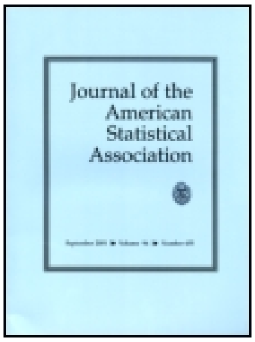

Journal of the American Statistical Association

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://amstat.tandfonline.com/loi/uasa20

Sufficient Dimension Reduction via Inverse Regression

R. Dennis Cook & Liqiang Ni

To cite this article: R. Dennis Cook & Liqiang Ni (2005) Sufficient Dimension Reduction via Inverse Regression, Journal of the American Statistical Association, 100:470, 410-428, DOI: 10.1198/016214504000001501

To link to this article: http://dx.doi.org/10.1198/016214504000001501

Published online: 01 Jan 2012.

Submit your article to this journal

Article views: 448

Citing articles: 106 View citing articles

Full Terms & Conditions of access and use can be found at
http://amstat.tandfonline.com/action/journalInformation?journalCode=uasa20

Download by: [University of Illinois, Chicago] **Date:** 06 October 2017, At: 07:20

Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach

R. Dennis COOK and Liqiang NI

A family of dimension-reduction methods, the inverse regression (IR) family, is developed by minimizing a quadratic objective function. An optimal member of this family, the inverse regression estimator (IRE), is proposed, along with inference methods and a computational algorithm. The IRE has at least three desirable properties: (1) Its estimated basis of the central dimension reduction subspace is asymptotically efficient, (2) its test statistic for dimension has an asymptotic chi-squared distribution, and (3) it provides a chi-squared test of the conditional independence hypothesis that the response is independent of a selected subset of predictors given the remaining predictors. Current methods like sliced inverse regression belong to a suboptimal class of the IR family. Comparisons of these methods are reported through simulation studies. The approach developed here also allows a relatively straightforward derivation of the asymptotic null distribution of the test statistic for dimension used in sliced average variance estimation.

KEY WORDS: Inverse regression estimator; Sliced average variance estimation; Sliced inverse regression; Sufficient dimension reduction.

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

1. INTRODUCTION

In full generality, the goal of a regression is to infer about the conditional distribution of the univariate response Y given the $p \times 1$ vector of predictors $\mathbf{X}$. Many different statistical contexts have been developed to address this issue, ranging from classical linear models to, relatively recently, support vector regression. In this article we consider *sufficient dimension reduction* (SDR), the basic idea of which is to replace the predictor vector with its projection onto a subspace of the predictor space without loss of information on the conditional distribution of $Y|\mathbf{X}$. Potential advantages accrue from working in the SDR context, because no prespecified model for $Y|\mathbf{X}$ is required and the curse of dimensionality that can hinder other nonparametric methods is often avoided.

SDR is based on a population meta-parameter, the *central subspace* (CS) (Cook 1996), represented by $\mathcal{S}_{Y|\mathbf{X}}$ and defined as the intersection of all subspaces $\mathcal{S} \subseteq \mathbb{R}^p$ having the property $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}$, where $\perp\!\!\!\perp$ indicates independence and $\mathbf{P}_{\mathcal{S}}$ is the orthogonal projection onto $\mathcal{S}$ in the usual inner product. The statement is that Y is independent of $\mathbf{X}$ given $\mathbf{P}_{\mathcal{S}}\mathbf{X}$, and, consequently, $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ carries all of the information that $\mathbf{X}$ has about Y. The CS is a uniquely defined subspace of $\mathbb{R}^p$ that allows reduction of the predictor from $\mathbf{X}$ to $\boldsymbol{\eta}^T\mathbf{X}$, where $\boldsymbol{\eta}$ is a $p \times \dim(\mathcal{S}_{Y|\mathbf{X}})$ matrix whose columns $\boldsymbol{\eta}_k$ form a basis for $\mathcal{S}_{Y|\mathbf{X}}$. We call $\boldsymbol{\eta}_k^T\mathbf{X}$ a *sufficient predictor*, $k = 1, \dots, \dim(\mathcal{S}_{Y|\mathbf{X}})$. (For background on the existence of the CS and its properties and related issues, see Cook 1998a, chap. 6, and the references therein.) Cook and Weisberg (1999) gave an introductory account of studying regressions via central subspaces. A linear transformation of the predictor $\mathbf{X}$ leads to a linear transformation of the CS. For example, define the standardized predictor $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$, where $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X}) > \mathbf{0}$. Then $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}$. When convenient for exposition, we work in the $\mathbf{Z}$-scale without loss of generality.

Other dimension-reduction methods estimate the *central mean subspace* (Cook and Li 2002), which is a subspace of the CS that captures the mean function. These include ordinary least squares and related methods based on convex objective functions, principal Hessian directions (Li 1992; Cook 1998b), iterative Hessian transformation (Cook and Li 2002), and minimum average variance estimation (Xia, Tong, Li, and Zhu 2002). In this article we are concerned only with the CS.

Methods for estimating the CS include sliced inverse regression (SIR) (Li 1991), sliced average variance estimation (SAVE) (Cook and Weisberg 1991), graphical regression (Cook 1994, 1998a), parametric inverse regression (Bura and Cook 2001a), and partial SIR (Chiaromonte, Cook, and Li 2002) when categorical predictors are present. Among these methods, SIR and its variants are perhaps the most widely used. The original SIR methodology proposed by Li (1991) can be based on the following logic, which is discussed in more detail in Section 5. Under mild conditions on the marginal distribution of the predictor vector, $\mathrm{E}(\mathbf{Z}|Y) \in \mathcal{S}_{Y|\mathbf{Z}}$ and thus $\mathrm{Span}(\mathbf{M}_{\mathrm{SIR}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$, where $\mathbf{M}_{\mathrm{SIR}} = \mathrm{cov}(\mathrm{E}(\mathbf{Z}|Y))$. If $\dim(\mathcal{S}_{Y|\mathbf{Z}})$ is known, if $\mathrm{Span}(\mathbf{M}_{\mathrm{SIR}}) = \mathcal{S}_{Y|\mathbf{Z}}$, and if we have a consistent estimate $\widehat{\mathbf{M}}_{\mathrm{SIR}}$ of $\mathbf{M}_{\mathrm{SIR}}$, then the span of the eigenvectors corresponding to the $\dim(\mathcal{S}_{Y|\mathbf{Z}})$ largest eigenvalues of $\widehat{\mathbf{M}}_{\mathrm{SIR}}$ is a consistent estimate of $\mathcal{S}_{Y|\mathbf{Z}}$. The sum of the smallest $p - m$ eigenvalues of $\widehat{\mathbf{M}}_{\mathrm{SIR}}$ can be used to test the dimension hypothesis $\dim(\mathcal{S}_{Y|\mathbf{Z}}) = m$. We call this the *spectral decomposition approach*, because it is based on a spectral decomposition of the kernel matrix $\widehat{\mathbf{M}}_{\mathrm{SIR}}$. SAVE and other SDR methods are based on the same logic but use different kernel matrices.

SIR has generated considerable interest since it was introduced, and there are many studies that elaborate on Li's original methodology. Hsing and Carroll (1992), Zhu and Ng (1995), and Zhu and Fang (1996) studied methods for estimating $\mathbf{M}_{\mathrm{SIR}}$. Li's original dimension test was based on the assumption of normally distributed predictors. Schott (1994) investigated inference methods for dimension when the predictors follow an elliptically contoured distribution. Chen and Li (1998) discussed how to round out SIR methodology to make it feel more like linear regression. Bura and Cook (2001b) proposed a weighted chi-squared test, which extended the SIR test for more general situations, while using the same test statistic. Following Cook and Weisberg (1991), Cook and Yin (2001) developed permutation procedures for inference about $\dim(\mathbf{M}_{\mathrm{SIR}})$.

R. Dennis Cook is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: *dennis@stat.umn.edu*). Liqiang Ni is Assistant Professor, Department of Statistics and Actuarial Science, University of Central Florida, Orlando, FL 32816 (E-mail: *lni@mail.ucf.edu*). This research was supported in part by National Science Foundation grants DMS-01-03983 and DMS-04-05360. The authors thank the editor for his helpful guidance and the referees for their critiques.

© 2005 American Statistical Association
Journal of the American Statistical Association
June 2005, Vol. 100, No. 470, Theory and Methods
DOI 10.1198/016214504000001501

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

Hsing (1999) used nearest-neighbor methods to develop a version of SIR that is applicable with multivariate responses. Naik and Tsai (2000) compared the performance of SIR to partial least squares in the context of a single-index model. Cook and Critchley (2000) showed that SDR methods in general and SIR in particular can be useful for identifying outliers and regression mixtures. Assuming that $\dim(\mathcal{S}_{Y|\mathbf{Z}})$ is known, Gather, Hilker, and Becker (2002) developed a robust version of SIR. Bura (2003) used linear smoothers to estimate the inverse mean function, $\mathrm{E}(\mathbf{X}|Y)$. Bura's approach covers the developments by Fung, He, Li, and Shi (2002), although her estimates were not based on a canonical correlation analysis. Many of these and other studies focused on varying the ingredients in the spectral decomposition approach that characterizes the SIR methodology.

In contrast, we develop a family of dimension-reduction methods by minimizing quadratic discrepancy functions, referring to these methods collectively as the *inverse regression* (*IR*) *family*. We derive the optimal member of this family, the *inverse regression estimator* (IRE). The IRE is optimal in terms of asymptotic efficiency, and its test statistic for dimension always has an asymptotic chi-squared distribution. We also show that SIR is a suboptimal member of the IR family.

We review IR and introduce minimum discrepancy estimation for dimension reduction in Section 2. We introduce the IRE in Section 3. We show in Section 3.4 how to test conditional independence hypotheses involving the predictors, using a general testing structure for SDR recently proposed by Cook (2004). We turn to a suboptimal class within the IR family in Section 4. We discuss an important member of this suboptimal class, SIR, in Section 5. In Section 6 we compare SIR and IR methods through simulation studies. In Section 7 we revisit SAVE, which uses second inverse moments to estimate the CS. Although it is generally recognized that SAVE and SIR are complementary methods, the development of SAVE has lagged behind, because of the difficulty in finding the asymptotic distribution of its test statistic for dimension. It turns out that the general approach developed in Section 3 permits a relatively straightforward solution, which we present in Section 7. To avoid interrupting the discussion, we defer many proofs to the Appendix.

There are several reasons why the results in this article might advance SDR. First, if it is found that SIR always performs essentially the same as the optimal estimator, then we will have important support for existing methodology. On the other hand, if it is concluded that SIR is inferior in plausible settings, then we will have shown that the IRE provides better SDR methodology. Either conclusion would represent an important advance. Second, regardless of the relative performance of SIR, the IR family allows for the straightforward development of various conditional independence tests (Sec. 3.4) that are relatively difficult and elusive in the spectral approach. These tests bring the capabilities of SDR much closer to those of model-based regression methodology. Third, the IR family greatly facilitates SDR advances in other directions as well. Our development of SAVE in Section 7 represents one instance of this; we discuss other possibilities in Section 8.

2. THE INVERSE REGRESSION FAMILY OF ESTIMATORS

2.1 Inverse Regression

IR relies on an assumption about the marginal distribution of $\mathbf{X}$. The *linearity condition* requires that $\mathrm{E}(\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}) = \mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}$. This condition connects the central subspace with the inverse regression of $\mathbf{Z}$ on Y. When it holds, $\mathrm{E}(\mathbf{Z}|Y) \in \mathcal{S}_{Y|\mathbf{Z}}$ and thus $\mathrm{Span}(\mathbf{M}_{\mathrm{SIR}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$, as stated previously. When Y is discrete, it is easy to construct a sample version of $\mathrm{E}(\mathbf{X}|Y = y)$ by averaging $\mathbf{X}$ for fixed $Y = y$. When Y is continuous, we follow Li (1991) and construct a discrete version $\tilde{Y}$ of Y by partitioning its range into h slices. It is always true that $\mathcal{S}_{\tilde{Y}|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ and, when h is sufficiently large, $\mathcal{S}_{\tilde{Y}|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}}$. Thus, without loss of generality, we assume that Y is discrete and has a finite support $\{1, 2, \dots, h\}$. A value y of Y is called a *slice*.

Define the working meta-parameter

$$\mathcal{S}_{\boldsymbol{\xi}} = \sum_{y=1}^{h} \mathrm{Span}(\boldsymbol{\xi}_y),$$

where

$$\boldsymbol{\xi}_y = \boldsymbol{\Sigma}^{-1}\big(\mathrm{E}(\mathbf{X}|Y = y) - \mathrm{E}(\mathbf{X})\big).$$

When the linearity condition holds, $\mathcal{S}_{\boldsymbol{\xi}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. This is often taken a step further in SDR by assuming the *coverage condition*, $\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{S}_{Y|\mathbf{X}}$. Let $d = \dim(\mathcal{S}_{\boldsymbol{\xi}})$ and let $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ denote a basis of $\mathcal{S}_{\boldsymbol{\xi}}$. Under the linearity and coverage conditions, an estimate of $\boldsymbol{\beta}$ provides an estimate of a CS basis, but inference about $\mathcal{S}_{\boldsymbol{\xi}}$ itself does not require either linearity or coverage. For this reason, we work in terms of $\mathcal{S}_{\boldsymbol{\xi}}$ as much as possible, using the linearity and coverage conditions only when necessary to connect the working subspace $\mathcal{S}_{\boldsymbol{\xi}}$ with the CS. Further discussion of the linearity and coverage conditions is offered in Section 3.2.

By definition, for each y, there exists a vector $\boldsymbol{\gamma}_y$ such that $\boldsymbol{\xi}_y = \boldsymbol{\beta}\boldsymbol{\gamma}_y$. Define

$$\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_h) = \boldsymbol{\beta}\boldsymbol{\gamma},$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_h)$. Let $\mathbf{f} = (f_1, f_2, \dots, f_h)^T$, where $f_y = \Pr(Y = y)$, and let $\mathbf{g} = \sqrt{\mathbf{f}}$. It is easy to see that $\boldsymbol{\xi}\mathbf{f} = \boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{f} = \mathbf{0}$, which we call the *intrinsic location constraint*.

2.2 Inverse Regression Estimators of $\mathcal{S}_{\boldsymbol{\xi}}$

In this section we introduce the IR family of methods for estimating $\mathcal{S}_{\boldsymbol{\xi}}$. To estimate $\mathcal{S}_{\boldsymbol{\xi}}$, we require both a method for estimating its dimension and a method for estimating a basis for a given dimension. We consider basis estimation first, assuming that d is known, and then later link with an inference method for d.

Starting with a simple random sample $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, n$, on $(\mathbf{X}, Y)$, we let $\mathbf{X}_{yj}$ denote the jth observation on $\mathbf{X}$ in slice y, $y = 1, \dots, h$, $j = 1, \dots, n_y$, and $\sum_y n_y = n$. Let $\bar{\mathbf{X}}_{\cdot\cdot}$ be the overall average of $\mathbf{X}$, and let $\bar{\mathbf{X}}_{y\cdot}$ be the average of the n_y points with $Y = y$. Let $\hat{f}_y = n_y/n$, $\hat{\mathbf{f}} = (\hat{f}_1, \dots, \hat{f}_h)^T$, and $\hat{\mathbf{g}} = \sqrt{\hat{\mathbf{f}}}$. Let $\hat{\boldsymbol{\Sigma}} > \mathbf{0}$ denote the usual sample covariance matrix for $\mathbf{X}$. The sample version of $\boldsymbol{\xi}_y$ is $\hat{\boldsymbol{\xi}}_y = \hat{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{X}}_{y\cdot} - \bar{\mathbf{X}}_{\cdot\cdot})$. Let

$$\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\xi}}_1, \dots, \hat{\boldsymbol{\xi}}_h) \in \mathbb{R}^{p\times h}. \tag{1}$$

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

It is natural to estimate $\mathcal{S}_{\boldsymbol{\xi}}$ with a d-dimensional subspace that is closest to the columns of $\hat{\boldsymbol{\xi}}$. There are many ways to define "closeness." Letting $\mathrm{vec}(\cdot)$ denote the operator that constructs a vector from a matrix by stacking its columns, we consider quadratic discrepancy functions of the form

$$F_d(\mathbf{B},\mathbf{C}) = \big(\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{R}_n) - \mathrm{vec}(\mathbf{BC})\big)^T \times \mathbf{V}_n\big(\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{R}_n) - \mathrm{vec}(\mathbf{BC})\big), \quad (2)$$

where $\mathbf{V}_n \in \mathbb{R}^{pl\times pl}$ is a positive-definite matrix. The columns of $\mathbf{B} \in \mathbb{R}^{p\times d}$ represent a basis for $\mathrm{Span}(\boldsymbol{\xi}\mathbf{R}_n)$, and $\mathbf{C} \in \mathbb{R}^{d\times l}$, which is used only in fitting, represents the coordinates of $\boldsymbol{\xi}\mathbf{R}_n$ relative to $\mathbf{B}$. The matrix $\mathbf{R}_n \in \mathbb{R}^{h\times l}$ decides how we organize the columns of $\hat{\boldsymbol{\xi}}$. The subspace of $\mathbb{R}^p$ spanned by a value of $\mathbf{B}$ that minimizes F_d provides an estimate of a subset of $\mathcal{S}_{\boldsymbol{\xi}}$, depending on $(\mathbf{R}_n, \mathbf{V}_n)$. One pair $(\mathbf{R}_n, \mathbf{V}_n)$ corresponds to a dimension-reduction method. We call these methods the IR family. Given $(\mathbf{R}_n, \mathbf{V}_n)$, solutions of this minimization are not unique because of the overparameterization of the setting. This nonidentifiability is not an issue, because any basis suffices to specify $\mathcal{S}_{\boldsymbol{\xi}}$. It is possible to impose constraints to make the parameterization unique, but the overparameterized setting is more intuitive and generally easier to treat analytically.

Let $\hat{F}_d$ denote the minimum value of $F_d(\mathbf{B},\mathbf{C})$. Under certain regularity conditions, $n\hat{F}_d$ has a well-defined limiting distribution that can be used to test hypotheses of the form $d = m$ versus $d > m$. If $n\hat{F}_m$ exceeds a selected quantile of the asymptotic distribution of $n\hat{F}_d$, then the hypothesis is rejected. A series of these tests can be used to estimate d; starting with $m = 0$, test the hypothesis $d = m$. If the hypothesis is rejected, then increment m by one and test again, stopping with the first nonsignificant result. The estimate of d is then the value of m in the last hypothesis tested. This type of procedure is used fairly commonly in estimating the dimension of a subspace (Li 1991; Rao 1965, p. 472).

In this section we introduced the IR family in terms of the scaled inverse first moments $\hat{\boldsymbol{\xi}}$. This family could be described more fully as the first moment IR family to indicate the use of inverse means. The same ideas apply for other $\hat{\boldsymbol{\xi}}$'s. In Section 7 we use the IR family with inverse second moments to derive the asymptotic distribution of SAVE's test statistic for dimension. Otherwise, we continue with inverse first moments, describing the optimal member of the IR family (2) in the next section.

3. INVERSE REGRESSION ESTIMATION

The choice of an optimal discrepancy function (2) depends on the choices of $\mathbf{R}_n$ and $\mathbf{V}_n$. It seems clear that there is no general advantage in allowing $\mathbf{R}_n$ to be singular, because then we may lose information. When $\mathbf{R}_n$ is nonsingular, we can write

$$\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{R}_n) - \mathrm{vec}(\mathbf{BC}) = (\mathbf{R}_n^T \otimes \mathbf{I}_p)\big(\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BCR}_n^{-1})\big).$$

Because we will eventually be minimizing $F_d(\mathbf{B},\mathbf{C})$, we can redefine $\mathbf{C}$ as $\mathbf{CR}_n^{-1}$ without loss of generality. In this way, we see that the IR family $(\mathbf{R}_n, \mathbf{V}_n)$ with $\mathbf{R}_n$ nonsingular is equivalent to the IR subfamily $(\mathbf{I}_h, \mathbf{V}_n)$. In other words, we can specify $\{\mathbf{R}_n\}$ as any convergent nonsingular sequence without limiting our search for an optimal member of the IR family.

