Statistica Sinica **32** (2022), 499-516
doi:https://doi.org/10.5705/ss.202019.0102

DIMENSION REDUCTION VIA ADAPTIVE SLICING

Tao Wang

Shanghai Jiao Tong University

Abstract: Sufficient dimension reduction often resorts to inverse regression, and most inverse regression methods rely on slicing a quantitative response. The choice of a particular slicing scheme is critical, but there are no current methods in the literature about how to select an optimal slicing scheme. We consider two popular slicing-based methods, namely, the sliced inverse regression and the sliced average variance estimation. By recasting the eigen-decomposition problem as a trace-optimization problem, we propose a penalized criterion for choosing an optimal slicing scheme. A dynamic programming algorithm is developed for numerical optimization. The theoretical properties are studied under mild conditions. Simulation examples show that our methods compare favorably with existing methods. An illustrative data analysis is also presented.

Key words and phrases: Nonlinear least squares, quantile slicing, optimal number of slices, SAVE, SIR, trace maximization.

1. Introduction

A fundamental concept in regression is dimension reduction, used to reduce the dimension of the predictor space without losing information on the regression (Cook (2007)). Many different contexts have been developed to achieve this. Among these, sufficient dimension reduction has received considerable interest in the past two decades (Cook (1998)). Consider the regression of a univariate response $Y \in \mathbb{R}$ on a p-dimensional predictor vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$. In full generality, sufficient dimension reduction seeks a set of linear combinations of $\boldsymbol{X}$, such that the conditional distribution of Y given $\boldsymbol{X}$ depends on $\boldsymbol{X}$ only through these linear combinations. More formally, if Y and $\boldsymbol{X}$ are independent given $\boldsymbol{\beta}^\top \boldsymbol{X}$, where $\boldsymbol{\beta}$ is a $p \times d$ matrix with $d \leq p$, then the column space of $\boldsymbol{\beta}$ is called a dimension-reduction subspace. Under mild assumptions, the intersection of all dimension-reduction subspaces is also a dimension-reduction subspace; in this case, it is called the central subspace for the regression of Y on $\boldsymbol{X}$, and is denoted by $\mathcal{S}_{Y|\boldsymbol{X}}$ (Cook (1998)).

Methods for estimating $\mathcal{S}_{Y|\boldsymbol{X}}$ include the sliced inverse regression (Li (1991)),

Corresponding author: Tao Wang, Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: neowangtao@sjtu.edu.cn.

sliced average variance estimation (Cook and Weisberg (1991)), minimum average variance estimation (Xia et al. (2002)), minimum discrepancy estimation (Cook and Ni (2005)), directional regression (Li and Wang (2007)), likelihood acquired directions (Cook and Forzani (2009)), and semiparametric estimation (Ma and Zhu (2012)); see Ma and Zhu (2013) for a review. Among these methods, perhaps the most widely used are the inverse regression methods, and in particular the sliced inverse regression (SIR) and sliced average variance estimation (SAVE). The inverse regression of $\boldsymbol{X}$ on Y, or $\boldsymbol{X} \mid Y$ for short, is composed of p regressions, $X_j \mid Y$, for $j = 1, \ldots, p$. Because Y is one dimensional, an inverse regression avoids the curse of dimensionality. In this study, we are concerned only with the SIR and SAVE, both of which rely on inverse conditional moments. See Section 2 for more details.

To estimate p inverse regressions, we can either use a smooth nonparametric method, such as a kernel regression (Zhu and Fang (1996)), or fit parametric curves using a linear regression (Bura and Cook (2001)). However, the usual routines for computing the SIR and SAVE use a simple nonsmooth nonparametric procedure introduced by Li (1991): partition the range of Y into a few slices, and compute the sample moments of $\boldsymbol{X}$ in each slice. We call this procedure slicing. Similar to the bandwidth in kernel smoothing, the slicing scheme is a tuning parameter that needs to be determined from the data. When Y is continuous, it is more convenient to use quantile slicing, that is, to slice the response according to its quantiles. Then, the choice of the number of slices is critical. To the best of our knowledge, there are no current methods in the literature for selecting the number of slices in quantile slicing, or the slicing scheme in general, which remains an open problem (Zhu et al. (2010)).

The performance of the SIR has been empirically observed to be robust to the choice of the number of slices. Zhu and Ng (1995) showed theoretically that the SIR estimator is $\sqrt{n}$-consistent, provided that the number of slices is between $\sqrt{n}$ and $n/2$, where n is the sample size. This is not true for the SAVE. Numerically, the SAVE is more sensitive to the number of slices than is the SIR (Zhu, Ohtaki and Li (2007)). Furthermore, it can be inconsistent when the number of observations in each slice is fixed and does not depend on n (Li and Zhu (2007)). As such, methods for adaptively choosing a slicing scheme are highly demanding.

To address the slicing problem, Zhu, Zhu and Feng (2010) proposed a cumulative slicing estimation. Similarly to the SIR and SAVE, they proposed a cumulative mean estimation and a cumulative variance estimation. The basic idea of a cumulative slicing estimation is to pool the collection of estimates of $\mathcal{S}_{Y|\boldsymbol{X}}$ from all possible slicing schemes with two slices. Rather than select the optimal number of slices, their method sidesteps the problem. More recently, Cook and Zhang (2014) developed a class of fused estimators. Like the cumulative slicing estimation, the general methodology can be applied to all dimension-reduction methods that rely on slicing a quantitative response. Two special cases are the fused SIR and the fused SAVE. However, fused estimators are not fully slicing-free, in the sense that the fusion is over a predefined set of slicing schemes. Consequently, if we adopt quantile slicing, then the number of slicing schemes has to be specified for each dimension-reduction method. However, the effect of this hyperparameter has not been studied systematically. Despite these advances, the problem of choosing an optimal slicing scheme remains.

In this study, we focus directly on the slicing problem and propose a practically useful solution. In Section 2, we review the SIR and SAVE in the usual dimension-reduction framework. In Section 3, we re-derive the SIR and SAVE estimates using a trace maximization principle. In Section 4.1, we propose a penalized criterion for selecting an optimal slicing scheme. An efficient algorithm is developed for numerical optimization in Section 4.2, and the theoretical properties of our methods are studied in Section 4.3. In Section 5, we compare the performance of our methods with that of existing methods by simulation. An illustrative data analysis is presented in Sections 6. We include a concluding discussion in Section 7. All proofs are given in the Supplementary Material.

For a matrix $\mathbf{M}$, span($\mathbf{M}$) denotes the subspace spanned by the columns of $\mathbf{M}$, and vec($\mathbf{M}$) is the operator that constructs a vector from $\mathbf{M}$ by stacking its columns. If $\mathbf{M}$ is a square matrix, trace($\mathbf{M}$) denotes the trace of $\mathbf{M}$. An identity matrix is denoted by $\mathbf{I}$ or $\mathbf{I}_p$, when it is necessary to indicate the order. A semi-orthogonal matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$, for $q < p$, has orthogonal columns; that is, $\mathbf{A}^\top\mathbf{A} = \mathbf{I}_q$.

2. Review of the SIR and SAVE

In keeping with the usual dimension-reduction protocol, we assume for now that the response Y has been discretized by constructing G slices. We continue to use Y to denote the sliced version with support $\{1, \ldots, G\}$. We also assume that an independent and identically distributed sample $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ from the joint distribution of $(\boldsymbol{X}, Y)$ is available.

