Counting process-based dimension reduction methods for censored outcomes

By QIANG SUN

Department of Statistical Sciences, University of Toronto, 100 St George Street, Toronto, Ontario M5S 3G3, Canada

qsun@utstat.toronto.edu

RUOQING ZHU

Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, Illinois 61820, U.S.A.

rqzhu@illinois.edu

TAO WANG

Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang District, Shanghai 200240, China

neowangtao@sjtu.edu.cn

AND DONGLIN ZENG

Department of Biostatistics, University of North Carolina at Chapel Hill, 3101 McGavran-Greenberg Hall, Chapel Hill, North Carolina 27599, U.S.A.

dzeng@email.unc.edu

SUMMARY

We propose counting process-based dimension reduction methods for right-censored survival data. Semiparametric estimating equations are constructed to estimate the dimension reduction subspace for the failure time model. Our methods address two limitations of existing approaches. First, using the counting process formulation, they do not require estimation of the censoring distribution to compensate for the bias in estimating the dimension reduction subspace. Second, the nonparametric estimation involved adapts to the structural dimension, so our methods circumvent the curse of dimensionality. Asymptotic normality is established for the estimators. We propose a computationally efficient approach that requires only a singular value decomposition to estimate the dimension reduction subspace. Numerical studies suggest that our new approaches exhibit significantly improved performance. The methods are implemented in the R package orthoDr.

Some key words: Estimating equation; Semiparametric inference; Sliced inverse regression; Sufficient dimension reduction; Survival analysis.

1. Introduction

Dimension reduction is important in regression analysis. Its goal is to extract a low-dimensional subspace from a p-dimensional covariate $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ in order to predict an outcome of interest T. The dimension reduction literature often assumes the multiple-index model

$$T = h(B^{\mathrm{T}}X, \epsilon), \tag{1}$$

where ϵ is a random error independent of X, $B \in \mathbb{R}^{p \times d}$ is a coefficient matrix with $d < p$, and $h(\cdot)$ is an unknown link function. This model is equivalent to assuming $T \perp\!\!\!\perp X \mid B^{\mathrm{T}}X$ (Li, 1991). Since any d linearly independent vectors in the linear space spanned by the columns of B also satisfy model (1) for some h, we denote this linear subspace by $\mathcal{S}(B)$. The intersection of all subspaces satisfying $T \perp\!\!\!\perp X \mid B^{\mathrm{T}}X$ is called the central subspace, $\mathcal{S}_{T|X}$, whose dimension is referred to as the structural dimension. According to Cook (2009), $\mathcal{S}_{T|X}$ is uniquely defined under mild conditions. The goal of sufficient dimension reduction is to determine the structural dimension and the central subspace using data.

There is an extensive literature on estimating the central subspace for completely observed data, including Li (1991), Cook & Weisberg (1991), Zhu et al. (2006), Li & Wang (2007), Xia (2007), and Ma & Zhu (2012). When T is subject to right censoring, model (1) includes many well-known survival models as special cases, such as the proportional hazards model (Cox, 1972), the accelerated failure-time model (Lin et al., 1998), and linear transformation models (Zeng & Lin, 2007).

There has been limited work on estimating the dimension reduction subspace in the presence of censored observations. Li et al. (1999) propose a modified sliced inverse regression method that uses the estimate of the conditional survival function to account for censored cases. Xia et al. (2010) propose to estimate the conditional hazard function nonparametrically and use its gradient to construct the dimension reduction directions. In Li et al. (1999), p-dimensional kernel estimation is used to compensate for the bias caused by censoring, while in Xia et al. (2010), the estimation procedure requires a p-dimensional kernel estimate of the hazard function to provide reliable initial values, and then gradually reduces the working dimension to d. These methods suffer from the curse of dimensionality. When p is not small, alternative approaches such as that of Lu & Li (2011) adopt an inverse probability weighting scheme, which implicitly requires the correct specification of the censoring mechanism.

In this paper, we propose a counting process-based dimension reduction framework that leads to four different approaches. The proposed methods address several limitations of the existing work. First, our framework is built upon a counting process representation of the underlying survival model. This framework allows construction of doubly robust estimating equations, and the resulting estimators are more stable than existing ones such as in Xia et al. (2010). Our formulation can avoid the linearity assumption (Li, 1991) and the estimation of any censoring distribution, which are necessary in Li et al. (1999) and Lu & Li (2011). Second, the proposed framework is adaptive to the structural dimension in the sense that the nonparametric estimation involved depends only on the dimension of $\mathcal{S}(B)$, which is usually small, thus circumventing the curse of dimensionality. To this end, the proposed method shares advantages similar to that in Xia et al. (2010). Computationally, we use an optimization technique (Wen & Yin, 2013) on the Stiefel manifold to solve the estimating equations, which is numerically stable and fast. Last, under restrictive assumptions, our method reduces to a singular value decomposition, which can directly estimate the dimension reduction subspace without nonparametric estimation.

2. Proposed methods

2.1. *Semiparametric estimating equations for the central subspace*

Throughout the paper, we denote the failure time by T and the censoring time by C. Let $Y = \min(T, C)$ and $\delta = I(T \leqslant C)$ be the observed event time and the censoring indicator. We assume that C is independent of T conditional on X. Let $N(u) = I(Y \leqslant u, \delta = 1)$ and $Y(u) = I(Y \geqslant u)$ denote the observed counting process and the at-risk process, respectively. Let $\lambda(u \mid X)$ be the conditional hazard for T given X. According to Xia et al. (2010), model (1) is equivalent to $\lambda(u \mid X) = \lambda(u \mid B^{\mathrm{T}}X)$. Let

$$\mathrm{d}M(u, X) = \mathrm{d}M(u, B^{\mathrm{T}}X) = \mathrm{d}N(u) - \lambda(u \mid B^{\mathrm{T}}X)Y(u)\mathrm{d}u$$

be the martingale increment process indexed by u. This paper considers constructing estimation equations that are based on the counting process representation of the survival model. To derive the estimating equations, we follow Bickel et al. (1993) and Tsiatis (2007) to obtain the ortho-complement of the nuisance tangent space at B as

$$\mathcal{E}^{\perp} = \left\{ \int \left\{ \alpha(u, X) - \alpha^{*}(u, B^{\mathrm{T}}X) \right\} \mathrm{d}M(u, X) : \alpha(u, X) \text{ is measurable in } X \text{ and } u \right. , \quad (2)$$

where $\alpha^{*}(u, B^{\mathrm{T}}X) = E\{\alpha(u, X) \mid \mathcal{F}_u, B^{\mathrm{T}}X\}$ and $\mathcal{F}_u$ is the filtration; see the Supplementary Material. To estimate B, we consider the unbiased estimating equations

$$E\left[\int \left\{ \alpha(u, X) - \alpha^{*}(u, B^{\mathrm{T}}X) \right\} \left\{ \mathrm{d}N(u) - \lambda(u \mid B^{\mathrm{T}}X)Y(u)\,\mathrm{d}u \right\} \right] = 0. \quad (3)$$

