

# Sufficient dimension reduction in regressions across heterogeneous subpopulations

Liqiang Ni

University of Central Florida, Orlando, USA

and R. Dennis Cook

University of Minnesota, St Paul, USA

[Received April 2004. Final revision September 2005]

**Summary.** Sliced inverse regression is one of the widely used dimension reduction methods. Chiaromonte and co-workers extended this method to regressions with qualitative predictors and developed a method, partial sliced inverse regression, under the assumption that the covariance matrices of the continuous predictors are constant across the levels of the qualitative predictor. We extend partial sliced inverse regression by removing the restrictive homogeneous covariance condition. This extension, which significantly expands the applicability of the previous methodology, is based on a new estimation method that makes use of a non-linear least squares objective function.

**Keywords:** General partial sliced inverse regression; Partial sliced inverse regression; Sliced inverse regression; Sufficient dimension reduction

## 1. Introduction

Consider the regression of a univariate response  $Y$  on a vector of predictors  $\mathbf{X} \in \mathbb{R}^p$ . In full generality, the goal of a regression is to infer about the conditional distribution of  $Y$  given  $\mathbf{X}$ , and many different statistical contexts have been developed to address this issue. In this paper we consider *sufficient dimension reduction*, the basic idea being to replace the predictor vector  $\mathbf{X}$  with its projection  $\mathbf{P}_S \mathbf{X}$  onto a subspace  $S$  of the predictor space without loss of information on  $Y|\mathbf{X}$ . More formally, interest is in subspaces  $S \subseteq \mathbb{R}^p$  with the property that  $Y$  is independent of  $\mathbf{X}$  given any value for  $\mathbf{P}_S \mathbf{X}$ :

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{P}_S \mathbf{X} \quad (1)$$

where ' $\perp\!\!\!\perp$ ' indicates independence. When the intersection of all subspaces satisfying condition (1) itself satisfies condition (1), it is called the *central space* (Cook 1994, 1996) of the regression and is denoted as  $S_{Y|\mathbf{X}}$  with dimension  $d = \dim(S_{Y|\mathbf{X}})$ . The central subspace, which is assumed to exist throughout this paper, is a population metaparameter that can be taken as the parsimonious target of a dimension reduction inquiry. When the central space is known the regression can be limited to  $d \leq p$  new *sufficient predictors*, expressed as linear combinations of the original ones:  $\eta_1^T \mathbf{X}, \dots, \eta_d^T \mathbf{X}$ , where the basis  $\{\eta_1, \dots, \eta_d\}$  for  $S_{Y|\mathbf{X}}$  is often chosen so that the sufficient predictors are uncorrelated. Cook and Weisberg (1999) gave an introductory account of studying regressions via central subspaces.

*Address for correspondence:* Liqiang Ni, Department of Statistics and Actuarial Science, University of Central Florida, Orlando, FL 32816-2370, USA.  
E-mail: lni@mail.ucf.edu

Most methods like sliced inverse regression (SIR) (Li, 1991) and sliced average variance estimation (Cook and Weisberg, 1991) for estimating the central space are limited to regressions with continuous or many-valued predictors because it is in such cases that linear combinations  $\mathbf{P}_{\mathcal{S}}\mathbf{X}$  of the predictors might provide an effective parsimonious summary. Chiaromonte *et al.* (2002) removed this limitation by extending the concept of sufficient dimension reduction to include multiple subpopulations identified by a random qualitative predictor  $W$ . Their extension is based on the idea of a *partial dimension reduction subspace* defined as any subspace  $\mathcal{S}$  that satisfies the conditional independence statement  $Y \perp\!\!\!\perp \mathbf{X} | (\mathbf{P}_{\mathcal{S}}\mathbf{X}, W)$ , and their resulting methodology is called *partial SIR*. In addition to the usual SIR assumptions, partial SIR requires that the conditional predictor variances  $\text{var}(\mathbf{X}|W)$  be homogeneous across subpopulations so that  $\text{var}(\mathbf{X}|W)$  is a constant matrix. Subsequent experience with partial SIR has shown that this *homogeneous covariance condition* is a restriction that should not be neglected in practice. Although partial SIR can be an effective method for pursuing sufficient dimension reduction in the presence of a qualitative predictor when the homogeneous covariance condition holds, it can also be misleading when the condition fails.

In this paper we extend partial SIR so that homogeneous covariances are no longer required, allowing partial SIR to be applied under the same general conditions as SIR. We review SIR and partial SIR in Section 2. We also include a brief illustration to help to fix ideas and set the stage for our extension, called *general partial SIR* (GPSIR), which is presented in Section 3. GPSIR requires a new approach to estimation and consequently a new computational algorithm that is presented in Section 4. In Section 5 we develop an asymptotic test for dimension under GPSIR. Simulation results are presented in Section 6 and an illustrative data analysis is given in Section 7. We conclude with a discussion of related issues in Section 8.

## 2. Sliced inverse regression and partial sliced inverse regression

Let  $\Sigma$  be the non-singular covariance matrix of  $\mathbf{X}$  and define the standardized predictor

$$\mathbf{Z} = \Sigma^{-1/2}\{\mathbf{X} - E(\mathbf{X})\}.$$

Then,  $S_{Y|\mathbf{X}} = \Sigma^{-1/2}S_{Y|\mathbf{Z}}$  and, without loss of generality, we may work on the  $\mathbf{Z}$ -scale, transforming back to the original  $\mathbf{X}$ -scale when necessary. The sample version of  $\mathbf{Z}$  is constructed by substituting the sample mean and covariance matrix for  $E(\mathbf{X})$  and  $\Sigma$ .

### 2.1. Sliced inverse regression

In keeping with the usual SIR protocol, we assume that the response is discrete or has been discretized by constructing  $h$  slices  $H_k$  and using ‘ $Y = k$ ’ to indicate the event that  $Y \in H_k$  with the implied sample space  $\{1, 2, \dots, h\}$ . SIR requires two conditions. The first, called the *linearity condition*, is on the marginal distribution of  $\mathbf{X}$  and not on  $Y|\mathbf{X}$  as is common in regression modelling. It requires that

$$E(\mathbf{Z}|\mathbf{P}_{S_{Y|\mathbf{Z}}}\mathbf{Z}) = \mathbf{P}_{S_{Y|\mathbf{Z}}}\mathbf{Z}.$$

When the linearity condition holds,

$$\text{span}\{E(\mathbf{Z}|Y = y), y = 1, \dots, h\} \subseteq S_{Y|\mathbf{Z}}$$

(Li, 1991). We take this a step further and assume the *coverage condition*

$$\sum_{y=1}^h \text{span}\{E(\mathbf{Z}|Y = y)\} = S_{Y|\mathbf{Z}},$$

which in part requires  $h > d$ . Thus, the subspace that is spanned by the inverse conditional means in the  $\mathbf{Z}$ -scale coincides with the central subspace. Further discussion of the linearity and coverage conditions is given in Section 8.

The linearity and coverage conditions imply that

$$\text{span}[\text{var}\{E(\mathbf{Z}|Y)\}] = S_{Y|\mathbf{Z}}.$$

This population identity is the basis for SIR's spectral decomposition approach to estimation: first, estimate  $E(\mathbf{Z}|Y = y)$  by using the sample mean  $\bar{\mathbf{Z}}_y$  of the sample standardized predictors in slice  $y$ . Then estimate  $\text{var}\{E(\mathbf{Z}|Y)\}$  by using

$$\widehat{\mathbf{M}}_{\text{SIR}} = \sum_{y=1}^h \frac{n_y}{n} \bar{\mathbf{Z}}_y \bar{\mathbf{Z}}_y^T \tag{2}$$

where  $n_y$  is the number of observations in slice  $y$  and  $n = \sum_y n_y$ . If  $d$  is known,  $S_{Y|\mathbf{Z}}$  is estimated by the span of the eigenvectors corresponding to the  $d$  largest eigenvalues of  $\widehat{\mathbf{M}}_{\text{SIR}}$ . Inference on  $d$  can be based on the eigenvalues of equation (2). Additional background on SIR is provided in subsequent sections of this paper. See also Li (1991), Chen and Li (1998) and Cook (1998).

SIR has been applied successfully in numerous applications. But there are also issues remaining. To highlight the main issue that is considered in this paper, we revisit one of the regressions that was discussed by Chiaromonte *et al.* (2002). For  $n = 202$  athletes at the Australian Institute of Sport, consider the regression of lean body mass  $L$  on  $p = 5$  continuous or many-valued predictors, the logarithms of height, weight, red blood cell count, white blood cell count and haemoglobin, represented by  $\mathbf{X}$  and gender indicated by  $W = m$  or  $W = f$ . The question is how to deal with the qualitative predictor  $W$ .

There is concern about interpretation and violation of the linearity condition when considering direct application of SIR to  $(\mathbf{X}, W)$ . To overcome this difficulty, Carroll and Li (1995) investigated dimension reduction through models of the form

$$Y = g\{\beta^T \mathbf{X} + \alpha J(W = m), \varepsilon\}, \tag{3}$$

where  $J$  is the indicator function,  $\beta$  is a  $p$ -dimensional vector,  $\alpha$  is a scalar and  $\varepsilon \perp\!\!\!\perp (\mathbf{X}, W)$  is an error term. Chiaromonte *et al.* (2002) noticed that

- (a) this model limits ‘the effect of  $W$  to an additive shift of  $\beta^T \mathbf{X}$  in the first argument of  $g$ ’, and the interpretation of linear combinations involving qualitative predictors may not be clear,
- (b) the same linear combination  $\beta^T \mathbf{X}$  of the continuous predictors is assumed for both sub-populations and
- (c) the lean body regression does not support a model in form (3).

To avoid these limitations, Chiaromonte *et al.* (2002) introduced a framework that focuses on a projection of  $\mathbf{X}$  that preserves information on  $Y|(\mathbf{X}, W)$ . We review this framework in the next section.

