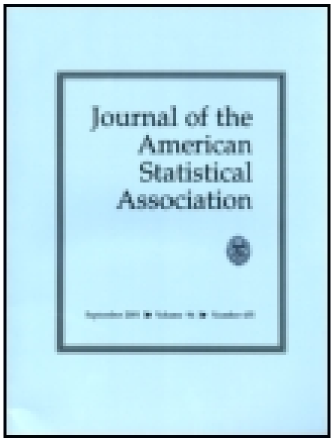

Journal of the American Statistical Association

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Fused Estimators of the Central Subspace in Sufficient Dimension Reduction

R. Dennis Cook & Xin Zhang

Accepted author version posted online: 03 Dec 2013.Published online: 13 Jun 2014.

To cite this article: R. Dennis Cook & Xin Zhang (2014) Fused Estimators of the Central Subspace in Sufficient Dimension Reduction, Journal of the American Statistical Association, 109:506, 815-827, DOI: 10.1080/01621459.2013.866563

To link to this article: http://dx.doi.org/10.1080/01621459.2013.866563

Fused Estimators of the Central Subspace in Sufficient Dimension Reduction

R. Dennis COOK and Xin ZHANG

When studying the regression of a univariate variable Y on a vector $\mathbf{x}$ of predictors, most existing sufficient dimension-reduction (SDR) methods require the construction of *slices* of Y to estimate moments of the conditional distribution of $\mathbf{X}$ given Y. But there is no widely accepted method for choosing the number of slices, while a poorly chosen slicing scheme may produce miserable results. We propose a novel and easily implemented fusing method that can mitigate the problem of choosing a slicing scheme and improve estimation efficiency at the same time. We develop two fused estimators—called FIRE and DIRE—based on an optimal inverse regression estimator. The asymptotic variance of FIRE is no larger than that of the original methods regardless of the choice of slicing scheme, while DIRE is less computational intense and more robust. Simulation studies show that the fused estimators perform effectively the same as or substantially better than the parent methods. Fused estimators based on other methods can be developed in parallel: fused sliced inverse regression (SIR), fused central solution space (CSS)-SIR, and fused likelihood-based method (LAD) are introduced briefly. Simulation studies of the fused CSS-SIR and fused LAD estimators show substantial gain over their parent methods. A real data example is also presented for illustration and comparison. Supplementary materials for this article are available online.

KEY WORDS: Cumulative mean estimation; Inverse regression estimator; Sliced average variance estimation; Sliced inverse regression.

1. INTRODUCTION

It is often useful in regression to concentrate the information about a univariate response Y in a low-dimensional function of the random predictor vector $\mathbf{X} \in \mathbb{R}^p$ without prespecifying a parsimonious parametric model. This may be useful for mitigating the effects of collinearity, facilitating model specification by allowing visualization of the regression in low dimensions (Cook 1998) and providing a relatively small set of "composite" predictors on which to base prediction or interpretation. When visualization is the goal, reducing the dimensionality of $\mathbf{X}$ may be useful when p exceeds 2 or 3 since these bounds represent the limits of our ability to view a dataset in full.

In particular, the goal of sufficient dimension reduction (SDR) is to replace $\mathbf{X}$ by a lower dimensional function $\boldsymbol{R}(\mathbf{X}) \in \mathbb{R}^q$, $q < p$, without loss of information on the regression. Usually, $\boldsymbol{R}(\mathbf{X})$ is constrained to be a linear function $\boldsymbol{\eta}^T\mathbf{X}$, where $\boldsymbol{\eta} \in \mathbb{R}^{p\times q}$, $q < p$. The space spanned by the columns of $\boldsymbol{\eta}$ is then called a dimension-reduction subspace for the regression of Y on $\mathbf{X}$. In most SDR contexts, we are interested in finding a population meta-parameter, the central subspaces (CSs), represented by $\mathcal{S}_{Y|\mathbf{X}}$ and define as the intersection of all dimension-reduction subspaces when this intersection is also a dimension-reduction subspace. It is well known that the CS exists under mild conditions (Cook 1998). Let $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$. Any basis $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$, $d < p$, of the CS satisfies three equivalent properties, (a) $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}$, (b) $\mathbf{X}|(Y, \boldsymbol{\beta}^T\mathbf{X}) \sim \mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}$, and (c) $Y|\mathbf{X} \sim Y|\boldsymbol{\beta}^T\mathbf{X}$. This implies that we can replace $\mathbf{X}$ by a lower dimensional projection $\boldsymbol{\beta}^T\mathbf{X}$ without loss of information on the regression, and that the CS is the smallest of the dimension-reduction subspaces.

Sliced inverse regression (SIR; Li 1991) and sliced average variance estimation (SAVE; Cook and Weisberg 1991) were the first methods proposed for SDR. For regressions with both quantitative and categorical predictors, Chiaromonte, Cook, and Li (2002) extended SIR to partial SIR for estimating a partial CS that reduces only the quantitative predictors. Ye and Weiss (2003) proposed to combine SIR and SAVE by using a bootstrap method to select the coefficients in a convex combination method. Li and Wang (2007) proposed directional regression (DR) as a dimension-reduction method derived from empirical directions. DR synthesizes SIR and SAVE but requires substantially less computation than the bootstrap method proposed by Ye and Weiss (2003). Cook and Ni (2005) found the optimal method in the inverse regression (IR) family in terms of asymptotic efficiency. Cook and Forzani (2009) developed a likelihood-based method (LAD), which inherits properties and methods from general likelihood theory. LAD is based on normality assumptions, but is $\sqrt{n}$ consistent and able to perform better than classical methods with deviations from normality.

Most of these and other dimension-z-reduction methods, including cluster-based methods (Setodji and Cook 2004) and methods that aim to estimate the central solution space (CSS; Li and Dong 2009; Dong and Li 2010), hinge on slicing a quantitative response, that is, on replacing Y with a discrete version $\tilde{Y}$ to approximate the conditional moments of $\mathbf{X}|Y$ from $\mathbf{X}|\tilde{Y}$. This in effect approximates the moments of $\mathbf{X}|Y$ with step functions. The effectiveness of a slicing scheme depends on the number of slices and on their arrangement, which in turn can depend on the sample size n, the dimension d of the CS, the joint distribution of the data, and other characteristics of the regression. How to choose a good slicing scheme is a theoretically difficult but

R. Dennis Cook is Professor (E-mail: *dennis@stat.umn.edu*), and Xin Zhang is Ph.D. student (E-mail: *zhan0648@umn.edu*), School of Statistics, University of Minnesota, Minneapolis, MN 55455. The authors are grateful to Yuexiao Dong and Bing Li for providing them the codes for computing the CSS estimators, and grateful to Liping Zhu, Lixing Zhu, and Zhenghui Feng for the cumulative slicing estimation codes. The authors also thank the editor, the associate editor, and two referees for those helpful comments that led to significant improvements in this article. Research for this article was supported in part by grant DMS-1007547 from the U.S. National Science Foundation.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

© 2014 American Statistical Association
Journal of the American Statistical Association
June 2014, Vol. 109, No. 506, Theory and Methods
DOI: 10.1080/01621459.2013.866563

crucial issue that has resisted practically useful solutions over the past 20 years and that is often a severe nag in applications. In this article, we propose a practically useful solution to the slicing dilemma.

The problem of choosing the number of slices was first noticed when Li (1991) established SIR. Nevertheless, it was not paid much attention, perhaps since Li claimed in the same article that even if the slice number is $n/2$, so that each slice contains only two observations, the resulting estimate will still be $\sqrt{n}$ consistent. Hsing and Carroll (1992) derived asymptotic properties for the special case where each slice has only two observations. This was generalized by Zhu and Ng (1995) to the case that each slice has an arbitrary but fixed number of observations, and to the case that the number of observations within each slice goes to infinity with increasing sample size. They derived the asymptotic variance of SIR's kernel matrix with a fixed number c of observation per slice and observed that if $c = O(n^{\alpha})$, $\alpha \in (0, \frac{1}{2})$, then the asymptotic variance could be reduced to a minimum. The choice of α depends on the distribution of $\mathbf{X}$ and on the smoothness of $\boldsymbol{m}(y) = \mathrm{E}(\mathbf{X}|Y = y)$. This can be interpreted as requiring that the number of observations per slice be large enough to yield efficient estimates, but still be relatively small when comparing with n. This result from Zhu and Ng (1995) suggests that slicing schemes with too many slices having too few observations per slice should be avoided. Nevertheless, it is still difficult to turn such qualitative conclusions into practically useful rules.

In practice, slicing Y according to its quantiles, so we have approximately the same number of observations in each slice, may produce satisfactory results. We will refer to this as *quantile slicing*. However, the performance of quantile slicing still depends on the number of slices especially for higher-order methods (Li and Zhu 2007). Such issues make choosing a good slicing scheme crucial, but there is no widely accepted rule for choosing the number of slices.

On the other hand, fusing information from different slicing schemes could effectively circumvent the pursuit of a best slicing scheme. There have been many studies on forecast combinations since Bates and Granger (1969), see Timmermann (2006) for an overview. However, forecast combination methods cannot be applied directly in SDR without a clear predictive goal. Recently, Zhu, Zhu, and Feng (2010) developed cumulative mean estimation (CUME), which uses a weighted average of SIR kernel matrices from all possible slicing schemes with two slices. CUME will be used for comparisons to our proposed method in simulation studies.

In this article, we propose a method of fusing quantile slicing schemes, which largely avoids the long-standing problem of selecting the number of slices and can result in substantially improved estimators of the CS. Our formal development is based on fusing through minimum discrepancy functions similar to the asymptotically optimal inverse regression estimator (IRE; Cook and Ni 2005), although the methodology can be applied straightforwardly to most SDR methods that rely on slicing a quantitative response, including SIR-, SAVE-, DR-, and CSS-based methods. Our fusing method not only recovers intraslice information and therefore mitigates the complexity of choosing a good slicing scheme, but can also substantially improve estimation efficiency over methods with single slicing schemes.

The rest of this article is organized as follows. In Section 2, the IR methodology is reviewed, and two fusing methods corresponding to different inner product matrices are introduced. Theoretical properties of our fusing methods are given in Section 3, where we compare the asymptotic variances of our fused estimators to estimators based on a single quantile slicing scheme. In Section 4, we discuss computing and the use of robust inner product matrices. Section 5 includes simulation studies and Section 6 gives a brief discussion. Proofs and additional simulations are contained in the supplement to this article.

Throughout this article, we will use bold symbols for matrices and vectors, and $\boldsymbol{\beta}$ for a semiorthogonal basis of the CS, so $\mathrm{span}(\boldsymbol{\beta}) = \mathcal{S}_{Y|\mathbf{X}}$. Also we will use $\boldsymbol{\Sigma}$ for the covariance matrix of the predictor vector $\mathbf{X}$, and $\boldsymbol{\Gamma}$ for covariance matrices in fusing methods. We use quantile slicing for all methods, unless otherwise specified.

2. METHODOLOGY AND ESTIMATION

Most of the existing dimension-reduction methods rely on an assumption about the marginal distribution of $\mathbf{X}$, called the linearity condition. The linearity condition requires that $\mathrm{E}(\mathbf{X}|\boldsymbol{\beta}^T\mathbf{X})$ is a linear function of $\boldsymbol{\beta}^T\mathbf{X}$. This condition holds for all elliptically distributed $\mathbf{X}$ (Eaton 1986) and it is also asymptotically true when the number of predictors $p \to \infty$ with $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ fixed (Hall and Li 1993). When it holds, $\boldsymbol{\Sigma}^{-1}\{\mathrm{E}(\mathbf{X}|Y = y) - \mathrm{E}(\mathbf{X})\} \in \mathcal{S}_{Y|\mathbf{X}}$ for all y. Thus, by varying y in the domain of Y and estimating $\boldsymbol{\Sigma}^{-1}\{\mathrm{E}(\mathbf{X}|Y = y) - \mathrm{E}(\mathbf{X})\}$, we can obtain directions in the CS. When Y is categorical, it is straightforward to construct a sample version of $\mathrm{E}(\mathbf{X}|Y = y)$ by averaging $\mathbf{X}$ within each class of Y. When Y is continuous, most SDR methods rely on nonparametric estimation of $\mathrm{E}(\mathbf{X}|Y = y)$. Such nonparametric estimation requires partitioning the range of Y into h slices and then constructing a discrete version $\tilde{Y}$ of Y to obtain directions in the CS. It is always true that $\mathcal{S}_{\tilde{Y}|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. The accuracy of estimating the CS via IR thus depends on slicing the response Y, as mentioned in the Introduction. At this point, given a slicing scheme, the IRE is an optimal member of the IR family, and it has at least three desirable properties: its estimated basis is asymptotically efficient given the slicing scheme; it has an asymptotic chi-squared statistic for testing hypotheses about d; and it provides for testing conditional independence hypotheses that the response is independent of a selected subset of predictors given the remaining predictors. Motivated by this, we introduce in this section a fusing method based on IRE objective functions following a review of the IRE method.

