Biostatistics (2016), **17**, 3, *pp.* 405–421
doi:10.1093/biostatistics/kxv051
Advance Access publication on December 24, 2015

Sufficient dimension reduction with additional information

HUNG HUNG*, CHIH-YEN LIU

Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan
hhung@ntu.edu.tw

HENRY HORNG-SHING LU

Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan

Summary

Sufficient dimension reduction is widely applied to help model building between the response Y and covariate X. In some situations, we also collect additional covariate W that has better performance in predicting Y, but has a higher obtaining cost, than X. While constructing a predictive model for Y based on (X, W) is straightforward, this strategy is not applicable since W is not available for future observations in which the constructed model is to be applied. As a result, the aim of the study is to build a predictive model for Y based on X only, where the available data is (Y, X, W). A naive method is to conduct analysis using (Y, X) directly, but ignoring W can cause the problem of inefficiency. On the other hand, it is not trivial to utilize the information of W to infer (Y, X), either. In this article, we propose a two-stage dimension reduction method for (Y, X) that is able to utilize the information of W. In the breast cancer data, the risk score constructed from the two-stage method can well separate patients with different survival experiences. In the Pima data, the two-stage method requires fewer components to infer the diabetes status, while achieving higher classification accuracy than the conventional method.

Keywords: Additional information; Efficiency; Envelopes; Sufficient dimension reduction.

1. Introduction

We consider the problem of inferring the association between the response $Y \in \mathbb{R}$ and the covariate $X \in \mathbb{R}^p$. A common situation in modern biomedical research is the high dimensionality of X, which complicates statistical model building for (Y, X). Sufficient dimension reduction (SDR) has been proposed to reduce the dimension of X while preserving its information for Y without requiring distributional assumption on (Y, X). It aims to search a matrix Γ such that $Y \perp X|\Gamma^T X$. The space span(Γ) is called the dimension reduction subspace for Y with respect to X, and the intersection of all span(Γ), denoted by $\mathcal{S}_{Y|X}$, is called the central subspace (CS) with the structural dimension $d = \dim(\mathcal{S}_{Y|X})$. Under mild conditions (Cook, 1998), $\mathcal{S}_{Y|X}$ exists and is unique. By definition, $\mathcal{S}_{Y|X}$ carries least but sufficient information of X regarding Y, and is the target of this research. The relationship of (Y, X) can be explored by the $(d + 1)$-dimensional

*To whom correspondence should be addressed.

plot of $(Y, \Gamma^{\mathrm{T}}X)$, which is helpful to model (Y, X). There are many methods developed to estimate $\mathcal{S}_{Y|X}$ since the work of Li (1991). We refer the readers to the review paper of Ma and Zhu (2013) for recent developments of linear SDR. See also Wu (2008), Yeh *and others* (2009), and Li *and others* (2011) for recent developments of nonlinear SDR.

In some studies, we also collect additional covariate W which has better performance in predicting Y, but has a higher obtaining cost, than X. While constructing a prediction model for Y based on (X, W) is straightforward, it is not applicable since W is not available for future observations in which the prediction model is to be applied. Thus, the research interest still lies on estimating $\mathcal{S}_{Y|X}$ to assist making inference about (Y, X), despite the information of W is available in the training stage only. In the breast cancer data, for example, Y is the survival time, X contains 30 real-valued features from digitized image of a fine needle aspirate, and W is the breast cancer stage. Obviously, the cancer stage provides more information about the survival experience, but it is an invasive process to collect W (which makes the patient unwilling to obtain W). It is thus of interest to construct a prediction model for Y based on the easily obtained X only. The invasive process is only suggested for those susceptible subjects (screened out by X) to obtain W, to help identify the true behavior of Y (with higher precision) for further treatment.

A naive approach is to estimate $\mathcal{S}_{Y|X}$ using (Y, X) directly, and any dimension reduction method can be applied. However, this simple strategy does not utilize W. As mentioned in our motivating example, W contains more information about Y than X, and ignoring W could cause the problem of inefficiency. This phenomenon can be partially observed via the following example.

EXAMPLE 1.1 Let $X \sim N(0, I_p)$ and $W|X \sim N(\beta^{\mathrm{T}}X, 1 - b^2)$ with $b = \|\beta\| < 1$. Let

$$Y|(X, W) \sim N(\gamma^{\mathrm{T}}X + aW, \sigma^2) \;\Rightarrow\; Y|X \sim N((a\beta + \gamma)^{\mathrm{T}}X, \sigma^2 + a^2(1 - b^2)), \tag{1.1}$$

where $\gamma \in \mathbb{R}^p$, $a \in \mathbb{R}^+$ controls the influence of W in explaining Y, b controls the correlation between (X, W), and σ^2 is the variance of Y given (X, W). It gives $\mathcal{S}_{Y|X} = \mathrm{span}(a\beta + \gamma)$.

One can observe from Example 1.1 that, without considering W, the conditional variance of Y is incremented by $a^2(1 - b^2)$, which is an increasing function of a. When a is large (i.e., W heavily affects Y), the estimation procedure using (Y, X, W) suffers less inherent variation, and hence, has a chance of being more efficient than using (Y, X) only.

Incorporating W into the estimation of $\mathcal{S}_{Y|X}$ is not trivial. One possibility is to use $\mathcal{S}_{Y|X}^{(W)}$, the partial central subspace (PCS) of Y on X given W, which is the intersection of all $\mathrm{span}(\Gamma_W)$ satisfying $Y \perp X|(\Gamma_W^T X, W)$. Chiaromonte *and others* (2002) show that either $W \perp X|\Gamma_W^T X$ or $W \perp Y|\Gamma_W^T X$ implies $\mathcal{S}_{Y|X} \subseteq \mathcal{S}_{Y|X}^{(W)}$, and $W \perp Y|X$ implies $\mathcal{S}_{Y|X}^{(W)} \subseteq \mathcal{S}_{Y|X}$. These results suggest the capability of using $\mathcal{S}_{Y|X}^{(W)}$ to partially estimate $\mathcal{S}_{Y|X}$. The stated conditions, however, are not easy to check, and $\mathcal{S}_{Y|X} \neq \mathcal{S}_{Y|X}^{(W)}$ in general. In Example 1.1, for instance, $\mathcal{S}_{Y|X}^{(W)} = \mathrm{span}(\gamma)$ differs from the target of interest $\mathcal{S}_{Y|X} = \mathrm{span}(a\beta + \gamma)$. The aim of this study is to propose an SDR method that targets $\mathcal{S}_{Y|X}$ correctly, while utilizing all the information of (Y, X, W).

2. REVIEWS OF DIMENSION REDUCTION METHODS

2.1 *Preliminary*

In the rest of discussion, Y is assumed to be discrete with support $\{1, \ldots, H\}$. For continuous response, the discretization procedure of Li (1991) is applied. For the sake of illustration, W is also assumed to be discrete with support $\{1, \ldots, C\}$. Extension to general W will be discussed. Let $Z = \Sigma^{-1/2}(X - \mu)$ with $\mu = E[X]$ and $\Sigma = \mathrm{cov}(X)$ be the standardized version of X. Since $\mathcal{S}_{Y|X} = \Sigma^{-1/2}\mathcal{S}_{Y|Z}$ and $\mathcal{S}_{Y|X}^{(W)} =$

$\Sigma^{-1/2}\mathcal{S}_{Y|Z}^{(W)}$, we work on Z-scale to introduce our method and transform back to X-scale. In practice, Z is replaced by $\hat{Z} = \hat{\Sigma}^{-1/2}(X - \hat{\mu})$, where $(\hat{\mu}, \hat{\Sigma})$ are moment estimators of (μ, Σ). Let $\{(Y_i, X_i, W_i)\}_{i=1}^n$ be random copies of (Y, X, W). Let P_A be the orthogonal projection matrix onto a space A, or span(A) when A is a matrix, and let $Q_A = I - P_A$. $I(\cdot)$ is the indicator function. For a population quantity θ, $\hat{\theta}$ or $\tilde{\theta}$ denotes its sample version. We assume all the structural dimensions (e.g., d) are known, and discuss their selections separately.

2.2 *Estimation of* $\mathcal{S}_{Y|Z}$

One branch of dimension reduction methods is to search a kernel matrix $K_{Y|Z}$ satisfying the equality span$(K_{Y|Z}) = \mathcal{S}_{Y|Z}$. The solutions from the maximization problem

$$\max_{\substack{\beta_s:\|\beta_s\|=1 \\ \beta_s^{\mathrm{T}}\beta_l=0\ \forall\ s\neq l}} \sum_{k=1}^{d} \beta_k^{\mathrm{T}} \hat{K}_{Y|Z} \beta_k \tag{2.1}$$

are used to estimate a basis of $\mathcal{S}_{Y|Z}$. Inverse regression-based methods usually rely on the linearity condition $E[Z|A^{\mathrm{T}}Z] = P_A Z$, such as the most widely applied sliced inverse regression (SIR) of Li (1991). The population kernel matrix of SIR is $K_{\mathrm{SIR}} = \mathrm{cov}(E[Z|Y])$. Its sample version is

$$\hat{K}_{\mathrm{SIR}} = \sum_{h=1}^{H} \frac{n_h}{n} m_h m_h^{\mathrm{T}}, \tag{2.2}$$

where $m_h = (1/n_h)\sum_{i=1}^n \hat{Z}_i I(Y_i = h)$ is the slice mean and $n_h = \sum_{i=1}^n I(Y_i = h)$ is the size of the slice h. Li (1991) shows that the leading d eigenvectors of $\hat{K}_{\mathrm{SIR}}$ is a $\sqrt{n}$-consistent estimator of $\mathcal{S}_{Y|Z}$.

