Published in final edited form as:

Biometrics. 2011 June ; 67(2): 513–523. doi:10.1111/j.1541-0420.2010.01490.x.

Sufficient Dimension Reduction for Censored Regressions

Wenbin Lu* and Lexin Li**

Department of Statistics, North Carolina State University, Raleigh, NC

SUMMARY

Methodology of sufficient dimension reduction (SDR) has offered an effective means to facilitate regression analysis of high dimensional data. When the response is censored, however, most existing SDR estimators can not be applied, or require some restrictive conditions. In this article we propose a new class of inverse censoring probability weighted SDR estimators for censored regressions. Moreover, regularization is introduced to achieve simultaneous variable selection and dimension reduction. Asymptotic properties and empirical performance of the proposed methods are examined.

Keywords

Censored data; Central subspace; Inverse censoring probability weighted estimation; Sliced inverse regression; Sufficient dimension reduction

1. Introduction

Regression analysis of high-dimensional data has recently prevailed in many scientific fields, while the high dimensionality often poses challenges to classical regression techniques. In many applications such as biomedical and economic studies, the analysis is further complicated by possible censoring in response variables. For instance, in survival data analysis, patients' survival times to death or disease recurrence are often right censored due to the termination of follow-up study or patients' drop-out; in employment studies, the times spent in the unemployed state are possibly censored by the cutoff nature of the sampling. Common methods for handling censored responses include Cox's proportional hazards model (Cox, 1972), proportional odds model (Bennett, 1983), accelerated failure time model (Cox and Oakes, 1984), among many others. Most of those methods require specification of some models. However, knowledge for selecting an appropriate model is often inadequate prior to the analysis, and model specification and diagnosis can become more elusive when the predictor dimension is high. Sufficient dimension reduction requires no model specification, it retains full regression information, and it provides a usually small set of composite variables upon which subsequent model formulation and prediction can be based. For these reasons SDR offers an appealing avenue to analysis of censored data with high-dimensional predictors.

Let Y denote a univariate response variable and $\mathbf{X}$ a $p \times 1$ predictor vector. SDR seeks a minimum subspace S in $\mathbb{R}^p$ such that $Y \perp\!\!\!\perp \mathbf{X} | P_S \mathbf{X}$, where $\perp\!\!\!\perp$ indicates independence and P_S denotes an orthogonal projection onto S . Under minor conditions, such a subspace, denoted

*lu@stat.ncsu.edu. **li@stat.ncsu.edu.

Supplementary Materials

The Web Appendix referenced in Sections 2.2 and 2.3 is available under the Paper Information link at the Biometrics website <http://www.tibs.org/biometrics>.

by $S_{Y|X}$, uniquely exists and is called the central subspace of regression of Y on X . $S_{Y|X}$ is the primary parameter of interest in our dimension reduction inquiry, because it carries all regression information of $Y|X$. There have been many SDR methods proposed to estimate $S_{Y|X}$ that impose no parametric assumptions on the conditional distribution $Y|X$. Among them, sliced inverse regression (SIR) is perhaps the most widely used. It amounts to a spectral decomposition of $\text{Cov}\{E(X|Y)\}$ with respect to $\Sigma = \text{Cov}(X)$. The first d eigenvectors corresponding to the largest d eigenvalues form an estimate of $S_{Y|X}$ under appropriate conditions. Operationally, SIR partitions Y into a fixed number of non-overlapping intervals, each of which is called a slice, and then takes the average of X within each slice to form an estimate of $E(X|Y)$.

Our discussion hereinafter will assume the context of survival data analysis, but the methodology applies to other censored regressions as well. Let Y denote the survival time, C be the corresponding censoring time, $\tilde{Y} = \min(Y, C)$ be the observed survival time, and $\delta = I(Y \leq C)$ be the censoring indicator. For identifiability, we assume $Y \perp\!\!\!\perp C|X$. When the response Y is censored, SIR and other SDR estimators can not be applied directly to estimate the central subspace $S_{Y|X}$. It is noted though, the bivariate response $(\tilde{Y}, \delta)$ is observable and is a function of (Y, C) , and as such $S_{(\tilde{Y}, \delta)|X} \subseteq S_{(Y, C)|X}$. Consequently, one may gain useful information about $S_{Y|X}$ through $S_{(Y, C)|X}$, which is in turn estimated through $S_{(\tilde{Y}, \delta)|X}$ from regression of $(\tilde{Y}, \delta)$ on X . Operationally this amounts to the slicing of the bivariate response $(\tilde{Y}, \delta)$, i.e., partitioning $\tilde{Y}$ within each subsample where $\delta = 0$ and $\delta = 1$, respectively. The remaining steps remain the same as a usual SIR. This procedure is called double slicing in Li et al. (1999). Mainly due to its simplicity, double slicing SIR has received wide applications; see Li and Li (2004) for analysis of a microarray gene expression data with censored phenotypes, and Li et al. (2007) for analysis of Tobit models in economics data. Although double slicing SIR is simple to use, the parameter of interest is $S_{Y|X}$ instead of $S_{(Y, C)|X}$. Actually, $S_{(Y, C)|X}$ is always a “larger” subspace than $S_{Y|X}$ in the sense that $S_{Y|X} \oplus S_{C|X} \subseteq S_{(Y, C)|X}$, where $\oplus$ denotes a direct sum between two subspaces. Li et al. (1999) examined a special case where $S_{Y|X} = S_{(Y, C)|X}$ by requiring that $(Y, X) \perp\!\!\!\perp C$. However this seems a too strong restriction. Cook (2002) relaxed this condition by requiring that $(Y, C) \perp\!\!\!\perp X|\Gamma^T X$, where Γ denotes a basis of $S_{Y|X}$. But given data this condition is very difficult to verify. Alternatively, Li et al. (1999) acknowledged that censoring can introduce substantial bias to the usual SIR estimator for the central subspace $S_{Y|X}$ of interest, and proposed a modification by utilizing information on the censoring mechanism to adjust for the censoring bias. The bias correction term, however, relies on a nonparametric estimate of the conditional survival function of Y given X , which itself is computationally complicated. Recently, Xia et al. (2010) developed a dimension reduction method for censored survival data based on kernel estimation of the conditional hazard function, which may suffer the curse of dimensionality when the number of predictors is large. In short, there seems a lack of an effective dimension reduction estimator of $S_{Y|X}$ given censored Y .

The main difficulty to apply the usual SDR estimators to censored response is concerned with the fact that most SDR methods rely on the inverse regression of X given Y . Take SIR as an example, it involves the first moment of inverse mean $E(X|Y)$, and when Y is censored, it is difficult to slice Y and thus difficult to obtain an unbiased estimate of $E(X|Y)$. To overcome this difficulty, we first investigate a re-formulation of SIR and recognize that one can estimate $S_{Y|X}$ through least squares regression of a set of transformation functions of Y given X . We then employ the inverse censoring probability weighted (ICPW) estimation method to handle censored responses. We further develop a variable selection strategy through regularized sparse estimation. Our proposed method contributes in at least three ways. First, it provides a useful addition to sufficient dimension reduction methodology and extends SDR to a large number of applications where the response is subject to censoring. An effective estimator is developed to target the primary parameter of interest $S_{Y|X}$ directly.