The sample versions based on n independent and identical copies $\{Y_i, \delta_i, X_i\}_{i=1}^{n}$ are

$$\frac{1}{n} \sum_{i=1}^{n} \left[\int \left\{ \alpha(u, X_i) - \alpha^{*}(u, B^{\mathrm{T}}X_i) \right\} \left\{ \mathrm{d}N_i(u) - \lambda(u \mid B^{\mathrm{T}}X_i)Y_i(u)\,\mathrm{d}u \right\} \right] = 0, \quad (4)$$

where the conditional hazard function will be estimated using data. For some particular choices of $\alpha(u, X)$, this can be implemented using the generalized method of moments (Hansen, 1982):

$$\arg\min_{B} \left\{ \psi_n(B)^{\mathrm{T}} \psi_n(B) \right\}, \quad (5)$$

where $\psi_n(B)$ is the vectorized left-hand side of (4). We estimate several quantities in $\psi_n(B)$ nonparametrically. For example, the conditional hazard function $\lambda(u \mid B^{\mathrm{T}}X_i)$ at any time-point u can be estimated by

$$\hat{\lambda}(u \mid B^{\mathrm{T}}X = z) = \frac{\sum_{i=1}^{n} K_b(Y_i - u)\delta_i K_h(B^{\mathrm{T}}X_i - z)}{\sum_{j=1}^{n} I(Y_j \geqslant u)K_h(B^{\mathrm{T}}X_j - z)} \quad (6)$$

for bandwidths b and h, where $K_h(\cdot)$ is a d-dimensional multivariate kernel function. We defer the details to § 3.

It is crucial to choose specific forms of $\alpha(u, X)$. Different choices may simplify the above formulation or may have theoretical or computational advantages. In the following two subsections, we present four different choices, which fall into two categories: the forward and inverse regression schemes. The main difference between the two schemes lies in whether the counting process $N(u)$ is used in the definition of $\alpha(u, X)$. The forward regression scheme is essentially the estimating equation approach, while the inverse regression scheme uses $N(u)$ to mimic the sliced inverse regression (Li, 1991) conceptually.

2.2. *Forward regression*

In the forward regression scheme, we choose $\alpha(u, X)$ to not depend on the observed failure process $N(u)$. Provided that $\alpha(u, X)$ depends at most on the at-risk process $Y(u)$, we can simplify the estimating equations in (3) to

$$E\left(\int \left[\alpha(u, X) - E\big\{\alpha(u, X) \mid Y(u) = 1, B^{\mathrm{T}}X\big\}\right] \mathrm{d}N(u) \right) = 0.$$

We now give one example of $\alpha(u, X)$ when the structural dimension $d = 1$. This requires only scalar nonparametric estimation.

Example 1. With $\alpha(u, X) = X$, the population versions of the p-dimensional estimating equations are

$$E\left(\int \left[X - E\big\{X \mid Y(u) = 1, B^{\mathrm{T}}X\big\}\right] \mathrm{d}N(u) \right) = 0, \tag{7}$$

which reduce to the efficient estimating equations for the proportional hazards model when the exponential link is known to be correct. This can also be used for the transformation models in Zeng & Lin (2007). For some simple extensions, we could let $\alpha(u, X) = E\{XY(u)\}X^{\mathrm{T}}$ to obtain p-by-p estimating equations, in order to handle the case of $d > 1$. To implement the forward regression method in (7), we can estimate $\psi_n(B)$ in (5) using

$$\hat{\psi}_n(B) = \frac{1}{n}\sum_{i=1}^{n}\big\{X_i - \hat{E}(X \mid Y \geqslant Y_i, B^{\mathrm{T}}X_i)\big\}\delta_i, \tag{8}$$

where for any given u and z,

$$\hat{E}(X \mid Y \geqslant u, B^{\mathrm{T}}X = z) = \frac{\sum_{i=1}^{n} X_i I(Y_i \geqslant u) K_h(B^{\mathrm{T}}X_i - z)}{\sum_{i=1}^{n} I(Y_i \geqslant u) K_h(B^{\mathrm{T}}X_i - z)} \tag{9}$$

for some choice of kernel function $K_h(\cdot)$ with bandwidth parameter h; see § 3.

2.3. *Inverse regression*

In this subsection, we focus on the inverse regression scheme. Our motivation is the following counting process representation of the model:

$$\big\{\mathrm{d}N(u) \mid Y(u) = 1, B^{\mathrm{T}}X\big\} \sim \mathrm{Ber}\big\{\lambda(u \mid B^{\mathrm{T}}X)\mathrm{d}u\big\},$$

where $dN(t) = N(t+dt) - N(t)$. Hence, we can consider the sliced conditional mean of X given the outcome of $dN(t)$ in the risk set, that is, $Y(t) = 1$. This leads to the construction of a local mean difference for the binary outcome $dN(u)$:

$$\varphi(u) = E\{X \mid dN(u) = 1, Y(u) = 1\} - E\{X \mid dN(u) = 0, Y(u) = 1\}. \tag{10}$$

The outcome $dN(u)$ conditioning on the event $Y(u) = 1$ depends only on $\lambda(u \mid B^{\mathrm{T}}X)$. Hence the inverse regression curve $\varphi(u)$ is contained within the central subspace $\mathcal{S}_{T|X}$. With this choice of $\varphi(u)$, we consider the function

$$\alpha(u, X) = X\varphi^{\mathrm{T}}(u). \tag{11}$$

Then

$$\alpha(u, X) - \alpha^*(u, B^{\mathrm{T}}X) = \left[X - E\{X \mid Y(u) = 1, B^{\mathrm{T}}X\}\right]\varphi^{\mathrm{T}}(u), \tag{12}$$

which can be estimated by combining the estimate of $E\{X \mid Y(u) = 1, B^{\mathrm{T}}X\}$ in (9) and that of $\varphi(u)$ in (10):

$$\hat{\varphi}(u) = \frac{\sum_{i=1}^n X_i I\left(u \leqslant Y_i < u + h, \delta_i = 1\right)}{\sum_{i=1}^n I\left(u \leqslant Y_i < u + h, \delta_i = 1\right)} - \frac{\sum_{i=1}^n X_i I\left(Y_i \geqslant u\right)}{\sum_{i=1}^n I\left(Y_i \geqslant u\right)}. \tag{13}$$

Based on this choice of α, we propose two approaches that use the estimating equations (3) and a computationally efficient approach that further simplifies the formula to a singular value decomposition.

Example 2. Replacing $\alpha(u, X) - \alpha^*(u, B^{\mathrm{T}}X)$ in (3) by (12) leads to estimating equations of a semiparametric inverse regression approach:

$$E\left(\int \left[X - E\{X \mid Y(u) = 1, B^{\mathrm{T}}X\}\right]\varphi^{\mathrm{T}}(u)\, dM(u)\right) = 0. \tag{14}$$

This approach consists of $p \times p$ estimating equations, and is able to handle the case of $d > 1$. However, the nonparametric estimation part is only d-dimensional, as reflected by $B^{\mathrm{T}}X$. Furthermore, this formulation enjoys the double robustness property, illustrated in the Supplementary Material. A similar phenomenon has been observed by Ma & Zhu (2012) in the setting without censoring. This suggests that if one of $E\{X \mid Y(u) = 1, B^{\mathrm{T}}X\}$ and $M(u)$ is estimated incorrectly, we can still obtain consistent estimators of the dimension reduction subspace. In our numerical experiment, we observe a numerical advantage of this approach over its simplified version, which is given in Example 3.