2.1 Review of IRE

Consider a univariate response variable Y, a $p \times 1$ vector of predictors $\mathbf{X}$ and the standardized predictor $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$, where $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X}) > 0$. Then $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}$ (Cook 1998, Proposition 6.1). Further assume that Y has a finite support $\{1, 2, \ldots, h\}$ after slicing, and then define

$$\boldsymbol{\xi}_y = \boldsymbol{\Sigma}^{-1}(\mathrm{E}(\mathbf{X}|Y = y) - \mathrm{E}(\mathbf{X})), \text{ and } \mathcal{S}_{\boldsymbol{\xi}} = \sum_{y=1}^{h} \mathrm{span}(\boldsymbol{\xi}_y).$$

The linearity condition implies $\mathcal{S}_{\boldsymbol{\xi}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. We further require the coverage condition to enforce the equality of the two subspaces $\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{S}_{Y|\mathbf{X}}$. The coverage condition requires that the target subspace $\mathcal{S}_{\boldsymbol{\xi}}$ has the same dimension as $\mathcal{S}_{Y|\mathbf{X}}$. See Cook and Ni (2005) and Li and Wang (2007) for discussion on the coverage condition. To focus our discussion on fusing, we assume that the dimension of the CS is known: $d = \dim(\mathcal{S}_{Y|\mathbf{X}}) = \dim(\mathcal{S}_{\boldsymbol{\xi}})$. By definition, for each y, there exists a vector $\boldsymbol{\gamma}_y$ such that $\boldsymbol{\xi}_y = \boldsymbol{\beta}\boldsymbol{\gamma}_y$. Define

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

and let $\mathbf{f} = (f_1, \ldots, f_h)^T$ with $f_y = \Pr(Y = y)$. It is easy to see that $\boldsymbol{\xi}\mathbf{f} = \boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{f} = 0$, called the intrinsic location constraint by Cook and Ni (2005).

After gathering a simple random sample $(\mathbf{X}_i, Y_i)$, $i = 1, 2, \ldots, n$, from $(\mathbf{X}, Y)$, we let $\mathbf{X}_{yj}$ denote the jth observation on $\mathbf{X}$ in slice y, $y = 1, \ldots, h$, $j = 1, \ldots, n_y$, and $\sum_{y=1}^{h} n_y = n$. Further, let $\bar{\mathbf{X}}$ be the overall average of $\mathbf{X}$, and let $\bar{\mathbf{X}}_y$ be the average of the n_y points with $Y = y$. Let $\hat{\mathbf{f}} = (n_1/n, \ldots, n_h/n)^T$, and let $\hat{\boldsymbol{\Sigma}} > 0$ be the usual sample covariance matrix for $\mathbf{X}$. The sample version of $\boldsymbol{\xi}$ is

$$\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\xi}}_1, \ \ldots, \ \hat{\boldsymbol{\xi}}_h) \in \mathbb{R}^{p\times h}, \ \hat{\boldsymbol{\xi}}_y = \hat{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}), y = 1, \ldots, h.$$

Let $\mathbf{D}_{\mathbf{f}}$ denote a diagonal matrix with the elements of the vector $\mathbf{f}$ on the diagonal. Because of the intrinsic location constraint, $\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{1}_h = \hat{\boldsymbol{\xi}}\hat{\mathbf{f}} = 0$. Construct $\mathbf{A} \in \mathbb{R}^{h\times(h-1)}$, such that $\mathbf{A}^T\mathbf{A} = \mathbf{I}_{h-1}$ and $\mathbf{A}^T\mathbf{1}_h = 0$, then use the reduced matrix $\boldsymbol{\zeta} \equiv \boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}}\mathbf{A} = \boldsymbol{\beta}\boldsymbol{\nu}$ and its sample version $\hat{\boldsymbol{\zeta}} = \hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{A}$ in the construction of discrepancy functions without loss of generality.

With these characters established, the quadratic discrepancy function of IRE is defined as

$$\mathrm{F}_d(\mathbf{B}, \mathbf{C}) = (\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC}))^T \hat{\boldsymbol{\Gamma}}_{\hat{\zeta}}^{-1}(\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC})), \tag{2.1}$$

where the columns of $\mathbf{B} \in \mathbb{R}^{p\times d}$ represent an orthogonal basis for $\mathrm{span}(\boldsymbol{\zeta})$, which equals $\mathcal{S}_{Y|\mathbf{X}}$ under the linearity and coverage conditions, but may be a proper subset of $\mathcal{S}_{Y|\mathbf{X}}$ with linearity but not coverage; $\mathbf{C} \in \mathbb{R}^{d\times(h-1)}$ is used only in fitting to represent the coordinates of $\boldsymbol{\zeta}$ relative to $\mathbf{B}$, and $\hat{\boldsymbol{\Gamma}}_{\hat{\zeta}} \in \mathbb{R}^{p(h-1)\times p(h-1)}$ is a consistent estimator of the asymptotic variance of $\mathrm{vec}(\hat{\boldsymbol{\zeta}})$. Replacing $\hat{\boldsymbol{\Gamma}}_{\hat{\zeta}}^{-1}$ in (2.1) with an arbitrary nonsingular matrix, we still obtain a $\sqrt{n}$ consistent estimator of $\mathcal{S}_{Y|\mathbf{X}} = \mathrm{span}(\boldsymbol{\zeta})$, but using $\hat{\boldsymbol{\Gamma}}_{\hat{\zeta}}^{-1}$ guarantees an asymptotically efficient estimator. For instance, Cook and Ni (2005) showed that SIR is a member of a suboptimal class using an inner product matrix different from $\hat{\boldsymbol{\Gamma}}_{\hat{\zeta}}^{-1}$. Minimizing $\mathrm{F}_d(\mathbf{B}, \mathbf{C})$ produces the IRE.

A distinct advantage of IRE is its ability to properly weight the slice means when the intraslice variance $\mathrm{var}(\mathbf{X}|Y = j)$ changes appreciably with the slice. In contrast, SIR and other methods like CUME that neglect the intraslice variances can produce poor results when there are substantial differences in variability across the slices.

However, the performance of IRE is still based on the tuning parameter h, the number of slices, and on the method for constructing the slices themselves. Instead of pursuing the "best" h, we propose a fusing approach in the following section where a set of possible choices of slice numbers $H = \{h_1, \ldots, h_K\}$ is used rather than a particular value h.

2.2 Fused Inverse Regression Estimators

With a set of $K \geq 2$ quantile slicing schemes with slice numbers $H = \{h_1, \ldots, h_K\}$, we have K different data matrices, denoted by

$$\hat{\boldsymbol{\xi}}^{(j)} \in \mathbb{R}^{p\times h_j}, \ \hat{\boldsymbol{\zeta}}^{(j)} \equiv \hat{\boldsymbol{\xi}}^{(j)}\mathbf{D}_j\mathbf{A}_j \in \mathbb{R}^{p\times(h_j-1)}, \ j = 1, \ldots, K,$$

where $\hat{\mathbf{D}}_j \equiv \mathrm{diag}(n_1^{(j)}/n, \ldots, n_{h_j}^{(j)}/n)$ is a diagonal matrix consisting of the sample proportions of slices in the jth slicing scheme. Constant matrices $\mathbf{A}_j \in \mathbb{R}^{h_j\times(h_j-1)}$, $j = 1, \ldots, K$, are associated with the intrinsic location constraint in each slicing scheme, so that $\mathbf{A}_j^T\mathbf{A}_j = \mathbf{I}_{h_j-1}$ and $\mathbf{A}_j^T\mathbf{1}_{h_j} = 0$. Here, the data matrices $\hat{\boldsymbol{\zeta}}^{(j)}$ converge in probability at rate $\sqrt{n}$ to

$$\boldsymbol{\zeta}^{(j)} \equiv \boldsymbol{\beta}\boldsymbol{\gamma}_j\mathbf{D}_j\mathbf{A}_j = \boldsymbol{\beta}\boldsymbol{\nu}_j \in \mathbb{R}^{p\times(h_j-1)}, \ j = 1, \ldots, K.$$

In this equation, $\hat{\mathbf{D}}_j$ converges to $\mathbf{D}_j$, and $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ is a basis for $\mathcal{S}_{Y|\mathbf{X}}$ such that the matrices $\boldsymbol{\zeta}^{(j)}$, $j = 1, 2, \ldots, K$, can be written as the products of $\boldsymbol{\beta}$ and the coordinate matrices $\boldsymbol{\nu}_j \in \mathbb{R}^{d\times(h_j-1)}$. The coordinate matrices $\boldsymbol{\nu}_j$ will share the same row dimension $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$, but different column dimension due to varied slice numbers.

To fuse information from various slicing schemes, the most fundamental step is to fuse the K data matrices by row, since they have the same number of rows and each row corresponds to one dimension in the CS. Therefore, the fused data matrix is $\hat{\boldsymbol{\zeta}}_F = (\hat{\boldsymbol{\zeta}}^{(1)}, \ldots, \hat{\boldsymbol{\zeta}}^{(K)}) \in \mathbb{R}^{d\times(\sum_{j=1}^{K} h_j - K)}$, which converges in probability to $\boldsymbol{\zeta}_F = (\boldsymbol{\zeta}^{(1)}, \ldots, \boldsymbol{\zeta}^{(K)}) = \boldsymbol{\beta}(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K)$. The fused IRE follows from minimizing a quadratic discrepancy function with the fused data matrix:

$$\begin{aligned}\mathcal{F}_d(\mathbf{B}, \mathbf{C}; \boldsymbol{\Psi}) = {} & (\mathrm{vec}(\hat{\boldsymbol{\zeta}}_F) - \mathrm{vec}(\mathbf{BC}))^T \\ & \times \hat{\boldsymbol{\Psi}}(\mathrm{vec}(\hat{\boldsymbol{\zeta}}_F) - \mathrm{vec}(\mathbf{BC})),\end{aligned} \tag{2.2}$$

where the columns of $\mathbf{B} \in \mathbb{R}^{p\times d}$ still represent an orthogonal basis of our estimate of the CS, $\mathbf{C} \in \mathbb{R}^{d\times\sum_j(h_j-1)} = \mathbb{R}^{d\times(\sum_j h_j-K)}$ is used only in fitting and it represents the coordinates of $\boldsymbol{\zeta}_F$ relative to $\mathbf{B}$. Moreover, $\hat{\boldsymbol{\Psi}}$ is a consistent estimator of a positive semidefinite matrix $\boldsymbol{\Psi} \in \mathbb{R}^{p(\sum_j h_j-K)\times p(\sum_j h_j-K)}$ and is essentially the inner product matrix of this weighted least-square route. The estimate of the CS is then constructed by minimizing the quadratic discrepancy function in (2.2), over $\mathbf{B}$ and $\mathbf{C}$. Since $\mathcal{F}_d(\mathbf{B}, \mathbf{C}; \boldsymbol{\Psi})$ is an overparameterized discrepancy function, the minimizer $(\hat{\mathbf{B}}, \hat{\mathbf{C}})$ is not unique: for any orthogonal matrix $\boldsymbol{\Lambda}$ that $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T = \mathbf{I}_d$, $(\hat{\mathbf{B}}\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^T\hat{\mathbf{C}})$ is also a minimizer. However, the product $\hat{\mathbf{B}}\hat{\mathbf{C}}$ is unique, as well as the CS estimator $\mathrm{span}(\hat{\mathbf{B}})$ and its projection $\hat{\mathbf{B}}\hat{\mathbf{B}}^T$.