Note that SIR is not able to identify $\mathcal{S}_{Y|Z}$ when $E[Z|Y]$ is degenerate. To solve this problem, Cook and Weisberg (1991) propose the sliced average variance estimates (SAVE) with kernel matrix $K_{\mathrm{SAVE}} = E\{I_p - \mathrm{cov}(Z|Y)\}^2$ to estimate $\mathcal{S}_{Y|Z}$ (but further requires the constant variance condition $\mathrm{cov}(Z|A^{\mathrm{T}}Z) = Q_A$). At sample level, $\mathcal{S}_{Y|Z}$ is estimated by the leading eigenvectors of

$$\hat{K}_{\mathrm{SAVE}} = \sum_{h=1}^{H} \frac{n_h}{n} \{I - \widehat{\mathrm{cov}}(Z|Y = h)\}^2, \tag{2.3}$$

where $\widehat{\mathrm{cov}}(Z|Y = h)$ is the sample covariance matrix of $\hat{Z}_i$ within the slice h.

2.3 *Estimation of* $\mathcal{S}_{Y|Z}^{(W)}$

To estimate $\mathcal{S}_{Y|Z}^{(W)}$, Chiaromonte *and others* (2002) propose the partial sliced inverse regression (PSIR). Let (Y_w, Z_w) denote the random variables (Y, Z) given $W = w$, and let $Z_w^* = \Sigma_w^{-1/2}(Z_w - \mu_w)$ with $\mu_w = E[Z_w]$ and $\Sigma_w = \mathrm{cov}(Z_w)$ be the standardized version of Z_w given $W = w$. The idea of PSIR is from the decomposition $\mathcal{S}_{Y|Z}^{(W)} = \bigoplus_{w=1}^C \mathcal{S}_{Y_w|Z_w} = \bigoplus_{w=1}^C \Sigma_w^{-1/2}\mathcal{S}_{Y_w|Z_w^*} = \Sigma_0^{-1/2}\bigoplus_{w=1}^C \mathcal{S}_{Y_w|Z_w^*}$, where the first equality is from Proposition 3.3 of Chiaromonte *and others* (2002) and the last equality holds under the equal covariance condition $\Sigma_w = \Sigma_0$, $w = 1, \ldots, C$. Σ_0 is estimated by $\hat{\Sigma}_0 = \sum_{w=1}^C (n_w/n)\hat{\Sigma}_w$, where n_w is the number of samples within $\{W = w\}$. Let

$$\hat{K}_{\mathrm{PSIR}} = \sum_{w=1}^{C} \frac{n_w}{n} \hat{K}_{\mathrm{SIR},w}, \tag{2.4}$$

where $\hat{K}_{\mathrm{SIR},w}$ is the SIR kernel matrix based on $(Y_w, \hat{Z}^*_w)$ with $\hat{Z}^*_w = \hat{\Sigma}_0^{-1/2}(\hat{Z}_w - \hat{\mu}_w)$. A basis of $\mathcal{S}^{(W)}_{Y|Z}$ can be estimated by $\hat{\Sigma}_0^{-1/2}$ multiplying the leading eigenvectors of $\hat{K}_{\mathrm{PSIR}}$.

3. Estimation of $\mathcal{S}_{Y|Z}$ with additional information

3.1 *The W-envelope subspace and a two-stage method*

The basic idea to utilizing W is to use (Y, Z, W) to construct the *W-envelope subspace* that encapsulates $\mathcal{S}_{Y|Z}$. With the confined searching space for $\mathcal{S}_{Y|Z}$, we have a chance to improve efficiency. Its construction is based on the fact $\mathcal{S}_{Y|Z} \subseteq \mathcal{S}_{(Y,W)|Z}$, since Y is a function of (Y, W).

Definition 3.1 The W-envelope subspace of $\mathcal{S}_{Y|Z}$ is defined to be $\mathcal{S}_{\mathrm{env}} = \mathcal{S}_{(Y,W)|Z}$ with the structural dimension $d_{\mathrm{env}} = \dim(\mathcal{S}_{\mathrm{env}})$.

Again, we assume that d_{env} is known and its selection will be discussed later. Another expression of $\mathcal{S}_{\mathrm{env}}$ via the concept of PCS is established below (see Appendix A for the proof in the supplementary materials available at *Biostatistics* online).

Proposition 3.1 $\mathcal{S}_{\mathrm{env}} = \mathcal{S}_{W|Z} \oplus \mathcal{S}^{(W)}_{Y|Z}$.

Although two expressions of $\mathcal{S}_{\mathrm{env}}$ are equivalent in the population level, we will see that the expression $\mathcal{S}_{\mathrm{env}} = \mathcal{S}_{W|Z} \oplus \mathcal{S}^{(W)}_{Y|Z}$ provides an adaptive method to construct $\mathcal{S}_{\mathrm{env}}$. Since $\mathcal{S}_{(Y,W)|Z}$ must exist, we always have the inclusion relationship

$$\mathcal{S}_{Y|Z} \subseteq \mathcal{S}_{\mathrm{env}}. \tag{3.1}$$

Take Example 1.1 to exemplify, where $\mathcal{S}_{Y|Z} = \mathrm{span}(a\beta + \gamma)$, $\mathcal{S}_{W|Z} = \mathrm{span}(\beta)$, and $\mathcal{S}^{(W)}_{Y|Z} = \mathrm{span}(\gamma)$. It can be seen that $\mathcal{S}_{Y|Z}$ is a proper subspace of $\mathcal{S}_{\mathrm{env}} = \mathrm{span}([\beta, \gamma])$.

Reasonably, it suffices to search $\mathcal{S}_{Y|Z}$ within $\mathcal{S}_{\mathrm{env}}$ due to (3.1). An improved estimation procedure is to search a basis of $\mathcal{S}_{Y|Z}$ via solving the maximization problem

$$\max_{\substack{\beta_s : \beta_s \in \mathcal{S}_{\mathrm{env}} \\ \|\beta_s\|=1, \beta_s^{\mathrm{T}}\beta_l = 0 \ \forall\ s \neq l}} \sum_{k=1}^{d} \beta_k^{\mathrm{T}} \hat{K}_{Y|Z} \beta_k. \tag{3.2}$$

Different from (2.1), the estimation criterion (3.2) incorporates the information of W via adding the constraints $\beta_s \in \mathcal{S}_{\mathrm{env}}$. Let B_{env} be a basis of $\mathcal{S}_{\mathrm{env}}$. Following Proposition 3 of Naik and Tsai (2005), the solutions of (3.2) are derived to be the leading d eigenvectors of $P_{B_{\mathrm{env}}} \hat{K}_{Y|Z} P_{B_{\mathrm{env}}}$. Since $\mathcal{S}_{\mathrm{env}}$ is rarely known a priori, $P_{B_{\mathrm{env}}}$ needs to be estimated from the data. Let K_{env} be a positive semi-definite kernel matrix satisfying $\mathrm{span}(K_{\mathrm{env}}) = \mathcal{S}_{\mathrm{env}}$, and let $\hat{B}_{\mathrm{env}}$ be the leading ν eigenvectors of $\hat{K}_{\mathrm{env}}$, where ν also needs to be selected from the data. The two-stage estimator for $\mathcal{S}_{Y|Z}$ is proposed to be the leading d eigenvectors of

$$P_{\hat{B}_{\mathrm{env}}} \hat{K}_{Y|Z} P_{\hat{B}_{\mathrm{env}}}. \tag{3.3}$$

From (3.3), instead of searching a basis of $\mathcal{S}_{Y|Z}$ from $\hat{K}_{Y|Z}$ directly, our two-stage method first projects $\hat{K}_{Y|Z}$ onto (the estimate of) the space $\mathcal{S}_{\mathrm{env}}$ within which we search a basis of $\mathcal{S}_{Y|Z}$. Obviously, the construction of $\hat{K}_{\mathrm{env}}$ plays the key role to the two-stage method, where the information of W should be properly used. This issue will be discussed in Section 3.2. Note that the idea of (3.3) is also used in Hung (2012),

although we derive it from (3.2). Moreover, as it will become clear from Theorem 3.2, our method is different in that we allow $\nu \geqslant d_{\text{env}}$ when forming (3.3), and a minimum variability criterion is proposed to select ν in Section 3.3.

We close this section by stating the consistency of the two-stage method, where its proof is deferred to Appendix A (see supplementary material available at *Biostatistics* online).

THEOREM 3.2 Assume the consistency of $\hat{K}_{Y|Z}$ and $\hat{K}_{\text{env}}$. Then, provided $\nu \geqslant d_{\text{env}}$, the two-stage kernel matrix $P_{\hat{B}_{\text{env}}} \hat{K}_{Y|Z} P_{\hat{B}_{\text{env}}}$ is a consistent estimator of $K_{Y|Z}$.

Theorem 3.2 implies that the two-stage method is a consistent estimator of $\mathcal{S}_{Y|Z}$. Moreover, the convergence rate can be $\sqrt{n}$, provided that $\hat{K}_{Y|Z}$ and $\hat{K}_{\text{env}}$ are $\sqrt{n}$-consistent estimators. It should be kept in mind that although (3.1) holds at the population level, this inclusion property is almost always not true at the sample level. That is, the probability that the leading eigenvectors of $\hat{K}_{Y|Z}$ lie in $\text{span}(\hat{B}_{\text{env}})$ is zero, except in the case of $\nu = p$ under which $\text{span}(\hat{B}_{\text{env}}) = \mathbb{R}^p$.