It is no longer necessary to gain knowledge of $S_{Y|X}$ through a larger subspace $S_{(Y,C)|X}$, and as such there is no need for conditions to force the equivalence between $S_{Y|X}$ and $S_{(Y,C)|X}$. Secondly, thanks to the flexible structure of the central subspace, the proposed method permits a variety of regression forms, for instance, the single-index model and the heteroscedastic model. Thirdly, the proposed solution achieves variable selection with censored responses, but requires no model pre-specification, and thus distinguishes itself from most of existing variable selection methods in censored regressions.

2. Sufficient Dimension Reduction with Censored Responses

2.1 Estimation of the central subspace through least squares

Without loss of generality, we assume $E(\mathbf{X}) = 0$ throughout this article. Define $\phi(y) \equiv \Sigma^{-1}E(\mathbf{X}|Y = y)$ for any y in the support of Y . Li (1991) effectively showed that

$$\phi(y) \in S_{Y|X}, \quad (1)$$

under the linearity condition such that $E(\mathbf{X}|\Gamma^T \mathbf{X} = \mathbf{x})$ is a linear function of $\mathbf{x}$, where Γ denotes a basis of $S_{Y|X}$. Equation (1) connects the inverse mean vector with the central subspace, and is the foundation of SIR and many elaborated variants of SIR. The linearity condition is imposed by SIR and most SDR estimation methods. It is interesting to note that it involves the distribution of $\mathbf{X}$ only, rather than $Y|\mathbf{X}$. The condition is satisfied when the distribution of $\mathbf{X}$ is elliptically symmetric (Eaton, 1986), e.g., multivariate normal. In practice, some caution may need to be taken if some predictors are discrete, since the elliptical symmetry is then not satisfied. But in general, this condition is often viewed as only a mild restriction, since it may be induced by coordinate-wise power transformation, and it holds to a reasonable approximation as p increases (Hall and Li, 1993). See also Cook and Ni (2006) for a further discussion on the linearity condition.

Proceeding from (1), we next discuss a re-formulation of SIR that is equivalent to its original form. SIR starts with constructing a $p \times h$ matrix $\mathbb{B} = (\beta_1, \dots, \beta_h)$, where $\beta_k = \Sigma^{-1}E\{\mathbf{X}J_k(Y)\}$, $J_k(Y) = 1$ if Y is in slice k and 0 otherwise, $k = 1, \dots, h$. Here the range of Y is partitioned into h non-overlapping intervals. Given (1), $\text{Span}(\mathbb{B} \subseteq S_{Y|X})$. SIR often takes one step further by assuming the coverage condition that $\text{Span}(\mathbb{B} = S_{Y|X})$ whenever $\text{Span}(\mathbb{B} \subseteq S_{Y|X})$, which often holds in practice; see also Cook and Ni (2006) for a discussion on the coverage condition. Then the eigenvectors of the matrix $\mathbb{B}^T$ form an estimate of a basis of $S_{Y|X}$. There are a number of variants of SIR that can all be formulated in a similar way, except that a different β_k is employed. For instance, Yin and Cook (2002) suggested using polynomial transformation $\beta_k = \Sigma^{-1}E\{\mathbf{X}Y^k\}$ up to power h ; Cook and Ni considered $\beta_k = \Sigma^{-1}E\{\mathbf{X}J_k(Y)Y\}$. Collectively, let $b_k(Y)$ denote a function of Y , SIR and its variants are all based upon the components

$$\beta_k = \Sigma^{-1}E\{\mathbf{X}b_k(Y)\}, k=1, \dots, h. \quad (2)$$

We make some key observations that will lead to our new estimator of the central subspace with censored responses. First, β_k is actually the least squares estimator of the slope vector of regressing $b_k(Y)$ on $\mathbf{X}$. Secondly, β_k involves no inverse regression, and thus it makes possible to adopt existing survival regression techniques, e.g., inverse censoring probability weighted estimation, to estimate β_k and subsequently $S_{Y|X}$. We also note that, although least squares estimator is used here, it does not imply that the linear regression model is true or even provides an adequate fit to the data.

Next we develop a new class of SDR estimators for censored regressions. Although the proposed estimation strategy applies to all transformation functions $b_k(Y)$, our empirical experiences have suggested that the polynomial function of Yin and Cook (2002) enjoys the best numerical performance, and thus it will be our choice of transformation in the subsequent development.

2.2 ICPW estimation of the central subspace

Define $G_{\mathbf{X}}(t) \equiv P(C \geq t|\mathbf{X})$ as the survival function of the censoring time given the predictors. We have

$$E \left\{ \frac{\delta}{G_{\mathbf{X}}(\tilde{Y})} b_k(\tilde{Y}) | \mathbf{X} \right\} = E \left[E \left\{ \frac{\delta}{G_{\mathbf{X}}(\tilde{Y})} b_k(\tilde{Y}) | Y, \mathbf{X} \right\} | \mathbf{X} \right] = E \left[E \left\{ \frac{I(Y \leq C)}{G_{\mathbf{X}}(Y)} b_k(Y) | Y, \mathbf{X} \right\} | \mathbf{X} \right] = E \{ b_k(Y) | \mathbf{X} \}.$$

Then $\beta_k = \Sigma^{-1} E\{ \mathbf{X} b_k(\tilde{Y}) \delta / G_{\mathbf{X}}(\tilde{Y}) \}$, $k = 1, \dots, h$. In light of this observation, one can obtain an unbiased estimate of β_k in (2) by introducing the inverse of $G_{\mathbf{X}}(\tilde{Y})$ as weight to the uncensored observations. This is the basic idea of ICPW estimation method that has been widely used in the survival analysis literature (Koul et al., 1981; Fan and Gijbels, 1994; Cheng et al., 1995, 1997; Fine et al., 1998; Kong et al., 2004; among others). It motivates us to consider the following ICPW estimator of the central subspace for censored responses.

Let $\{(\tilde{Y}_i, \delta_i, \mathbf{X}_i), i = 1, \dots, n\}$ denote the n i.i.d. random observations of the triplet $(\tilde{Y}, \delta, \mathbf{X})$, and let $\hat{G}_{\mathbf{X}}(t)$ denotes an estimate of $G_{\mathbf{X}}(t)$. We propose to estimate β_k in (2) by solving the following weighted least squares estimating equation,

$$\mathbf{U}_k(\beta) \equiv \sum_{i=1}^n \frac{\delta_i}{\hat{G}_{\mathbf{X}_i}(\tilde{Y}_i)} \mathbf{X}_i \{ b_k(\tilde{Y}_i) - \beta^\top \mathbf{X}_i \} = 0, k=1, \dots, h. \quad (3)$$

There is a closed form solution for (4), i.e.

$$\hat{\beta}_k = \left\{ \sum_{i=1}^n \frac{\delta_i}{\hat{G}_{\mathbf{X}_i}(\tilde{Y}_i)} \mathbf{X}_i \mathbf{X}_i^\top \right\}^{-1} \sum_{i=1}^n \frac{\delta_i}{\hat{G}_{\mathbf{X}_i}(\tilde{Y}_i)} \mathbf{X}_i b_k(\tilde{Y}_i), k=1, \dots, h.$$

We then form the $p \times h$ matrix $\hat{\mathbf{B}} = (\hat{\beta}_1, \dots, \hat{\beta}_h)$, and perform spectral decomposition of $\hat{\mathbf{B}} \hat{\mathbf{B}}^\top$. Let $\{\hat{\gamma}_1, \dots, \hat{\gamma}_p\}$ denote the resulting eigenvectors corresponding to its eigenvalues in descending order. We call $\text{Span}(\hat{\gamma}_1, \dots, \hat{\gamma}_d)$ an *ICPW estimator* of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$. Here we first assume the structural dimension $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ is known, and will develop a consistent estimator of d later.