Let $\mathbf{D}_{\mathbf{v}}$ denote a diagonal matrix with the elements of the vector $\mathbf{v}$ on the diagonal, and construct a nonstochastic matrix $\mathbf{A} \in \mathbb{R}^{h\times(h-1)}$ such that $\mathbf{A}^T\mathbf{A} = \mathbf{I}_{h-1}$ and $\mathbf{A}^T\mathbf{1}_h = \mathbf{0}$. Then $\mathbf{D}_{\hat{\mathbf{f}}}(\mathbf{A}, \mathbf{1}_h) \in \mathbb{R}^{h\times h}$ is nonsingular and can be used as our choice for $\mathbf{R}_n$. However, because of the intrinsic location constraint, $\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{1}_h = \mathbf{0}$ and, consequently, $\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}(\mathbf{A}, \mathbf{1}_h) = (\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{A}, \mathbf{0})$. Because the last column is always $\mathbf{0}$, we will lose no generality by using the reduced data matrix $\hat{\boldsymbol{\zeta}} \equiv \hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{A}$ in the construction of discrepancy functions,

$$F_d(\mathbf{B},\mathbf{C}) = \big(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC})\big)^T\mathbf{V}_n(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC})), \quad (3)$$

where $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{C} \in \mathbb{R}^{d\times(h-1)}$, and $\mathbf{V}_n > \mathbf{0}$ has yet to be specified. The optimal choice of $\mathbf{V}_n$ in this version of the discrepancy function depends on the asymptotic distribution of $\mathrm{vec}(\hat{\boldsymbol{\zeta}})$. Here $\hat{\boldsymbol{\zeta}}$ converges in probability to

$$\boldsymbol{\zeta} \equiv \boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}}\mathbf{A} = \boldsymbol{\beta}\boldsymbol{\nu}, \quad (4)$$

where $\boldsymbol{\nu} = \boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}}\mathbf{A}$. As stated formally in the next section, $\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu}))$ converges to a normal variable with mean $\mathbf{0}$ and nonsingular covariance matrix $\boldsymbol{\Gamma}_{\hat{\boldsymbol{\zeta}}} \in \mathbb{R}^{p(h-1)\times p(h-1)}$. Additionally, the optimal version F_d^{ire} of F_d is obtained by setting $\mathbf{V}_n$ equal to a consistent estimate $\widehat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{\zeta}}}^{-1}$ of $\boldsymbol{\Gamma}_{\hat{\boldsymbol{\zeta}}}^{-1}$,

$$F_d^{\mathrm{ire}}(\mathbf{B},\mathbf{C}) = \big(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC})\big)^T \times \widehat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{\zeta}}}^{-1}\big(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC})\big). \quad (5)$$

This discrepancy function has two desirable properties, as we demonstrate in the next section. The distribution of $n\hat{F}_d^{\mathrm{ire}}$ has an asymptotic chi-squared distribution with $(p-d)(h-d-1)$ degrees of freedom, and its estimate of $\mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu})$ is asymptotically efficient. A function of $(\boldsymbol{\beta}, \boldsymbol{\nu})$ is uniquely defined only when it is a function of $\mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu})$, so only functions of $\mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu})$ are estimable. The asymptotic efficiency that we consider here means that the estimate of any function of $\mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu})$ obtained from F_d^{ire} has the smallest asymptotic variance among estimates from all possible $\mathbf{V}_n$. The estimate of $\mathcal{S}_{\boldsymbol{\xi}}$ constructed by minimizing (5) is called the IRE.

3.1 Asymptotic Normality and Optimality

We need to find the asymptotic distribution of $\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu}))$ to establish the optimality of F_d^{ire}. Because $\mathbf{A}$ is a constant matrix, we need consider only $\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}}))$. Some preparation is needed before we can report the results. Define h random variables J_y such that J_y equals 1 if an observation is in the yth slice and 0 otherwise, $y = 1, 2, \ldots, h$. Then $\mathrm{E}(J_y) = f_y$. Also, define the random vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_h)^T$, where its elements,

$$\varepsilon_y = J_y - \mathrm{E}(J_y) - \mathbf{Z}^T\mathrm{E}(\mathbf{Z}J_y), \qquad y = 1, 2, \ldots, h, \quad (6)$$

are the population residuals from the ordinary least squares fit of J_y on $\mathbf{Z}$. The required asymptotic distributions are reported in the following theorem, the proof of which is given in Appendix B.

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

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

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

where $\boldsymbol{\Gamma} = \mathrm{cov}(\mathrm{vec}(\boldsymbol{\Sigma}^{-1/2}\mathbf{Z}\boldsymbol{\varepsilon}^T)) \in \mathbb{R}^{ph\times ph}$. Consequently,

$$\sqrt{n}\big(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu})\big) \xrightarrow{\mathcal{D}} \mathrm{Normal}(\mathbf{0}, \boldsymbol{\Gamma}_{\hat{\zeta}}),$$

where $\boldsymbol{\Gamma}_{\hat{\zeta}} = (\mathbf{A}^T \otimes \mathbf{I})\boldsymbol{\Gamma}(\mathbf{A} \otimes \mathbf{I})$ is nonsingular.

Asymptotic properties are given in the following theorem, where we use the $p(h-1) \times (p+h-1)d$ matrix

$$\boldsymbol{\Delta}_{\zeta} = (\boldsymbol{\nu}^T \otimes \mathbf{I}_p, \mathbf{I}_{h-1} \otimes \boldsymbol{\beta}), \tag{7}$$

which is the Jacobian matrix

$$\boldsymbol{\Delta} = \left(\frac{\partial\,\mathrm{vec}(\mathbf{BC})}{\partial\,\mathrm{vec}(\mathbf{B})}, \frac{\partial\,\mathrm{vec}(\mathbf{BC})}{\partial\,\mathrm{vec}(\mathbf{C})}\right) \tag{8}$$

evaluated at $(\mathbf{B} = \boldsymbol{\beta}, \mathbf{C} = \boldsymbol{\nu})$.

Theorem 2. Assume that the data $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, are a simple random sample of $(\mathbf{X}, Y)$ with finite fourth moments. Let $\mathcal{S}_{\xi} = \sum_{y=1}^{h} \mathrm{Span}\{\boldsymbol{\xi}_y\}$, let $d = \dim(\mathcal{S}_{\xi})$ and let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\nu}}) = \arg_{\mathbf{B},\mathbf{C}} \min F_d^{\mathrm{ire}}(\mathbf{B}, \mathbf{C})$. Then the following results hold:

1. $\mathrm{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}})$ is asymptotically efficient, and $\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu}))$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Delta}_{\zeta}(\boldsymbol{\Delta}_{\zeta}^T \boldsymbol{\Gamma}_{\hat{\zeta}}^{-1} \boldsymbol{\Delta}_{\zeta})^{-} \boldsymbol{\Delta}_{\zeta}^T$.
2. $n\hat{F}_d^{\mathrm{ire}}$ has an asymptotic chi-squared distribution with degrees of freedom $(p-d)(h-d-1)$.
3. $\mathrm{Span}(\hat{\boldsymbol{\beta}})$ is a consistent estimator of $\mathcal{S}_{\xi}$.

The minimization of F_d in (3) always provides a consistent estimate of $\mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu})$ for any sequence of $\mathbf{V}_n > \mathbf{0}$ that converges to $\mathbf{V} > \mathbf{0}$. But the particular choice of $\mathbf{V}_n = \widehat{\boldsymbol{\Gamma}}_{\hat{\zeta}}^{-1}$ makes the estimate have the smallest asymptotic covariance. The proof of Theorem 2 is given in Appendix C. It is easy to verify that incorporating a right nonsingular transformation of $\hat{\boldsymbol{\zeta}}$ provides the same test statistic and the same asymptotic efficiency.

Theorems 1 and 2 are related when $d = p$. In that case $\boldsymbol{\beta}$ is any basis for $\mathbb{R}^p$, and it ceases to become a parameter to be estimated. Without loss of generality, we can set $\boldsymbol{\beta} = \mathbf{I}_p$. Then $\hat{\boldsymbol{\nu}} = \hat{\boldsymbol{\zeta}}$, $\boldsymbol{\Delta}_{\zeta} = (\boldsymbol{\nu}^T \otimes \mathbf{I}_p, \mathbf{I}_{h-1} \otimes \mathbf{I}_p)$, and the asymptotic covariance matrix in the first conclusion of Theorem 2 reduces to $\boldsymbol{\Gamma}_{\hat{\zeta}}$ of Theorem 1.

3.2 Linearity and Coverage

Theorem 2 is quite general, requiring neither linearity nor coverage. However, without some additional considerations, $\mathcal{S}_{\xi}$ might not be a useful parameter. If the linearity condition holds, then $\mathcal{S}_{\xi} \subseteq \mathcal{S}_{Y|\mathbf{X}}$, and we are able to use Theorem 2 to infer about a possibly proper subset of the CS. When $\mathcal{S}_{\xi}$ is a proper subset of $\mathcal{S}_{Y|\mathbf{X}}$, inverse methods can still yield important information on the regression, as demonstrated throughout the SDR literature. If the linearity and coverage conditions both hold, then $\mathcal{S}_{\xi} = \mathcal{S}_{Y|\mathbf{X}}$, and we can use Theorem 2 to infer about the full CS.

The linearity condition is equivalent to requiring that $\mathrm{E}(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ be a linear function of $\boldsymbol{\eta}^T\mathbf{X}$ (Cook 1998a, prop. 4.2), where the columns of the matrix $\boldsymbol{\eta}$ still form a basis for $\mathcal{S}_{Y|\mathbf{X}}$. Li's (1991) design condition is equivalent to the linearity condition, which applies to the marginal distribution of the predictors and not to the conditional distribution of $Y|\mathbf{X}$ as is common in regression modeling. Consequently, we are free to use experimental design, one-to-one predictor transformations $\boldsymbol{\tau}$, or reweighting (Cook and Nachtsheim 1994) to induce the condition when necessary without suffering complications when inferring about $Y|\mathbf{X}$. Because we are not assuming a model for $Y|\mathbf{X}$, these adaptation methods need not change the fundamental issues in the regression. For example, because $Y|(\mathbf{X} = \mathbf{x})$ has the same distribution as $Y|(\boldsymbol{\tau}(\mathbf{X}) = \boldsymbol{\tau}(\mathbf{x}))$, predictor transformations just change the way in which the conditional distribution of $Y|\mathbf{X}$ is indexed. The linearity condition holds for elliptically contoured predictors. Additionally, Hall and Li (1993) showed that as p increases with $\dim(\mathcal{S}_{Y|\mathbf{X}})$ fixed, the linearity condition holds to a reasonable approximation in many problems.

To gain intuition about the roles of the linearity and coverage conditions beyond that available in the literature, define the predictor subspace

$$\mathcal{S}_{\eta} = \mathrm{Span}\big\{\boldsymbol{\Sigma}^{-1}\big(\mathrm{E}(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X}) - \mathrm{E}(\mathbf{X})\big)\big\}$$

as $\boldsymbol{\eta}^T\mathbf{X}$ varies in its marginal sample space. Because every vector in $\mathcal{S}_{\xi}$ can be written as an average of vectors in $\mathcal{S}_{\eta}$, it follows that $\mathcal{S}_{\xi} \subseteq \mathcal{S}_{\eta}$. Next, every column of $\boldsymbol{\eta}$ can be written as an average of vectors in $\mathcal{S}_{\eta}$ and, consequently, $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{\eta}$. In the absence of the linearity and coverage conditions then, we always know that $\mathcal{S}_{\xi}$ and $\mathcal{S}_{Y|\mathbf{X}}$ are both subsets of $\mathcal{S}_{\eta}$. The linearity condition alone forces $\mathcal{S}_{Y|\mathbf{X}}$ to equal the upper bound, and thus $\mathcal{S}_{\xi} \subseteq \mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{\eta}$. Adding the coverage condition results in equality of the three subspaces $\mathcal{S}_{\xi} = \mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{\eta}$.

Another route to ordering the three subspaces is to first assume the *generalized coverage condition*,

$$\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{\xi} \subseteq \mathcal{S}_{\eta}. \tag{9}$$

Inference about $\mathcal{S}_{\xi}$ then provides an upper bound on the CS. The generalized coverage condition may often hold when $\mathcal{S}_{\xi} \neq \mathcal{S}_{Y|\mathbf{X}}$. For example, suppose that $p = 2$, $X_2 = X_1^2 + \delta$, and $Y = X_1 + \varepsilon$, where $(X_1, \delta, \varepsilon)^T \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_3)$. Then $\mathcal{S}_{Y|\mathbf{X}} = \mathrm{Span}((1,0)^T) \subset \mathcal{S}_{\xi} = \mathcal{S}_{\eta} = \mathbb{R}^2$. If we impose the linearity condition on top of generalized coverage, then the three subspaces under consideration are again equal. In short, generalized coverage is sufficient to guarantee that $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{\xi}$, so that we infer about an upper bound on the CS without the linearity condition.

Assuming generalized coverage, $\dim(\mathcal{S}_{\xi}) = 0$ if and only if $\dim(\mathcal{S}_{Y|\mathbf{X}}) = 0$, and $\dim(\mathcal{S}_{\xi}) = 1$ implies that $\dim(\mathcal{S}_{Y|\mathbf{X}}) = 1$. Consequently, inferring that $\dim(\mathcal{S}_{\xi}) = 0$ or 1 provides the same inference on $\mathcal{S}_{Y|\mathbf{X}}$. If $\dim(\mathcal{S}_{\xi}) = k > 1$, then $\dim(\mathcal{S}_{Y|\mathbf{X}}) \leq k$, and there is a possibility of overestimation. Such a possibility can often be assessed graphically when $k = 2$ or 3, using the visualization methods discussed by Cook (1998a).

Finally, the linearity condition has desirable consequences in contexts other than the present one, including linear regression (Li 1997; Cook 1998a, p. 226). Comments on interpretation without reference to linearity and coverage are given in Section 5.1.

3.3 Computation

To make IR estimation practical, we need address two issues. The first issue is how to construct a consistent estimate $\widehat{\boldsymbol{\Gamma}}_{\hat{\zeta}}^{-1}$ of $\boldsymbol{\Gamma}_{\hat{\zeta}}^{-1}$. We know that $\boldsymbol{\Gamma}_{\hat{\zeta}} = (\mathbf{A}^T \otimes \mathbf{I})\boldsymbol{\Gamma}(\mathbf{A} \otimes \mathbf{I})$, where

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

$\mathbf{A}$ is nonstochastic. Thus we need only plug in a consistent sample version of $\boldsymbol{\Gamma} = \operatorname{cov}(\operatorname{vec}(\boldsymbol{\Sigma}^{-1/2}\mathbf{Z}\boldsymbol{\varepsilon}^T))$ (cf. Thm. 1). Such an estimate is easily constructed by substituting sample versions of population quantities, noting that $\mathrm{E}(\operatorname{vec}(\boldsymbol{\Sigma}^{-1/2}\mathbf{Z}\boldsymbol{\varepsilon}^T)) = 0$. For instance, we constructed a sample version of $\boldsymbol{\varepsilon}$ by substituting sample means for $\mathrm{E}(J_y)$ and $\mathrm{E}(\mathbf{Z}J_y)$ for each of its elements ε_y defined in (6), giving the sample version $\hat{\varepsilon}_{yi} = J_{yi} - \hat{f}_y - (\mathbf{X}_i - \bar{\mathbf{X}}_{\cdot\cdot})^T \hat{\boldsymbol{\xi}}_y \hat{f}_y$, $i = 1, \dots, n$.

The second issue is the minimization of the discrepancy function given $\mathbf{V}_n$. Any discrepancy function of the form given in (3), which covers (5) with $\mathbf{V}_n = \widehat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{\zeta}}}^{-1}$, can be minimized by treating it as a separable nonlinear least squares problem (Ruhe and Wedin 1980). We have separate sets of parameters, $\mathbf{B}$ and $\mathbf{C}$. The value of $\operatorname{vec}(\mathbf{C})$ 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 $\mathbf{V}_n^{1/2}\operatorname{vec}(\hat{\boldsymbol{\zeta}})$ on $\mathbf{V}_n^{1/2}(\mathbf{I}_{h-1} \otimes \mathbf{B})$. On the other hand, fixing $\mathbf{C}$, consider minimization with respect to one column $\mathbf{b}_k$ of $\mathbf{B}$, given the remaining columns of $\mathbf{B}$ and subject to the length constraint $\|\mathbf{b}_k\| = 1$ and the orthogonality constraint $\mathbf{b}_k^T\mathbf{B}_{(-k)} = 0$, where $\mathbf{B}_{(-k)}$ is the matrix that remains after taking away $\mathbf{b}_k$ from $\mathbf{B}$. For this partial minimization problem, the discrepancy function can be reexpressed as

$$F^*(\mathbf{b}) = \bigl(\boldsymbol{\alpha}_k - (\mathbf{c}_k^T \otimes \mathbf{I}_p)\mathbf{Q}_{\mathbf{B}_{(-k)}}\mathbf{b}\bigr)^T \mathbf{V}_n \bigl(\boldsymbol{\alpha}_k - (\mathbf{c}_k^T \otimes \mathbf{I}_p)\mathbf{Q}_{\mathbf{B}_{(-k)}}\mathbf{b}\bigr),$$

where $\boldsymbol{\alpha}_k = \operatorname{vec}(\hat{\boldsymbol{\zeta}} - \mathbf{B}_{(-k)}\mathbf{C}_{(-k)}) \in \mathbb{R}^{p(h-1)}$, $\mathbf{c}_k$ is the kth row of $\mathbf{C}$, $\mathbf{C}_{(-k)}$ consists of all but the kth row of $\mathbf{C}$, and $\mathbf{Q}_{\mathbf{B}_{(-k)}}$ projects onto the orthogonal complement of $\operatorname{Span}(\mathbf{B}_{(-k)})$ in the usual inner product. This is a linear regression problem again.

We are now in a position to describe an algorithm for the minimization of (3), which we call the *alternating least squares method*. This method uses the special features of the objective function, and thus it is probably more efficient than a general optimization algorithm. (See Kiers 2002 for background on alternating least squares optimization algorithms.)

Outline of the Algorithm:

1. Choose an initial $\mathbf{B} \leftarrow (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_d)$. Constant initial starting vectors are often a good choice. Our experience is that the initial values do not generally affect the ultimate result.
2. Holding $\mathbf{B}$ fixed, calculate the least squares coefficients

   $$\operatorname{vec}(\mathbf{C}) = [(\mathbf{I}_{h-1} \otimes \mathbf{B}^T)\mathbf{V}_n(\mathbf{I}_{h-1} \otimes \mathbf{B})]^{-1} \times (\mathbf{I}_{h-1} \otimes \mathbf{B}^T)\mathbf{V}_n \operatorname{vec}(\hat{\boldsymbol{\zeta}}).$$

   Assign $e_0 \leftarrow F_d(\mathbf{B}, \mathbf{C})$ and *iter* $\leftarrow 0$.
3. (a) For $k = 1, 2, \dots, d$:
   - At the current step, $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_{k-1}, \mathbf{b}_k, \mathbf{b}_{k+1}, \dots, \mathbf{b}_d)$. Assign

     $$\boldsymbol{\alpha}_k \leftarrow \operatorname{vec}\bigl(\hat{\boldsymbol{\zeta}} - \mathbf{B}_{(-k)}\mathbf{C}_{(-k)}\bigr),$$

     which is a residual vector with $\mathbf{b}_k$ excluded. Find a new $\mathbf{b}_k$ minimizing the function with the constraint that it is orthogonal to $\mathbf{B}_{(-k)}$ and has length 1,

     $$\hat{\mathbf{b}}_k = \mathbf{Q}_{\mathbf{B}_{(-k)}}\bigl[\mathbf{Q}_{\mathbf{B}_{(-k)}}(\mathbf{c}_k^T \otimes \mathbf{I}_p)\mathbf{V}_n(\mathbf{c}_k \otimes \mathbf{I}_p)\mathbf{Q}_{\mathbf{B}_{(-k)}}\bigr]^- \times \mathbf{Q}_{\mathbf{B}_{(-k)}}(\mathbf{c}_k^T \otimes \mathbf{I}_p)\mathbf{V}_n\boldsymbol{\alpha}_k,$$

     $$\hat{\mathbf{b}}_k \leftarrow \hat{\mathbf{b}}_k / \|\hat{\mathbf{b}}_k\|.$$

   - Update

     $$\mathbf{B} \leftarrow (\mathbf{b}_1, \dots, \mathbf{b}_{k-1}, \hat{\mathbf{b}}_k, \mathbf{b}_{k+1}, \dots, \mathbf{b}_d)$$

     and

     $$\mathbf{C} \leftarrow \arg_{\mathbf{C}^*} \min F_d(\mathbf{B}, \mathbf{C}^*).$$

   (b) $e_1 \leftarrow F_d(\mathbf{B}, \mathbf{C})$ and *iter* $\leftarrow$ *iter* $+ 1$.
4. Return to step 3 until e_1 no longer decreases, then assign $\widetilde{\mathbf{B}} \leftarrow \mathbf{B}$ and exit.

At termination, $\operatorname{Span}(\widetilde{\mathbf{B}})$ is an estimate of $\mathcal{S}_{\boldsymbol{\xi}}$. After one iteration of step 3, the algorithm produces a monotonically decreasing series of evaluations and thus is guaranteed to converge because $F_d \geq 0$.