3.2 *The construction of* $\hat{K}_{\text{env}}$

We now proceed to the construction of $\hat{K}_{\text{env}}$, the core of the two-stage method. One idea is to use the definition $\mathcal{S}_{\text{env}} = \mathcal{S}_{(Y,W)|Z}$ directly. Let $K_{(Y,W)|Z}$ be a positive semi-definite kernel matrix satisfying $\text{span}(K_{(Y,W)|Z}) = \mathcal{S}_{(Y,W)|Z}$. One can then use $\hat{K}_{\text{env}} = \hat{K}_{(Y,W)|Z}$ in the two-stage method, and any dimension reduction method can be applied. For example, $K_{(Y,W)|Z}$ can be chosen to be the SIR kernel matrix

$$K^*_{\text{SIR}} = \text{cov}(E[Z|Y, W]). \tag{3.4}$$

At the sample level, one can still use (2.2) to construct $\hat{K}^*_{\text{SIR}}$, except that (Y, W) are now treated as the response to do slicing. A naive two-stage estimator of $\mathcal{S}_{Y|Z}$ is proposed below.

DEFINITION 3.3 The two-stage estimator of $\mathcal{S}_{Y|Z}$ from (3.3) with $\hat{B}_{\text{env}}$ being the leading ν eigenvectors of $\hat{K}_{\text{env}} = \hat{K}_{(Y,W)|Z}$ is defined to be $\hat{B}_0(\nu)$.

The efficiency gain of $\hat{B}_0(d_{\text{env}})$ is stated below (see Appendix A for the proof in the supplementary materials available at *Biostatistics* online), which supports the superiority of the two-stage method. Let $\tilde{B}$ be the direct estimator of $\mathcal{S}_{Y|Z}$ based on (Y, Z).

PROPOSITION 3.2 Assume the linearity and constant variance conditions. Then, $\hat{B}_0(d_{\text{env}})$ using $\hat{K}_{Y|Z} = \hat{K}_{\text{SIR}}$ in (2.2) and $\hat{K}_{(Y,W)|Z} = \hat{K}^*_{\text{SIR}}$ in (3.4) is asymptotically more efficient than $\tilde{B}$ in estimating $\mathcal{S}_{Y|Z}$, provided that $\text{span}(K_{\text{SIR}}) \bigcap \text{span}(K^*_{\text{SIR}} - K_{\text{SIR}}) \neq \{0\}$.

One advantage of using $\hat{K}_{\text{env}} = \hat{K}_{(Y,W)|Z}$ is its simple implementation, which can be conducted by existing algorithms. However, this method ignores the relative importance between $\mathcal{S}^{(W)}_{Y|Z}$ and $\mathcal{S}_{W|Z}$ when forming $\mathcal{S}_{\text{env}}$ to encapsulate $\mathcal{S}_{Y|Z}$. In Example 1.1, for instance, $\mathcal{S}_{Y|Z} = \text{span}(a\beta + \gamma)$, $\mathcal{S}_{W|Z} = \text{span}(\beta)$, and $\mathcal{S}^{(W)}_{Y|Z} = \text{span}(\gamma)$. When a is small, $a\beta + \gamma \approx \gamma$, and $\mathcal{S}^{(W)}_{Y|Z}$ should play a more important role in constructing $\mathcal{S}_{Y|Z}$ than $\mathcal{S}_{W|Z}$. Using $\hat{K}_{\text{env}} = \hat{K}_{(Y,W)|Z}$ ignores this information and may not be optimal in estimating $\mathcal{S}_{Y|Z}$.

The problem can be overcome by constructing $\hat{K}_{\text{env}}$ via the expression $\mathcal{S}_{\text{env}} = \mathcal{S}_{W|Z} \oplus \mathcal{S}^{(W)}_{Y|Z}$. Let $K^{(W)}_{Y|Z}$ and $K_{W|Z}$ be two positive semi-definite kernel matrices satisfying $\text{span}(K^{(W)}_{Y|Z}) = \mathcal{S}^{(W)}_{Y|Z}$ and $\text{span}(K_{W|Z}) =$

$\mathcal{S}_{W|Z}$. From Proposition 3.1, we have, for any $\omega \in (0, 1)$, that

$$\text{span}\left(\omega \cdot K_{W|Z} + (1-\omega)\cdot K_{Y|Z}^{(W)}\right) = \mathcal{S}_{\text{env}}. \tag{3.5}$$

An improved method is to construct $\hat{K}_{\text{env}}$ as the hybrid kernel matrix

$$\hat{K}(\omega) = \omega \cdot \hat{K}_{W|Z} + (1-\omega)\cdot \hat{K}_{Y|Z}^{(W)}, \tag{3.6}$$

where ω is the weight that controls the relative importance between $\hat{K}_{W|Z}$ and $\hat{K}_{Y|Z}^{(W)}$. Note that the construction of $\hat{K}_{W|Z}$ is still a dimension reduction problem for Z, where W is now treated as a response. Hence, we can apply SIR to construct $\hat{K}_{W|Z}$ as in (2.2) with Y being replaced by W. As to $\hat{K}_{Y|Z}^{(W)}$, it is nothing but a partial dimension reduction problem, and PSIR can be applied. An adaptive two-stage estimator of $\mathcal{S}_{Y|Z}$ is proposed below.

DEFINITION 3.4 The two-stage estimator of $\mathcal{S}_{Y|Z}$ from (3.3) with $\hat{B}_{\text{env}}$ being the leading ν eigenvectors of $\hat{K}_{\text{env}} = \hat{K}(\omega)$ is defined to be $\hat{B}(\nu, \omega)$.

Obviously, the selection of ω is critical to the performance of $\hat{B}(\nu, \omega)$. Although the choice of ω will not affect the consistency (due to (3.5)), it will affect the variation of $\hat{B}(\nu, \omega)$. Moreover, the performance of $\hat{B}(\nu, \omega)$ also depends on the value of ν. A data-adaptive method to select (ν, ω) is proposed in Section 3.3.

REMARK 3.5 For multivariate W, one can still apply PSIR to construct $\hat{K}_{Y|Z}^{(W)}$ by using W to do slicing. As to $\hat{K}_{(Y,W)|Z}$ or $\hat{K}_{W|Z}$, this is a dimension reduction problem with multivariate response, and the projective resampling technique of Li *and others* (2008) can be applied. The same two-stage procedure is ready to estimate $\mathcal{S}_{Y|Z}$ by using the modified $\hat{K}_{\text{env}}$.

3.3 *Tuning*

The performances of $\hat{B}(\nu, \omega)$ and $\hat{B}_0(\nu)$ depend on ν and ω, which should be properly determined. Below we discuss this issue for $\hat{B}(\nu, \omega)$. The case of $\hat{B}_0(\nu)$ is similar.

From Theorem 3.2, $\hat{B}(\nu, \omega)$ is a consistent estimator of $\mathcal{S}_{Y|Z}$ for any $\omega \in (0, 1)$ and $\nu \geqslant d_{\text{env}}$. Thus, the main influence of (ν, ω) should be on the variation part. It motivates us to determine (ν, ω) by the minimal variability criterion of Ye and Weiss (2003). Let $q(B_1, B_2) = \|P_{B_1} - P_{B_2}\|_F^2$ be a distance between $\text{span}(B_1)$ and $\text{span}(B_2)$. The variability of $\hat{B}(\nu, \omega)$ is constructed to be

$$v(\nu, \omega) = \frac{1}{m}\sum_{b=1}^{m} q(\hat{B}^{(b)}(\nu, \omega), \hat{B}(\nu, \omega)), \tag{3.7}$$

where $\{\hat{B}^{(b)}(\nu, \omega) : b = 1, \ldots, m\}$ are the bootstrapped two-stage estimates. The data-adaptive two-stage estimator of $\mathcal{S}_{Y|Z}$ is then proposed to be

$$\hat{B} = \hat{B}(\nu^*, \omega^*) \text{ with } (\nu^*, \omega^*) = \underset{\nu \geqslant d_{\text{env}},\ \omega \in \Omega}{\text{argmin}}\ v(\nu, \omega), \tag{3.8}$$

where Ω is a predetermined finite subset of $(0, 1)$. To facilitate the selection of $\omega \in \Omega$, the kernel matrices $\hat{K}_{Y|Z}^{(W)}$ and $\hat{K}_{W|Z}$ are normalized so that the sum of eigenvalues equals 1. The normalization will not affect the order of eigenvalues, but will make $\hat{K}(\omega)$ representative equally over $\omega \in \Omega$. Without any confusion,

we still use the notation $\hat{K}_{Y|Z}^{(W)}$ and $\hat{K}_{W|Z}$ to represent their normalized versions when forming $\hat{K}(\omega)$. Note that we allow the searching range $\{\nu \geqslant d_{\text{env}}\}$ to construct $\hat{B}$. This makes the two-stage estimator general to include the direct method $\hat{B}$ as a special case when $\nu^* = p$, since $P_{\hat{B}_{\text{env}}} = I_p$ in this situation. See Section 3.5 for further details. As to the case of $\hat{B}_0(\nu)$, we similarly propose

$$\hat{B}_0 = \hat{B}_0(\nu^*) \text{ with } \nu^* = \underset{\nu \geqslant d_{\text{env}}}{\text{argmin}}\, v_0(\nu), \tag{3.9}$$

where $v_0(\nu)$ is the variability of $\hat{B}_0(\nu)$, which is similarly constructed as (3.7). We remind the readers that (ν, ω) can also be tuned by the cross-validation (CV) type criterion when there is a direct measure of the performance (e.g., the classification accuracy). This method has the advantages of easy implementation and being related to the underlying problem directly.