Next we discuss a number of ways to estimate the survival function $G_{\mathbf{X}}(\cdot)$ of the censoring time under different scenarios, and discuss the strength and limitations of each method. First, if the censoring times are i.i.d. and independent of the covariates, a consistent estimator of $G_{\mathbf{X}}(\cdot)$ can be obtained by the Kaplan-Meier method. It requires no model assumption but the condition that $C \perp \mathbf{X}$ can be restrictive. Secondly, if C does depend on $\mathbf{X}$, one can posit a semiparametric model, for instance, a proportional hazards model, and then estimate $G_{\mathbf{X}}(\cdot)$ based on the fitted model. This approach is very simple to use, and as long as the model is correctly specified, the resulting estimator of survival function is consistent. On the other hand, it loses some nonparametric flavor of the SDR estimation since it requires a model specification on the censoring time. In our implementation, we adopt the

semiparametric approach by positing a proportional hazards model due to its simplicity as well as the high dimensional nature of our targeting problems. To study its performance under possible model misspecification, we conduct an intensive sensitivity analysis in Section 3.2, as is often done in the ICPW literature. Our experiences have suggested that it works reasonably well for a moderate to large sample size even when the model for the censoring distribution is misspecified. We also comment that, operationally, one may estimate $G_{\mathbf{X}}(\cdot)$ nonparametrically using, e.g., the kernel-weighted local Kaplan-Meier estimator (Beran, 1981), although it does not satisfy the conditions (C3) and (C4) given below. This nonparametric solution does not require any model and is thus robust to model misspecification. However, it is subject to the curse of dimensionality and works best only when the number of predictors is relatively small. We will briefly examine the numerical performance of this nonparametric estimation strategy in Section 3.

To establish the asymptotic properties of our proposed ICPW estimator, we assume the following regularity conditions and the properties of the estimator $\hat{G}_{\mathbf{X}}(\cdot)$:

- (C1) The vector of covariates $\mathbf{X}$ has a convex support;
- (C2) There exists a constant $\tau > 0$ such that $P(C = \tau) > 0$ and $P(C > \tau) = 0$;
- (C3) $\sup_{t \in [0, \tau]} |\hat{G}_{\mathbf{X}}(t) - G_{\mathbf{X}}(t)| = o_p(1)$ for all possible values of $\mathbf{X}$, where τ is a pre-specified positive constant;
- (C4) For $0 \leq t \leq \tau$,

$$\sqrt{n} \left\{ \hat{G}_{\mathbf{X}}(t) - G_{\mathbf{X}}(t) \right\} = \frac{1}{\sqrt{n}} \sum_{j=1}^n \left\{ \psi_1(t, \mathbf{X}) \int_0^t \psi_2(s) dM_j(s) + \psi_3(t, \mathbf{X}) \psi_{4,j} \right\} + o_p(1),$$

where $M_j(t)$'s are independent mean-zero martingale processes, $\psi_{4,j}$'s are i.i.d. random variables with mean zero and finite variance, and the functions ψ_1 , ψ_2 and ψ_3 are bounded and satisfy the Lipschitz conditions.

The regularity condition (C2) is the same as the condition (C1) in Peng and Fine (2009), which is used to simplify theoretical arguments and is satisfied in many clinical studies with administrative censoring. In practice, τ can be chosen as the maximum of follow-up. To improve the ICPW estimator, especially for the heavy censoring case, Fine et al. (1998) proposed an effective modification by introducing an artificial censoring time L to truncate data at the right tail, i.e. $C^* = \min(C, L)$. Then condition (C2) is automatically satisfied with the truncated censoring time C^* . The same strategy can be adopted in our context as well. Conditions (C3) and (C4) are assumed to ensure the uniform consistency and the asymptotic representation of $\hat{G}_{\mathbf{X}}(\cdot)$, which in turn leads to the consistency and asymptotic normality of the ICPW estimator of $S_{Y|\mathbf{X}}$. If a proportional hazards model is correctly specified for C , the uniform consistency and asymptotic representation of the estimator can be easily derived (Andersen and Gill, 1982). For other correctly specified semiparametric models, similar results can be obtained. We next state the main asymptotic results.

Theorem 1: Under conditions (C1)–(C4), we have

1. $\sqrt{n} \left\{ \text{vec}(\hat{\mathbb{B}}) - \text{vec}(\mathbb{B}) \right\}$ converges in distribution to a normal random vector with mean zero and a covariance matrix $\mathbf{\Lambda} \in \mathbb{R}^{\text{ph} \times \text{ph}}$, as $n \rightarrow \infty$.
2. Furthermore, if the linearity and coverage conditions hold, then $\text{Span}(\hat{\gamma}_1, \dots, \hat{\gamma}_d)$ is a $\sqrt{n}$ -consistent estimator of the central subspace $S_{Y|\mathbf{X}}$.

Here $\text{vec}(\cdot)$ denotes the matrix operator that stacks all columns of a matrix to a vector. The proof of Theorem 1 is given in the Web Appendix. Basically, Theorem 1 first establishes the $\sqrt{n}$ -consistency of the ICPW estimator $\hat{\mathbb{B}}$ of $\mathbb{B}$. With the additional linearity and coverage conditions, $\text{Span}(\hat{\mathbb{B}})$ is further connected with the central subspace of interest. As such, $\text{Span}(\hat{\gamma}_1, \dots, \hat{\gamma}_d)$ provides a $\sqrt{n}$ -consistent estimator of $S_{Y|X}$.

To estimate the structural dimension $d = \dim(S_{Y|X})$, we employ the criterion proposed by Zhu et al. (2006), which estimates d via the number of nonzero eigenvalues of the matrix $\hat{\mathbb{B}}\hat{\mathbb{B}}^\top$. More specifically, let $\hat{v}_1, \dots, \hat{v}_p$ denote the eigenvalues of $\hat{\mathbb{B}}\hat{\mathbb{B}}^\top + \mathbf{I}_p$, and κ denotes the number of $\hat{v}_i$'s that are greater than one. Then an estimate of d is taken as the maximizer of the criterion,

$$\left\{ \frac{n}{2} \sum_{k=1+\min(\kappa, m)}^p (\log(\hat{v}_k) + 1 - \hat{v}_k) - \frac{C_n m(2p - m + 1)}{2} \right\}, \quad (4)$$

for $m \in \{0, 1, \dots, p-1\}$. In (4), C_n is a penalty constant, and Zhu et al. (2006) recommended $C_n = O_p(n^a)$. We choose $a = 0.1$ in our implementation, and our experience suggests that it estimates d very well. Based on Theorem 2 of Zhu et al. (2006) and the $\sqrt{n}$ -consistency of $\hat{\mathbb{B}}$ established in Theorem 1, we have the consistency of $\hat{d}$.

Corollary 1: Assume that $\lim_{n \rightarrow \infty} C_n/n = 0$ and $\lim_{n \rightarrow \infty} C_n = \infty$, then $\hat{d} \rightarrow d$ almost surely as $n \rightarrow \infty$.

2.3 Sparse estimation of the central subspace

Variable selection is another important aspect of dimension reduction in survival data analysis, since it usually leads to a better health risk assessment and an easier model interpretation. In this section, we develop a regularized sparse estimation of the central subspace for simultaneous variable selection and reduction projection basis estimation with censored responses. The idea is similar to Bondell and Li (2009) but with a different aim. Here we focus on variable selection with censored responses.