The following two assumptions are common in sufficient dimension reduction: (C1) $\mathrm{E}(\boldsymbol{X} \mid \boldsymbol{\beta}^\top\boldsymbol{X})$ is a linear function of $\boldsymbol{\beta}^\top\boldsymbol{X}$, and (C2) $\mathrm{Cov}(\boldsymbol{X} \mid \boldsymbol{\beta}^\top\boldsymbol{X})$ is constant, where the columns of the matrix $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ form a basis for $\mathcal{S}_{Y|\boldsymbol{X}}$. Both conditions apply to the marginal distribution of $\boldsymbol{X}$, and not to the conditional distribution of Y given $\boldsymbol{X}$, and are widely regarded as mild. See Li and Wang (2007) for a discussion.

For ease of exposition, we often work in terms of the standardized predictor

$$\boldsymbol{Z} = \{\mathrm{Cov}(\boldsymbol{X})\}^{-1/2}\{\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\},$$

with the sample version given by $\boldsymbol{z}_i = \mathbf{S}^{-1/2}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$, where $\bar{\boldsymbol{x}} = \sum_{i=1}^n \boldsymbol{x}_i/n$ is the sample mean of $\boldsymbol{x}_i$, and $\mathbf{S} = \sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top/n$ is the sample covariance matrix. This involves no loss of generality, because $\mathcal{S}_{Y|\boldsymbol{X}} = \{\mathrm{Cov}(\boldsymbol{X})\}^{-1/2}\mathcal{S}_{Y|\boldsymbol{Z}}$ (Cook (1998)).

The SIR, SAVE, and many other methods for estimating $\mathcal{S}_{Y|\boldsymbol{Z}}$ are based on the following general procedure. Suppose $\mathbf{M} \in \mathbb{R}^{p\times p}$ is a kernel matrix with the property that $\mathrm{span}(\mathbf{M}) \subseteq \mathcal{S}_{Y|\boldsymbol{Z}}$, and $\hat{\mathbf{M}}$ is a consistent estimate of $\mathbf{M}$. Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ be the eigenvalues of $\hat{\mathbf{M}}$, and let $\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2, \ldots, \hat{\boldsymbol{\eta}}_p$ be the corresponding eigenvectors. We use $\mathrm{span}(\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_d)$ to estimate $\mathcal{S}_{Y|\boldsymbol{Z}}$. We then use $\mathrm{span}\{\mathbf{S}^{-1/2}(\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_d)\}$ to estimate $\mathcal{S}_{Y|\boldsymbol{X}}$.

The SIR is based on a fundamental result by Li (1991): if condition (C1) holds, then the conditional mean $\mathrm{E}(\boldsymbol{Z} \mid Y) \in \mathcal{S}_{Y|\boldsymbol{Z}}$. Then, $\mathrm{span}(\mathbf{M}_{SIR}) \subseteq \mathcal{S}_{Y|\boldsymbol{Z}}$, where $\mathbf{M}_{SIR} = \mathrm{Cov}\{\mathrm{E}(\boldsymbol{Z} \mid Y)\}$ is the SIR kernel matrix.

The SAVE uses the conditional variance $\mathrm{Cov}(\boldsymbol{Z} \mid Y)$. Define the SAVE kernel matrix as $\mathbf{M}_{SAVE} = \mathrm{E}[\{\mathbf{I}_p - \mathrm{Cov}(\boldsymbol{Z} \mid Y)\}^2]$. Given conditions (C1) and (C2), the column space of $\mathbf{M}_{SAVE}$ is contained in $\mathcal{S}_{Y|\boldsymbol{Z}}$ (Cook and Weisberg (1991)).

It is well known that $\mathrm{span}(\mathbf{M}_{SIR}) \subseteq \mathrm{span}(\mathbf{M}_{SAVE})$. Specifically, let $\boldsymbol{\mu}_g = \mathrm{E}(\boldsymbol{Z} \mid Y = g)$ and $\boldsymbol{\Sigma}_g = \mathrm{Cov}(\boldsymbol{Z} \mid Y = g)$. Then, $\mathrm{span}(\mathbf{M}_{SIR}) = \mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G)$, and by Proposition 6 of Cook and Critchley (2000),

$$\mathrm{span}(\mathbf{M}_{SAVE}) = \mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G - \boldsymbol{\Sigma}_{G-1}).$$

Let $n_g = \sum_{i=1}^n I(y_i = g)$, where $I(\cdot)$ is the indicator function. Let $\hat{\boldsymbol{\mu}}_g = \sum_{i:y_i=g} \boldsymbol{z}_i/n_g$ and $\hat{\boldsymbol{\Sigma}}_g = \sum_{i:y_i=g}(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_g)(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_g)^\top/n_g$. We estimate $\mathbf{M}_{SIR}$ and $\mathbf{M}_{SAVE}$ by

$$\hat{\mathbf{M}}_{SIR} = \sum_{g=1}^G \frac{n_g}{n}\hat{\boldsymbol{\mu}}_g\hat{\boldsymbol{\mu}}_g^\top$$

and

$$\hat{\mathbf{M}}_{SAVE} = \sum_{g=1}^G \frac{n_g}{n}(\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_g)^2,$$

respectively.

3. The Trace-Maximization Principle

Instead of computing the eigen decomposition of a kernel matrix, we can solve a trace-optimization problem (Chen, Zou and Cook (2010)).

Lemma 1. *Let* $\mathbf{A} \in \mathbb{R}^{p\times p}$ *be a symmetric matrix, and* $\mathbf{B}$ *be a* $p \times d$ *semi-orthogonal matrix. Denote by* $\boldsymbol{\eta}_1(\mathbf{A}), \ldots, \boldsymbol{\eta}_p(\mathbf{A})$ *the eigenvectors of* $\mathbf{A}$*, ordered from the largest to the smallest eigenvalue* $\lambda_j(\mathbf{A})$*. We have trace*$(\mathbf{B}^\top\mathbf{A}\mathbf{B}) \leq \sum_{j=1}^d \lambda_j(\mathbf{A})$*, with equality if and only if* $\mathbf{B} = [\boldsymbol{\eta}_1(\mathbf{A}), \ldots, \boldsymbol{\eta}_d(\mathbf{A})]\mathbf{U}$*, where* $\mathbf{U}$ *is any* $d \times d$ *orthogonal matrix.*

From this lemma, the SIR is equivalent to the criterion

$$\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top\boldsymbol{\alpha}=\mathbf{I}_d} \text{trace}(\boldsymbol{\alpha}^\top\hat{\mathbf{M}}_{SIR}\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top\boldsymbol{\alpha}=\mathbf{I}_d} \sum_{g=1}^G \frac{n_g}{n}\text{trace}(\boldsymbol{\alpha}^\top\hat{\boldsymbol{\mu}}_g\hat{\boldsymbol{\mu}}_g^\top\boldsymbol{\alpha}),$$

and the SAVE is equivalent to the criterion

$$\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top\boldsymbol{\alpha}=\mathbf{I}_d} \text{trace}(\boldsymbol{\alpha}^\top\hat{\mathbf{M}}_{SAVE}\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top\boldsymbol{\alpha}=\mathbf{I}_d} \sum_{g=1}^G \frac{n_g}{n}\text{trace}\{\boldsymbol{\alpha}^\top(\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_g)^2\boldsymbol{\alpha}\}.$$

We can also obtain the SIR or SAVE estimate as the solution to a nonlinear least squares problem (Cook and Ni (2005)). For any $\mathbf{B} \in \mathbb{R}^{p\times d}$, define

$$L_{SIR}(\mathbf{B}, \mathbf{C}) = \sum_{g=1}^G \frac{n_g}{n}\|\hat{\boldsymbol{\mu}}_g - \mathbf{B}\mathbf{C}_g\|_2^2,$$

where $\mathbf{C}_g \in \mathbb{R}^{d\times 1}$ and $\mathbf{C} = (\mathbf{C}_1, \ldots, \mathbf{C}_G)$. In addition, define

$$L_{SAVE}(\mathbf{B}, \mathbf{F}) = \sum_{g=1}^G \frac{n_g}{n}\|\text{vec}(\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_g) - \text{vec}(\mathbf{B}\mathbf{F}_g)\|_2^2,$$

where $\mathbf{F}_g \in \mathbb{R}^{d\times p}$ and $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_G)$.