3. ASYMPTOTIC PROPERTIES

We study the asymptotic properties of the fused IREs in this section. We rewrite the quadratic discrepancy function of IRE with slicing scheme j as

$$\begin{aligned}\mathrm{F}_d^{(j)}(\mathbf{B}, \mathbf{C}) = {} & \left(\mathrm{vec}\left(\hat{\boldsymbol{\zeta}}^{(j)}\right) - \mathrm{vec}(\mathbf{BC})\right)^T \\ & \times \hat{\boldsymbol{\Gamma}}_j^{-1}\left(\mathrm{vec}\left(\hat{\boldsymbol{\zeta}}^{(j)}\right) - \mathrm{vec}(\mathbf{BC})\right), j = 1, 2,\end{aligned} \tag{3.1}$$

which is just (2.1) with specific slicing schemes, and where $\hat{\boldsymbol{\Gamma}}_j \equiv \hat{\boldsymbol{\Gamma}}_{\hat{\xi}_j}$ is a consistent estimator of the nonsingular positive definite matrix $\boldsymbol{\Gamma}_j \equiv \boldsymbol{\Gamma}_{\hat{\xi}_j}$. Without loss of generality, we will discuss fusing only two schemes, that is, $K = 2$ and $H = \{h_1, h_2\}$, along with comparisons to IRE with $h = h_1$. The results hold for $K > 2$ automatically.

3.1 Asymptotic Normality

Prior to deriving the asymptotic properties of the fused estimators, we need to find the asymptotic distribution of the fused data matrix $\hat{\boldsymbol{\xi}}_F$. As stated in Theorem 1 from Cook and Ni (2005), $\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\zeta}}^{(j)}) - \mathrm{vec}(\boldsymbol{\zeta}^{(j)}))$, $j = 1, 2$, converges in distribution to a normal with mean 0 and nonsingular covariance matrix $\boldsymbol{\Gamma}_j$. For each slicing scheme, define h_j random variables $J_y^{(j)}$ such that $J_y^{(j)}$ equals 1 if an observation is in the yth slice of jth slicing scheme and 0 otherwise, $j = 1, 2$, $y = 1, \ldots, h_j$. Then $\mathrm{E}(J_y^{(j)}) = (\mathbf{D}_j)_{yy} = f_y^{(j)}$. Also define the random vector $\boldsymbol{\epsilon}^{(j)} = (\epsilon_1^{(j)}, \ldots, \epsilon_{h_j}^{(j)})^T$ with elements

$$\epsilon_y^{(j)} = J_y^{(j)} - \mathrm{E}\big(J_y^{(j)}\big) - \mathbf{Z}^T\mathrm{E}\big(\mathbf{Z}J_y^{(j)}\big)$$

that are the population residuals from the ordinary least-square fit of $J_y^{(j)}$ on $\mathbf{Z}$. Denote the fused population residual vector by

$$\mathbf{E} \equiv \big(\boldsymbol{\epsilon}^{(1)^T}, \boldsymbol{\epsilon}^{(2)^T}\big)^T = \big(\epsilon_1^{(1)}, \ldots, \epsilon_{h_1}^{(1)}, \epsilon_1^{(2)}, \ldots, \epsilon_{h_2}^{(2)}\big)^T \in \mathbb{R}^{(h_1+h_2)\times 1}. \tag{3.2}$$

The following theorem gives the asymptotic distribution of the data matrix $\hat{\boldsymbol{\xi}}_F$ that we have used in the fused estimators.

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

$$\sqrt{n}\big(\mathrm{vec}\big(\hat{\boldsymbol{\xi}}^{(1)}\hat{\mathbf{D}}_1, \hat{\boldsymbol{\xi}}^{(2)}\hat{\mathbf{D}}_2\big) - \mathrm{vec}\big(\boldsymbol{\xi}^{(1)}\mathbf{D}_1, \boldsymbol{\xi}^{(2)}\mathbf{D}_2\big)\big) \longrightarrow N(0, \boldsymbol{\Gamma}_H),$$

where $\boldsymbol{\Gamma}_H = \mathrm{cov}(\mathrm{vec}(\boldsymbol{\Sigma}^{-1/2}\mathbf{Z}\mathbf{E}^T))$ and $\mathbf{E}$ is given in (3.2). Consequently,

$$\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\xi}}_F) - \mathrm{vec}(\boldsymbol{\zeta}_F)) \longrightarrow N(0, \boldsymbol{\Gamma}_F),$$

where $\boldsymbol{\Gamma}_F = (\mathbf{A}_F^T \otimes \mathbf{I}_p)\boldsymbol{\Gamma}_H(\mathbf{A}_F \otimes \mathbf{I}_p)$, and $\mathbf{A}_F = \mathrm{diag}(\mathbf{A}_1, \mathbf{A}_2)$ is a nonstochastic block diagonal matrix.

The inverse of $\boldsymbol{\Gamma}_F$ will be used later, but $\boldsymbol{\Gamma}_F$ is possibly singular even though all $\boldsymbol{\Gamma}_j$, $j = 1, \ldots, K$, are assumed to be nonsingular. The singularity of $\boldsymbol{\Gamma}_F$ occurs when one of the K data matrices $\hat{\boldsymbol{\xi}}^{(j)}$ can be expressed as a linear combination of the other data matrices. For example, suppose we have a 5-slice scheme and a 10-slice scheme obtained by splitting each of the five slices. Then every element of the data matrix in the 5-slice scheme will be a linear combination of the data matrix in the 10-slice scheme. Fusing these two will lead to a singular $\boldsymbol{\Gamma}_F$ because of the five redundant slices. Singularity of $\boldsymbol{\Gamma}_F$ can be avoided by using quantile slicing schemes that do not share the same slice boundaries. We can fuse a 5-slice scheme and a 11-slice scheme without encounter singular $\boldsymbol{\Gamma}_F$. Unless indicated otherwise, we will assume that the numbers of slices $H = \{h_1, \ldots, h_K\}$ are properly chosen such that $\boldsymbol{\Gamma}_F$ is nonsingular. See Section 5 for more discussion on choosing the set of numbers of slices $H = \{h_1, \ldots, h_K\}$.

3.2 Consistency of the Fused Estimators

With the asymptotic distribution of the data matrix $\hat{\boldsymbol{\xi}}_F$, we can establish the $\sqrt{n}$ consistency of the fused estimator from (2.2). Let $(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\nu}}_j)$ be minimizer of the IRE discrepancy function with $h = h_j$, that is, (3.1); let $(\hat{\boldsymbol{\beta}}_F, \hat{\boldsymbol{\omega}}_F) \equiv (\hat{\boldsymbol{\beta}}_F, \hat{\boldsymbol{\omega}}_1, \hat{\boldsymbol{\omega}}_2)$ be a minimizer of the fused discrepancy function (2.2). As discussed in Section 2, the minimizers $(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\nu}}_j)$, $j = 1, 2$, and $(\hat{\boldsymbol{\beta}}_F, \hat{\boldsymbol{\omega}}_F)$ are not unique. But we have uniqueness in the products: $\hat{\boldsymbol{\beta}}_j\hat{\boldsymbol{\nu}}_j$ and $\hat{\boldsymbol{\beta}}_F\hat{\boldsymbol{\omega}}_F$. Moreover, the two products: $\hat{\boldsymbol{\beta}}_j\hat{\boldsymbol{\nu}}_j$ and $\hat{\boldsymbol{\beta}}_F\hat{\boldsymbol{\omega}}_j$ both converge to the matrix $\boldsymbol{\zeta}^{(j)}$ for $j = 1, 2$. Therefore, we can judge our fused estimators by comparing asymptotic behaviors of $\hat{\boldsymbol{\beta}}_F\hat{\boldsymbol{\omega}}_j$ to $\hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\nu}}_1$. The result is summarized in the following theorem.

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 and assume that $\hat{\boldsymbol{\Psi}}$ converges in probability to $\boldsymbol{\Psi} > 0$ as n goes to infinity. Then $\hat{\boldsymbol{\beta}}_F\hat{\boldsymbol{\omega}}_F = (\hat{\boldsymbol{\beta}}_F\hat{\boldsymbol{\omega}}_1, \hat{\boldsymbol{\beta}}_F\hat{\boldsymbol{\omega}}_2)$ and $(\hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\nu}}_1, \hat{\boldsymbol{\beta}}_2\hat{\boldsymbol{\nu}}_2)$ are both $\sqrt{n}$-consistent estimators for $\boldsymbol{\zeta}_F = (\boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)})$ such that for $j = 1, 2$,

$$\sqrt{n}\big(\mathrm{vec}(\hat{\boldsymbol{\beta}}_j\hat{\boldsymbol{\nu}}_j) - \mathrm{vec}\big(\boldsymbol{\zeta}^{(j)}\big)\big) \longrightarrow N(0, \mathbf{V}_j),$$
$$\sqrt{n}\big(\mathrm{vec}(\hat{\boldsymbol{\beta}}_F\hat{\boldsymbol{\omega}}_j) - \mathrm{vec}\big(\boldsymbol{\zeta}^{(j)}\big)\big) \longrightarrow N(0, \mathbf{W}_j),$$

where the explicit expressions for $\mathbf{V}_j$ and $\mathbf{W}_j$ are given in the Appendix equation (A.2). In particular, if $\boldsymbol{\Psi} = \boldsymbol{\Gamma}_F^{-1}$, then the fused estimator is asymptotically efficient with respect to the fused data $\hat{\boldsymbol{\xi}}_F$ and the asymptotic variances satisfy

$$\mathbf{W}_j \leq \mathbf{V}_j, \quad j = 1, 2. \tag{3.3}$$

All the results of Theorem 2 extend straightforwardly to slicing schemes with $K \geq 2$. The estimators of the CS constructed from minimizing $\mathcal{F}(\mathbf{B}, \mathbf{C}; \boldsymbol{\Gamma}_F^{-1})$ are called FIRE.

A direct consequence of (3.3) is that the asymptotic variances of $\mathrm{vec}(\hat{\boldsymbol{\beta}}_F\hat{\boldsymbol{\beta}}_F^T)$ for FIRE will be no larger than the asymptotic variances of $\mathrm{vec}(\hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_1^T)$ or $\mathrm{vec}(\hat{\boldsymbol{\beta}}_2\hat{\boldsymbol{\beta}}_2^T)$. This shows that FIRE is asymptotically at least as efficient at estimating the CS as IRE with the corresponding slicing schemes. Another straightforward generalization of this theorem is obtained by considering $H = \{h_1, \ldots, h_K\}$ and $H_1 \subset H$, a proper subset of H. The asymptotic variances of FIRE based on H are smaller than or equal to the asymptotic variances of FIRE based on H_1. Notice that IRE is asymptotically efficient for only a single data matrix $\hat{\boldsymbol{\zeta}}^{(j)}$, but not necessarily asymptotically efficient with respect to the fused data matrix $\hat{\boldsymbol{\xi}}_F$. On the other hand, Theorem 2 states FIRE is asymptotically efficient with respect to $\hat{\boldsymbol{\xi}}_F$, because this method is essentially fusing a set of quantile slicing schemes H to a single fused slicing scheme.

Although the boundaries of a fused slicing scheme are at quantiles, these quantiles are not arranged so each slice has the same number of observations and thus fused slicing is not the same as a single quantile slicing scheme. When the number of observations in each slice goes to infinity, this fused slicing scheme has asymptotic variance no larger than that for IRE.

An example may help fix ideas. Assume we have 120 iid samples with distinct Y_i's and $H = \{4, 5\}$. Let y_p denote the pth quantile of Y, that is, $y_p = \inf\{y : \sum_{i=1}^{n} I(Y_i \leq y)/n \leq p\}$. The quantile slicing scheme for the IRE estimator with $h = 4$

(similarly for $h = 5$) is obtained from $\{y_{0.25},\ y_{0.5},\ y_{0.75}\}$ with four slices consist of 30 observations each. Therefore, the fused data matrix will consist of four slices with 30 observations each and 5 slices with 24 observations each. FIRE uses the full inner product matrix. A full rank linear transformation of the fused data matrix $\hat{\boldsymbol{\xi}}_F$ will result in a corresponding linear transformation of the inner product matrix. Since there is a linear transformation between the fused data matrix and a single data matrix obtained from eight slices with slice boundaries $\{y_{0.2},\ y_{0.25},\ y_{0.4},\ y_{0.5},\ y_{0.6},\ y_{0.75},\ y_{0.8}\}$, FIRE is exactly the same as IRE with this (nonquantile) fused slicing scheme having eight slices with 24, 6, 18, 12, 12, 18, 6, and 24 observations per slice. However, IRE with this fused slicing scheme is different from IRE with a single quantile slicing scheme having $h = 8$ with 15 observations per slice. In our simulation studies (Section 5), we see clear advantages of fused slicing schemes over quantile slicing schemes with the same numbers of slices.

