Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 1961–1978
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Optimal sufficient dimension reduction in regressions with categorical predictors

Xuerong Wen[a,*], R. Dennis Cook[b,1]

[a] *Department of Mathematics and Statistics, University of Missouri, Rolla, MO 65409, USA*
[b] *School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA*

Received 14 June 2005; received in revised form 3 May 2006; accepted 16 May 2006
Available online 24 August 2006

Abstract

Though partial sliced inverse regression (partial SIR: Chiaromonte et al. [2002. Sufficient dimension reduction in regressions with categorical predictors. Ann. Statist. 30, 475–497]) extended the scope of sufficient dimension reduction to regressions with both continuous and categorical predictors, its requirement of homogeneous predictor covariances across the subpopulations restricts its application in practice. When this condition fails, partial SIR may provide misleading results. In this article, we propose a new estimation method via a minimum discrepancy approach without this restriction. Our method is optimal in terms of asymptotic efficiency and its test statistic for testing the dimension of the partial central subspace always has an asymptotic chi-squared distribution. It also gives us the ability to test predictor effects. An asymptotic chi-squared test of the conditional independence hypothesis that the response is independent of a selected subset of the continuous predictors given the remaining predictors is obtained.

Keywords: Inverse regression; Minimum discrepancy approach; Partial central subspace; Partial SIR

1. Introduction

For a typical regression problem with a scalar response Y and a vector of random predictors $\mathbf{X} \in \mathbb{R}^p$, the goal is to understand how the conditional distribution of $Y|\mathbf{X}$ depends on the value of $\mathbf{X}$. The spirit of sufficient dimension reduction (Cook, 1994, 1998) is to reduce the dimension of $\mathbf{X}$ without loss of information on the original regression and without requiring a pre-specified parametric model. The basic idea is to replace $\mathbf{X}$ by a minimal set of linear combinations of $\mathbf{X}$ without loss of information on $Y|\mathbf{X}$. These linear combinations of $\mathbf{X}$ are called the *sufficient predictors*. More formally, we seek subspaces $\mathscr{S} \subseteq \mathbb{R}^p$ such that

$$Y \perp\!\!\!\perp \mathbf{X} | P_{\mathscr{S}}\mathbf{X},$$

where $\perp\!\!\!\perp$ indicates independence, and $P_{(\cdot)}$ stands for a projection operator with respect to the standard inner product. Such an $\mathscr{S}$ is called a *dimension reduction subspace*. When the intersection of all dimension reduction subspaces itself

* Corresponding author.
E-mail address: wenx@umr.edu (X. Wen).
[1] This work was supported in part by National Science Foundation Grant DMS-0405360.

0378-3758/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2006.05.008

is also a dimension reduction subspace, it is called the *central subspace* (CS; Cook, 1994, 1998) of the regression and denoted as $\mathscr{S}_{Y|\mathbf{X}}$. The dimension of $\mathscr{S}_{Y|\mathbf{X}}$ is called the *structural dimension* of the regression. A regression with a structural dimension of m is called an mD regression.

Although the CS does not always exist, it does exist for a wide class of regressions (Cook, 1998), and then it is uniquely defined. We assume that the CS exists throughout this article. To ease exposition, we often work with the standardized predictors $\mathbf{Z}$ instead of $\mathbf{X}$, where $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - E(\mathbf{X}))$, $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}) > 0$. Since $\mathscr{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathscr{S}_{Y|\mathbf{Z}}$, this involves no loss of generality.

Sufficient dimension reduction may provide an effective starting point for regression analyses, since it can facilitate data visualization. The structural dimension is at most three in many applications and this allows a fully informative and direct visualization of the original regression through a plot of Y versus the sufficient predictors. In this sense, sufficient dimension reduction provides a foundation for regression graphics, as argued by Cook (1998) and Chiaromonte and Cook (2002). And unlike other nonparametric approaches, sufficient dimension reduction can often avoid the curse of dimensionality (Friedman, 1994). Though sufficient dimension reduction does require assumptions on the marginal distribution of $\mathbf{X}$, these are mild, can usually be induced in practice, and are certainly much less restrictive than requiring a model.

There are two general approaches to estimating the CS, the *spectral decomposition approach* and the *minimum discrepancy approach*. Many methods, including the two most popular ones, *sliced inverse regression* (SIR; Li, 1991) and *sliced average variance estimation* (Cook and Weisberg, 1991; Cook, 2000), take the first approach using the following logic. First, find a symmetric population kernel matrix $\mathbf{M}$, which satisfies the property that $\mathrm{Span}(\mathbf{M}) \subseteq \mathscr{S}_{Y|\mathbf{Z}}$. Then, spectrally decompose $\widehat{\mathbf{M}}$, a consistent estimate of $\mathbf{M}$, and use the span of eigenvectors corresponding to the $\dim(\mathscr{S}_{Y|\mathbf{Z}})$ largest eigenvalues of $\widehat{\mathbf{M}}$ to estimate $\mathrm{Span}(\mathbf{M})$. The eigenvalues provide a test statistic for hypotheses on the structural dimension. This is called the spectral decomposition approach since it is based on a spectral decomposition of the sample kernel matrix $\widehat{\mathbf{M}}$. Recently, Cook and Ni (2005) introduced an innovative method to estimate the CS via a minimum discrepancy approach. They developed a family of dimension reduction methods by minimizing quadratic inference functions. An optimal member of this family, the *inverse regression estimator* was proposed. They also showed that many current methods like SIR belong to a sub-optimal class of this family.

Over the past decade sufficient dimension reduction has been mostly limited to regressions with continuous or many-valued predictors because in such cases that the linear combinations $P_{\mathscr{S}}\mathbf{X}$ of the predictors might provide an effective parsimonious summary. Chiaromonte, Cook and Li (2002, hereinafter CCL) introduced the *partial central subspace* (partial CS), to facilitate dimension reduction in regressions with both continuous and categorical predictors. The partial CS is defined as the intersection of all subspaces $\mathscr{S}$ satisfying

$$Y \perp\!\!\!\perp \mathbf{X} \mid (P_{\mathscr{S}}\mathbf{X}, W), \tag{1}$$

where $W \in \{1, \ldots, K\}$ is a categorical predictor. The partial CS, which is assumed to exist and is denoted as $\mathscr{S}^{(W)}_{Y|\mathbf{X}}$, allows for reduction of the vector $\mathbf{X}$ of continuous predictors simultaneously across all subpopulations determined by W. CCL developed a method called *partial SIR* to estimate the partial CS under various mild conditions and the relatively restrictive condition that $\mathrm{Cov}(\mathbf{X}|W)$ is a nonrandom matrix. Experience has shown that this *homogeneous covariance condition* restricts application of partial SIR in practice, and that its failure can result in misleading conclusions.

In this article, we will combine the ideas from CCL and Cook and Ni (2005), taking the minimum discrepancy approach to develop an asymptotically optimal method to estimate the partial CS without the homogeneous covariance constraint. Additionally, because the new method is optimal in a sense described later, we expect that it will prove to be superior to partial SIR even when the homogeneous covariance condition holds. In Section 2, we review inverse regression and the minimum discrepancy approach. The new method, *optimal partial inverse regression estimation* (OPIRE), is proposed in Section 3, where we discuss optimality and computation, and show that this approach allows simple asymptotic chi-squared tests of hypotheses on the dimension of the partial CS. We also re-derive partial SIR via the minimum discrepancy approach. We present asymptotic chi-squared tests for testing various hypotheses about the effects of the continuous predictors in Section 4. Representative simulation results are reported in Section 5 and a final discussion is given in Section 6. To keep the flow of the discussion, details of the proofs appear in the Appendix.

2. Inverse regression and the minimum discrepancy approach

2.1. Preliminary results

Unlike traditional regression modeling, inverse regression relies on an assumption about the marginal distribution of $\mathbf{Z}$ instead of the conditional distribution of $Y|\mathbf{Z}$. The so-called *linearity condition* requires that

$$E(\mathbf{Z}|P_{\mathscr{S}_{Y|\mathbf{Z}}}\mathbf{Z}) = P_{\mathscr{S}_{Y|\mathbf{Z}}}\mathbf{Z}.$$

Let the columns of $\boldsymbol{\rho} \in \mathbb{R}^{p\times q}$ be a basis for $\mathscr{S}_{Y|\mathbf{Z}}$, where $q = \dim(\mathscr{S}_{Y|\mathbf{Z}})$. This condition is equivalent to requiring that $E(\mathbf{Z}|\boldsymbol{\rho}^{\mathrm{T}}\mathbf{Z})$ be a linear function of $\boldsymbol{\rho}^{\mathrm{T}}\mathbf{Z}$. We are free to use experimental design, and one-to-one predictor transformations to induce the linearity condition. Cook and Nachtsheim (1994) proposed a re-weighting method to force this condition when necessary without suffering complications when inferring about $Y|\mathbf{Z}$. Since no model is assumed for $Y|\mathbf{Z}$, these methods will not change the fundamental issues in the regression. The linearity condition holds for elliptically contoured predictors. Additionally, Hall and Li (1993) showed that as p increases with q fixed the linearity condition holds to a reasonable approximation in many problems.

The linearity condition connects the CS with the inverse regression of $\mathbf{Z}$ on Y. Li (1991) showed that $E(\mathbf{Z}|Y) \in \mathscr{S}_{Y|\mathbf{Z}}$ when it holds. Defining

$$\mathscr{S}_{E(\mathbf{Z}|Y)} \equiv \mathrm{Span}\{E(\mathbf{Z}|Y=y),\ \ y \text{ varies}\} = \mathrm{Span}\{\boldsymbol{\Sigma}^{-1/2}(E(\mathbf{X}|Y=y) - E(\mathbf{X})),\ \ y \text{ varies}\},$$

one then has $\mathscr{S}_{E(\mathbf{Z}|Y)} \subseteq \mathscr{S}_{Y|\mathbf{Z}}$ and consequently an estimate of $\mathscr{S}_{E(\mathbf{Z}|Y)}$ provides an estimate of at least a part of the CS. To estimate $\mathscr{S}_{E(\mathbf{Z}|Y)}$, we need to estimate its spanning vectors $\boldsymbol{\Sigma}^{-1/2}(E(\mathbf{X}|Y=y) - E(\mathbf{X}))$. The mean $E(\mathbf{X})$ and covariance matrix $\boldsymbol{\Sigma}$ can be estimated using their sample versions. Conditional sample means can be used to estimate $E(\mathbf{X}|Y=y)$ when Y is discrete or categorical. When Y is continuous, Li (1991) proposed estimating $E(\mathbf{X}|Y=y)$ by replacing Y with a discrete version constructed by partitioning the range of Y into h fixed slices. Accordingly, we follow standard methodology and assume that Y takes values in $\{1, 2, \ldots, h\}$.

Besides the linearity condition, the *coverage condition* $\mathscr{S}_{E(\mathbf{Z}|Y)} = \mathscr{S}_{Y|\mathbf{Z}}$, is also often assumed. With both linearity and coverage conditions, an estimate of $\mathscr{S}_{E(\mathbf{Z}|Y)}$ is also an estimate of the CS.

Alternatively, we can connect $\mathscr{S}_{E(\mathbf{Z}|Y)}$ and $\mathscr{S}_{Y|\mathbf{X}}$ via a generalized coverage condition, following Cook and Ni (2005). Before we introduce the definition, a little setup is necessary. Define the following predictor subspace:

$$\mathscr{S}_{\boldsymbol{\rho}} \equiv \mathrm{Span}\{E(\mathbf{Z}|\boldsymbol{\rho}^{\mathrm{T}}\mathbf{Z} = v),\ \ v \text{ varies}\},$$

$\mathscr{S}_{\boldsymbol{\rho}}$ is constructed the same way as $\mathscr{S}_{E(\mathbf{Z}|Y)}$ except that the expectation is conditioned on the sufficient predictor $\boldsymbol{\rho}^{\mathrm{T}}\mathbf{Z}$ instead of Y. Then

Lemma 1.

$$\mathscr{S}_{E(\mathbf{Z}|Y)} \subseteq \mathscr{S}_{\boldsymbol{\rho}} \quad \textit{and} \quad \mathscr{S}_{Y|\mathbf{Z}} \subseteq \mathscr{S}_{\boldsymbol{\rho}}.$$

Lemma 1says that $\mathscr{S}_{\boldsymbol{\rho}}$ always provides an upper bound for both $\mathscr{S}_{Y|\mathbf{Z}}$ and $\mathscr{S}_{E(\mathbf{Z}|Y)}$, without requiring either the linearity or coverage condition. The linearity condition alone will force $\mathscr{S}_{Y|\mathbf{Z}} = \mathscr{S}_{\boldsymbol{\rho}}$. We then have $\mathscr{S}_{E(\mathbf{Z}|Y)} \subseteq \mathscr{S}_{Y|\mathbf{Z}} = \mathscr{S}_{\boldsymbol{\rho}}$. Adding the coverage condition will result in equality of all the three subspaces $\mathscr{S}_{E(\mathbf{Z}|Y)} = \mathscr{S}_{Y|\mathbf{Z}} = \mathscr{S}_{\boldsymbol{\rho}}$.