For SIR and other SDR methods, estimated basis directions are ordered by the eigenvalues of a sample kernel matrix. This algorithm will not necessarily produce an analogous ordering. However, we can construct an ordered basis for $\operatorname{Span}\{\widetilde{\mathbf{B}}\}$ with respect to the amount by which directions decrease $F_d(\mathbf{B}, \mathbf{C})$. For example, the most important direction is

$$\hat{\mathbf{b}}_1 = \arg_{\mathbf{b}} \min\bigl(\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}(\mathbf{b}\mathbf{C})\bigr)^T \mathbf{V}_n \bigl(\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}(\mathbf{b}\mathbf{C})\bigr),$$

where the minimization is over $\mathbf{C} \in \mathbb{R}^{1\times(h-1)}$ and $\mathbf{b} \in \operatorname{Span}\{\widetilde{\mathbf{B}}\}$ with $\|\mathbf{b}\| = 1$. The second direction is

$$\hat{\mathbf{b}}_2 = \arg_{\mathbf{b}} \min\bigl(\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}([\hat{\mathbf{b}}_1, \mathbf{b}]\mathbf{C})\bigr)^T \times \mathbf{V}_n\bigl(\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}([\hat{\mathbf{b}}_1, \mathbf{b}]\mathbf{C})\bigr),$$

where the minimization is over $\mathbf{C} \in \mathbb{R}^{2\times(h-1)}$, $\mathbf{b} \in \operatorname{Span}\{\widetilde{\mathbf{B}}\}$ with $\|\mathbf{b}\| = 1$ and $\mathbf{b}^T\hat{\boldsymbol{\Sigma}}\hat{\mathbf{b}}_1 = 0$, and so on.

3.4 Testing

Tests for predictor effects are often important in model-based regression but have received little attention in SDR. Cook (2004) recently introduced a general formulation for testing predictors in SDR and developed a specific implementation based on SIR. In this section we use Cook's formulation to develop predictor tests in the minimum discrepancy approach to SDR.

In the context of SDR, we seek tests of the conditional independence hypothesis

$$Y \perp\!\!\!\perp \mathbf{P}_{\mathcal{H}}\mathbf{X} | \mathbf{Q}_{\mathcal{H}}\mathbf{X}, \tag{10}$$

where $\mathcal{H}$ is an r-dimensional user-selected subspace of the predictor space. The hypothesis is certainly false if $r > p - \dim(\mathcal{S}_{Y|\mathbf{X}})$, so for nontrivial applications, we should have $r \leq p - \dim(\mathcal{S}_{Y|\mathbf{X}})$. Partitioning $\mathbf{X}^T = (\mathbf{X}_r^T, \mathbf{X}_{-r}^T)$, we imagine a typical application to test the hypothesis that r selected predictors $\mathbf{X}_r$ do not contribute to the regression, in which case $\mathcal{H} = \operatorname{Span}(\mathbf{H})$ with basis $\mathbf{H} = (\mathbf{I}_r, \mathbf{0})^T$. The conditional independence hypothesis (10) is equivalent to the hypothesis $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = \mathcal{O}_p$ (Cook 2004), where $\mathcal{O}_p$ indicates the origin in $\mathbb{R}^p$. Assuming that $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{\boldsymbol{\xi}}$, which is implied by the linearity and coverage conditions, we then have the following chain of equivalent hypotheses:

$$Y \perp\!\!\!\perp \mathbf{P}_{\mathcal{H}}\mathbf{X} | \mathbf{Q}_{\mathcal{H}}\mathbf{X} \quad \Leftrightarrow \quad \mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = \mathcal{O}_p \quad \Leftrightarrow \quad \mathbf{P}_{\mathcal{H}}\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{O}_p.$$

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

Consequently, we can test (10) by testing the corresponding hypothesis about the working subspace $\mathcal{S}_{\boldsymbol{\xi}}$. Assuming generalized coverage (9) but not linearity, we have

$$\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathbf{P}_{\mathcal{H}}\mathcal{S}_{\boldsymbol{\xi}}, \tag{11}$$

and $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{O}_p$ implies that $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = \mathcal{O}_p$. Consequently, lacking information to reject $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{O}_p$ supports the hypothesis (10), but rejecting does not necessarily imply dependence, because the containment in (11) may be proper.

With the introduction of predictor hypotheses, we now have five hypothesis forms that might be considered, depending on application-specific requirements:

- Marginal dimension hypotheses: $d = m$ versus $d > m$.
- Marginal predictor hypotheses: $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = \mathcal{O}_p$ versus $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} \neq \mathcal{O}_p$.
- Joint hypotheses: $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = \mathcal{O}_p$ and $d = m$ versus $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} \neq \mathcal{O}_p$ or $d > m$.
- Conditional predictor hypotheses: Given d, $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = \mathcal{O}_p$ versus $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} \neq \mathcal{O}_p$.
- Conditional dimension hypotheses: Given $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = \mathcal{O}_p$, $d = m$ versus $d > m$.

Marginal dimension hypotheses are considered extensively in the SDR literature. Here they can be tested using the distribution of $n\hat{F}_d^{\text{ire}}$ given in Theorem 2. Tests for the next three hypotheses are developed in the following three sections as applications of Theorems 1 and 2. Tests for conditional dimension hypotheses are easily preformed by deleting the linear combinations of the predictors $\mathbf{P}_{\mathcal{H}}\mathbf{X}$ under the hypotheses and then using the marginal dimension tests on the reduced data. Cook (2004) developed marginal and conditional predictor tests in the context of SIR and illustrated how they can be used in practice, but did not consider joint tests.

3.4.1 Marginal Predictor Hypotheses. The marginal predictor hypothesis $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{O}_p$, which does not require specification of d, is equivalent to the hypothesis $\mathbf{H}^T\boldsymbol{\zeta} = \mathbf{0}$, where $\mathbf{H}$ is a $p \times r$ basis for $\mathcal{H}$ and $\boldsymbol{\zeta}$ is the population limit of $\hat{\boldsymbol{\zeta}}$ as defined previously in (4). Theorem 1 provides a method for testing $\mathbf{H}^T\boldsymbol{\zeta} = \mathbf{0}$ by using the Wald test statistic,

$$\begin{aligned} T(\mathcal{H}) &= n\,\text{vec}(\mathbf{H}^T\hat{\boldsymbol{\zeta}})^T \\ &\quad \times \{(\mathbf{I}_{h-1} \otimes \mathbf{H}^T)\widehat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{\zeta}}}(\mathbf{I}_{h-1} \otimes \mathbf{H})\}^{-1}\,\text{vec}(\mathbf{H}^T\hat{\boldsymbol{\zeta}}). \end{aligned} \tag{12}$$

It follows immediately from Theorem 1 and Slutsky's theorem that under the null hypothesis, $T(\mathcal{H})$ is distributed asymptotically as a chi-squared random variable with $r(h-1)$ degrees of freedom. Perhaps, as expected, $T(\mathcal{H})$ is invariant with respect to the choice of basis for $\mathcal{H}$. In addition,

$$T(\mathcal{H}) = \min_{\mathbf{C}} nF_p^{\text{ire}}(\mathbf{I}_p, \mathbf{C}),$$

subject to the constraint $\mathbf{H}^T\mathbf{C} = \mathbf{0}$. A marginal predictor hypothesis, $\mathbf{H}^T\boldsymbol{\zeta} = \mathbf{0}$, is therefore tested without specifying d by setting $\mathbf{B} = \mathbf{I}_p$, so that the hypothesis becomes $\mathbf{H}^T\boldsymbol{\nu} = \mathbf{0}$, and then fitting $F_p(\mathbf{I}_p, \mathbf{C})$ subject to the constraint $\mathbf{H}^T\mathbf{C} = \mathbf{0}$. The proposed statistic $T(\mathcal{H})$ is then n times the minimum value of the constrained objective function.

Under linearity and coverage, $T(\mathcal{H})$ can be used to test $Y \perp\!\!\!\perp X_k|\mathbf{X}_{-k}$, where $\mathbf{X}_{-k}$ indicates the predictors left after taking away X_k. Letting $\mathbf{e}_k$ be the $p \times 1$ vector with a 1 in the kth position and 0's elsewhere, we have in this case $\mathbf{H} = \mathbf{e}_k$ and

$$\begin{aligned} T_k &\equiv T(\text{Span}(\mathbf{e}_k)) \\ &= n\mathbf{e}_k^T\hat{\boldsymbol{\zeta}}\{(\mathbf{I}_{h-1} \otimes \mathbf{e}_k^T)\widehat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{\zeta}}}(\mathbf{I}_{h-1} \otimes \mathbf{e}_k)\}^{-1}\hat{\boldsymbol{\zeta}}^T\mathbf{e}_k. \end{aligned} \tag{13}$$

We have found it informative to give the statistics T_k, $k = 1, \ldots, p$, as default computer output, because they can be used much like the t statistics for the coefficients in linear regression.

3.4.2 Joint Dimension Predictor Hypotheses. The predictor part $\mathbf{H}^T\boldsymbol{\zeta} = \mathbf{0}$ of a joint hypothesis is equivalent to the statement $\boldsymbol{\zeta} = \mathbf{Q}_{\mathcal{H}}\boldsymbol{\zeta}$. Adding the dimension part $d = m$, we have $\boldsymbol{\zeta} = \mathbf{Q}_{\mathcal{H}}\boldsymbol{\beta}\boldsymbol{\nu} = \mathbf{H}_0\boldsymbol{\beta}_{\mathbf{H}_0}\boldsymbol{\nu}$, where $\boldsymbol{\beta} \in \mathbb{R}^{p\times m}$, $\boldsymbol{\nu} \in \mathbb{R}^{m\times(h-1)}$, and $\boldsymbol{\beta}_{\mathbf{H}_0} \in \mathbb{R}^{(p-r)\times m}$ holds the coordinates of $\boldsymbol{\beta}$ represented in terms of the basis $\mathbf{H}_0 \in \mathbb{R}^{p\times(p-r)}$ for $\text{Span}(\mathbf{Q}_{\mathcal{H}})$. Without loss of generality, we can take $\mathbf{H}_0$ to be an orthonormal basis, so that $\mathbf{H}_0^T\mathbf{H}_0 = \mathbf{I}_{p-r}$. For example, consider the hypothesis $Y \perp\!\!\!\perp \mathbf{X}_r|\mathbf{X}_{-r}$, where $\mathbf{X}_r$ contains r predictors and $\mathbf{X}_{-r}$ contains the remaining predictors. Then we can construct $\mathbf{H}$ to extract the rows of $\boldsymbol{\beta}$ corresponding to $\mathbf{X}_r$ and construct $\mathbf{H}_0$ to extract the remaining rows. Because the columns of both $\mathbf{H}$ and $\mathbf{H}_0$ are subsets of the columns of $\mathbf{I}_p$, we have $\mathbf{H}^T\mathbf{H} = \mathbf{I}_r$ and $\mathbf{H}_0^T\mathbf{H}_0 = \mathbf{I}_{p-r}$.

We can fit under a joint hypothesis by minimizing the constrained optimal discrepancy function

$$\begin{aligned} F_{m,\mathbf{H}}^{\text{ire}}(\mathbf{B}, \mathbf{C}) &= \big(\text{vec}(\hat{\boldsymbol{\zeta}}) - \text{vec}(\mathbf{H}_0\mathbf{B}\mathbf{C})\big)^T \\ &\qquad \times \widehat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{\zeta}}}^{-1}\big(\text{vec}(\hat{\boldsymbol{\zeta}}) - \text{vec}(\mathbf{H}_0\mathbf{B}\mathbf{C})\big) \end{aligned}$$

over $\mathbf{B} \in \mathbb{R}^{(p-r)\times m}$ and $\mathbf{C} \in \mathbb{R}^{m\times(h-1)}$, which can be accomplished by adapting the general algorithm given in Section 3.3. Values of $\mathbf{B}$ and $\mathbf{C}$ that minimize this discrepancy function are estimates of $\boldsymbol{\beta}_{\mathbf{H}_0}$ and $\boldsymbol{\nu}$, and under the joint hypothesis, the test statistic $n\hat{F}_{m,\mathbf{H}}^{\text{ire}}$ is distributed asymptotically as a chi-squared random variable with $(p-m)(h-m-1) + mr$ degrees of freedom. This distributional result can be demonstrated as Theorem 2, noting that the Jacobian matrix is

$$\begin{aligned} \boldsymbol{\Delta}_{\boldsymbol{\zeta},\mathbf{H}} &= (\mathbf{I}_{h-1} \otimes \mathbf{H}_0)(\boldsymbol{\nu}^T \otimes \mathbf{I}_{p-r}, \mathbf{I}_{h-1} \otimes \boldsymbol{\beta}_{\mathbf{H}_0}) \\ &\qquad \in \mathbb{R}^{p(h-1)\times m(p+h-r-1)}. \end{aligned}$$

The degrees of freedom are then found by calculating $p(h-1) - \text{rank}(\boldsymbol{\Delta}_{\boldsymbol{\zeta},\mathbf{H}})$.

3.4.3 Conditional Predictor Hypotheses. A conditional predictor hypothesis, $\mathbf{P}_{\mathcal{H}}\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{O}_p$ given d, might be useful when d is specified as a modeling device or when inference on d using marginal dimension tests results in a firm estimate. We expect that tests based on knowledge of d will have greater power than the marginal tests discussed in Section 3.4.1.

The difference in minimum discrepancies,

$$T(\mathcal{H}|d) = n\hat{F}_{d,\mathbf{H}}^{\text{ire}} - n\hat{F}_d^{\text{ire}}, \tag{14}$$

can be used to test a conditional predictor hypothesis. To find its asymptotic null distribution, we first note that Lemmas A.3 and A.4 in Appendix C can be used to justify replacing the sample inner-product matrices with population versions.

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

Then, setting $\mathbf{V} = \boldsymbol{\Gamma}_{\hat{\xi}}^{-1}$, it follows from conclusion 1 of Proposition A.1 (App. C) that $T(\mathcal{H}|d)$ is asymptotically equivalent to $\mathbf{U}^T(\mathbf{P}_{\zeta} - \mathbf{P}_{\zeta,\mathbf{H}})\mathbf{U}$, where $\mathbf{U} \in \mathbb{R}^{p(h-1)}$ is a standard normal random vector and $\mathbf{P}_{\zeta}$ and $\mathbf{P}_{\zeta,\mathbf{H}}$ are the projections with respect to the usual inner product onto $\mathrm{Span}(\boldsymbol{\Gamma}_{\hat{\xi}}^{-1/2}\boldsymbol{\Delta}_{\zeta})$ and $\mathrm{Span}(\boldsymbol{\Gamma}_{\hat{\xi}}^{-1/2}\boldsymbol{\Delta}_{\zeta,\mathbf{H}})$. Next, it can be shown that $\mathrm{Span}(\boldsymbol{\Delta}_{\zeta,\mathbf{H}}) \subseteq \mathrm{Span}(\boldsymbol{\Delta}_{\zeta})$, and thus $\mathrm{Span}(\boldsymbol{\Gamma}_{\hat{\xi}}^{-1/2}\boldsymbol{\Delta}_{\zeta,\mathbf{H}}) \subseteq \mathrm{Span}(\boldsymbol{\Gamma}_{\hat{\xi}}^{-1/2}\boldsymbol{\Delta}_{\zeta})$. Consequently, $(\mathbf{P}_{\zeta} - \mathbf{P}_{\zeta,\mathbf{H}})$ is a projection with rank

$$\begin{aligned}
&\mathrm{rank}(\boldsymbol{\Delta}_{\zeta}) - \mathrm{rank}(\boldsymbol{\Delta}_{\zeta,\mathbf{H}}) \\
&\quad = d(p+h-d-1) - d(p-r+h-d-1) \\
&\quad = rd,
\end{aligned}$$

and $T(\mathcal{H}|d)$ is asymptotically distributed as a chi-squared random variable with rd degrees of freedom under the null hypothesis. Additionally, it follows from this argument that the conditional predictor test statistic $T(\mathcal{H}|d)$ and the marginal dimension statistic $n\hat{F}_d^{\mathrm{ire}}$ are asymptotically independent. In the context of a conditional predictor hypothesis, the marginal dimension test provides an asymptotically independent check on the validity of the specification of d. For ease of reference, we summarize these results in the following corollary.

Corollary 1. Under the conditional predictor hypothesis, the test statistic $T(\mathcal{H}|d)$ defined in (14) has an asymptotic chi-squared distribution with rd degrees of freedom. Additionally, $T(\mathcal{H}|d) \perp\!\!\!\perp \hat{F}_d^{\mathrm{ire}}$ asymptotically.

In addition to $T(\mathcal{H}|d)$, we used Theorem 2 to construct a Wald test statistic for conditional predictor hypotheses. Although this Wald statistic is equivalent asymptotically to $T(\mathcal{H}|d)$, we found through simulation studies that under the null hypothesis, the small-sample behavior of $T(\mathcal{H}|d)$ is clearly better than that of the Wald statistic, with the estimated level of $T(\mathcal{H}|d)$ tests typically being much closer to the nominal level. This is in line with recognized characteristics of Wald statistics in nonlinear regression (see, e.g., Donaldson and Schnabel 1987). Thus we recommend using of $T(\mathcal{H}|d)$ in practice. We find it useful as a computer default to display $T_{k|d} \equiv T(\mathrm{Span}(\mathbf{e}_k)|d)$ for $k = 1, \ldots, p$.

As mentioned previously, we expect the power of $T(\mathcal{H}|d)$ to be greater than the power of $T(\mathcal{H})$. However, $T(\mathcal{H}|d)$ can lead to misleading results if d is misspecified. Cook (2004) provided a discussion of possibilities in this regard.

4. SUBOPTIMAL INVERSE REGRESSION

The IRE takes into account two important issues: the intrinsic location constraint and the covariance of the limiting distribution of $\hat{\boldsymbol{\zeta}}$. In this section we consider a suboptimal class that does not acknowledge either issue. At least two important consequences are associated with this negligence: We might lose asymptotically efficiency, and the asymptotic distribution of the test statistic $n\hat{F}_m$ for dimension is generally not chi-squared under the hypothesis $d = m$, but rather is a more complicated linear combination of chi-squares. Nevertheless, considering the suboptimal class allows us to show that SIR and related methodology are covered by the IR family, and to build a more comprehensive foundation for future studies.

The suboptimal class that we consider is defined by $\mathbf{R}_n = \mathbf{D}_{\hat{\mathbf{f}}}$ and $\mathbf{V}_n = \mathrm{diag}\{\mathbf{V}_{ny}\}$ being a positive-definite block-diagonal matrix, with $p \times p$ blocks $\mathbf{V}_{ny}$, that converges in probability to $\mathbf{V} > \mathbf{0}$,

$$\begin{aligned}
&F_m^{\mathrm{sopt}}(\mathbf{B}, \mathbf{C}) \\
&\quad = \big(\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \mathrm{vec}(\mathbf{BC})\big)^T \mathbf{V}_n \big(\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \mathrm{vec}(\mathbf{BC})\big) \\
&\quad = \sum_{y=1}^{h} (\hat{f}_y\hat{\boldsymbol{\xi}}_y - \mathbf{BC}_y)^T \mathbf{V}_{ny} (\hat{f}_y\hat{\boldsymbol{\xi}}_y - \mathbf{BC}_y),
\end{aligned} \tag{15}$$

where $\mathbf{B} \in \mathbb{R}^{p\times m}$ and $\mathbf{C}_y$ is the yth column of $\mathbf{C} \in \mathbb{R}^{m\times h}$. Because $\mathbf{D}_{\hat{\mathbf{f}}}$ is nonsingular, the restriction $\mathbf{R}_n = \mathbf{D}_{\hat{\mathbf{f}}}$ does not by itself limit the class of discrepancy functions. However, the restrictions on $\mathbf{V}_n$ do limit the class, and this accounts for its suboptimality. We consider SIR after addressing asymptotic properties and computation for this suboptimal class.

To report the asymptotic distribution of $n\hat{F}_d^{\mathrm{sopt}}$ we need the $ph \times (p+h)d$ matrix

$$\boldsymbol{\Delta}_{\boldsymbol{\xi}} = (\mathbf{D}_{\mathbf{f}}\boldsymbol{\gamma}^T \otimes \mathbf{I}_p, \mathbf{I}_h \otimes \boldsymbol{\beta}), \tag{16}$$

which is the Jacobian matrix (8) evaluated at $(\boldsymbol{\beta}, \boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}})$. Let $\tilde{\boldsymbol{\xi}}_y = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\xi}}_y = \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}_{y\cdot} - \bar{\mathbf{X}}_{\cdot\cdot})$, and

$$\tilde{\boldsymbol{\xi}} = (\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_h). \tag{17}$$

Let

$$\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}} = (\mathbf{D}_{\mathbf{g}}\mathbf{Q}_{\mathbf{g}} \otimes \boldsymbol{\Sigma}^{-1})\,\mathrm{diag}\{\boldsymbol{\Sigma}_{\mathbf{X}|y}\}(\mathbf{Q}_{\mathbf{g}}\mathbf{D}_{\mathbf{g}} \otimes \boldsymbol{\Sigma}^{-1}), \tag{18}$$

where $\boldsymbol{\Sigma}_{\mathbf{X}|y} = \mathrm{cov}(\mathbf{X}|Y = y)$. Finally, letting $\boldsymbol{\Phi} = \mathbf{V}^{1/2}\boldsymbol{\Delta}_{\boldsymbol{\xi}}$ and $\boldsymbol{\Omega} = \mathbf{V}^{1/2}\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}\mathbf{V}^{1/2}$, the asymptotic distribution of $n\hat{F}_d^{\mathrm{sopt}}$ is given in the following theorem.