The improvement from fusing over simply increasing the number of quantile slices is remarkable. Slicing on Y and then using the discrete approximation of $\mathrm{E}(\mathbf{X}|Y)$ in the objective function is essentially an approximation to some integral with respect to the probability density function of Y. Results from Luceno (1999) imply that a properly chosen slicing scheme with h slices could guarantee that the discrete version $\tilde{Y}$ has the same rth moments as Y for $r = 0, \ldots, 2h - 1$. Since high-order moments are not important, composite rules with small numbers of grid points in numerical integration are often more accurate than directly applying the integration formula with a large number of grid points. Fusing estimators are analogous to composite rules in numerical integration. Fusing several crude slicing schemes with small values of h often leads to more accurate basis estimation than IRE with a large h. We will demonstrate this result in Section 5.

3.3 Inner Product Matrix $\boldsymbol{\Psi}$

Our estimation based on minimizing the quadratic discrepancy function defined in (2.2) is largely related to the choice of $\boldsymbol{\Psi}$. FIRE uses the full inverse of the asymptotic covariance matrix of the vectorized fused data matrix $\mathrm{vec}(\hat{\boldsymbol{\xi}}_F) = \mathrm{vec}(\hat{\boldsymbol{\xi}}^{(1)}, \hat{\boldsymbol{\xi}}^{(2)})$, which makes itself asymptotically efficient. However, estimation of such inner product matrix is difficult because of its dimensionality. Because a poor estimate of the inner product matrix could lead to poor performance, we propose simpler alternatives to the full inner product matrix $\boldsymbol{\Psi} = \boldsymbol{\Gamma}_F^{-1}$ in this section.

The first natural simplification is by taking $\boldsymbol{\Psi} = \mathrm{diag}\{\boldsymbol{\Gamma}_1^{-1}, \boldsymbol{\Gamma}_2^{-1}\} \equiv \boldsymbol{\Gamma}_D^{-1}$. The fused estimators constructed from minimizing $\mathcal{F}(\mathbf{B}, \mathbf{C}; \boldsymbol{\Gamma}_D^{-1})$ are called DIRE. The block diagonal inner product matrix is obtained by pretending that the slicing schemes to be fused are independent, which reduces the number of parameters in the inner product matrix and may improve estimation accuracy. Additionally, using a block diagonal inner product matrix corresponds to simply adding the objective functions for the individual slicing schemes to obtain the fused objective function, which is a procedure that can be used straightforwardly to fuse quantile slicing schemes for any method based on an objective function. Moreover, we demonstrate in Section 6 that this block diagonal fusing approach is actually a composite likelihood approach (Lindsay 1988) in likelihood-based dimension reduction (Cook and Forzani 2009).

For small-sample problems, Ni and Cook (2007) introduced a robust version of the IRE inner product matrix, which requires only second-order moments rather than fourth-order moments. They showed that the robust IRE produces a $\sqrt{n}$ consistent basis estimator of the CS and was quite robust in small-sample simulations. Therefore, we will use this robust version of the inner product matrices in our fused methods for small-sample problems. Replacing the previous residual variables $\epsilon_y^{(j)} = J_y^{(j)} - \mathrm{E}(J_y^{(j)}) - \mathbf{Z}^T\mathrm{E}(\mathbf{Z}J_y^{(j)})$ by $\varepsilon_y^{(j)} = J_y^{(j)} - \mathrm{E}(J_y^{(j)})$, we will obtain the robust version of inner product matrices:

$$\mathbf{G}_j = \big(\mathbf{A}_j^T \otimes \boldsymbol{\Sigma}^{-1/2}\big)\mathrm{cov}\big\{\mathbf{Z}\big(\boldsymbol{\varepsilon}^{(j)}\big)^T\big\}(\mathbf{A}_j \otimes \boldsymbol{\Sigma}^{-1/2}), \quad (3.4)$$

where $\boldsymbol{\varepsilon}^{(j)} = (\epsilon_1^{(j)}, \ldots, \epsilon_{h_j}^{(j)})^T \in \mathbb{R}^{h_j}$ is the vector of modified residuals. Then we similarly define the fused inner product matrix $\mathbf{G}_F^{-1}$ and $\mathbf{G}_D^{-1}$ in responding to $\boldsymbol{\Gamma}_F^{-1}$ and $\boldsymbol{\Gamma}_D^{-1}$. The estimators of the CS obtained from minimizing $\mathcal{F}_d(\mathbf{B}, \mathbf{C}; \mathbf{G}_F^{-1})$ and $\mathcal{F}_d(\mathbf{B}, \mathbf{C}; \mathbf{G}_D^{-1})$ will be called the robust FIRE and robust DIRE. And their $\sqrt{n}$-consistency follows directly from Theorem 2. We will show simulation results in Section 5 regarding both the original fused estimators and the robust fused estimators.

FIRE preserves asymptotic optimality, while DIRE and the robust estimators are all asymptotically suboptimal. However, in a finite sample problem, DIRE often performs better than both FIRE and IRE because estimating $\boldsymbol{\Gamma}_D^{-1}$ well requires no more observations than estimating $\boldsymbol{\Gamma}_j$ with the largest $h_j \in H$. Simulation studies and real data application in Section 5 also indicate that DIRE is better than or close to FIRE in most situations. FIRE is preferred over DIRE only when n is very large comparing to p, see Section 5.3 for a real data example where $n = 1030$ and $p = 8$. Moreover, it is straightforward to generalize DIRE to other methods by pretending independent slicing schemes. At the same time, the $\sqrt{n}$ consistency and asymptotic variance properties of DIRE could be maintained. We illustrate this by adapting DIRE for fusing SIR estimators in the next section.

3.4 Degenerate Slicing Schemes and Fused SIR

From Section 3.2, we know that FIRE is asymptotically at least as efficient as IRE with any single slicing scheme in H. As a referee pointed out, it would be helpful to guide practice if we know when a single slicing scheme IRE could be asymptotically efficient with respect to the fused data matrix. Then IRE based on this slicing scheme would have the same asymptotic performance as FIRE.

Corollary 1. Under the conditions of Theorem 2, the following statements are equivalent.

1. IRE with slicing scheme h_k is asymptotic efficient with respect to $\hat{\boldsymbol{\xi}}_F$.
2. $\mathrm{avar}(\sqrt{n}\hat{\boldsymbol{\beta}}_k\hat{\boldsymbol{\beta}}_k^T) = \mathrm{avar}(\sqrt{n}\hat{\boldsymbol{\beta}}_F\hat{\boldsymbol{\beta}}_F^T)$.
3. $\boldsymbol{\Gamma}_j - \boldsymbol{\Gamma}_{kj}^T\boldsymbol{\Gamma}_k^{-1}\boldsymbol{\Gamma}_{kj} = 0$ for all $j \neq k$.
4. $\boldsymbol{\Gamma}_F = \mathbf{O}\left(\begin{smallmatrix}\boldsymbol{\Gamma}_k & 0\\ 0 & 0\end{smallmatrix}\right)\mathbf{O}^T$ for some full rank transformation $\mathbf{O}$.

The situation in Corollary 1 happens when H is a set of poorly chosen degenerated slicing schemes, as discussed in Section 3.1. Another possible cause of such situation is that $\mathbf{X}$ depends on Y

only through some latent slices of Y and one slicing scheme h_k happens to coincide with that latent structure. In our first simulation example in Section 5, we simulated data in this way and observed that FIRE had approximately the same performance to IRE with the latent slicing scheme, while DIRE, being asymptotically suboptimal, had even improved finite performance.

To elaborate on the above discussion, we use the IR model of Cook et al. (2012). In addition to the linearity and the coverage conditions, they assumed the covariance condition that $\operatorname{var}(\mathbf{X}|Y)$ is nonstochastic, which led to the model

$$\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\eta}\mathbf{f}(Y_i) + \boldsymbol{\varepsilon}_i, \ i = 1, \ldots, n, \tag{3.5}$$

where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$ and $\boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\Phi})$ and is independent of Y and the vector-valued functions $\mathbf{f}(Y_i) \in \mathbb{R}^d$. Since we are focusing on slicing instead of a general functional estimation, we assume $\mathbf{f}(Y_i)$ is a linear function of slice indicators $J_y^{(k)}$, $y = 1, \ldots, h_k$, of some particular single slicing scheme h_k. Neither $\boldsymbol{\eta}$ nor $\mathbf{f}(Y_i)$ is identified in model (3.5), because for any nonsingular $d \times d$ matrix $\mathbf{A}$, $\boldsymbol{\eta}\mathbf{A}\mathbf{A}^{-1}\mathbf{f}(Y_i) = \boldsymbol{\eta}\mathbf{f}(Y_i)$ leads to a different parameterization. We could add additional constraint $\boldsymbol{\eta}^T\boldsymbol{\eta} = \mathbf{I}_d$ to get uniqueness. Nevertheless, the CS is affected only by $\operatorname{span}(\boldsymbol{\eta}\mathbf{f}(Y_i))$.

Lemma 1. Under model (3.5), IRE with slicing scheme h_k is asymptotically efficient with respect to $\hat{\boldsymbol{\xi}}_F$, where H is any set of slicing schemes such that $h_k \in H$. Moreover, if the errors in (3.5) are isotropic, $\boldsymbol{\Phi} = \sigma^2\mathbf{I}_p$, then IRE is asymptotically equivalent to SIR, which is the maximum-likelihood estimation (MLE) under these assumptions.

Lemma 1 implies that we could judge the necessity of fusing based on model (3.5): if this model holds for some known h_k, then fusing would not be necessary. However, in practice, it is difficult to check such conditions and almost impossible to find such h_k for a continuous Y, and then fusing will alleviate the burden of choosing slicing schemes.

Lemma 1 also implies that SIR could be used instead of IRE for isotropic errors. SIR is based on the spectral decomposition of a kernel matrix $\widehat{\mathbf{M}}_{\text{SIR}} = \sum_{y=1}^{h} \hat{f}_y \bar{\mathbf{Z}}_y \bar{\mathbf{Z}}_y^T$, where $\bar{\mathbf{Z}}_y = \hat{\boldsymbol{\Sigma}}^{-1/2}(\bar{\mathbf{X}}_y - \bar{\mathbf{X}})$ is the sample average of the n_y points in slice y computed in the $\mathbf{Z}$-scale. Cook and Ni (2005) showed that SIR is covered by the IR family with the following objective function:

$$\begin{aligned}\mathrm{F}_d^{\mathrm{SIR}}(\mathbf{B}, \mathbf{C}; h) = {} & (\operatorname{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \operatorname{vec}(\mathbf{B}\mathbf{C}))^T \\ & \times \operatorname{diag}\{\hat{f}_y^{-1}\hat{\boldsymbol{\Sigma}}\}(\operatorname{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \operatorname{vec}(\mathbf{B}\mathbf{C})),\end{aligned} \tag{3.6}$$

where the columns of $\mathbf{B} \in \mathbb{R}^{p \times d}$ represent a basis of SIR's estimator of the CS and $\mathbf{C} \in \mathbb{R}^{d \times h}$ is again a coordinate matrix used only in fitting. Let $(\hat{\boldsymbol{b}}, \hat{\boldsymbol{c}})$ be the minimizer of (3.6). Then $\operatorname{span}(\hat{\boldsymbol{b}})$ is the same as the span of the d eigenvectors of $\widehat{\mathbf{M}}_{\text{SIR}}$ corresponding to its largest d eigenvalues.

Let $H = \{h_1, \ldots, h_K\}$ denote a set of K quantile slicing schemes. Analogous to DIRE, we define the fused SIR estimator via its objective function

$$\mathrm{F}_d^{\mathrm{FSIR}}(\mathbf{B}, \mathbf{C}) = \sum_{j=1}^{K} \mathrm{F}_d^{\mathrm{sir}}(\mathbf{B}, \mathbf{C}_j; h_j),$$

where $\mathbf{C} = (\mathbf{C}_1, \ldots, \mathbf{C}_K) \in \mathbb{R}^{d \times \sum h_j}$. Let $(\hat{\boldsymbol{b}}_1, \hat{\boldsymbol{c}}_1)$ be minimizer of $\mathrm{F}_d^{\mathrm{sir}}(\mathbf{B}, \mathbf{C}_1; h_1)$; let $(\hat{\boldsymbol{b}}_F, \hat{\boldsymbol{c}}_F) \equiv (\hat{\boldsymbol{b}}_F, \hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_K)$ be a minimizer of the fused SIR objective function $\mathrm{F}_d^{\mathrm{FSIR}}(\mathbf{B}, \mathbf{C})$. In practice, it is unnecessary to numerically minimize the objective function $\mathrm{F}_d^{\mathrm{FSIR}}(\mathbf{B}, \mathbf{C})$ since we can obtain an explicit form of the minimizer. The next lemma provides a spectral decomposition approach for fused SIR.