Now we are ready to define the *generalized coverage condition* $\mathscr{S}_{E(\mathbf{Z}|Y)} = \mathscr{S}_{\boldsymbol{\rho}}$. This condition provides another route to ordering the three subspaces. It alone guarantees that $\mathscr{S}_{Y|\mathbf{Z}} \subseteq \mathscr{S}_{E(\mathbf{Z}|Y)}$ so that we infer about an upper bound on the CS without the linearity condition. Adding the linearity condition to the generalized coverage condition will again result in equality of all the three subspaces. The generalized coverage condition may often hold even when $\mathscr{S}_{Y|\mathbf{Z}} \neq \mathscr{S}_{\boldsymbol{\rho}}$. Cook and Ni (2005) gave an example in this regard.

With the generalized coverage condition alone, the following useful observations can be made:

1. $\dim(\mathscr{S}_{E(\mathbf{Z}|Y)}) = 0 \Longleftrightarrow \dim(\mathscr{S}_{Y|\mathbf{Z}}) = 0$;
2. $\dim(\mathscr{S}_{E(\mathbf{Z}|Y)}) = 1 \Longrightarrow \dim(\mathscr{S}_{Y|\mathbf{Z}}) = 1$;
3. $\dim(\mathscr{S}_{E(\mathbf{Z}|Y)}) = k \geqslant 2 \Longrightarrow 1 \leqslant \dim(\mathscr{S}_{Y|\mathbf{Z}}) \leqslant k$.

Judging from experience, many regression problems have 1D or 2D structure. The first two observations imply that for 0D and 1D regressions, assuming the generalized coverage condition, we can restrict inference for the CS to $\mathcal{S}_{E(\mathbf{Z}|Y)}$. Also, observation 1 can be used for model diagnostics, e.g., considering Y as residuals, where independence between Y and $\mathbf{Z}$ is equivalent to $\dim(\mathcal{S}_{E(\mathbf{Z}|Y)}) = 0$. In order to make use of observation 3, we first introduce the following lemma.

Lemma 2. *If $\mathcal{S}$ is a dimension reduction subspace for $Y|\mathbf{Z}$, then $\mathcal{S}_{Y|P_{\mathcal{S}}\mathbf{Z}} = \mathcal{S}_{Y|\mathbf{Z}}$.*

Lemma 2 tells us that as long as we can find one dimension reduction subspace $\mathcal{S}$ for the regression of $Y|\mathbf{Z}$, we can always use $\mathcal{S}_{Y|P_{\mathcal{S}}\mathbf{Z}}$ to infer about $\mathcal{S}_{Y|\mathbf{Z}}$. The new regression problem $Y|P_{\mathcal{S}}\mathbf{Z}$ will be easier to handle if $P_{\mathcal{S}}\mathbf{Z}$ has a lower dimension than that of the original predictors, which is often the case in practice. Since $\mathcal{S}_{\boldsymbol{\rho}}$ is a dimension reduction subspace, $\mathcal{S}_{Y|P_{\mathcal{S}_{\boldsymbol{\rho}}}\mathbf{Z}} = \mathcal{S}_{Y|\mathbf{Z}}$. Assuming the generalized coverage condition, we can check the linearity condition on $P_{\mathcal{S}_{E(\mathbf{Z}|Y)}}\mathbf{Z}$. If $\dim(\mathcal{S}_{E(\mathbf{Z}|Y)}) < p$, the complexity of this checking process may be reduced relative to that of checking the linearity condition directly on $\mathbf{Z}$. If the linearity condition holds, then $\dim(\mathcal{S}_{E(\mathbf{Z}|Y)}) = k \Longrightarrow \dim(\mathcal{S}_{Y|\mathbf{Z}}) = k$. And $\dim(\mathcal{S}_{Y|\mathbf{Z}}) \leqslant k$ if it fails.

2.2. Minimum discrepancy approach in partial sufficient dimension reduction

We now consider partial dimension reduction following CCL, who used $(\mathbf{X}_w, Y_w)$ to indicate a generic pair distributed like $(\mathbf{X}, Y)|(W = w)$. For example, $\mathcal{S}_{Y_w|\mathbf{X}_w}$ is the CS in subpopulation $W = w$, and $\mathbf{Z}_w = \boldsymbol{\Sigma}_w^{-1/2}(\mathbf{X}_w - E(\mathbf{X}_w))$, where $\boldsymbol{\Sigma}_w = \mathrm{Var}(\mathbf{X}_w) > 0$.

Define the working meta-parameters

$$\mathcal{S}_{\boldsymbol{\xi}_w} \equiv \sum_{y=1}^{h_w} \mathrm{Span}(\boldsymbol{\xi}_{wy}), \quad \mathcal{S}_{\boldsymbol{\xi}} \equiv \sum_{w=1}^{K} \mathcal{S}_{\boldsymbol{\xi}_w},$$

where

$$\boldsymbol{\xi}_{wy} = \boldsymbol{\Sigma}_w^{-1}(E(\mathbf{X}_w|Y_w = y) - E(\mathbf{X}_w)),$$

h_w is the number of slices in subpopulation w and $h = \sum_w h_w$.

Lemma 3. *Assuming that the generalized coverage condition holds for all subpopulations, we then have*

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} \subseteq \mathcal{S}_{\boldsymbol{\xi}}.$$

Hence, inference about $\mathcal{S}_{\boldsymbol{\xi}}$ provides an upper bound for $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$. Under both the generalized coverage and linearity conditions for all subpopulations, $\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{S}_{Y|\mathbf{X}}^{(W)}$. Because inference about $\mathcal{S}_{\boldsymbol{\xi}}$ does not require these conditions, we will as far as possible work in terms of $\mathcal{S}_{\boldsymbol{\xi}}$, using these conditions only when necessary. Let $d = \dim(\mathcal{S}_{\boldsymbol{\xi}})$ and let $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ denote a basis for $\mathcal{S}_{\boldsymbol{\xi}}$. By definition, for each $\boldsymbol{\xi}_{wy}$, we can find a vector $\boldsymbol{\gamma}_{wy} \in \mathbb{R}^d$ such that $\boldsymbol{\xi}_{wy} = \boldsymbol{\beta}\boldsymbol{\gamma}_{wy}$. Define

$$\boldsymbol{\xi}_w = (\boldsymbol{\xi}_{w1}, \ldots, \boldsymbol{\xi}_{wh_w}) = \boldsymbol{\beta}\boldsymbol{\gamma}_w, \quad \boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K) = \boldsymbol{\beta}\boldsymbol{\gamma} \in \mathbb{R}^{p\times h},$$

where $\boldsymbol{\gamma}_w = (\boldsymbol{\gamma}_{w1}, \ldots, \boldsymbol{\gamma}_{wh_w})$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K) \in \mathbb{R}^{d\times h}$. Let $p_w = \Pr(W = w)$, and let $\mathbf{f}_w = (f_{w1}, \ldots, f_{wh_w})^{\mathrm{T}}$, where $f_{wy} = \Pr(Y_w = y)$. This structure implies the following *intrinsic location constraints*:

$$\boldsymbol{\xi}_w\mathbf{f}_w = \boldsymbol{\beta}\boldsymbol{\gamma}_w\mathbf{f}_w = 0, \quad w = 1, \ldots, K.$$

Suppose we have a 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 $\bar{\mathbf{X}}_{w\bullet\bullet}$ be the average in subpopulation w, and let $\bar{\mathbf{X}}_{wy\bullet}$ be the average of the n_{wy} points with $Y_w = y$. Let $\hat{p}_w = n_w/n$, $\hat{f}_{wy} = n_{wy}/n_w$, $\hat{\mathbf{f}}_w = (\hat{f}_{w1}, \ldots, \hat{f}_{wh_w})^{\mathrm{T}}$. And let $\hat{\boldsymbol{\Sigma}}_w > 0$ denote the sample covariance matrix for $\mathbf{X}$ in subpopulation w. The sample versions of $\boldsymbol{\xi}_w$ and $\boldsymbol{\xi}$ are

$$\hat{\boldsymbol{\xi}}_w = (\hat{\boldsymbol{\xi}}_{w1}, \ldots, \hat{\boldsymbol{\xi}}_{wh_w}), \quad \hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_K) \in \mathbb{R}^{p\times h},$$

where $\hat{\boldsymbol{\xi}}_{wy} = \hat{\boldsymbol{\Sigma}}_w^{-1}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet})$.

Because $\boldsymbol{\xi} = \boldsymbol{\beta}\boldsymbol{\gamma}$, we consider estimating a basis for $\mathscr{S}_{\boldsymbol{\xi}}$ by using quadratic discrepancy functions of the form

$$(\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{A}) - \mathrm{vec}(\mathbf{BC}))^{\mathrm{T}}\mathbf{V}_n(\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{A}) - \mathrm{vec}(\mathbf{BC})), \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{h\times(h-K)}$ is nonstochastic and introduced to remove intrinsic location constraints, $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{C} \in \mathbb{R}^{d\times(h-K)}$, and $\mathbf{V}_n \in \mathbb{R}^{p(h-K)\times p(h-K)} > 0$. The subspace of $\mathbb{R}^p$ spanned by a value of $\mathbf{B}$ that minimizes (2) provides an estimate of $\mathscr{S}_{\boldsymbol{\xi}}$. Discrepancy function (2) represents a family of methods, with an individual method being determined by the choice of $\mathbf{V}_n$.

3. Optimal partial inverse regression estimation

Let $\mathbf{D}_{\mathbf{u}}$ denote a diagonal matrix with the elements of the vector $\mathbf{u}$ on the diagonal, and let $\mathrm{diag}\{\mathbf{M}_j\}$ denote a positive definite block diagonal matrix with blocks $\mathbf{M}_j \in \mathbb{R}^{p_j\times p_j}$, where both j and p_j are positive integers. As we discussed before, the columns of $\hat{\boldsymbol{\xi}}$ provide redundant information due to the intrinsic location constraints. In order to reduce the redundancy, we construct a series of nonstochastic matrices $\mathbf{A}_w \in \mathbb{R}^{h_w\times(h_w-1)}$ such that $\mathbf{A}_w^{\mathrm{T}}\mathbf{A}_w = I_{h_w-1}$ and $\mathbf{A}_w^{\mathrm{T}}\mathbf{1}_{h_w} = 0$. Without loss of generality, we will use the reduced data matrices

$$\hat{\boldsymbol{\zeta}}_w \equiv \boldsymbol{\xi}_w\mathbf{D}_{\hat{\mathbf{f}}_w}\mathbf{A}_w \in \mathbb{R}^{p\times(h_w-1)} \quad \text{and} \quad \hat{\boldsymbol{\zeta}} \equiv (\hat{\boldsymbol{\zeta}}_1, \ldots, \hat{\boldsymbol{\zeta}}_K) \in \mathbb{R}^{p\times(h-K)} \tag{3}$$

to construct the discrepancy function

$$F_d(\mathbf{B}, \mathbf{C}) \equiv (\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC}))^{\mathrm{T}}\mathbf{V}_n(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC})), \tag{4}$$

where $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{C} \in \mathbb{R}^{d\times(h-K)}$, and $\mathbf{V}_n > 0$ has yet to be specified. Let

$$\boldsymbol{\nu}_w = \boldsymbol{\gamma}_w\mathbf{D}_{\mathbf{f}_w}\mathbf{A}_w, \quad \boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K) = (\gamma_1, \ldots, \gamma_K)\mathrm{diag}\{\mathbf{D}_{\mathbf{f}_w}\}\mathrm{diag}\{\mathbf{A}_w\}.$$

Then $\hat{\boldsymbol{\zeta}}$ converges in probability to $\boldsymbol{\zeta} = \boldsymbol{\beta}\boldsymbol{\nu}$, $\boldsymbol{\nu} \in \mathbb{R}^{d\times(h-K)}$.

3.1. Asymptotic normality

As discussed by Ferguson (1958), Lindsay and Qu (2003), and Shapiro (1986), a well-known choice of $\mathbf{V}_n$ is a consistent estimate of the inverse of the asymptotic covariance of $\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu}))$. In order to study the asymptotic properties of $n\hat{F}_d$ with this choice, we need to find the asymptotic distribution of $\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu}))$, where $\hat{F}_d$ is the minimum value of $F_d(\mathbf{B}, \mathbf{C})$. First we will study the asymptotic distribution of $\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}}))$.

Conditioning on subpopulation w, define the random variable $J_{wy} = I\{Y_w = y\}$. Then given w, $E(J_{wy}) = \Pr(Y_w = y) = f_{wy}$. Define $\mathbf{J}_w = (J_{w1}, \ldots, J_{wh_w})^{\mathrm{T}}$ and $\boldsymbol{\epsilon}_w = (\boldsymbol{\varepsilon}_{w1}, \ldots, \boldsymbol{\varepsilon}_{wh_w})^{\mathrm{T}}$, where its elements $\boldsymbol{\varepsilon}_{wy} = J_{wy} - f_{wy} - \mathbf{Z}_w^{\mathrm{T}}E(\mathbf{Z}_w J_{wy})$, $y = 1, \ldots, h_w$, are the population residuals from the ordinary least-squares fit of J_{wy} on $\mathbf{Z}_w$. We then have the following theorem.