Theorem 3. Assume that the data $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, are a simple random sample of $(\mathbf{X}, Y)$ with finite fourth moments. Let $\mathcal{S}_{\boldsymbol{\xi}} = \sum_{y=1}^{h} \mathrm{Span}\{\boldsymbol{\xi}_y\}$, let $d = \dim(\mathcal{S}_{\boldsymbol{\xi}})$, and let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \arg_{\mathbf{B},\mathbf{C}} \min F_d^{\mathrm{sopt}}(\mathbf{B}, \mathbf{C})$. Then the following results hold:

1. $\mathrm{Span}(\hat{\boldsymbol{\beta}})$ is a consistent estimator of $\mathcal{S}_{\boldsymbol{\xi}}$.
2. As $n \to \infty$, $n\hat{F}_d^{\mathrm{sopt}} \overset{\mathcal{D}}{\to} \sum_{i=1}^{ph} \lambda_i\chi_i^2(1)$, where $\{\chi_i^2(1)\}$ are independent chi-squared random variables each with 1 degree of freedom and $\{\lambda_1 \ge \cdots \ge \lambda_{ph}\}$ are the eigenvalues of $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$.

Like Theorem 2, this theorem is quite general, requiring none of the regularity conditions discussed previously. It is valid for a general $\mathbf{V}_n > \mathbf{0}$, of which the block-diagonal $\mathbf{V}_n$ used in the suboptimal class is a special case. A value $\hat{\boldsymbol{\beta}}$ of $\mathbf{B}$ that minimizes F_d^{sopt} always provides a consistent estimate of a basis for $\mathcal{S}_{\boldsymbol{\xi}}$, and this theorem allows us to test hypotheses about its dimension. The proof of Theorem 3 is given in Appendix D.

To use Theorem 3 in practice, we need to replace $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$ with a consistent estimate under the null hypothesis $d = m$. The $ph \times (p+h)m$ Jacobian matrix $\boldsymbol{\Delta}_{\boldsymbol{\xi}}$ can be estimated consistently by substituting solutions to the minimization of (15), $\hat{\boldsymbol{\Delta}}_{\boldsymbol{\xi}} = (\mathbf{D}_{\hat{\mathbf{f}}}\hat{\boldsymbol{\gamma}}^T \otimes \mathbf{I}_p, \mathbf{I}_h \otimes \hat{\boldsymbol{\beta}})$. To estimate $\mathbf{V}$, we use $\mathbf{V}_n$. We also can estimate $\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}$ with $(\mathbf{D}_{\hat{\mathbf{g}}}\mathbf{Q}_{\hat{\mathbf{g}}} \otimes \hat{\boldsymbol{\Sigma}}^{-1})\,\mathrm{diag}\{\hat{\boldsymbol{\Sigma}}_{\mathbf{X}|y}\}(\mathbf{Q}_{\hat{\mathbf{g}}}\mathbf{D}_{\hat{\mathbf{g}}} \otimes \hat{\boldsymbol{\Sigma}}^{-1})$, where $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}|y}$ is the sample covariance matrix for the yth slice.

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

These estimates are then substituted to yield an estimate of $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$ 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-squared random variables (see Field 1993 for an introduction).

5. SLICED INVERSE REGRESSION IN THE INVERSE REGRESSION FAMILY

As reviewed in Section 1, SIR is based on the spectral decomposition of a sample version $\widehat{\mathbf{M}}_{\text{SIR}}$ of the kernel matrix $\mathbf{M}_{\text{SIR}} = \text{cov}(\text{E}(\mathbf{Z}|Y))$. Given the dimension d and the linearity and coverage conditions, the eigenvectors corresponding to its largest d eigenvalues are used to estimate a basis of the CS.

The implementation of SIR originally proposed by Li (1991) is as follows: Standardize $\mathbf{X}$ to $\hat{\mathbf{Z}} = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X} - \bar{\mathbf{X}}_{..})$, then construct the $p \times p$ kernel matrix $\widehat{\mathbf{M}}_{\text{SIR}} = \sum_{y=1}^{h}\hat{f}_y\bar{\mathbf{Z}}_{y\cdot}\bar{\mathbf{Z}}_{y\cdot}^T$, where $\bar{\mathbf{Z}}_{y\cdot}$ is the average of $\hat{\mathbf{Z}}$ in the yth slice. Next, construct the eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$ and corresponding eigenvectors $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \ldots, \hat{\boldsymbol{\mu}}_p$ of $\widehat{\mathbf{M}}_{\text{SIR}}$. Given $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ ($d < \min\{p, h\}$), $\hat{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\mu}}_j$, $j = 1, 2, \ldots, d$, is the estimated basis of $\mathcal{S}_{Y|\mathbf{X}}$. The test statistic $\Lambda_m = n\sum_{i=m+1}^{p}\hat{\lambda}_i$ is used to estimate d, following the procedure described in at the end of Section 2.2.

5.1 SIR's Suboptimal Discrepancy Function

SIR is the member of the suboptimal class (15) with $\mathbf{V}_n = \text{diag}\{\hat{f}_y^{-1}\hat{\boldsymbol{\Sigma}}\}$,

$$\begin{aligned} F_d^{\text{sir}}(\mathbf{B}, \mathbf{C}) &= \sum_{y=1}^{h}(\hat{f}_y\hat{\boldsymbol{\xi}}_y - \mathbf{BC}_y)^T\hat{f}_y^{-1}\hat{\boldsymbol{\Sigma}}(\hat{f}_y\hat{\boldsymbol{\xi}}_y - \mathbf{BC}_y) \\ &= \sum_{y=1}^{h}\left(\sqrt{\hat{f}_y}\bar{\mathbf{Z}}_{y\cdot} - \hat{f}_y^{-1/2}\hat{\boldsymbol{\Sigma}}^{1/2}\mathbf{BC}_y\right)^T \\ &\quad \times \left(\sqrt{\hat{f}_y}\bar{\mathbf{Z}}_{y\cdot} - \hat{f}_y^{-1/2}\hat{\boldsymbol{\Sigma}}^{1/2}\mathbf{BC}_y\right). \end{aligned} \qquad (19)$$

Based on Lemma A.1 in Appendix A, $\text{Span}(\hat{\boldsymbol{\Sigma}}^{1/2}\hat{\boldsymbol{\beta}})$ is the space spanned by the d eigenvectors corresponding to $\widehat{\mathbf{M}}_{\text{SIR}}$'s d largest eigenvalues, where $\hat{\boldsymbol{\beta}}$ is a value of $\mathbf{B}$ that minimizes (19). Thus the estimate of $\mathcal{S}_{Y|\mathbf{X}}$ is $\text{Span}(\hat{\boldsymbol{\Sigma}}^{-1/2}(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \ldots, \hat{\boldsymbol{\mu}}_d))$. Because the diagonal blocks in $\mathbf{V}_n$ differ by only a scalar, the minimization of (19) reduces to a spectral decomposition problem.

Chen and Li (1998) developed population interpretations of SIR that might be helpful in some applications without considering the relationship between $\mathcal{S}_{\boldsymbol{\xi}}$ and $\mathcal{S}_{Y|\mathbf{X}}$. It follows from Theorem 3 that SIR and IRE both estimate $\mathcal{S}_{\boldsymbol{\xi}}$. Consequently, the IR framework allows for the Chen–Li interpretations.

5.2 SIR's Chi-Squared Test for Dimension

It follows from Lemma A.1 in Appendix A that $\Lambda_m = n\hat{F}_m^{\text{sir}}$. We know from Theorem 3 that Λ_d is distributed asymptotically as a linear combination of independent chi-squared random variables. Li (1991) showed that it has an asymptotic chi-squared distribution when $\mathbf{X}$ is normally distributed. Cook (1998a) proved that a weaker *marginal covariance condition* suffices for Λ_d to have the same asymptotic chi-squared distribution. These earlier results can be deduced from Theorem 3 by using their special conditions to simplify $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$. We restate Cook's result here for completeness.

Corollary 2 (Cook 1998a, prop. 11.5). Assume the following:

1. The linearity condition: $\text{E}(\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}) = \mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}$
2. The marginal covariance condition: $\text{cov}(\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}) = \mathbf{Q}_{\mathcal{S}_{Y|\mathbf{Z}}}$
3. The coverage condition: $\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{S}_{Y|\mathbf{X}}$.

Suppose that $\dim(\mathcal{S}_{Y|\mathbf{X}}) = d$. Then $\Lambda_d \overset{\mathcal{D}}{\to} \chi^2_{(p-d)(h-d-1)}$, as $n \to \infty$.

An often important hypothesis is $d = 0$ versus $d > 0$. When Y is independent of $\mathbf{X}$, the within-slice covariance $\boldsymbol{\Sigma}_{\mathbf{X}|y} = \boldsymbol{\Sigma}$. Thus $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}} = (\boldsymbol{\Gamma}_{22}^T\mathbf{Q}_{\mathbf{g}}\boldsymbol{\Gamma}_{22} \otimes \mathbf{I}_{p-d})$ is idempotent with a trace $(p-d)(h-1)$, and the limiting distribution of $n\hat{F}_0^{\text{sir}}$ is a chi-squared distribution with $p(h-1)$ degrees of freedom without any special conditions.

If, instead of SIR's $\mathbf{V}_n = \text{diag}\{\hat{f}_y^{-1}\hat{\boldsymbol{\Sigma}}\}$, we let $\hat{\text{E}}(\boldsymbol{\Sigma}_{\mathbf{X}|Y}) = \sum_{y=1}^{h}\hat{f}_y\hat{\boldsymbol{\Sigma}}_{\mathbf{X}|y}$ and then use

$$\mathbf{V}_n^* = \text{diag}\left\{\hat{f}_y^{-1}\hat{\boldsymbol{\Sigma}}\,(\hat{\text{E}}(\boldsymbol{\Sigma}_{\mathbf{X}|Y}))^{-1}\hat{\boldsymbol{\Sigma}}\right\}$$

in the suboptimal discrepancy function (15) while preserving the other structure, then the asymptotic results are the same as those in Corollary 2. Because $\boldsymbol{\Sigma} \geq \text{E}(\boldsymbol{\Sigma}_{\mathbf{X}|Y})$, the test statistic $n\hat{F}_m^*$ based in $\mathbf{V}_n^*$ is never smaller than SIR's test statistic Λ_m, although both statistics converge to the same chi-squared distribution when $m = d$. Such situations seem to reflect the nonoptimal properties of SIR.

5.3 Weighted Chi-Squared Test for Dimension

Bura and Cook (2001b) proved that in general, Λ_d is distributed asymptotically as a linear combination of independent chi-squared random variables, and showed how to construct consistent estimates of the weights for use in practice. Their resulting *weighted chi-square test* (WCT) holds for any predictor distribution, with linearity and coverage still required to equate the working subspace and CS. Additionally, it follows immediately from Theorem 3 that Λ_d is distributed asymptotically as a linear combination of chi-squared random variables. By investigating $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$ in detail, it can be shown after a fair bit of algebra that the Bura–Cook weighted chi-squared test follows as a direct result of Theorem 3, providing another connection between existing methodology and IR.

6. SIMULATION RESULTS AND DATA ANALYSIS

SIR and the IRE both provide consistent estimates of $\mathcal{S}_{\boldsymbol{\xi}}$, and both use the same sequential paradigm for estimating $\dim(\mathcal{S}_{\boldsymbol{\xi}})$ based on their own marginal dimension tests. SIR can be used with the WCT or the more restrictive chi-squared test (CT). Table 1 summarizes these methods in terms of $(\mathbf{R}_n, \mathbf{V})$, which identifies them as members of the IR family, and the asymptotic distributions of their test statistics for dimension. In this section we report selected results to support our general conclusions from simulation studies to compare the methods summarized in Table 1.

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

Table 1. Summary of SDR Methods

	R_n	V	*Asymptotic distribution*
SIR: CT	$\mathbf{D}_{\hat{\mathbf{f}}}$	$\mathbf{D}_{\mathbf{f}}^{-1} \otimes \boldsymbol{\Sigma}$	$\chi^2_{(p-d)(h-d-1)}$
SIR: WCT	$\mathbf{D}_{\hat{\mathbf{f}}}$	$\mathbf{D}_{\mathbf{f}}^{-1} \otimes \boldsymbol{\Sigma}$	$\sum_{i=1}^{ph} \lambda_i \chi_i^2(1)$
IRE	$\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{A}$	$\boldsymbol{\Gamma}_{\hat{\xi}}^{-1}$	$\chi^2_{(p-d)(h-d-1)}$

6.1 Estimation of $\mathcal{S}_\xi$ Given d

Our first goal is to use a simple simulation model to illustrate potential differences between the IRE and the SIR estimator of $\mathcal{S}_\xi$ when d is known. These estimators are identical when only two slices are used, and, consequently, there must be at least three slices for informative results. The SIR estimator is based on the sample slice means and frequencies and on the marginal sample covariance of the predictors. The IRE depends on these statistics, on the intraslice covariance matrices, and on higher sample moments present in its inner-product matrix. Consequently, different intraslice covariances are needed to fully demonstrate the potential differences. To allow easy control of these distinguishing features, we used p conditionally normal predictors and simulated an IR with a categorical response having $h = 3$ levels.

Model A. $\mathbf{X}_{yj} = \sigma_y \mathbf{Z}_{yj} + \mu_y \mathbf{1}_p$, $j = 1, \ldots, n_y$, $y = 1, 2, 3$, where $\mathbf{1}_p$ indicates the $p \times 1$ vector of 1's, $\mathbf{Z} \in \mathbb{R}^p$ is a vector of independent standard normal variates, and $\mathcal{S}_\xi = \operatorname{Span}(\mathbf{1}_p)$. To facilitate later description, we use a reference model with $p = 5$, $n_1 = n_2 = n_3 = 200$, $\mu_1 = 1$, $\mu_2 = .7$, $\mu_3 = .3$, $\sigma_1 = 5$, and $\sigma_2 = \sigma_3 = .5$. Parameters not explicitly specified in a simulation configuration should be understood to be the same as those in this reference model.

Figure 1(a) shows the angles in degrees between $\mathcal{S}_\xi$ and its estimates from data following Model A with $\sigma_1 = 7$, with each point corresponding to one simulated dataset. The angle between two vectors $\mathbf{a}$ and $\mathbf{c}$ was computed as $180\cos^{-1}(|\mathbf{a}^T\mathbf{c}|/\|\mathbf{a}\|\|\mathbf{c}\|)/\pi$. In Model A simulations, $\mathbf{a}$ was set equal to a basis vector for $\mathcal{S}_\xi$, $\mathbf{a} = \mathbf{1}_p$, and $\mathbf{c} = \hat{\mathbf{b}}_1$, where $\hat{\mathbf{b}}_1$ is the IRE computed using the algorithm described in Section 3.3. The IRE was closer to $\mathcal{S}_\xi$ in all 200 replications of Figure 1. As we decreased σ_1, the point cloud in Figure 1(a) moved across the diagonal line. Figure 1(b) shows the results from 200 simulation runs with $\sigma_1 = 5$. In this case the IRE did better than SIR 94% of the time. Results with $\sigma_1 = .5$ are shown in Figure 2(a). The three groups now have equal frequencies and equal variation, and we found nothing to distinguish between the methods.

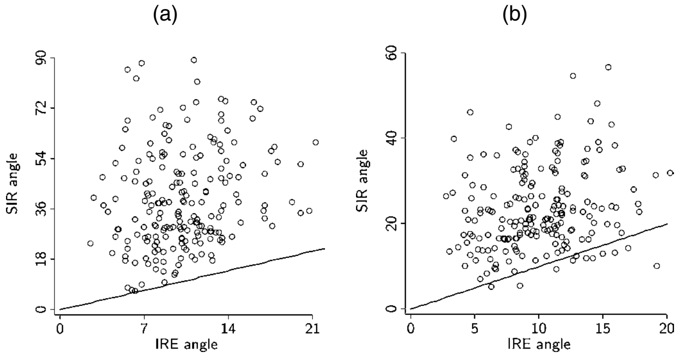

Figure 1. Angles in Degrees Between $\mathcal{S}_\xi$ and the SIR Estimate and the IRE for 200 Replications of Model A With (a) $\sigma_1 = 7$ and (b) $\sigma_1 = 5$. The lines of equal angles were added for reference, and the axis scales differ.

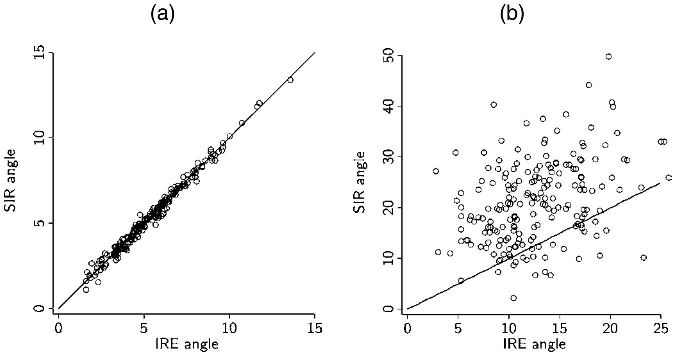

Figure 2. Angles in Degrees Between $\mathcal{S}_\xi$ and the SIR Estimate and the IRE for 200 Replications of Two Versions of Model A: (a) Equal $\sigma_y = .5$ and (b) Unequal σ_y's and Unequal n_y's.

Figure 2(b) shows the angles from 200 replications of Model A with $n_1 = 100$, $n_2 = 470$, $n_3 = 30$, $\sigma_1 = 2$, $\sigma_2 = 1$, and $\sigma_3 = .5$. In this case, unequal frequencies plus mildly different σ_y's still resulted in clear differences in the estimators, with the IRE angle being less than the SIR angle in 87% of the trials.

As anticipated, we found that the quality of both estimators deteriorated as p increased. For example, again starting with the reference version of Model A, Figure 3(a) shows plots of average angles at nine values of p, with each average based on 200 datasets (200 runs). Figure 3(b) shows plots of the average SIR and IRE angles from 200-run simulations at values of σ_1 between .5 and 9. The IRE is little affected by changes in σ_1, whereas SIR is quite sensitive to them.

Figure 4(a) shows the average SIR and IRE angles from 200-run simulations in which we varied the equal group sample size n_y. The anticipated failure for small sample sizes is evident in the figure. As the sample size grows, the average angle should converge to 0 for both methods, but this is not evident in the figure. From another simulation with $n_y = 800$, the average SIR and IRE angles were 11.6 and 4.9 degrees; thus n_y must be large for the angles to be very close.

These simulation results depend on our choice of the inverse means $\mu_y \mathbf{1}_p$, and changing them can result in differences between SIR and IRE that are greater or less than those demon-

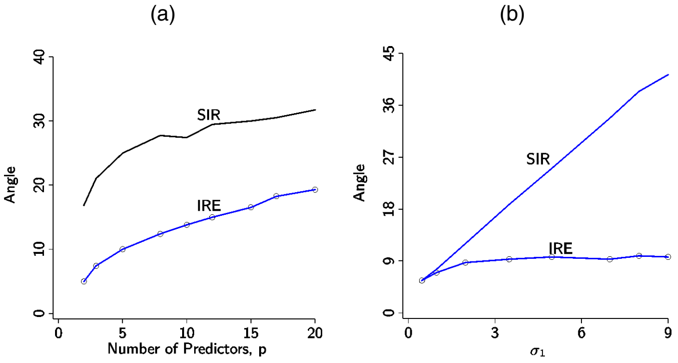

Figure 3. Simulation Results as p and σ_1 Are Varied in Model A: (a) Average Angle versus p; (b) Average Angle versus σ_1.