### 2.2. Partial sliced inverse regression

We consider regressions in which a single sample yields observations on  $Y$ ,  $\mathbf{X}$  and a qualitative predictor  $W$  with  $K$  levels represented in generic situations by  $\{1, 2, \dots, K\}$ . A partial dimension reduction subspace  $\mathcal{S}$  satisfies the statement  $Y \perp\!\!\!\perp \mathbf{X} | (\mathbf{P}_{\mathcal{S}} \mathbf{X}, W)$ . If the intersection of all partial

dimension reduction subspaces is itself a partial dimension reduction subspace it is called the *partial central subspace* (PCS) and is denoted as  $\mathcal{S}_{Y|X}^{(W)}$ . If the PCS is known then the regression can again be limited to new sufficient predictors expressed as linear combinations of the original ones:  $\beta^T \mathbf{X} = (\beta_1^T \mathbf{X} \dots \beta_d^T \mathbf{X})^T$ , where now the columns of the matrix  $\beta = (\beta_1 \dots \beta_d)$  form a basis for  $\mathcal{S}_{Y|X}^{(W)}$  and  $d = \dim(\mathcal{S}_{Y|X}^{(W)})$ .

For notational simplicity, we shall follow Chiaromonte *et al.* (2002) and use  $(\mathbf{X}_w, Y_w)$  to indicate a generic pair distributed like  $(\mathbf{X}, Y)|W = w$  so, for example,  $\mathcal{S}_{Y_w|\mathbf{X}_w}$  is the central subspace given  $W = w$  and

$$\mathbf{Z}_w = \Sigma_w^{-1/2} \{\mathbf{X}_w - E(\mathbf{X}_w)\},$$

where  $\Sigma_w = \text{var}(\mathbf{X}_w) > 0$ . The PCS is constructed so that the predictors  $\beta^T \mathbf{X}$  are sufficient for every subpopulation, but they might not all be necessary for any single subpopulation. In particular, Chiaromonte *et al.* (2002) showed that there is a close connection between the *conditional central subspaces*  $\mathcal{S}_{Y_w|\mathbf{X}_w}$  and the PCS:

$$\mathcal{S}_{Y|X}^{(W)} = \sum_{w=1}^K \mathcal{S}_{Y_w|\mathbf{X}_w}. \tag{4}$$

This identity, which requires only the existence of the  $\mathcal{S}_{Y_w|\mathbf{X}_w}$ , suggests that  $\mathcal{S}_{Y|X}^{(W)}$  can be estimated by combining dimension reduction across subpopulations. They used equation (4) to develop *partial SIR* for inference about the PCS by imposing the condition that the subpopulation covariance matrices are constant,  $\Sigma_w = \Sigma_{\text{pool}}$  for all  $w$ . Under this *homogeneous covariance condition*, and assuming essentially that the linearity and coverage conditions hold within each subpopulation, they based their partial SIR estimate of  $\mathcal{S}_{Y|X}^{(W)}$  on the implied identity

$$\mathcal{S}_{Y|X}^{(W)} = \Sigma_{\text{pool}}^{-1/2} \text{span}[\text{var}\{E(\mathbf{Z}_w|Y_w)\}]. \tag{5}$$

A spectral analysis of a sample version of  $\text{var}\{E(\mathbf{Z}_w|Y_w)\} = \text{var}\{E(\mathbf{Z}|Y, W)\}$  can now be used to infer about  $\mathcal{S}_{Y|X}^{(W)}$  in the same way that SIR is used to infer about  $\mathcal{S}_{Y|X}$  by using a spectral analysis of a sample version of  $\text{var}\{E(\mathbf{Z}|Y)\}$ .

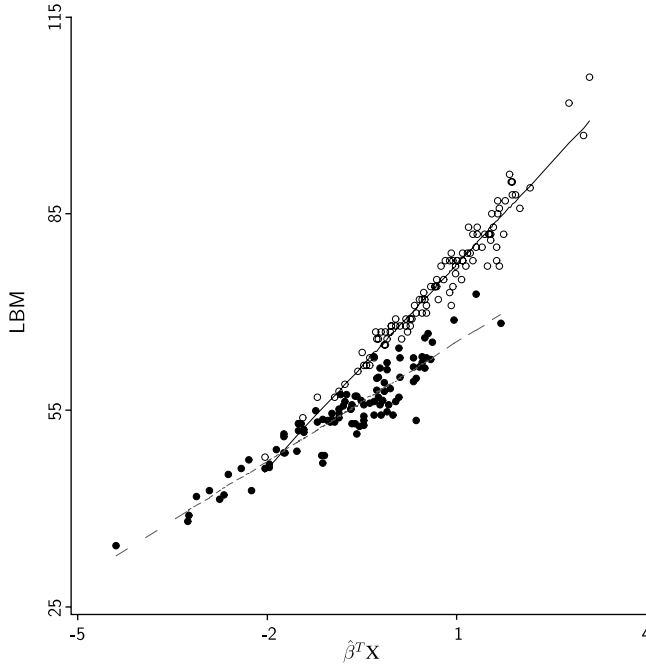
In addition to developing partial SIR methodology, Chiaromonte *et al.* (2002) established a general relationship between  $\mathcal{S}_{Y|X}$  and  $\mathcal{S}_{Y|X}^{(W)}$ :

$$\mathcal{S}_{Y|X} \subseteq \mathcal{S}_{W|X} + \mathcal{S}_{Y|X}^{(W)},$$

where  $\mathcal{S}_{W|X}$  is the central subspace for the regression of  $W$  on  $\mathbf{X}$ . This relationship implies that the marginal central subspace  $\mathcal{S}_{Y|X}$  may often include directions from  $\mathcal{S}_{W|X}$ , which are not of interest in the present context, along with directions from  $\mathcal{S}_{Y|X}^{(W)}$ , which are of interest.

Let us return to the lean body regression. The homogeneous covariance assumption is reasonable with a  $p$ -value of 0.48 obtained from the test of  $\Sigma_m = \Sigma_f$  (Anderson (1984), chapter 10). Using partial SIR as proposed by Chiaromonte *et al.* (2002) we inferred that  $\dim(\mathcal{S}_{Y|X}^{(W)}) = 1$ . A plot of  $L$  versus the estimated sufficient predictor  $\hat{\beta}^T \mathbf{X}$  is shown in Fig. 1. The ordinary least squares (OLS) fits are shown for males and females as visual aids. The interpretation of the plot is that although males and females have different regressions they both depend on one and the same linear combination of the predictors  $\mathbf{X}$ . Also, Fig. 1 suggests that we cannot pass from one gender to the other by adding a constant to  $\hat{\beta}^T \mathbf{X}$ , which rules out model (3).

Partial SIR can be quite effective for dimension reduction when the homogeneous covariance condition is reasonable as in the lean body regression, but it can produce misleading results when the condition fails. Additionally, Chiaromonte *et al.* (2002) found that departures from



**Fig. 1.** Summary plot from application of partial SIR to the lean body mass regression:  $\circ$ , males;  $\bullet$ , females

this condition introduce scaling issues that are not easily resolved in the spectral approach to estimation. But we can construct useful estimates of  $\mathcal{S}_{Y|X}^{(W)}$  without homogeneous covariances by abandoning the pursuit of SIR-type spectral decompositions and basing estimation instead on a non-linear least squares objective function. We describe the new method of estimation, called GPSIR, in the next section and show that it reduces to SIR in the absence of subpopulations and to partial SIR when multiple subpopulations are present and the homogeneous covariance condition holds. In effect, GPSIR allows the logic of partial SIR to be applied under the same conditions as SIR, without the need for the additional condition of homogeneous covariances.

### 3. General partial sliced inverse regression

Suppose that we have a random sample of size  $n$  for  $(\mathbf{X}, Y, W)$  from the total population. There are  $n_w$  points in subpopulation  $w$ , among which  $n_{wy}$  points have  $Y_w = y$ . Let  $p_w = \Pr(W = w)$  and let  $\hat{p}_w = n_w/n$  be the observed fraction for subpopulation  $w$ . Similarly, let  $f_{wy} = \Pr(Y_w = y)$  and let  $\hat{f}_{wy} = n_{wy}/n_w$  denote the corresponding observed fraction. Let  $h_w$  denote the number of slices in subpopulation  $w$  and for consistency with previous notation we let  $h = \sum_w h_w$ . Using notation that is often associated with an analysis of variance, let  $\mathbf{X}_{wyi}$  denote the  $i$ th observation on  $\mathbf{X}$  in slice  $y$  of subpopulation  $w$ , let  $\bar{\mathbf{X}}_{w..}$  be the average in subpopulation  $w$ ,

$$\bar{\mathbf{X}}_{w..} = \frac{1}{n_w} \sum_{y=1}^{h_w} \sum_{i=1}^{n_{wy}} \mathbf{X}_{wyi},$$

and let  $\bar{\mathbf{X}}_{wy.}$  be the average of  $n_{wy}$  points in slice  $y$  of subpopulation  $w$ . Let  $\hat{\Sigma}_w$  denote the sample covariance of  $\mathbf{X}$  in subpopulation  $w$ .

Assuming that the linearity and coverage conditions hold for any subpopulation  $w$ , we have

$$\mathcal{S}_{Y_w|X_w} = \sum_{y=1}^{h_w} \text{span}(\boldsymbol{\xi}_{wy}),$$

where

$$\begin{aligned} \boldsymbol{\xi}_{wy} &= \boldsymbol{\Sigma}_w^{-1/2} E(\mathbf{Z}_w | Y_w = y) \\ &= \boldsymbol{\Sigma}_w^{-1} \{E(\mathbf{X}_w | Y_w = y) - E(\mathbf{X}_w)\}. \end{aligned}$$

A sample version of  $\boldsymbol{\xi}_{wy}$  can be represented as  $\hat{\boldsymbol{\xi}}_{wy} = \hat{\boldsymbol{\Sigma}}_w^{-1} (\bar{\mathbf{X}}_{wy} - \bar{\mathbf{X}}_{w\cdot})$ . It follows from equation (4) that

$$\mathcal{S}_{Y|X}^{(W)} = \sum_{w=1}^K \sum_{y=1}^{h_w} \text{span}(\boldsymbol{\xi}_{wy}). \quad (6)$$

The condition  $h_w > \dim(\mathcal{S}_{Y_w|X_w})$ , which implies that  $h > d$ , helps to avoid violating coverage. Now, recalling that the columns of the  $p \times d$  matrix  $\boldsymbol{\beta}$  form a basis for  $\mathcal{S}_{Y|X}^{(W)}$ , equation (6) implies that for each  $\boldsymbol{\xi}_{wy}$  we can find a vector  $\boldsymbol{\gamma}_{wy}$  so that  $\boldsymbol{\xi}_{wy} = \boldsymbol{\beta}\boldsymbol{\gamma}_{wy}$ . This relationship suggests that a basis for  $\mathcal{S}_{Y|X}^{(W)}$  might be estimated by minimizing an average discrepancy between  $\hat{\boldsymbol{\xi}}_{wy}$  and the estimate of  $\boldsymbol{\beta}\boldsymbol{\gamma}_{wy}$ . For later use, we let  $\boldsymbol{\gamma}_w = (\boldsymbol{\gamma}_{w1}, \dots, \boldsymbol{\gamma}_{wh_w})$  and define the  $d \times h$  matrix  $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_K)$ .