We close this section by using Example 1.1 to demonstrate the role of ω in the efficiency gain of the two-stage method. For simplicity, assume $\beta = \gamma$ such that $\mathcal{S}_{\text{env}} = \mathcal{S}_{Y|X} = \text{span}(\beta)$. In this situation, the conventional LSE $\tilde{\beta}$ based on (Y, X) is a consistent estimator of $\mathcal{S}_{Y|X}$. Paralleling to the idea of $\hat{K}(\omega)$ in (3.6), for any fixed $\omega \in (0, 1)$, $\hat{\beta}_\omega$ from the minimization problem

$$(\hat{\beta}_\omega, \hat{a}) = \underset{\beta, a}{\text{argmin}} \left\{ \omega \cdot \sum_{i=1}^{n} (W_i - \beta^{\mathrm{T}} X_i)^2 + (1-\omega) \cdot \sum_{i=1}^{n} (Y_i - \beta^{\mathrm{T}} X_i - a W_i)^2 \right\} \tag{3.10}$$

is a consistent estimator of $\mathcal{S}_{\text{env}}$ and, hence, also a consistent estimator of $\mathcal{S}_{Y|X}$. One can treat $\tilde{\beta}$ as the direct method for $\mathcal{S}_{Y|X}$, and treat $\hat{\beta}_\omega$ as the two-stage method for $\mathcal{S}_{Y|X}$ since the information of W is utilized in (3.10). To calculate the efficiency gain from using $\hat{\beta}_\omega$ over $\tilde{\beta}$, we first deduce from conventional arguments that

$$\sqrt{n} \begin{bmatrix} \hat{\beta}_\omega - \beta \\ \hat{a} - a \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \dfrac{1}{1-(1-\omega)b^2} P_\beta + Q_\beta & \dfrac{-1}{1-(1-\omega)b^2} \beta \\ \dfrac{-1}{1-(1-\omega)b^2} \beta^{\mathrm{T}} & \dfrac{1}{(1-\omega)\{1-(1-\omega)b^2\}} \end{bmatrix} N_0, \tag{3.11}$$

where N_0 follows a zero mean multivariate normal distribution with covariance matrix

$$\begin{bmatrix} \{\omega^2(1-b^2) + (1-\omega)^2\sigma^2\} I_p & (1-\omega)^2\sigma^2\beta \\ (1-\omega)^2\sigma^2\beta^{\mathrm{T}} & (1-\omega)^2\sigma^2 \end{bmatrix}. \tag{3.12}$$

By the delta method, $\sqrt{n}\text{vec}(P_{\hat{\beta}_\omega} - P_\beta) \xrightarrow{d} N(0, C_\omega)$ with $C_\omega = \sigma_\omega^2 \cdot (I_{p^2} + K_{p,p})(P_\beta \otimes Q_\beta)(I_{p^2} + K_{p,p})$, where $\sigma_\omega^2 = \{\omega^2(1-b^2) + (1-\omega)^2\sigma^2\}/b^2$, $\otimes$ is the Kronecker product, and $K_{p,p}$ is the commutation matrix (Henderson and Searle, 1979). Similarly, we deduce that $\sqrt{n}\text{vec}(P_{\tilde{\beta}} - P_\beta) \xrightarrow{d} N(0, C_0)$, where $C_0 = \sigma_0^2 \cdot (I_{p^2} + K_{p,p})(P_\beta \otimes Q_\beta)(I_{p^2} + K_{p,p})$ with $\sigma_0^2 = \{\sigma^2 + a^2(1-b^2)\}/\{(1+a)^2 b^2\}$. Comparing two asymptotic covariance matrices C_ω and C_0, the effect of ω can be observed via $\delta_\omega = \sigma_0^2 - \sigma_\omega^2$, where $\delta_\omega > 0$ indicates an efficiency gain from using $\hat{\beta}_\omega$ in estimating $\mathcal{S}_{Y|X}$. Direct calculation gives $\omega^* = \text{argmax}_\omega\, \delta_\omega = \sigma^2/\{\sigma^2 + (1-b^2)\}$ (not surprisingly, ω^* reflects the relative magnitudes of $\text{var}(Y|X, W) = \sigma^2$ and $\text{var}(W|X) = 1 - b^2$) and $\delta_{\omega^*} = \{a(1-b^2) - \sigma^2\}^2/[b^2(1+a)^2\{\sigma^2 + (1-b^2)\}] \geqslant 0$. Provided $a \neq \sigma^2/(1-b^2)$, $\delta_{\omega^*} > 0$; that is, there exists an open subset of ω within which the efficiency gain of $\hat{\beta}_\omega$ is guaranteed.

The implications are 2-fold. First, the selection of ω is important to the success of the two-stage method. It also supports choosing ω via minimizing the variability measure in (3.8). Secondly, with the properly

chosen ω, the efficiency gain of the two-stage method is expected except in the restrictive situations (e.g., $a = \sigma^2/(1-b^2)$ under which $\hat{\beta}_{\omega^*}$ and $\tilde{\beta}$ are asymptotically equivalent).

3.4 *Determination of* (d, d_{env})

We adopt the bayesian information criterion (BIC)-type criterion of Zhu *and others* (2010) to determine (d, d_{env}). Let

$$\hat{d}_{\text{env}}(\omega) = \underset{k=1,\ldots,p}{\text{argmax}} \left\{ \frac{n \sum_{j=1}^{k} \{\ln(\hat{\lambda}_j + 1) - \hat{\lambda}_j\}}{2 \sum_{j=1}^{p} \{\ln(\hat{\lambda}_j + 1) - \hat{\lambda}_j\}} - 2C_n \frac{k(k-1)}{2p} \right\}, \tag{3.13}$$

where $\hat{\lambda}_j$ is the jth eigenvalue of $\hat{K}_{\text{env}}$ and C_n is a predetermined penalty. Note that $\hat{d}_{\text{env}}(\omega)$ can be a function of ω when using $\hat{K}_{\text{env}} = \hat{K}(\omega)$. To integrate out the effect of ω, we propose $\hat{d}_{\text{env}} = \text{median}\{\hat{d}_{\text{env}}(\omega) : \omega \in \Omega\}$ to estimate d_{env}. The same idea is used to determine d. Let

$$\hat{d}(\omega) = \underset{k=1,\ldots,\hat{d}_{\text{env}}(\omega)}{\text{argmax}} \left\{ \frac{n \sum_{j=1}^{k} \{\ln(\hat{\lambda}_j^* + 1) - \hat{\lambda}_j^*\}}{2 \sum_{j=1}^{p} \{\ln(\hat{\lambda}_j^* + 1) - \hat{\lambda}_j^*\}} - 2C_n \frac{k(k-1)}{2p} \right\}, \tag{3.14}$$

where $\hat{\lambda}_j^*$ is the jth eigenvalue of $P_{\hat{B}_{\text{env}}} \hat{K}_{Y|Z} P_{\hat{B}_{\text{env}}}$ with the dimension ν of $\hat{B}_{\text{env}}$ being set to be $\hat{d}_{\text{env}}(\omega)$. We then propose to estimate d by $\hat{d} = \text{median}\{\hat{d}(\omega) : \omega \in \Omega\}$.

For any fixed ω, the consistency of $\hat{d}_{\text{env}}(\omega)$ and $\hat{d}(\omega)$ can be similarly derived as Theorem 4 of Zhu *and others* (2010), provided $C_n/n \to 0$ and $C_n \to \infty$ as $n \to \infty$. Since Ω is finite, the consistency of $(\hat{d}, \hat{d}_{\text{env}})$ is a direct consequence.

3.5 *Evaluation of* W

An important issue is to evaluate W, to guide when the two-stage method gains efficiency. Showing the efficiency gain requires the joint asymptotic normality of $(\hat{K}_{Y|Z}, \hat{K}_{\text{env}})$, which depends on the chosen dimension reduction methods. To make the inference procedure flexible to adapt to various dimension reduction methods, we alternatively propose to estimate the performance of W directly. Note that the aim of using W is to reduce the variation in estimating $\mathcal{S}_{Y|Z}$, and it is natural to evaluate W via the variability measure. Recall that $\tilde{B}$ is the direct estimate of $\mathcal{S}_{Y|Z}$. Similar to (3.7), the variability of $\tilde{B}$ is constructed to be $\tilde{v} = (1/m) \sum_{b=1}^{m} q(\tilde{B}^{(b)}, \tilde{B})$, where $\tilde{B}^{(b)}$ is the bootstrapped version of $\tilde{B}$. One can then compare $\tilde{v}$ with $\hat{v} = v(\nu^*, \omega^*)$ from (3.8) to evaluate W, via the proportion of improvement $\xi_1 = (\tilde{v} - \hat{v})/\tilde{v}$.

Although using ξ_1 is straightforward, a drawback is that we need to obtain $\tilde{B}$ first, which involves determining its dimension, say $\tilde{d}$, from $\hat{K}_{Y|Z}$. We should emphasize that $\hat{d}$ and $\tilde{d}$ need not be the same. Since the distance measure q is sensitive to the dimension of the space, comparing $\hat{v}$ with $\tilde{v}$ is inappropriate when $\hat{d} \neq \tilde{d}$. We propose another criterion to evaluate W, which is based on the two-stage method solely. Recall that $P_{\hat{B}_{\text{env}}} = I_p$ when $\nu = p$, i.e., W is useless in estimating $\mathcal{S}_{Y|Z}$. It is also detected in our numerical studies that ν^* from (3.8) tends to approach p when W is less informative. In the most extreme case where $W \perp Y|Z$, we usually have $\nu^* = p$. This is reasonable since, in this case, W cannot improve estimating $\mathcal{S}_{Y|Z}$, and it suffices to estimate $\mathcal{S}_{Y|Z}$ from $\hat{K}_{Y|Z}$ directly. Another index to evaluate W is proposed to be $\xi_2 = I(\nu^* < p)$. A value of $\xi_2 = 1$ indicates an efficiency gain from using W. Note that ξ_2 is only able to indicate if W is useful or not, while the value of ξ_1 can quantify how informative W is. Both (ξ_1, ξ_2) are suggested to evaluate W when properly applied.