To implement this method, we estimate $\psi_n(B)$ in (5) by

$$\text{vec}\left[\frac{1}{n}\sum_{i=1}^n \sum_{\substack{j=1 \\ \delta_j=1}}^n \left\{X_i - \hat{E}(X \mid Y \geqslant Y_j, B^{\mathrm{T}}X_i)\right\} \hat{\varphi}^{\mathrm{T}}(Y_j) \left\{\delta_i I(j = i) - \hat{\lambda}(Y_j \mid B^{\mathrm{T}}X_i)\right\}\right], \tag{15}$$

where $\hat{E}\{X \mid Y \geqslant u, B^{\mathrm{T}}X = z\}$ and $\hat{\varphi}^{\mathrm{T}}(u)$ are given in (9) and (13), respectively, and the conditional hazard function can be estimated by (6). We finally apply the generalized method of moments to estimate B.

Example 3. Similar to Example 1, our choice of $\alpha(u, X)$ in (11) depends on at most the at-risk process $Y(u)$. Hence, the estimating functions in (14) can be simplified to

$$E\left\{\int\left[X - E\{X \mid Y(u) = 1, B^{\mathrm{T}}X\}\right]\varphi^{\mathrm{T}}(u)\,\mathrm{d}N(u)\right\}. \tag{16}$$

Replacing $\mathrm{d}M(u)$ with $\mathrm{d}N(u)$ greatly reduces the computational burden. This can be seen from (15), where a conditional hazard function $\hat{\lambda}(Y_j \mid B^{\mathrm{T}}X_i)$ needs to be evaluated at each observed failure time-point j for all observations i. Using this simplification, we lose the double robustness property. The implementation of this approach is a simplified version of that in Example 2 with

$$\hat{\psi}_n(B) = \mathrm{vec}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{X_i - \hat{E}(X \mid Y \geqslant Y_i, B^{\mathrm{T}}X_i)\right\}\delta_i\hat{\varphi}^{\mathrm{T}}(Y_i)\right], \tag{17}$$

where the estimators of the nonparametric components are the same as before.

Example 4. With additional assumptions, B can be estimated without nonparametric smoothing. We need the following definitions.

DEFINITION 1. *For any $\alpha \in \mathbb{R}^p$ and any $u > 0$, the linearity condition (Li, 1991) is satisfied conditioning on the event $\{Y(u) = 1\}$, i.e.,*

$$E\{\alpha^{\mathrm{T}}X \mid Y(u) = 1, B^{\mathrm{T}}X = z\} = c_0(u) + c^{\mathrm{T}}(u)z, \tag{18}$$

where $c_0(u)$ and $c(u)$ are constants that possibly depend on u. Furthermore, the time-invariant covariance condition requires

$$\mathrm{cov}\{X \mid Y(u) = 1\} = c_1(u)\Sigma, \tag{19}$$

where $c_1(u)$ is some constant depending on u.

After centring X at time u, if (18) and (19) are satisfied, we have

$$E\{X \mid Y(u) = 1, B^{\mathrm{T}}X\} - E\{X \mid Y(u) = 1\} = P[X - E\{X \mid Y(u) = 1\}], \tag{20}$$

where $P = \Sigma B(B^{\mathrm{T}}\Sigma B)^{-1}B^{\mathrm{T}}$ and the constant term $c_1(u)$ vanishes. By (19), P remains the same across all time-points. Hence, inserting (20) into (16) leads to

$$Q E\left\{\int\left[X - E\{X \mid Y(u) = 1\}\right]\varphi^{\mathrm{T}}(u)\,\mathrm{d}N(u)\right\} = 0,$$

where $Q = I - P$. This is equivalent to deriving the left-singular space of the covariance matrix

$$E\left\{\int\left[X - E\{X \mid Y(u) = 1\}\right]\varphi^{\mathrm{T}}(u)\,\mathrm{d}N(u)\right\}. \tag{21}$$

The computation of this approach is extremely simple. Since $\mathrm{d}N(u)$ takes value 1 at no more than one time-point in the entire time domain, the covariance matrix can be estimated directly. Then we perform singular value decomposition on this sample covariance matrix and obtain its leading left-singular vectors as our final estimators. Details are provided in Algorithm 1.

Remark 1. The two conditions imposed in Example 4 are restrictive and do not always hold. For example, since $Y(u)$ is a process that depends on both the failure and the censoring distribution, as long as the censoring distribution depends on structures beyond $B^\mathrm{T}X$, the conditions could be violated. Nevertheless, many recent papers argue that sliced inverse regression performs well empirically even when the linearity condition fails (Li & Dong, 2009; Dong & Li, 2010). Hence, this does not prevent the method from serving as an exploratory tool. The method is also practically useful since it provides an initial value for solving our other estimation approaches.

3. IMPLEMENTATION AND ALGORITHMS

Implementation of the method in (21) is straightforward. Algorithm 1 summarizes the estimation procedure.

Algorithm 1. Algorithm for the computationally efficient approach.
Input: $\{(X_i, \delta_i, Y_i), 1 \leqslant i \leqslant n\}$, $h > 0$, $k > 0$.
Step 1: For each Y_i such that $\delta_i = 1$, calculate $\hat{\varphi}(Y_i)$ using equation (13) and calculate
$\hat{E}(X \mid Y > Y_i)$ using $\hat{E}(X \mid Y > u) = \{\sum_{i=1}^n I(Y_i > u)\}^{-1} \{\sum_{i=1}^n X_i I(Y_i > u)\}$.
Step 2: Calculate $\hat{M} = n^{-1} \sum_{\delta_i=1} \{X_i - \hat{E}(X \mid Y_i)\} \hat{\varphi}^\mathrm{T}(Y_i)$.
Step 3: Perform the singular value decomposition, $\hat{M} = \hat{U}\hat{D}\hat{V}^\mathrm{T}$.
Output: $\hat{B}$ as the first k columns of $\hat{U}$.

Advanced numerical optimization techniques are needed to solve the estimating equations of the forward regression approach in (7) and the two inverse regression approaches in (14) and (16). For all three, we solve for the minimizer of $\hat{\psi}_n(B)^\mathrm{T}\hat{\psi}_n(B)$, where $\hat{\psi}_n(B)$ is specified in (8), (15) and (17) respectively. Existing methods use general-purpose optimization tools such as the Newton–Raphson algorithm to solve for the minimizer, but dimension reduction methods create an additional difficulty because B is not uniquely defined and this causes numerical instability. To tackle this, Ma & Zhu (2012) propose to take a selected set of d rows of B to be an identity matrix and solve for the other parameters. This approach requires knowledge of locations of important variables. Instead, we propose an orthogonality-constrained optimization approach to solve our semiparametric estimating equations within the Stiefel manifold (Edelman et al., 1998):

$$\begin{aligned} \text{minimize} \quad & \hat{\psi}_n(B)^\mathrm{T}\hat{\psi}_n(B) \\ \text{subject to} \quad & B^\mathrm{T}B = I_{d \times d}. \end{aligned}$$

This optimization approach preserves the rank d of the column space of B while not directly restricting its entries.