For fixed $\mathbf{B}$, let $\hat{\mathbf{C}}_\mathbf{B}$ be the value of $\mathbf{C}$ that minimizes $L_{SIR}(\mathbf{B}, \mathbf{C})$, and let $\hat{\mathbf{F}}_\mathbf{B}$ be the value of $\mathbf{F}$ that minimizes $L_{SAVE}(\mathbf{B}, \mathbf{F})$. Let $\mathcal{G}_{p,d} = \{\mathbf{B} \in \mathbb{R}^{p\times d} : \mathbf{B}^\top\mathbf{B} = \mathbf{I}_d\}$. The following proposition gives the connection between the least squares formulation and the trace optimization problem.

Proposition 1. *Minimizing* $L_{SIR}(\mathbf{B}, \hat{\mathbf{C}}_\mathbf{B})$ *over* $\mathbf{B} \in \mathcal{G}_{p,d}$ *is equivalent to maximizing trace*$(\boldsymbol{\alpha}^\top\hat{\mathbf{M}}_{SIR}\boldsymbol{\alpha})$ *over* $\boldsymbol{\alpha} \in \mathcal{G}_{p,d}$*. Furthermore, minimizing* $L_{SAVE}(\mathbf{B}, \hat{\mathbf{F}}_\mathbf{B})$ *over* $\mathbf{B} \in \mathcal{G}_{p,d}$ *is equivalent to maximizing trace*$(\boldsymbol{\alpha}^\top\hat{\mathbf{M}}_{SAVE}\boldsymbol{\alpha})$ *over* $\boldsymbol{\alpha} \in \mathcal{G}_{p,d}$*.*

Remarkably, the results in this section hold even when conditions (C1) and (C2) fail. Nevertheless, they may be of little practical importance if there is no useful connection between the subspace estimated by the SIR or the SAVE and the subspace we would like to estimate, namely, $\mathcal{S}_{Y|\boldsymbol{X}}$.

Under the normal model, we can check whether a subspace is a dimension-reduction subspace (Cook and Forzani (2009)). Let $\mathcal{S}_{SIR} = \{\mathrm{Cov}(\boldsymbol{X})\}^{-1/2}\mathrm{span}(\mathbf{M}_{SIR})$ be the SIR subspace in the $\boldsymbol{X}$-scale. Similarly, let $\mathcal{S}_{SAVE} = \{\mathrm{Cov}(\boldsymbol{X})\}^{-1/2}$ span$(\mathbf{M}_{SAVE})$.

Proposition 2. *Let $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$ be a semi-orthogonal matrix, and let $\boldsymbol{\eta}_0$ be an orthogonal complement of $\boldsymbol{\eta}$, such that $(\boldsymbol{\eta}, \boldsymbol{\eta}_0)$ is $p \times p$ orthogonal. Assume that $\boldsymbol{X} \mid (Y = g) \sim N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. Let $\boldsymbol{\theta} = \mathrm{E}(\boldsymbol{X})$ and $\boldsymbol{\Delta} = \mathrm{E}\{\mathrm{Cov}(\boldsymbol{X} \mid Y)\}$. If*

(i) $\boldsymbol{\eta}^\top \boldsymbol{X} \mid (Y = g) \sim N(\boldsymbol{\eta}^\top\boldsymbol{\theta} + \boldsymbol{\eta}^\top\boldsymbol{\Delta}\boldsymbol{\eta}\boldsymbol{v}_g, \boldsymbol{\eta}^\top\boldsymbol{\Sigma}_g\boldsymbol{\eta})$, for some $\boldsymbol{v}_g \in \mathbb{R}^d$, and

(ii) $\boldsymbol{\eta}_0^\top \boldsymbol{X} \mid (\boldsymbol{\eta}^\top\boldsymbol{X} = \boldsymbol{\eta}^\top\boldsymbol{x}, Y = g) \sim N\{\boldsymbol{\eta}_0^\top\boldsymbol{\theta} + \boldsymbol{\eta}_0^\top\boldsymbol{\Delta}\boldsymbol{\eta}(\boldsymbol{\eta}^\top\boldsymbol{\Delta}\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^\top(\boldsymbol{x} - \boldsymbol{\theta}), (\boldsymbol{\eta}_0^\top\boldsymbol{\Delta}^{-1}\boldsymbol{\eta}_0)^{-1}\}$,

for all $g \in \{1, \ldots, G\}$, then, span$(\boldsymbol{\eta}) = \mathcal{S}_{Y|\boldsymbol{X}} = \mathcal{S}_{SAVE}$. If, in addition, $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_G$, then span$(\boldsymbol{\eta}) = \mathcal{S}_{Y|\boldsymbol{X}} = \mathcal{S}_{SIR} = \mathcal{S}_{SAVE}$.

4. Dimension Reduction via Adaptive Slicing

The developments so far have been based on a fixed slicing scheme: the range of the response Y has been partitioned into G slices, indexed by $g = 1, \ldots, G$. In practice, the slicing scheme is an important tuning parameter, and the optimal slicing scheme should be chosen adaptively from the data.

For ease of discourse, in this section, we assume that the response Y has a finite support $\mathbb{Y} = \{1, \ldots, K\}$. There is no loss of generality implied by this restriction, because when Y is continuous, we can construct a discrete version $\tilde{Y}$ by dividing its range into K intervals, and it is known that $\mathcal{S}_{\tilde{Y}|\boldsymbol{X}} \subseteq \mathcal{S}_{Y|\boldsymbol{X}}$ with equality when K is sufficiently large.

Denote $\mathcal{S}$ as a generic slicing scheme and $|\mathcal{S}|$ as the cardinality of $\mathcal{S}$. Mathematically, we can write $\mathcal{S} = \{\mathcal{B}_g \subseteq \mathbb{Y}, g = 1, \ldots, |\mathcal{S}|\}$, where the subsets $\mathcal{B}_g$ satisfy $\cup_{g=1}^G \mathcal{B}_g = \mathbb{Y}$ and $\mathcal{B}_g \cap \mathcal{B}_{g'} = \emptyset$, for all $g \neq g'$. Without loss of generality, we assume that the slices in $\mathcal{S}$ are sorted: if $g < g'$, then $y < y'$, for any $y \in \mathcal{B}_g$ and $y' \in \mathcal{B}_{g'}$. We call $\mathcal{B}_g$ the gth slice of $\mathcal{S}$.