Lemma 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 second moments. Let $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_p$ be the eigenvectors of $\widehat{\mathbf{M}}_{\text{FSIR}} = \sum_{j=1}^{K} \sum_{y=1}^{h_j} \hat{f}_y^{(j)} \bar{\mathbf{Z}}_y^{(j)} \bar{\mathbf{Z}}_y^{(j)T}$ corresponding to eigenvalues $\hat{\lambda}_1 > \cdots > \hat{\lambda}_d > \hat{\lambda}_{d+1} \geq \cdots \geq \hat{\lambda}_p$, where $\bar{\mathbf{Z}}_y^{(j)}$ is the sample average in the yth slice of slicing scheme h_j. Then the minimizer $\operatorname{span}(\hat{\boldsymbol{b}}_F) = \operatorname{span}(\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_d)$.

The kernel matrix $\widehat{\mathbf{M}}_{\text{FSIR}}$ is the sum of K kernel matrices for SIR, which is very intuitive by treating the K slicing schemes independently. In the same way, fused estimators of SAVE (Cook and Weisberg 1991) and DR (Li and Wang 2007) can be constructed by spectral decomposing the sum of kernel matrices with different slicing schemes.

4. ESTIMATION PROCEDURE AND COMPUTING ALGORITHM

We will use the data matrix $\hat{\boldsymbol{\xi}}_F$ and a consistent estimator of $\boldsymbol{\Gamma}_F^{-1}$ or a consistent estimator of $\boldsymbol{\Gamma}_D^{-1}$ for $\mathcal{F}_d(\mathbf{B}, \mathbf{C}; \boldsymbol{\Psi})$ in (2.2). Also, we adopt the same computing algorithm, the alternating least-square algorithm, as in Cook and Ni (2005), to minimize the objective function. Since $\boldsymbol{\Gamma}_D^{-1}$ is block diagonal with $\boldsymbol{\Gamma}_j^{-1}$ on the diagonal, we can simply replace it by the consistent estimator used in the IRE, $\hat{\boldsymbol{\Gamma}}_j^{-1}$. So that $\operatorname{diag}(\hat{\boldsymbol{\Gamma}}_1^{-1}, \ldots, \hat{\boldsymbol{\Gamma}}_K^{-1})$ is a consistent estimator of $\boldsymbol{\Gamma}_D^{-1}$. Similarly for $\boldsymbol{\Gamma}_F^{-1} = [(\mathbf{A}_F^T \otimes \mathbf{I}_p)\boldsymbol{\Gamma}_H(\mathbf{A}_F \otimes \mathbf{I}_p)]^{-1}$, we only need to estimate $\boldsymbol{\Gamma}_H = \operatorname{cov}(\operatorname{vec}(\boldsymbol{\Sigma}^{-1/2}\mathbf{Z}\mathbf{E}^T))$ by substituting the sample version $\hat{\mathbf{E}}_i = (\hat{\epsilon}_{1i}^{(1)}, \ldots, \hat{\epsilon}_{h_1 i}^{(1)}, \ldots, \hat{\epsilon}_{1i}^{(K)}, \ldots, \hat{\epsilon}_{h_K i}^{(K)})^T$ and $\hat{\epsilon}_{yi}^{(j)} = J_{yi}^{(j)} - n_y^{(j)}/n - (\mathbf{X}_i - \bar{\mathbf{X}}_{..})^T \hat{\boldsymbol{\xi}}_y^{(j)} n_y^{(j)}/n, i = 1, \ldots, n$, $y = 1, \ldots, h_j$, for all $j = 1, \ldots, K$. Then a consistent estimate of $\boldsymbol{\Gamma}_F^{-1}$ follows by noticing $\mathbf{A}_F = \operatorname{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_K)$ is a block diagonal nonstochastic matrix. For the robust FIRE and robust DIRE, we apply the same estimation procedure except that $\hat{\epsilon}_{yi}^{(j)}$ was replaced by $\hat{\varepsilon}_{yi}^{(j)} = J_{yi}^{(j)} - n_y^{(j)}/n$.

We discussed in Section 2 that we could guarantee the nonsingularity of $\boldsymbol{\Gamma}_F$ by choosing slicing schemes with different slice boundaries. However, the $\sum_{j=1}^{K}(h_j - 1)p$ by $\sum_{j=1}^{K}(h_j - 1)p$ matrix $\hat{\boldsymbol{\Gamma}}_F$ is possibly ill-conditioned and thus we use the Moore–Penrose generalized inverse of $\hat{\boldsymbol{\Gamma}}_F$ instead of the regular inverse matrix in our algorithm. In Section 8, we discuss the possibility of using regularized estimators of $\hat{\boldsymbol{\Gamma}}_F^{-1}$ to replace the generalized inverse.

5. NUMERICAL STUDIES

In this section, we compare our fused IREs to CUME, IRE, and SIR using quantile slicing with various values of h. The simulation results based on various settings support our theoretical conclusions. We then apply our method to a concrete compressive strength dataset (Yeh 1998).

Table 1. Comparisons of the fused estimators and other methods from 5-slice inverse model (5.1), with four different error structures (a)–(d)

	(a)		(b)		(c)		(d)	
	r^2	q^2	r^2	q^2	r^2	q^2	r^2	q^2
SIR $h = 5$	0.971	0.943	0.969	0.938	0.657	0.352	0.644	0.332
SIR $h = 7$	0.944	0.890	0.934	0.869	0.411	0.117	0.459	0.141
IRE $h = 5$	0.947	0.894	0.958	0.916	0.957	0.916	0.936	0.874
IRE $h = 7$	0.912	0.826	0.917	0.838	0.946	0.894	0.930	0.863
CUME	0.948	0.898	0.921	0.846	0.541	0.194	0.546	0.200
DIRE	0.961	0.922	0.968	0.936	0.949	0.900	0.942	0.886
FIRE	0.957	0.915	0.965	0.930	0.943	0.889	0.933	0.869
s.d.	0.0010	0.0020	0.0008	0.0016	0.0027	0.0052	0.0034	0.0064

5.1 Inverse Regression

We calculated the vector correlation coefficient q^2 (Hotelling 1936), the trace correlation r^2 (Hooper 1959), and angles $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$ to measure the closeness between the estimated CS $\operatorname{span}(\hat{\boldsymbol{\beta}})$ and the true CS $\mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}(\boldsymbol{\beta})$:

$$q^2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \prod_{i=1}^{d} \rho_i, \quad r^2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{1}{d} \sum_{i=1}^{d} \rho_i,$$

where $\rho_1, \rho_2, \ldots, \rho_d$ are ordered eigenvalues of $\boldsymbol{\beta}^T \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^T \boldsymbol{\beta}$. Both $q^2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and $r^2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ range from 0 to 1, with values closer to 1 indicating that the two subspace $\operatorname{span}(\hat{\boldsymbol{\beta}})$ and $\mathcal{S}_{Y|\mathbf{X}}$ are closer. Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_d) \in \mathbb{R}^{p \times d}$. Then the angle θ_i is defined as the angle between $\hat{\boldsymbol{\beta}}_i$ and its projection onto $\mathcal{S}_{Y|\mathbf{X}}$. In the following simulation examples, we report the averaged values of q^2, r^2, and $\boldsymbol{\theta}$ and their standard errors.

To imitate the asymptotic performance of the proposed two fused IREs, we generated data with sample size $n = 600$, dimension $p = 6$ from a model with dimension $d = 2$. The model was constructed with a clear optimal choice of slicing scheme: five slices with equal number of observations in each slice. Y was simulated from the uniform distribution on the interval [0, 5] and

$$\mathbf{X} = \boldsymbol{\Phi} \boldsymbol{\eta} \mathbf{f}(Y) + c(Y) \boldsymbol{\epsilon}, \tag{5.1}$$

where $\boldsymbol{\epsilon}$ is a multivariate normal random variable with mean 0, covariance matrix $\boldsymbol{\Sigma} > 0$, and is independent of Y. The vector-valued function $\mathbf{f}(Y) \in \mathbb{R}^{r \times 1}$ was designed to consist of indicator functions of Y:

$$\mathbf{f}(Y) = (J_1, \ldots, J_5)^T, \ J_k = I(k-1 < Y \leq k), \ k = 1, \ldots, 5,$$

which is why five slices are optimal for model (5.1). We used $c(Y) \in \mathbb{R}^1$ to construct either homoscedastic or heteroscedastic error structure. The elements of both constant matrices $\boldsymbol{\Phi} \in \mathbb{R}^{p \times d}$ and $\boldsymbol{\eta} \in \mathbb{R}^{d \times r}$ were generated as a random sample from the standard normal distribution and then standardized to be semiorthogonal. From Cook and Forzani (2008), we know that for this setting $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}^{-1} \operatorname{span}(\boldsymbol{\Phi})$.

Four different error structures were employed to study the fusing methods:

(a) isotropic error structure $c(Y) = 0.1$ and $\boldsymbol{\Sigma} = \mathbf{I}_p$;
(b) autoregressive error structure $c(Y) = 0.1$, $(\boldsymbol{\Sigma})_{ij} = 0.5^{|i-j|}$;
(c) heteroscedastic error structure $c(Y) = 0.01 J_1 + 0.05(J_2 + J_3) + J_4 + 10 J_5$, $\boldsymbol{\Sigma} = \mathbf{I}_p$;
(d) heteroscedastic error structure $c(Y) = 0.01 J_1 + 0.05(J_2 + J_3) + J_4 + 10 J_5$, $(\boldsymbol{\Sigma})_{ij} = 0.5^{|i-j|}$.

Table 1 shows the performances of SIR, IRE, CUME, FIRE, and DIRE. For SIR and IRE, we reported two choices of numbers of slices, $h = 5$ and $h = 7$. This is because we know $h = 5$ is the optimal choice in this model, while $h = 7$ is just an arbitrary choice to represent a typical situation when we have no prior information about how to choose the slicing scheme. FIRE and DIRE were obtained from fusing a set of 13 quantile slicing schemes $H = \{3, \ldots, 15\}$. This set likely covers all the sensible slicing schemes, so we have no responsibility in picking the "best" numbers of slices.

As can be seen in Table 1, in all the four error structure settings, both fused estimators outperformed CUME, SIR with $h = 7$ and IRE with $h = 7$. While at the same time, DIRE always outperformed FIRE. In a great majority of comparisons, DIRE was the best method overall. By comparing it to SIR with $h = 5$ and IRE with $h = 5$, the two classical method with the optimal number of slices, we notice that it was at least as good as if not better than IRE with $h = 5$. However, the fused methods lost to SIR with $h = 5$ in the simulation model with isotropic error structure because this simplest case is designed to best fit in the SIR context. SIR with five slices outperformed all other methods in the isotropic error structures (a) and (b), because SIR with five slices gives the MLE for this problem (see Cook and Forzani 2008), while IRE and the fused estimators had more parameters to estimate and lost some efficiency. Even with such isotropic error structure, the user still has to choose the correct number of slices for SIR to outperform our fused estimators. Although SIR with the best choice of slices could be more efficient than the fused estimators, the efficiency gain by SIR is very little. But in more complex situations, SIR as well as CUME may fail to capture the CS when IRE and fused estimators still work well. In the more challenging settings (c) and (d), where the error structures are heteroscedastic, SIR and CUME fail to capture the CS.