Assuming that the dimension  $d$  of  $\mathcal{S}_{Y|X}^{(W)}$  is known, we propose to estimate  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}_{wy}$  by minimizing the non-linear discrepancy function

$$F_d(\mathbf{B}, \mathbf{C}) \equiv \sum_{w=1}^K \hat{p}_w \sum_{y=1}^{h_w} \hat{f}_{wy} (\hat{\boldsymbol{\xi}}_{wy} - \mathbf{B}\mathbf{C}_{wy})^T \hat{\boldsymbol{\Sigma}}_w (\hat{\boldsymbol{\xi}}_{wy} - \mathbf{B}\mathbf{C}_{wy}) \quad (7)$$

so that

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \arg \min_{\mathbf{B}, \mathbf{C}} \{F_d(\mathbf{B}, \mathbf{C})\} \quad (8)$$

where the minimization is over  $\mathbf{C} \in \mathbb{R}^{d \times h}$  with columns  $\mathbf{C}_{wy}$ , and  $\mathbf{B} \in \mathbb{R}^{p \times d}$  subject to  $\mathbf{B}^T \mathbf{B} = \mathbf{I}$ . We call this method GPSIR.

The discrepancy function  $F_d(\mathbf{B}, \mathbf{C})$  converges almost surely to its population version

$$\begin{aligned} \tilde{F}_d(\mathbf{B}, \mathbf{C}) &\equiv \sum_{w=1}^K p_w \sum_{y=1}^{h_w} f_{wy} (\boldsymbol{\xi}_{wy} - \mathbf{B}\mathbf{C}_{wy})^T \boldsymbol{\Sigma}_w (\boldsymbol{\xi}_{wy} - \mathbf{B}\mathbf{C}_{wy}) \\ &= E(\boldsymbol{\xi}_{WY} - \mathbf{B}\mathbf{C}_{WY})^T \boldsymbol{\Sigma}_W (\boldsymbol{\xi}_{WY} - \mathbf{B}\mathbf{C}_{WY}). \end{aligned}$$

Under the linearity and coverage conditions,  $(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \arg \min \{\tilde{F}_d(\mathbf{B}, \mathbf{C})\}$  so GPSIR provides a Fisher consistent estimate of a basis for  $\mathcal{S}_{Y|X}^{(W)}$ . The minimizers of  $\tilde{F}_d$  are not unique, but that is not a problem since any basis for  $\mathcal{S}_{Y|X}^{(W)}$  will suffice. In the algorithm that is described later in Section 4 we handle the uniqueness issue by imposing orthogonality and length constraints on the columns of  $\hat{\boldsymbol{\beta}}$ .

The GPSIR estimate of  $\mathcal{S}_{Y|X}^{(W)}$  reduces to the partial SIR estimate when the pooled covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{\text{pool}} = \sum_{w=1}^K \hat{p}_w \hat{\boldsymbol{\Sigma}}_w$$

is used in place of  $\hat{\Sigma}_w$  in  $\hat{\xi}_{wy}$  and in the inner product of the discrepancy function. To show this, we replace  $\hat{\Sigma}_w$  with  $\hat{\Sigma}_{\text{pool}}$  in equation (8). Then after a little algebra we find that  $(\hat{\beta}, \hat{\gamma}_{wy})$  can be obtained from the value of  $(\mathbf{B}, \mathbf{C}_{wy})$  that minimizes

$$\sum_{w=1}^K \sum_{y=1}^{h_w} \frac{n_{wy}}{n} (\bar{\mathbf{Z}}_{wy} - \mathbf{B}\mathbf{C}_{wy})^T (\bar{\mathbf{Z}}_{wy} - \mathbf{B}\mathbf{C}_{wy}),$$

where now  $\bar{\mathbf{Z}}_{wy} = \hat{\Sigma}_{\text{pool}}^{-1/2} (\bar{\mathbf{X}}_{wy} - \bar{\mathbf{X}}_{w..})$  and  $\mathbf{B} = \hat{\Sigma}_{\text{pool}}^{1/2} \mathbf{B}$ . After minimizing over  $\mathbf{C}_{wy}$  for a fixed  $\mathbf{B}$ , we have

$$\begin{aligned} \hat{\beta} &= \hat{\Sigma}_{\text{pool}}^{-1/2} \arg \min_{\mathbf{B}} \left\{ \sum_{w,y} \frac{n_{wy}}{n} \|\bar{\mathbf{Z}}_{wy} - \mathbf{B}(\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \bar{\mathbf{Z}}_{wy}\|^2 \right\} \\ &= \hat{\Sigma}_{\text{pool}}^{-1/2} \arg \min_{\mathbf{B}} \{ \text{tr}(\widehat{\mathbf{M}}_{\text{PSIR}} \mathbf{Q}_{\mathbf{B}}) \}, \end{aligned}$$

where  $\mathbf{Q}_{(\cdot)} = \mathbf{I} - \mathbf{P}_{(\cdot)}$  and

$$\widehat{\mathbf{M}}_{\text{PSIR}} = \sum_{w=1}^K \hat{p}_w \sum_{y=1}^{h_w} \hat{f}_{wy} \bar{\mathbf{Z}}_{wy} \bar{\mathbf{Z}}_{wy}^T. \quad (9)$$

is the pooled sample covariance matrix of the slice means, which is the estimate of  $\text{var}\{E(\mathbf{Z}_w | Y_w)\}$  (see equation (5)) that was used by Chiaromonte *et al.* (2002). Thus,  $\hat{\beta} = \hat{\Sigma}_{\text{pool}}^{-1/2} (\hat{\mu}_1, \dots, \hat{\mu}_d)$ , where the  $\hat{\mu}_j$ s are the eigenvectors of  $\widehat{\mathbf{M}}_{\text{PSIR}}$  corresponding to its  $d$  largest eigenvalues. It follows immediately from this result that GPSIR reduces to SIR,  $\widehat{\mathbf{M}}_{\text{SIR}} = \widehat{\mathbf{M}}_{\text{PSIR}}$ , when there is only one subpopulation ( $K = 1$ ). In this case it follows from the results leading to equation (9) that the minimum value  $\hat{F}_d$  of  $F_d$  is

$$\hat{F}_d = \min_{\mathbf{B}} \{ \text{tr}(\widehat{\mathbf{M}}_{\text{SIR}} \mathbf{Q}_{\mathbf{B}}) \} = \sum_{j=d+1}^p \hat{\lambda}_j \quad (10)$$

where  $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p$  are the eigenvalues of  $\widehat{\mathbf{M}}_{\text{SIR}}$  defined in equation (2).

There are two essential tasks left to develop GPSIR as methodology. The first is to describe a numerical algorithm for the minimization in equation (8). The other is to find an appropriate test statistic for dimensionality and its asymptotic distribution. The algorithm is discussed in Section 4. Inference is addressed in Section 5.

#### 4. Algorithm

Like many other dimension reduction methods, SIR and partial SIR adopt a spectral approach based on finding a consistently estimable kernel matrix that spans either  $\mathcal{S}_{Y|X}$  or  $\mathcal{S}_{Y|X}^{(W)}$ . The eigenvectors of the sample kernel matrix ( $\widehat{\mathbf{M}}_{\text{SIR}}$  or  $\widehat{\mathbf{M}}_{\text{PSIR}}$ ) corresponding to its eigenvalues that are inferred to be non-zero in the population form the estimate of the target subspace. However, when we have heterogeneous subpopulations, the minimization of the discrepancy function (8) no longer reduces to a spectral decomposition problem, which is a generalization that allows us to avoid the limitations of the previous approach.

From equation (8), GPSIR is based on the minimization of the discrepancy function

$$\begin{aligned} F_d(\mathbf{B}, \mathbf{C}) &= \sum_{w,y} \frac{n_{wy}}{n} (\hat{\xi}_{wy} - \mathbf{B}\mathbf{C}_{wy})^T \hat{\Sigma}_w (\hat{\xi}_{wy} - \mathbf{B}\mathbf{C}_{wy}) \\ &= \sum_{w,y} (\hat{\mathbf{V}}_{wy} \hat{\xi}_{wy} - \hat{\mathbf{V}}_{wy} \mathbf{B}\mathbf{C}_{wy})^T (\hat{\mathbf{V}}_{wy} \hat{\xi}_{wy} - \hat{\mathbf{V}}_{wy} \mathbf{B}\mathbf{C}_{wy}), \end{aligned} \quad (11)$$

where  $\hat{\mathbf{V}}_{wy} = \{(n_{wy}/n)\hat{\Sigma}_w\}^{1/2}$ . Let  $\tilde{\mathbf{B}}$  denote the  $\mathbf{B}$  that minimizes equation (11). Thus,  $\tilde{\mathbf{B}}$  is an estimate of  $\beta$ . This minimization is a special case of finding the values of  $\mathbf{B}$  and  $\mathbf{C}$  that minimize a generic discrepancy function of the form

$$H(\mathbf{B}, \mathbf{C}) = \sum_{j=1}^h (\alpha_j - \mathbf{S}_j \mathbf{B} \mathbf{C}_j)^T (\alpha_j - \mathbf{S}_j \mathbf{B} \mathbf{C}_j), \quad (12)$$