4. Numerical studies

4.1 *Simulation settings*

We consider two models for simulations. The first one is model (1.1) in Example 1.1 with $\beta = b \cdot (0, 0, 1, 1, 0_{p-4}^{\mathrm{T}})^{\mathrm{T}}/\sqrt{2}$ and $\gamma = (1, 1, 0, 0, 0_{p-4}^{\mathrm{T}})^{\mathrm{T}}/\sqrt{2}$, which gives $\mathcal{S}_{Y|X} = \mathrm{span}(a\beta + \gamma)$ and $\mathcal{S}_{\mathrm{env}} = \mathrm{span}([\beta, \gamma])$. The second model is constructed as below. Let $X \sim N(0, I_p)$ and $W = (W_1, W_2)^{\mathrm{T}}$ be generated from $W_1|X \sim N(\beta_1^{\mathrm{T}}X, 1 - \|\beta_1\|^2)$ and $W_2|X \sim N(\beta_2^{\mathrm{T}}X, 1 - \|\beta_2\|^2)$, where $\beta_1 = b \cdot (0, 0, 1, 1, 0_{p-4}^{\mathrm{T}})^{\mathrm{T}}/\sqrt{2}$ and $\beta_2 = b \cdot (1, 1, 0, 0, 0_{p-4}^{\mathrm{T}})^{\mathrm{T}}/\sqrt{2}$. Then, Y is generated from

$$\begin{aligned} &Y|(X, W) \sim N((1 + \alpha^{\mathrm{T}}X) \cdot \{a(W_1 + W_2) + \gamma^{\mathrm{T}}X\}, \sigma^2) \\ \Rightarrow\ &Y|X \sim N((1 + \alpha^{\mathrm{T}}X) \cdot \{a(\beta_1 + \beta_2) + \gamma\}^{\mathrm{T}}X, \sigma^2 + 2a^2(1 - b^2)(1 + \alpha^{\mathrm{T}}X)^2), \end{aligned} \tag{4.1}$$

which gives $\mathcal{S}_{Y|X}^{(W)} = \mathrm{span}([\alpha, \gamma])$, $\mathcal{S}_{W|X} = \mathrm{span}([\beta_1, \beta_2])$, and $\mathcal{S}_{Y|X} = \mathrm{span}([\alpha, a(\beta_1 + \beta_2) + \gamma])$. We consider $\alpha = (0, 0, 0, 0, 1, 1, 0_{p-6}^{\mathrm{T}})^{\mathrm{T}}/\sqrt{2}$ and $\gamma = (1, 1, 2, 2, 0_{p-4}^{\mathrm{T}})^{\mathrm{T}}/\sqrt{10}$ so that $\gamma \in \mathrm{span}([\beta_1, \beta_2])$ and, hence, $\mathcal{S}_{Y|X}^{(W)}$ and $\mathcal{S}_{W|X}$ have overlap. It further gives $\mathcal{S}_{\mathrm{env}} = \mathrm{span}([\alpha, \beta_1, \beta_2])$. In both models, a controls the ability of W to explain Y, and b controls the correlation between (W, Z).

Both $\hat{B}$ and $\hat{B}_0$ are constructed to compare with the direct method $\tilde{B}$. We use SIR to construct $\hat{K}_{Y|Z}$ by categorizing Y into 10 slices. For $\hat{B}$, we use SIR to construct $\hat{K}_{W|Z}$ by using W to categorize subjects into 4 (for $n = 150$) or 9 (for $n = 250$) slices, and use PSIR to construct $\hat{K}_{Y|Z}^{(W)}$ by further categorizing Y into 3 slices within each slice of W. For $\hat{B}_0$, SIR is used to construct $\hat{K}_{(Y,W)|Z}$ using the same slicing as $\hat{B}$. We also try other settings of slicing, and the simulation results are placed in Appendix B (see supplementary material available at *Biostatistics* online). For the two-stage method, we use $C_n = n^{1/4}$ and $\Omega = \{\frac{5}{50}, \frac{6}{50}, \ldots, \frac{45}{50}\}$ to determine (υ^*, ω^*). The *trace correlation coefficient* $r = \sqrt{\mathrm{tr}(P_{B_1} P_{B_2})/d}$ is used to measure the distance between two d-dimensional subspaces with bases B_1 and B_2. The value of r belongs to $[0, 1]$, and $r = 1$ indicates $\mathrm{span}(B_1) = \mathrm{span}(B_2)$. Simulation results are reported under $\sigma = 0.5$ and different combinations of $(n, p) = \{(150, 9), (250, 25)\}$, $a = (0, 0.5, \ldots, 3)$ and $b = (0.1, 0.3)$, based on $m = 300$ bootstraps and 500 replicates.

4.2 *Simulation results*

Simulation results under models (1.1) and (4.1) are placed in Figure 1(a) and (b), which show the means of the trace correlation $\hat{r}$ and $\tilde{r}$ of $\hat{B}$ and $\tilde{B}$, respectively. The means of the performance measures (ξ_1, ξ_2) of $\hat{B}$ are shown in Figure 1(c) and (d). Recall that a controls the influence of W on Y (which further affects $\mathrm{var}(Y|Z)$; see (1.1)), and b controls the correlation between (Z, W). Thus, when a is large (i.e., W is an important covariate), ignoring W gives the direct method $\tilde{B}$ a small $\tilde{r}$ value. For any fixed a, a high value of $\tilde{r}$ is detected for large b, since in this situation Z can well predict Y through its connection with W. Similar patterns are also observed for $\hat{r}$.

The magnitude of improvement $(\hat{r} - \tilde{r})$ shows a different behavior. One can see that $(\hat{r} - \tilde{r})$ increases as a increases. The more information W contains (i.e., large a), the more improvement $\hat{B}$ can achieve. On the other hand, $(\hat{r} - \tilde{r})$ increases as b decreases. Note that when b is small, Z can hardly be a surrogate of W and, hence, the two-stage method benefits more from utilizing W. Overall, $\hat{B}$ outperforms $\tilde{B}$ in all settings, and the measures (ξ_1, ξ_2) appropriately reflect the performance of $\hat{B}$. Even in the most extreme case of $a = 0$, i.e., $W \perp Y|Z$, $\hat{B}$ can perform not worse than $\tilde{B}$. Note that, in this situation, it suffices to estimate $\mathcal{S}_{Y|Z}$ based on (Y, Z). Thus, υ^* tends to approach p so that $\hat{B}$ reduces to $\tilde{B}$ as mentioned below (3.8). This fact also shows the applicability of the two-stage method in various situations of W.

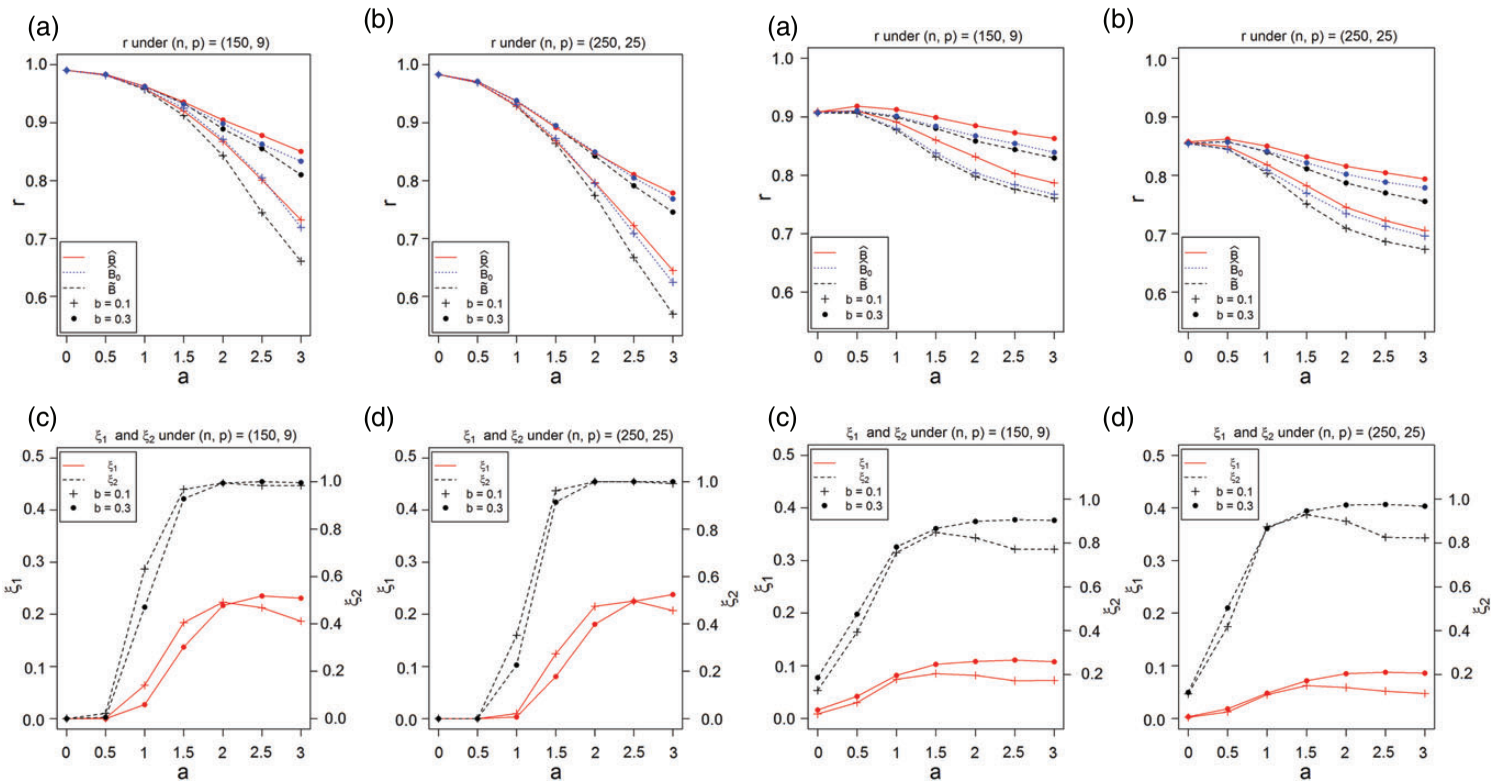

Fig. 1. Simulation results for model (1.1) (the left panel) and model (4.1) (the right panel), under $(n, p) = (150, 9)$ in (a) and (c), and under $(n, p) = (250, 25)$ in (b) and (d). (a) and (b) The trace correlations r of $\hat{B}$ (real line), $\hat{B}_0$ (dot line), and $\tilde{B}$ (dash line) under $b = 0.1$ (lines with $+$) and $b = 0.3$ (lines with $\bullet$); (c) and (d) The performance measures ξ_1 (real line) and ξ_2 (dash line) of $\hat{B}$ under $b = 0.1$ (lines with $+$) and $b = 0.3$ (lines with $\bullet$). Note that the scale of y-axis in the left of subplot (c) or (d) is for ξ_1, while the one in the right is for ξ_2.