The main machinery of the algorithm evolved from a first-order descent algorithm proposed by Wen & Yin (2013), which preserves the update of the parameters within the manifold. In particular, let the gradient matrix be defined as

$$G = \frac{\partial \, \hat{\psi}_n(B)^\mathrm{T}\hat{\psi}_n(B)}{\partial B}.$$

Then, utilizing the Cayley transformation, we can update B to

$$B(\tau_0) = \left(I + \frac{\tau_0}{2}A\right)^{-1}\left(I - \frac{\tau_0}{2}A\right)B,$$

where $A = GB^T - BG^T$ is a skew-symmetric matrix and τ_0 is a step size. In practice, τ_0 can be chosen using inexact line search by incorporating the Wolfe conditions (Nocedal & Wright, 2006). It can easily be verified that if $B^T B = I$, then $B(\tau_0)^T B(\tau_0) = I$ for any $\tau_0 > 0$. In this way, the algorithm preserves the constraint exactly. As with classical dimensional reduction methods, our method recovers the column space of B rather than treating each entry as a fixed parameter. Moreover, if an upper block-diagonal version is desired, we can easily convert the obtained solutions through linear transformations. However, in this case, we can select the largest entries in the estimated $\hat{B}$ as the location of the diagonal matrix, instead of prespecifying the locations. Algorithm 2 summarizes the details.

Algorithm 2. The orthogonality constrained optimization algorithm.
Input: ε_0, $\{(X_i, \delta_i, Y_i), 1 \leqslant i \leqslant n\}$.
Initialize: Obtain $B^{(0)}$ from the computationally efficient approach in Algorithm 1.
For $k = 1$ to $k = $ max.iter:
　　Numerically approximate the gradient matrix G at $B^{(k)}$.
　　Compute the skew-symmetric matrix $A = GB^T - BG^T$.
　　Perform line search for τ_0 on the path $B(\tau_0) = \left(I + \frac{\tau_0}{2}A\right)^{-1}\left(I - \frac{\tau_0}{2}A\right)B$.
　　Update $B^{(k+1)} = B(\tau_0)$.
　　Stop if $\left\|B^{(k+1)} - B^{(k)}\right\|_2 \leqslant \varepsilon_0$.
Output: $\hat{B} = B^{(k+1)}$.

The iteration is stopped when a prespecified optimization precision ε_0 is reached. To estimate the nonparametric components (6) and (9), we exploit a multivariate Gaussian kernel with a diagonal bandwidth matrix such that the bandwidth for the jth variable is taken as $h = \left\{4/(d + 2)\right\}^{1/(d+4)} n^{-1/(d+4)} \hat{\sigma}_j$ (Silverman, 1986), where $\hat{\sigma}_j$ is the sample standard deviation of the jth variable in $B^T X$. In our numerical implementation, we simply standardize all d coordinates of $B^T X$, so that $\hat{\sigma}_j = 1$ for all j. We implemented the algorithm in the R package orthoDr (Zhao et al., 2017; R Development Core Team, 2019).

4. ASYMPTOTIC NORMALITY

We prove asymptotic normality of the proposed estimators. Without loss of generality, we focus on the semiparametric inverse regression approach but with general $\alpha(u, X)$ and $\alpha^*(u, B^T X)$, in which we obtain $\hat{B}$ by solving

$$\frac{1}{n}\text{vec}\left[\sum_{i=1}^{n}\int_0^{\tau}\left\{\alpha(u, X_i) - \hat{\alpha}^*(u, \hat{B}^T X_i)\right\}d\hat{M}(u, \hat{B}^T X_i)\right] = 0.$$

To address the identifiability of B, we restrict our attention to matrices in the form of $B = (B_u^T, B_\ell^T)^T$, where the upper submatrix $B_u = I_d \in \mathbb{R}^{d \times d}$ is the $d \times d$ identity matrix. In this manner, we can view B_ℓ as the unique parameterization of the subspace $\mathcal{S}(B)$. We then write $\beta_\ell = \text{vecl}(B) = \text{vec}(B_\ell)$, the vector concatenating all free parameters in B. We need the following regularity assumptions.

Assumption 1. There exists τ such that $0 < \tau < \infty$ and $\text{pr}(Y > \tau \mid X) > 0$.

Assumption 2. Let $f_{B^{\mathrm{T}}X}(z)$ be the density function of $B^{\mathrm{T}}X$ evaluated at $z = B^{\mathrm{T}}x$, let $f(t,z)$ be the density of T given $B^{\mathrm{T}}X = z$, and let $S(t,x) = \mathrm{pr}(T \geqslant t \mid X = x)$ and $S_c(t,x) = \mathrm{pr}(C \geqslant t \mid X = x)$. Assume that $f(t,z), f_{B^{\mathrm{T}}X}(z), S(t,z)$ and $E\{S_c(t,X) \mid z\}$ are bounded and have bounded first and second derivatives with respect to t and z, and that $S(t,z)$ is bounded away from zero.

Assumption 3. The univariate kernel function $K(x)$ is symmetric with $\int x^2 K(x)\,\mathrm{d}x < \infty$. The d-dimensional kernel function is a product of d univariate kernel functions; that is, $K(u) = \prod K(u_j)$ for $u = (u_1, \ldots, u_d)^{\mathrm{T}}$.

Assumptions 1 and 2 are standard in survival analysis. Assumption 3 is commonly used in kernel estimation. Based on these assumptions, we provide convergence rates for our estimators of the conditional hazard function and its derivative. It is easy to see that the Silverman formula implemented in our numerical approach leads to consistent estimators.

LEMMA 1. *Under Assumptions 2 and 3, and assuming that the bandwidths satisfy $h, b \to 0$ and $nbh^{d+2} \to \infty$, we have that, uniformly for all t and z,*

$$\hat{\lambda}(t \mid z) = \lambda(t,z) + O_{\mathrm{p}}\left\{\left(nbh^d\right)^{-1/2} + h^2 + b^2\right\},$$

$$\frac{\partial}{\partial z}\hat{\lambda}(t \mid z) = \frac{\partial}{\partial z}\lambda(t,z) + O_{\mathrm{p}}\left\{\left(nbh^{d+2}\right)^{-1/2} + h^2 + b^2\right\}.$$

Before presenting our main theorem, we need the convergence of the α^* functions. However, we do not want the theoretical result to be limited to the choice in equation (2), so we provide results for any valid α^*, provided the following condition is satisfied.