where  $\mathbf{B} \in \mathbb{R}^{p \times d}$ ,  $\mathbf{C} = (\mathbf{C}_1, \dots, \mathbf{C}_h) \in \mathbb{R}^{d \times h}$ ,  $\alpha_j \in \mathbb{R}^p$  and  $\mathbf{S}_j \in \mathbb{R}^{p \times p}$ . The  $\mathbf{S}_j$ s are positive definite. All  $\alpha_j$  and  $\mathbf{S}_j$  are fixed in the minimization algorithm. The function  $H$  can be minimized by treating it as a separable non-linear least squares problem (see Ruhe and Wedin (1980)). We have separate sets of parameters,  $\mathbf{B}$  and  $\mathbf{C}$  in equation (12). Given  $\mathbf{B}$ , the minimization with respect to  $\mathbf{C}$  is straightforward: we only need to solve  $h$  independent linear regressions of  $\alpha_j$  on  $\mathbf{S}_j \mathbf{B}$ ,  $j = 1, \dots, h$ . In contrast, consider minimizing  $H$  with respect to one column  $\mathbf{b}_k$  of  $\mathbf{B}$ , given  $\mathbf{C}$  and the remaining columns of  $\mathbf{B}$  and subject to the orthogonality constraint  $\mathbf{b}_k^T \mathbf{B}_{(-k)} = 0$ , where  $\mathbf{B}_{(-k)}$  is the matrix that is left after taking away  $\mathbf{b}_k$  from  $\mathbf{B}$ . For this partial minimization problem,  $H$  can be re-expressed as

$$H^*(\mathbf{b}_k) = \sum_{j=1}^h (\alpha_j^{(k)} - c_{jk} \mathbf{S}_j \mathbf{b}_k)^T (\alpha_j^{(k)} - c_{jk} \mathbf{S}_j \mathbf{b}_k),$$

where  $\alpha_j^{(k)} = \alpha_j - \mathbf{S}_j \mathbf{B}_{(-k)} \mathbf{C}_{j(-k)}$ ,  $c_{jk}$  is the  $k$ th element of  $\mathbf{C}_j$  and  $\mathbf{C}_{j(-k)}$  consists of all except the  $k$ th element of  $\mathbf{C}_j$ . For ease of reference, the solution to this partial minimization problem is stated in the following lemma; its justification is sketched in Appendix A.

*Lemma 1.* The argument that minimizes  $H^*(\mathbf{b}_k)$  subject to the constraint  $\mathbf{b}_k^T \mathbf{B}_{(-k)} = 0$  is

$$\hat{\mathbf{b}}_k = \mathbf{W}_2^{-1} (\mathbf{I} - \mathbf{B}_{(-k)} (\mathbf{B}_{(-k)}^T \mathbf{W}_2^{-1} \mathbf{B}_{(-k)})^{-1} \mathbf{B}_{(-k)}^T \mathbf{W}_2^{-1}) \mathbf{W}_1,$$

where  $\mathbf{W}_1 = \sum_{j=1}^h c_{jk} \mathbf{S}_j^T \alpha_j^{(k)}$  and  $\mathbf{W}_2 = \sum_{j=1}^h c_{jk}^2 \mathbf{S}_j^T \mathbf{S}_j$ .

An *alternating least squares method* (Cook and Ni (2005), section 3.3) can be easily adapted for equation (12) by minimizing one column of  $\mathbf{B}$  and  $\mathbf{C}$  in turn. Xia *et al.* (2002), section 2.3, also used similar ideas in the algorithm of minimum average variance estimation.

In the spectral approach of SIR and partial SIR, estimated basis directions are ordered by the eigenvalues of the sample kernel matrix. This algorithm will not necessarily produce an analogous ordering. However, we can construct an ordered basis for  $\text{span}(\tilde{\mathbf{B}})$  with respect to the amount by which directions decrease  $H(\mathbf{B}, \mathbf{C})$ . For example, the most important direction is

$$\hat{\mathbf{b}}_1 = \arg \min_{\mathbf{b}} \left\{ \sum_{j=1}^h (\mathbf{Q}_{\mathbf{S}_j \mathbf{b}} \alpha_j)^T (\mathbf{Q}_{\mathbf{S}_j \mathbf{b}} \alpha_j) \right\},$$

where the minimization is over  $\mathbf{b} \in \text{span}(\tilde{\mathbf{B}})$  with  $\|\mathbf{b}\| = 1$ . The second direction is

$$\hat{\mathbf{b}}_2 = \arg \min_{\mathbf{b}} \left\{ \sum_{j=1}^h (\mathbf{Q}_{\mathbf{S}_j[\mathbf{b}, \hat{\mathbf{b}}_1]} \alpha_j)^T (\mathbf{Q}_{\mathbf{S}_j[\mathbf{b}, \hat{\mathbf{b}}_1]} \alpha_j) \right\},$$

where the minimization is over  $\mathbf{b} \in \text{span}(\tilde{\mathbf{B}})$  with  $\|\mathbf{b}\| = 1$  and  $\mathbf{b}^T \hat{\mathbf{b}}_1 = 0$ , and so on.

There is one key issue remaining, which is how to infer about the dimension of the PCS when using GPSIR. Here we can also benefit from the minimum discrepancy approach because the minimum value can be used to construct a test statistic for dimensionality. The construction of test statistics and inference are the topics of the next section.

## 5. Inference about $d = \dim(S_{Y|X}^{(W)})$ in general partial sliced inverse regression

We saw at the end of Section 3 that, when there is only a single subpopulation ( $K = 1$ ) and  $d = \dim(S_{Y|X})$  is known, minimization of the discrepancy function  $F_d(\mathbf{B}, \mathbf{C})$  (7) results in the SIR estimate of  $S_{Y|X}$ , and the minimum value (10) of  $F_d$  is  $\hat{F}_d = \sum_{j=d+1}^p \hat{\lambda}_j$  where  $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p$  are the eigenvalues of  $\widehat{\mathbf{M}}_{\text{SIR}}$  defined in equation (2). The usual SIR test statistic for testing  $d = m$  versus  $d > m$ , where  $m < p$ , is simply  $n\hat{F}_m$ , with relatively large values resulting in rejection. Assuming that  $\mathbf{X}$  has a multivariate normal distribution and implicitly assuming the coverage condition, Li (1991) proved that the null distribution of  $n\hat{F}_d$  is asymptotically  $\chi^2$  with  $(p-d)(h-d-1)$  degrees of freedom. Bura and Cook (2001) proved that in general  $n\hat{F}_d$  is distributed as a weighted sum of independent  $\chi^2$  random variables and showed how to construct consistent estimates of the weights for use in practice.

There are parallel results for partial SIR. The partial SIR statistic that was proposed by Chiaromonte *et al.* (2002) for the hypothesis  $d = m$  versus  $d > m$  is again proportional to the minimum value of  $F_m: \hat{F}_m = \sum_{j=m+1}^p \hat{\alpha}_j$ , where  $\hat{\alpha}_1 \geq \dots \geq \hat{\alpha}_p$  are the eigenvalues of  $\widehat{\mathbf{M}}_{\text{PSIR}}$  defined in equation (9). Chiaromonte *et al.* (2002) showed under conditions that are implied by having normal predictors that  $n\hat{F}_d$  is asymptotically distributed as a  $\chi^2$  random variable with  $(p-d)(h-d-K)$  degrees of freedom. Although it was not emphasized in the main part of Chiaromonte *et al.* (2002), they also showed in an appendix that in general  $n\hat{F}_d$  is distributed asymptotically as a linear combination of independent  $\chi^2$  random variables.

In SIR and partial SIR applications,  $d$  is often estimated by using a sequence of hypothesis tests with the statistic  $n\hat{F}_m$ : starting with  $m = 0$ , test the hypothesis  $d = m$  versus  $d > m$ . If the test is rejected, increment  $m$  by 1 and test again, stopping with the first non-significant result. This type of procedure is fairly common for estimating the dimension of a subspace (see, for example, Rao (1973), page 556). Since  $n\hat{F}_m$  is a generalized version of the test statistics for SIR and partial SIR, we propose to use it to test the hypothesis  $d = m$  versus  $d > m$  in GPSIR. This requires the asymptotic distribution of  $n\hat{F}_d$  or perhaps a nonparametric alternative. Here we follow the asymptotic route.

### 5.1. Asymptotic distribution of $n\hat{F}_d$ in general partial sliced inverse regression

A little set-up is necessary before we can report the asymptotic distribution of  $n\hat{F}_d$  in GPSIR. Conditioning on subpopulation  $w$  for the time being, define the random variable  $J_{wy}$  to equal 1 if  $Y_w = y$  and 0 otherwise. Given  $w$ ,  $E(J_{wy}) = \Pr(Y_w = y) = f_{wy}$ . Define

$$\varepsilon_{wy} = J_{wy} - f_{wy} - \mathbf{Z}_w^T E(\mathbf{Z}_w J_{wy})$$

to be the population residuals from the OLS fit of  $J_{wy}$  on  $\mathbf{Z}_w$ . Let  $\varepsilon_w = (\varepsilon_{w1}, \dots, \varepsilon_{wh_w})^T$  denote the  $h_w \times 1$  vector of residuals, one for each slice, for a typical observation from subpopulation  $w$ , and let  $\mathbf{f}_w = (f_{w1}, f_{w2}, \dots, f_{wh_w})^T$  and  $\mathbf{D}_{\mathbf{f}_w} \equiv \text{diag}(f_{wy})$  be the  $h_w \times h_w$  diagonal matrix with the elements of  $\mathbf{f}_w$  on the diagonal. With this notation we can now define the following  $ph_w \times ph_w$  covariance matrix for subpopulation  $w$ :

$$\Omega_w = \text{var}(\mathbf{D}_{\mathbf{f}_w}^{-1/2} \varepsilon_w \otimes \mathbf{Z}_w). \quad (13)$$

We then arrange these covariance matrices in a  $ph \times ph$  block diagonal matrix  $\Omega \equiv \text{diag}(\Omega_w)$ , which is one component that we need to describe the asymptotic distribution of  $n\hat{F}_d$ .

We also need the  $ph \times ph$  block diagonal matrix

$$\mathbf{V} \equiv \text{diag}(p_w \mathbf{D}_{\mathbf{f}_w}^{-1} \otimes \Sigma_w) \quad (14)$$

and the  $ph \times (p+h)d$  matrix

$$\Delta \equiv (\nu^T \otimes I_p, I_h \otimes \beta), \tag{15}$$

which is the Jacobian matrix for a vectorized version of the discrepancy function,

$$\Delta = \left( \frac{\partial\{\text{vec}(\mathbf{BC})\}}{\partial\{\text{vec}(\mathbf{B})\}}, \frac{\partial\{\text{vec}(\mathbf{BC})\}}{\partial\{\text{vec}(\mathbf{C})\}} \right),$$

evaluated at  $(\beta, \nu)$ , where  $\mathbf{B} \in \mathbb{R}^{p \times d}$ ,  $\mathbf{C} \in \mathbb{R}^{d \times h}$ ,  $\beta$  is a basis for  $\mathcal{S}_{Y|X}^{(W)}$  as defined previously,  $\nu = \gamma \text{diag}(\mathbf{D}_{f_w})$  and  $\gamma$  is as defined following equation (6). Also,  $\mathbf{V}$  is the inner product matrix for the same vectorized version of  $F_d$ . Finally, letting  $\Phi = \mathbf{V}^{1/2} \Delta$ , the asymptotic distribution of  $n\hat{F}_d$  is given in the following theorem.