Theorem 1. *Assume that the data* $(Y_i, \mathbf{X}_i, W_i)$, $i = 1, \ldots, n$, *is a simple random sample of* $(Y, \mathbf{X}, W)$ *with finite fourth moments. Then*

$$\sqrt{n_w}(\mathrm{vec}(\boldsymbol{\xi}_w\mathbf{D}_{\hat{\mathbf{f}}_w}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}_w\mathbf{D}_{\mathbf{f}_w})) \xrightarrow{\mathscr{D}} \mathrm{Normal}(0, \boldsymbol{\Gamma}_w),$$

where $\boldsymbol{\Gamma}_w = \mathrm{Cov}(\mathrm{vec}(\boldsymbol{\Sigma}_w^{-1/2}\mathbf{Z}_w\boldsymbol{\epsilon}_w^{\mathrm{T}})) \in \mathbb{R}^{ph_w\times ph_w}$.

And,

$$\sqrt{n_w}(\mathrm{vec}(\hat{\boldsymbol{\zeta}}_w) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu}_w)) \xrightarrow{\mathscr{D}} \mathrm{Normal}(0, \boldsymbol{\Gamma}_{\zeta_w}),$$

where $\boldsymbol{\Gamma}_{\zeta_w} = (\mathbf{A}_w^{\mathrm{T}} \otimes I_p)\boldsymbol{\Gamma}_w(\mathbf{A}_w \otimes I_p) \in \mathbb{R}^{p(h_w-1)\times p(h_w-1)}$ *is nonsingular, and* $\otimes$ *denotes the Kronecker product.*

Based on the above theorem, we reach the following corollary by Slutsky's Theorem.

Corollary 1. *Assume that the data* $(Y_i, \mathbf{X}_i, W_i)$, $i = 1, \ldots, n$, *are a simple random sample of* $(Y, \mathbf{X}, W)$ *with finite fourth moments. Then*

$$\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{v})) \xrightarrow{\mathscr{D}} \mathrm{Normal}(0, \boldsymbol{\Gamma}_\zeta),$$

where $\boldsymbol{\Gamma}_\zeta = (\mathrm{diag}(\mathbf{A}_w^{\mathrm{T}}) \otimes I_p)(\mathrm{diag}(\boldsymbol{\Gamma}_w / p_w))(\mathrm{diag}(\mathbf{A}_w) \otimes I_p)$.

3.2. Asymptotic optimality

To obtain the optimal version of $\hat{F}_d$, we choose $\mathbf{V}_n = \hat{\boldsymbol{\Gamma}}_\zeta^{-1}$, where $\hat{\boldsymbol{\Gamma}}_\zeta$ is a consistent estimate of $\boldsymbol{\Gamma}_\zeta$ given in Corollary 1. The sample version of $\boldsymbol{\Gamma}_\zeta^{-1}$ is one choice for $\mathbf{V}_n$. Define then the new discrepancy function,

$$F_d^{\mathrm{opt}}(\mathbf{B}, \mathbf{C}) = (\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{B}\mathbf{C}))^{\mathrm{T}} \hat{\boldsymbol{\Gamma}}_\zeta^{-1} (\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{B}\mathbf{C})). \tag{5}$$

Solutions of the minimization of (5) are not unique due to the over-parameterization of the setting. This nonidentifiability is not an issue since any estimator of a basis can specify $\mathscr{S}_\xi$. For example, we can minimize (5) subject to the constraint that $\mathbf{B}^{\mathrm{T}}\mathbf{B} = I_d$. The estimate of $\mathscr{S}_\xi$ constructed by minimizing (5) is called the *optimal partial inverse regression estimation* (OPIRE) estimator.

Let $\boldsymbol{\Delta}_\zeta \equiv (\boldsymbol{v}^{\mathrm{T}} \otimes I_p,\ I_{h-K} \otimes \boldsymbol{\beta})$, which is the Jacobian matrix

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

evaluated at $(\mathbf{B} = \boldsymbol{\beta}, \mathbf{C} = \boldsymbol{v})$. The following theorem provides the asymptotic properties of OPIRE.

Theorem 2. *Assume that the data* $(Y_i, \mathbf{X}_i, W_i)$, $i = 1, \ldots, n$, *is a simple random sample of* $(Y, \mathbf{X}, W)$ *with finite fourth moments. Let* $\mathscr{S}_\xi = \sum_{w=1}^{K} \sum_{y=1}^{h_w} \mathrm{Span}(\boldsymbol{\xi}_{wy})$, *let* $d = \dim(\mathscr{S}_\xi)$ *and let* $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{v}}) = \arg_{\mathbf{B},\mathbf{C}} \min F_d^{\mathrm{opt}}(\mathbf{B}, \mathbf{C})$ *as defined in* (5). *Then*,

1. *The estimate* $\mathrm{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{v}})$ *is asymptotically efficient, and*

$$\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{v}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{v})) \xrightarrow{\mathscr{D}} \mathrm{Normal}(0, \boldsymbol{\Delta}_\zeta (\boldsymbol{\Delta}_\zeta^{\mathrm{T}} \boldsymbol{\Gamma}_\zeta^{-1} \boldsymbol{\Delta}_\zeta)^{-} \boldsymbol{\Delta}_\zeta^{\mathrm{T}}).$$

2. $n\hat{F}_d^{\mathrm{opt}}$ *has an asymptotic chi-squared distribution with degrees of freedom* $(p - d)(h - d - K)$.
3. $\mathrm{Span}(\hat{\boldsymbol{\beta}})$ *is a consistent estimator of* $\mathscr{S}_\xi$.

The optimality of F_d^{opt} stems from the asymptotic normality of $\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\beta}v))$ given in Theorem 1. The asymptotic efficiency in Theorem 2 means that the estimate of any function of $\mathrm{vec}(\boldsymbol{\beta}v)$ obtained via the OPIRE method has the smallest asymptotic variance among estimates from all possible $\mathbf{V}_n$'s. This kind of estimate was called a best generalized least-squares estimator by Browne (1984), and a best asymptotically normal estimator by Ferguson (1958). Simulation studies show that OPIRE can easily dominate partial SIR in terms of estimation of the partial CS. Also, its test statistic for testing hypotheses about the partial structural dimension performs at least as well as that of partial SIR and sometimes a lot better, as illustrated by the simulation studies shown in Section 5.

3.3. Computation for OPIRE

Since all the matrices $\mathbf{A}_w$'s are nonstochastic, in order to get a consistent estimate $\hat{\boldsymbol{\Gamma}}_\zeta$ of $\boldsymbol{\Gamma}_\zeta$, we need only plug in a consistent sample version of $\boldsymbol{\Gamma}_w = \mathrm{Cov}(\mathrm{vec}(\boldsymbol{\Sigma}_w^{-1/2}\mathbf{Z}_w\boldsymbol{\epsilon}_w^{\mathrm{T}})) \in \mathbb{R}^{ph_w \times ph_w}$, $w = 1, 2, \ldots, K$. Such estimates can be constructed easily by substituting sample versions of population quantities, noting that $E(\mathrm{vec}(\boldsymbol{\Sigma}_w^{-1/2}\mathbf{Z}_w\boldsymbol{\varepsilon}_w^{\mathrm{T}})) = 0$ for all subpopulations.

We then adapt the *alternating least-squares method* proposed by Cook and Ni (2005) to minimize (5) for *this* $\mathbf{V}_n$. For any given $\mathbf{V}_n$ we can always treat the discrepancy function (5) as a separable nonlinear least-squares problem (Ruhe and Wedin, 1980). The value of $\mathrm{vec}(\mathbf{C})$ that minimizes $F_d(\mathbf{B}, \mathbf{C})$ for a given $\mathbf{B}$ can be constructed as the coefficient

vector from the least-squares fit of $\mathbf{V}_n^{1/2}\,\mathrm{vec}(\hat{\boldsymbol{\zeta}})$ on $\mathbf{V}_n^{1/2}(I_{h-K}\otimes\mathbf{B})$. On the other hand, fixing $\mathbf{C}$, consider minimization with respect to one column $\mathbf{b}_j$ of $\mathbf{B}$, given the remaining columns of $\mathbf{B}$ and subject to the length constraint $\|\mathbf{b}_j\|=1$ and the orthogonality constraint $\mathbf{b}_j^{\mathrm{T}}\mathbf{B}_{(-j)}=0$, where $\mathbf{B}_{(-j)}$ is the matrix that is left after taking away $\mathbf{b}_j$ from $\mathbf{B}$. For this partial minimization problem, the discrepancy function can be re-expressed as

$$F^*(\mathbf{b})=(\boldsymbol{\alpha}_j-(\mathbf{c}_j^{\mathrm{T}}\otimes I_p)\mathbf{Q}_{\mathbf{B}_{(-j)}}\mathbf{b})^{\mathrm{T}}\mathbf{V}_n(\boldsymbol{\alpha}_j-(\mathbf{c}_j^{\mathrm{T}}\otimes I_p)\mathbf{Q}_{\mathbf{B}_{(-j)}}\mathbf{b}),$$

where $\boldsymbol{\alpha}_j=\mathrm{vec}(\hat{\boldsymbol{\zeta}}-\mathbf{B}_{(-j)}\mathbf{C}_{(-j)})\in\mathbb{R}^{p(h-K)}$, $\mathbf{c}_j$ is the jth row of $\mathbf{C}$, $\mathbf{C}_{(-j)}$ consists of all but the jth row of $\mathbf{C}$, and $\mathbf{Q}_{\mathbf{B}_{(-j)}}$ projects onto the orthogonal complement of $\mathrm{Span}(\mathbf{B}_{(-j)})$ in the usual inner product. It becomes a linear regression problem again. In the end, the algorithm provides $\mathrm{Span}(\hat{\boldsymbol{\beta}})$, an estimate of $\mathscr{S}_{\boldsymbol{\xi}}$. The linear combinations $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}$ are the estimated sufficient predictors.

As suggested by the form of $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\zeta}}$, the inner-product matrix in F_d^{opt} depends on $\mathscr{S}_{\boldsymbol{\xi}}$. We could apply an iterative algorithm to reduce the variability of the inner-product matrix $\mathbf{V}_n$. Here is a sketch of this idea. First, get $\mathrm{Span}(\hat{\boldsymbol{\beta}})$, an estimate of $\mathscr{S}_{\boldsymbol{\xi}}$ via the above alternating least-squares method. Second, obtain a new estimate of $\mathbf{V}_n$ using $\mathrm{Span}(\hat{\boldsymbol{\beta}})$. Then, re-run the above algorithm to update $\mathrm{Span}(\hat{\boldsymbol{\beta}})$ applying this new $\mathbf{V}_n$. Carroll and Ruppert (1988) recommended at least two cycles. Our experiences with OPIRE suggest that another cycle of iteration provides only minimal improvement. We hence use a one-cycle iterative computation algorithm for OPIRE.

3.4. Partial SIR via the minimum discrepancy approach

As discussed in Section 1, partial SIR is one of the methods taking the spectral decomposition approach. The kernel matrix that partial SIR used is $\widehat{\mathbf{M}}_{\mathrm{psir}}=\sum_{w=1}^{K}\sum_{y=1}^{h_w}\hat{p}_w\hat{f}_{wy}\bar{\mathbf{Z}}_{wy\bullet}\bar{\mathbf{Z}}_{wy\bullet}^{\mathrm{T}}$. Let $\hat{\boldsymbol{\Sigma}}_{\mathrm{pool}}\equiv\sum_w(n_w/n)\hat{\boldsymbol{\Sigma}}_w$ be the pooled covariance matrix. Then partial SIR can be re-derived by considering a minimum discrepancy function (2) setting $\mathbf{A}=I_h$ and $\mathbf{V}_n=\mathrm{diag}(\mathbf{D}_{\hat{\mathbf{f}}_w})\otimes(\hat{p}_w\hat{\boldsymbol{\Sigma}}_{\mathrm{pool}})$. Let $\hat{\boldsymbol{\eta}}\in\mathbb{R}^{p\times d}$ be a $\mathbf{B}$ minimizer of the discrepancy function. Then $\mathrm{Span}(\hat{\boldsymbol{\Sigma}}_{\mathrm{pool}}{}^{1/2}\hat{\boldsymbol{\eta}})$is the space spanned by the d eigenvectors corresponding to $\widehat{\mathbf{M}}_{\mathrm{psir}}$'s d largest eigenvalues. Also, it can be shown that the minimum value of the discrepancy function is the summation of n times the $n-d$ smallest eigenvalues of $\widehat{\mathbf{M}}_{\mathrm{psir}}$, which is exactly the test statistic proposed by CCL for testing the partial structural dimension using partial SIR.