For each $k \in \mathbb{Y}$, let $n_k = \sum_{i=1}^n I(y_i = k)$ and $f_k = n_k/n$. For a generic slice $\mathcal{B} \subseteq \mathbb{Y}$, let $f_{\mathcal{B}} = \sum_{k\in\mathcal{B}} f_k$, $\hat{\boldsymbol{\mu}}_{\mathcal{B}} = \sum_{k\in\mathcal{B}}\sum_{i:y_i=k} \boldsymbol{z}_i / \sum_{k\in\mathcal{B}} n_k$, and $\hat{\boldsymbol{\Sigma}}_{\mathcal{B}} = \sum_{k\in\mathcal{B}}\sum_{i:y_i=k}(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_{\mathcal{B}})(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_{\mathcal{B}})^\top / \sum_{k\in\mathcal{B}} n_k$. To emphasize the dependence of a kernel matrix on the slicing scheme, we write $\hat{\mathbf{M}} = \hat{\mathbf{M}}(\mathcal{S})$. Then, the criteria become

$$\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top\boldsymbol{\alpha}=\mathbf{I}_d} \text{trace}\{\boldsymbol{\alpha}^\top\hat{\mathbf{M}}_{SIR}(\mathcal{S})\boldsymbol{\alpha}\} = \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top\boldsymbol{\alpha}=\mathbf{I}_d} \sum_{g=1}^{|\mathcal{S}|} f_{\mathcal{B}_g}\text{trace}(\boldsymbol{\alpha}^\top\hat{\boldsymbol{\mu}}_{\mathcal{B}_g}\hat{\boldsymbol{\mu}}_{\mathcal{B}_g}^\top\boldsymbol{\alpha})$$

and

$$\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top\boldsymbol{\alpha}=\mathbf{I}_d} \text{trace}\{\boldsymbol{\alpha}^\top\hat{\mathbf{M}}_{SAVE}(\mathcal{S})\boldsymbol{\alpha}\} = \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top\boldsymbol{\alpha}=\mathbf{I}_d} \sum_{g=1}^{|\mathcal{S}|} f_{\mathcal{B}_g}\text{trace}\{\boldsymbol{\alpha}^\top(\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_g})^2\boldsymbol{\alpha}\}.$$

4.1. Penalized trace maximization

A key ingredient of a slicing scheme, $\mathcal{S}$, is the number of slices, $|\mathcal{S}|$. If $|\mathcal{S}|$ is smaller than d, the dimension of $\mathcal{S}_{Y|\boldsymbol{Z}}$, then all methods will miss some directions. On the other hand, if we partition the range of the response into too many slices, the accuracy of the intra-slice estimates can suffer. To select an optimal slicing scheme, we consider the penalized trace optimization problem

$$\max_{(\boldsymbol{\alpha},\mathcal{S}):\boldsymbol{\alpha}^\top\boldsymbol{\alpha}=\mathbf{I}_d} \left[\text{trace}\{\boldsymbol{\alpha}^\top\hat{\mathbf{M}}(\mathcal{S})\boldsymbol{\alpha}\} - \frac{\log(n)}{n} \times \text{df}_0 \times |\mathcal{S}|\right], \tag{4.1}$$

where df_0, to be specified, is a complexity factor for introducing an additional slice. Given $\boldsymbol{\alpha}$, this amounts to using the Bayesian information criterion (BIC) to choose a slicing scheme (Schwarz (1978)).

We motivate the penalty term as follows. In addition to the number of slices, a slicing scheme must consider the arrangement of the slices. Following Jiang, Ye and Liu (2015), we assign a prior on the slicing scheme, and then penalize the trace using this prior. Specifically, we assume that $|\mathcal{S}| - 1$ follows a Poisson distribution, with rate parameter $\exp(-\tau_n)$, and that given the partition size $|\mathcal{S}|$, the conditional distribution on the slice widths (normalized to sum to one) is Dirichlet$(1, \ldots, 1)$. Then, a maximum a posteriori estimation results in the penalty term $\tau_n \times (|\mathcal{S}| - 1)$, and setting $\tau_n = \log(n) \times \text{df}_0$ gives the BIC. Penalized test statistics of this form have been studied recently in the K-sample problem and in the independence problem; see Jiang, Ye and Liu (2015) and Heller et al. (2016) for details.

4.2. Algorithms

We can solve (4.1) using an alternating optimization procedure: fix $\mathcal{S}$ and estimate $\boldsymbol{\alpha}$, then fix $\boldsymbol{\alpha}$ and estimate $\mathcal{S}$, and iterate between these two steps until the algorithm converges.

We start by specifying the complexity factor df_0. Assume, for the moment, that $\mathcal{S}$ is given and $\mathcal{B}$ is a slice in $\mathcal{S}$. From Section 2, we know the SIR uses slice means $\boldsymbol{\mu}_{\mathcal{B}} = \mathrm{E}(\boldsymbol{Z} \mid Y \in \mathcal{B})$, and the SAVE uses slice means and slice covariances $\boldsymbol{\Sigma}_{\mathcal{B}} = \mathrm{Cov}(\boldsymbol{Z} \mid Y \in \mathcal{B})$. Under condition (C1), we have $\boldsymbol{\mu}_{\mathcal{B}} \in \mathcal{S}_{Y|\boldsymbol{Z}}$, or equivalently, $\boldsymbol{\mu}_{\mathcal{B}} = \boldsymbol{\eta}\boldsymbol{v}_{\mathcal{B}}$, for some $\boldsymbol{v}_{\mathcal{B}} \in \mathbb{R}^d$. Here, $\boldsymbol{\eta}$ is a basis matrix for $\mathcal{S}_{Y|\boldsymbol{Z}}$. On the other hand, if both conditions (C1) and (C2) hold, then $\mathbf{I}_p - \boldsymbol{\Sigma}_{\mathcal{B}} \in \mathcal{S}_{Y|\boldsymbol{Z}}$, or equivalently, $\mathbf{I}_p - \boldsymbol{\Sigma}_{\mathcal{B}} = \boldsymbol{\eta}\mathbf{A}_{\mathcal{B}}\boldsymbol{\eta}^\top$, where $\mathbf{A}_{\mathcal{B}}$ is a $d \times d$ symmetric matrix. Now, if the number of slices is incremented by one, then the number of free parameters is incremented by $\mathrm{df}_0 = d$ for the SIR, and by $\mathrm{df}_0 = d + d(d+1)/2$ for the SAVE.

We treat the SIR and SAVE separately. For the SIR, the corresponding problem is

$$\max_{(\boldsymbol{\alpha},\mathcal{S}):\boldsymbol{\alpha}^\top\boldsymbol{\alpha}=\mathbf{I}_d} \left\{ \sum_{g=1}^{|\mathcal{S}|} f_{\mathcal{B}_g} \mathrm{trace}(\boldsymbol{\alpha}^\top \hat{\boldsymbol{\mu}}_{\mathcal{B}_g} \hat{\boldsymbol{\mu}}_{\mathcal{B}_g}^\top \boldsymbol{\alpha}) - \frac{\log(n)}{n} d|\mathcal{S}| \right\}. \tag{4.2}$$

The optimization procedure is outlined in Algorithm 1. For the SAVE, the problem becomes

$$\max_{(\boldsymbol{\alpha},\mathcal{S}):\boldsymbol{\alpha}^\top\boldsymbol{\alpha}=\mathbf{I}_d} \left[\sum_{g=1}^{|\mathcal{S}|} f_{\mathcal{B}_g} \mathrm{trace}\{\boldsymbol{\alpha}^\top (\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_g})^2 \boldsymbol{\alpha}\} - \frac{\log(n)}{n} \frac{d(d+3)}{2} |\mathcal{S}| \right]. \tag{4.3}$$

The optimization procedure is outlined in Algorithm 2.