*Theorem 1.* Assume that the data  $(\mathbf{X}_i, Y_i, W_i)$ ,  $i = 1, \dots, n$ , are a random sample of  $(\mathbf{X}, Y, W)$ . Let

$$\mathcal{S}_\xi = \sum_{w=1}^K \sum_{y=1}^{h_w} \text{span}(\xi_{wy}),$$

let  $d = \dim(\mathcal{S}_\xi)$  and let  $(\hat{\beta}, \hat{\gamma}) = \arg_{\mathbf{B}, \mathbf{C}} \min\{F_d(\mathbf{B}, \mathbf{C})\}$  as defined previously in equation (8). Then

- (a)  $\text{span}(\hat{\beta})$  is a consistent estimator of  $\mathcal{S}_\xi$  and
- (b) as  $n \rightarrow \infty$

$$n\hat{F}_d \xrightarrow{\mathcal{D}} \sum_{i=1}^{ph} \lambda_i \chi_i^2(1)$$

where  $\{\chi_i^2(1)\}$  are independent  $\chi^2$  random variables each with 1 degree of freedom and  $\{\lambda_1 \geq \dots \geq \lambda_{ph}\}$  are the eigenvalues of  $\mathbf{Q}_\Phi \Omega \mathbf{Q}_\Phi$ .

Theorem 1 is quite general, requiring none of the special conditions that were discussed previously. The value  $\hat{\beta}$  of  $\mathbf{B}$  that minimizes the discrepancy function  $F_d(\mathbf{B}, \mathbf{C})$  always provides a consistent estimate of a basis for  $\mathcal{S}_\xi$ , and this theorem allows us to test a hypothesis about its dimension. However, without some of the special conditions,  $\mathcal{S}_\xi$  might not be a useful population parameter and therefore tests on its dimension might not be of interest.

If the linearity condition holds within subpopulations then  $\mathcal{S}_\xi \subseteq \mathcal{S}_{Y|X}^{(W)}$ . The subspace that is spanned by  $\hat{\beta}$  is still a consistent estimate of  $\mathcal{S}_\xi$ , which is now a subspace of the PCS. In this case we can use theorem 1 to infer about a possibly proper subset of the PCS. If the linearity and coverage conditions hold, then we are back to the main line that was introduced at the beginning of Section 3. In this case, as previously pointed out in equation (6),  $\mathcal{S}_\xi = \mathcal{S}_{Y|X}^{(W)}$ , and we can use theorem 1 to infer about the full PCS. A proof of theorem 1 is given in Appendix A. We next summarize the computations that are necessary to implement the tests that are available as a result of theorem 1.

### 5.2. Computations

To use theorem 1 in practice, we need to replace  $\mathbf{Q}_\Phi \Omega \mathbf{Q}_\Phi$  with a consistent estimate under the null hypothesis. Under the hypothesis  $d = m$ , the  $ph \times (p+h)m$  Jacobian matrix  $\Delta$  can be estimated consistently by substituting the corresponding estimates for  $\beta$  and  $\nu$ :  $\hat{\Delta} = (\hat{\nu}^T \otimes I_p, I_h \otimes \hat{\beta})$ . The remaining unknowns are moments that do not depend on the hypothesis and can be estimated consistently by substituting the usual sample versions. To estimate  $\mathbf{V}$  we use  $\hat{\mathbf{V}} = \text{diag}(\hat{\rho}_w \mathbf{D}_{f_w}^{-1} \otimes \hat{\Sigma}_w)$ . Because  $\varepsilon_w$  contains residuals from the population OLS fit of  $J_{wy}$  on

$\mathbf{Z}_w$  within subpopulation  $w$ , it is uncorrelated with  $\mathbf{Z}_w$ . Consequently,

$$\mathbf{\Omega}_w = (\mathbf{D}_{\mathbf{f}_w}^{-1/2} \otimes I_p) E(\varepsilon_w \varepsilon_w^T \otimes \mathbf{Z}_w \mathbf{Z}_w^T) (\mathbf{D}_{\mathbf{f}_w}^{-1/2} \otimes I_p)$$

which suggests the estimate

$$\hat{\mathbf{\Omega}}_w = (\mathbf{D}_{\hat{\mathbf{f}}_w}^{-1/2} \otimes I_p) \left\{ \frac{1}{n_w} \sum_{j=1}^{n_w} (\hat{\varepsilon}_{wj} \hat{\varepsilon}_{wj}^T \otimes \hat{\mathbf{Z}}_{wj} \hat{\mathbf{Z}}_{wj}^T) \right\} (\mathbf{D}_{\hat{\mathbf{f}}_w}^{-1/2} \otimes I_p)$$

where  $\hat{\mathbf{Z}}_{wj}$  is the sample version of  $\mathbf{Z}_w$  and  $\hat{\varepsilon}_w$  contains residuals from the sample OLS fit of  $J_{wy}$  on  $\hat{\mathbf{Z}}_w$ . These estimates are then substituted to yield an estimate of  $\mathbf{Q}_{\Phi} \mathbf{\Omega} \mathbf{Q}_{\Phi}$  from which sample eigenvalues  $\hat{\lambda}_j$  are obtained. The statistic  $n \hat{F}_m$  is then compared with the percentage points of the distribution of  $\sum_{i=1}^{ph} \hat{\lambda}_i \chi_i^2(1)$  to obtain a  $p$ -value. There is a substantial literature on computing tail probabilities of the distribution of a linear combination of  $\chi^2$  random variables. See Field (1993) for an introduction.

## 6. Simulation results

In this section, we present results from a simulation study to investigate properties of partial SIR and GPSIR. Results are based on two models in which the predictor  $\mathbf{X} \in \mathbb{R}^p$  is sampled from one of two normal populations that are indicated by  $W$ ,  $\mathbf{X}|W \sim \text{normal}(0, \mathbf{\Sigma}_w)$ :

$$Y = \exp\{-(X_1 + X_2 + 2X_3)\} + 0.5\varepsilon, \quad (16)$$

$$Y = \frac{X_1}{0.5 + (X_2 + 1.5)^2} + 0.5\varepsilon, \quad (17)$$

where in each model  $\varepsilon \sim \text{normal}(0, 1)$ . The first is a one-dimensional model and the second is a two-dimensional model. Since  $\mathbf{X}|W$  is normal, the linearity condition required is satisfied. For each sampling configuration we generated half of the sample from each subpopulation and ran 1000 simulations. For simulations with heterogeneous subpopulations, we set  $\mathbf{\Sigma}_1 = I_p$  and generated

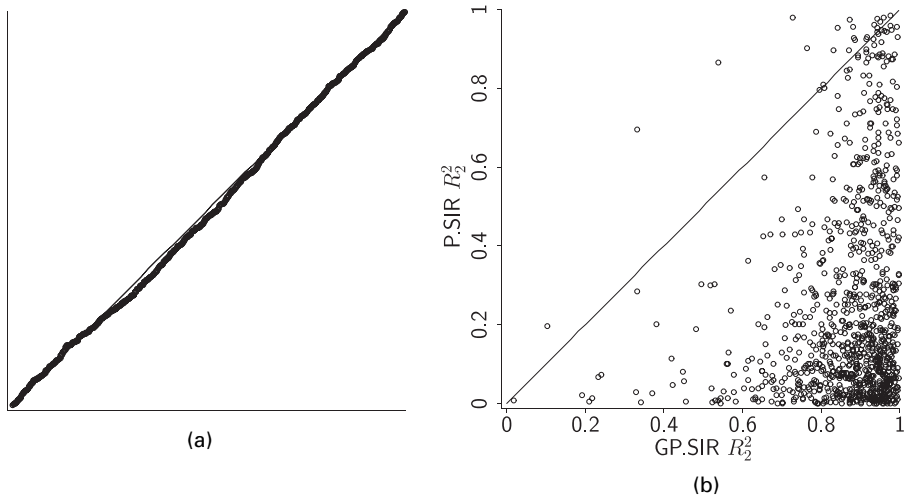
$$\mathbf{\Sigma}_2 = \mathbf{A}_2^T \mathbf{A}_2 / p, \quad (18)$$

where the elements of  $\mathbf{A}_2 \in \mathbb{R}^{p \times p}$  were sampled once from a standard normal population. The difference between  $\mathbf{\Sigma}_w$ s that are generated in this way was easily detected by using the test in Anderson (1984), chapter 10.

SIR is appropriate when  $p = 5$  and  $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = I_5$  since the two subpopulations are identical. Using  $h_w = 4$  slices per subpopulation, our simulations confirmed the asymptotic results that were developed previously. The performances of SIR, partial SIR and GPSIR were very similar. We next consider heterogeneous subpopulations,  $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ , for models (16) and (17), generating  $\mathbf{\Sigma}_w$  as indicated in equation (18) with  $p = 5$ . With unequal covariance matrices, SIR typically responds to directions other than those in the PCS (Cook and Critchley, 2000) and thus it was not included in this part of the study. We set  $h_w = 4$  slices per subpopulation.

### 6.1. Dimension test

At sample size  $n = 400$  we estimated the actual levels of the partial SIR and GPSIR dimension test. Partial SIR fell apart testing  $H_0: d = 1$  for model (16) since its maximum  $p$ -value over the 1000 replications was less than 0.001. It failed again testing  $H_0: d = 2$  for model (17) with its average  $p$ -value being 0.0076. These results support our observation that partial SIR tends to



**Fig. 2.** Model (17): (a) uniform quantile plot of  $p$ -values from GPSIR testing  $H_0 : d = 2$  versus  $H_a : d > 2$  and (b) accuracy of estimation partial SIR and GPSIR

confuse differences between the  $\Sigma_w$ s with the PCS. Meanwhile GPSIR tests held their levels very well. A uniform quantile plot of the  $p$ -values from GPSIR is shown in Fig. 2(a) for model (17); the corresponding plot for model (16) is quite similar.