4. Testing dimension and coordinates hypotheses in partial sufficient dimension reduction

In traditional model-based regression, tests for predictor effects are often important. While, in sufficient dimension reduction, little attention has been paid to this kind of problems until very recently. Cook (2004) introduced a general formulation and specific implementation based on SIR for testing predictor effects in sufficient dimension reduction. Following Cook's formulation, we develop predictor tests via the minimum discrepancy approach with both continuous and categorical predictors.

We shall study tests of the conditional independence hypothesis

$$Y\perp\!\!\!\perp P_{\mathscr{H}}\mathbf{X}\mid(Q_{\mathscr{H}}\mathbf{X},W),\tag{6}$$

where $\mathscr{H}$ is a r-dimensional user-selected subspace of the quantitative predictor space, and $Q_{\mathscr{H}}=I-P_{\mathscr{H}}$. Since the hypothesis is certainly false if $r>p-\dim(\mathscr{S}_{Y|\mathbf{X}}^{(W)})$, it is reasonable to restrict that $r\leqslant p-\dim(\mathscr{S}_{Y|\mathbf{X}}^{(W)})$. For example, if we partition $\mathbf{X}^{\mathrm{T}}=(\mathbf{X}_1^{\mathrm{T}},\mathbf{X}_2^{\mathrm{T}})$, we can set $\mathscr{H}=\mathrm{Span}((I_r0)^{\mathrm{T}})$ to test the hypothesis of no intra-subpopulation effects for the first r predictors $\mathbf{X}_1$, $Y\perp\!\!\!\perp\mathbf{X}_1\mid(\mathbf{X}_2,W)$. CCL showed that $\mathscr{S}_{Y|\mathbf{X}}\subseteq\mathscr{S}_{W|\mathbf{X}}+\mathscr{S}_{Y|\mathbf{X}}^{(W)}$, where $\mathscr{S}_{W|\mathbf{X}}$ is the CS for the regression of the categorical predictor W on $\mathbf{X}$. Hence, (6) differs from the hypotheses that Cook (2004) studied. For instance, in the *lean body mass* example of CCL, they inferred that $Y\perp\!\!\!\perp\mathbf{X}\mid(\mathbf{v}_1^{\mathrm{T}}\mathbf{X},\mathbf{v}_2^{\mathrm{T}}\mathbf{X})$ and $Y\perp\!\!\!\perp\mathbf{X}\mid(\mathbf{v}_1^{\mathrm{T}}\mathbf{X},W)$.

Proposition 1. *Assume that the partial CS* $\mathscr{S}_{Y|\mathbf{X}}^{(W)}$ *exists. Then* (6) *is true if and only if* $P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)}=\mathscr{O}_p$, *where* $\mathscr{O}_p$ *indicates the origin in* $\mathbb{R}^p$.

This proposition indicates that instead of dealing with (6) directly, we could test the corresponding hypothesis about the working subspace $\mathscr{S}_\xi$ when $\mathscr{S}_{Y|\mathbf{X}}^{(W)} = \mathscr{S}_\xi$. Assuming the generalized coverage condition but not the linearity condition, we have

$$P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} \subseteq P_{\mathscr{H}}\mathscr{S}_\xi \tag{7}$$

and consequently $P_{\mathscr{H}}\mathscr{S}_\xi = \mathscr{O}_p$ implies $P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} = \mathscr{O}_p$. Since the containment in (7) may be proper, lacking information to reject $P_{\mathscr{H}}\mathscr{S}_\xi = \mathscr{O}_p$ supports the hypothesis (6), while rejecting does not necessarily imply dependence.

We are now ready to consider the following five hypothesis forms, depending on application-specific requirements:

1. *Marginal dimension hypotheses*: $d = m$ vs. $d > m$.
2. *Marginal predictor hypotheses*: $P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} = \mathscr{O}_p$ vs. $P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} \neq \mathscr{O}_p$.
3. *Joint dimension–predictor hypotheses*: $P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} = \mathscr{O}_p$ and $d = m$ vs. $P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} \neq \mathscr{O}_p$ or $d > m$.
4. *Conditional predictor hypotheses*: $P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} = \mathscr{O}_p$ vs. $P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} \neq \mathscr{O}_p$ given d.
5. *Conditional dimension hypotheses*: $d = m$ vs. $d > m$ given $P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} = \mathscr{O}_p$.

Marginal dimension hypotheses were considered by CCL. Here they can be tested using $n\hat{F}_d^{\text{opt}}$, which has an asymptotic chi-squared distribution with degrees of freedom $(p-d)(h-d-K)$ under the null hypotheses. Our method for testing dimensions has the advantages that it provides greater power, and its test statistic has a simpler asymptotic null distribution compared to that developed by CCL. Simulation studies show that OPIRE can easily win over partial SIR when testing marginal dimension hypotheses with nonconstant $\text{Cov}(\mathbf{X}|W)$, and can do at least as well as partial SIR when the homogeneous covariance condition holds.

Given $P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} = \mathscr{O}_p$, testing a dimension hypothesis is equivalent to testing a marginal dimension hypothesis with new coordinates. To help fix the idea, here is an example. Suppose that $\mathbf{X} = (X_1, X_2, X_3)^{\mathrm{T}} \in \mathbb{R}^3$, $Y \perp\!\!\!\perp X_3 | (X_1, X_2, W)$. Defining $\mathbf{X}_{\text{new}} = (X_1, X_2)^{\mathrm{T}}$, the conditional dimension hypothesis is equivalent to testing the marginal dimension hypothesis $\dim(\mathscr{S}_{Y|\mathbf{X}_{\text{new}}}^{(W)}) = m$. In the following sections, we will use Theorems 1 and 2 to develop test statistics for hypotheses 2–4 based on the working meta-parameter $\mathscr{S}_\xi$.

4.1. Marginal predictor hypotheses

Let $\boldsymbol{\alpha} \in \mathbb{R}^{p\times r}$ be an orthonormal basis for $\mathscr{H}$. Testing $P_{\mathscr{H}}\mathscr{S}_\xi = \mathscr{O}_p$ is equivalent to testing $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\zeta} = 0$, where $\boldsymbol{\zeta}$ is the population limit of $\hat{\boldsymbol{\zeta}}$ as defined previously in (3). We can test $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\zeta} = 0$ by measuring how far $\boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{\zeta}}$ is from 0. Theorem 1 provides a method for doing this by using the Wald test statistic

$$T(\mathscr{H}) = n\,\text{vec}(\boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{\zeta}})^{\mathrm{T}}\{(I_{h-K}\otimes\boldsymbol{\alpha}^{\mathrm{T}})\hat{\boldsymbol{\Gamma}}_\zeta(I_{h-K}\otimes\boldsymbol{\alpha})\}^{-1}\,\text{vec}(\boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{\zeta}}). \tag{8}$$

By Theorem 1 and Slutsky's theorem, we can easily show that, under the null hypothesis, $T(\mathscr{H})$ has an asymptotic chi-squared distribution with degrees of freedom $r(h-K)$. Also,

$$T(\mathscr{H}) = \min_{\mathbf{C}} n F_p^{\text{opt}}(I_p, \mathbf{C})$$

subject to the constraint $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{C} = 0$. Thus, we can test a marginal hypothesis $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\zeta} = 0$ by setting $\boldsymbol{\beta} = I_p$. In this case, the hypothesis becomes $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\nu} = 0$. We then fit $F_p^{\text{opt}}(I_p, \mathbf{C})$ subject to the constraint $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{C} = 0$ and take n times the minimum value of the constrained objective function as the test statistic, which is the same as (8).

The test statistic $T(\mathscr{H})$ is invariant with respect to the choice of basis for $\mathscr{H}$. Applying generalized coverage and then linearity conditions, we can use $T(\mathscr{H})$ to test $Y \perp\!\!\!\perp X_j \mid (\mathbf{X}_{-j}, W)$, where $\mathbf{X}_{-j}$ indicates the predictors left after taking away X_j.

As we stated before, assuming only the generalized coverage condition, hypotheses about the working subspace $\mathscr{S}_\xi$ are not equivalent to hypotheses about the partial CS unless $\dim(\mathscr{S}_\xi) = 0$ or 1.

4.2. Joint dimension–predictor hypotheses

Under the null hypothesis, $P_{\mathscr{H}}\mathscr{S}_{\xi} = \mathscr{O}_p$ and $d = m$, $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\zeta} = 0$ is equivalent to $\boldsymbol{\zeta} = Q_{\mathscr{H}}\boldsymbol{\zeta} = Q_{\mathscr{H}}\boldsymbol{\beta}\boldsymbol{v} = \boldsymbol{\alpha}_0\boldsymbol{\beta}_0\boldsymbol{v}$, where $\boldsymbol{\beta} \in \mathbb{R}^{p\times m}$ is an orthonormal basis for $\mathscr{S}_{\xi}$, $\boldsymbol{v} \in \mathbb{R}^{m\times(h-K)}$, $\boldsymbol{\alpha}_0 \in \mathbb{R}^{p\times(p-r)}$ is an orthonormal basis for $\mathrm{Span}(Q_{\mathscr{H}})$, $\boldsymbol{\beta}_0$ contains the coordinates of $\boldsymbol{\beta}$ represented in terms of the basis $\boldsymbol{\alpha}_0$. We then can fit under the joint hypothesis by minimizing the following constrained optimal discrepancy function:

$$F^{\mathrm{opt}}_{m,\boldsymbol{\alpha}}(\mathbf{B}, \mathbf{C}) = (\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\alpha}_0\mathbf{B}\mathbf{C}))^{\mathrm{T}}\hat{\boldsymbol{\Gamma}}_{\zeta}^{-1}(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\alpha}_0\mathbf{B}\mathbf{C})) \tag{9}$$

over $\mathbf{B} \in \mathbb{R}^{(p-r)\times m}$ and $\mathbf{C} \in \mathbb{R}^{m\times(h-K)}$. Values of $\mathbf{B}$ and $\mathbf{C}$ that minimize (9) provide estimates of $\boldsymbol{\beta}_0$ and $\boldsymbol{v}$. Following the result of Theorem 2, we know that, under the null hypothesis, the test statistic $n\hat{F}^{\mathrm{opt}}_{m,\boldsymbol{\alpha}}$ has an asymptotic chi-squared distribution with degrees of freedom $(p-m)(h-m-K)+mr$. Note that the Jacobian matrix is

$$\boldsymbol{\Delta}_{\zeta,\boldsymbol{\alpha}} \equiv (I_{h-K} \otimes \boldsymbol{\alpha}_0)(\boldsymbol{v}^{\mathrm{T}} \otimes I_{p-r}, I_{h-K} \otimes \boldsymbol{\beta}_0) \in \mathbb{R}^{p(h-K)\times m(p+h-r-K)}. \tag{10}$$

The degrees of freedom are then found by calculating $p(h-K) - \mathrm{rank}(\boldsymbol{\Delta}_{\zeta,\boldsymbol{\alpha}}) = p(h-K) - m(p+h-r-K-m) = (p-m)(h-m-K)+mr$.

4.3. Conditional predictor hypotheses

When d is specified as a modeling device, or when inference on d using marginal dimension tests results in a firm estimate, we might consider the conditional hypothesis $P_{\mathscr{H}}\mathscr{S}_{\xi} = \mathscr{O}_p$ given d. Since information on d is used, these conditional tests are expected to provide greater power than the marginal tests discussed in Section 4.1.

We use the difference in minimum discrepancies

$$T(\mathscr{H}|d) = n\hat{F}^{\mathrm{opt}}_{d,\boldsymbol{\alpha}} - n\hat{F}^{\mathrm{opt}}_{d} \tag{11}$$

to test a conditional predictor hypothesis. It follows from Shapiro (1986) and Cook and Ni (2005) that $T(\mathscr{H}|d)$ is asymptotically equivalent to

$$\mathbf{U}^{\mathrm{T}}(P_{\zeta} - P_{\zeta,\boldsymbol{\alpha}})\mathbf{U},$$

where $\mathbf{U} \in \mathbb{R}^{p(h-K)}$ is a standard normal random vector, P_{ζ} and $P_{\zeta,\boldsymbol{\alpha}}$ are the orthogonal projections with respect to the usual inner product onto $\mathrm{Span}(\boldsymbol{\Gamma}_{\zeta}^{-1/2}\boldsymbol{\Delta}_{\zeta})$ and $\mathrm{Span}(\boldsymbol{\Gamma}_{\zeta}^{-1/2}\boldsymbol{\Delta}_{\zeta,\boldsymbol{\alpha}})$. Since $\mathrm{Span}(\boldsymbol{\Delta}_{\zeta,\boldsymbol{\alpha}}) \subseteq \mathrm{Span}(\boldsymbol{\Delta}_{\zeta})$, $\mathrm{Span}(\boldsymbol{\Gamma}_{\zeta}^{-1/2}\boldsymbol{\Delta}_{\zeta,\boldsymbol{\alpha}}) \subseteq \mathrm{Span}(\boldsymbol{\Gamma}_{\zeta}^{-1/2}\boldsymbol{\Delta}_{\zeta})$. Therefore, $P_{\zeta} - P_{\zeta,\boldsymbol{\alpha}}$ is an orthogonal projection with rank

$$\mathrm{rank}(\boldsymbol{\Delta}_{\zeta}) - \mathrm{rank}(\boldsymbol{\Delta}_{\zeta,\boldsymbol{\alpha}}) = d(p+h-d-K) - d(p+h-r-d-K) = rd,$$

and $T(\mathscr{H}|d)$ has an asymptotic chi-squared distribution with degrees of freedom rd. Following this argument, it is easy to derive the asymptotic independence of the conditional predictor test statistic $T(\mathscr{H}|d)$ and the marginal dimension test statistic $n\hat{F}^{\mathrm{opt}}_{d}$.