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

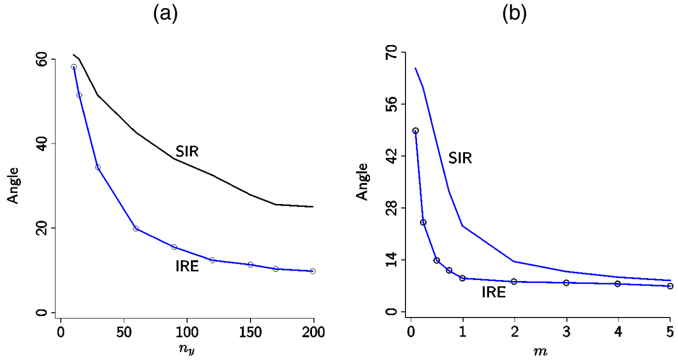

Figure 4. Simulation Results as n_y and the Mean Multiplier m Are Varied in Model A: (a) Average Angle versus n_y; (b) Average Angle versus Mean Multiplier m.

strated so far. For instance, consider replacing μ_y with $m\mu_y$, $y = 1, 2, 3$, where we refer to m as the *mean multiplier*, because it multiplies the mean vector $\mu_y \mathbf{1}_p$ of the reference model. Figure 4(b) shows average SIR and IRE angles for values of m between .1 and 5, with the averages again computed from 200 runs. At $m = .1$, the average SIR angle is close to that expected between $\mathcal{S}_\xi$ and a randomly chosen vector, about 66 degrees. The greatest angular difference occurred around $m = .25$ and the greatest ratio (SIR/IRE) occurred around $m = .5$. The value $m = 1$ in the reference model then is not the most extreme case. The IRE responded faster than SIR to increasing m, and the methods ended at roughly the same place when $m = 5$, suggesting that heterogeneity in the regression may not matter much for estimation if the signal for $\mathcal{S}_\xi$ is sufficiently strong.

Increasing the number of groups or using a continuous response with slicing while maintaining the essential heteroscedastic characteristics of Model A did not have much of an impact. For example, for 100 observations in each of six groups, with the μ_y's spaced equally between 1 and .3 and $\sigma_y = .5$, except when $\mu_y = 1$ and then $\sigma_y = 5$, we observed results similar to those in Figure 1(b). The IRs used in these simulations could be "inverted" and data generated on the forward scale, but in principle there seems little more to be learned by pursuing that approach.

Notable variation in the conditional variances $\text{cov}(\mathbf{X}|Y)$ relative to the magnitude of the signal for $\mathcal{S}_\xi$ seems to be an essential qualitative characteristic of our examples that highlight the differences between SIR and IRE. Variation in $\text{cov}(\mathbf{X}|Y)$ alone has been found to be important in past studies of characteristics of SIR. For instance, Bura and Cook (2001b) emphasized that SIR's CT for dimension is robust "provided there is no significant nonconstant variance in the $\mathbf{X}$ conditionals." It seems now that, with normal predictors, SIR's estimates are near optimal when variation in $\text{cov}(\mathbf{X}|Y)$ is small relative to the signal for $\mathcal{S}_\xi$. First results with nonnormal predictors indicate that skewness can exacerbate the problems caused by different conditional variances, leading us to suspect that the results here may give a generally optimistic view of SIR's performance.

We performed several simulations in homogeneous cases like the case illustrated in Figure 2(a), but never detected any cost in estimation of using the IRE relative to SIR. In these and all other simulations, we tried to avoid situations in which estimated covariance matrices were singular or nearly so. Dimension reduction in regressions with sparse data is an important problem, particularly in bioinformatics, but is outside the scope of this report.

We conclude this section by reporting simulation results from a forward regression model that we also use in Section 6.2.

Model B. $Y = 1.5(5 + X_1)(2 + X_2 + X_3) + .5\epsilon$, where ϵ is a standard normal random variable, $X_1 = W_1$, $X_2 = V_1 + W_2/2$, $X_3 = -V_1 + W_2/2$, $X_4 = V_2 + V_3$, and $X_5 = V_2 - V_3$. The V_i's and W_j's are mutually independent, with the V_i's drawn from a $t_{(5)}$ distribution and the W_j's drawn from a gamma(.2) distribution. We used 5 degrees of freedom for the t-distribution to guarantee the existence of fourth moments. Versions of this model were used by Li (1991), Velilla (1998), and Bura and Cook (2001b) in simulation studies related to the performance of SIR. The predictors are quite skewed and prone to outliers, and the conditional variance functions are not constant, as can be seen in the scatterplots provided by Bura and Cook (2001b). Thus we would expect SIR and the IRE to give noticeably different results for this model.

We used $h = 4$ slices in every simulation and tabulated the results over 500 runs. Because the regression is two-dimensional, we used a different method to track the results. We assessed estimation accuracy for each method by computing the absolute sample correlations $|r|$ between each of the sufficient predictors, X_1 and $X_2 + X_3$, and their fitted values from the linear regressions on the first two estimated sufficient predictors. The curves shown in Figure 5(a) are plots by a sufficient predictor of the percentage of runs in which $|r_{\text{ire}}| > |r_{\text{sir}}|$ versus sample size n, for values of n between 200 and 1,600. For instance, IRE did better than SIR about 75% of the time in estimating the direction of $X_2 + X_3$ with $n = 200$ observations. This plot reflects the asymptotic efficiency of the IRE, because its performance relative to SIR improves as n increases. When $n = 1{,}600$, the IRE did better than SIR for both directions in $\mathcal{S}_\xi$ on about 85% of the trials. This result is qualitatively similar to that shown in Figure 2(b). Figure 5(b) is discussed in Section 6.2.

6.2 Estimation of *d*

Tests of dimension hypotheses can be of interest on their own, but perhaps are most frequently used to estimate d sequentially, as described in Section 2.2. Reasoning in the context of Model B 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

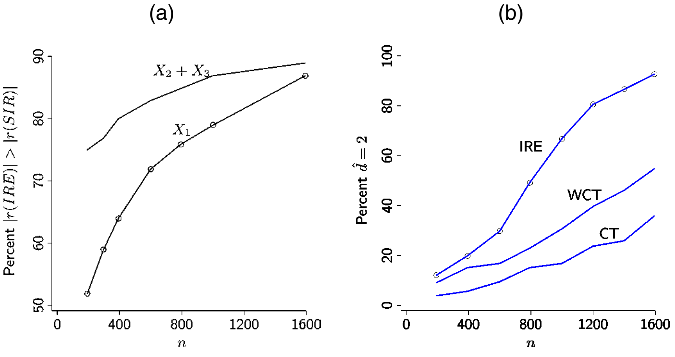

Figure 5. Comparison of SIR and IRE for Model B. (a) Percentage of runs in which $|r_{ire}| > |r_{sir}|$ versus sample size; (b) percentage of runs in which $\hat{d} = 2$ versus sample size for IRE and for SIR's CT and WCT.

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

from the level α of the test of $d = 2$, resulting in estimates $\hat{d} = 2$ with probability $1 - \alpha$ and $\hat{d} > 2$ with probability α. 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.

As mentioned previously, Bura and Cook (2001b) emphasized that variation in $\mathrm{cov}(\mathbf{X}|Y)$ can significantly affect the actual test levels of SIR's CT for dimension, and that the WCT levels are relatively robust to such variation. Our results confirm these conclusions, but also indicate that the nominal and actual levels for these tests can still be far apart. An extreme instance is shown in Figure 6(a). With a minimum p value of about .93 over the 1,000 replications, the performance the CT is much worse than that of the WCT shown in Figure 6(a). As Bura and Cook indicated, the WCT is more robust that the CT, but in this extreme case the variation among conditional variances is too large for it to work well. In contrast, the actual and nominal levels of the IRE test statistic $n\hat{F}_m^{\mathrm{ire}}$ are always quite close, as illustrated in Figure 6(b).

Comparing the power of dimension tests observed to have similar levels, we concluded that $n\hat{F}_d^{\mathrm{ire}}$ performs at least as well as SIR's CT and WCT and frequently much better. It perhaps should not be surprising that there are sizeable differences in power for Models A and B, because we have already demonstrated differences in estimation for these models. However, we also found differences in power for simulation models in which we found no notable differences in estimation or testing level.

These general observations on test level and power suggest that there can be substantial differences between the IR and SIR estimators of d. For example, Figure 5(b) shows percentages of correct dimension estimates $\hat{d} = 2$ for various sample sizes in 500 simulated datasets from Model B using nominal level $\alpha = .05$. None of the methods do very well at $n = 200$. At $n = 1{,}600$, the IR estimator is close to the best rate of correct decisions (in this case 95%), whereas SIR's CT and WCT fall substantially below this value.

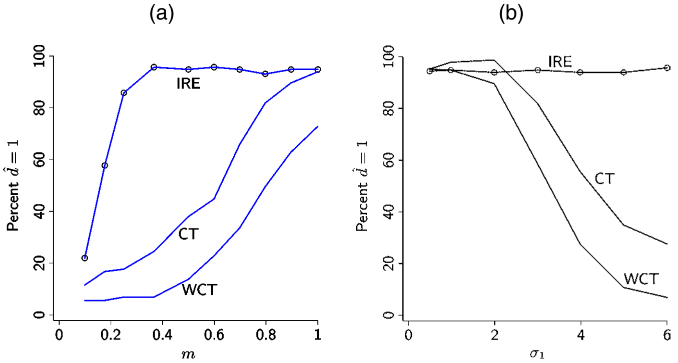

Figure 7. Percentage of Correct Estimates $\hat{d} = 1$ in Model A Using 5% Tests versus: (a) Signal Represented by the Mean Multiplier m and (b) Noise Represented by σ_1 With Mean Multiplier $m = 1/2$.

Figure 7 shows percentages of correct dimension estimates $\hat{d} = 1$ in 500 datasets simulated from Model A at various values of the mean multiplier m and σ_1, again using $\alpha = .05$. From Figure 7(a) shows that the IR estimator responds much faster to increasing signal than either of the SIR estimators, and that its frequency of correct decisions remains close to 95% for $m > .3$, an indication that the nominal and actual error rates for testing $d = 1$ are close. The CT estimator reaches about the same frequency of correct decisions with $m = 1$. However, recall from Section 6.1 that even given d, there can be substantial differences between the IR and SIR estimators of $\mathcal{S}_{\boldsymbol{\xi}}$. The relationship between the CT and the WCT estimators seems to arise because the WCT tends to be conservative in this simulation, leading to fewer rejections than the liberal CT. Finally, Figure 7(b) shows that the IR estimator of d is immune to changes in σ_1, while the CT and WCT estimates are sensitive to such changes. Because of its liberal nature in this example, the frequency of correct decisions for the CT estimate is greater than 95% for smaller values of σ_1, but this effect is quickly overtaken by its loss of power. The behavior of the IR estimator in Figure 7(b) is another indication that the IR test level is close to its nominal value.

We conclude from our simulations in this and the previous section that IR methods can easily dominate SIR methods in estimation, particularly when the variation in the intraslice covariance of the predictors is not small relative to the signal for $\mathcal{S}_{\boldsymbol{\xi}}$.

6.3 Predictor Tests

We also compared the nominal and estimated levels of the predictor tests discussed in Section 3.4. Like the IR dimension tests, the simulation results were in useful agreement with the theory, as illustrated in Figure 6(c), so here we present results that reflect other aspects of the behavior of these tests

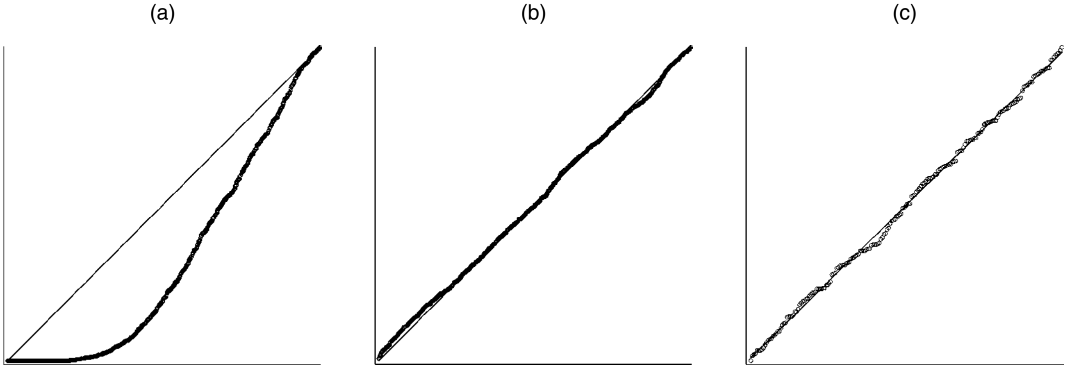

Figure 6. Uniform Quantile Plots of p Values From (a) the WCT and (b) the IR Statistic $n\hat{F}^{ire}$ for the Hypothesis d = 1 From 1,000 Replications of Model A With $\sigma_1 = 20$, $\mu_2 = -.3$, and $\mu_3 = -.7$ and (c) 500 Replications of the Marginal Predictor Test for X_5 in Model B With n = 200.

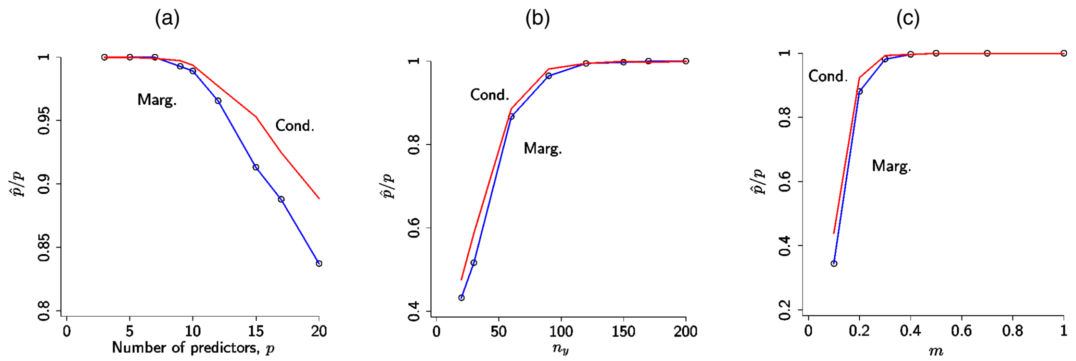

Figure 8. Performance of Marginal and Conditional Predictor Tests as p, n_y, and the Group Means $m\mu_y$ Are Varied in a Version of Model A. (a) Average $\hat{p}/p$ versus p; (b) average $\hat{p}/p$ versus n_y; (c) average $\hat{p}/p$ versus m.

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

under Model A with $\mu_2 = 0$ and $\mu_3 = -1$. Let $\hat{p}$ denote the number of "significant predictors" determined by applying a marginal or conditional test to each predictor in turn. Because $\mathcal{S}_\xi = \mathrm{Span}(\mathbf{1}_p)$, and because the predictors were generated independently, it is reasonable to hope that $\hat{p}$ is close to p. Using 5% tests, Figure 8 shows averages of $\hat{p}/p$ over 200 simulated datasets for various values of p, n_y, and $m\mu_y$. For example, we see from Figure 8(a) that with 10 predictors, the marginal predictor test declared about 9.8 significant predictors on the average, whereas the conditional tests did a bit better. We conclude from these and other simulation results that (a) as anticipated, the conditional test is generally more powerful than the marginal test, and (b) both tests respond well to increasing signal [Fig. 8(c)]. In simulations not represented in Figure 8, we found that varying σ_1 between .5 and 20 had little impact on the tests, with the average $\hat{p}/p$ remaining close to 1. The results for $p = 20$ in Figure 8(a) suggest that information on the regression becomes more important as p increases. Of course, we could increase or decrease the power of the marginal and conditional tests by increasing or decreasing the signal for $\mathcal{S}_\xi$, as suggested by Figure 8(c), or increasing or decreasing the sample size.

A potential advantage of marginal predictor tests is that they may identify the active predictors in a regression without inferring about its dimension. For example, in the 500 simulations of Model B with $n = 200$ summarized in Figure 5(b), the IR estimator produced the correct value $\hat{d} = 2$ in only about 15% of the runs. Nevertheless, the 5% marginal predictor test applied to each predictor resulted in the correct identification of all three active predictors (X_1, X_2, X_3) in 96.8% of the runs. The estimated level for the inactive predictors (X_4, X_5) was close to the nominal rate of 5%. Keeping the simulation model the same but adding five independent standard normal predictors to the calculations did not make a notable difference in these results, with the three active predictors being detected in 95.4% of the runs.

6.4 Lean Body Mass Regression

We use measurements on 202 athletes from the Australian Institute of Sport to illustrate aspects of data analysis with IRE (see Cook 2004 for a corresponding analysis with SIR and additional background). The response is lean body mass (Y), and the predictors are the logarithms of height (Ht), weight (Wt), sum of skin folds (SSF), red cell count (RCC), white cell count (WCC), plasma ferritin concentration (PFC), hematocrit (Hc), and hemoglobin (Hg).

The p values for IRE's marginal dimension test of $d = m$, $m = 0, \ldots, 4$ are 0, 0, .001, .067, and .842, hinting that the CS might be four-dimensional. IRE requires estimation of fourth predictor moments and thus can be sensitive to outliers. One case with a relatively large red cell count stands out in a scatterplot matrix of the predictors. Deleting this case results in only minor perturbations of the p values, except for the p value for $m = 3$, which changed to .16. It seems, therefore, that $d = 3$ could be a reasonable inference. Nevertheless, we continue with the full data by considering marginal predictor tests, because they do not require a dimension specification.

The second column of Table 2 gives the p values from the marginal predictor test (13) applied to each predictor in turn. For instance, assuming linearity and coverage, the PFC p value of .022 is for the test that Y is independent of PFC given the other seven predictors. We see that the first three predictors would normally be judged to contribute significantly, whereas the fourth (RCC) is marginal. As in linear regression, here the interpretation of the results for the last four predictors $\mathbf{X}_{L4}$ is problematic, because two correlated predictors might both have large p values, whereas deleting either of them causes the remaining one to decrease substantially. This is a possibility here, because the correlation between log(RCC) and log(Hc) is about .92. Letting $\mathbf{X}_{F4}$ denote the vector of the first four predictors in Table 2, the marginal predictor test (12) of the hypothesis $Y \perp\!\!\!\perp \mathbf{X}_{L4} | \mathbf{X}_{F4}$ gives a p value of .311.

Starting with the second column of Table 2, we next applied backward elimination, at each step deleting the predictor with the largest p value greater than .05. Columns 3 and 4 of Table 2

Table 2. p Values From Marginal Predictor Tests Applied to Lean Body Mass Regression via Backward Elimination

X	*Step 0*	*Step 3*	*Step 4*
SSF	.000	.000	.000
Wt	.000	.000	.000
PFC	.022	.031	.032
RCC	.059	.049	.001
Hc	.121	.205	*
Ht	.399	*	*
Hg	.809	*	*
WCC	.898	*	*

NOTE: Asterisks indicate deleted predictors.

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

Table 3. Estimated Sufficient Predictors

$\boldsymbol{X}$	$\hat{\eta}_1^T\boldsymbol{X}$	$\hat{\eta}_2^T\boldsymbol{X}$	$\hat{\eta}_3^T\boldsymbol{X}$
log (PFC)	.007	−.304	.541
log (RCC)	.013	−.426	−.748
log (SSF)	−.341	.690	−.328
log (Wt)	.940	.500	.198

give the p values for the final two steps in this process. Note that the p value for RCC decreased substantially after removal of Hc, due perhaps to their high sample correlation. We checked on that possibility by testing the marginal hypothesis that the last five predictors listed are independent of the response given the first three predictors. The p value of .002 supports the conjecture.

At an operational level, an IR analysis of predictor contributions using marginal or conditional predictor tests can proceed like a traditional analysis of a homoscedastic linear regression model using F tests, although here no model for $Y|\mathbf{X}$ is assumed. We used marginal tests in this illustration, but conditional tests could have been used instead. For example, given $d = 3$, the conditional test [(14) and Corollary 1] of the hypothesis $Y \perp\!\!\!\perp \mathbf{X}_{L4}|\mathbf{X}_{F4}$ resulted in a p value of .463, which agrees with the results of the marginal test reported earlier. As a check on the final conclusion, we tested the joint (Sec. 3.4.2) hypothesis $d = 3$ and $Y \perp\!\!\!\perp \mathbf{X}_{L4}|\mathbf{X}_{F4}$, obtaining a p value of .12.

Finally, Table 3 presents the estimated sufficient predictors under the inference that $d = 3$ and $Y \perp\!\!\!\perp \mathbf{X}_{L4}|\mathbf{X}_{F4}$, scaling the predictors marginally so that each has a sample standard deviation of 1. The interpretation of the estimated predictors is as discussed at the end of the algorithm in Section 3.3. Given the results to this point, the analysis could be continued in various ways, depending on application-specific requirements. Illustrations on sufficient dimension reduction are available throughout the literature (see, e.g., Chen and Li 1998; Li, Cook, and Chiaromonte 2003).