Assumption 4. For some $\kappa < 1/2$, the convergence rate for the following conditional nonparametric estimation holds uniformly over all u and z:

$$\mathrm{vec}\left\{\hat{\alpha}^*(u,z) - \alpha^*(u,z)\right\} = O_{\mathrm{p}}\left(n^{-1/2+\kappa}\right),$$

$$\frac{\partial}{\partial z}\mathrm{vec}\left\{\hat{\alpha}^*(u,z) - \alpha^*(u,z)\right\} = O_{\mathrm{p}}\left(n^{-1/2+\kappa}\right).$$

For most choices, such as a kernel estimator of the conditional density, when the dimension d is fixed, the rate in Lemma 1 can be achieved for $\hat{\alpha}^*(u,z)$, while for conditional expectation estimation, the classical rate of $O_{\mathrm{p}}\{(nh^d)^{-1/2} + h^2\}$ can be attained. Hence, with a proper choice of the bandwidth, the rates in Assumption 4 can usually be guaranteed. We present the main theorem.

THEOREM 1 (Asymptotic normality). *Under Assumptions 1–4 and the choice of bandwidths specified in Lemma 1, the estimator $\mathrm{vecl}(\hat{B})$ is asymptotically normal, that is, $n^{1/2}(\hat{\beta}_\ell - \beta_\ell) \to N(0, \Sigma)$, where $\Sigma = (G^{\mathrm{T}}G)^{-1}G\Sigma_A G^{\mathrm{T}}(G^{\mathrm{T}}G)^{-1}$, with*

$$\Sigma_A = \mathrm{cov}\{A(\tau)\} = \mathrm{cov}\left[\int_0^\tau \mathrm{vec}\left\{\alpha(u,X) - \alpha^*(u, B^{\mathrm{T}}X)\right\}\mathrm{d}M(u, B^{\mathrm{T}}X)\right],$$

$$G = E\left(\frac{\partial}{\partial \beta_\ell}\int_0^\tau \mathrm{vec}\left[\left\{\alpha(u,X_i) - \alpha^*(u, B^{\mathrm{T}}X)\right\}\right]\mathrm{d}M(u, B^{\mathrm{T}}X)\right).$$

5. NUMERICAL EXAMPLES

5.1. *Simulation studies*

We examine the finite-sample performance of our proposed methods via numerical experiments. We estimate the dimension reduction subspace using the forward regression approach (7), the semiparametric inverse regression approach (14), the counting process inverse regression approach (16), and the computationally efficient approach (21). All of our methods are implemented in the `orthoDr` package in R. Four alternative approaches are considered: a naive approach that performs sliced inverse regression on the failure observations, carried out using the `dr` package (Weisberg, 2002); the double slicing approach (Li et al., 1999) using the R package `censorSIR` provided by Wu et al. (2008); the minimal average variance estimation based on hazard functions in Xia et al. (2010); and the inverse probability-of-censoring weighted approach based on Lu & Li (2011). Xia et al. (2010)'s approach is implemented through Matlab, provided at Prof. Xia's website. We carry out Lu & Li (2011)'s approach ourselves by using a proportional hazards model to estimate the censoring weights and obtain the reduced space by using the `dr` package with subject weights.

We consider four different settings. Setting 1 is a classical proportional hazards model. Setting 2 is set up with structural dimension $d = 2$ and with directions in the hazard function changing over time. Setting 3 has the structural dimension equal to 2, with the two directions interacting with each other. Setting 4 also has two interacting structural dimensions, while the failure and censoring variables overlap. For each setting, we consider $p = 6$, 12 and 18. Each experiment is repeated 200 times with sample size $n = 400$.

In Setting 1, the true survival time T and the censoring time C are generated from exponential distributions with rates $\exp(\beta^{\mathrm{T}} X)$ and $\exp(X_4 + X_5 - 1)$ respectively, where $\beta = (1, 0.5, 0, \ldots, 0)^{\mathrm{T}}$ and X_j is the jth element of X, for $1 \leqslant j \leqslant p$. The covariate X follows the multivariate normal distribution with mean zero and covariance $\Sigma = \left(0.5^{|i-j|}\right)_{ij}$. The overall censoring rate is around 35%.

In Setting 2, we generate T_1 and T_2 from exponential distributions with rates $\exp(\beta_1^{\mathrm{T}} X)$ and $\exp(\beta_2^{\mathrm{T}} X)$ respectively, where $\beta_1 = (1, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$ and $\beta_2 = (0, 1, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$. The true survival time $T = T_1 I(T_1 < 0.4) + (T_2 + 0.4) I(T_1 \geqslant 0.4)$. The censoring time C is generated from exponential distributions with rate $\exp(X_5 - X_6 - 2)$. The covariate X follows the same distribution as in Setting 1. The overall censoring rate is around 35%.

In Setting 3, the true survival time T is generated from a Weibull distribution with shape parameter 5 and scale parameter $\exp\{4\beta_2^{\mathrm{T}} X (\beta_1^{\mathrm{T}} X - 1)\}$, where $\beta_1 = (1, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$ and $\beta_2 = (0, 1, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$. The censoring time C is generated uniformly from 0 to $3 \exp(X_5 - X_6 + 0.5)$. We further draw X such that the X_js follow the standard uniform distribution $U(0, 1)$ independently. The overall censoring rate is around 34%.

In Setting 4, the true survival time T is generated from a proportional hazards model with $\log(T) = -2.5 + \beta_1^{\mathrm{T}} X + 0.5 \beta_1^{\mathrm{T}} X \beta_2^{\mathrm{T}} X + 0.25 \log\{-\log(1 - u)\}$ and $\log(C) = -0.5 + \beta_3^{\mathrm{T}} X + \log\{-\log(1 - u)\}$, where the us are independent and identically uniformly distributed, $\beta_1 = (1, 1, 0, \ldots, 0)^{\mathrm{T}}$, $\beta_2 = (0, 0, 1, -1, 0, \ldots, 0)^{\mathrm{T}}$, and $\beta_3 = (0, 1, 0, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$. The covariate X follows the same distribution as in Setting 1, except that $\Sigma = \left(0.25^{|i-j|}\right)$. The overall censoring rate is around 26%.

We investigate the statistical performance using the Frobenius norm distance between the projection matrix P and its estimator $\hat{P}$, where $P = B(B^{\mathrm{T}} B)^{-1} B^{\mathrm{T}}$, the trace correlation $\mathrm{tr}(P\hat{P})/d$, where d is the structural dimension, and the canonical correlation between $B^{\mathrm{T}} X$ and $\hat{B}^{\mathrm{T}} X$. The results are summarized in Table 1.