The power of $T(\mathscr{H}|d)$ is expected to be greater than the power of $T(\mathscr{H})$. However, when d is misspecified, $T(\mathscr{H}|d)$ may lead to misleading results. This speculation is confirmed by simulation studies.

5. Simulation results

In this section, selected simulation results are reported to support the results developed in previous sections. The performance of OPIRE and partial SIR were compared regarding estimation accuracies and actual testing levels.

5.1. Estimation of $\mathscr{S}_{\xi}$ with d known

When d, the dimension of $\mathscr{S}_{\xi}$, is known and the sample size is large, OPIRE is expected to provide a better estimate of $\mathscr{S}_{\xi}$ than partial SIR. We will study several simple simulation models to demonstrate this.

Model A: We first consider a 2D model for $K=2$ subpopulations indicated by $W \in \{1, 2\}$. Within each subpopulation, Y takes values in $\{1, 2, 3\}$, and we generate n_{wy} points for each value of Y. For $W=1$, $\mathbf{X}_{iy} = \sigma_y\mathbf{Z}_{iy} + \mu_y\mathbf{e}_1$; for $W=2$,

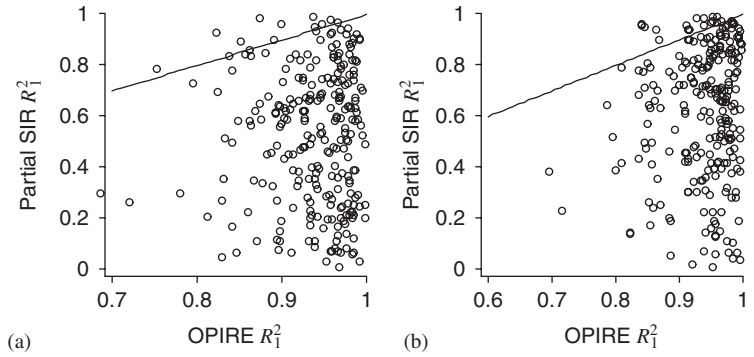

Fig. 1. Model A: Estimation accuracy of OPIRE and partial SIR with (a) $\sigma_1 = 7$ and (b) $\sigma_1 = 5$. The lines indicating equal R^2's were added for reference.

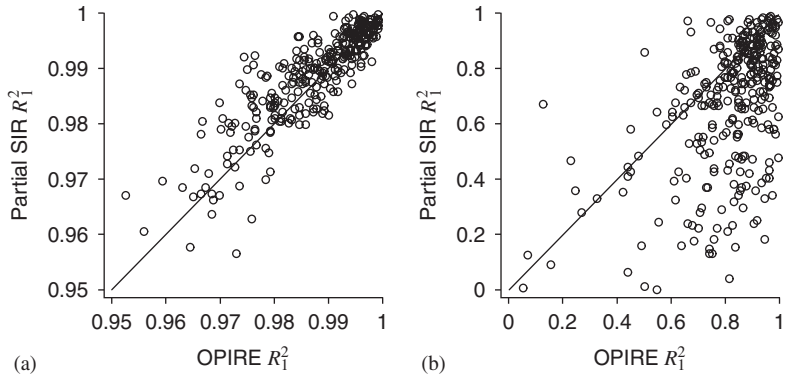

Fig. 2. Model A: Estimation accuracy of OPIRE and partial SIR for equal and unequal σ_y's. The lines indicating equal R^2's were added for reference. (a) Equal $\sigma_y = 0.5$; (b) Unequal σ_y's and unequal n_{wy}'s

$\mathbf{X}_{iy} = \sigma_y \mathbf{Z}_{iy} + \mu_y \mathbf{e}_2$, where $\mathbf{e}_j \in \mathbb{R}^p$ is a vector whose jth element is 1 and elsewhere is 0, and $\mathbf{Z} \in \mathbb{R}^p$ is a vector of independent standard normal variates. For ease of discussion, we will use the *reference model* with $p = 5$, $n_1 = n_2 = n_3 = 200$, $\mu_1 = 1$, $\mu_2 = 0.7$, $\mu_3 = 0.3$, $\sigma_1 = 5$ and $\sigma_2 = \sigma_3 = 0.5$. Parameters not explicitly specified in a simulation configuration are the same as those in this reference model. All estimates were constructed with $d = 2$.

We calculate R_1^2 and R_2^2, the R^2 values from the regressions of X_1 and X_2 on the $d = 2$ estimated sufficient predictors $\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{X}$ and $\hat{\boldsymbol{\beta}}_2^{\mathrm{T}}\mathbf{X}$, to measure estimation accuracy. Since the results for R_2^2 are essentially the same as those for R_1^2, we will use just R_1^2 in the following discussion. As shown in Fig. 1a, OPIRE definitely won over partial SIR with $\sigma_1 = 7$. The average of R_1^2 from 250 replications was 0.935 from OPIRE, and 0.530 from partial SIR. Also, the R_1^2 from OPIRE exceeded the R_1^2 from partial SIR 95.3% of the time. Shown in Fig. 1b are the results from 250 simulation runs with $\sigma_1 = 5$. OPIRE still did a better job than partial SIR over 94% of the time. When $\sigma_1 = 0.5$, the three groups within each subpopulation have equal frequencies and equal variation, and we observed no difference between the performance of OPIRE and that of partial SIR, as shown in Fig. 2a.

Let $n_1 = 100,\ n_2 = 470,\ n_3 = 30,\ \sigma_1 = 2,\ \sigma_2 = 1,\ \sigma_3 = 0.5$. The unequal frequencies plus mildly different σ_y's still resulted in clear differences in the estimators as shown in Fig. 2b. The R_1^2 from OPIRE exceeded the R_1^2 from partial SIR about 80% of the time.

Fig. 3a shows plots of the average R_1^2 from OPIRE and partial SIR from 1000 simulation runs at values of σ_1 between 0.5 and 10. The changes in σ_1 did not affect the performance of OPIRE much. In contrast, partial SIR was quite sensitive to them. As shown in Fig. 3b, we also varied the equal sample size n_{wy} within each subpopulation from 10 to 400, and again OPIRE proved to be the better method. In some simulations, we also changed the signal by replacing μ_y with

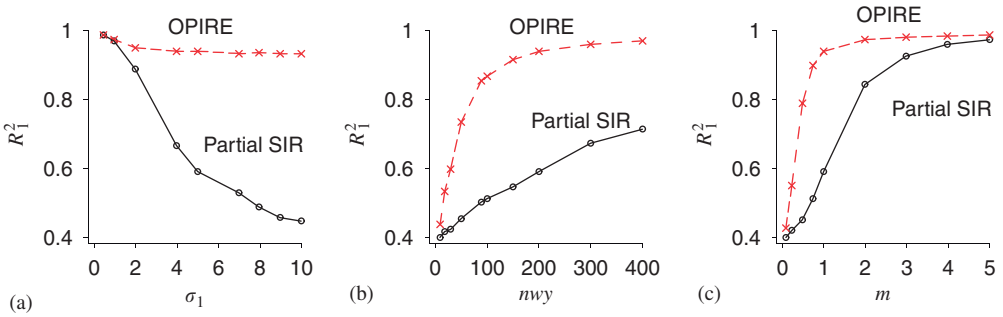

Fig. 3. Model A: Estimation accuracy of OPIRE and partial SIR as (a) σ_1, (b) n_{wy} and (c) m vary.

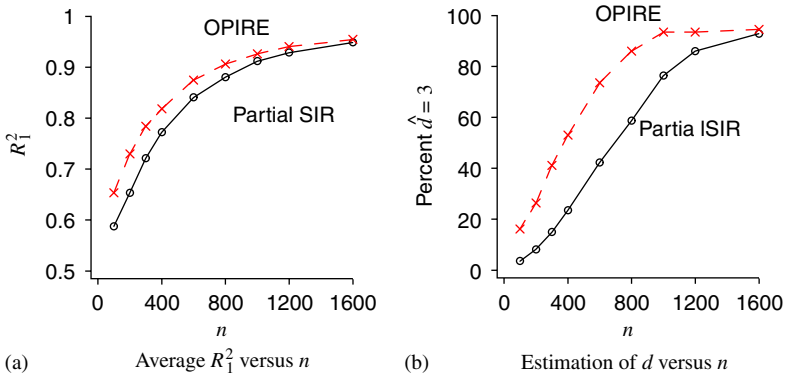

Fig. 4. Model B-3D: (a) Estimation accuracy of Model B versus n. (b) Percentage of runs in which $\hat{d} = 3$ versus n.

$m\mu_y$, for $y = 1, 2, 3$. Fig. 3c shows the average R_1^2 from 1000 simulation replications from partial SIR and OPIRE with m taking value between 0.1 and 5. For sufficiently large m, there is little difference between these two methods. Hence, if we have a very strong signal, the heterogeneity in the regression might be ignored.

In addition to the results reported here, we also tried versions of Model A with $K > 2$ subpopulations and found similar results.

Model B: We consider two versions of Model B. Version 1 is a 3D model for $K = 2$ subpopulations indicated by $W \in \{1, 2\}$. $Y = 1.5(5 + X_1)(2 + X_2 + X_3) + 0.5\varepsilon$ when $W = 1$; $Y = 1.5(5 + X_4)(2 + X_2 + X_3) + 0.5\varepsilon$ when $W = 2$, where ε is a standard normal random variate, $X_1 = W_1$, $X_2 = V_1 + W_2/2$, $X_3 = -V_1 + W_2/2$, $X_4 = V_2 + V_3$, and $X_5 = V_2 - V_3$. The V_i's and W_j's are independent with V_i's drawn from a $t_{(5)}$ distribution and the W_j's from gamma(.2) distribution. Version 2 is a 2D model constructed by replacing X_4 with X_1 in the generation of Y when $W = 2$. Versions of this model were also used by Li (1991), Velilla (1986) and others in simulation studies related to the performance of SIR. The predictors are quite skewed and prone to outliers.

For each subpopulation, we generated n points. In each simulation, we used $h = 4$ slices within each subpopulation and compared the results over 1000 runs. Estimation accuracy was assessed for each method by computing the R^2's between each of the sufficient predictors, say, X_1, $X_2 + X_3$ and X_4 for the version 1 of Model B, and X_1, $X_2 + X_3$ for the version 2 of Model B, and their fitted values from the linear regressions on the d estimated sufficient predictors.

The curves shown in Fig. 4a are plots of the average R_1^2's from 1000 replications for the 3D model versus the subpopulation sample size n, for values of n between 100 and 1600. The plots for R_{23}^2 and R_4^2's look similar. As we can see, OPIRE always did better in estimation of the partial CS, although not by a lot in this example. Fig. 5a gives the average of R_1^2 from 1000 simulation runs for the 2D model versus the subpopulation sample size n. OPIRE showed obvious advantages over partial SIR in this case. Figs. 4b and 5b are discussed in the next section.

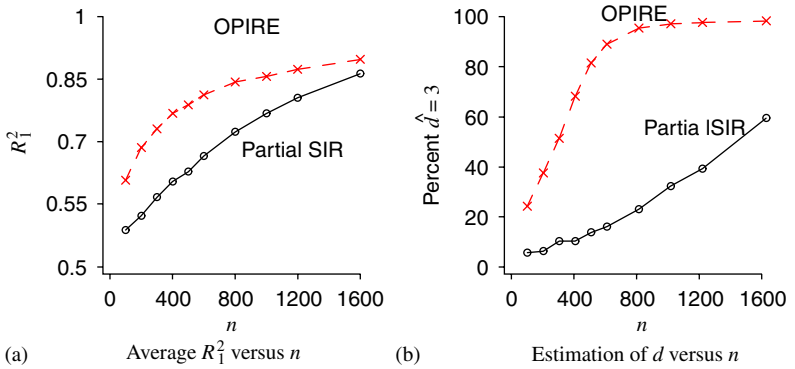

Fig. 5. Model B-2D: (a) Estimation accuracy of Model B versus n. (b) Percentage of runs in which $\hat{d} = 2$ versus n.