Algorithm 1 SIR with Adaptive Slicing (SIR-AS).

```
1: Take an initial guess for $\boldsymbol{\alpha}$, for example, a SIR with a quantile slicing scheme.
2: Adaptive slicing. Given $\boldsymbol{\alpha}$, compute the optimal slicing scheme $\{\mathcal{B}_g\}$ using the adaptive slicing algorithm (Algorithm 3).
3: Given $\{\mathcal{B}_g\}$, compute the SIR estimate of $\boldsymbol{\alpha}$.
4: Iterate steps 2 and 3 until convergence.
```

Algorithm 2 SAVE with Adaptive Slicing (SAVE-AS).

```
1: Take an initial guess for $\boldsymbol{\alpha}$, for example, a SAVE with a quantile slicing scheme.
2: Adaptive slicing. Given $\boldsymbol{\alpha}$, compute the optimal slicing scheme $\{\mathcal{B}_g\}$ using the adaptive slicing algorithm (Algorithm 4).
3: Given $\{\mathcal{B}_g\}$, compute the SAVE estimate of $\boldsymbol{\alpha}$.
4: Iterate steps 2 and 3 until convergence.
```

In both algorithms, the first step conducts dimension reduction (SIR or SAVE) on a fixed slicing scheme $\mathcal{S}$, and the second step uses a dynamic programming algorithm called *adaptive slicing* to find an optimal slicing scheme (see Algorithms 3 and 4). Note that adaptive slicing is a variant of the Viterbi algorithm, and is similar to the procedure of Jiang, Ye and Liu (2015), who investigated nonparametric K-sample testing from the perspective of inverse modeling. Viewing the K-sample testing problem as a test of independence between a continuous random variable and a categorical random variable, Jiang, Ye and Liu (2015) proposed a test statistic by slicing the continuous variable, deriving the likelihood ratio, and then including a term regularizing the number of slices; see Heller et al. (2016) for more on this idea. The computational complexity of the adaptive slicing algorithm is $O(n^2p)$ for the SIR, and $O(n^2p^2)$ for the SAVE. One way to speed up the algorithm is to pre-allocate observations into bins, and then to restrict the slicing to these bins.

Algorithm 3 Adaptive Slicing for SIR in the $\boldsymbol{Z}$-scale.

```
1: Rank the observed responses, and re-express the data as $(y_{(i)}, \boldsymbol{z}_{(i)}), i = 1, \ldots, n$. To
   ease notation, assume that the observations have been sorted; that is, $y_{(i)} = y_i$ and
   $\boldsymbol{z}_{(i)} = \boldsymbol{z}_i$.
2: For $i = 1, \ldots, n$ and $s = 1, \ldots, i$, compute
       $\hat{\boldsymbol{\mu}}^{(s:i)} = \frac{1}{i-s+1} \sum_{i'=s}^{i} \boldsymbol{z}_{i'}.$
3: Set $v_0 = 0$. Fill in entries of two vectors $(v_1, \ldots, v_n)^\top$ and $(s_1, \ldots, s_n)^\top$ recursively
   as follows:
       $v_i = \max_{s \in \{1,\ldots,i\}} \left\{ v_{s-1} + \frac{i-s+1}{n} \mathrm{trace}(\boldsymbol{\alpha}^\top \hat{\boldsymbol{\mu}}^{(s:i)} \hat{\boldsymbol{\mu}}^{(s:i)\top} \boldsymbol{\alpha}) - \frac{\log(n)}{n} d \right\},$
       $s_i = \mathrm{argmax}_{s \in \{1,\ldots,i\}} \left\{ v_{s-1} + \frac{i-s+1}{n} \mathrm{trace}(\boldsymbol{\alpha}^\top \hat{\boldsymbol{\mu}}^{(s:i)} \hat{\boldsymbol{\mu}}^{(s:i)\top} \boldsymbol{\alpha}) - \frac{\log(n)}{n} d \right\}.$
4: Trace back the vector $(s_1, \ldots, s_n)^\top$ as follows. Let $e_0 = n$. Compute $e_g = s_{e_{g-1}} - 1$
   recursively for $g \geq 1$ until $e_G = 0$, for some integer $G$. Then, the slicing scheme is
   given by $y_i \in \mathcal{B}_{G-g+1}$, for $e_g + 1 \leq i \leq e_{g-1}$ and $1 \leq g \leq G$, with $G$ the number of
   slices.
```

4.3. Theoretical properties

Before we can get started, we need a few definitions. We restrict our discussion to the inverse regression. Here, $\mathcal{S}$ is called an optimal slicing scheme in location if $\mathrm{E}(\boldsymbol{Z} \mid Y)$ takes $|\mathcal{S}|$ values and is constant within each slice. Furthermore, $\mathcal{S}$ is called an optimal slicing scheme in scale if $\mathrm{E}(\boldsymbol{Z} \mid Y)$ is constant, and $\mathrm{Cov}(\boldsymbol{Z} \mid Y)$ takes $|\mathcal{S}|$ values and is constant within each slice.

Throughout this section, we assume that the optimal slicing scheme, either

Algorithm 4 Adaptive Slicing for SAVE in the $\boldsymbol{Z}$-scale.

```
Rank the observed responses, and re-express the data as $(y_{(i)}, \boldsymbol{z}_{(i)}), i = 1, \ldots, n$. To ease notation, assume that the observations have been sorted; that is, $y_{(i)} = y_i$ and $\boldsymbol{z}_{(i)} = \boldsymbol{z}_i$.
For $i = 1, \ldots, n$ and $s = 1, \ldots, i$, compute
    $$\hat{\boldsymbol{\mu}}^{(s:i)} = \frac{1}{i-s+1} \sum_{i'=s}^{i} \boldsymbol{z}_{i'},$$
    and
    $$\hat{\boldsymbol{\Sigma}}^{(s:i)} = \frac{1}{i-s+1} \sum_{i'=s}^{i} (\boldsymbol{z}_{i'} - \hat{\boldsymbol{\mu}}^{(s:i)})(\boldsymbol{z}_{i'} - \hat{\boldsymbol{\mu}}^{(s:i)})^\top.$$
Set $\mathrm{df}_0 = d(d+3)/2$ and $v_0 = 0$. Fill in entries of two vectors $(v_1, \ldots, v_n)^\top$ and $(s_1, \ldots, s_n)^\top$ recursively as follows:
    $$v_i = \max_{s \in \{1,\ldots,i\}} \left[ v_{s-1} + \frac{i-s+1}{n} \mathrm{trace}\{\boldsymbol{\alpha}^\top (\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}^{(s:i)})^2 \boldsymbol{\alpha}\} - \frac{\log(n)}{n} \mathrm{df}_0 \right],$$
    $$s_i = \operatorname*{argmax}_{s \in \{1,\ldots,i\}} \left[ v_{s-1} + \frac{i-s+1}{n} \mathrm{trace}\{\boldsymbol{\alpha}^\top (\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}^{(s:i)})^2 \boldsymbol{\alpha}\} - \frac{\log(n)}{n} \mathrm{df}_0 \right].$$
Trace back the vector $(s_1, \ldots, s_n)^\top$ as follows. Let $e_0 = n$. Compute $e_g = s_{e_{g-1}} - 1$ recursively for $g \geq 1$ until $e_G = 0$, for some integer $G$. Then, the slicing scheme is given by $y_i \in \mathcal{B}_{G-g+1}$, for $e_g + 1 \leq i \leq e_{g-1}$ and $1 \leq g \leq G$, with $G$ the number of slices.
```

in location or in scale, exists, and is denoted by $\mathcal{S}_0$. For a continuous response, if $\boldsymbol{Z}$ depends on Y only through some latent slices of Y, then the optimal slicing scheme coincides with that latent structure (Cook and Zhang (2014)). As in the previous section, we assume that Y is discrete and has a finite support $\mathbb{Y} = \{1, \ldots, K\}$. In this case, the existence is guaranteed. Note that the above definition is not well defined for a forward regression. The adaptive slicing algorithm, however, is not restricted to an inverse regression.