Table 1. *Simulation results: mean ($\times 10^2$) and standard deviation ($\times 10^2$, in parentheses) of the Frobenius norm distance, trace correlation and canonical correlation*

Setting 1 ($d = 1$)	$p = 6$			$p = 12$			$p = 18$		
	Frob	Tr	CCor	Frob	Tr	CCor	Frob	Tr	CCor
Naive	54 (12)	85 (6)	94 (3)	66 (12)	78 (8)	92 (3)	73 (11)	73 (8)	91 (3)
DS	33 (10)	94 (4)	98 (2)	46 (11)	89 (5)	97 (2)	53 (10)	85 (5)	96 (2)
IPCW-SIR	64 (13)	78 (9)	91 (4)	75 (11)	71 (9)	89 (4)	80 (11)	68 (9)	89 (4)
hMave	68 (12)	76 (8)	86 (5)	73 (11)	73 (8)	86 (5)	79 (10)	68 (8)	84 (5)
Forward	21 (6)	98 (1)	99 (0)	33 (8)	94 (3)	99 (1)	39 (7)	92 (3)	98 (1)
CP-SIR	26 (9)	96 (3)	99 (1)	40 (10)	91 (4)	98 (1)	49 (9)	88 (4)	97 (1)
IR-CP	23 (7)	97 (2)	99 (0)	35 (8)	94 (3)	98 (1)	41 (7)	91 (3)	98 (1)
IR-Semi	23 (8)	97 (2)	99 (0)	37 (8)	93 (3)	98 (1)	44 (8)	90 (4)	98 (1)
Setting 2 ($d = 2$)	$p = 6$			$p = 12$			$p = 18$		
	Frob	Tr	CCor	Frob	Tr	CCor	Frob	Tr	CCor
Naive	67 (19)	88 (6)	96 (3)	87 (18)	80 (8)	91 (6)	106 (17)	71 (9)	87 (6)
DS	44 (13)	95 (3)	98 (1)	68 (15)	88 (5)	94 (3)	84 (11)	82 (5)	92 (3)
IPCW-SIR	83 (19)	82 (8)	94 (3)	98 (17)	75 (9)	90 (6)	114 (16)	67 (9)	86 (8)
hMave	114 (31)	65 (16)	74 (16)	139 (19)	51 (12)	64 (12)	151 (14)	43 (10)	59 (10)
Forward	102 (1)	49 (1)	100 (0)	105 (2)	48 (1)	99 (0)	107 (2)	46 (1)	99 (0)
CP-SIR	37 (11)	96 (2)	98 (1)	61 (12)	90 (4)	96 (2)	78 (10)	85 (4)	93 (2)
IR-CP	49 (19)	93 (6)	96 (3)	73 (20)	86 (8)	92 (4)	90 (17)	79 (8)	89 (5)
IR-Semi	39 (14)	96 (3)	98 (2)	65 (16)	89 (6)	94 (3)	83 (15)	82 (6)	91 (3)
Setting 3 ($d = 2$)	$p = 6$			$p = 12$			$p = 18$		
	Frob	Tr	CCor	Frob	Tr	CCor	Frob	Tr	CCor
Naive	72 (23)	86 (9)	96 (5)	99 (22)	74 (11)	88 (11)	116 (18)	66 (10)	82 (13)
DS	40 (14)	95 (3)	99 (1)	60 (13)	91 (4)	97 (3)	73 (15)	86 (6)	95 (5)
IPCW-SIR	113 (26)	66 (13)	81 (14)	129 (15)	58 (9)	74 (12)	133 (11)	55 (7)	74 (12)
hMave	40 (18)	95 (6)	99 (3)	66 (27)	87 (12)	94 (9)	89 (29)	78 (14)	89 (12)
Forward	100 (0)	50 (0)	100 (0)	100 (0)	50 (0)	100 (0)	101 (0)	50 (0)	100 (0)
CP-SIR	34 (11)	97 (2)	99 (1)	55 (11)	92 (3)	97 (2)	67 (11)	88 (4)	96 (3)
IR-CP	30 (14)	97 (3)	99 (1)	46 (14)	94 (4)	99 (1)	58 (15)	91 (5)	97 (4)
IR-Semi	19 (8)	99 (1)	100 (0)	29 (8)	98 (1)	100 (0)	40 (11)	96 (2)	99 (1)
Setting 4 ($d = 2$)	$p = 6$			$p = 12$			$p = 18$		
	Frob	Tr	CCor	Frob	Tr	CCor	Frob	Tr	CCor
Naive	33 (9)	97 (2)	99 (1)	52 (10)	93 (3)	97 (1)	66 (10)	89 (4)	95 (2)
DS	49 (12)	94 (3)	95 (3)	62 (11)	90 (4)	94 (3)	71 (11)	87 (4)	93 (3)
IPCW-SIR	35 (9)	97 (2)	98 (1)	52 (10)	93 (3)	97 (1)	64 (10)	89 (3)	95 (2)
hMave	142 (3)	50 (2)	59 (4)	145 (5)	47 (4)	57 (4)	149 (7)	45 (6)	55 (6)
Forward	101 (0)	50 (0)	100 (0)	102 (1)	49 (0)	99 (0)	102 (1)	49 (0)	99 (0)
CP-SIR	36 (7)	97 (1)	98 (1)	51 (8)	93 (2)	97 (1)	63 (8)	90 (3)	95 (1)
IR-CP	22 (9)	99 (2)	99 (1)	42 (15)	95 (4)	98 (2)	57 (17)	91 (5)	96 (3)
IR-Semi	13 (5)	99 (0)	100 (0)	24 (7)	98 (1)	99 (0)	34 (10)	97 (2)	99 (1)

DS, method of Li et al. (1999); IPCW-SIR, method of Lu & Li (2011); hMave, method of Xia et al. (2010); Forward, forward regression; CP-SIR, the computationally efficient approach; IR-CP, the counting process inverse regression approach; IR-Semi, the semiparametric inverse regression approach; Frob, Frobenius norm distance; Tr, trace correlation; CCor, canonical correlation.

Overall, the two inverse regression methods achieve the best performance, followed by the computationally efficient approach. When no nonparametric approximation is required, the latter outperforms existing methods in almost all settings. Among the competing methods, double slicing performs best in general, while the methods of Xia et al. (2010) and Lu & Li (2011) outperform

double slicing in Settings 3 and 4, respectively. Regarding the three error measurements, the Frobenius norm distance is the most informative, while the trace and canonical correlations are less sensitive to the performances.

Of the two inverse regression methods, the semiparametric version is slightly better in Settings 3 and 4. The main advantage of the semiparametric version compared with the counting process version is its double robustness, which ensures consistency even when the conditional expectations are not estimated correctly. However, this theoretical advantage does not translate into strong numerical improvements in Settings 1 and 2, especially when p is large. This is possibly due to the variations in the hazard function estimation, which introduces numerical instability. In Setting 1, forward regression achieves the best performance. As discussed in Example 1, this method mimics the efficient estimating equations used in the proportional hazards model and is thus the most efficient method in this setting. In Setting 2, the computationally efficient approach performs similarly to the two inverse regression approaches and even outperforms them for large p. This demonstrates the potential of this approach in higher dimensional settings when nonparametric estimation may not be preferred.

One major challenge in solving the estimating equations is the computational burden, especially for equations with nonparametric components. Our method adds difficulties due to the extra orthogonality constraints $B^{\mathrm{T}}B = I_{d \times d}$. However, by combining the first-order algorithm with the Rcpp interface, our implementation can solve the estimating equations very efficiently. Also, parallel computing through OpenMP is used to approximate the gradient for each entry of B numerically. In Setting 2 with $p = 6$, the mean computational time for the inverse regression counting process approach is only 1.62 seconds, while the time for the semiparametric version is 8.01 seconds. The Supplementary Material summarizes the computational costs. All simulations were done on an Intel Xeon E5-2680v4 processor with five parallel threads.