Given the ICPW estimator, $\hat{\beta}_k, k = 1, \dots, h$, we propose to minimize the following objective function over a $p \times 1$ shrinkage vector $\alpha = (\alpha_1, \dots, \alpha_p)^\top$,

$$\frac{1}{nh} \sum_{k=1}^h \sum_{i=1}^n \frac{\delta_i}{\hat{G}_{X_i}(\tilde{Y}_i)} \left[b_k(\tilde{Y}_i) - \{\text{diag}(\alpha) \hat{\beta}_k\}^\top \mathbf{X}_i \right]^2 \quad \text{subject to} \quad \sum_{j=1}^p |\alpha_j| \leq \tau, \quad (5)$$

where $\tau \geq 0$ is a regularization parameter. Letting $\hat{\alpha}$ denote the resulting minimizer of (5), we then construct the $p \times h$ matrix $\hat{\mathbb{B}} = \text{diag}(\hat{\alpha}) \hat{\mathbb{B}}$, and call $\text{Span}(\hat{\mathbb{B}})$ the *sparse ICPW estimator* of the central subspace $S_{Y|X}$.

We first note that the shrinkage estimation given in (5) is closely related to the nonnegative garrote estimation (Breiman, 1995) and the adaptive-Lasso method (Zou, 2006) in the multivariate response setup. When $\tau \geq p$, $\hat{\alpha}_j = 1$ for all $j = 1, \dots, p$, and one gets back a usual ICPW solution. As τ gradually decreases, however, some $\hat{\alpha}_j$'s are shrunk to exact zero, which indicates exclusion of the corresponding predictors X_j 's. As such variable selection is achieved. Secondly, we note that the structural dimension d of the central subspace is not required if variable selection is the only purpose for (5). This dimension information is needed only when one aims to obtain a simultaneous basis estimation of $S_{Y|X}$.

We next turn to numerical optimization of (5). Rewrite (5) as,

$$\frac{1}{nh} \|\mathbf{U} - \mathbf{V}\boldsymbol{\alpha}\|^2, \text{ subject to } \sum_{j=1}^p |\alpha_j| \leq \tau,$$

where

$$\mathbf{U} = \begin{pmatrix} w_1^{1/2} b_1(\tilde{Y}_1) \\ \vdots \\ w_n^{1/2} b_1(\tilde{Y}_n) \\ \vdots \\ w_1^{1/2} b_h(\tilde{Y}_1) \\ \vdots \\ w_n^{1/2} b_h(\tilde{Y}_n) \end{pmatrix}_{nh \times 1}, \text{ and } \mathbf{V} = \begin{pmatrix} w_1^{1/2} \widehat{\beta}_{11} X_{11} & \cdots & w_1^{1/2} \widehat{\beta}_{p1} X_{1p} \\ \vdots & \vdots & \vdots \\ w_n^{1/2} \widehat{\beta}_{11} X_{n1} & \cdots & w_n^{1/2} \widehat{\beta}_{p1} X_{np} \\ \vdots & \vdots & \vdots \\ w_1^{1/2} \widehat{\beta}_{1h} X_{11} & \cdots & w_1^{1/2} \widehat{\beta}_{ph} X_{1p} \\ \vdots & \vdots & \vdots \\ w_n^{1/2} \widehat{\beta}_{1h} X_{n1} & \cdots & w_n^{1/2} \widehat{\beta}_{ph} X_{np} \end{pmatrix}_{nh \times p}$$

$w_i = \delta_i / \widehat{G}_{\mathbf{X}_i}(\tilde{Y}_i)$, $i = 1, \dots, n$, $\widehat{\beta}_{jk}$ is the j -th element of $\widehat{\beta}_k$, $j = 1, \dots, p$, $k = 1, \dots, h$, and X_{ij} is the j -th element of $\mathbf{X}_i$. Then estimation of $\boldsymbol{\alpha}$ becomes a standard lasso problem and can be obtained by any existing lasso algorithm (Tibshirani, 1996, Fu, 1998, Efron et al., 2004). To choose the regularization parameter τ , we adopt a BIC-type criterion (Wang and Leng, 2007) by minimizing

$$\frac{\|\mathbf{U} - \mathbf{V}\widehat{\boldsymbol{\alpha}}(\tau)\|^2}{\widehat{\sigma}^2} + \log(n)p_e,$$

where $\widehat{\boldsymbol{\alpha}}(\tau)$ denotes the minimizer of (5) given τ , $\widehat{\sigma}^2$ is the usual variance estimator obtained from the least squares fit of $\mathbf{U}$ on $\mathbf{V}$, and p_e is the effective number of parameters,

$$2 \sum_{j=1}^p \mathbf{1}(\widehat{\alpha}_j(\tau) > 0) + \sum_{j=1}^p \widehat{\alpha}_j(\tau)(h-2), \text{ as given in Yuan and Lin (2006, equation 6.5).}$$

We next continue to examine the asymptotic properties of the proposed sparse estimation. Following Bondell and Li (2009), variable selection in our context is to seek the smallest subset of the predictors $\mathbf{X}_{\mathcal{A}}$, with partition $\mathbf{X} = (\mathbf{X}_{\mathcal{A}}^T, \mathbf{X}_{\mathcal{A}^c}^T)^T$, such that $Y \perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}$. This subset $\mathcal{A}$ uniquely exists when the central subspace $S_{Y|\mathbf{X}}$ exists. As implied by Cook (2004, Proposition 1), $\mathcal{A}$ is directly connected to $\mathbf{B}$ that spans $S_{Y|\mathbf{X}}$. That is, if we partition $\mathbf{B}$ to two submatrices $\mathbf{B}_{\mathcal{A}}$ and $\mathbf{B}_{\mathcal{A}^c}$ with the same number of columns as $\mathbf{B}$ following the partition of $\mathbf{X}$, then $\mathbf{B}_{\mathcal{A}^c} = \mathbf{0}$. Consequently, we can define $\mathcal{A}$ as $\mathcal{A} = \{j : \beta_{kj} \neq 0 \text{ for some } k, 1 \leq j \leq p, 1 \leq k \leq h\}$ and β_{kj} is the (k, j) -th element of $\mathbf{B}$. Given the sparse estimate from (5), we partition $\widehat{\mathbf{B}}$ in a similar way as $\mathbf{B}$ yielding $\widehat{\mathbf{B}}_{\mathcal{A}}$ and $\widehat{\mathbf{B}}_{\mathcal{A}^c}$, and denote the sparse estimate of $\mathcal{A}$ as $\widehat{\mathcal{A}} = \{j : \widehat{\alpha}_j \neq 0\}$. We further consider the equivalent Lagrangian formulation of (5),

$$\frac{1}{nh} \sum_{k=1}^h \sum_{i=1}^n \frac{\delta_i}{\widehat{G}_{\mathbf{X}_i}(\tilde{Y}_i)} \left[b_k(\tilde{Y}_i) - \{\text{diag}(\boldsymbol{\alpha}) \widehat{\boldsymbol{\beta}}_k\}^T \mathbf{X}_i \right]^2 + \lambda \sum_{j=1}^p |\alpha_j|, \quad (6)$$

where λ is a penalty constant. Then we have the following theorem.