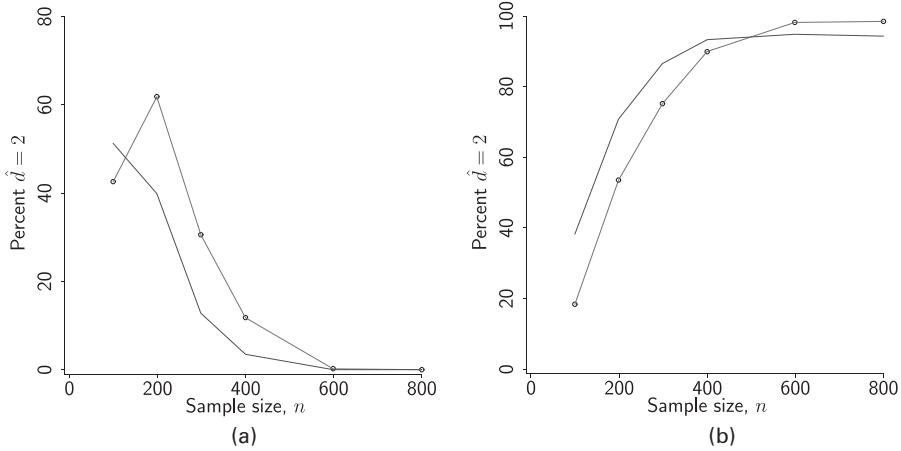
### 6.2. Estimation of $d$

Tests of dimension hypotheses can be of interest on their own but are perhaps most frequently used to estimate  $d$  sequentially, as described in Section 5. Reasoning in the context of model (17) with true dimension  $d = 2$ , if the leading tests of  $d = 0$  and  $d = 1$  have power 1, then all of the estimation error arises from the level  $\alpha$  of the test of  $d = 2$ , resulting in estimates  $\hat{d} = 2$  with probability  $1 - \alpha$  and  $\hat{d} > 2$  with probability  $\alpha$ . This is perhaps the best that we can expect when using the same level for all tests. Clearly, properties of the sequential estimator  $\hat{d}$  of  $d$  depend on the level and power of the dimension tests. Here we examine the percentage of correct dimension decisions by using automated tests with fixed nominal level.

First we consider model (16). Since partial SIR confuses the difference between  $\Sigma_1$  and  $\Sigma_2$  as part of the PCS, it tends to overestimate the dimension. For example, at  $n = 400$  and  $p = 5$ , partial SIR did not make any correct decisions in 1000 replications, whereas GPSIR gave  $\hat{d} = 1$  98.6% of the time with nominal 1%-tests, and 92.8% of the time with nominal 5%-tests. In another set of simulations, we fixed  $n = 800$  and varied  $p$  between 5 and 20, still with 1000 replications of each sampling configuration. Again, partial SIR did not make a single correct decision. Using nominal 1%-tests, GPSIR estimated  $d = 1$  about 99% of the time with  $p = 5$  and about 97% of the time with  $p = 20$ .

Turning to model (17), Fig. 3 shows the percentage of replications in which  $\hat{d} = 2$  with nominal 1%- and 5%-tests. Partial SIR's tendency to overestimate dimension compensated for a weak signal at the smaller sample sizes, which partly explains the relatively high percentage of correct decisions around  $n = 200$ . When the signal becomes stronger with larger sample sizes, partial SIR falls into the pitfall of overestimation again. Meanwhile, the percentage of correct decisions from GPSIR is very close to the best possible even with moderate sample sizes.

In these and other unreported simulations we essentially always found that the estimated true level of a nominal 1%-test was at most 5% when  $n_w > 10p$ . Thus, we recommend the use of



**Fig. 3.** Percentage of correct decisions for model (17) at  $p = 5$  (·····, nominal 1%; —, nominal 5%): (a) partial SIR; (b) GPSIR

nominal 1%-tests when there are at least 10 observations per predictor in each subpopulation. The agreement between the actual and nominal levels seems to improve significantly when  $n_w > 20p$ , although such sample sizes may not be necessary for practically useful results.

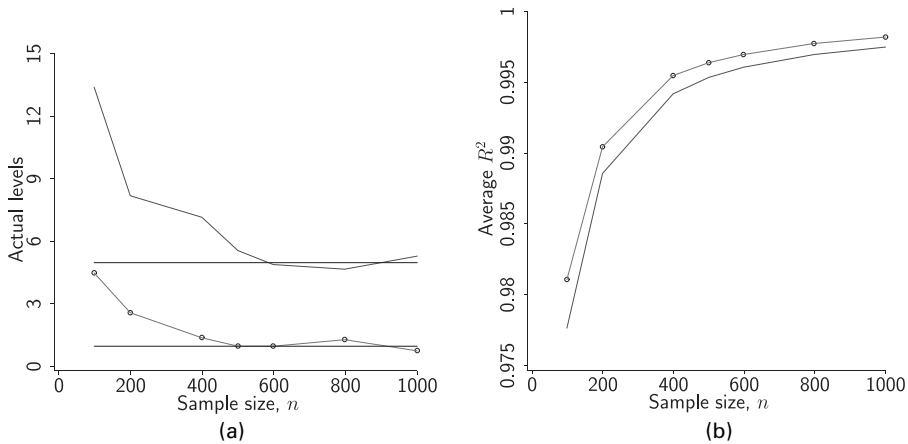
**6.3. Estimation accuracy**

For model (16) we measured the accuracy of estimation through the average values over 1000 replications of  $R^2$  between  $X_1 + X_2 + 2X_3$  and the estimated sufficient predictor  $\hat{\beta}^T \mathbf{X}$ . We already know that with unequal  $\Sigma_w$ s partial SIR can fail to detect the dimension  $d = 1$ . None-the-less, we compared the accuracy of estimation assuming  $d$  to be known, finding that 64.8% of the time  $R^2$  from GPSIR was larger than that from partial SIR. For model (17) we calculated  $R_1^2$  and  $R_2^2$ , the  $R^2$ -values from the regressions of  $X_1$  and  $X_2$  on the two estimated sufficient predictors  $\hat{\beta}_1^T \mathbf{X}$  and  $\hat{\beta}_2^T \mathbf{X}$ . The signal from  $X_1$  is strong in model (17), and the average of  $R_1^2$  was 0.981 from GPSIR and 0.978 from partial SIR. However, the average of GPSIR's  $R_2^2$  was 0.872, which is much larger than partial SIR's average  $R_2^2$  of 0.249. Additionally, GPSIR's  $R_2^2$  exceeded partial SIR's  $R_2^2$  97.8% of the time, as shown in Fig. 2(b).

**6.4. Varying  $n$  and  $p$**

First we consider model (16) with heterogeneous subpopulations and varying sample size  $n$ . For testing  $H_0 : d = 1$  versus  $H_a : d > 1$ , Fig. 4(a) shows the estimated actual levels of GPSIR for nominal 1%- and 5%-tests. The simulation results with increasing sample sizes support the asymptotic calculations and, when judged against the simulation standard errors, suggest that the difference between the nominal and actual levels may not be worrisome in practice. The accuracy of estimation also improves with increasing sample sizes as shown in Fig. 4(b), where  $R^2$  is the squared correlation between  $X_1 + X_2 + 2X_3$  and the estimated sufficient predictor  $\hat{\beta}^T \mathbf{X}$ .

With sample size fixed at  $n = 800$ , we increased  $p$  in model (16) by adding independent standard normal variates. As expected, the actual levels gradually deviate from the nominal levels as  $p$  increases. For example, the observed level of the nominal 1%-test was about 1.2% with five predictors but 2.5% with 20 predictors. The accuracy of estimation also deteriorated. However, the average  $R^2$  for GPSIR was greater than that for partial SIR for all values of  $p$  considered and, even at  $p = 20$ , both average  $R^2$ s were above 0.99.



**Fig. 4.** Model (16) with  $p = 5$ : (a) actual levels from GPSIR ( $\circ$ , nominal 1%;  $\square$ , nominal 5%); (b) accuracy of estimation ( $\circ$ , GPSIR;  $\square$ , partial SIR)

**7. Economic indices**

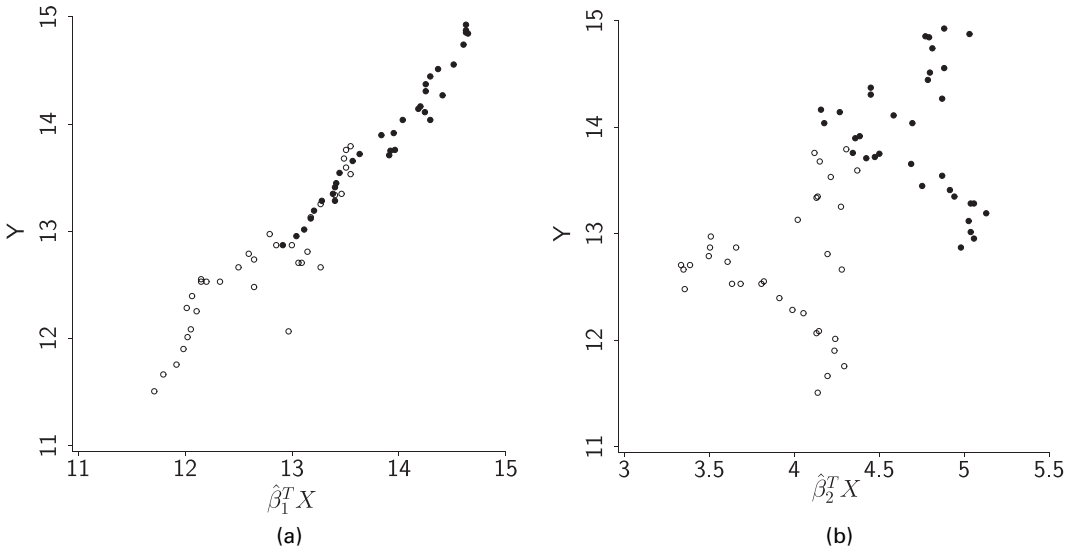
In this section we consider data on relationships between economic indices from O’Donnell *et al.* (2001). The data consist of annual indices from 1960 to 1993 on agricultural production and prices in the state of Oklahoma and the state of Texas, for a total of  $n = 68$  observations. The response  $Y$  is the logarithm of the price index for labour. The five continuous predictors are the logarithms of quantity indices for labour, capital and materials, and the logarithms of price indices for capital and materials. There is one categorical predictor  $W = 1$  for Oklahoma and  $W = 2$  for Texas. We applied partial SIR and GPSIR with  $h = 6$ . On the basis of results in Table 1, GPSIR indicated that  $d = 2$  linear combinations of predictors are needed to characterize the regression, whereas partial SIR clearly indicated that  $d > 2$ . Because the estimated covariance matrices for the two states are quite different, we expect that partial SIR overestimated  $d$ .