We further investigate the variance of the proposed methods. Due to the complicated variance formula, we instead use bootstrap to estimate the standard deviations of the proposed estimators. Using an upper block-diagonal version of the parameter of interest, we estimate the standard deviations based on 100 bootstrap samples and also report the 95% confidence intervals. The results show that in Setting 1, the bootstrap estimators of all the proposed methods approximate the standard deviations well. In the other settings, the approximations for the computationally efficient and counting process inverse regression approaches still achieve good performances, while the approximation for the semiparametric inverse regression slightly overestimates the standard deviation, leading to slight overcoverage, around 98%.

5.2. *Skin cutaneous melanoma data analysis*

We apply the proposed method to The Cancer Genome Atlas (https://cancergenome.nih.gov/) skin cutaneous melanoma dataset, which provides comprehensive profiling data on more than thirty cancer types. We acquired 20 531 items of mRNA expression and clinical data on a total of 469 patients, with 156 observed failures. To produce biologically meaningful results, we preselect the top 20 genes highly associated with cutaneous melanoma based on meta-analyses of over 145 papers (Chatzinasiou et al., 2011). A list of these genes can be found at http://bioserver-3.bioacademy.gr/Bioserver/MelGene/. We further include age at diagnosis as a clinical control variable. All covariates are pre-processed to have unit variance and zero mean.

Selecting the number of structural dimensions can be challenging, especially with right-censored survival data (Xia et al., 2010), and we adopt the validated information criterion (Ma & Zhang, 2015), which is particularly suited to our generalized method of moments framework. The
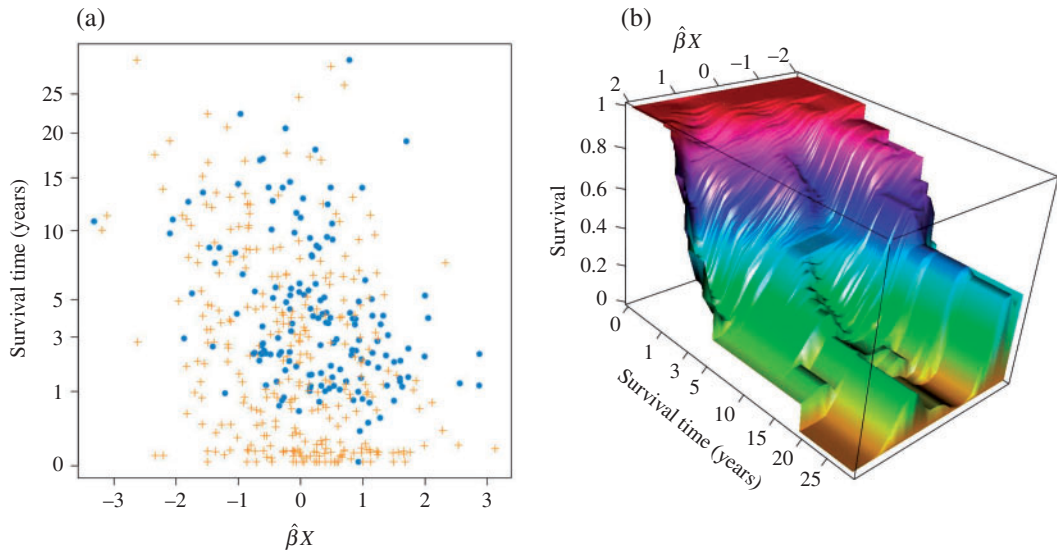
Fig. 1. Fitted direction and survival function of the semiparametric inverse regression: (a) the projected direction versus the observed failure times (blue dots) and censoring times (orange plus signs); (b) a nonparametric estimate of the survival function based on the projected direction.

validated information criterion is constructed by penalizing the quadratic form of the objective function. When we apply this method to all of our proposed estimating equation approaches, $d = 1$ always yields the best fit. Hence we present the results for all methods with $d = 1$. As a demonstration of the fitted model, we project the design matrix onto the estimated direction of the semiparametric inverse regression approach and plot the survival outcome against the projection. A nonparametric estimator of the conditional survival function based on this projection is also produced. We can see a clear trend that subjects with larger values of the projection have a lower survival rate. For comparisons, we look at the competing methods with one structural dimension, and the results can be found in the Supplementary Material. It seems that double slicing obtains a similar direction with monotone effects on the risk of failure, while the other directions obtained by other methods are nonmonotone.

We observe both similarities and differences between the different methods for the identified genes. Figure 1 suggests that a higher survival rate is observed for a smaller derived direction. This further indicates that younger patients tend to have a higher survival rate. This finding is consistent with double slicing, which identifies the variable Age with loading 0.47. However, other methods do not assign large loading to Age. We summarize in Table 2 the results with a positive loading of Age, and the directions are multiplied by -1 otherwise. Another important predictor for which all methods agree in signs is MTAP. This gene has been previously reported to have a negative correlation with the progression of melanocytic tumours (Behrmann et al., 2003), which justifies the sizeable negative value estimated by the proposed methods. However, the magnitudes in alternative methods are small. Other common genes identified by the proposed methods are MYH7B and CASP8. Li et al. (2008) genotyped putatively functional polymorphisms of CASP8 and found a significant association with lower risk of cutaneous melanoma. The result therein supports the large negative loading of the CASP8 gene in our fitted model. For differences across methods, TYR is identified by the alternative methods except for the double slicing method, with large loadings up to -0.78. The enzyme encoded by this gene controls the production of melanin and has been shown to be strongly associated with melanoma (Gudbjartsson et al., 2008). Although the estimated directions are dominated by this gene, we did not observe a monotone

Table 2. *Skin cutaneous melanoma data analysis results: the loading vectors $(\times 10^2)$ of the first structural dimension*

	Naive	DS	IPCW-SIR	hMave	Forward	CP-SIR	IR-CP	IR-Semi
Age	16	47	10	0	60	59	53	54
TYRP1	−16	−5	−9	24	18	11	39	30
OCA2	18	17	14	−6	21	19	22	5
TYR	−60	−9	−65	−78	−19	−27	−19	9
SLC45A2	11	24	23	14	30	28	16	17
CDKN2A	6	−28	−2	−12	−9	−7	−2	−11
MX2	2	−2	−2	−12	−19	−13	−30	−27
MTAP	−15	−8	−10	−14	−31	−36	−35	−30
MITF	56	−9	43	5	−13	−12	2	−27
VDR	5	−18	−9	10	−10	−6	−4	2
CCND1	−20	35	−21	−5	16	17	18	16
MYH7B	10	−27	5	−4	−29	−32	−30	−48
ATM	−16	−22	2	28	−4	7	0	6
PLA2G6	−22	−16	−21	7	4	−5	−11	−3
CASP8	15	−39	21	−13	−26	−24	−18	−14
AFG3L1	12	26	18	−15	17	10	−6	−9
CDK10	3	8	2	25	−7	−1	9	8
PARP1	−9	3	−22	17	14	18	8	18
CLPTM1L	−8	−5	17	2	−6	−6	−2	−6
ERCC5	−14	25	−17	−7	12	13	22	3
FTO	−3	−3	−8	14	15	17	7	5

DS, method of Li et al. (1999); IPCW-SIR, method of Lu & Li (2011); hMave, method of Xia et al. (2010); Forward, forward regression; CP-SIR, the computationally efficient approach; IR-CP, the counting process inverse regression approach; IR-Semi, the semiparametric inverse regression approach.

effect of the directions. We provide a comparison of the different methods with respect to the Frobenius norm distance in the Supplementary Material.