Table 1. *Selection proportions of* $\hat{d}$ *and* $\tilde{d}$ (*at values* $1, 2, 3,$ *and* > 3) *under models* (1.1) *and* (4.1) *with different combinations of* (a, b, C_n). *The columns corresponding to the true dimension* d *are marked as bold*

			$C_n = 0.5n^{1/4}$				$C_n = n^{1/4}$				$C_n = 2n^{1/4}$			
Model	(a, b)		1	2	3	>3	1	2	3	>3	1	2	3	>3
(1.1)	(1, 0.1)	$\hat{d}$	**0.365**	0.625	0.010	0.000	**0.855**	0.145	0.000	0.000	**0.985**	0.015	0.000	0.000
		$\tilde{d}$	**0.000**	0.010	0.500	0.490	**0.000**	0.205	0.735	0.060	**0.005**	0.790	0.205	0.000
	(1, 0.3)	$\hat{d}$	**0.470**	0.520	0.010	0.000	**0.885**	0.115	0.000	0.000	**0.995**	0.005	0.000	0.000
		$\tilde{d}$	**0.000**	0.005	0.540	0.455	**0.000**	0.205	0.750	0.045	**0.005**	0.870	0.125	0.000
	(3, 0.1)	$\hat{d}$	**0.015**	0.490	0.485	0.010	**0.075**	0.855	0.070	0.000	**0.480**	0.520	0.000	0.000
		$\tilde{d}$	**0.000**	0.000	0.030	0.970	**0.000**	0.000	0.265	0.735	**0.000**	0.080	0.715	0.205
	(3, 0.3)	$\hat{d}$	**0.010**	0.650	0.335	0.005	**0.165**	0.795	0.040	0.000	**0.660**	0.340	0.000	0.000
		$\tilde{d}$	**0.000**	0.000	0.055	0.945	**0.000**	0.005	0.475	0.520	**0.000**	0.135	0.810	0.055
(4.1)	(1, 0.1)	$\hat{d}$	0.000	**0.480**	0.520	0.000	0.010	**0.905**	0.085	0.000	0.110	**0.890**	0.000	0.000
		$\tilde{d}$	0.000	**0.005**	0.175	0.820	0.000	**0.010**	0.625	0.365	0.000	**0.210**	0.765	0.025
	(1, 0.3)	$\hat{d}$	0.000	**0.710**	0.290	0.000	0.000	**0.985**	0.015	0.000	0.105	**0.895**	0.000	0.000
		$\tilde{d}$	0.000	**0.000**	0.220	0.780	0.000	**0.045**	0.660	0.295	0.000	**0.370**	0.615	0.015
	(3, 0.1)	$\hat{d}$	0.005	**0.535**	0.450	0.010	0.085	**0.845**	0.070	0.000	0.545	**0.455**	0.000	0.000
		$\tilde{d}$	0.000	**0.000**	0.125	0.875	0.000	**0.010**	0.570	0.420	0.000	**0.260**	0.715	0.025
	(3, 0.3)	$\hat{d}$	0.000	**0.535**	0.465	0.000	0.015	**0.895**	0.090	0.000	0.185	**0.815**	0.000	0.000
		$\tilde{d}$	0.000	**0.000**	0.105	0.895	0.000	**0.010**	0.530	0.460	0.000	**0.190**	0.775	0.035

The trace correlation $\hat{r}_0$ of $\hat{B}_0$ is also shown in Figure 1(a) and (b). Obviously, the winner is still $\hat{B}$, followed by $\hat{B}_0$ and then $\tilde{B}$. These results further indicate the benefit of using the hybrid method to estimate $\mathcal{S}_{\text{env}}$. Indeed, we rarely know the relative importance of $\mathcal{S}_{Y|Z}^{(W)}$ and $\mathcal{S}_{W|Z}$, which is totally ignored when directly estimating $\mathcal{S}_{\text{env}}$ by $\hat{K}_{(Y,W)|Z}$. By using the hybrid kernel matrix $\hat{K}(\omega)$, it allows the data to select ω with minimal variability, to adapt to various relationships between $\mathcal{S}_{Y|Z}^{(W)}$ and $\mathcal{S}_{W|Z}$, and a good performance of $\hat{B}$ is achieved.

The selection proportions of $\hat{d}$ using $\hat{K}_{\text{env}} = \hat{K}(\omega)$ and different C_n values are placed in Table 1. The results of $\tilde{d}$, the estimator of d from using $\hat{K}_{Y|Z}$ directly, are also provided. One can see that $\hat{d}$ achieves higher accuracies than $\tilde{d}$ over a wide range of C_n. By using W, we cannot only improve estimating $\mathcal{S}_{Y|Z}$, but also improve estimating the structural dimension d.

5. Data analysis

Two data sets downloaded from the UCI Machine Learning Repository are analyzed in this section. Let $\hat{S}_j$ and $\tilde{S}_j$ be the jth SDR components of $\hat{S} = \hat{B}^{\mathrm{T}}\hat{Z}$ and $\tilde{S} = \tilde{B}^{\mathrm{T}}\hat{Z}$, respectively.

5.1 *The breast cancer Wisconsin prognostic data*

The data contains 253 breast cancer cases. In our analysis, X is the first 10 principal components of the 30 real-valued features from digitized image of a fine needle aspirate, W is the staging of breast cancer which is defined by tumor size and the lymph node status, and Y is the survival time in months (the censoring rate is 81%). Empirically, W is strongly associated with the survival time, but an invasive surgery is required to

obtain W. It is of importance to build a pre-screening model based on X only, to identify patients in high risk for further diagnosis.

Since the survival time is subject to censoring, we can only observe (Y^*, δ) instead of Y, where Y^* is the last observed time and δ is the censoring status. To estimate $\mathcal{S}_{Y|Z}$, under the assumption of independent censorship, Li, Wang, and Chen (1999) propose censored SIR with the kernel matrix $K_{Y|Z} = \text{cov}(E[Z|Y^*, \delta])$, and estimate it by (2.2) except that (Y^*, δ) is used to do slicing. To estimate $\mathcal{S}_{Y|Z}^{(W)}$, the same double-slicing procedure by (Y^*, δ) is used to modify PSIR to construct $\hat{K}_{Y|Z}^{(W)}$. The BIC-type criterion gives $(\hat{d}, \hat{d}_{\text{env}}) = (2, 3)$, which further gives $(\nu^*, \omega^*) = (5, 0.1)$ for $\hat{B}$. Figure 2(a) and (b) shows the scatter plots of (Y^*, δ) with respect to $(\hat{S}_1, \hat{S}_2)$. One can observe that $\hat{S}_2$ is correlated to event time (with correlation coefficient 0.4145). While $\hat{S}_1$ tends to correlate with censoring time, the variation of event time becomes large when $\hat{S}_1 > 0$. It suggests to fit the stratified Cox model $h(t|\hat{S}_1, \hat{S}_2) = h_g(t)\exp(-0.51 \cdot \hat{S}_2)$ with $g = I(\hat{S}_1 < 0)$, where $h_g(t)$ denotes the baseline hazard function of group g, $g = 0, 1$. Figure 3(a) gives the estimate of the cumulative hazard function $H_g(t) = \int_0^t h_g(u)\,\text{d}u$. One can see that $H_1(t)$ dominates $H_0(t)$ for $t > 40$, which supports the stratification by $\hat{S}_1$.

To evaluate the performances of $\hat{S}$, we divide subjects into three risk groups (Low, Medium, High) based on the 33% and 66% quantiles of the survival probability $\exp\{-H_g(t)\exp(-0.51 \cdot \hat{S}_2)\}$ at $t = 3$. To avoid over-fitting, each subject is excluded from model fitting when being classified. The Kaplan–Meier (KM) curves of three risk groups are placed in Figure 3. For comparison, the analysis results from using $\tilde{S}$ are also placed in Figures 2 and 3. One can see that the KM curves from $\hat{S}$ are well separated, while crossed KM curves are detected when using $\tilde{S}$ (Figure 3(c) and (d)). The log-rank test yields p-value of 0.0203 for $\hat{S}$, while it is 0.2412 for $\tilde{S}$. Our analysis indicates a better separability for the risk score constructed by the two-stage method.

We also consider other survival models besides the stratified Cox model, and the analysis results still convey the same message that the two-stage method produces more separated KM curves than the direct method. See Appendix C for details (see supplementary material available at *Biostatistics* online).

5.2 *The Pima Indians diabetes data*

The Pima data contain females of Pima Indian heritage, each with 7 biological covariates (X), the family disease history (W), and an indicator of diabetes status (Y). Although the family disease history is shown to strongly associate with diabetes status, missingness is very likely to occur. It is of interest to construct a prediction rule for the diabetes status based solely on X.