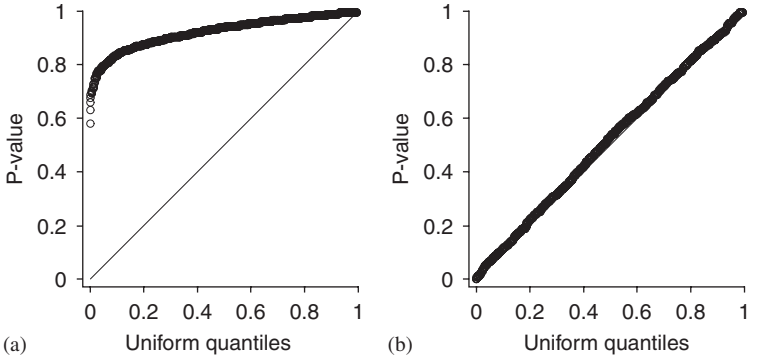

Fig. 6. Uniform quantile plot of p-values testing $H_0 : d = 2$ for Model A. (a) Partial SIR; (b) OPIRE

5.2. Estimation of d

Estimation of d is often based on testing a sequence of hypotheses $\mathrm{H}_0 : d = m$ versus $\mathrm{H}_a : d > m$, with m incremented by 1 until the hypothesis is not rejected. At which point $\hat{d}$ is the last value of m tested. First we consider the actual levels of the test under the null. Fig. 6 shows uniform quantile plots of p-values for testing $\mathrm{H}_0 : d = 2$ in reference Model A with 1000 replications. The sampling distribution of OPIRE's test statistics is much closer to the asymptotic one, suggesting a close agreement between the actual and nominal levels.

Fig. 7 shows the percentage of correct estimates $\hat{d} = 2$ in Model A from 1000 replications versus varying sample sizes, the mean multiplier m and σ_1 at test level 0.05. With $\sigma_1 = 5$ and $m = 1$, partial SIR did much worse than OPIRE. Partial SIR only gave 12% of correct dimension estimates $\hat{d} = 2$ when $n_y = 400$. Meanwhile, the frequency of correct decisions for the OPIRE estimate was close to 95% for large n_{wy}, which indicates that the OPIRE test level is close to its nominal value. From Fig. 7b we see that the OPIRE estimator responded much faster to increasing signal than did partial SIR. The frequency of correct decisions for partial SIR is greater than 95% for large values of m, indicating that partial SIR's actual test levels are less than the nominal level in this example. We can also see that the OPIRE estimator was immune to changes in σ_1, while the partial SIR estimator was very sensitive to such changes.

For Model B, the curves of Figs. 4b and 5b show that OPIRE did a lot better on estimating d. For the 3D model, when $n = 800$, with a test level of 0.05, OPIRE got the right dimension about 87% of the time, while partial SIR got the right answers only 59% of the time. For the 2D model, OPIRE made close to 95% correct decisions for large n, which is much better than the performance of partial SIR.

The above simulation results suggest that there can be substantial differences between the OPIRE and partial SIR estimator of d, and OPIRE outperformed partial SIR.

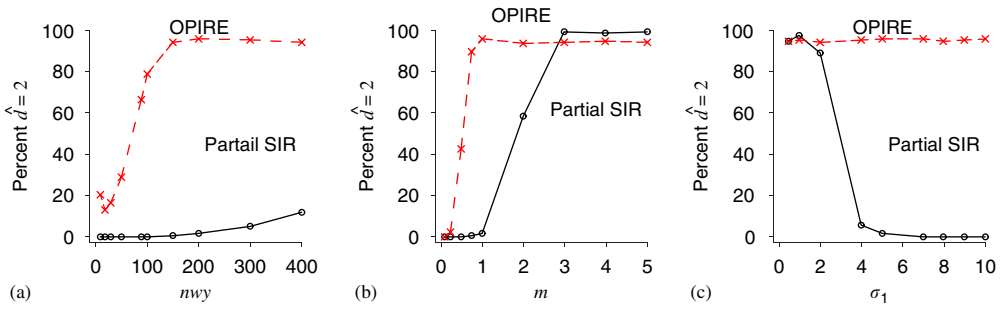

Fig. 7. Percentage of runs in which $\hat{d}=2$ versus (a) n_{wy}, (b) m and (c) σ_1 for Model A.

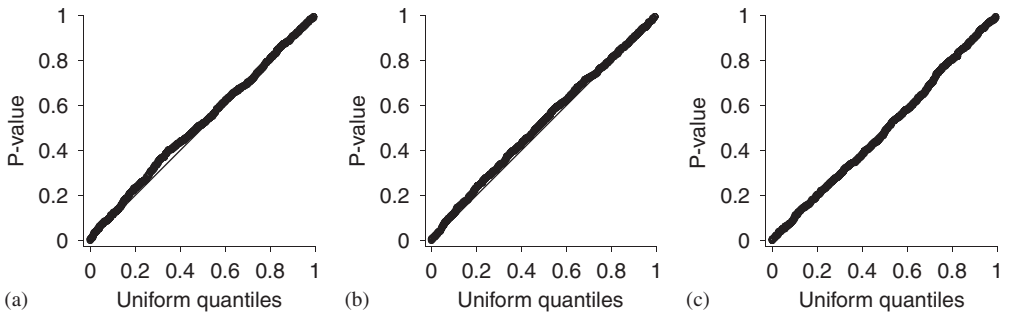

Fig. 8. Uniform quantile plots of p-values for predictor tests about X_5 in Model A. (a) Marginal; (b) Joint, with $d=2$ and (c) Conditional, given $d=2$

5.3. Predictor test

As we discussed in Section 4, one advantage of OPIRE is that it gives us the ability to test the predictor effects. As shown in Fig. 8, the simulation results agreed with our theory nicely. All uniform quantile plots of the p-values from testing the relevance of X_5 in the reference version of Model A with $n_y=200$ lie close to the straight line, indicating that the OPIRE predictor test levels are close to the nominal values. Simulation results from Model B which are not shown also gave good support to our theory.

6. Discussion

In the context of sufficient dimension reduction, a minimum discrepancy approach is introduced for regressions with a mix of continuous and categorical predictors. We proposed an optimal method, optimal partial inverse regression estimation, which removes the *homogeneous covariance condition* required by partial SIR. This new approach gives an optimal estimate of the partial CS, and it always provides an asymptotic chi-squared distribution for testing the dimension of the partial CS. It also allows us to consider testing the effects of the predictors. We also re-derived partial SIR method via the minimum discrepancy approach.

The development of sliced average variance estimation (SAVE) has lagged behind SIR due to the technical difficulty of dealing with the distribution of eigenvalues of quadratic functions of covariance matrices. Cook and Ni (2005) re-derived SAVE via the minimum discrepancy approach. They developed an asymptotic test statistic for testing SAVE's dimension. We can apply our approach directly to SAVE for partial sufficient dimension reduction to obtain an optimal version of partial SAVE. Another possible method is to combine information from first and second moments via the minimum discrepancy approach to get an optimal method for sufficient dimension reduction with categorical predictors.

It will be interesting to investigate methods for dealing with categories having a factorial structure. In this case, though we can arrange the various combinations of levels into a single variable W and apply the same methodology, with many factors, this procedure will require a large overall sample size, and the results may be difficult to interpret. One way around such difficulties would be to limit considerations to "additive effects", similar in spirit to additive ANOVA models. We may extend the partial one-dimensional models of Cook and Weisberg (2004) to include factorial structure and multiple linear combinations. Research along these lines is underway.

Appendix

Proof of Lemma 1. Let $\boldsymbol{\rho} \in \mathbb{R}^{p\times q}$ be an orthonormal basis for $\mathcal{S}_{Y|\mathbf{Z}}$, where $q = \dim(\mathcal{S}_{Y|\mathbf{Z}})$.

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

So, every vector in $\mathcal{S}_{E(\mathbf{Z}|Y)}$ can be written as an average of vectors in $\mathcal{S}_{\boldsymbol{\rho}}$. Thus, $\mathcal{S}_{E(\mathbf{Z}|Y)} \subseteq \mathcal{S}_{\boldsymbol{\rho}}$.

Also,

$$\boldsymbol{\rho} = E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\rho}) = E[E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\rho}|\boldsymbol{\rho}^{\mathrm{T}}\mathbf{Z})] = E[E(\mathbf{Z}|\boldsymbol{\rho}^{\mathrm{T}}\mathbf{Z})\mathbf{Z}^{\mathrm{T}}\boldsymbol{\rho}].$$

Thus, every vector in $\mathcal{S}_{Y|\mathbf{Z}}$ can be written as an average of vectors in $\mathcal{S}_{\boldsymbol{\rho}}$ too. Hence, $\mathcal{S}_{Y|\mathbf{Z}} \subseteq \mathcal{S}_{\boldsymbol{\rho}}$. □

Proof of Lemma 2. Let $\boldsymbol{\rho}$ be an orthonormal basis for $\mathcal{S}_{Y|\mathbf{Z}}$, and $\boldsymbol{\gamma}$ be an orthonormal basis for $\mathcal{S}_{Y|P_{\mathcal{S}}\mathbf{Z}}$.

$$Y \perp\!\!\!\perp \mathbf{Z} \mid \boldsymbol{\rho}^{\mathrm{T}}\mathbf{Z} \Rightarrow Y \perp\!\!\!\perp P_{\mathcal{S}}\mathbf{Z} \mid \boldsymbol{\rho}^{\mathrm{T}}P_{\mathcal{S}}\mathbf{Z} \Rightarrow \mathcal{S}_{Y|P_{\mathcal{S}}\mathbf{Z}} \subseteq \mathcal{S}_{Y|\mathbf{Z}}.$$

Also,

$$\begin{aligned} Y \perp\!\!\!\perp P_{\mathcal{S}}\mathbf{Z} \mid (\boldsymbol{\gamma}^{\mathrm{T}}(P_{\mathcal{S}}\mathbf{Z})) \quad \text{and} \quad Y \perp\!\!\!\perp \mathbf{Z} \mid P_{\mathcal{S}}\mathbf{Z} &\Rightarrow Y \perp\!\!\!\perp \mathbf{Z} \mid (\boldsymbol{\gamma}^{\mathrm{T}}(P_{\mathcal{S}}\mathbf{Z})) \\ &\Rightarrow Y \perp\!\!\!\perp \mathbf{Z} \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z} \\ &\Rightarrow \mathcal{S}_{Y|\mathbf{Z}} \subseteq \mathcal{S}_{Y|P_{\mathcal{S}}\mathbf{Z}}. \end{aligned}$$

Therefore, $\mathcal{S}_{Y|P_{\mathcal{S}}\mathbf{Z}} = \mathcal{S}_{Y|\mathbf{Z}}$. □

Proof of Lemma 3. Conditioning on subpopulation w, let $\boldsymbol{\eta}_w$ be a basis for $\mathcal{S}_{Y_w|\mathbf{X}_w}$. Define the following predictor subspace:

$$\mathcal{S}_{\boldsymbol{\eta}_w} \equiv \mathrm{Span}\{\boldsymbol{\Sigma}_w^{-1}(E(\mathbf{X}|\boldsymbol{\eta}_w^{\mathrm{T}}\mathbf{X} = v) - E(\mathbf{X})),\ v \text{ varies}\}.$$

Since every vector in $\mathcal{S}_{\boldsymbol{\xi}_w}$ can be written as an average of vectors in $\mathcal{S}_{\boldsymbol{\eta}_w}$, we have $\mathcal{S}_{\boldsymbol{\xi}_w} \subseteq \mathcal{S}_{\boldsymbol{\eta}_w}$. Also, every column of $\boldsymbol{\eta}_w$ can be written as an average of vectors in $\mathcal{S}_{\boldsymbol{\eta}_w}$ and consequently $\mathcal{S}_{Y_w|\mathbf{X}_w} \subseteq \mathcal{S}_{\boldsymbol{\eta}_w}$. Under the generalized coverage condition, $\mathcal{S}_{\boldsymbol{\xi}_w} = \mathcal{S}_{\boldsymbol{\eta}_w}$, we then have $\mathcal{S}_{Y_w|\mathbf{X}_w} \subseteq \mathcal{S}_{\boldsymbol{\xi}_w}$.

CCL established that $\mathcal{S}^{(W)}_{Y|\mathbf{X}} = \sum_{w=1}^{K} \mathcal{S}_{Y_w|\mathbf{X}_w}$. Also, by definition, $\mathcal{S}_{\boldsymbol{\xi}} \equiv \sum_{w=1}^{K} \mathcal{S}_{\boldsymbol{\xi}_w}$. Thus, with the generalized coverage condition, we have $\mathcal{S}^{(W)}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{\boldsymbol{\xi}}$. □

Proof of Theorem 1. Conditioning on subpopulation w, we will decompose $\sqrt{n_w}(\mathrm{vec}(\hat{\boldsymbol{\xi}}_w\mathbf{D}_{\hat{\mathbf{f}}_w}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}_w\mathbf{D}_{\mathbf{f}_w}))$ as a summation of i.i.d. observations plus a remainder converging to 0 in probability, and then apply the central limit theorem. In order to proceed with the decomposition, we need the following lemma that decomposes the difference between the inverse of a sample covariance matrix and its population value within each subpopulation. This lemma follows from Li et al. (2003). □

Lemma 4. *Suppose a random vector* $\mathbf{X}_w \in \mathbb{R}^p$ *has covariance matrix* $\boldsymbol{\Sigma}_w > 0$. *Then*,

$$\hat{\boldsymbol{\Sigma}}_w^{-1} - \boldsymbol{\Sigma}_w^{-1} = -n_w^{-1}\boldsymbol{\Sigma}_w^{-1/2}\sum_{j=1}^{n_w}(\mathbf{Z}_w^{(j)}\mathbf{Z}_w^{(j)\mathrm{T}} - I)\boldsymbol{\Sigma}_w^{-1/2} + \mathrm{O}_p(n_w^{-1}),$$

where $\hat{\boldsymbol{\Sigma}}_w$ *is the sample covariance calculated from a sample of size* n_w *and* $\mathbf{Z}_w = \boldsymbol{\Sigma}_w^{-1/2}(\mathbf{X}_w - E(\mathbf{X}_w))$ *is the standardized version of* $\mathbf{X}_w$.