Fig. 5 shows plots of  $Y$  versus  $\beta_1^T \mathbf{X}$  and  $\beta_2^T \mathbf{X}$ , where  $\beta_1^T \mathbf{X}$  has a correlation of 0.998 with the simple OLS fit. On the basis of a graphical analysis of the data, we conjectured that the ‘<-’-like pattern for each state in Fig. 5(b) might be caused by a two-phase linear regression, with a changepoint around 1978, 1 year after the Food and Agriculture Act of 1977 was passed by the US Congress. Accordingly, let  $Ind$  indicate that the measurement year is earlier than 1978. Then from an OLS fit of the two-phase linear model

$$Y_W = Ind(\alpha_{1W} + \eta_1^T \mathbf{X}) + (1 - Ind)(\alpha_{2W} + \eta_2^T \mathbf{X}) + \varepsilon \tag{19}$$

**Table 1.** Dimension tests for the data on economic indices

$H_0: d = m$	Results for partial SIR			Results for GPSIR		
	$n\hat{F}_m$	Degrees of freedom	$p$ -value	$n\hat{F}_m$	Trace	$p$ -value
0	98.87	20	0.00	86.61	20.00	0.00
1	46.49	12	0.00	33.94	5.50	0.00
2	16.35	6	0.00	3.34	2.73	0.29



**Fig. 5.** Response versus estimated predictors for the data on economic indices:  $\circ$ , Oklahoma;  $\bullet$ , Texas

we found that the multiple correlation of  $\hat{\eta}_1^T \mathbf{X}$  with  $(\hat{\beta}_1^T \mathbf{X}, \hat{\beta}_2^T \mathbf{X})$  is 0.989 and for  $\hat{\eta}_2^T \mathbf{X}$  the correlation is 0.992, suggesting that  $(\hat{\beta}_1, \hat{\beta}_2)$  and  $(\hat{\eta}_1, \hat{\eta}_2)$  are estimating the same subspace.

### 8. Discussion

This paper covers two advances in sufficient dimension reduction. The first is the removal of the homogeneous covariance condition in partial SIR. This significantly extends the range of application for partial SIR and means that it may now be applied when the usual SIR conditions hold in each of the subpopulations. The second advance is the rederivation of partial SIR in terms of a non-linear least squares discrepancy function. This moves partial SIR closer to standard optimization-based methodology and allows for generalizations that might not have been apparent in the spectral approach. For example, robust versions of partial SIR might be developed by, in part, replacing the least squares objective function with an objective function that is more resistant to outliers.

Like SIR and partial SIR, GPSIR requires the linearity and coverage conditions. Because the linearity condition involves only the marginal distribution of  $\mathbf{X}$ , we are free to use experimental design, one-to-one predictor transformations or reweighting (Cook and Nachtsheim, 1994) to induce the condition when necessary. The linearity condition holds for elliptically contoured predictors. Additionally, Hall and Li (1993) showed that as  $p$  increases with  $d$  fixed the linearity condition holds to a reasonable approximation. For the coverage condition to fail, we must have at least one direction  $\eta \in \mathcal{S}_{Y|\mathbf{X}}^{(W)}$  such that  $E(\eta^T \mathbf{X}_w | Y_w)$  is constant for all subpopulations, and this requires highly symmetric regressions. Even if the coverage condition does not hold, GPSIR estimates a subspace of PCS which still provides relevant information for the regression. A few dimension reduction methods have been proposed to bypass the linearity condition by exploiting local features of the regression (Hristache *et al.*, 2001; Xia *et al.*, 2002). These methods did not address the unique nature of qualitative predictors *per se*, but it will be of interest to incorporate these developments in the framework of partial dimension reduction.

The asymptotic tests for SIR, partial SIR and GPSIR are all based on fixed slices so the expected number of observations per slice grows in proportion to the sample size  $n$ . For any

fixed  $n$ , increasing the number of slices sufficiently can cause nominal characteristics of the asymptotic tests to fail significantly. By keeping the number of slices small, but at least  $d + 1$ , we lose no information in the population while keeping the number of observations per slice relatively large, so the asymptotics might provide a useful approximation.

We assumed throughout this paper that the data are a random sample from  $(\mathbf{X}, Y, W)$ . Theory can be modified straightforwardly to deal with situations in which the fraction  $n_w/n$  of observations from subpopulation  $w$  is fixed by design and held constant as  $n$  grows,  $w = 1, \dots, K$ . However, this modified sampling plan does not result in any changes in the methodology that is discussed here.

A referee suggested that we compare an alternative spectral approach using the kernel matrix

$$\mathbf{M}_{\text{alt}} = \sum_{w=1}^K \hat{p}_w \sum_{y=1}^{h_w} \hat{f}_{wy} \hat{\boldsymbol{\xi}}_{wy} \hat{\boldsymbol{\xi}}_{wy}^T$$

with the approach proposed based on the non-linear objective function (7). This alternative is equivalent to replacing the middle matrix  $\hat{\boldsymbol{\Sigma}}_w$  in expression (7) with the identity, and the asymptotic distribution of the corresponding test statistic for dimension can be found by replacing  $\mathbf{V}$  in equation (14) with  $\text{diag}(p_w \mathbf{D}_{f_w}^{-1} \otimes I_p)$ . However, this approach is not as efficient as GPSIR. Intuitively, the inverse of

$$\text{var}(\boldsymbol{\Sigma}^{-1} \bar{\mathbf{X}}_y) = \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{X|y} \boldsymbol{\Sigma}^{-1}$$

is much closer to  $\boldsymbol{\Sigma}$  than the identity matrix. This could be one of the reasons that the original SIR kernel matrix was constructed in the  $\mathbf{Z}$ -scale, which corresponds to using  $\boldsymbol{\Sigma}$  as the middle matrix for the one-population case. A simple simulation study confirmed this intuition. Consider model (16) with  $\boldsymbol{\Sigma}_1 = I_5$  and  $\boldsymbol{\Sigma}_2 = \text{diag}(1, 1, 1, \sigma, \sigma)$ . For each  $\sigma$ , we generate 200 data sets, each with  $n = 400$ . In Table 2, we show the average squared correlation between the true and estimated sufficient predictor, and the percentage of cases with  $\hat{d} = 1$  based on nominal 5%-tests.

Finally, GPSIR can be applied straightforwardly when there is more than one qualitative predictor in a factorial structure like gender  $\times$  species. In such situations the various combinations of levels can be arranged into a single qualitative variable  $W$  and the methodology applied without modification. However, this procedure will require a large overall sample size if there are many factors, and the results may be difficult to interpret. One way around such difficulties would be to develop an extension that limits consideration to ‘additive effects’, which is similar in spirit to additive analysis-of-variance models.

**Table 2.** Comparison of GPSIR and  $\mathbf{M}_{\text{alt}}$

$\sigma$	Average $R^2$		% $\hat{d} = 1$	
	$\mathbf{M}_{\text{alt}}$	GPSIR	$\mathbf{M}_{\text{alt}}$	GPSIR
0.050	0.259	0.996	0.0	92.5
0.075	0.491	0.996	0.5	91.5
0.10	0.679	0.995	6.0	95.5
0.15	0.932	0.995	14.5	95.5
0.50	0.994	0.995	93.5	90.5
1	0.995	0.995	95.5	96.0

## Acknowledgements

This work was supported in part by grant DMS 04-05360 from the US National Science Foundation. We are grateful to the Joint Editor and a referee for their helpful comments.

## Appendix A

### A.1. Proof of lemma 1

For notational convenience, let  $\mathbf{L} = \mathbf{B}_{(-k)}$ . The minimization with the orthogonality constraint is equivalent to minimizing the function

$$k(\mathbf{g}) = \sum_{j=1}^h (\alpha_j^{(k)} - c_{jk} \mathbf{S}_j \mathbf{Q}_L \mathbf{g})^T (\alpha_j^{(k)} - c_{jk} \mathbf{S}_j \mathbf{Q}_L \mathbf{g})$$

where  $\mathbf{g} \in \mathbb{R}^p$  and  $\mathbf{Q}_L$  is the projection on the orthogonal complement of  $\text{span}(\mathbf{L})$ . Since  $k(\mathbf{g})$  is the sum of squares of residuals of a multivariate OLS fit,

$$\begin{aligned} \hat{\mathbf{g}} &= \arg \min_{\mathbf{g}} \{k(\mathbf{g})\} \\ &= (\mathbf{Q}_L \mathbf{W}_2 \mathbf{Q}_L)^- \mathbf{Q}_L \mathbf{W}_1 \end{aligned}$$

and

$$\begin{aligned} \hat{\mathbf{b}}_k &= \mathbf{Q}_L \hat{\mathbf{g}} \\ &= \mathbf{W}_2^{-1/2} \mathbf{W}_2^{1/2} \mathbf{Q}_L (\mathbf{Q}_L \mathbf{W}_2^{1/2} \mathbf{W}_2^{1/2} \mathbf{Q}_L)^- \mathbf{Q}_L \mathbf{W}_2^{1/2} \mathbf{W}_2^{-1/2} \mathbf{W}_1 \\ &= \mathbf{W}_2^{-1/2} \mathbf{Q}_L \mathbf{W}_2^{-1/2} \mathbf{L} \mathbf{W}_2^{-1/2} \mathbf{W}_1 \\ &= \mathbf{W}_2^{-1} \{I - \mathbf{L}(\mathbf{L}^T \mathbf{W}_2^{-1} \mathbf{L})^- \mathbf{L}^T \mathbf{W}_2^{-1}\} \mathbf{W}_1. \end{aligned} \quad (20)$$

We know that

$$(\mathbf{W}_2^{1/2} \mathbf{Q}_L)^T \mathbf{W}_2^{-1/2} \mathbf{L} = 0$$

and that

$$\text{rank}(\mathbf{W}_2^{1/2} \mathbf{Q}_L) + \text{rank}(\mathbf{W}_2^{-1/2} \mathbf{L}) = p.$$

Therefore,

$$\mathbf{P}_{\mathbf{W}_2^{1/2} \mathbf{Q}_L} + \mathbf{P}_{\mathbf{W}_2^{-1/2} \mathbf{L}} = I_p$$

and the equality in expression (20) holds.