Theorem 2: Suppose that the same conditions of Theorem 1 hold and that $n\lambda \rightarrow \infty$ and $\sqrt{n}\lambda \rightarrow 0$, then as $n \rightarrow \infty$,

- a. Consistency in variable selection: $P(\hat{\mathcal{A}} = \mathcal{A}) \rightarrow 1$.
- b. Asymptotic normality: $\sqrt{n} \{ \text{vec}(\hat{\mathbb{B}}_{\mathcal{A}}) - \text{vec}(\mathbb{B}_{\mathcal{A}}) \} \rightarrow N(0, \mathbf{\Omega})$, for some $\mathbf{\Omega} > 0$.

The proof of Theorem 2 is given in the Web Appendix. Theorem 2 shows that our sparse estimate can select all truly relevant predictors in $\mathcal{A}$ with probability approaching one. In addition, the basis estimate of those relevant predictors remains $\sqrt{n}$ -consistent.

3. Simulations and Real Data Analysis

3.1 Dimension reduction basis estimation

We first examine the finite sample performance of the ICPW estimator of $S_{Y|\mathbf{X}}$ when the censoring time distribution is correctly specified. We also compare with the double slicing SIR (DSIR) estimator of Li et al. (1999) that is commonly used in practice.

The true survival time is generated as follows:

$$Y = \exp \left\{ -2.5 + \sin(0.1\pi\gamma_1^\top \mathbf{X}) + 0.1(\gamma_1^\top \mathbf{X} + 2)^2 + 0.25\varepsilon \right\}, \quad (7)$$

$$Y = \exp \left\{ -2.5 + \gamma_1^\top \mathbf{X} + 0.5 \gamma_1^\top \mathbf{X} \times \gamma_2^\top \mathbf{X} + 0.25\varepsilon \right\}, \quad (8)$$

where $\mathbf{X}$ is a 10-dimensional multivariate normal vector with mean zero and identity covariance matrix. Different choice of the error ε leads to some commonly used survival models: the extreme value distribution $\varepsilon = \log\{-\log(1-u)\}$ for the proportional hazards (PH) model, and the logistic distribution $\varepsilon = \log\{u/(1-u)\}$ for the proportional odds (PO) model, where u is uniform(0,1). For model (7), $S_{Y|\mathbf{X}} = \text{Span}(\gamma_1)$, with $\gamma_1 = (1, 1, 1, 0, \dots, 0)^\top$, and the structural dimension $d = 1$; for model (8), $S_{Y|\mathbf{X}} = \text{Span}(\gamma_1, \gamma_2)$, with $\gamma_1 = (1, 0, \dots, 0)^\top$, $\gamma_2 = (0, \dots, 0, 1, -1)^\top$, and $d = 2$. We generate the censoring time C as

$$C = \exp(c + \gamma_c^\top \mathbf{X} + \varepsilon_c), \quad (9)$$

where c is a constant that controls the censoring proportion, $\gamma_c = (-1, 0, 0, 1, 0, \dots, 0)^\top$ and ε_c first takes the extreme value distribution, yielding a PH model for C .

The vector correlation coefficient (Hotelling, 1936) is employed to evaluate the discrepancy between the estimated and the true central subspaces. This measure ranges between 0 and 1 with a larger value indicating a better estimation accuracy. We examine both the PH model and the PO model for (7) and (8), each with two censoring proportions, 20% and 40%. Since the results show very similar qualitative pattern, we only report the PH model of (7) and the PO model of (8) with 40% censoring. Figure 1 shows the median vector correlation out of 100 data replications. The solid line denotes the covariance SDR estimator of Yin and Cook (2002) without any censoring, which serves as a benchmark. The dashed line denotes the ICPW estimator where $G_{\mathbf{X}}(\cdot)$ is estimated based on a PH model fitting of C given all the covariates $\mathbf{X}$. The estimation accuracy of the ICPW estimator quickly approaches the benchmark when the sample size increases. We also examine two variants of the ICPW estimator, one estimating $G_{\mathbf{X}}(\cdot)$ using the usual Kaplan-Meier method ignoring the covariate

information (dot-dashed line), and the other using a local Kaplan-Meier method based on the estimates from double slicing SIR (dotted line) to partly overcome the curse of dimensionality. For the local Kaplan-Meier method, we use the normal kernel for each of the derived covariates from DSIR and choose the optimal bandwidth parameter $h_x = 4^{1/3} \sigma_x n^{-1/3}$ (Jones, 1990), where σ_x is the sample standard deviation of a derived covariate. From the plot, we see that the ICPW estimator with the censoring time distribution estimated by a PH model uniformly dominates the two variants. However, this is not surprising since in the current setting the censoring distribution is correctly specified by the fitted PH model. In the subsequent analysis, we will concentrate on using the PH model for $G_{\mathbf{X}}(\cdot)$. Finally, we examine the double slicing SIR estimator, which is denoted by the long-dashed line in the plot, and in all cases, our ICPW estimator outperforms DSIR uniformly. In addition, we also compare the two estimators when the censoring time is generated independent of $\mathbf{X}$, so DSIR is supposed to work well. Our results (not shown here) indicate that the ICPW estimator achieves almost the same accuracy as DSIR in that setup.

3.2 Sensitivity analysis

In the ICPW literature, sensitivity analyses have often been employed to evaluate the effect of misspecification of the censoring time distribution. In this section, we conduct an intensive sensitivity analysis by varying the censoring distribution specified in (9).

Specifically, the hazard function of ε_c in (9) is given by

$$\lambda_{\varepsilon_c}(t, r) = \frac{\exp(t)}{1 + r \exp(t)}, \quad (10)$$

where the constant r controls the level of deviation from a PH model. That is, when $r = 0$, it yields a PH model, but as r increases, it leads to a model deviant from PH, and when $r = 1$, it corresponds to a PO model for the censoring time distribution. In our sensitivity analysis, we consider $r = 0.25, 0.5, 0.75$ and 1.0 . In addition, we also examine a log-normal model for C by generating ε_c from a standard normal distribution (denoted by log normal), and a PH model ($r = 0$) but with an interaction term of covariates, i.e., $C = \exp(c - X_1 + X_4 + 0.5X_1X_2 + \varepsilon_c)$ (denoted by mis-spec. PH). Recall that for our ICPW estimator, the censoring time distribution is obtained by always fitting a PH model with all linear terms of covariates, and as such the model is misspecified for all above cases considered. The true survival time is still generated from models (7) and (8), respectively.

Table 1 summarizes the results in terms of the median vector correlation based on 100 data replications, with 40% censoring. As a comparison, the results for the DSIR estimator are also reported. It is seen that the performance of ICPW degrades as the censoring time distribution deviates away from the PH model; however, it maintains a competitive accuracy throughout. More importantly, it always outperforms DSIR, often by a large margin, even when the censoring distribution is misspecified. Such sensitivity analyses, though by no means comprehensive, clearly suggest that the proposed ICPW estimator provides a useful solution to sufficient dimension reduction estimation, and is a better alternative to DSIR that is widely used in the current literature. For practical application of the ICPW estimator, sensitivity analyses using various censoring time distributions are always recommended.