7. SLICED AVERAGE VARIANCE ESTIMATION

It has long been recognized that the coverage condition fails in certain highly symmetric regressions. For example, if $Y = X_1 + X_2^2 + \epsilon$, where the predictors and the error are independent standard normal variables, then $\mathrm{E}(X_2|Y) = 0$, and IR methods based on $\mathrm{E}(\mathbf{X}|Y)$ will fail to find the X_2 direction. Cook and Weisberg (1991) proposed the SAVE method to deal with such possibilities. Although SAVE has been considered a useful complement to SIR, its development has lagged because of the technical difficulty in dealing with the distribution of eigenvalues of quadratic functions of covariance matrices. In this section, we advance the data-analytic capabilities of SAVE by using IR to derive the large-sample distribution of its test statistic for dimension. These results might be particularly useful in recent bioinformatics applications of SAVE (Bura and Pfeiffer 2003; Antoniadis, Lambert-Lacroix, and Leblack 2003).

SAVE is like SIR, but uses a different kernel matrix. Recall that $\boldsymbol{\Sigma}_{\mathbf{Z}|Y} = \mathrm{cov}(\mathbf{Z}|Y)$ and define the kernel matrix $\mathbf{M}_{\mathrm{SAVE}} = \mathrm{E}\{(\mathbf{I} - \boldsymbol{\Sigma}_{\mathbf{Z}|Y})^2\}$. Then, under the linearity and marginal covariance (Corollary 2) conditions, $\mathrm{Span}(\mathbf{M}_{\mathrm{SAVE}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. Consequently, if $\dim(\mathcal{S}_{Y|\mathbf{Z}})$ is known and if there is coverage $\mathrm{Span}(\mathbf{M}_{\mathrm{SAVE}}) = \mathcal{S}_{Y|\mathbf{Z}}$, then the subspace spanned by the eigenvectors corresponding to the $\dim(\mathcal{S}_{Y|\mathbf{Z}})$ largest eigenvalues of the sample version $\widehat{\mathbf{M}}_{\mathrm{SAVE}}$ of $\mathbf{M}_{\mathrm{SAVE}}$ is a consistent estimate of $\mathcal{S}_{Y|\mathbf{Z}}$. The SAVE test statistic for the dimension hypothesis $\dim(\mathcal{S}_{Y|\mathbf{Z}}) = m$ is $\Psi_m = n\sum_{i=m+1}^{p}\hat{\psi}_i$, where $\hat{\psi}_1 > \cdots > \hat{\psi}_p > 0$ are the ordered eigenvalues of $\widehat{\mathbf{M}}_{\mathrm{SAVE}}$. This statistic can in principle be used to estimate $\dim(\mathcal{S}_{Y|\mathbf{Z}})$ in the manner described in Section 2.2, but its asymptotic null distribution is unavailable. A permutation test for $\dim(\mathcal{S}_{Y|\mathbf{Z}}) = m$ is available (Cook and Yin 2001), but this procedure is computationally time-consuming and requires additional assumptions.

The conditions needed to ensure that $\mathrm{Span}(\mathbf{M}_{\mathrm{SAVE}}) = \mathcal{S}_{Y|\mathbf{Z}}$ are not needed for the asymptotic distribution of Ψ_m under the null hypothesis $\dim(\mathbf{M}_{\mathrm{SAVE}}) = m$. Thus in the next section we proceed without reference to these conditions.

7.1 SAVE in the IR Family

The sample version of $\mathbf{M}_{\mathrm{SAVE}}$ can be represented as

$$\widehat{\mathbf{M}}_{\mathrm{SAVE}} = \sum_{y=1}^{h}\hat{f}_y(\mathbf{I} - \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}|y})^2$$

$$= \sum_{y=1}^{h}\hat{f}_y\{\hat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}} - \hat{\boldsymbol{\Sigma}}_{\mathbf{Z}|y})\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\Sigma}}^{-1/2}\}^2,$$

where $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}|y}$ is the sample covariance matrix for the sample version of $\mathbf{Z}$, $\hat{\mathbf{Z}} = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X} - \bar{\mathbf{X}}_{\cdot\cdot})$. Let $d = \mathrm{rank}(\mathbf{M}_{\mathrm{SAVE}})$. According to Lemma A.1 in Appendix A, the span of the eigenvectors corresponding to the d largest eigenvalues of $\widehat{\mathbf{M}}_{\mathrm{SAVE}}$ can be found by minimizing

$$G_d(\mathbf{A}, \mathbf{K}) = \sum_{y=1}^{h}\big\|\mathrm{vec}\{\hat{f}_y^{1/2}(\mathbf{I} - \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}|y})\} - \mathrm{vec}(\mathbf{A}\mathbf{K}_y)\big\|^2$$

over $\mathbf{A} \in \mathbb{R}^{p\times d}$ and $\mathbf{K}_y \in \mathbb{R}^{d\times p}$, with $\mathbf{K} = (\mathbf{K}_1, \dots, \mathbf{K}_h)$. Let $\hat{\mathbf{A}}$ and $\hat{\mathbf{K}}$ be values of $\mathbf{A}$ and $\mathbf{K}$ that minimize G_d, and let $\hat{G}_d = G_d(\hat{\mathbf{A}}, \hat{\mathbf{K}})$. Then $\mathrm{Span}(\hat{\mathbf{A}})$ equals the space spanned by the d largest eigenvectors of $\widehat{\mathbf{M}}_{\mathrm{SAVE}}$, and $n\hat{G}_d = \Psi_d$. The minimizing value $\hat{\mathbf{A}}$ is not necessarily unique, but $\mathrm{Span}(\hat{\mathbf{A}})$ and $\hat{G}_d$ are unique. The eigenvectors corresponding to the d largest eigenvalues of $\widehat{\mathbf{M}}_{\mathrm{SAVE}}$ always provide one solution, $\hat{\mathbf{A}}$. Here G_d is used only to aid in deriving the distribution of Ψ_d; computations at the end will still be based on the spectral decomposition of $\widehat{\mathbf{M}}_{\mathrm{SAVE}}$.

Finding the distribution of $n\hat{G}_d$ is facilitated by working in terms of a related discrepancy function that has the same minimum value as G_d. Let $\mathbf{S}_y = \hat{f}_y(\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}} - \hat{\boldsymbol{\Sigma}}_{\mathbf{Z}|y})$, $\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_h) \in \mathbb{R}^{p\times ph}$, and $\mathbf{V}_n = \mathrm{diag}\{\mathbf{V}_{ny}\} \in \mathbb{R}^{p^2h\times p^2h}$, where

$$\mathbf{V}_{ny} = \hat{f}_y^{-1}\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^{1/2} \otimes \boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^{1/2} \in \mathbb{R}^{p^2\times p^2}.$$

Then $\hat{G}_d$ is equal to the minimum of

$$F_d(\mathbf{B}, \mathbf{C}) = \{\mathrm{vec}(\mathbf{S}) - \mathrm{vec}(\mathbf{B}\mathbf{C})\}^T\mathbf{V}_n\{\mathrm{vec}(\mathbf{S}) - \mathrm{vec}(\mathbf{B}\mathbf{C})\} \quad (20)$$

over $\mathbf{B} \in \mathbb{R}^{p\times d}$, and $\mathbf{C} = (\mathbf{C}_1, \dots, \mathbf{C}_h)$ with $\mathbf{C}_y \in \mathbb{R}^{d\times p}$. The advantage of F_d is that it is written in terms of the unobservable variable $\mathbf{Z}$ instead of its sample counterpart $\hat{\mathbf{Z}}$, which facilitates asymptotic calculations. Because the minimum values

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

$\hat{G}_d$ and $\hat{F}_d$ of G_d and F_d are the same, $n\hat{G}_d = n\hat{F}_d = \Psi_d$ and we can pursue the distribution of Ψ_d through $n\hat{F}_d$. Notice also that $\mathbf{B}$ and $\mathbf{C}_y$ are defined differently from $\mathbf{A}$ and $\mathbf{K}_y$, with $\mathbf{B} = \boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}^{1/2}\mathbf{A}$. The next step is to find the asymptotic distribution of $\operatorname{vec}(\mathbf{S})$.

7.2 Asymptotic Distribution of S

We know that $\hat{f}_y$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$, and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}|y}$ converge to f_y, $\mathbf{I}$, and $\boldsymbol{\Sigma}_{\mathbf{Z}|y}$ in probability. Thus $\mathbf{S}_y \to f_y(\mathbf{I} - \boldsymbol{\Sigma}_{\mathbf{Z}|y})$ in probability as $n_y \to \infty$. We want to show that $\mathbf{S}_y$ has asymptotic normal distribution with mean $\boldsymbol{\xi}_y = f_y(\mathbf{I} - \boldsymbol{\Sigma}_{\mathbf{Z}|y})$.

Let $\boldsymbol{\mu}_y = \mathrm{E}(\mathbf{Z}|Y = y)$, and recall that J_y is the slice indicator, as defined near (6). The first term of $\mathbf{S}_y$, $\hat{f}_y\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$, can be expanded as

$$\begin{aligned}\hat{f}_y\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}} &= f_y\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}} + (\hat{f}_y - f_y)\mathbf{I} + (\hat{f}_y - f_y)(\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}} - \mathbf{I})\\ &= \frac{1}{n}\sum_{i=1}^{n} f_y(\mathbf{Z}_i - \bar{\mathbf{Z}}_{\cdot\cdot})(\mathbf{Z}_i - \bar{\mathbf{Z}}_{\cdot\cdot})^T\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(J_{yi} - f_y)\mathbf{I} + O_p(1/n)\\ &= \frac{1}{n}\sum_{i=1}^{n}\{f_y\mathbf{Z}_i\mathbf{Z}_i^T + (J_{yi} - f_y)\mathbf{I}\} + O_p(1/n).\end{aligned}$$

The second term of $\mathbf{S}_y$, $\hat{f}_y\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}|y}$, has the expansion

$$\begin{aligned}\hat{f}_y\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}|y} &= \frac{1}{n}\sum_{j=1}^{n_y}(\mathbf{Z}_{yj} - \bar{\mathbf{Z}}_{y\cdot})(\mathbf{Z}_{yj} - \bar{\mathbf{Z}}_{y\cdot})^T\\ &= \frac{1}{n}\sum_{i=1}^{n} J_{yi}(\mathbf{Z}_i - \boldsymbol{\mu}_y)(\mathbf{Z}_i - \boldsymbol{\mu}_y)^T + O_p(1/n).\end{aligned}$$

Therefore, letting $\mathbf{T}_y = f_y\mathbf{Z}\mathbf{Z}^T - J_y(\mathbf{Z} - \boldsymbol{\mu}_y)(\mathbf{Z} - \boldsymbol{\mu}_y)^T - (J_y - f_y)\mathbf{I} - \boldsymbol{\xi}_y$ with $\mathrm{E}(\mathbf{T}_y) = \mathbf{0}$, we have

$$\mathbf{S}_y - \boldsymbol{\xi}_y = \frac{1}{n}\sum_{i=1}^{n}\mathbf{T}_{yi} + O_p(1/n),$$

where $\mathbf{T}_{yi}$ is the ith realization of $\mathbf{T}_y$. Because $(J_{yi}, \mathbf{X}_i)$, $i = 1, \ldots, n$, are iid observations, the $\mathbf{T}_{yi}$ are iid as well. Therefore, $\sqrt{n}\{\operatorname{vec}(\mathbf{S}_y) - \operatorname{vec}(\boldsymbol{\xi}_y)\}$ converges in law to a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Gamma}_y = \operatorname{cov}\{\operatorname{vec}(\mathbf{T}_y)\}$. Letting $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_h)$ and $\mathbf{T} = (\mathbf{T}_1, \ldots, \mathbf{T}_h)$, it follows that $\sqrt{n}\{\operatorname{vec}(\mathbf{S}) - \operatorname{vec}(\boldsymbol{\xi})\}$ converges in law to a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Gamma} = \operatorname{cov}\{\operatorname{vec}(\mathbf{T})\}$.

7.3 Asymptotic Distribution of Test Statistic $n\hat{F}_d$

Letting $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ be a basis for $\operatorname{Span}(\boldsymbol{\xi}) = \operatorname{Span}(\mathbf{M}_{\text{SAVE}})$, we have

$$\boldsymbol{\xi} = (\boldsymbol{\beta}\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\beta}\boldsymbol{\gamma}_h) = \boldsymbol{\beta}\boldsymbol{\gamma},$$

where $\boldsymbol{\gamma}_y \in \mathbb{R}^{d\times p}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{d\times ph}$ are implicitly defined. Define the positive-definite matrix $\mathbf{V} = \operatorname{diag}\{f_y^{-1}\mathbf{I}_{p^2}\}$, which is the limit of inner-product matrix $\mathbf{V}_n$, and let $\boldsymbol{\Omega} = \mathbf{V}^{1/2}\boldsymbol{\Gamma}\mathbf{V}^{1/2}$. Finally, let $\boldsymbol{\Phi} = \mathbf{V}^{1/2}\boldsymbol{\Delta}$, where $\boldsymbol{\Delta} = (\boldsymbol{\gamma}^T \otimes \mathbf{I}_p, \mathbf{I}_{ph} \otimes \boldsymbol{\beta})$ is the Jacobian matrix for discrepancy function (20) evaluated at $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. The asymptotic distribution of $n\hat{F}_d = \Psi_d$ is given in the following theorem, the proof of which follows that of Theorem 2 and thus is omitted here.

Theorem 4. Assume that the data $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, are a simple random sample of $(\mathbf{X}, Y)$ with finite fourth moments. Let $d = \operatorname{rank}(\mathbf{M}_{\text{SAVE}})$ and let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \arg_{\mathbf{B},\mathbf{C}}\min F_d(\mathbf{B}, \mathbf{C})$ as defined previously in (20). Then, as $n \to \infty$, $n\hat{F}_d \overset{\mathcal{D}}{\to} \sum_{i=1}^{p^2h}\lambda_i\chi_i^2(1)$, where $\{\chi_i^2(1)\}$ are independent chi-squared random variables each with 1 degree of freedom and $\lambda_1 \geq \cdots \geq \lambda_{ph}$ are the eigenvalues of $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$.

Like Theorems 2 and 3, this theorem is quite general, requiring none of the regularity conditions discussed previously. If the linearity and marginal covariance conditions hold, then $\operatorname{Span}(\boldsymbol{\xi}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. In this case we can use Theorem 4 to infer about a possibly proper subset of the central subspace. Under the linearity, coverage, and marginal covariance conditions, $\operatorname{Span}(\mathbf{M}_{\text{SAVE}}) = \mathcal{S}_{Y|\mathbf{Z}}$. In a special case, if Y is independent of $\mathbf{Z}$, then $\boldsymbol{\beta}\boldsymbol{\gamma} = \mathbf{0}$ and $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}} = \boldsymbol{\Omega}$.

7.4 Computation

To use Theorem 4 in practice, we need to replace $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$ with a consistent estimate under the null hypothesis $d = m$. We have $\boldsymbol{\Omega} = \mathbf{V}^{1/2}\boldsymbol{\Gamma}\mathbf{V}^{1/2}$. We estimate $\boldsymbol{\Gamma}$ with

$$\hat{\boldsymbol{\Gamma}}_y = \frac{1}{n}\sum_{i=1}^{n}\operatorname{vec}(\hat{\mathbf{T}}_i)\operatorname{vec}(\hat{\mathbf{T}}_i)^T,$$

where $\hat{\mathbf{T}}_i = (\hat{\mathbf{T}}_{1i}, \ldots, \hat{\mathbf{T}}_{hi})$ and

$$\hat{\mathbf{T}}_{yi} = \hat{f}_y\hat{\mathbf{Z}}_i\hat{\mathbf{Z}}_i^T + J_{yi}\mathbf{I} + J_{yi}(\hat{\mathbf{Z}}_i - \hat{\mathbf{Z}}_{y\cdot})(\hat{\mathbf{Z}}_i - \hat{\mathbf{Z}}_{y\cdot})^T - 2\hat{f}_y\mathbf{I} + \hat{f}_y\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}|y}.$$

Here $\hat{\mathbf{Z}}_{y\cdot}$ is the average of $\hat{\mathbf{Z}}$ within the yth slice.

Meanwhile,

$$\boldsymbol{\Phi} = \left[(f_1^{-1/2}\boldsymbol{\gamma}_1, \ldots, f_h^{-1/2}\boldsymbol{\gamma}_h)^T \otimes \mathbf{I}_p, \operatorname{diag}\{f_y^{-1/2}\mathbf{I}_p \otimes \boldsymbol{\beta}\}\right],$$

and we need estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_y$ to estimate $\mathbf{Q}_{\boldsymbol{\Phi}}$. We estimate $\boldsymbol{\beta}$ using the matrix $\hat{\boldsymbol{\beta}}$ whose columns are the first m eigenvectors of $\widehat{\mathbf{M}}_{\text{SAVE}}$ and estimate $\boldsymbol{\gamma}_y$ using $\hat{\boldsymbol{\gamma}}_y = \hat{f}_y\hat{\boldsymbol{\beta}}^T(\mathbf{I} - \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}|y})$, then construct the sample eigenvalues $\hat{\psi}_j$ of $\mathbf{Q}_{\hat{\boldsymbol{\Phi}}}\hat{\boldsymbol{\Omega}}\mathbf{Q}_{\hat{\boldsymbol{\Phi}}}$. We then compare the statistic $n\hat{F}_m = \Psi_m$ to the percentage points of the distribution of $\sum_{i=1}^{p^2h}\hat{\psi}_i\chi_i^2(1)$ to obtain a p value, treating the $\hat{\psi}_j$ as fixed.

We conducted a variety of simulations to support the asymptotic calculations and reaffirm the recognized differences between methods based on first inverse moments (SIR and IRE) and SAVE, which is based on second inverse moments. In all cases our numerical results agreed well with the asymptotic predictions.

8. DISCUSSION

In this article we have developed a general approach to SDR based on the IR family, leading to the IRE and various conditional independence tests. One important conclusion is that the IRE always performed at least as well as SIR. We studied many different regressions via simulation, but did not find one in which SIR clearly dominated. On the other hand, as suggested by the simulations of Section 6, it was not hard to find situations in which IR methods dominated the corresponding SIR methods.

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

In theory, this approach subsumes much of the past work on SIR and promises to cover other methods as well. There is a parallel IR family that includes SAVE, and our first use of it led to a relatively straightforward derivation of the asymptotic distribution of SAVE's dimension test statistic, a problem that has apparently eluded solution since the introduction of SAVE by Cook and Weisberg (1991). Recently, Ye and Weiss (2003) proposed bootstrap methods to combine information from SIR and SAVE. The general approach developed here has considerable potential for deriving methods that optimally combine information from first and second inverse moments.

The alternating least squares algorithm (Sec. 3.3) seems to be quite stable, and, although we can guarantee convergence to only a local minimum, our simulation results suggest that it typically converges to the global minimum. We reached this conclusion in part because our simulation results agree well with theoretical predictions [see, e.g., Fig. 6(b)], which we would not expect if convergence to a local minimum were a prevalent problem. Nevertheless, a variety of perhaps randomly selected starting vectors could be used to safeguard against getting trapped at a local minimum. The computational cost of the alternating least squares algorithm is greater than that of SIR's relatively simple spectral algorithm. For example, running IRE alone, the response time for 200 simulated datasets from Model A was about twice that for running SIR alone. Such differences can be annoying when performing thousands of simulations, but are not noticeable for a single dataset. The response time increases with p for both SIR and IRE, but even for $p = 20$ the difference seems small with a 1.25-GHz processor.

Nevertheless, faster algorithms are possible. Statistical properties of IREs and their computational complexity depend on the middle matrix $\mathbf{V}_n$ in (3). At one extreme, we might let $\mathbf{V}_n = \mathbf{I}_{h-1} \otimes \hat{\boldsymbol{\Sigma}}$ to obtain an SIR-type spectral decomposition solution, which is easy to compute and guarantees reaching the global minimum. However, like SIR, this solution will be asymptotically inefficient and its asymptotic distributions in general will be more complicated than chi-squared distributions. Meanwhile, the IRE uses $\mathbf{V}_n = \widehat{\boldsymbol{\Gamma}}_{\hat{\xi}}^{-1}$ to achieve asymptotic efficiency at the expense of the computational algorithm. We might consider intermediate estimators with middle matrices "between" these extremes. In particular, a faster version might be constructed in part by approximating $\boldsymbol{\Gamma}$ with a matrix that requires only second moments. One possibility is $\tilde{\boldsymbol{\Gamma}} = \operatorname{cov}(\operatorname{vec}(\boldsymbol{\Sigma}^{-1/2}\mathbf{Z}\tilde{\boldsymbol{\varepsilon}}^T)) \in \mathbb{R}^{ph\times ph}$, where $\tilde{\boldsymbol{\varepsilon}} \in \mathbb{R}^h$ with elements $\tilde{\varepsilon}_y = J_y - \mathrm{E}(J_y)$, $y = 1, 2, \dots, h$. Like $n\hat{F}_d^{\mathrm{ire}}$, the asymptotic distribution of $n\hat{F}_d$ using the plug-in estimate of $\tilde{\boldsymbol{\Gamma}}$ is chi-squared with $(p-d)(h-d-1)$ degrees of freedom. Although it is faster than IRE, we may lose significant asymptotic efficiency, and the corresponding asymptotic distributions for the predictor tests come from linear combinations of chi-squares instead of just chi-squares. Because this IR estimator requires only second moments instead of the fourth moments as IRE does, it might have better convergence properties and be more robust to outliers. At this point, it is not obvious how strike an effective balance between speed, convergence to the global minimum, robustness, and statistical efficiency, but work along this line is in progress.