Let $G_0 = |\mathcal{S}_0|$ and write $\mathcal{S}_0 = \{\mathcal{B}_{0g}, g = 1, \ldots, G_0\}$. Here, $\mathcal{S}$ is said to be over-slicing if it divides one or more slices in $\mathcal{S}_0$ into sub-slices; that is, $\mathcal{S}_0 \setminus \mathcal{S} \neq \emptyset$, and for each $\mathcal{B}_0 \in \mathcal{S}_0 \setminus \mathcal{S}$, there is a nontrivial partition $\mathcal{B}_0 = \cup_l \mathcal{B}_0^l$, such that $\mathcal{B}_0^l \in \mathcal{S}$, for all l. In addition, $\mathcal{S}$ is under-slicing if one slice contains elements from two or more slices in $\mathcal{S}_0$; that is, there exists a slice $\mathcal{B} \in \mathcal{S}$ and some $1 \leq g \leq G_0$, such that $\mathcal{B} \cap \mathcal{B}_{0g} \neq \emptyset$ and $\mathcal{B} \cap \mathcal{B}_{0(g+1)} \neq \emptyset$. According to whether $\mathcal{S}$ is over-slicing or under-slicing, we set $\mathfrak{S}_+ = \{\mathcal{S} : \mathcal{S} \text{ is over-slicing}\}$ and $\mathfrak{S}_- = \{\mathcal{S} : \mathcal{S} \text{ is under-slicing}\}$.

Define

$$\mathrm{BIC}_1(\mathcal{S};\boldsymbol{\alpha}) = \sum_{g=1}^{|\mathcal{S}|} f_{\mathcal{B}_g}\mathrm{trace}(\boldsymbol{\alpha}^\top \hat{\boldsymbol{\mu}}_{\mathcal{B}_g}\hat{\boldsymbol{\mu}}_{\mathcal{B}_g}^\top \boldsymbol{\alpha}) - \frac{\log(n)}{n} \times d \times |\mathcal{S}|$$

and

$$\mathrm{BIC}_2(\mathcal{S};\boldsymbol{\alpha}) = \sum_{g=1}^{|\mathcal{S}|} f_{\mathcal{B}_g}\mathrm{trace}\{\boldsymbol{\alpha}^\top (\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_g})^2 \boldsymbol{\alpha}\} - \frac{\log(n)}{n} \times \frac{d(d+3)}{2} \times |\mathcal{S}|.$$

Let $\hat{\mathcal{S}}_1(\boldsymbol{\alpha}) = \mathrm{argmax}_{\mathcal{S}}\mathrm{BIC}_1(\mathcal{S};\boldsymbol{\alpha})$ and $\hat{\mathcal{S}}_2(\boldsymbol{\alpha}) = \mathrm{argmax}_{\mathcal{S}}\mathrm{BIC}_2(\mathcal{S};\boldsymbol{\alpha})$. For each $k \in \mathbb{Y}$, let $\pi_k = P(Y = k)$. We have the following theorems.

Theorem 1. *Assume that (1) $\pi_k > 0$, for all $k \in \mathbb{Y}$, and (2) $\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}^\top = \boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^\top + O_p(n^{-1/2})$, where $\boldsymbol{\alpha}_0$ is a basis matrix for span$(\mathbf{M}_{SIR})$ and $\tilde{\boldsymbol{\alpha}}$ is an initial estimator of $\boldsymbol{\alpha}_0$. Then, as $n \to \infty$, $\hat{\mathcal{S}}_1(\tilde{\boldsymbol{\alpha}})$ converges in probability to $\mathcal{S}_0$, the optimal slicing scheme in location.*

Theorem 2. *Assume that (1) $\pi_k > 0$, for all $k \in \mathbb{Y}$, and (2) $\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}^\top = \boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^\top + O_p(n^{-1/2})$, where $\boldsymbol{\alpha}_0$ is a basis matrix for span$(\mathbf{M}_{SAVE})$ and $\tilde{\boldsymbol{\alpha}}$ is an initial estimator of $\boldsymbol{\alpha}_0$. Then, as $n \to \infty$, $\hat{\mathcal{S}}_2(\tilde{\boldsymbol{\alpha}})$ converges in probability to $\mathcal{S}_0$, the optimal slicing scheme in location or in scale.*

The SAVE and SIR are very different. At the population level, the SAVE is exhaustive under mild conditions (i.e., $\mathcal{S}_{SAVE} = \mathcal{S}_{Y|\boldsymbol{X}}$), but the SIR is not (Li and Wang (2007)). For a fixed slicing scheme, the asymptotic behavior of the SAVE differs from that of the SIR (Li and Zhu (2007)). For adaptive slicing, the difference remains. It is evident from the proof that the theory for the SAVE is more challenging, because it requires the optimal slicing scheme $\mathcal{S}_0$ be in location or in scale. In practice, $\mathcal{S}_0$ could be in both location and scale, that is, the collection of $\mathrm{E}(\boldsymbol{Z} \mid Y)$ and $\mathrm{Cov}(\boldsymbol{Z} \mid Y)$ takes $|\mathcal{S}|$ values and is constant within each slice. Although the SAVE is more comprehensive than the SIR, we are not able to give the general theory for the SAVE. A bias correction might be useful (Li and Zhu (2007)), but this is beyond the scope of this study. Finally, note that the consistency of the subspace estimation is guaranteed by the slicing consistency (Wang and Zhu (2015)). However, as noted by a referee, this consistency is quite different from that of a conventional estimation, because the selected slicing scheme is random rather than fixed. Nevertheless, our experience suggests that this randomness introduces some uncertainty into the estimates. It would be interesting to investigate the impact of adaptive slicing on the asymptotic distribution of the subspace estimator.

5. Simulation Results

We conducted simulation studies to evaluate the performance of the SIR-AS and SAVE-AS. We considered both inverse-regression and forward-regression models. To measure the closeness between $\mathcal{S}_{Y|\boldsymbol{X}}$ and its estimate, we used the vector correlation coefficient (Ye and Weiss (2003)). Let $\mathbf{B}$ and $\hat{\mathbf{B}}$ be basis matrices for the true and estimated subspaces, respectively. The vector correlation coefficient is defined as the positive square root of the product of the eigenvalues of $\hat{\mathbf{B}}^\top \mathbf{B}\mathbf{B}^\top \hat{\mathbf{B}}$. For each simulation example, we took $n = 400$ and $p = 10$, and tabulated the results over 200 replications. We treated the SIR-AS and SAVE-AS separately.