In Section 2.2, we define $\bar{\mathbf{X}}_{wy\bullet}$ as the average of the n_{wy} observations in the Y_wth slice and $\bar{\mathbf{X}}_{w\bullet\bullet}$ as the average of all n_w observations in subpopulation w. Let $\boldsymbol{\mu}_{wy} = E(\bar{\mathbf{X}}_{wy\bullet})$, $\boldsymbol{\mu} = E(\bar{\mathbf{X}}_{w\bullet\bullet})$, consider

$$\begin{aligned}
&\sqrt{n_w}(\hat{f}_{wy}\hat{\boldsymbol{\xi}}_{wy} - f_{wy}\boldsymbol{\xi}_{wy})\\
&\quad= \sqrt{n_w}\hat{f}_{wy}\hat{\boldsymbol{\Sigma}}_w^{-1}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet}) - \sqrt{n_w}f_{wy}\boldsymbol{\Sigma}_w^{-1}(\boldsymbol{\mu}_{wy} - \boldsymbol{\mu}_w)\\
&\quad= \sqrt{n_w}(\hat{\boldsymbol{\Sigma}}_w^{-1} - \boldsymbol{\Sigma}_w^{-1})f_{wy}(\boldsymbol{\mu}_{wy} - \boldsymbol{\mu}_w) + \sqrt{n_w}\boldsymbol{\Sigma}_w^{-1}[\hat{f}_{wy}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet}) - f_{wy}(\boldsymbol{\mu}_{wy} - \boldsymbol{\mu}_w)]\\
&\qquad+ \sqrt{n_w}(\hat{\boldsymbol{\Sigma}}_w^{-1} - \boldsymbol{\Sigma}_w^{-1})[\hat{f}_{wy}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet}) - f_{wy}(\boldsymbol{\mu}_{wy} - \boldsymbol{\mu}_w)]\\
&\quad= \sqrt{n_w}(\hat{\boldsymbol{\Sigma}}_w^{-1} - \boldsymbol{\Sigma}_w^{-1})f_{wy}(\boldsymbol{\mu}_{wy} - \boldsymbol{\mu}_w) + \sqrt{n_w}\boldsymbol{\Sigma}_w^{-1}[\hat{f}_{wy}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet}) - f_{wy}(\boldsymbol{\mu}_{wy} - \boldsymbol{\mu}_w)]\\
&\qquad+ \mathrm{O}_p(n_w^{-1/2}).
\end{aligned} \tag{12}$$

We next use Lemma 4 to rewrite the first term in (12) as

$$\begin{aligned}
\sqrt{n_w}(\hat{\boldsymbol{\Sigma}}_w^{-1} - \boldsymbol{\Sigma}_w^{-1})f_{wy}(\boldsymbol{\mu}_{wy} - \boldsymbol{\mu}_w) &= -n_w^{-1/2}\boldsymbol{\Sigma}_w^{-1/2}\sum_{j=1}^{n_w}(\mathbf{Z}_w^{(j)}\mathbf{Z}_w^{(j)\mathrm{T}} - I)\boldsymbol{\Sigma}_w^{-1/2}f_{wy}(\boldsymbol{\mu}_{wy} - \boldsymbol{\mu}_w) + \mathrm{O}_p(n_w^{-1/2})\\
&= -n_w^{-1/2}\boldsymbol{\Sigma}_w^{-1/2}\sum_{j=1}^{n_w}(\mathbf{Z}_w^{(j)}\mathbf{Z}_w^{(j)\mathrm{T}} - I)E(\mathbf{Z}_w J_{wy}) + \mathrm{O}_p(n_w^{-1/2}).
\end{aligned} \tag{13}$$

Denote $J_{wy}^{(j)}$ be the value of J_{wy} for the jth observation in subpopulation w, $j = 1, 2, \ldots, n_w$, then

$$\begin{aligned}
\hat{f}_{wy}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet}) &= \frac{1}{n_w}\sum_{j=1}^{n_w}[(\mathbf{X}_w^{(j)} - \bar{\mathbf{X}}_{w\bullet\bullet})J_{wy}^{(j)}]\\
&= \frac{1}{n_w}\sum_{j=1}^{n_w}[(\mathbf{X}_w^{(j)} - \bar{\mathbf{X}}_{w\bullet\bullet})(J_{wy}^{(j)} - E(J_{wy}))]\\
&= \frac{1}{n_w}\sum_{j=1}^{n_w}[(\mathbf{X}_{jw} - \boldsymbol{\mu}_w)(J_{wy}^{(j)} - E(J_{wy}))] - \frac{1}{n_w}\sum_{j=1}^{n_w}[(\bar{\mathbf{X}}_{w\bullet\bullet} - \boldsymbol{\mu}_w)(J_{wy}^{(j)} - E(J_{wy}))]\\
&= \frac{1}{n_w}\sum_{j=1}^{n_w}[(\mathbf{X}_w^{(j)} - \boldsymbol{\mu}_w)(J_{wy}^{(j)} - E(J_{wy}))] - \frac{1}{n_w}(\bar{\mathbf{X}}_{w\bullet\bullet} - \boldsymbol{\mu}_w)\sum_{j=1}^{n_w}(J_{wy}^{(j)} - E(J_{wy}))\\
&= \frac{1}{n_w}\sum_{j=1}^{n_w}[(\mathbf{X}_w^{(j)} - \boldsymbol{\mu}_w)(J_{wy}^{(j)} - E(J_{wy}))] + \mathrm{O}_p(n_w^{-1}).
\end{aligned}$$

Now, we can simplify the second term in (12) as

$$\begin{aligned}
&\sqrt{n_w}\boldsymbol{\Sigma}_w^{-1}[\hat{f}_{wy}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet}) - f_{wy}(\boldsymbol{\mu}_{wy} - \boldsymbol{\mu}_w)]\\
&\quad= n_w^{-1/2}\boldsymbol{\Sigma}_w^{-1/2}\sum_{j=1}^{n_w}[\boldsymbol{\Sigma}_w^{-1/2}(\mathbf{X}_w^{(j)} - \boldsymbol{\mu}_w)(J_{wy}^{(j)} - E(J_{wy}))] - \sqrt{n_w}\boldsymbol{\Sigma}_w^{-1}f_{wy}(\boldsymbol{\mu}_{wy} - \boldsymbol{\mu}_w) + \mathrm{O}_p(n_w^{-1/2})\\
&\quad= n_w^{-1/2}\boldsymbol{\Sigma}_w^{-1/2}\sum_{j=1}^{n_w}[\mathbf{Z}_w^{(j)}(J_{wy}^{(j)} - E(J_{wy}))] - \sqrt{n_w}\boldsymbol{\Sigma}_w^{-1/2}E(\mathbf{Z}_w J_{wy}) + \mathrm{O}_p(n_w^{-1/2})\\
&\quad= n_w^{-1/2}\boldsymbol{\Sigma}_w^{-1/2}\sum_{j=1}^{n_w}[\mathbf{Z}_w^{(j)}(J_{wy}^{(j)} - E(J_{wy})) - E(\mathbf{Z}_w J_{wy})] + \mathrm{O}_p(n_w^{-1/2}).
\end{aligned} \tag{14}$$

Plugging (13) and (14) into (12), we get

$$\begin{aligned}\sqrt{n_w}(\hat{f}_{wy}\hat{\boldsymbol{\xi}}_{wy} - f_{wy}\boldsymbol{\xi}_{wy}) &= n_w^{-1/2}\boldsymbol{\Sigma}_w^{-1/2}\sum_{j=1}^{n_w}[\mathbf{Z}_w^{(j)}(J_{wy}^{(j)} - E(J_{wy})) - E(\mathbf{Z}_w J_{wy})\\ &\quad - (\mathbf{Z}_w^{(j)}\mathbf{Z}_w^{(j)\mathrm{T}} - I)E(\mathbf{Z}_w J_{wy})] + \mathrm{O}_p(n_w^{-1/2})\\ &= n_w^{-1/2}\boldsymbol{\Sigma}_w^{-1/2}\sum_{j=1}^{n_w}[\mathbf{Z}_w^{(j)}(J_{wy}^{(j)} - E(J_{wy}) - \mathbf{Z}_w^{(j)\mathrm{T}}E(\mathbf{Z}_w J_{wy}))] + \mathrm{O}_p(n_w^{-1/2})\\ &= n_w^{-1/2}\boldsymbol{\Sigma}_w^{-1/2}\sum_{j=1}^{n_w}[\mathbf{Z}_w^{(j)}\varepsilon_{wy}^{(j)}] + \mathrm{O}_p(n_w^{-1/2}),\end{aligned}$$

where $\varepsilon_{wy}^{(j)} = J_{wy}^{(j)} - E(J_{wy}) - \mathbf{Z}_w^{(j)\mathrm{T}}E(\mathbf{Z}_w J_{wy})$ is the jth value for ε_{wy}. Let $\boldsymbol{\epsilon}_w^{(j)} = [\varepsilon_{w_1}^{(j)}, \ldots, \varepsilon_{w_{hw}}^{(j)}]^{\mathrm{T}}$ be the jth value for the random vector $\boldsymbol{\epsilon}_w$. We have

$$\sqrt{n_w}(\mathrm{vec}(\hat{\boldsymbol{\xi}}_w\mathbf{D}_{\hat{\mathbf{f}}_w}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}_w\mathbf{D}_{\mathbf{f}_w})) = n_w^{-1/2}\sum_{j=1}^{n_w}\mathrm{vec}(\boldsymbol{\Sigma}_w^{-1/2}\mathbf{Z}_w^{(j)}(\boldsymbol{\epsilon}_w^{(j)})^{\mathrm{T}}) + \mathrm{O}_p(n_w^{-1/2}),$$

where $\mathrm{vec}(\boldsymbol{\Sigma}_w^{-1/2}\mathbf{Z}_w^{(j)}\boldsymbol{\epsilon}_w^{(j)})$ are i.i.d. random vectors. Thus,

$$\sqrt{n_w}(\mathrm{vec}(\hat{\boldsymbol{\xi}}_w\mathbf{D}_{\hat{\mathbf{f}}_w}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}_w\mathbf{D}_{\mathbf{f}_w})) \xrightarrow{\mathscr{D}} \mathrm{Normal}(0, \boldsymbol{\Gamma}_w),$$

where

$$\boldsymbol{\Gamma}_w = \mathrm{Cov}(\mathrm{vec}(\boldsymbol{\Sigma}_w^{-1/2}\mathbf{Z}_w\boldsymbol{\epsilon}_w^{\mathrm{T}})).$$

The second part of Theorem 1 can be easily derived following Slutsky's Theorem.

Proof of Theorem 2. The proof of Theorem 2 hinges on Shapiro's (1986) results on asymptotics of over-parameterized discrepancy functions and two supplemental lemmas (Cook and Ni, 2005). The discrepancy functions that Shaprio considered are

$$H(\boldsymbol{\tau}_n, g(\theta)) = (\boldsymbol{\tau}_n - g(\theta))^{\mathrm{T}}\mathbf{V}(\boldsymbol{\tau}_n - g(\theta)), \tag{15}$$

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

Instead of using a constant inner-product matrix as required in Shapiro's results, we adapt his results to include random inner-product matrices. The following two lemmas make that possible. Lemma 5 deals with the asymptotic distribution of the minimum discrepancy value. Lemma 6 gives the asymptotic properties of the estimate of Span($\boldsymbol{\beta}$) based on a generalized result of the modified χ^2 method in Ferguson (1958).