We have also briefly compared the ICPW estimators based on the semiparametric PH model estimation and the nonparametric local Kaplan-Meier estimate as considered in the previous section under the censoring time mis-specification, where the true censoring times are from a proportional odds model. The results, which are reported in the supplementary Appendix for brevity of the paper, show that the nonparametric estimator performs slightly better than

the semiparametric estimator for model (7), while the two estimators have very comparable performance for model (8). One possible explanation is that when the true number of sufficient predictors from both the survival time and the censoring time is relatively small, (e.g., in model (7), this number equals 2), the nonparametric solution is to have a competitive performance, but when the true dimension is relative large (e.g., in model (8), this number equals 3), the kernel-based nonparametric estimator is to suffer. Moreover, in the current simulations, we treat the true number of sufficient predictors as known. In practice, it is not uncommon that this number is overestimated, which would adversely affect the nonparametric estimator. By contrast, the semiparametric estimator is expected to be less sensitive to the overestimation of the true dimension since it always fits a PH model for the censoring times with all the covariates.

3.3 Model assessment and comparison

In Sections 3.1 and 3.2, we have focused our attention on evaluating the estimation accuracy of the dimension reduction basis, while no model fitting is conducted after reduction. From a complete modeling point of view, it is also informative to study the performance of the dimension reduction aided modeling of survival time, and to compare with some common survival modeling practices in the literature. This is to be the focus of this section.

Toward that end, we employ the mean squared error (MSE) of the estimated covariate

effects as the evaluation criterion, i.e. $MSE = (1/n) \sum_{i=1}^n \{\hat{f}(\mathbf{X}_i) - f_0(\mathbf{X}_i)\}^2$, where $f_0(\cdot)$ is the true covariates effect and $\hat{f}(\cdot)$ is the estimated covariates effect. We continue to employ models (7) and (8) and compare four solutions. The first is SDR-aided Cox modeling. That is, we first use ICPW to obtain the estimated sufficient predictors $\hat{v}_1 = \hat{\gamma}_1^\top \mathbf{X}$ for model (7), and $\hat{v}_1 = \hat{\gamma}_1^\top \mathbf{X}$, $\hat{v}_2 = \hat{\gamma}_2^\top \mathbf{X}$ for model (8). We then fit a Cox PH model with saturated covariate effects, i.e., $\lambda(t) = \lambda_0(t) \exp(\theta_1 \hat{v}_1 + \theta_2 \hat{v}_1^2)$ for (7), and

$\lambda(t) = \lambda_0(t) \exp(\theta_1 \hat{v}_1 + \theta_2 \hat{v}_2 + \theta_3 \hat{v}_1^2 + \theta_4 \hat{v}_2^2 + \theta_5 \hat{v}_1 \times \hat{v}_2)$ for (8). Since SDR effectively reduces the number of covariates from $p = 10$ to $d = 1$ or 2 , it actually affords us flexibility to consider more complex models such as a quadratic or higher-order polynomial than just a linear model. The second solution is similar to the first one, but the reduced covariates are obtained by DSIR, with the dimension of $d+1$ to incorporate the directions in both the survival time distribution and the censoring time distribution. The third solution is a standard linear Cox model with the original covariates $\mathbf{X}$. The last is a Cox model based on the cubic spline bases of $\mathbf{X}$ with three inner knots, so to accommodate nonlinear covariate effects.

The median MSEs out of 100 replications for the three methods are reported in Table 2. It is seen that the Cox model based on the ICPW estimator achieves the smallest MSEs among all solutions for all the scenarios, suggesting that the proposed dimension reduction method would be useful to help build a better model than the standard solutions.

3.4 Variable selection

We next evaluate the sparse estimation of the central subspace in terms of variable selection. We employ the simulation setup in Section 3.1, but vary the number of active predictors in the basis of $S_{Y|\mathbf{X}}$. More specifically, for model (7), $\gamma_1 = (1, \dots, 1, 0, \dots, 0)^\top$, with the number of ones equal to 3, 6, and 10; and for model (8), γ_1 takes the same form but the number of ones equals 1, 4, and 8, while γ_2 remains the same as before. As a result, the number of zeros equal 7, 4, and 0 for both models, which correspond roughly to very sparse, medium sparse, and non-sparse scenarios. In addition, we introduce correlation among the predictors, with $\text{Corr}(X_i, X_j) = 0.5^{|i-j|}$, $1 \leq i, j \leq p$.

For (7) and (8), both the PH and PO models are examined, and we only report the median vector correlation coefficients for the PO model of (7) and the PH model of (8) with 40% censoring in Table 3, while the results for the other two cases are very similar. We show for both the non-regularized ICPW estimator and its regularized version (denoted by $ICPW_s$ in the table). Evaluation criteria include the number of correctly identified zeros (cor 0) and the number of incorrectly identified zeros (inc 0), both of which measure accuracy of variable selection, and the vector correlation coefficient (vcr), which measures estimation accuracy of the sparse central subspace basis. It is seen from the table that for the regularized ICPW estimator ($ICPW_s$), the number of incorrect zeros are 0 or very small, indicating that nearly all truly active predictors have been selected; while the number of correct zeros is slightly smaller than the true, indicating that false positive selection is low. Furthermore, a generally large value of vector correlation coefficient suggests that sparse basis estimate is also accurate.

3.5 PBC data

To illustrate the proposed method, we first apply it to the primary biliary cirrhosis data collected at the Mayo Clinic. This data has been analyzed in the literature. In particular, Zhang and Lu (2007) concentrated on 276 subjects with complete observations and concluded that 8 covariates, age, albumin (alb), serum bilirubin (bil), urine copper (cop), prothrombin time (prot), liver enzyme (sgot), presence of oedema (oed) and stage, are important in modeling subjects' survival time. Following Zhang and Lu (2007), we focus on these 8 covariates in our illustration. Among them, the variable oed takes three possible values, 0, 0.5 and 1, while the variable stage takes integer value between 1 and 4. For simplicity, we treat all variables as continuous. Among the 276 subjects, 147 subjects have their responses censored.

We apply the ICPW estimation to this data, and the criterion (4) concludes that $\hat{d} = 2$. We next obtain the two SDR variates, $\hat{v}_1 = \tilde{\gamma}_1^T \mathbf{X}$ and $\hat{v}_2 = \tilde{\gamma}_2^T \mathbf{X}$ where $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are the sparse ICPW estimators. Figure 2 (a) shows the solution path of the shrinkage vector α as the regularization parameter τ increases. The vertical line indicates the value of τ selected by BIC, and all covariates but age and liver enzyme are selected. It is also seen that when τ equals the number of predictors, all elements of $\hat{\alpha}$ become 1. To complete the analysis, we next fit a Cox proportional hazards model with full quadratic effects, i.e.

$\lambda(t) = \lambda_0(t) \exp(\theta_1 \hat{v}_1 + \theta_2 \hat{v}_2 + \theta_3 \hat{v}_1^2 + \theta_4 \hat{v}_2^2 + \theta_5 \hat{v}_1 \times \hat{v}_2)$, and find the main effect term $\hat{v}_2$ and the interaction term $\hat{v}_1 \times \hat{v}_2$ are significant. We then re-fit the model with these two terms, yielding $\lambda(t) = \lambda_0(t) \exp(-1.15 \hat{v}_2 - 0.18 \hat{v}_1 \times \hat{v}_2)$, with the corresponding p-values for both coefficients less than $1e-4$. Figure 2 (b) shows the Kaplan-Meier estimate of the survival curves for the three risk groups determined by the estimated risk scores (divided by the 33% and 66% quantiles). A good separation of the three risk curves is observed, suggesting that the model fits the data reasonably well.