5.1. SIR-AS

Example 1. Inverse regression. We first simulated Y uniformly on the interval $[0, 5]$. Given $Y = y$, we then generated $\boldsymbol{X}$ from the model

$$\boldsymbol{X} = \boldsymbol{\beta}\mathbf{C}\boldsymbol{h}(y) + 0.5\boldsymbol{\varepsilon} + 0.3\boldsymbol{\beta}\epsilon, \tag{5.1}$$

where $\boldsymbol{\beta} = (1, 1, 0, \dots, 0)^\top \in \mathbb{R}^{p\times 1}, \mathbf{C} = (2, -2, \dots, 2, -2) \in \mathbb{R}^{1\times G_0}$, $\boldsymbol{h}(y) \in \mathbb{R}^{G_0 \times 1}$ is a vector of slice indicator functions, and $(\boldsymbol{\varepsilon}^\top, \epsilon)^\top \in \mathbb{R}^{p+1}$ is multivariate Gaussian with zero mean and an identity covariance matrix and is independent of Y. We set $G_0 = 10$ and constructed $\boldsymbol{h}$ using quantile slicing of the observed responses with G_0 slices. By Proposition 2, $\mathcal{S}_{Y|\boldsymbol{X}} = \mathcal{S}_{SIR} = \text{span}(\boldsymbol{\beta})$. In this example, there is an optimal slicing scheme in location: G_0 slices, with an equal number of observations in each slice.

Example 2. Forward regression. We first generated $\boldsymbol{X}$ from a multivariate Gaussian distribution with mean vector zero and covariance matrix $\boldsymbol{\Sigma} = (\Sigma_{ij})$, with $\Sigma_{ij} = 0.5^{|i-j|}$. We then generated Y according to the following model:

$$Y = \boldsymbol{\beta}_1^\top \boldsymbol{X}(\boldsymbol{\beta}_2^\top \boldsymbol{X} + 0.5) + 0.3\epsilon, \tag{5.2}$$

where $\boldsymbol{\beta}_1 = (1, 0, \dots, 0)^\top \in \mathbb{R}^{p\times 1}, \boldsymbol{\beta}_2 = (0, 1, 0, \dots, 0)^\top \in \mathbb{R}^{p\times 1}$, and ϵ is standard normal and is independent of $\boldsymbol{X}$. In this example, $\mathcal{S}_{Y|\boldsymbol{X}} = \mathcal{S}_{SIR} = \text{span}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. The optimal slicing scheme is not well defined.

In addition to the SIR-AS, we examined the performance of the original SIR of Li (1991), cumulative mean estimation (CUME) of Zhu, Zhu and Feng (2010), and fused sliced inverse regression (FSIR) of Cook and Zhang (2014). Whereas the SIR uses a single slicing scheme, the CUME and the FSIR extract information from multiple slicing schemes. The kernel matrix for the CUME is the sum of

Table 1. Means and standard deviations (in parentheses) of the vector correlation coefficient for the SIR-AS and its various competitors, based on 200 data applications, for Examples 1 and 2.

	SIR				FSIR			
Model	$G = 5$	$G = 10$	$G = 20$	CUME	$H = 10$	$H = 20$	$H = 30$	SIR-AS
(5.1)	0.010	0.979	0.978	0.226	0.794	0.906	0.916	0.979
	(0.008)	(0.008)	(0.009)	(0.075)	(0.078)	(0.038)	(0.032)	(0.008)
(5.2)	0.679	0.706	0.652	0.747	0.740	0.750	0.734	0.786
	(0.159)	(0.156)	(0.209)	(0.110)	(0.130)	(0.136)	(0.149)	(0.128)

the SIR kernel matrices from all slicing schemes with two slices. In this sense, the CUME is slicing-free. The kernel matrix for the FSIR is the sum of the SIR kernel matrices for a set of predefined slicing schemes. We used quantile slicing for the SIR and FSIR. To explore the sensitivity of the SIR to the number of slices G, and of the FSIR to the set of slice numbers $\mathcal{H}$, we took $G \in \{5, 10, 20\}$ and $\mathcal{H} = \{2, \ldots, H\}$, with $H \in \{10, 20, 30\}$.

The simulation results for these two examples are summarized in Table 1. Overall, the SIR-AS performed best, followed by the FSIR. Consider first the inverse regression model (5.1). We see that the CUME performed poorly, and that the FSIR is sensitive to the choice of H. Furthermore, the SIR with $G = 10$ and the SIR-AS performed best. For the SIR, over-slicing ($G = 20$) did not affect the performance, but under-slicing ($G = 5$) deteriorated the performance dramatically. The success of the SIR with $G = 10$ is expected, because for this model, the quantile slicing scheme with 10 slices is optimal. To understand the reason for the success of the SIR-AS, we calculated the percentage of choosing the optimal slicing scheme. It turns out that the SIR-AS always made the correct decision. However, this is expected from Theorem 1. Our conclusion is that when there is an optimal slicing scheme, the SIR and CUME can fail, and the FSIR can be unstable. We now turn to the forward regression model (5.2). Here, the SIR-AS outperformed its competitors. The user-specified parameter, G for the SIR and H for the FSIR, had only minor effects on the results.

5.2. SAVE-AS

Example 3. Inverse regression. The setup is the same as in Example 2.

Example 4. Forward regression. We first generated $\boldsymbol{X}$ from a multivariate Gaussian distribution with mean vector zero and identity covariance matrix. We then generated Y according to the following model:

$$Y = (\boldsymbol{\beta}_1^\top \boldsymbol{X})^2 + 3\sin\left(\frac{\boldsymbol{\beta}_2^\top \boldsymbol{X}}{4}\right) + 0.2\epsilon, \tag{5.3}$$

where $\boldsymbol{\beta}_1 = (1,1,1,0,\ldots,0)^\top \in \mathbb{R}^{p\times 1}, \boldsymbol{\beta}_2 = (1,0,0,0,1,3,0,\ldots,0)^\top \in \mathbb{R}^{p\times 1}$, and ϵ is standard normal and is independent of $\boldsymbol{X}$. In this example, $\mathcal{S}_{Y|\boldsymbol{X}} = \mathcal{S}_{SAVE} = \text{span}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. The optimal slicing scheme is not well defined.

We compared the SAVE-AS with the original SAVE of Cook and Weisberg (1991), cumulative variance estimation (CUVE) of Zhu, Zhu and Feng (2010), and fused sliced average variance estimation (FSAVE) of Cook and Zhang (2014). The kernel matrix of the CUVE is the sum of the SAVE kernel matrices from all slicing schemes with two slices, and the kernel matrix for the FSAVE is the sum of the SAVE kernel matrices for a predefined set of slicing schemes. We used quantile slicing for the SAVE and FSAVE. For the SAVE, we considered the number of slices $G \in \{5, 10, 20\}$, and for the FSAVE, we took the set of slice numbers $\mathcal{H} = \{2, \ldots, H\}$ with $H \in \{10, 20, 30\}$.

The simulation results for these two examples are summarized in Table 2. Overall, the SAVE-AS performed well. In the inverse regression model (5.1), the CUVE and FSAVE performed poorly. The SAVE with $G = 10$ and the SAVE-AS performed best, followed by the SAVE with $G = 20$. For this model, the quantile slicing scheme with 10 slices is optimal. This explains the success of the SAVE with $G = 10$. To understand the reason for the success of the SAVE-AS, we calculated the percentage of choosing the optimal slicing scheme. It turns out that the SAVE-AS made the correct decision every time. However, this is expected from Theorem 2. We also see that for the SAVE, under-slicing ($G = 5$) degraded the performance severely. We thus conclude that when there is an optimal slicing scheme, the SAVE, CUVE, and FSAVE can all fail. We now consider the forward regression model (5.3). We see that the SAVE-AS outperformed the CUVE, and the FSAVE is sensitive to the choice of H. Furthermore, the SAVE was strongly affected by the number of slices. This is in consistent with the theoretical behavior of the SAVE (Li and Zhu (2007)).

6. An Illustration

In this section, we illustrate the proposed methodology using a real-data example.