Figure 1 shows the failure of SIR and CUME with heteroscedastic error structure (d) in terms of the angles θ_1 and θ_2, and shows the optimality of the fused estimators with respect to IRE using different numbers of slices. The dimension $d = \dim(\mathcal{S}_{Y|\mathbf{X}}) = 2$, so we used two angles to visually summarize the estimation results. In this summary plot, the horizontal

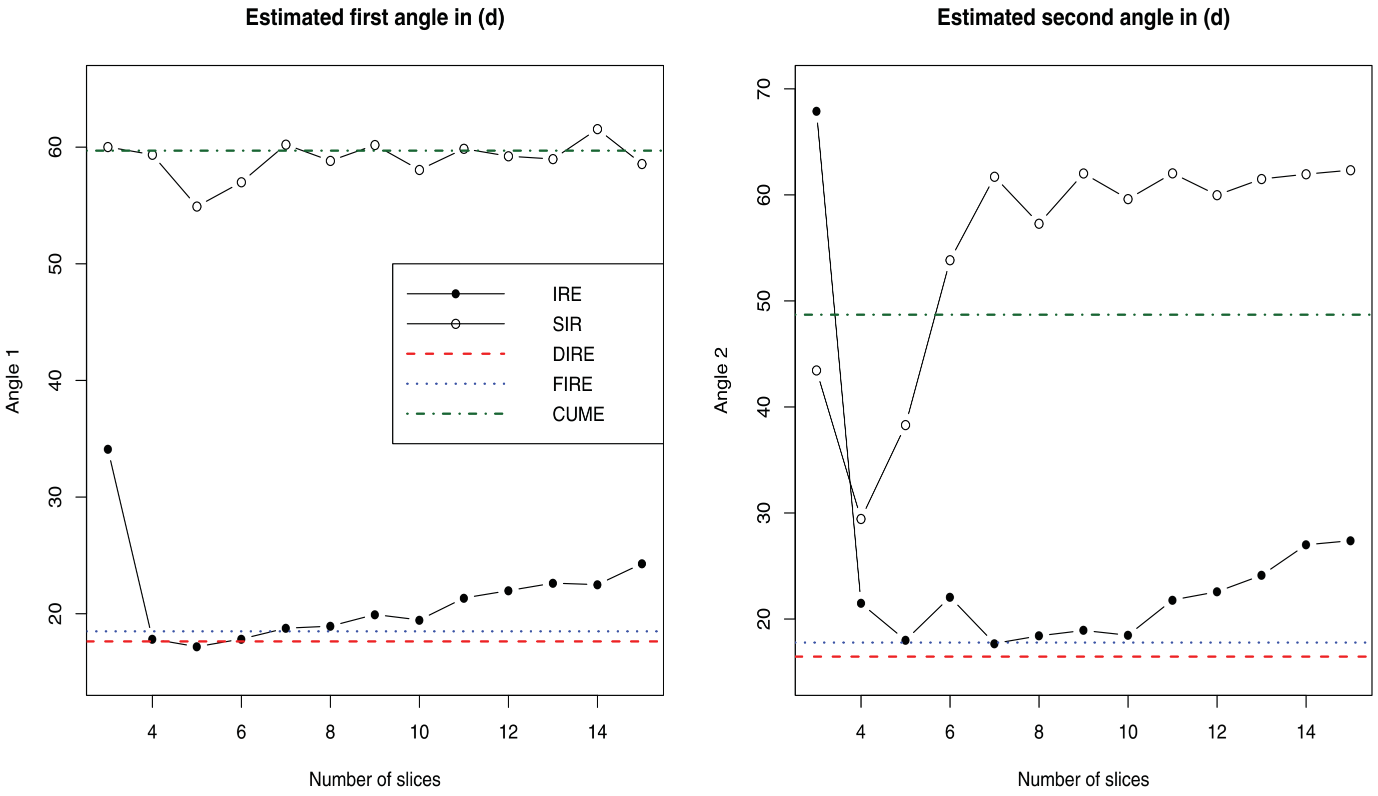

Figure 1. 5-slice inverse model (5.1) with heteroscedastic error structure (d).

axis represents the numbers of slices used in IRE and SIR. Because of the nature of this model, $h = 5$ produced the smallest angles for both IRE and SIR. As the graphs show, DIRE delivered more accurate estimates than FIRE and CUME.

To better show that the two fused estimators have smaller asymptotic variances in estimating the CS than the asymptotic variances of IRE with arbitrary number of slices $h \in H = \{3, \ldots, 15\}$, we plotted for the isotropic case (a) the two angles between the estimated CS and the true CS in Figure 2. From this plot, we can see clearly that DIRE has significantly smaller first and second angles than IRE even with five slices, while FIRE has similar first and second angles to IRE with five slices. And a small change in the numbers of slices can lead to substantial change in the output of IRE. For example, the second angle of IRE with $h = 6$ is about twice as that of $h = 5$. Fortunately, FIRE and DIRE are unaffected.

Figure 3 shows the behavior of the IRE estimators with h ranging from $h = 3$ to $h = 100$. In the 5-slice inverse model with the heteroscedastic error structure (d), the trace correlations of the IRE estimators are almost monotonically decreasing as the number of slices increases. This result was anticipated by Zhu and Ng (1995): too many slices with too few observations per slice are not good for the variances of estimates. In general, we find that the performance of IRE is usually monotonically decreasing in h after a moderate number. However, the fused slicing scheme of FIRE in this case has approximately 100 slices but can still perform well. Actually, FIRE and DIRE did much better than IRE with $h = 100$: from Figure 3 the trace correlation for IRE is about 0.70, while the trace correlations for DIRE and FIRE from Table 1 are 0.942 and 0.933. As discussed in Section 3.2, fusing slicing schemes with small values of h, that is, $h = 3, \ldots, 15$, may lead to much more accurate basis estimation than IRE with a large value $h = 100$.

5.2 Forward Regression Models

We have also studied many forward regression models and observed the following qualitative finite sample performance. (1) DIRE always outperforms FIRE and IRE with any h, unless the sample size is quite large and then FIRE performs better than IRE and a bit better than DIRE. We expect that this happens because the cost of estimating the FIRE inner product matrix can be substantially greater than the cost of estimating the DIRE inner product matrix. (2) When the dimension-reduction subspace is easy to find, as is the case with a homoscedastic forward linear regression model, SIR with a reasonable number of slices can outperform FIRE, DIRE, and IRE by a small amount. Of course, if we knew we had a simple forward model at the outset, then SDR methods would be unnecessary. (3) In challenging forward regressions where the CS is not easy to find, FIRE and DIRE are substantially better than SIR and CUME and better than IRE by an amount that depends on the number of slices. One example of this behavior is described below; additional forward regression simulation are given in the supplement.

Data $\{X_{i1}, \ldots, X_{i15}, Y_i\}$, $i = 1, \ldots, 400$, were generated as follows. First, we sampled $\mathbf{X}_i \in \mathbb{R}^{15}$ from a mixture of normal distributions

$$\mathbf{X} \sim \frac{1}{4}N(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1) + \frac{1}{2}N(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2) + \frac{1}{4}N(\boldsymbol{\mu}_3, \boldsymbol{\sigma}_3),$$

where $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_3 = (1, 0, \ldots, 0)^T$, $\boldsymbol{\mu}_2 = (2, 0, \ldots, 0)^T$, $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 = \sqrt{0.1}\boldsymbol{I}_{15}$, and $\boldsymbol{\sigma}_3 = \sqrt{10}\boldsymbol{I}_{15}$. Then the Y_i's were

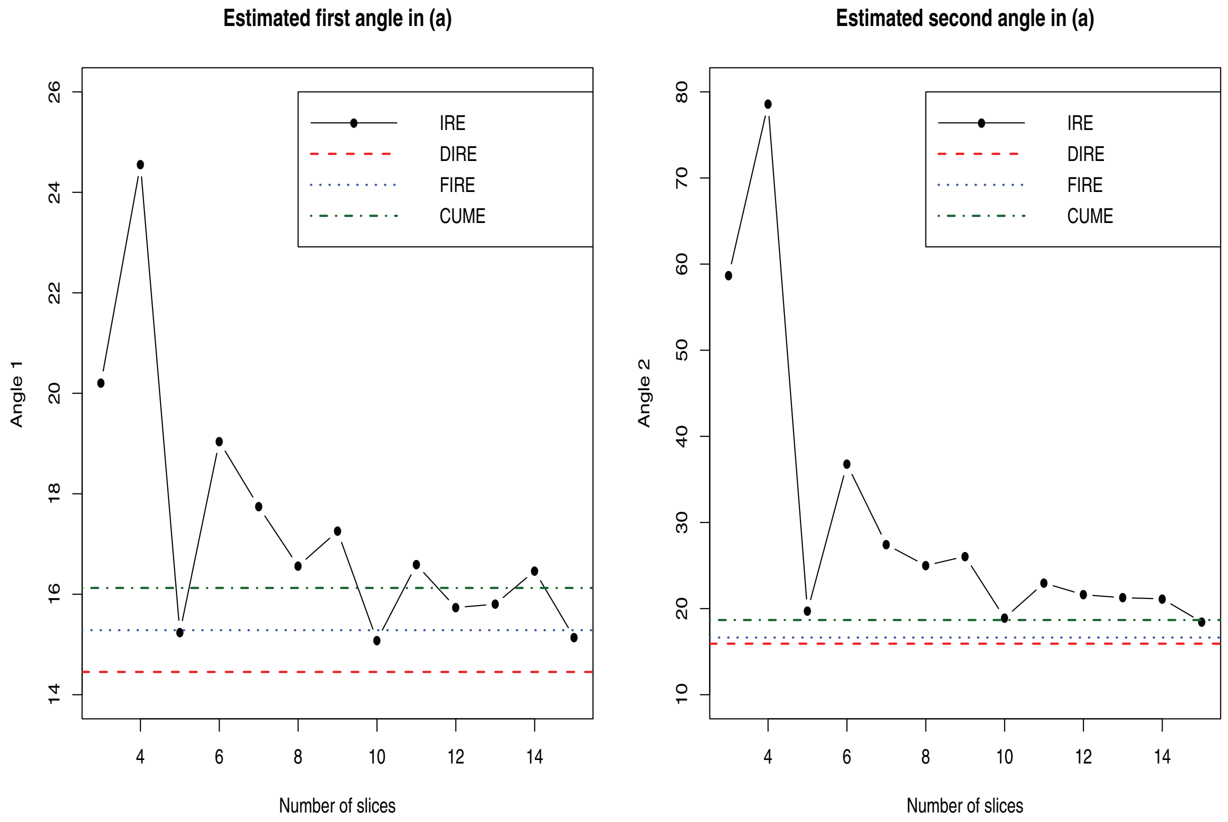

Figure 2. 5-slice inverse model (5.1) with isotropic error structure (a). SIR fluctuates around the line of CUME and is not included in the plot for better visualization.

generated as

$$Y = |\sin(X_1)| + 0.2\epsilon, \tag{5.2}$$

where ϵ is a uniform (0,1) random variable. Thus, $\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{(1, 0, \ldots, 0)^T\}$. Again, we simulated 500 datasets and summarized the angle θ between $\mathcal{S}_{Y|\mathbf{X}}$ and its estimate in the left panel of Figure 4, where we used $H = \{3, \ldots, 15\}$ for both fused estimators. We adopted the robust version of the inner product matrices (Ni and Cook 2007) for both IRE and the fused estimators. Since the intraslice variances var($\mathbf{X}|Y$) are nonconstant, SIR and CUME both suffered. These two methods completely failed to estimate the CS. On the other hand, FIRE and DIRE automatically adjusted weights on each slices through their inner product matrices $\hat{\boldsymbol{\Gamma}}_F^{-1}$ and $\hat{\boldsymbol{\Gamma}}_D^{-1}$, and thus estimated

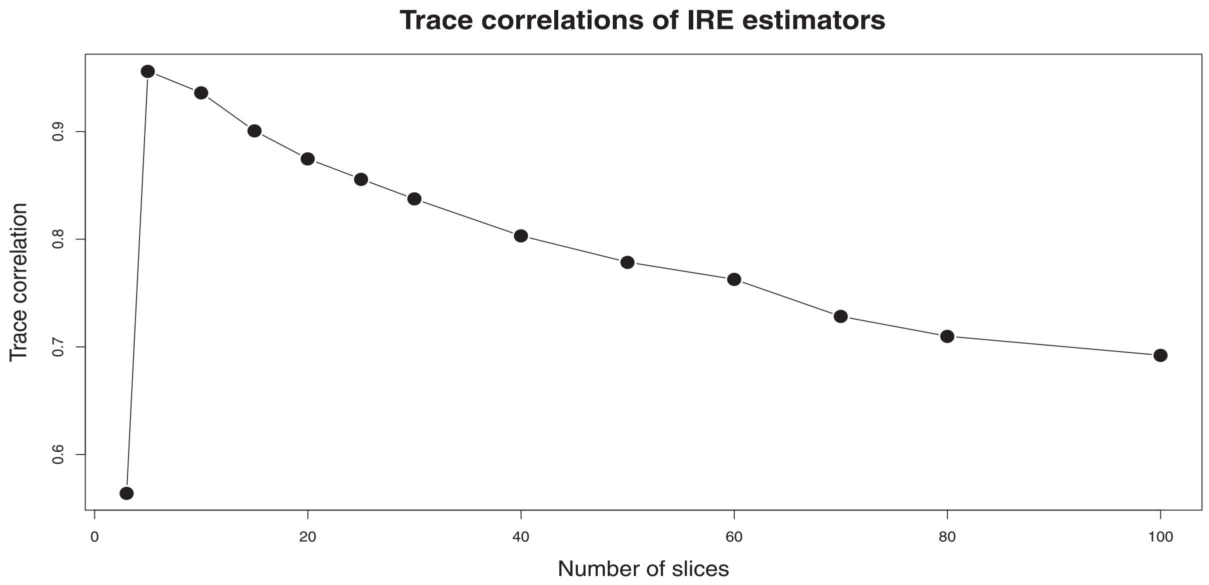

Figure 3. 5-slice inverse model with heteroscedastic error structure (d): trace correlations of the IRE estimators with different numbers of slices.