3.6 DLBCL gene expression data

We next analyze the large B-cell lymphoma (DLBCL) microarray gene expression data of Rosenwald et al. (2002). The data consist of the survival time of 240 DLBCL patients, along with expression levels of 7399 genes for each patient. Following the practice of Li and Luan (2005) and Lu and Li (2008), we focus on the top 50 genes ranked by the univariate Cox scores in our analysis. The data was pre-partitioned by Rosenwald et al. (2002) to a training group of 160 samples and a testing group of 80 samples. Moreover, we remove five subjects whose observed survival time equal to exactly zero from further analysis, leaving 156 training samples and 79 testing samples. Among 235 patients, 102 patients have censored responses.

The ICPW estimator concludes that $\hat{d} = 3$ based on the training samples. Three SDR variates are constructed, $\hat{v}_j = \hat{\gamma}_j^\top \mathbf{X}$, $j = 1, 2, 3$, where $\hat{\gamma}_j$'s are obtained from the sparse ICPW estimation. 10 genes out of 50 are selected. We next fit a Cox proportional hazards model with full quadratic effects based on $\hat{v}_j$'s and find that all three main effects and one interaction term are significant. The final model is $\lambda(t) = \lambda_0(t) \exp(-1.12 \hat{v}_1 - 0.427 \hat{v}_2 + 1.182 \hat{v}_3 + 0.757 \hat{v}_2 \times \hat{v}_3)$, with the corresponding p-values equal to 0.0003, 0.018, 1.9×10^{-7} , and 0.032, respectively. Figure 3 (a) shows the Kaplan-Meier estimate of the survival curves for the same three risk groups determined by the estimated risk scores for the training data. The three groups are well separated, again suggesting a good model fitting. We also briefly evaluate the predictive performance of the fitted model, and Figure 3 (b) shows the survival curves of the three risk groups for the testing data. The separation is reasonably good, suggesting a satisfactory prediction performance of the fitted survival model.

4. Discussion

In this article, we have developed a sufficient dimension reduction estimator for regressions with censored responses. It is based on a least squares formulation of sliced inverse regression, which is then coupled with inverse censoring probability weighted estimation. The main motivation is that, when the dimension of the predictors is high, the failure time can depend on the covariates in a very complicated fashion and a direct modeling can be very challenging. We believe a dimension reduction step that reduces the high dimensional covariates to a few summary predictors would greatly facilitate modeling of such type of data. Both our theoretical and numerical studies have offered some support to this belief.

Our new estimator requires modeling the censoring time distribution, and its asymptotic properties hinge on a correct estimation of $G_{\mathbf{X}}(t) = P(C \geq t | \mathbf{X})$. This task itself is by no means easy when the dimension of $\mathbf{X}$ is high. However, several reasons lead us to employ the ICPW strategy in our dimension reduction context. First, ICPW has been widely used in the survival analysis literature. It has been shown to be an effective solution for both estimation and inference. Second, in many well designed and followed clinical trials, the censoring distribution may often be relatively easier to model compared to the failure time distribution of interest. Third, we always recommend to carry out a sensitivity analysis in real practice to examine the performance of the ICPW estimator when the censoring distribution is misspecified. This is indeed a common practice in the ICPW literature. Our intensive sensitivity analysis has suggested that the proposed ICPW estimator maintains a competitive accuracy under censoring time model misspecification, and more importantly, it always outperforms, often by a large margin, the currently used double sliced inverse regression estimator. As such, we feel our ICPW estimator would provide a useful and better alternative to SDR with censored responses.

The least squares formulation used in this article also points to possible future development of SDR estimator for censored regressions. One option is the rank estimation that has been widely studied in the survival analysis literature (Tsiatis, 1990; Ying, 1993; among others). In the context of standard linear regression with right censored data, the rank estimator makes use of the well-studied counting process based martingale theory, and is known to be generally more efficient than the ICPW estimator. In our context of central subspace estimation, one may consider to estimate β_k in (2) by solving the weighted log-rank estimating equations. However, unlike the ICPW estimator, the asymptotic properties of the rank estimator of central subspace are hard to obtain since they require a linear model for the transformed survival time that is generally not true for the dimension reduction context. A careful study of the rank estimator in terms of both asymptotic properties and finite sample performance will be our future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the associate editor and two referees for their comments that substantially improved the presentation of the article. This work was partially supported by NIH/NCI grant R01 CA-140632 (WL) and NSF grant DMS-0706919 (LL).

REFERENCES

- Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Annals of Statistics*. 1982; 10:1100–1120.
- Bennett S. Analysis of survival data by the proportional odds model. *Statistics in Medicine*. 1983; 2:273–277. [PubMed: 6648142]
- Beran, R. Technical Report. Berkeley: University of California; 1981. Nonparametric regression with randomly censored survival data.
- Breiman L. Better subset regression using the nonnegative garrote. *Technometrics*. 1995; 37:373–384.
- Bondell HD, Li L. Shrinkage inverse regression estimation for model free variable selection. *Journal of the Royal Statistical Society, Series B*. 2009; 71:287–299.
- Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. *Biometrika*. 1995; 82:835–845.
- Cheng SC, Wei LJ, Ying Z. Prediction of survival probabilities with semi-parametric transformation models. *Journal of the American Statistical Association*. 1997; 92:227–235.
- Cook RD. Dimension reduction and graphical exploration in regression including survival analysis. *Statistics in Medicine*. 2002; 22:1399–1413. [PubMed: 12704605]
- Cook RD. Testing predictor contributions in sufficient dimension reduction. *Annals of Statistics*. 2004; 32:1061–1092.
- Cook RD, Ni L. Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika*. 2006; 93:65–74.
- Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*. 1972; 34:187–220.
- Cox, DR.; Oakes, D. *Analysis of Survival Data*. New York: Chapman and Hall; 1984.
- Eaton M. A characterization of spherical distributions. *Journal of Multivariate Analysis*. 1986; 20:272–276.
- Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. *Annals of Statistics*. 2004; 32:407–451.
- Fan J, Gijbels I. Censored regression: nonparametric techniques and their applications. *Journal of the American Statistical Association*. 1994; 89:560–570.
- Fine J, Ying Z, Wei LJ. On the linear transformation model for censored data. *Biometrika*. 1998; 85:980–986.
- Fu WJ. Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*. 1998; 7:397–416.
- Hall P, Li KC. On almost linearity of low dimensional projections from high dimensional data. *Annals of Statistics*. 1993; 21:867–889.
- Hotelling H. Relations between two sets of variates. *Biometrika*. 1936; 28:321–377.
- Kong L, Cai J, Sen PK. Weighted estimating equations for semiparametric transformation model with censored data from a case-cohort design. *Biometrika*. 2004; 91:305–319.
- Jones MC. The performance of kernel density functions in kernel distribution function estimation. *Statistics and Probability Letters*. 1990; 9:129–132.
- Koul H, Susarla V, van Ryzin J. Regression analysis with randomly right-censored data. *Annals of Statistics*. 1981; 9:1276–1288.