Backward elimination in linear regression has long been known to give practically useful results, and the variable elimination procedure in Section 6.4 was used in this spirit. It is possible to develop consistent variable selection methods in SDR by adding an appropriate penalty to the discrepancy function, much like the penalties in model selection procedures like the Akaike information criterion and the Bayes information criterion. Similar comments apply to inference on dimension. Work along these lines is in progress as well.

APPENDIX A: LEMMA FOR SPECTRAL DECOMPOSITION

Lemma A.1. Suppose that $\mathbf{A} = \sum_{i=1}^n \mathbf{a}_i\mathbf{a}_i^T$ with a spectral decomposition $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where $\mathbf{a}_i \in \mathbb{R}^p$, $\boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_i\}$, and $\mathbf{U} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_p)$ are its eigenvectors corresponding to eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$. Let $\kappa_m = \sum_{j=m+1}^p \lambda_j$, $m = 0, 1, \dots, p-1$. Then

$$\kappa_m = \min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times m}, \boldsymbol{\gamma}_i\in\mathbb{R}^m} \sum_{i=1}^n (\mathbf{a}_i - \boldsymbol{\beta}\boldsymbol{\gamma}_i)^T(\mathbf{a}_i - \boldsymbol{\beta}\boldsymbol{\gamma}_i)$$

and $\operatorname{Span}\{\hat{\boldsymbol{\beta}}\} = \operatorname{Span}\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_m\}$. Here $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that minimizes $\sum_{i=1}^n (\mathbf{a}_i - \boldsymbol{\beta}\boldsymbol{\gamma}_i)^T(\mathbf{a}_i - \boldsymbol{\beta}\boldsymbol{\gamma}_i)$.

Proof. The proof seems straightforward and thus is omitted.

APPENDIX B: PROOF OF THEOREM 1

The strategy for showing asymptotic normality is to decompose $\sqrt{n}(\operatorname{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \operatorname{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}}))$ as a summation of iid observations plus a remainder converging to 0 in probability. Then we obtain the desired results by the central limit theorem. In this decomposition process, we need the following lemma, which decomposes the difference between the inverse of a sample covariance matrix and its population value. It follows straightforwardly from equation 35 of Li et al. (2003).

Lemma A.2. Suppose that a random vector $\mathbf{X}$ has covariance matrix $\boldsymbol{\Sigma} > \mathbf{0}$. Then

$$\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1} = -n^{-1}\boldsymbol{\Sigma}^{-1/2}\sum_{j=1}^n (\mathbf{Z}_j\mathbf{Z}_j^T - \mathbf{I})\boldsymbol{\Sigma}^{-1/2} + O_p(n^{-1}).$$

Here $\hat{\boldsymbol{\Sigma}}$ is the sample covariance calculated from a sample of size n, and $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \mathrm{E}[\mathbf{X}])$ is the standardized version of $\mathbf{X}$.

Recall that $\bar{\mathbf{X}}_{y\cdot}$ is the average of the n_y observations in the yth slice and $\bar{\mathbf{X}}_{\cdot\cdot}$ is the average of all n observations. Letting $\boldsymbol{\mu}_y = \mathrm{E}[\bar{\mathbf{X}}_{y\cdot}]$ and $\boldsymbol{\mu} = \mathrm{E}[\bar{\mathbf{X}}_{\cdot\cdot}]$, consider

$$\begin{aligned}
&\sqrt{n}(\hat{f}_y\hat{\boldsymbol{\xi}}_y - f_y\boldsymbol{\xi}_y)\\
&\quad = \sqrt{n}\hat{f}_y\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{X}}_{y\cdot} - \bar{\mathbf{X}}_{\cdot\cdot}) - \sqrt{n}f_y\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu})\\
&\qquad + \sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})[\hat{f}_y(\bar{\mathbf{X}}_{y\cdot} - \bar{\mathbf{X}}_{\cdot\cdot}) - f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})]\\
&\quad = \sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})\\
&\qquad + \sqrt{n}\boldsymbol{\Sigma}^{-1}[\hat{f}_y(\bar{\mathbf{X}}_{y\cdot} - \bar{\mathbf{X}}_{\cdot\cdot}) - f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})]\\
&\qquad + O_p(n^{-1/2}). \qquad\qquad \text{(B.1)}
\end{aligned}$$

By Lemma A.2, we have

$$\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1} = -n^{-1}\boldsymbol{\Sigma}^{-1/2}\sum_{j=1}^n (\mathbf{Z}_j\mathbf{Z}_j^T - \mathbf{I})\boldsymbol{\Sigma}^{-1/2} + O_p(n^{-1}).$$

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

Therefore, the first term in (B.1) can be simplified as

$$
\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) \\
&\qquad = -n^{-1/2}\boldsymbol{\Sigma}^{-1/2}\sum_{j=1}^{n}(\mathbf{Z}_j\mathbf{Z}_j^T - \mathbf{I})\mathrm{E}[\mathbf{Z}J_y] + O_p(n^{-1/2}).
\end{aligned} \tag{B.2}
$$

Meanwhile, letting J_{yj} denote the value of J_y for the jth observation, $j = 1, 2, \ldots, n$, we have

$$
\begin{aligned}
&\hat{f}_y(\bar{\mathbf{X}}_{y\cdot} - \bar{\mathbf{X}}_{\cdot\cdot}) \\
&\quad = \frac{1}{n}\sum_{j=1}^{n}[(\mathbf{X}_j - \bar{\mathbf{X}}_{\cdot\cdot})J_{yj}] \\
&\quad = \frac{1}{n}\sum_{j=1}^{n}\big[(\mathbf{X}_j - \boldsymbol{\mu})(J_{yj} - \mathrm{E}[J_y])\big] - \frac{1}{n}(\bar{\mathbf{X}}_{\cdot\cdot} - \boldsymbol{\mu})\sum_{j=1}^{n}(J_{yj} - \mathrm{E}[J_y]) \\
&\quad = \frac{1}{n}\sum_{j=1}^{n}\big[(\mathbf{X}_j - \boldsymbol{\mu})(J_{yj} - \mathrm{E}[J_y])\big] + O_p(n^{-1}).
\end{aligned}
$$

Therefore, the second term in (B.1) can be simplified as

$$
\begin{aligned}
&\sqrt{n}\boldsymbol{\Sigma}^{-1}[\hat{f}_y(\bar{\mathbf{X}}_{y\cdot} - \bar{\mathbf{X}}_{\cdot\cdot}) - f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})] \\
&\quad = n^{-1/2}\boldsymbol{\Sigma}^{-1/2}\sum_{j=1}^{n}\big[\boldsymbol{\Sigma}^{-1/2}(\mathbf{X}_j - \boldsymbol{\mu})(J_{yj} - \mathrm{E}[J_y])\big] \\
&\qquad - \sqrt{n}\boldsymbol{\Sigma}^{-1}f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) + O_p(n^{-1/2}) \\
&\quad = n^{-1/2}\boldsymbol{\Sigma}^{-1/2}\sum_{j=1}^{n}\big[\mathbf{Z}_j(J_{yj} - \mathrm{E}[J_y]) - \mathrm{E}[\mathbf{Z}J_y]\big] + O_p(n^{-1/2}).
\end{aligned} \tag{B.3}
$$

Plugging (B.2) and (B.3) into (B.1), we obtain

$$
\begin{aligned}
&\sqrt{n}(\hat{f}_y\hat{\boldsymbol{\xi}}_y - f_y\boldsymbol{\xi}_y) \\
&\quad = n^{-1/2}\boldsymbol{\Sigma}^{-1/2}\sum_{j=1}^{n}\big[\mathbf{Z}_j(J_{yj} - \mathrm{E}[J_y]) - \mathrm{E}[\mathbf{Z}J_y] - (\mathbf{Z}_j\mathbf{Z}_j^T - \mathbf{I})\mathrm{E}[\mathbf{Z}J_y]\big] \\
&\qquad + O_p(n^{-1/2}) \\
&\quad = n^{-1/2}\boldsymbol{\Sigma}^{-1/2}\sum_{j=1}^{n}[\mathbf{Z}_j\varepsilon_{yj}] + O_p(n^{-1/2}),
\end{aligned}
$$

where $\varepsilon_{yj} = J_{yj} - \mathrm{E}[J_y] - \mathbf{Z}_j^T\mathrm{E}[\mathbf{Z}J_y]$ is the jth value for ε_y. Let $\boldsymbol{\epsilon}_j = [\varepsilon_{1j}, \ldots, \varepsilon_{hj}]^T$ be the jth value for the random vector $\boldsymbol{\varepsilon}$ [cf. (6)]. We then have

$$
\begin{aligned}
&\sqrt{n}\big(\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}})\big) \\
&\qquad\qquad = n^{-1/2}\sum_{j=1}^{n}\mathrm{vec}(\boldsymbol{\Sigma}^{-1/2}\mathbf{Z}_j\boldsymbol{\epsilon}_j^T) + O_p(n^{-1/2}),
\end{aligned}
$$

where $(\mathbf{Z}_j, \boldsymbol{\epsilon}_j)$ are iid random vectors. Thus

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

where $\boldsymbol{\Gamma} = \mathrm{cov}(\mathrm{vec}(\boldsymbol{\Sigma}^{-1/2}\mathbf{Z}\boldsymbol{\varepsilon}^T))$.

APPENDIX C: PROOF OF THEOREM 2

C.1 Preparations

The proof of Theorem 2 hinges on Shapiro's (1986) results on the asymptotics of overparameterized discrepancy functions and two supplemental lemmas. We first give these results, then show how they can be used to prove the theorem.

Proposition A.1 (Shapiro 1986, props. 3.1, 3.2, 4.1, and 5.1). Suppose that $\boldsymbol{\theta}$ is a q-dimensional parameter vector that lies in an open and connected parameter space $\Theta \subseteq \mathbb{R}^q$. Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$. Define $\mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_m(\boldsymbol{\theta}))^T : \Theta \to \mathbb{R}^m$, where $g_i(\boldsymbol{\theta})$ is twice continuously differentiable on Θ, $i = 1, \ldots, m$. The Jacobian matrix $\boldsymbol{\Delta} = \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ need not be of full rank, so g can be overparameterized. Also assume the following:

1. $\boldsymbol{\tau}_n$ is an asymptotically normal estimate of the population value $\mathbf{g}(\boldsymbol{\theta}_0) : \sqrt{n}(\boldsymbol{\tau}_n - \mathbf{g}(\boldsymbol{\theta}_0)) \xrightarrow{\mathcal{D}} \mathrm{Normal}(\mathbf{0}, \boldsymbol{\Gamma})$, where n is the sample size.
2. For a known inner-product matrix $\mathbf{V}$, the discrepancy function
$$H(\boldsymbol{\tau}_n, \mathbf{g}(\boldsymbol{\theta})) = (\boldsymbol{\tau}_n - \mathbf{g}(\boldsymbol{\theta}))^T\mathbf{V}(\boldsymbol{\tau}_n - \mathbf{g}(\boldsymbol{\theta}))$$
satisfies the following properties:
p1. $H(\mathbf{a}, \mathbf{b}) \geq 0 \; \forall \, \mathbf{a}, \; \mathbf{b} \in \mathbb{R}^m$.
p2. $H(\mathbf{a}, \mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$.
p3. H is at least twice continuously differentiable in $\mathbf{a}$ and $\mathbf{b}$.
p4. There are positive constants δ and ϵ such that $H(\mathbf{a}, \mathbf{b}) \geq \epsilon$ whenever $\|\mathbf{a} - \mathbf{b}\| \geq \delta$, where $\|\cdot\|$ represents ordinary Euclidean distance.
3. The point $\boldsymbol{\theta}_0$ is regular.
4. $\mathrm{rank}(\boldsymbol{\Delta}) = \mathrm{rank}(\boldsymbol{\Delta}^T\mathbf{V}\boldsymbol{\Delta})$.

Then the following results hold:

1. Letting $\hat{H} = H(\boldsymbol{\tau}_n, \mathbf{g}(\hat{\boldsymbol{\theta}}))$ denote the value of the discrepancy function minimized over Θ, the asymptotic distribution of $n\hat{H}$ is the same as the distribution of the quadratic form $\mathbf{W}^T\mathbf{U}\mathbf{W}$, where $\mathbf{W} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{\Gamma})$,
$$\mathbf{U} = \mathbf{V} - \mathbf{V}\boldsymbol{\Delta}(\boldsymbol{\Delta}^T\mathbf{V}\boldsymbol{\Delta})^{-}\boldsymbol{\Delta}^T\mathbf{V} = \mathbf{V}^{1/2}\mathbf{Q}_{\boldsymbol{\Phi}}\mathbf{V}^{1/2},$$
and $\boldsymbol{\Phi} = \mathbf{V}^{1/2}\boldsymbol{\Delta}$.
2. If $\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}$, then $n\hat{H} \xrightarrow{\mathcal{D}} \chi^2_D$, where the degrees of freedom $D = \mathrm{trace}(\mathbf{U}\boldsymbol{\Gamma})$.
3. The estimate $\mathbf{g}(\hat{\boldsymbol{\theta}})$ that minimizes the discrepancy function is a consistent estimator of $\mathbf{g}(\boldsymbol{\theta}_0)$ and $\sqrt{n}(\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}(\boldsymbol{\theta}_0))$ has an asymptotically normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}^{-1/2}\mathbf{P}_{\boldsymbol{\Phi}}\mathbf{V}^{1/2}\boldsymbol{\Gamma}\mathbf{V}^{1/2}\mathbf{P}_{\boldsymbol{\Phi}}\mathbf{V}^{-1/2}$.
4. When $\boldsymbol{\Gamma}$ is nonsingular, $\mathbf{g}(\hat{\boldsymbol{\theta}})$ is asymptotically efficient and $n\hat{H} \xrightarrow{\mathcal{D}} \chi^2_k$, where the degrees of freedom $k = m - \mathrm{rank}(\boldsymbol{\Delta})$, if and only if $\mathbf{V} = (\boldsymbol{\Gamma} + \boldsymbol{\Delta}\mathbf{D}\boldsymbol{\Delta}^T)^{-1}$, where $\mathbf{D}$ is an arbitrary symmetric matrix.

In our adaptation of Shapiro's results, the inner-product matrix $\mathbf{V}$ is often random rather than fixed, as required in Proposition A.1. The next two lemmas allow us to connect minimum discrepancy functions with fixed inner products to those with random inner products. Lemma A.3 deals with the asymptotic distribution of the minimum discrepancy value. Lemma A.4 covers asymptotic properties of the estimate of Span($\boldsymbol{\beta}$).

Lemma A.3. 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 that $\{\mathbf{V}_n > \mathbf{0}\}$ is a sequence of $s \times s$ matrices that converges to $\mathbf{V} > \mathbf{0}$ in probability. If

$$n\hat{H}_{\mathbf{V}} = \min_{\xi \in \Xi} n(\mathbf{Y}_n - \boldsymbol{\xi})^T\mathbf{V}(\mathbf{Y}_n - \boldsymbol{\xi}) \xrightarrow{\mathcal{D}} \Psi,$$

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

then $n\hat{H}_{\mathbf{V}_n} = \min_{\boldsymbol{\xi}\in\Xi} n(\mathbf{Y}_n - \boldsymbol{\xi})^T\mathbf{V}_n(\mathbf{Y}_n - \boldsymbol{\xi})$ also converges in distribution to Ψ and vice versa.

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

Proof. Because $\mathbf{V}_n \to \mathbf{V}$ in probability, $\Pr[(1-\epsilon)\mathbf{V} < \mathbf{V}_n < (1+\epsilon)\mathbf{V}] \to 1\ \forall \epsilon > 0$. For any $\boldsymbol{\xi} \in \Xi$, if $(1-\epsilon)\mathbf{V} < \mathbf{V}_n < (1+\epsilon)\mathbf{V}$, then

$$(\mathbf{Y}_n - \boldsymbol{\xi})^T(1-\epsilon)\mathbf{V}(\mathbf{Y}_n - \boldsymbol{\xi}) \le (\mathbf{Y}_n - \boldsymbol{\xi})^T\mathbf{V}_n(\mathbf{Y}_n - \boldsymbol{\xi}) \le (\mathbf{Y}_n - \boldsymbol{\xi})^T(1+\epsilon)\mathbf{V}(\mathbf{Y}_n - \boldsymbol{\xi}).$$

Hence the minimum of these functions keeps the same ordering,

$$(1-\epsilon)n\hat{H}_{\mathbf{V}} \le n\hat{H}_{\mathbf{V}_n} \le (1+\epsilon)n\hat{H}_{\mathbf{V}}.$$

Therefore, $\Pr[|\frac{n\hat{H}_{\mathbf{V}_n}}{n\hat{H}_{\mathbf{V}}} - 1| \le \epsilon] \to 1$, that is, $\frac{\hat{H}_{\mathbf{V}_n}}{\hat{H}_{\mathbf{V}}} \xrightarrow{p} 1$. By Slutsky's theorem, $n\hat{H}_{\mathbf{V}_n} \xrightarrow{\mathcal{D}} \Psi$.

Furthermore, because $n\hat{H}_{\mathbf{V}} = n\|\mathbf{V}^{1/2}\mathbf{Y}_n - \mathbf{V}^{1/2}\hat{\boldsymbol{\xi}}_1\|^2 \to \Psi\ \forall \epsilon > 0$,

$$\lim_{n\to\infty} \Pr\bigl[\|\mathbf{V}^{1/2}\mathbf{Y}_n - \mathbf{V}^{1/2}\hat{\boldsymbol{\xi}}_1\|^2 > \epsilon\bigr] = \lim \Pr[n\hat{H}_{\mathbf{V}} > n\epsilon] = \lim \Pr[\Psi > n\epsilon] = 0.$$

Because $\mathbf{V}_n^{1/2} \xrightarrow{p} \mathbf{V}^{1/2}$, $\lim_{n\to\infty} \mathbf{V}^{1/2}\mathbf{Y}_n = \lim \mathbf{V}_n^{1/2}\hat{\boldsymbol{\xi}}_1 = \lim \mathbf{V}_n^{1/2}\hat{\boldsymbol{\xi}}_2$ if the first limit exits.

Lemma A.4. Let $\mathcal{X}_n$ 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 $\boldsymbol{\theta}$ in $\Theta \subseteq \mathbb{R}^k$. Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$. Assume the following:

1. Θ is an open set.
2. The mapping $\mathbf{p}(\boldsymbol{\theta})$ from Θ into $\mathbb{R}^s$ is one-to-one, bicontinuous, and twice continuously differentiable. Let $\mathbf{D}(\boldsymbol{\theta}) = \frac{\partial \mathbf{p}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{s\times k}$ and $\mathbf{D}_0 = \mathbf{D}(\boldsymbol{\theta}_0)$.
3. $\mathbf{Y}_n = \mathbf{Y}_n(\mathcal{X}_n) \in \mathbb{R}^s$ is a consistent estimate of $p(\boldsymbol{\theta}_0)$ with
$$\sqrt{n}(\mathbf{Y}_n - \mathbf{p}(\boldsymbol{\theta}_0)) \xrightarrow{\mathcal{D}} \text{Normal}(\mathbf{0}, \boldsymbol{\Gamma}).$$
4. $\mathbf{V}_n = \mathbf{V}_n(\mathcal{X}_n)$ is a positive-definite matrix that converges to a constant matrix $\mathbf{V}$ in probability.