Since SIR can only find one direction for binary response, we use SAVE to construct $\hat{K}_{Y|Z}$. As to the estimation of $\mathcal{S}_{\text{env}}$, we apply PSIR to construct $\hat{K}_{Y|Z}^{(W)}$, and apply SIR to construct $\hat{K}_{W|Z}$. In this analysis, we choose (ν, ω) of the two-stage method via maximizing the leave-one-out classification accuracy (CA) from quadratic discriminant analysis, which gives $\hat{B} = \hat{B}(4, 0.2)$. The maximum leave-one-out CA is 0.7959 for $(\hat{S}_1, \hat{S}_2)$, while it is 0.7041 for $(\tilde{S}_1, \tilde{S}_2)$, and is 0.7781 for $(\tilde{S}_1, \tilde{S}_2, \tilde{S}_3)$. It indicates an efficiency gain from using W. This fact can be further observed from the scatter plots of $\hat{S}_j$ and $\tilde{S}_j$ in Figure 4. One can see that $(\tilde{S}_1, \tilde{S}_2)$ tend to separate diabetes status by variation, while different locations of two groups are detected for $\tilde{S}_3$. Interestingly, $(\hat{S}_1, \hat{S}_2, \hat{S}_3)$ demonstrate different behaviors in that patients with different diabetes status tend to have different variations of $\hat{S}_1$, while different locations are observed for $\hat{S}_2$. Moreover, $\hat{S}_3$ is useless in separating the diabetes status, suggesting that one only requires $(\hat{S}_1, \hat{S}_2)$ to infer the diabetes status. However, it requires $(\tilde{S}_1, \tilde{S}_2, \tilde{S}_3)$ when ignoring W. By using W, the order of the "location-separating component" is also changed from $\tilde{S}_3$ to $\hat{S}_2$.

We also implement the quadratic discriminant analysis to classify subjects by directly using W together with the leading two components of $\hat{K}_{Y|Z}^{(W)}$. This result is treated as the benchmark since W is directly used in the classification process. The resulting leave-one-out CA for the benchmark method is 0.7985. By using

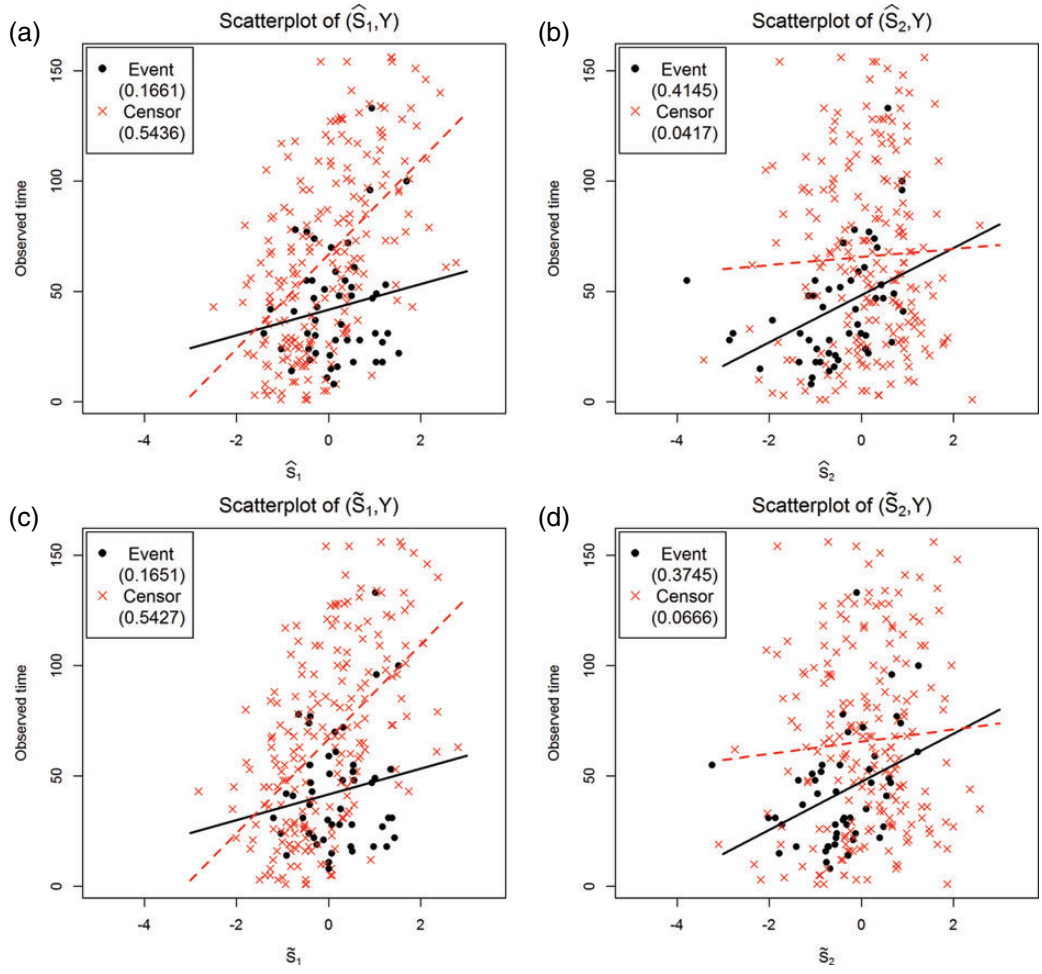

Fig. 2. (a) and (b) The scatter plots of the event time and the censoring time with respect to $(\hat{S}_1, \hat{S}_2)$; (c) and (d). The scatter plots of the event time and the censoring time with respect to $(\tilde{S}_1, \tilde{S}_2)$. Here • and × denote event and censoring, respectively, and the real (dash) line is the fitted regression line for event (censoring) time. The Pearson correlation coefficients are shown in the parentheses.

the two-stage procedure, we only require two SDR components of biological measurements to infer the behavior of diabetes status, and can achieve comparable performance with the benchmark method.

6. Discussion

We develop a two-stage method that utilizes the information of W to improve estimating $\mathcal{S}_{Y|X}$. Note that "W contains more information than X" is not a necessary condition to apply our method, although it is satisfied in our motivating examples. As a result, the two-stage method is applicable to the more general situation of covariates (X, W), where W is subject to missing for future observations, and the aim is to estimate $\mathcal{S}_{Y|X}$ for those observations without W.

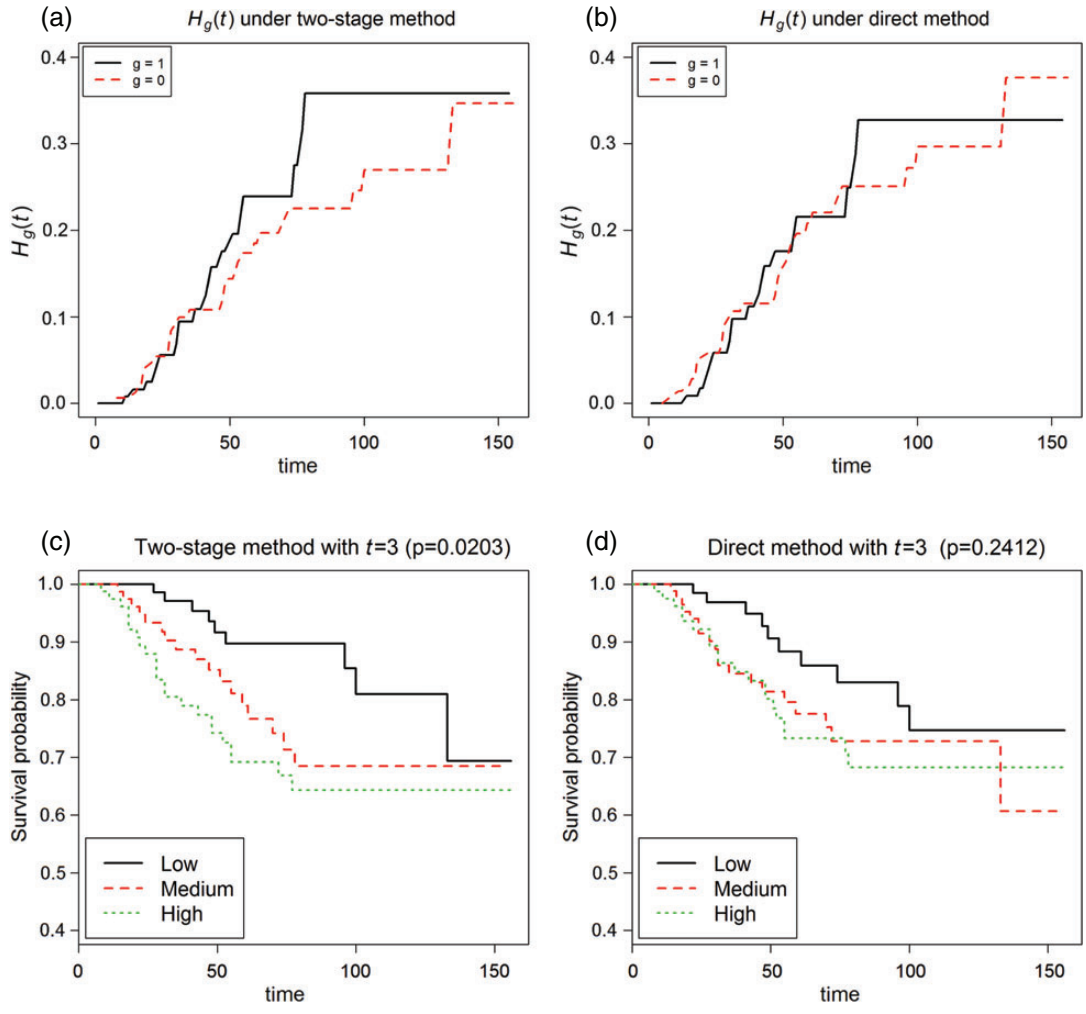

Fig. 3. (a) and (b) The estimates of the cumulative baseline hazard function $H_g(t)$; (c) and (d) The KM curves of the three risk groups (Low, Medium, High) classified by the stratified Cox model at $t = 3$, where the p-value of log-rank test is shown in the parentheses. (a) and (c) The two-stage method; (b) and (d) The direct method.