- Li H, Luan Y. Boosting proportional hazards models using smoothing spline, with application to high-dimensional microarray data. *Bioinformatics*. 2005; 21:2403–2409. [PubMed: 15713732]
- Li KC. Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*. 1991; 86:316–327.
- Li KC, Wang JL, Chen CH. Dimension reduction for censored regression data. *The Annals of Statistics*. 1999; 27:1–23.
- Li L, Li H. Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*. 2004; 20:3406–3412. [PubMed: 15256406]
- Li L, Simonoff JS, Tsai CL. Tobit model estimation and sliced inverse regression. *Statistical Modelling*. 2007; 7:107–123.
- Lu W, Li L. Boosting method for nonlinear transformation models with censored survival data. *Biostatistics*. 2008; 9:658–667. [PubMed: 18344565]
- Peng L, Fine J. Competing Risks Quantile Regression. *Journal of the American Statistical Association*. 2009; 104:1440–1453.
- Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Staudt LM. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*. 2002; 346:1937–1947. [PubMed: 12075054]
- Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
- Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics*. 1990; 18:354–372.
- Wang H, Leng C. Unified lasso estimation via least squares approximation. *Journal of the American Statistical Association*. 2007; 102:1039–1048.
- Xia Y, Zhang D, Xu J. Dimension reduction and semiparametric estimation of survival models. *Journal of the American Statistical Association*. 2010; 105:278–290.
- Yin X, Cook RD. Dimension reduction for the conditional k th moment in regression. *Journal of the Royal Statistical Society, Series B*. 2002; 64:159–175.
- Ying Z. A large sample study of rank estimation for censored regression data. *Annals of Statistics*. 1993; 21:76–99.
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*. 2006; 68:49–67.
- Zhang HH, Lu W. Adaptive-LASSO for Cox's proportional hazards model. *Biometrika*. 2007; 94:1–13.
- Zhu LX, Miao B, Peng H. On sliced inverse regression with large dimensional covariates. *Journal of the American Statistical Association*. 2006; 101:630–643.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.

Figure 1.

The ICPW estimation of $S_{Y|X}$: median vector correlation coefficient versus the sample size. The solid line denotes the benchmark with no censoring; the dashed line denotes the ICPW estimator with the censoring distribution estimated by a PH model; the dot-dashed line denotes the ICPW estimator with the censoring distribution estimated by the Kaplan-Meier method without covariate information; the dotted line denotes the ICPW estimator with the censoring distribution estimated by the local Kaplan-Meier method based on DSIR; the long-dashed line denotes the DSIR estimator.

Figure 2.

Analysis of the PBC data: the left panel shows the entire solution path, where the vertical line denotes the selected tuning parameter. The right panel shows the Kaplan-Meier estimate of survival curves for the three risk groups of patients.

(a) training data

(b) testing data

Figure 3.

Analysis of the DLBCL data: the left panel shows the Kaplan-Meier estimate of survival curves for the three risk groups of patients in the training data, and the right panel shows for the testing data.

Table 1

Sensitivity analysis: median vector correlation coefficient for various censoring time distributions different than the proportional hazards model. For each scenario, we report the results when the censoring time distribution follows model (10) with a varying value of r , a log-normal model, and a PH model with interaction terms.

Model	ε	ξ_c	$n = 400$			$n = 800$			$n = 1200$		
			ICPW	DSIR	ICPW	DSIR	ICPW	DSIR	ICPW	DSIR	DSIR
(7)	PH	$r = 0.25$	0.969	0.835	0.976	0.838	0.980	0.853			
		$r = 0.50$	0.968	0.852	0.969	0.858	0.971	0.875			
		$r = 0.75$	0.964	0.868	0.969	0.879	0.965	0.891			
		$r = 1.00$	0.964	0.888	0.957	0.900	0.958	0.905			
		log normal	0.962	0.725	0.967	0.736	0.964	0.748			
		mis-spec. PH	0.969	0.857	0.976	0.857	0.981	0.870			
(7)	PO	$r = 0.25$	0.968	0.782	0.977	0.780	0.981	0.794			
		$r = 0.50$	0.967	0.800	0.972	0.803	0.972	0.819			
		$r = 0.75$	0.960	0.824	0.967	0.821	0.964	0.844			
		$r = 1.00$	0.959	0.831	0.958	0.841	0.958	0.856			
		log normal	0.958	0.663	0.966	0.674	0.967	0.678			
		mis-spec. PH	0.967	0.813	0.975	0.818	0.980	0.829			
(8)	PH	$r = 0.25$	0.757	0.491	0.829	0.562	0.846	0.593			
		$r = 0.50$	0.729	0.484	0.789	0.559	0.833	0.589			
		$r = 0.75$	0.695	0.525	0.780	0.587	0.803	0.618			
		$r = 1.00$	0.687	0.554	0.771	0.614	0.805	0.650			
		log normal	0.716	0.393	0.727	0.496	0.776	0.511			
		mis-spec. PH	0.761	0.449	0.812	0.529	0.850	0.566			
(8)	PO	$r = 0.25$	0.750	0.440	0.821	0.519	0.850	0.555			
		$r = 0.50$	0.710	0.461	0.816	0.546	0.849	0.575			
		$r = 0.75$	0.697	0.475	0.775	0.561	0.813	0.593			
		$r = 1.00$	0.694	0.507	0.777	0.570	0.811	0.608			
		log normal	0.715	0.367	0.737	0.475	0.751	0.492			
		mis-spec. PH	0.768	0.449	0.811	0.522	0.856	0.563			

Model assessment and comparison: median mean squared error of the estimated covariate effects. Under comparison are the Cox model fittings based on the ICPW estimate, the DSIR estimate, the original predictors, and the spline basis of the original predictors, respectively.

Table 2

Model	ε	Method	$n = 400$	$n = 800$	$n = 1200$
(7)	PH	Cox-ICPW	0.656	0.619	0.603
		Cox-DSIR	0.797	0.643	0.591
		Cox-Linear	1.200	1.293	1.351
		Cox-Spline	0.864	0.926	0.990
(7)	PO	Cox-ICPW	0.888	0.865	0.858
		Cox-DSIR	1.091	0.978	0.936
		Cox-Linear	1.321	1.369	1.435
		Cox-Spline	1.046	1.124	1.175
(8)	PH	Cox-ICPW	0.511	0.403	0.353
		Cox-DSIR	1.052	1.017	0.977
		Cox-Linear	0.864	0.890	0.932
		Cox-Spline	0.755	0.796	0.818
(8)	PO	Cox-ICPW	0.624	0.548	0.523
		Cox-DSIR	1.122	1.104	1.131
		Cox-Linear	0.911	0.938	0.975
		Cox-Spline	0.816	0.881	0.900

Table 3

Variable selection: estimation accuracy in terms of variable selection (correctly identified zeros and incorrectly identified zeros) and basis estimation (vector correlation coefficient).

Model	ϵ	# of true 0	$n = 400$				$n = 800$			
			cor 0	inc 0	ver	cor 0	inc 0	ver	cor 0	ver
(7)	odds	7	ICPW	0.00	0.00	0.942	0.00	0.00	0.00	0.967
			ICPW _s	5.87	0.00	0.958	6.11	0.00	0.00	0.979
		4	ICPW	0.00	0.00	0.916	0.00	0.00	0.00	0.953
			ICPW _s	3.00	0.00	0.902	4.00	0.00	0.00	0.957
(8)		0	ICPW	0.00	0.00	0.908	0.00	0.00	0.00	0.941
			ICPW _s	0.00	1.00	0.858	0.00	0.00	0.00	0.918
	hazards	7	ICPW	0.00	0.00	0.805	0.00	0.00	0.00	0.850
			ICPW _s	4.63	0.64	0.652	4.02	0.33	0.33	0.779
		4	ICPW	0.00	0.00	0.729	0.00	0.00	0.00	0.845
			ICPW _s	4.00	0.00	0.718	4.00	0.00	0.00	0.881
		0	ICPW	0.00	0.00	0.683	0.00	0.00	0.00	0.819
			ICPW _s	0.00	1.00	0.620	0.00	0.00	0.00	0.808