Example 5. The concrete compressive strength data (Yeh (1998)). Concrete is one of the most important materials in civil engineering. This data set records the compressive strength of 1030 concrete mixtures, together with

Table 2. Means and standard deviations (in parentheses) of the vector correlation coefficient for the SAVE-AS and its various competitors, based on 200 data applications, for Examples 3 and 4.

	SAVE				FSAVE			
Model	$G=5$	$G=10$	$G=20$	CUVE	$H=10$	$H=20$	$H=30$	SAVE-AS
(5.1)	0.209	0.677	0.643	0.213	0.252	0.411	0.385	0.677
	(0.138)	(0.061)	(0.095)	(0.140)	(0.180)	(0.215)	(0.214)	(0.061)
(5.3)	0.936	0.851	0.466	0.721	0.941	0.855	0.689	0.793
	(0.082)	(0.173)	(0.263)	(0.248)	(0.071)	(0.152)	(0.236)	(0.220)

their age and seven ingredients. These ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate. The data set is available at `http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength`. Here, we regress the concrete compressive strength on the other variables.

Dimension-reduction analyses of this data set are largely based on first conditional moments (Zhou and He (2008); Cook and Zhang (2014)). As such, we restrict our attention to the SIR-AS and its competitors: the SIR with $G \in \{5, 10, 20\}$, CUME, and FSIR with $\mathcal{H} = \{2, \ldots, H\}$ and $H \in \{10, 20, 30\}$. Previous results suggest that we can take $d = 2$ as the dimension of $\mathcal{S}_{SIR}$. An accurate estimation of d is important and interesting, but is beyond the scope of this study.

In real-data problems, we do not know the true dimension-reduction subspace. This makes a comparison of different methods difficult. Because all methods considered here are unbiased for estimating $\mathcal{S}_{SIR}$, we can pick the one with the minimum variance (Ye and Weiss (2003)). To assess the variability, we used the resampling technique. Specifically, we generated 400 subsamples randomly from the observed data, with sample size 800. For each method, we calculated the vector correlation coefficient between the full sample estimate $\hat{\mathbf{B}}$ and the subsample estimate $\hat{\mathbf{B}}^{(s)}$, for $s = 1, \ldots, 400$. The results are summarized in Table 3. The results show that the SIR-AS outperformed all other methods. The comparison of the SIR-AS and the SIR with a fixed slicing scheme and that of the SIR-AS and the FSIR with a predefined set of slicing schemes are not completely fair, because adaptive slicing introduces additional uncertainty into the estimates. Nevertheless, the results suggest that for this data set the SIR-AS is very competitive.

Table 3. Means of the vector correlation coefficient between the full sample estimate $\hat{\mathbf{B}}$ and the subsample estimate $\hat{\mathbf{B}}^{(s)}$ for the SIR-AS and its various competitors, based on 400 random subsamples, for the concrete compressive strength data.

SIR ($G = 5$)	0.757
SIR ($G = 10$)	0.709
SIR ($G = 20$)	0.599
CUME	0.775
FSIR ($H = 10$)	0.734
FSIR ($H = 20$)	0.739
FSIR ($H = 30$)	0.738
SIR-AS	0.845

7. Discussion

We have considered the long-standing problem of how to choose a good slicing scheme for the SIR and SAVE. We re-derived the SIR and SAVE using the trace maximization principle, and then proposed two procedures, the SIR with adaptive slicing and the SAVE with adaptive slicing, by penalizing the respective traces. We developed a dynamic programming algorithm for numerical optimization. Our simulation results show that, on average, adaptive slicing outperforms cumulative slicing and fusion, both of which indirectly address the slicing problem. We have implemented the procedures in **R**, and the computer program can be requested from the authors directly. The general methodology of adaptive slicing can be applied to other dimension-reduction methods that involve slicing a quantitative response, such as a directional regression and linear combinations of these methods (Ye and Weiss (2003)).

Almost all traditional dimension-reduction methods rely on the traditional asymptotic reasoning for support, letting the sample size $n \to \infty$ with the number of predictors p fixed. When n is not sufficiently large, they encounter estimation problems. Proposals such as screening and selection have been proposed to carry out sufficient dimension reduction in high-dimensional regressions. Investigations of adaptive slicing in high-dimensional settings are interesting and important. This is left to future work.

The structural dimension is assumed to be known throughout the paper. In practice, it is unknown. To learn the dimension from the data, sequential tests (Cook and Weisberg (1991)) and information criteria (Zhu, Miao and Peng (2006)) are commonly used in the dimension-reduction literature. One advantage of information criteria is that the selection consistency follows from the estimation consistency. Because adaptive slicing explicitly accounts for the structural dimension, addressing slicing and dimension selection simultaneously deserves further investigation. A simple approach, as employed in the real-data example, is to choose the dimension based on a rough slicing scheme, and then to base adaptive slicing on the chosen dimension.

Supplementary Material

The online Supplementary Material contains additional simulations and all proofs.

Acknowledgments

Tao Wang was supported, in part, by the National Natural Science Foundation of China (11971017), National Key R&D Program of China (2018YFC091050 0), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), SJTU Trans-med Awards Research Young Faculty Grant (YG2019QNA26, YG2 019QNA37), and Neil Shen's SJTU Medical Research Fund.

References

Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *J. Roy. Statist. Soc. Ser. B. Stat. Methodol.* **63**, 393–410.

Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38**, 3696–3723.

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

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22**, 1–26.

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

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *J. Amer. Statist. Assoc.* **104**, 197–208.

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

Cook, R. D. and Weisberg, S. (1991). Comment. *J. Amer. Statist. Assoc.* **86**, 328–332.

Cook, R. D. and Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *J. Amer. Statist. Assoc.* **109**, 815–827.

Heller, R., Heller, Y., Kaufman, S., Brill, B. and Gorfine, M. (2016). Consistent distribution-free K-sample and independence tests for univariate random variables. *J. Mach. Learn. Res.* **17**, 1–54.

Jiang, B., Ye, C. and Liu, J. S. (2015). Nonparametric K-sample tests via dynamic slicing. *J. Amer. Statist. Assoc.* **110**, 642–653.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316–327.

Li, Y. and Zhu, L. (2007). Asymptotics for sliced average variance estimation. *Ann. Statist.* **35**, 41–69.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc.* **107**, 168–179.

Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *Int. Stat. Rev.* **81**, 134–150.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Wang, T. and Zhu, L. (2015). A distribution-based lasso for a general single-index model. *Sci. China Math.* **58**, 109–130.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B. Stat. Methodol.* **64**, 363–410.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.* **98**, 968–979.

Yeh, I. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cem. Concr. Res.* **28**, 1797–1808.

Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36**, 1649–1668.

Zhu, L. and Fang, K. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **24**, 1053–1068.

Zhu, L., Miao, B. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Amer. Statist. Assoc.* **101**, 630–643.

Zhu, L. and Ng, K. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5**, 727–736.

Zhu, L., Ohtaki, M. and Li, Y. (2007). On hybrid methods of inverse regression-based algorithms. *Comput. Statist. Data Anal.* **51**, 2621–2635.

Zhu, L., Wang, T., Zhu, L. and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika* **97**, 295–304.

Zhu, L., Zhu, L. and Feng, Z. (2010). Dimension reduction in regressions through cumulative slicing estimation. *J. Amer. Statist. Assoc.* **105**, 1455–1466.

Tao Wang

Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai 200240, China.

SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University, Shanghai 200240, China.

MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China.

E-mail: neowangtao@sjtu.edu.cn

(Received March 2019; accepted July 2020)