s are orthogonal, $\sum_{w=1}^{C} \mathbf{F}_w\mathbf{F}'_w$ is also a projection onto a subspace of $\text{Span}(\mathbf{\Psi}_0)$. With this it follows by straightforward calculation that $\mathbf{\Omega}$ is a symmetric idempotent matrix with trace $(H - d - C)(p - d)$, which is the desired conclusion. □

REFERENCES

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

CHIAROMONTE, F. and COOK, R. D. (2002). Sufficient dimension reduction and graphics in regression. *Ann. Inst. Statist. Math.* To appear.

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

COOK, R. D. and CRITCHLEY, F. (2000). Identifying regression outliers and mixtures graphically. *J. Amer. Statist. Assoc.* **95** 781–794.

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

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

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

COOK, R. D. and WEISBERG, S. (1999b). *Applied Regression Including Computing and Graphics.* Wiley, New York.

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

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

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** 1009–1052.

NEVILL, A. M. and HOLDER, R. L. (1995). Body mass index: A measure of fatness or leanness? *British Journal of Nutrition* **73** 507–516.

SCHOTT, J. (1994). Determining the dimensionality in sliced inverse regression. *J. Amer. Statist. Assoc.* **89** 141–148.

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

F. CHIAROMONTE
DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
326 THOMAS BUILDING
UNIVERSITY PARK, PENNSYLVANIA 16802
E-MAIL: chiaro@stat.psu.edu

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

B. LI
DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
326 THOMAS BUILDING
UNIVERSITY PARK, PENNSYLVANIA 16802
E-MAIL: bing@stat.psu.edu