Define a discrepancy function as

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

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

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \text{Normal}\bigl(\mathbf{0}, (\mathbf{D}_0^T\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0^T\mathbf{V}\boldsymbol{\Gamma}\mathbf{V}\mathbf{D}_0(\mathbf{D}_0^T\mathbf{V}\mathbf{D}_0)^{-1}\bigr)$$

and

$$\sqrt{n}\bigl(\mathbf{p}(\hat{\boldsymbol{\theta}}) - \mathbf{p}(\boldsymbol{\theta}_0)\bigr) \xrightarrow{\mathcal{D}} \text{Normal}\bigl(\mathbf{0}, \mathbf{D}_0(\mathbf{D}_0^T\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0^T\mathbf{V}\boldsymbol{\Gamma}\mathbf{V}\mathbf{D}_0(\mathbf{D}_0^T\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0\bigr).$$

Proof. Define the function $\mathbf{G}(\mathcal{X}_n, \boldsymbol{\theta}) = \mathbf{D}^T(\hat{\boldsymbol{\theta}})\mathbf{V}_n(\mathbf{Y}_n - \mathbf{p}(\boldsymbol{\theta}))$. Let $\mathbf{G}_{\boldsymbol{\theta}}(\mathcal{X}_n, \boldsymbol{\theta}) = -\mathbf{D}^T(\hat{\boldsymbol{\theta}})\mathbf{V}_n\mathbf{D}(\boldsymbol{\theta})$ be the partial derivative of $\mathbf{G}$ with respect to $\boldsymbol{\theta}$. We expand $\mathbf{G}(\mathcal{X}_n, \boldsymbol{\theta}_0)$ about the point $\hat{\boldsymbol{\theta}}$,

$$\mathbf{G}(\mathcal{X}_n, \boldsymbol{\theta}_0) = \mathbf{G}(\mathcal{X}_n, \hat{\boldsymbol{\theta}}) + \biggl[\int_0^1 \mathbf{G}_{\boldsymbol{\theta}}\{\mathcal{X}_n, \hat{\boldsymbol{\theta}} + \lambda(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\}\,d\lambda\biggr](\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}), \quad \text{(C.1)}$$

where the integral of a matrix proceeds elementwise. According to the definition of $\hat{\boldsymbol{\theta}}$, we have

$$\left.\frac{\partial F(\mathbf{Y}_n, \mathbf{p}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^T}\right|_{\hat{\boldsymbol{\theta}}} = -2\mathbf{D}^T(\hat{\boldsymbol{\theta}})\mathbf{V}_n(\mathbf{Y}_n - \mathbf{p}(\hat{\boldsymbol{\theta}})) = \mathbf{0},$$

that is, $\mathbf{G}(\mathcal{X}_n, \hat{\boldsymbol{\theta}}) = \mathbf{0}$. We multiply both sides of (C.1) by $\sqrt{n}$,

$$\sqrt{n}\mathbf{D}^T(\hat{\boldsymbol{\theta}})\mathbf{V}_n(\mathbf{Y}_n - \mathbf{p}(\boldsymbol{\theta}_0)) = -\sqrt{n}\biggl[\int_0^1 \mathbf{G}_{\boldsymbol{\theta}}\{\mathcal{X}_n, \hat{\boldsymbol{\theta}} + \lambda(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\}\,d\lambda\biggr](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

By Lemma A.3, we know that $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$. Thus for any λ,

$$\mathbf{G}_{\boldsymbol{\theta}}\{\mathcal{X}_n, \hat{\boldsymbol{\theta}} + \lambda(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\} = -\mathbf{D}^T(\hat{\boldsymbol{\theta}})\mathbf{V}_n\mathbf{D}\bigl(\hat{\boldsymbol{\theta}} + \lambda(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\bigr)$$

converges to $\mathbf{D}_0^T\mathbf{V}\mathbf{D}_0$ in probability. Therefore, we have

$$\lim_{n\to\infty}\int_0^1 \mathbf{G}_{\theta}\{\mathcal{X}_n, \hat{\boldsymbol{\theta}} + \lambda(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\}\,d\lambda = \int_0^1 \lim_{n\to\infty} \mathbf{G}_{\boldsymbol{\theta}}\{\mathcal{X}_n, \hat{\boldsymbol{\theta}} + \lambda(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\}\,d\lambda = \mathbf{D}_0\mathbf{V}\mathbf{D}_0^T,$$

by the bounded convergence theorem. By Slutsky's theorem and the delta method, we reach the conclusions.

Remark. Lemma A.4 is similar to the modified chi-squared method studied by Ferguson (1958). However, Ferguson considered only the cases where $\mathbf{V}_n$ is a function of the statistic $\mathbf{Y}_n(\mathcal{X}_n)$. Lemma A.4 generalizes the result for all $\mathbf{V}_n$ are general functions of the whole sample $\mathcal{X}_n$.

Lemma A.4 shows that the asymptotic distribution of $\mathbf{p}(\hat{\boldsymbol{\theta}})$ is the same for different series of $\mathbf{V}_n$ as long as $\mathbf{V}_n$ converges to the same $\mathbf{V}$ in probability. It is easy to see that parameterization $\boldsymbol{\theta}$ does not affect the asymptotic properties of $\mathbf{p}(\hat{\boldsymbol{\theta}})$. Thus, for simplicity, we can impose $\mathbf{V}_n = \mathbf{V}$ when considering asymptotic properties of $\mathbf{p}(\hat{\boldsymbol{\theta}})$, if we can show that there exists one parameterization that satisfies the conditions in the statement of the lemma.

C.2 Proof of Theorem 2

In Theorem 2 we consider a discrepancy function,

$$F_d(\mathbf{B}, \mathbf{C}) = \bigl(\text{vec}(\hat{\boldsymbol{\zeta}}) - \text{vec}(\mathbf{BC})\bigr)^T\mathbf{V}_n\bigl(\text{vec}(\hat{\boldsymbol{\zeta}}) - \text{vec}(\mathbf{BC})\bigr).$$

We first address the issue that the inner-product matrix in Proposition A.1 is assumed to be known, whereas the inner-product matrix in F_d is estimated. Because $\mathbf{V}_n$ converges to $\mathbf{V}$ in probability, it follows from Lemma A.3 that the asymptotic distribution of $n\hat{F}_d$ is the same as that of $n\hat{H}_d$, where

$$H_d(\mathbf{B}, \mathbf{C}) = \bigl(\text{vec}(\hat{\boldsymbol{\zeta}}) - \text{vec}(\mathbf{BC})\bigr)^T\mathbf{V}\bigl(\text{vec}(\hat{\boldsymbol{\zeta}}) - \text{vec}(\mathbf{BC})\bigr).$$

Furthermore, we want to show the asymptotic distribution of $\text{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}})$ of $F_d(\mathbf{B}, \mathbf{C})$ is the same as that of $H_d(\mathbf{B}, \mathbf{C})$. Based on the remarks about Lemma A.4, we need only show that there is one parameterization that satisfies the conditions in the statement of Lemma A.4. We can use any full-rank reparameterization of $(\boldsymbol{\beta}, \boldsymbol{\nu})$. For instance, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, 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; otherwise, we need only change the order of the elements in $\mathbf{X}$. Then

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

Therefore, we can set $\boldsymbol{\beta}_1 = \mathbf{I}_d$, leading to the new parameters $\boldsymbol{\beta}_2 \in \mathbb{R}^{(p-d)\times d}$ and $\boldsymbol{\nu} \in \mathbb{R}^{d\times(h-1)}$, which together correspond to the $\boldsymbol{\theta}$ in Lemma A.4. This new parameterization brings a full-rank Jacobian matrix and an open parameter space in $\mathbb{R}^{d(h+p-d-1)}$, thus satisfying the conditions in Lemma A.4. At same time, it affects neither our algorithm for minimization nor our asymptotic results. From now on, we need to prove only the conclusions for H_d.

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017

$H_d(\mathbf{B}, \mathbf{C})$ is in the form of Shapiro's discrepancy function H. This can be seen by setting

$$\boldsymbol{\theta} = \begin{pmatrix} \mathrm{vec}(\mathbf{B}) \\ \mathrm{vec}(\mathbf{C}) \end{pmatrix} \in \mathbb{R}^{d(p+h-1)},$$

$$\mathbf{g}(\boldsymbol{\theta}) = \mathrm{vec}(\mathbf{BC}) \in \mathbb{R}^{p(h-1)},$$

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

and

$$\mathbf{g}(\boldsymbol{\theta}_0) = \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu}),$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ is in general a basis for $\mathcal{S}_{\boldsymbol{\xi}}$ and $\boldsymbol{\nu} \in \mathbb{R}^{d\times(h-1)}$. With these associations, condition 1 follows immediately from Theorem 1, and, by straightforward algebra, $\boldsymbol{\Delta}_{\boldsymbol{\zeta}} = (\boldsymbol{\nu}^T \otimes \mathbf{I}_p, \mathbf{I}_{h-1} \otimes \boldsymbol{\beta})$, as defined previously in (7). Because $\mathbf{V} > \mathbf{0}$, conditions 2 (including properties p1–p4) and 4 of Proposition A.1 are met. Condition 3 is also met, because $\mathbf{g}(\boldsymbol{\theta})$ is analytic. (See Shapiro 1986 for details about regular points.) Because $\mathbf{V} = \boldsymbol{\Gamma}_{\hat{\boldsymbol{\zeta}}}^{-1}$, based on conclusions 3 and 4 in Proposition A.1, $\mathrm{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}})$ of $H_d(\mathbf{B}, \mathbf{C})$ is asymptotically efficient with

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

which leads to the conclusion 1 of the theorem. Meanwhile, $n\hat{H} \xrightarrow{\mathcal{D}} \chi_k^2$, where the degrees of freedom $k = p(h-1) - \mathrm{rank}(\boldsymbol{\Delta}_{\boldsymbol{\zeta}})$. Because

$$\begin{aligned} \mathrm{rank}(\boldsymbol{\Delta}_{\boldsymbol{\zeta}}) &= \mathrm{rank}([\boldsymbol{\nu}^T \otimes \mathbf{Q}_{\boldsymbol{\beta}}, \mathbf{I}_{h-1} \otimes \boldsymbol{\beta}]) \\ &= d(p-d) + d(h-1) = d(h+p-d-1), \end{aligned}$$

we have $k = (p-d)(h-d-1)$. Thus conclusion 2 is proved. The consistency of $\mathrm{Span}(\hat{\boldsymbol{\beta}})$ in conclusion 3 follows directly from conclusion 1.

APPENDIX D: PROOF OF THEOREM 3

The suboptimal discrepancy function F_d^{sopt} [see (15)] can be written as

$$F_d(\widetilde{\mathbf{B}}, \widetilde{\mathbf{C}}) = \big(\mathrm{vec}(\tilde{\boldsymbol{\xi}}\mathbf{D}_{\mathbf{f}}) - \mathrm{vec}(\widetilde{\mathbf{B}}\widetilde{\mathbf{C}})\big)^T \widetilde{\mathbf{V}}_n \big(\mathrm{vec}(\tilde{\boldsymbol{\xi}}\mathbf{D}_{\mathbf{f}}) - \mathrm{vec}(\widetilde{\mathbf{B}}\widetilde{\mathbf{C}})\big),$$

where $\widetilde{\mathbf{B}} = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\mathbf{B}$, $\widetilde{\mathbf{C}} = \mathbf{C}\mathbf{D}_{\hat{\mathbf{f}}}^{-1}\mathbf{D}_{\mathbf{f}}$, $\widetilde{\mathbf{V}}_n = (\mathbf{D}_{\mathbf{f}}^{-1}\mathbf{D}_{\hat{\mathbf{f}}} \otimes \boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}) \times \mathbf{V}_n(\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{D}_{\mathbf{f}}^{-1} \otimes \hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma})$, and $\tilde{\boldsymbol{\xi}}$ is defined in (17). Because $\widetilde{\mathbf{V}}_n$ converges to $\mathbf{V}$ as $\mathbf{V}_n$ does, it follows from Lemma A.3 that the asymptotic distribution of $n\hat{F}_d$ is the same as that of $n\hat{H}_d$, where

$$H_d(\mathbf{B}, \mathbf{C}) = \big(\mathrm{vec}(\tilde{\boldsymbol{\xi}}\mathbf{D}_{\mathbf{f}}) - \mathrm{vec}(\mathbf{BC})\big)^T \mathbf{V} \big(\mathrm{vec}(\tilde{\boldsymbol{\xi}}\mathbf{D}_{\mathbf{f}}) - \mathrm{vec}(\mathbf{BC})\big).$$

This H_d is a version of Shapiro's discrepancy function H. Noting that the Jacobian matrix is $\boldsymbol{\Delta}_{\boldsymbol{\xi}} = (\mathbf{D}_{\mathbf{f}}\boldsymbol{\gamma}^T \otimes \mathbf{I}_p, \mathbf{I}_h \otimes \boldsymbol{\beta})$, Theorem 3 can be proven in the same way as we proved Theorem 2.

Briefly, it is easy to see that $\mathrm{E}[\tilde{\boldsymbol{\xi}}] = \boldsymbol{\beta}\boldsymbol{\gamma}$ regardless of $\hat{\mathbf{g}}$. Thus

$$\begin{aligned} &\mathrm{cov}(\mathrm{vec}(\tilde{\boldsymbol{\xi}})) \\ &\quad = \mathrm{E}\big[\mathrm{cov}\big(\mathrm{vec}(\tilde{\boldsymbol{\xi}})|\hat{\mathbf{g}}\big)\big] \\ &\quad = \frac{1}{n}\mathrm{E}\big[(\mathbf{D}_{\hat{\mathbf{g}}}^{-1}\mathbf{Q}_{\hat{\mathbf{g}}} \otimes \boldsymbol{\Sigma}^{-1})\,\mathrm{diag}\{\boldsymbol{\Sigma}_{\mathbf{X}|y}\}(\mathbf{Q}_{\hat{\mathbf{g}}}\mathbf{D}_{\hat{\mathbf{g}}}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\big] \\ &\quad = \frac{1}{n}(\mathbf{D}_{\mathbf{g}}^{-1}\mathbf{Q}_{\mathbf{g}} \otimes \boldsymbol{\Sigma}^{-1})\,\mathrm{diag}\{\boldsymbol{\Sigma}_{\mathbf{X}|y}\}(\mathbf{Q}_{\mathbf{g}}\mathbf{D}_{\mathbf{g}}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) + o\left(\frac{1}{n}\right). \end{aligned}$$

Therefore,

$$\sqrt{n}\big(\mathrm{vec}(\tilde{\boldsymbol{\xi}}\mathbf{D}_{\mathbf{f}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}})\big) \xrightarrow{\mathcal{D}} \mathrm{Normal}(\mathbf{0}, \boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}).$$

It now follows from conclusion 1 of Proposition A.1 that the asymptotic distribution of $n\hat{F}_d$ is the same as that of $\|\mathbf{Q}_{\boldsymbol{\Phi}}\mathbf{V}^{1/2}\mathbf{W}\|^2$, where $\mathbf{W}$ is normal with mean 0 and covariance matrix $\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}$ and $\boldsymbol{\Phi} = \mathbf{V}^{1/2}\boldsymbol{\Delta}_{\boldsymbol{\xi}}$ as defined for the statement of Theorem 3. Consequently, $n\hat{F}_d$ is asymptotically distributed as a linear combination of independent chi-squared random variables each with 1 degree of freedom. The coefficients of the chi-squared variables are the eigenvalues of $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$, where $\boldsymbol{\Omega} = \mathbf{V}^{1/2}\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}\mathbf{V}^{1/2}$ is as defined for the statement of the theorem. Finally, consistency follows from conclusion 3 of Proposition A.1 in combination with Lemma A.3.

[Received August 2003. Revised August 2004.]

REFERENCES

Antoniadis, A., Lambert-Lacroix, S., and Leblack, F. (2003), "Effective Dimension Reduction Methods for Tumor Classification Using Gene Expression Data," *Bioinformatics*, 19, 563–570.

Bura, E. (2003), "Using Linear Smoothers to Assess the Structural Dimension of Regressions," *Statistica Sinica*, 13, 143–162.

Bura, E., and Cook, R. D. (2001a), "Estimating Structural Dimensions of Regressions via Parametric Inverse Regressions," *Journal of the Royal Statistical Society*, Ser. B, 63, 393–410.

——— (2001b), "Extending Sliced Inverse Regression: The Weighted Chi-Squared Test," *Journal of the American Statistical Association*, 96, 996–1003.

Bura, E., and Pfeiffer, R. M. (2003), "Graphical Methods for Class Prediction Using Dimension Reduction Techniques on DNA Microarray Data," *Bioinformatics*, 19, 1252–1258.

Chen, C.-H., and Li, K. C. (1998). "Can SIR Be as Popular as Multiple Linear Regression?" *Statistica Sinica*, 8, 289–316.

Chiaromonte, F., Cook, R. D., and Li, B. (2002), "Sufficient Dimension Reduction in Regressions With Categorical Predictors," *The Annals of Statistics*, 30, 475–497.

Cook, R. D. (1994), "On the Interpretation of Regression Plots," *Journal of the American Statistical Association*, 89, 177–189.

——— (1996), "Graphics for Regressions With a Binary Response," *Journal of the American Statistical Association*, 91, 983–992.

——— (1998a), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: Wiley.

——— (1998b), "Principal Hessian Directions Revisited," *Journal of the American Statistical Association*, 93, 84–100.

——— (2004), "Testing Predictor Contributions in Sufficient Dimension Reduction," *The Annals of Statistics*, 32, 1062–1092.

Cook, R. D., and Critchley, F. (2000), "Identifying Regression Outliers and Mixtures Graphically," *Journal of the American Statistical Association*, 95, 781–794.

Cook, R. D., and Li, B. (2002), "Dimension Reduction for Conditional Mean in Regression," *The Annals of Statistics*, 30, 455–474.

Cook, R. D., and Nachtsheim, C. J. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *Journal of the American Statistical Association*, 89, 592–599.

Cook, R. D., and Weisberg, S. (1991), Discussion of "Sliced Inverse Regression for Dimension Reduction," by K. C. Li, *Journal of the American Statistical Association*, 86, 328–332.

——— (1999), *Applied Regression Including Computing and Graphics*, New York: Wiley.

Cook, R. D., and Yin, X. (2001), "Dimension Reduction and Visualization in Discriminant Analysis," *Australian and New Zealand Journal of Statistics*, 43, 147–199.

Donaldson, J. R., and Schnabel, R. B. (1987), "Computational Experience With Confidence Regions and Confidence Intervals in Nonlinear Least Squares," *Technometrics*, 29, 76–82.

Ferguson, T. (1958), "A Method of Generating Best Asymptotically Normal Estimates With Application to the Estimation of Bacterial Densities," *The Annals of Mathematical Statistics*, 29, 1046–1062.

Field, C. (1993), "Tail Areas of Linear Combinations of Chi-Squares and Noncentral Chi-squares," *Journal of Statistical Computation and Simulation*, 45, 243–248.

Fung, W. K., He, X., Li, L., and Shi, P. (2002), "Dimension Reduction Based on Canonical Correlation," *Statistical Sinica*, 12, 1093–1113.

Gather, U., Hilker, T., and Becker, C. (2002), "A Note on Outlier Sensitivity of Sliced Inverse Regression," *Statistics*, 13, 271–281.

Hall, P., and Li, K. C. (1993), "On Almost Linearity of Low-Dimensional Projections From High-Dimensional Data," *The Annals of Statistics*, 21, 867–889.

Hsing, T. (1999), "Nearest-Neighbor Inverse Regression," *The Annals of Statistics*, 27, 697–731.

Hsing, T., and Carroll, R. J. (1992), "An Asymptotic Theory for Sliced Inverse Regression," *The Annals of Statistics*, 20, 1040–1061.

Kiers, H. A. L. (2002), "Setting Up Alternating Least Squares and Iterative Majorization Algorithms for Solving Various Matrix Optimization Problems," *Computational Statistics & Data Analysis*, 41, 157–170.

Li, B., Cook, R. D., and Chiaromonte, F. (2003), "Dimension Reduction for the Conditional Mean in Regressions With Categorical Predictors," *The Annals of Statistics*, 31, 1636–1668.

Li, K.-C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–342.

——— (1992), "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma," *Journal of the American Statistical Association*, 87, 1025–1039.

——— (1997), "Nonlinear Confounding in High-Dimensional Regression," *The Annals of Statistics*, 57, 577–612.

Naik, P., and Tsai, C.-L. (2000), "Partial Least Squares Estimator for Single-Index Models," *Journal of the Royal Statistical Society*, Ser. B, 62, 763–771.

Rao, C. R. (1965), *Linear Statistical Inference and Its Applications*, New York: Wiley.

Ruhe, A., and Wedin, P. A. (1980), "Algorithms for Separable Nonlinear Least Squares Problems," *SIAM Review*, 22, 318–337.

Schott, J. R. (1994), "Determining the Dimensionality in Sliced Inverse Regression," *Journal of the American Statistical Association*, 89, 141–148.

Shapiro, A. (1986), "Asymptotic Theory of Overparameterized Structural Models," *Journal of the American Statistical Association*, 81, 142–149.

Velilla, S. (1998), "Assessing the Number of Linear Components in a General Regression Problem," *Journal of the American Statistical Association*, 93, 1088–1098.

Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002), "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society*, Ser. B, 64, 363–410.

Ye, Z., and Weiss, R. E. (2003), "Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods," *Journal of the American Statistical Association*, 98, 968–979.

Zhu, L.-X., and Fang, K.-T. (1996), "Asymptotics for Kernel Estimate of Sliced Inverse Regression," *The Annals of Statistics*, 24, 1053–1068.

Zhu, L.-X., and Ng, K. W. (1995), "Asymptotics of Sliced Inverse Regression," *Statistica Sinica*, 5, 727–736.

Downloaded by [University of Illinois, Chicago] at 07:20 06 October 2017