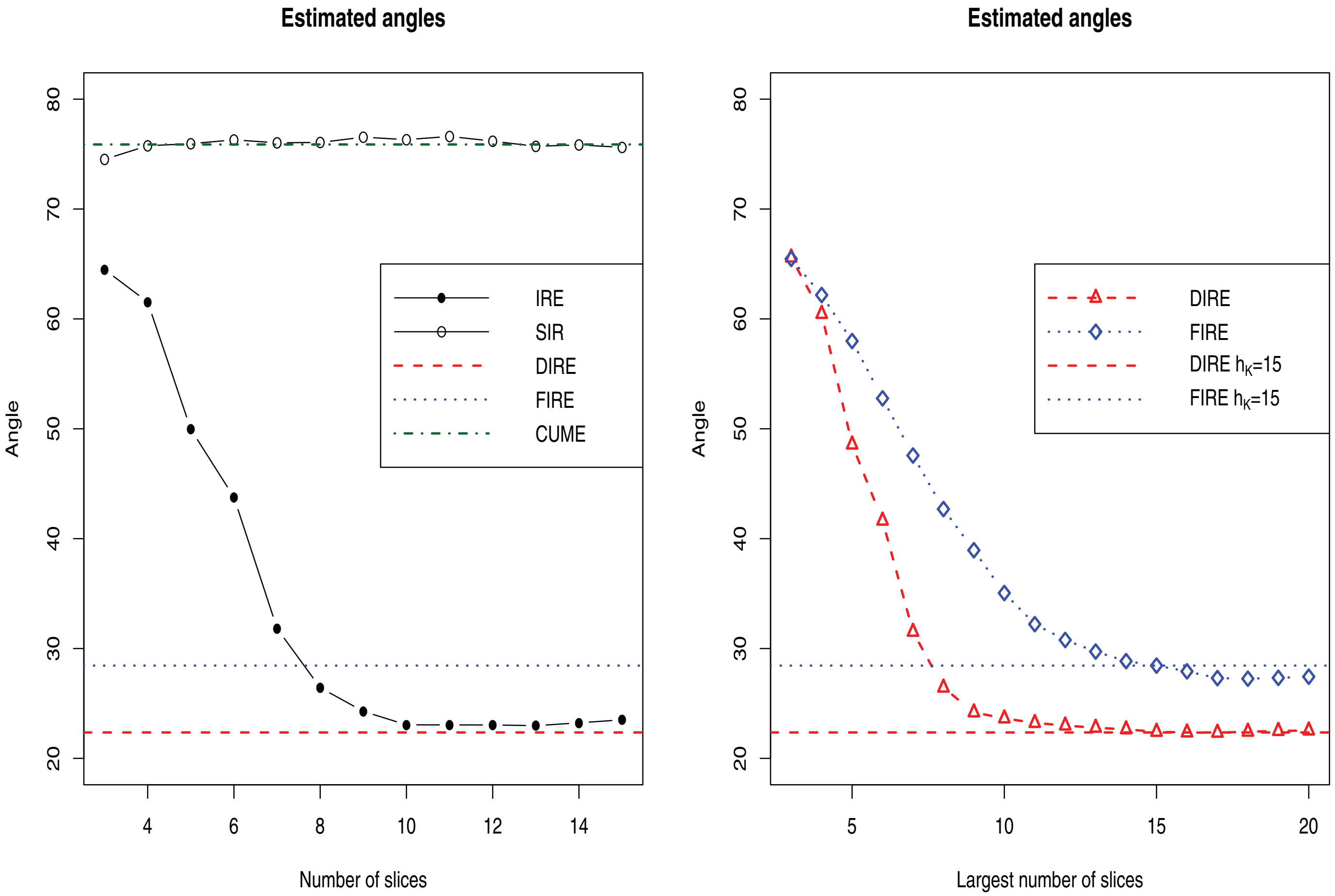

Figure 4. Estimated angle θ by different methods in the forward model (5.2). Left panel: θ versus h in a single quantile slicing scheme with the fused estimators and CUME superimposed. Right panel: θ versus the largest number of slices h_K in the fused estimators with $H = \{3, 4, \dots, h_K\}$.

the CS efficiently. The left panel of Figure 4 shows that DIRE dominated FIRE in estimation, and outperformed IRE with any number of slices. Although it is a little difficult to see from the plot, the IRE angle starts to increase after about 10 slices, similar to the phenomenon shown in Figure 3.

The right panel of Figure 4 was constructed to illustrate the behavior of FIRE and DIRE as a function of h_K, the largest slice number in the fused estimators. Specifically, we plotted the average angle θ against h_K with $H = \{3, 4, \dots, h_K\}$. The behavior in right panel of Figure 4 is typical from our experience: the fused estimators rapidly accumulate information as h_K increases through the small integers, eventually flattening with very little change thereafter. Because of such results we typically choose $10 \le h_K \le 20$ when fusing. We have not systematically studied the behavior of the fused estimators with $h_K > 20$.

5.3 Concrete Compressive Strength Dataset

For this illustration (Yeh 1998, *http://archive.ics.uci.edu/ml/datasets.html*), we want to predict the compressive strength of high-performance concrete (Mpa, Y) from seven ingredients ($X_1, \dots, X_7$) and age (days, X_8). These ingredients are cement (X_1), blast furnace slag (X_2), fly ash (X_3), water (X_4), superplasticizer (X_5), coarse aggregate (X_6), and fine aggregate (X_7) all measured in kg/m^3. We used all 1030 observations in estimation.

We applied SIR and IRE with various numbers of slices, and the usual dimension tests always indicated $d = 2$ with significant level 0.05. IRE dimension tests with different quantile slicing schemes typically result in the same conclusion for d. However, because of the discrete nature of the decision, it seems inevitable that the conclusion about d will vary with the slicing scheme in some datasets (see, e.g., Ye and Weiss 2003), and then additional analysis in the context of the fused estimator may be necessary to guide the final choice. A graphical study, the general permutation test proposed by Cook and Yin (2001) and some form of cross-validation may all be useful in this regard, depending on application specific goals. It is also possible to develop dimension estimation based directly on the fused estimator, but that is outside the scope of this report.

We then plotted Y versus the first two orthogonal directions obtained from CUME, DIRE, and FIRE in Figure 5, where we again chose $H = \{3, \dots, 15\}$ for DIRE and FIRE. We randomly sampled and plotted 200 out of the 1030 observations and a LOWESS smooth line based on all the 1030 observations with the same smoothing parameter. The first directions of FIRE, DIRE, and CUME are essentially the same. However, the second directions of FIRE and DIRE gave clear linear patterns while the second direction of CUME apparently provided little information regarding the response. The plots (not included) for

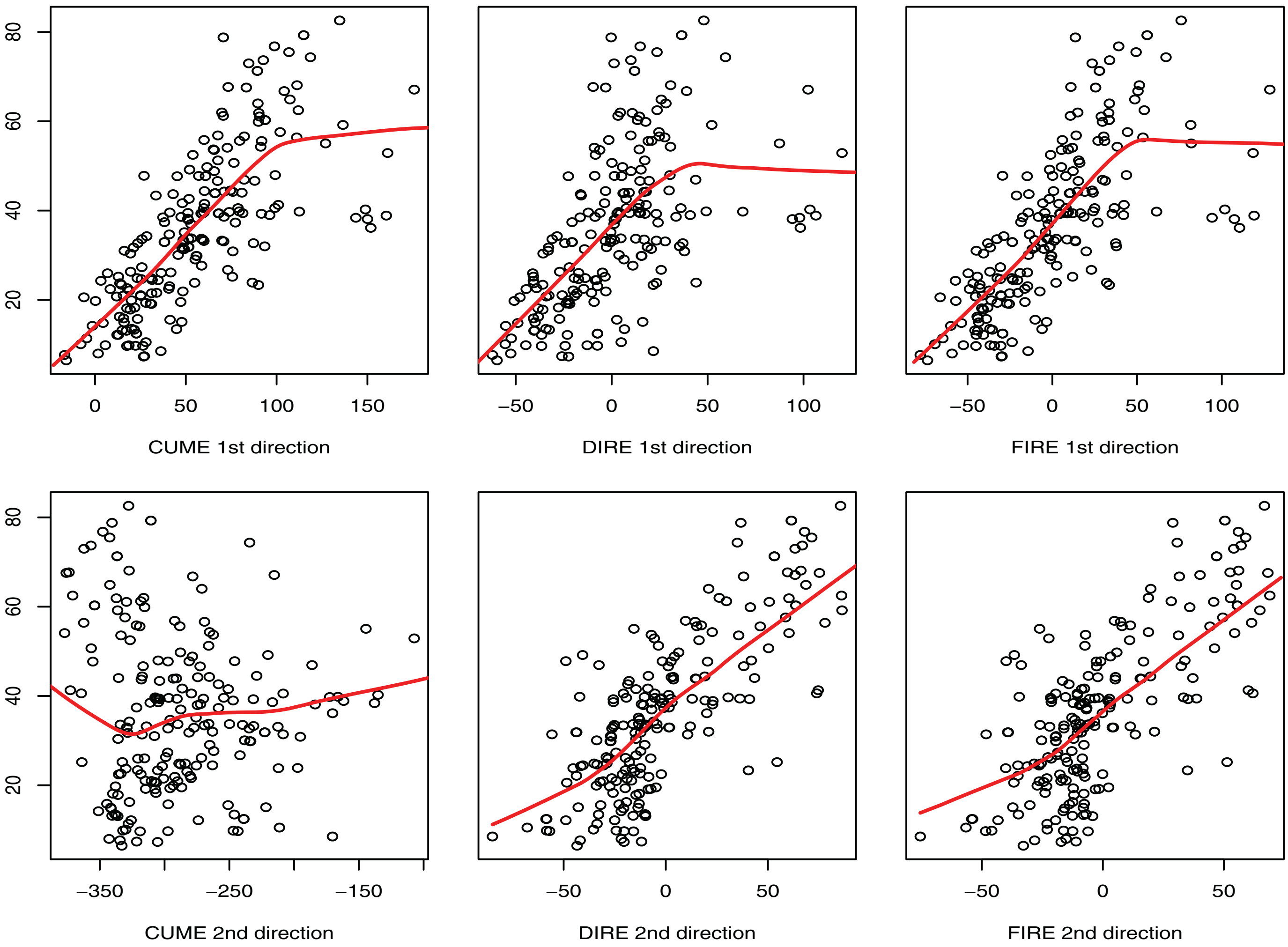

Figure 5. Concrete compressive strength dataset: scatterplots of the response (vertical axes) and the first two orthogonal directions (horizontal axes), with mean smooths added to aid visualization.

SIR and IRE with various slicing schemes showed similar characteristics in the first direction and vague patterns in the second direction.

To further assess the estimation of the CS, we used the bootstrap criterion developed by Ye and Weiss (2003) and used by Zhu, Zhu, and Feng (2010). We used the bootstrap to generate 100 dataset with sample size $n = 400$ and 100 dataset with sample size $n = 1030$. With $d = 2$, we then computed the trace correlation coefficient $r^2(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^b)$, $b = 1, \ldots, 100$, between the originally estimated CS span$(\hat{\boldsymbol{\beta}})$ and the bootstrap estimate span$(\hat{\boldsymbol{\beta}}^b)$. The results are summarized in Table 2. From the table, we can see that FIRE has the largest r^2, followed by DIRE. Therefore, FIRE is the best method for this dataset, which might have been anticipated since FIRE is the asymptotically optimal fused estimator.

6. COMPOSITE LIKELIHOODS AND FUSED LAD

A good choice of slicing scheme may become more crucial in second-order methods such as SAVE (Cook and Weisberg 1991), DR (Li and Wang 2007), and LAD (Cook and Forzani 2009). This is because the second-order methods rely on higher moment of the conditional variable $\mathbf{X}_y \equiv (\mathbf{X}|J_y = 1)$, $y = 1, \ldots, h$. In this section, we describe a fused version of LAD.