6. DISCUSSION

In this paper, we have proposed a counting process-based dimension reduction framework for censored outcomes. A family of generalized methods of moments approaches is constructed for estimating the dimension reduction subspace. The main advantage of the proposed methods is that they are adaptive to the structural dimension and free of the modelling of censoring mechanisms. This circumvents the difficulties of many existing methods and improves the efficiency when the ambient dimension p is too large for kernel methods. Our simulation study suggests that the proposed method outperforms existing methods in a variety of settings. To efficiently solve the proposed estimating equations, we further introduce an orthogonality constrained optimization algorithm that solves the parameters within a Stiefel manifold. With implementations in the R package orthoDr through C++, the counting process version of the estimators can be solved within a few seconds. However, the martingale version requires significantly more calculation due to the local estimation of the hazard function, hence requiring a few minutes to solve. We believe that there is still room to improve the computational performance. Besides, our computationally

efficient approach requires only a singular value decomposition and has satisfactory performance. However, it does not enjoy the same theoretical guarantee without conditions on the covariates. Further relaxation of these conditions is of interest.

Our framework may be extended. First, by imposing penalization on the estimating equations, it is possible to extend the framework to high-dimensional data. Sparse estimation of the B parameter may help interpretations and improves the prediction accuracy of subsequent nonparametric models. Another direction is to search for alternative α functions. In our inverse regression framework, we used the φ function, which is motivated by the inverse regression curve of a binary outcome. It would be interesting to investigate the possibility of a sliced average variance estimation-type (Cook & Weisberg, 1991) function that may deal with more complicated model structure. We can also consider using $\alpha(u, X) = B^{\mathrm{T}} X \varphi^{\mathrm{T}}(u)$. It would also be interesting to derive an α function that achieves semiparametric efficiency. Lastly, it would be interesting to extend this framework to a time-varying coefficient setting, where the dimension reduction space $\mathcal{S}(t)$ changes over time t.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes a derivation of the ortho-complement of the nuisance tangent space, proof of the double robustness property for the semiparametric inverse regression approach, proofs of Lemma 1 and Theorem 1, and additional simulation and data analysis results.

ACKNOWLEDGEMENT

REFERENCES

BEHRMANN, I., WALLNER, S., KOMYOD, W., HEINRICH, P. C., SCHUIERER, M., BUETTNER, R. & BOSSERHOFF, A.-K. (2003). Characterization of methylthioadenosin phosphorylase (MTAP) expression in malignant melanoma. *Am. J. Pathol.* **163**, 683–90.

BICKEL, P. J., KLAASSEN, C. A., RITOV, Y. & WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, Maryland: Johns Hopkins University Press.

CHATZINASIOU, F., LILL, C. M., KYPREOU, K., STEFANAKI, I., NICOLAOU, V., SPYROU, G., EVANGELOU, E., ROEHR, J. T., KODELA, E., KATSAMBAS, A. et al. (2011). Comprehensive field synopsis and systematic meta-analyses of genetic association studies in cutaneous melanoma. *J. Nat. Cancer Inst.* **103**, 1227–35.

COOK, R. D. (2009). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. New York: John Wiley & Sons.

COOK, R. D. & WEISBERG, S. (1991). Discussion of sliced inverse regression for dimension reduction. *J. Am. Statist. Assoc.* **86**, 328–32.

COX, D. R. (1972). Regression models and life-tables. *J. R. Statist. Soc. B* **34**, 187–220.

DONG, Y. & LI, B. (2010). Dimension reduction for non-elliptically distributed predictors: Second-order methods. *Biometrika* **97**, 279–94.

EDELMAN, A., ARIAS, T. A. & SMITH, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**, 303–53.

GUDBJARTSSON, D. F., SULEM, P., STACEY, S. N., GOLDSTEIN, A. M., RAFNAR, T., SIGURGEIRSSON, B., BENEDIKTSDOTTIR, K. R., THORISDOTTIR, K., RAGNARSSON, R., SVEINSDOTTIR, S. G. et al. (2008). ASIP and TYR pigmentation variants associate with cutaneous melanoma and basal cell carcinoma. *Nature Genet.* **40**, 886–91.

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–54.

LI, B. & DONG, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *Ann. Statist.* **37**, 1272–98.

LI, B. & WANG, S. (2007). On directional regression for dimension reduction. *J. Am. Statist. Assoc.* **102**, 997–1008.

LI, C., ZHAO, H., HU, Z., LIU, Z., WANG, L.-E., GERSHENWALD, J. E., PRIETO, V. G., LEE, J. E., DUVIC, M., GRIMM, E. A. et al. (2008). Genetic variants and haplotypes of the caspase-8 and caspase-10 genes contribute to susceptibility to cutaneous melanoma. *Hum. Mutat.* **29**, 1443–51.

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Am. Statist. Assoc.* **86**, 316–27.

LI, K.-C., WANG, J.-L. & CHEN, C.-H. (1999). Dimension reduction for censored regression data. *Ann. Statist.* **27**, 1–23.

LIN, D., WEI, L. & YING, Z. (1998). Accelerated failure time models for counting processes. *Biometrika* **85**, 605–18.

LU, W. & LI, L. (2011). Sufficient dimension reduction for censored regressions. *Biometrics* **67**, 513–23.

MA, Y. & ZHANG, X. (2015). A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika* **102**, 409–20.

MA, Y. & ZHU, L. (2012). A semiparametric approach to dimension reduction. *J. Am. Statist. Assoc.* **107**, 168–79.

NOCEDAL, J. & WRIGHT, S. J. (2006). *Numerical Optimization*. New York: Springer.

R DEVELOPMENT CORE TEAM (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Boca Raton, Florida: CRC Press.

TSIATIS, A. (2007). *Semiparametric Theory and Missing Data*. New York: Springer.

WEISBERG, S. (2002). Dimension reduction regression in R. *J. Statist. Software* **7**, 1–22.

WEN, Z. & YIN, W. (2013). A feasible method for optimization with orthogonality constraints. *Math. Program.* **142**, 397–434.

WU, T., SUN, W., YUAN, S., CHEN, C.-H. & LI, K.-C. (2008). A method for analyzing censored survival phenotype with gene expression data. *BMC Bioinformatics* **9**, 417.

XIA, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.* **35**, 2654–90.

XIA, Y., ZHANG, D. & XU, J. (2010). Dimension reduction and semiparametric estimation of survival models. *J. Am. Statist. Assoc.* **105**, 278–90.

ZENG, D. & LIN, D. (2007). Maximum likelihood estimation in semiparametric regression models with censored data (with Discussion). *J. R. Statist. Soc.* B **69**, 507–64.

ZHAO, R., ZHANG, J. & ZHU, R. (2017). *orthoDr: An Orthogonality Constrained Optimization Approach for Semi-Parametric Dimension Reduction Problems*. R package version 0.3.0.

ZHU, L., MIAO, B. & PENG, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Am. Statist. Assoc.* **101**, 630–43.