When W is not available, a common strategy is to build a model for $E(W|X)$, to impute the missing W by the observable X. The analysis results, however, can be heavily biased when $E(W|X)$ is incorrectly modeled. When $X \perp W$, there is even no hope to predict W from X, and the imputation could fail. On the other hand, the two-stage method does not need to model $E(W|X)$, and should have better performance than the imputation method when modeling $E(W|X)$ is difficult. See Appendix D for more discussion and simulation studies (see supplementary material available at *Biostatistics* online).

We remind the readers that the two-stage method can be applied regardless of the missing mechanism of W in the evaluation stage, since it is X that is required only to infer Y based on $\mathcal{S}_{Y|X}$. On the other hand, W may also be subject to missing in the training stage, and the observed data become $\{(Y_i, X_i, W_i R_i, R_i)\}_{i=1}^n$, where $R_i = 0$ indicates that W_i is missing. It is possible to extend our method to adapt to this data structure, by applying the technique of *pmf imputation* (Ding and Wang, 2011) under

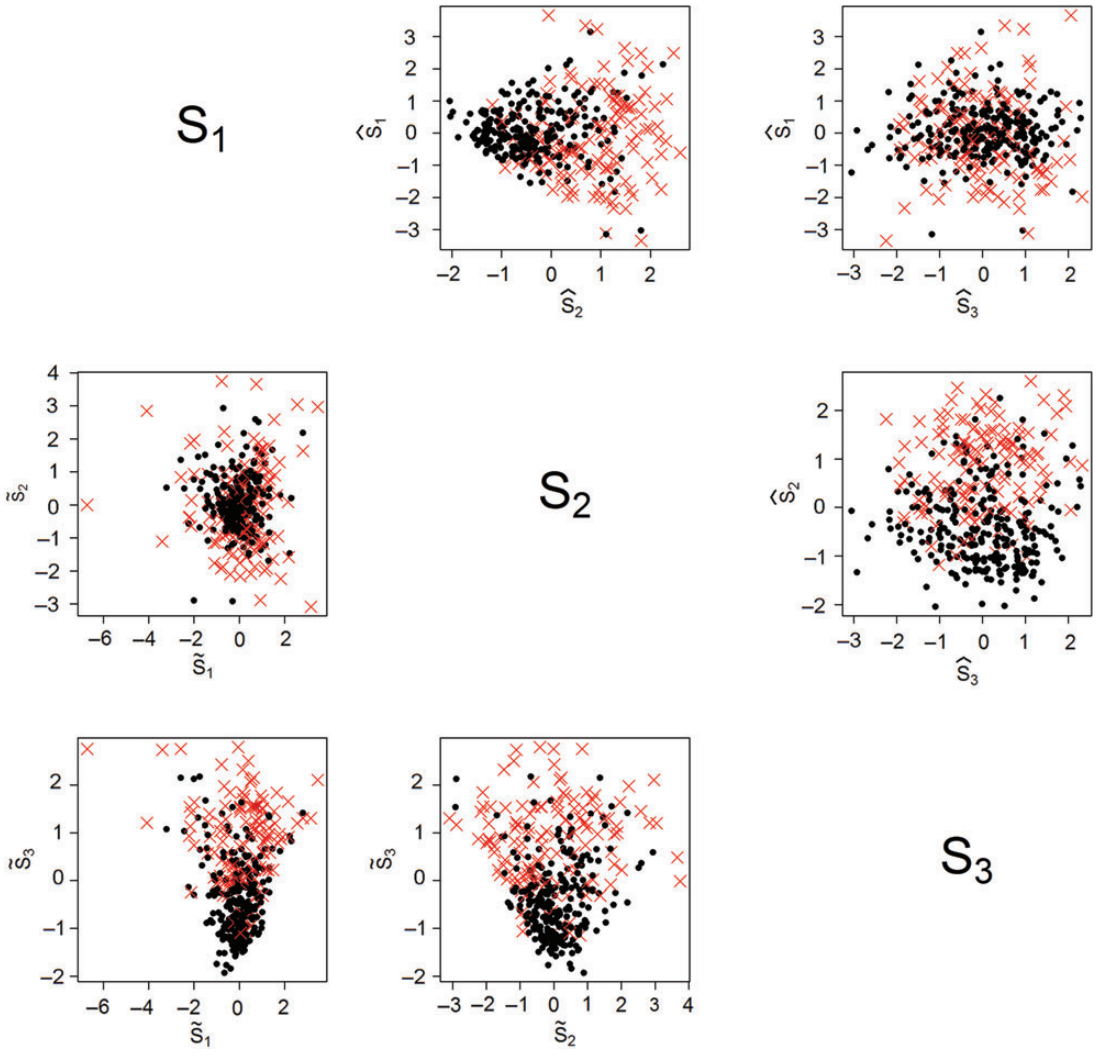

Fig. 4. The scatter plot matrix of the leading three SDR components from the two-stage method in the upper triangular panels, and from the direct method in the lower triangular panels. Here • and × indicate the normal and diabetes patients, respectively.

the assumption of missing at random: $R \perp W|(Y, X)$. Different from the conventional multiple imputation (Rubin, 1996), pmf imputation conducts the analysis "once" based on the pseudo-data $\mathcal{D}_s = \{(Y_i, X_i, w) : i = 1, \ldots, n, w = 1, \ldots, C\}$ with the weight $p_{iw} = R_i I(w = W_i) + (1 - R_i)P(W = w|Y = Y_i, X = X_i)$ for the ith subject having $W_i = w$. Ding and Wang (2011) further propose a smoothing estimator for $P(W = w|Y = Y_i, X = X_i)$. Consequently, the modified two-stage method constructs $\hat{B}$ as (3.8) to estimate $\mathcal{S}_{Y|X}$, except that the weighted versions of SDR methods (e.g., SIR, PSIR, and SAVE) using the pseudo-data $\mathcal{D}_s$ with weights $\{p_{iw} : i = 1, \ldots, n, w = 1, \ldots, C\}$ are implemented instead.

The idea of utilizing additional information to improve statistical inference has been discussed in the word of the SDR literatures, and can be formulated into the framework of this work. In Ding and Wang

(2011), the authors aim to use (Y, Z, δ) to enhance estimating $\mathcal{S}_{Y|Z}$, where δ is the missing indicator of Y. In this case, $\mathcal{S}_{Y|Z} \subseteq \mathcal{S}_{(Y,\delta)|Z}$, and the information of δ can be utilized via using $\mathcal{S}_{\text{env}} = \mathcal{S}_{(Y,\delta)|Z}$. In Hung (2012), the author aims to use (Y, Z) to more efficiently estimate $\mathcal{S}_{g(Y)|Z}$ for a given function g. In this case, $\mathcal{S}_{g(Y)|Z} \subseteq \mathcal{S}_{Y|Z}$ suggests using $\mathcal{S}_{\text{env}} = \mathcal{S}_{Y|Z}$ to utilize the whole information of Y. We note that the success of the two-stage method also relies on a good estimate of the underlying $\mathcal{S}_{\text{env}}$. One of our contributions is to develop the hybrid kernel matrix $\hat{K}(\omega)$ to estimate $\mathcal{S}_{\text{env}}$ and, via the tuning of ω, $\hat{B}$ with $\hat{K}_{\text{env}} = \hat{K}(\omega)$ can adapt to various situations of W. The adaptive method also has limitation. For example, we can hardly define the role of W in Hung (2012), i.e., the "residual" response of Y after removing $g(Y)$. It is of interest to extend the idea of $\hat{K}(\omega)$ to adapt to this situation.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors thank the editor, associate editor, and two anonymous referees for valuable comments that substantially improve the paper, and thank Dr Su-Yun Huang for helpful discussion. *Conflict of Interest*: None declared.

FUNDING

H.H. is supported by the Ministry of Science and Technology of Taiwan (102-2628-M-002-005-MY2).

REFERENCES

CHIAROMONTE, F., COOK, R. D. AND LI, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics* **30**, 475–497.

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

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.

DING, X. AND WANG, Q. (2011). Fusion-refinement procedure for dimension reduction with missing response at random. *Journal of the American Statistical Association* **106**, 1193–1207.

HENDERSON, H. V. AND SEARLE, S. R. (1979). Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canadian Journal of Statistics* **7**, 65–81.

HUNG, H. (2012). A two-stage dimension reduction method for transformed response and its applications. *Biometrika* **99**, 865–877.

LI, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86**, 316–327.

LI, B., ARTEMIOU, A. AND LI, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics* **39**, 3182–3210.

LI, K. C., WANG, J. L. AND CHEN, C. (1999). Dimension reduction for censored regression data. *The Annals of Statistics* **27**, 1–23.

Li, B., Wen, S. and Zhu, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of the Acoustical Society of America* **103**, 1177–1186.

Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review* **81**, 134–150.

Naik, P. A. and Tsai, C. L. (2005). Constrained inverse regression for incorporating prior information. *Journal of the American Statistical Association* **100**, 204–211.

Rubin, D. R. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489.

Wu, H. M. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics* **17**, 590–610.

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.

Yeh, Y. R., Huang, S. Y. and Lee, Y. Y. (2009). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Transactions on Knowledge and Data Engineering* **21**, 1590–1603.

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

[*Received April 10, 2015; revised November 2, 2015; accepted for publication November 15, 2015*]