Within each slice y, the LAD model assumes conditional normality of $\mathbf{X}_y$: $\mathbf{X}_y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y)$ with arbitrary mean function $\boldsymbol{\mu}_y$ and covariance $\boldsymbol{\Delta}_y > 0$. Let $\boldsymbol{\Delta} = \mathrm{E}(\boldsymbol{\Delta}_Y) \equiv \mathrm{E}\{\mathrm{var}(\mathbf{X}|Y)\}$ and

Table 2. Concrete compressive strength dataset: averaged r^2 using bootstrap samples. The standard errors for the averaged r^2 are between 0.012 and 0.016 for $n = 1030$, and are between 0.010 and 0.018 for $n = 400$

	SIR			IRE					
	$h = 5$	$h = 10$	$h = 15$	$h = 5$	$h = 10$	$h = 15$	CUME	DIRE	FIRE
$n = 1030$	0.699	0.685	0.674	0.730	0.779	0.643	0.721	0.820	0.915
$n = 400$	0.618	0.596	0.613	0.703	0.701	0.623	0.628	0.771	0.806

assume the existence of a d-dimensional CS. Then the MLE for the CS is obtained by maximizing the following likelihood-based objective function (Cook and Forzani 2009):

$$L_d(\boldsymbol{\beta}; h) = \frac{n}{2}\log|\boldsymbol{\beta}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta}| - \sum_{y=1}^{h}\frac{n_y}{2}\log|\boldsymbol{\beta}^T\hat{\boldsymbol{\Delta}}_y\boldsymbol{\beta}|, \qquad (6.1)$$

where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix of $\mathbf{X}$, $\hat{\boldsymbol{\Delta}}_y$ is the sample covariance matrix of the data within slice y, and the maximization is over the Grassmann manifold $\mathcal{G}_{(p,d)}$.

The objective function $L_d(\boldsymbol{\beta}; h)$ is derived by partially maximizing the log-likelihood function $\ell(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y, \boldsymbol{\beta}; \mathbf{X}_1, \ldots, \mathbf{X}_h) = \sum_{y=1}^{h}\ell(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y, \boldsymbol{\beta}; \mathbf{X}_y)$ over the nuisance parameters $\boldsymbol{\mu}_y$ and $\boldsymbol{\Delta}_y$, $y = 1, \ldots, h$. By the definition of minimum discrepancy function (Shapiro 1986), the likelihood-based objective function $L_d(\boldsymbol{\beta}; h)$ can be reparameterized into a minimum discrepancy function (Cook and Forzani 2009, appendix. A.5). Analogous to FIRE and DIRE, we consider fusing a set of slicing schemes $H = \{h_1, \ldots, h_K\}$ via the full likelihood of $\mathbf{X}_{yj}$, $y = 1, \ldots, h_j$ and $j = 1, \ldots K$ and the composite likelihoods.

From the discussion in Section 3.2, it is not surprising to find that the full likelihood approach leads to the same objective function as LAD with the fused slicing scheme. However, unlike FIRE or IRE that uses $\hat{\boldsymbol{\zeta}}_F$ in its objective function, LAD objective function (6.1) depends largely on sample covariance matrices $\tilde{\boldsymbol{\Delta}}_y$ of each slice, which involves many more parameters. And ensuring well-conditioned $\tilde{\boldsymbol{\Delta}}_y$'s require a reasonable number of observations per slice. Hence, the full likelihood approach of fusing LAD can be impractical.

Fortunately, we can adapt the idea of DIRE in the likelihood-based SDR context. By treating the slicing schemes independently, we write the fused objective function as $L_d^{\mathrm{FLAD}}(\boldsymbol{\beta}; H) \equiv \sum_{j=1}^{K} L_d(\boldsymbol{\beta}; h_j)$, where $L_d(\boldsymbol{\beta}; h_j)$ is the LAD objective function defined in (6.1). This is exactly the composite likelihood approach proposed by Lindsay (1988). See Varin, Reid, and Firth (2011) for an overview of recent development in the theory and application of composite likelihoods. More specifically, our fusing method corresponds to the block composite log-likelihood, where the whole set of variables $\{\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_{h_1}^{(1)}, \ldots, \mathbf{X}_{h_K}^{(K)}\}$ was divided into K subsets and the block composite log-likelihood was the sum of the log-likelihoods for every subset of variables.

The $\sqrt{n}$-consistency of the block-composite-fused LAD then follows from the theory of composite likelihoods. The simulation results for fused LAD are very encouraging, as illustrated in the supplementary materials.

7. FUSING TO ESTIMATE THE CENTRAL SOLUTION SPACE

Li and Dong (2009) introduced the notion of the central solution subspace (CSS) and developed methods based on CSS that relax the linearity condition and allow nonlinear conditional means $\mathrm{E}(\mathbf{X}|\boldsymbol{\beta}^T\mathbf{X})$.

We adapted the idea of DIRE to one of the CSS methods, CSS-SIR, which is based on a quadratic objective function that depends on slicing (see Li and Dong 2009 for details). For a fixed number of slices h, we write the objective function of CSS-SIR as $L_h(\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ represents a basis for the CSS. We define the fused objective function over a slicing set $H = \{h_1, \ldots, h_K\}$ as $L^{\mathrm{F}}(\boldsymbol{\eta}) = \sum_{j=1}^{K} L_{h_j}(\boldsymbol{\eta})$. The fused estimator $\hat{\boldsymbol{\eta}}^{\mathrm{F}}$ is obtained by minimizing this objective function. The $\sqrt{n}$-consistency of $\hat{\boldsymbol{\eta}}^{\mathrm{F}}$ follows from Li and Dong (2009, Corollary 6.1).

To illustrate the performance of this fused estimator, we conducted simulation studies based on the three simulation models used by Li and Dong (2009). The results given in the supplement show that CSS-SIR can be quite sensitive to the number of slices. Moreover, the fused estimators accumulate information rapidly and have little variability for all $10 \le h_K \le 20$. Fusing provided significant improvement over the original CSS-SIR method in all three models. The gain from the fused estimator is the greatest in the most challenging model among the three models, suggesting that fusing is particularly needed when finding the CSS is not easy.

8. DISCUSSION

We proposed a practically useful solution to the important, long-standing problem of choosing a slicing scheme in SDR methodology. FIRE is the asymptotically optimal methods but, owing to the cost of estimating the inner product matrix, a large sample size may be needed to manifest this property. DIRE may generally be the method of choice. Although it is not asymptotically optimal, its performance was better than that of IRE unless the sample size was quite large and it overshadowed all other competitors.

An alternative approach is to directly obtain a regularized estimate of the inner product matrix $\boldsymbol{\Gamma}_F^{-1}$. For instance, the sparse permutation invariant covariance estimator (Rothman et al. 2008) might offer improvement over DIRE, which uses only the block diagonal parts of $\hat{\boldsymbol{\Gamma}}_F$.

Following Cook and Ni (2005) and Ni and Cook (2007), we could build the theoretical foundation of tests for predictor effects and the CS dimensionality. Let $\widehat{\mathcal{F}}_d(\mathbf{B}, \mathbf{C}; \boldsymbol{\Psi}) = \mathcal{F}_d(\hat{\mathbf{B}}, \hat{\mathbf{C}}; \boldsymbol{\Psi})$ denote the minimized objective function value for the fused estimators with inner product matrix $\boldsymbol{\Psi}$. Then under the same condition as Theorem 2, $n\widehat{\mathcal{F}}_d(\mathbf{B}, \mathbf{C}; \boldsymbol{\Gamma}_F^{-1})$ and $n\widehat{\mathcal{F}}_d(\mathbf{B}, \mathbf{C}; \mathbf{G}_F^{-1})$ both follow an asymptotic χ^2-squared distribution with degree of freedom $(p - d)(\sum_{j=1}^{K}(h_j - 1) - d)$. Then the asymptotic χ^2-tests follow straightforwardly.

Recently, Ma and Zhu (2012) developed semiparametric SDR methods that do not require the linearity condition. It is also possible to develop fused semiparametric methods to relax linearity condition. The methods proposed by Ma and Zhu (2012) require nonparametric estimation of conditional moments, where the kernel bandwidth selection problem can be alleviated by fused estimator.

SUPPLEMENTARY MATERIALS

The supplementary materials contain detailed proofs and additional simulations.

[Received January 2012. Revised October 2013.]

REFERENCES

Bates, J. M., and Granger, C. W. J. (1969), "The Combination of Forecasts," *Operations Research Quarterly*, 20, 451–468. [816]

Chiaromonte, F., Cook, R. D., and Li, B. (2002), "Sufficient Dimension Reduction in Regressions with Categorical Predictors," *The Annals of Statistics*, 30, 475–497. [815]

Cook, R. D. (1998), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: Wiley. [815,816]

Cook, R. D., and Forzani, L. (2008), "Principal Fitted Components for Dimension Reduction in Regression," *Statistical Science*, 23, 485–501. [821]

——— (2009), "Likelihood-Based Sufficient Dimension Reduction," *Journal of the American Statistical Association*, 104, 197–208. [815,819,824]

Cook, R. D., Forzani, L., and Rothman, A. J. (2012), "Estimating Sufficient Reductions of the Predictors in Abundant High-Dimensional Regressions," *The Annals of Statistics*, 40, 353–384. [820]

Cook, R. D., and Ni, L. (2005), "Sufficient Dimension Reduction Via Inverse Regression: A Minimum Discrepancy Approach," *Journal of the American Statistical Association*, 100, 410–428. [815,816,817,818,820,826]

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. [815,820,824]

Cook, R. D., and Yin, X. (2001), "Dimension Reduction and Visualization in Discriminant Analysis" (with discussion), *Australia/New Zealand Journal of Statistics*, 43, 147–199. [823]

Dong, Y., and Li, B. (2010), "Dimension Reduction for Nonelliptically Distributed Predictors: Second Order Methods," *Biometrika*, 97, 279–294. [815]

Hall, P., and Li, K. C. (1993), "On Almost Linearity of Low Dimensional Projection From High Dimensional Data," *The Annals of Statistics*, 21, 867–889. [816]

Hooper, J. (1959), "Simultaneous Equations and Canonical Correlation Theory," *Econometrica*, 27, 245–256. [820]

Hotelling, H. (1936), "Relations Between Two Sets of Variates," *Biometrika*, 28, 321–377. [820]

Hsing, T., and Carroll, R. J. (1992), "An Asymptotic Theory for Sliced Inverse Regression," *The Annals of Statistics*, 20, 1040–1061. [816]

Li, B., and Dong, Y. (2009), "Dimension Reduction for Nonelliptically Distributed Predictors," *The Annals of Statistics*, 37, 1272–1298. [815,825]

Li, B., and Wang, S. (2007), "On Directional Regression for Dimension Reduction," *Journal of the American Statistical Association*, 102, 997–1008. [815,816,820,824]

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–342. [815,816]

Lindsay, B. (1988), "Composite Likelihood Methods," *Contemporary Mathematics*, 80, 220–239. [819,825]

Luceno, A. (1999), "Discrete Approximations to Continuous Univariate Distributions–and Alternative to Simulation," *Journal of the Royal Statistical Society*, Series B, 61, 345–352. [819]

Ma, Y., and Zhu, L. (2012), "A Semiparametric Approach to Dimension Reduction," *Journal of the American Statistical Association*, 107, 168–179. [826]

Ni, L., and Cook, R. D. (2007), "A Robust Inverse Regression Estimator," *Statistics and Probability Letters*, 77, 343–349. [819,823,826]

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515. [826]

Setodji, C., and Cook, R. D. (2004), "K-Means Inverse Regression," *Technometrics*, 46, 421–429. [815]

Shapiro, A. (1986), "Asymptotic Theory of Overparameterized Structural Models," *Journal of the American Statistical Association*, 81, 142–149. [824]

Timmermann, A. G. (2006), "Forecast Combinations," in *Handbook of Economic Forecasting*, eds. G. Elliott, C. W. J. Granger, and A. Timmerman, Amsterdam: North-Holland. [816]

Varin, C., Reid, N., and Firth, D. (2011), "An Overview of Composite Likelihood Methods," *Statistica Sinica*, 21, 5–42. [825]

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. [815,823,824]

Zhu, L., and Ng, K. W. (1995), "Asymptotics of Sliced Inverse Regression," *Statistica Sinica*, 5, 727–736. [816,823]

Zhu, L., Zhu, L., and Feng, Z. (2010), "Dimension Reduction in Regressions Through Cumulative Slicing Estimation," *Journal of the American Statistical Association*, 105, 1455–1466. [816,824]