Lemma 5. *Let $\{\mathbf{Y}_n\} \in \mathbb{R}^s$ be a sequence of random vectors, and let $\boldsymbol{\xi} \in \Xi \subseteq \mathbb{R}^s$. Suppose $\{\mathbf{V}_n > 0\}$ is a sequence of $s \times s$ matrices that converges to $\mathbf{V} > 0$ in probability. If $n\hat{H}_{\mathbf{V}} = \min_{\xi\in\Xi} n(\mathbf{Y}_n - \boldsymbol{\xi})^{\mathrm{T}}\mathbf{V}(\mathbf{Y}_n - \boldsymbol{\xi})$ converges to a random variable Ψ in probability, then so does the $n\hat{H}_{\mathbf{V}_n} = \min_{\xi\in\Xi} n\,(\mathbf{Y}_n - \boldsymbol{\xi})^{\mathrm{T}}\mathbf{V}_n(\mathbf{Y}_n - \boldsymbol{\xi})$ and vice versa.*

Moreover, let $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ be the values of $\boldsymbol{\xi}$ which reach $n\hat{H}_{\mathbf{V}}$ and $n\hat{H}_{\mathbf{V}_n}$, respectively. If $\mathbf{V}^{1/2}\mathbf{Y}_n \xrightarrow{p} \boldsymbol{\alpha}$, then both $\mathbf{V}^{1/2}\boldsymbol{\xi}_1$ and $\mathbf{V}_n^{1/2}\boldsymbol{\xi}_2$ converge to $\boldsymbol{\alpha}$ in probability.

Lemma 6. *Let $\mathscr{X}_n$ denote a simple random sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, where $\mathbf{X}_i$ can be a scalar or a vector. The distribution of $\mathbf{X}$ depends on parameters that include a vector θ in $\Theta \subseteq \mathbb{R}^k$. Let θ_0 be the true value of θ. Assume that*

1. *Θ is an open set.*
2. *The mapping $p(\theta)$ from Θ into $\mathbb{R}^s$ is one-to-one, bicontinuous, and twice continuously differentiable. Let $\mathbf{D}(\theta) = \partial p(\theta)/\partial\theta \in \mathbb{R}^{s\times k}$, and $\mathbf{D}_0 = \mathbf{D}(\theta_0)$.*

3. $\mathbf{Y}_n = \mathbf{Y}_n(\mathscr{X}_n) \in \mathbb{R}^s$ *is a consistent estimate of* $p(\theta_0)$ *with*

$$\sqrt{n}(\mathbf{Y}_n - p(\theta_0)) \xrightarrow{\mathscr{D}} \text{Normal}(0, \mathbf{\Gamma}).$$

4. $\mathbf{V}_n = \mathbf{V}_n(\mathscr{X}_n)$ *is a positive definite matrix that converges to a constant matrix* $\mathbf{V}$ *in probability.*

Define a discrepancy function as

$$F(\mathbf{Y}_n, p(\theta)) = (\mathbf{Y}_n - p(\theta))^{\mathrm{T}}\mathbf{V}_n(\mathbf{Y}_n - p(\theta)).$$

Let $\hat{\theta} = \hat{\theta}(\mathscr{X}_n)$ *be the value of* θ *that minimizes F. Then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathscr{D}} \text{Normal}(0, (\mathbf{D}_0^{\mathrm{T}}\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0^{\mathrm{T}}\mathbf{V}\mathbf{\Gamma}\mathbf{V}\mathbf{D}_0(\mathbf{D}_0^{\mathrm{T}}\mathbf{V}\mathbf{D}_0)^{-1}),$$

and

$$\sqrt{n}(p(\hat{\theta}) - p(\theta_0)) \xrightarrow{\mathscr{D}} \text{Normal}(0, \mathbf{D}_0(\mathbf{D}_0^{\mathrm{T}}\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0^{\mathrm{T}}\mathbf{V}\mathbf{\Gamma}\mathbf{V}\mathbf{D}_0(\mathbf{D}_0^{\mathrm{T}}\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0).$$

We now begin the proof of Theorem 2. Based on Lemma 5, the asymptotic distribution of n times the minimum value of (5) is the same as the asymptotic distribution of n times the minimum value of H_d, where

$$H_d(\mathbf{B}, \mathbf{C}) = (\text{vec}(\hat{\boldsymbol{\zeta}}) - \text{vec}(\mathbf{BC}))^{\mathrm{T}}\mathbf{\Gamma}_{\hat{\boldsymbol{\zeta}}}^{-1}(\text{vec}(\hat{\boldsymbol{\zeta}}) - \text{vec}(\mathbf{BC})).$$

Cook and Ni dealt with sufficient dimension reduction with only continuous predictors using Lemma 6. The problem we are studying is more intricate due to the introduction of categorical predictors. Based on Lemma 6, we know that the asymptotic distribution of $\text{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}})$ of (5) is the same as that of $H_d(\mathbf{B}, \mathbf{C})$. Hence, we need only to show that there is one parameterization that satisfies the conditions in the statement of Lemma 6. We are free to use any full rank re-parameterization of $(\boldsymbol{\beta}, \boldsymbol{\nu})$. Here is one possible choice: $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_1 \in \mathbb{R}^{d\times d}$, $\boldsymbol{\beta}_2 \in \mathbb{R}^{(p-d)\times d}$. Without loss of generality, we assume that $\boldsymbol{\beta}_1$ is nonsingular. Then,

$$\boldsymbol{\beta}\boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \boldsymbol{\nu} = \begin{pmatrix} I_d \\ \boldsymbol{\beta}_2\boldsymbol{\beta}_1^{-1} \end{pmatrix} \boldsymbol{\beta}_1\boldsymbol{\nu}.$$

Thus, we can set $\boldsymbol{\beta}_1 = I_d$, leading to the new parameters $\boldsymbol{\beta}_2 \in \mathbb{R}^{(p-d)\times d}$ and $\boldsymbol{\nu} \in \mathbb{R}^{d\times(h-K)}$, which together corresponds to the θ in Lemma 6. This new parameterization leads to a full rank Jacobian matrix and an open parameter space in $\mathbb{R}^{d(h+p-d-K)}$, thus satisfying the conditions in Lemma 6. Meanwhile, it affects neither our algorithm for minimization or asymptotic results. Hence, it suffices to prove the conclusions for H_d.

The following setting makes it clear that $H_d(\mathbf{B}, \mathbf{C})$ is in the form of Shapiro's discrepancy function H:

$$\theta = \begin{pmatrix} \text{vec}(\mathbf{B}) \\ \text{vec}(\mathbf{C}) \end{pmatrix} \in \mathbb{R}^{d(p+h-K)},$$

$$h(\theta) = \text{vec}(\mathbf{BC}) \in \mathbb{R}^{p(h-K)},$$

$$\boldsymbol{\tau}_n = \text{vec}(\hat{\boldsymbol{\zeta}}),$$

$$h(\theta_0) = \text{vec}(\boldsymbol{\beta}\boldsymbol{\nu}),$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ is in general a basis for $\mathscr{S}_{\boldsymbol{\xi}}$ and $\boldsymbol{\nu} \in \mathbb{R}^{d\times(h-K)}$. Following from Shapiro (1986), we then have $\text{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}})$ of $H_d(\mathbf{B}, \mathbf{C})$ is asymptotically efficient with

$$\sqrt{n}(\text{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}}) - \text{vec}(\boldsymbol{\beta}\boldsymbol{\nu})) \xrightarrow{\mathscr{D}} \text{Normal}(0, \mathbf{\Delta}_{\boldsymbol{\zeta}}(\mathbf{\Delta}_{\boldsymbol{\zeta}}^{\mathrm{T}}\mathbf{V}\mathbf{\Delta}_{\boldsymbol{\zeta}})^{-}\mathbf{\Delta}_{\boldsymbol{\zeta}}),$$

which leads to the conclusion 1 of Theorem 2. And $n\hat{H}$ has an asymptotic chi-squared distribution with degrees of freedom $p(h-K) - \text{rank}(\mathbf{\Delta}_{\boldsymbol{\zeta}})$, where

$$\begin{aligned} \text{rank}(\mathbf{\Delta}_{\boldsymbol{\zeta}}) &= \text{rank}(\boldsymbol{\nu}^{\mathrm{T}} \otimes \mathbf{Q}_{\boldsymbol{\beta}}, I_{h-K} \otimes \boldsymbol{\beta}) \\ &= d \times (p-d) + d \times (h-K) \\ &= d(p+h-d-K). \end{aligned}$$

Therefore, the degrees of freedom are $p(h-K)-d(p+h-d-K)=(p-d)(h-d-K)$. Thus, conclusion 2 is proved. It is easy to get consistency of Span($\hat{\boldsymbol{\beta}}$) in conclusion 3 following the conclusion 1. □

Proof of Proposition 6. *Let the columns of the $p\times r$ matrix $\boldsymbol{\alpha}$ be a basis for $\mathscr{H}$, and let the columns of $p\times d$ matrix $\boldsymbol{\eta}$ be a basis for $\mathscr{S}_{Y|\mathbf{X}}^{(W)}$.*

"$\Longleftarrow$"

$Y \perp\!\!\!\perp P_{\mathscr{H}}\mathbf{X} \mid (Q_{\mathscr{H}}\mathbf{X}, W) \Rightarrow Y \perp\!\!\!\perp (P_{\mathscr{H}}\mathbf{X}, Q_{\mathscr{H}}\mathbf{X}) \mid (Q_{\mathscr{H}}\mathbf{X}, W) \Rightarrow Y \perp\!\!\!\perp \mathbf{X} \mid (Q_{\mathscr{H}}\mathbf{X}, W)$. Therefore, Span($Q_{\mathscr{H}}$) satisfies (1). Since the partial CS is assumed to exist, $\text{Span}(Q_{\mathscr{H}}) \supseteq \mathscr{S}_{Y|\mathbf{X}}^{(W)}$. Therefore, $P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} = \mathscr{O}_p$.

"$\Longrightarrow$"

$$Y \perp\!\!\!\perp \mathbf{X} \mid (\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}, W) \Leftrightarrow Y \perp\!\!\!\perp (P_{\mathscr{H}}\mathbf{X}, Q_{\mathscr{H}}\mathbf{X}) \mid (\boldsymbol{\eta}^{\mathrm{T}}(Q_{\mathscr{H}}\mathbf{X} + P_{\mathscr{H}}\mathbf{X}), W)$$

$$P_{\mathscr{H}}\mathscr{S}_{Y|\mathbf{X}}^{(W)} = \mathscr{O}_p \Rightarrow Y \perp\!\!\!\perp (P_{\mathscr{H}}\mathbf{X}, Q_{\mathscr{H}}\mathbf{X}) \mid (\boldsymbol{\eta}^{\mathrm{T}}Q_{\mathscr{H}}\mathbf{X}, W).$$

Thus, $Y \perp\!\!\!\perp P_{\mathscr{H}}\mathbf{X} \mid (Q_{\mathscr{H}}\mathbf{X}, W)$. □

References

Browne, M.W., 1984. Asymptotic distribution-free methods for the analysis of covariance structures. British J. Math. Statist. Psychology 37, 62–83.

Carroll, R., Ruppert, D., 1988. Transformation and Weighting in Regression. Chapman & Hall, London.

Chiaromonte, F., Cook, R.D., 2002. Sufficient dimension reduction and graphics in regression. Ann. Inst. Statist. Math. 54, 768–795.

Chiaromonte, F., Cook, R.D., Li, B., 2002. Sufficient dimension reduction in regressions with categorical predictors. Ann. Statist. 30, 475–497.

Cook, R.D., 1994. Using dimension-reduction subspaces to identify important inputs in models of physical systems. In: Proceedings of Section on Physical and Engineering Sciences.American Statistical Association, Alexandria, VA, pp. 18–25.

Cook, R.D., 1998. Regression Graphics. Wiley, New York, NY.

Cook, R.D., 2000. SAVE: a method for dimension reduction and graphics in regression. Comm. Statist.: Theory Methods 29, 161–175.

Cook, R.D., 2004. Testing predictor contributions in sufficient dimension reduction. Ann. Statist. 32, 1062–1092.

Cook, R.D., Nachtsheim, C.J., 1994. Reweighting to achieve elliptically contoured covariates in regression. J. Amer. Statist. Assoc. 89, 592–599.

Cook, R.D., Ni, L., 2005. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. J. Amer. Statist. Assoc. 100, 410–428.

Cook, R.D., Weisberg, S., 1991. Discussion of sliced inverse regression for dimension reduction. J. Amer. Statist. Assoc. 86, 316–342.

Cook, R.D., Weisberg, S., 2004. Partial one-dimensional regression models. Amer. Statist. 58, 110–116.

Ferguson, T., 1958. A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. Ann. Math. Statist. 29, 1046–1062.

Friedman, J.H., 1994. An overview of predictive learning and function approximation, from statistics to neural networks. In: Cherkassy, V., Friedman, J.H., Wechsler, H. (Eds.), NATO ASI Series F, vol. 136. Springer, New York.

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

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

Li, B., Cook, R.D., Chiaromonte, F., 2003. Dimension reduction for the conditional mean in regressions with categorical predictors. Ann. Statist. 31, 1636–1668.

Lindsay, B., Qu, A., 2003. Inference functions and quadratic score tests. Statist. Sci. 18, 394–410.

Ruhe, A., Wedin, P.A., 1980. Algorithms for separable nonlinear least squares problems. SIAM Review, 22, 318–337.

Shapiro, A., 1986. Asymptotic theory of overparameterized structural model. J. Amer. Statist. Assoc. 81, 142–149.

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