### A.2. Proof of theorem 1

The proof of theorem 1 hinges on Shapiro's (1986) results on the asymptotics of overparameterized discrepancy functions. Propositions 3.1 and 4.1 in Shapiro (1986) are based on a class of discrepancy functions

$$H\{\tau_n, g(\theta)\} = (\tau_n - g(\theta))^T \mathbf{V} (\tau_n - g(\theta)), \quad (21)$$

where  $\tau_n$  is an asymptotically normal estimate of the population value  $g(\theta_0)$  and  $\mathbf{V}$  is a known inner product matrix.

To use Shapiro's results we first write our discrepancy function  $F_d$  defined at equation (7) in the form of the general discrepancy function (21). Using the definitions of  $\xi_{wy}$  and  $\hat{\xi}_{wy}$  that were established at the beginning of Section 3, define  $\zeta_{wy} = f_{wy} \xi_{wy}$  with corresponding sample version  $\hat{\zeta}_{wy} = \hat{f}_{wy} \hat{\xi}_{wy}$ . Define also  $\zeta_w = (\zeta_{w1}, \dots, \zeta_{wh_w})$  and  $\zeta = (\zeta_1, \dots, \zeta_K)$  with corresponding sample versions  $\hat{\zeta}_w$  and  $\hat{\zeta}$ . Then for fixed dimension  $d$  the GPSIR discrepancy function can be written as

$$F_d(\mathbf{B}, \mathbf{C}) = (\text{vec}(\hat{\zeta}) - \text{vec}(\mathbf{BC}))^T \hat{\mathbf{V}} (\text{vec}(\hat{\zeta}) - \text{vec}(\mathbf{BC})) \quad (22)$$

where  $\hat{\mathbf{V}} = \text{diag}(\hat{p}_w \mathbf{D}_{\hat{\mathbf{r}}_w}^{-1} \otimes \hat{\Sigma}_w)$ , as defined previously in Section 5.2. The argument  $\mathbf{C}$  that is used here corresponds to the argument  $\mathbf{C}$  in equation (8) times  $\text{diag}(\mathbf{D}_{\hat{\mathbf{r}}_w})$ . The same relationship holds between  $\nu$  and

$\gamma$  mentioned following equation (15):  $\nu = \gamma \text{diag}(\mathbf{D}_{f_w})$ . The argument  $\mathbf{B}$  that stands for a basis of  $\mathcal{S}_{Y|X}^{(W)}$  is the same in both versions. That  $F_d$  is a version of Shapiro’s discrepancy function  $H$  can now be seen by setting

$$\begin{aligned} \theta &= \begin{pmatrix} \text{vec}(\mathbf{B}) \\ \text{vec}(\mathbf{C}) \end{pmatrix} \in \mathbb{R}^{d(p+h)}, \\ g(\theta) &= \text{vec}(\mathbf{BC}) \in \mathbb{R}^{ph}, \\ \tau_n &= \text{vec}(\hat{\zeta}), \\ g(\theta_0) &= \text{vec}(\beta\nu), \end{aligned}$$

where  $\beta \in \mathbb{R}^{p \times d}$  is in general a basis for  $\mathcal{S}_\zeta$  and  $\nu \in \mathbb{R}^{d \times h}$ . With these associations it is straightforward to verify that the Jacobian matrix  $\Delta = (\nu^T \otimes I_p, I_h \otimes \beta)$  as defined previously in equation (15).

The inner product matrix in Shapiro (1986) is assumed to be known, whereas the inner product matrix in  $F_d$  is estimated. However, it can be shown that since  $\hat{\mathbf{V}}$  converges to  $\mathbf{V}$  in probability, where  $\mathbf{V}$  is as defined in equation (14), the asymptotic distribution of  $n\hat{F}_d$  is the same whether we use  $\mathbf{V}$  or  $\hat{\mathbf{V}}$  as the inner product matrix. It is easy to verify all other conditions of Shapiro (1986), except for asymptotic normality. The strategy to showing asymptotic normality is to decompose  $n^{1/2}\{\text{vec}(\hat{\zeta}) - \text{vec}(\beta\nu)\}$  as a summation of independent and identically distributed observations plus a remainder converging to 0 in probability. Then, by the central limit theorem, we obtain the conclusion.

We next focus on a generic  $w$ th population. For notational simplicity, we drop  $w$  from the subscripts and will restore it when we reach the conclusion. The subscript  $y$  still denotes a slice in the subpopulation. Define  $h$  random variables  $J_y$  such that  $J_y$  equals 1 if the points is in the  $y$ th slice and 0 otherwise,  $y = 1, 2, \dots, h$ . Define a random vector  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_h)^T$ , where

$$\varepsilon_y = J_y - E(J_y) - \mathbf{Z}^T E(\mathbf{Z}J_y).$$

Cook and Ni (2005), page 425, showed that

$$n^{1/2}\{\text{vec}(\hat{\zeta}) - \text{vec}(\zeta)\} = n^{-1/2} \sum_{j=1}^n \text{vec}(\Sigma^{-1/2} \mathbf{Z}_j \varepsilon_j^T) + O_p(n^{-1/2}),$$

where  $(\mathbf{Z}_j, \varepsilon_j)$  are independent and identically distributed observations. Restoring  $w$  in subscripts we have

$$n_w^{1/2}\{\text{vec}(\hat{\zeta}_w) - \text{vec}(\beta\nu_w)\} \rightarrow \text{normal}[0, \text{var}\{\text{vec}(\Sigma_w^{-1/2} \mathbf{Z}_w \varepsilon_w^T)\}].$$

By Slutsky’s theorem, we conclude that  $n^{1/2}\{\text{vec}(\hat{\zeta}) - \text{vec}(\beta\nu)\}$  converges to a normal vector with mean 0 and covariance matrix

$$\Gamma = \text{diag} \left[ \frac{1}{p_w} \text{var}\{\text{vec}(\Sigma_w^{-1/2} \mathbf{Z}_w \varepsilon_w^T)\} \right].$$

It now follows from Shapiro (1986) that the asymptotic distribution of  $n\hat{F}_d$  is the same as that of  $\|\mathbf{Q}_\Phi \mathbf{V}^{1/2} \mathbf{W}\|^2$  where  $\mathbf{W}$  is normal with mean 0 and covariance matrix  $\Gamma$  and  $\Phi = \mathbf{V}^{1/2} \Delta$  as defined for the statement of theorem 1. Consequently,  $n\hat{F}_d$  is asymptotically distributed as a linear combination of independent  $\chi^2$  random variables each with 1 degree of freedom. The coefficients of the  $\chi^2$ -variables are the eigenvalues of

$$\mathbf{Q}_\Phi \mathbf{V}^{1/2} \Gamma \mathbf{V}^{1/2} \mathbf{Q}_\Phi = \mathbf{Q}_\Phi \Omega \mathbf{Q}_\Phi,$$

where  $\Omega = \mathbf{V}^{1/2} \Gamma \mathbf{V}^{1/2}$  is as defined for the statement of theorem 1. Finally, consistency follows from Shapiro (1986) and the fact that  $\hat{\mathbf{V}}$  converges to  $\mathbf{V}$  in probability.

## References

Anderson, T. W. (1984) *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.  
 Bura, E. and Cook, R. D. (2001) Extending sliced inverse regression: the weighted chi-squared test. *J. Am. Statist. Ass.*, **96**, 996–1003.  
 Carroll, R. J. and Li, K. C. (1995) Binary regressors in dimension reduction models: a new look at treatment comparisons. *Statist. Sin.*, **5**, 667–688.  
 Chen, C.-H. and Li, K. C. (1998) Can SIR be as popular as multiple linear regression? *Statist. Sin.*, **8**, 289–316.

- Chiaromonte, F., Cook, R. D. and Li, B. (2002) Sufficient dimension reduction in regressions with categorical predictors. *Ann. Statist.*, **30**, 475–497.
- Cook, R. D. (1994) On the interpretation of regression plots. *J. Am. Statist. Ass.*, **89**, 177–189.
- Cook, R. D. (1996) Graphics for regressions with a binary response. *J. Am. Statist. Ass.*, **91**, 983–992.
- Cook, R. D. (1998) *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.
- Cook, R. D. and Critchley, F. (2000) Identifying regression outliers and mixtures graphically. *J. Am. Statist. Ass.*, **95**, 781–794.
- Cook, R. D. and Nachtsheim, C. J. (1994) Reweighting to achieve elliptically contoured covariates in regression. *J. Am. Statist. Ass.*, **89**, 592–599.
- Cook, R. D. and Ni, L. (2005) Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J. Am. Statist. Ass.*, **100**, 410–428.
- Cook, R. D. and Weisberg, S. (1991) Discussion on ‘Sliced inverse regression for dimension reduction’ (by K.-C. Li). *J. Am. Statist. Ass.*, **86**, 328–332.
- Cook, R. D. and Weisberg, S. (1999) *Applied Regression including Computing and Graphics*. New York: Wiley.
- Field, C. (1993) Tail areas of linear combinations of chi-squares and non-central chi-squares. *J. Statist. Computn Simuln*, **45**, 243–248.
- Hall, P. and Li, K.-C. (1993) On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.*, **21**, 867–889.
- Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2001) Structure adaptive approach for dimension reduction. *Ann. Statist.*, **29**, 1537–1566.
- Li, K.-C. (1991) Sliced inverse regression for dimension reduction. *J. Am. Statist. Ass.*, **86**, 316–342.
- O’Donnell, C. J., Rambaldi, A. N. and Doran, H. E. (2001) Estimating economic relationships subject to firm- and time-varying equality and inequality constraints. *J. Appl. Econometr.*, **16**, 709–726.
- Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*, 2nd edn. New York: Wiley.
- Ruhe, A. and Wedin, P. A. (1980) Algorithms for separable nonlinear least squares problems. *SIAM Rev.*, **22**, 318–337.
- Shapiro, A. (1986) Asymptotic theory of overparameterized structural models. *J. Am. Statist. Ass.*, **81**, 142–149.
- Xia, Y., Tong, H., Li, W. K. and Zhu, L.-X. (2002) An adaptive estimation of dimension reduction space (with discussion). *J. R. Statist. Soc. B*, **64**, 363–410.