The Annals of Statistics
1999, Vol. 27, No. 1, 1–23

DIMENSION REDUCTION FOR CENSORED REGRESSION DATA

BY KER-CHAU LI,[1] JANE-LING WANG[2] AND CHUN-HOUH CHEN

University of California, Los Angeles, University of California, Davis and Academia Sinica, Taiwan

Without parametric assumptions, high-dimensional regression analysis is already complex. This is made even harder when data are subject to censoring. In this article, we seek ways of reducing the dimensionality of the regressor before applying nonparametric smoothing techniques. If the censoring time is independent of the lifetime, then the method of sliced inverse regression can be applied directly. Otherwise, modification is needed to adjust for the censoring bias. A key identity leading to the bias correction is derived and the root-n consistency of the modified estimate is established. Patterns of censoring can also be studied under a similar dimension reduction framework. Some simulation results and an application to a real data set are reported.

1. Introduction. Survival data are often subject to censoring. When this occurs, the incompleteness of the observed data may induce a substantial bias in the sample. Several approaches have been suggested to overcome the associated difficulties in regression, including the accelerated failure time model, censored linear regression, the Cox proportional hazard model and many others. Survival analysis becomes even more intricate when the dimension of the regressor increases. To apply any of the aforementioned methods, users are required to specify a functional form which relates the outcome variables to the input ones. However, in reality, knowledge needed for an appropriate model specification is often inadequate. As a matter of fact, the acquisition of such information may well turn out to be one of the primary goals of the study itself. Under such circumstances, it seems preferable to have exploratory tools that rely less on such model specification. This is the issue to be addressed in this article. The dimension reduction approach of Li (1991) will be extended to settings which allow for censoring in the data. We shall offer methods of finding low-dimensional projections of the data for visually examining the censoring pattern. We shall show how censored regression data can still be analyzed without assuming the functional form a priori.

Received January 1997; revised September 1998.

[1] Supported in part by an NSF grant and a Guggenheim Fellowship. Part of the research was carried out while visiting the Institute of Statistical Science, Academia Sinica, Taiwan, in 1994 with support from the National Science Council, Taiwan, R.O.C.

[2] Supported in part by NSF Grant DMS-93-12170.

AMS 1991 *subject classifications*. 62G05, 62J20.

Key words and phrases. accelated failure time model, censored linear regression, Cox model, curse of dimensionality, hazard function, Kaplan-Meier estimate, regression graphics, sliced inverse regression, survival analysis

Dimensionality sets a severe limitation even in the exploratory stage of data analysis. This is true even without the presence of censoring. For example, when the dimension is one or two, a two-dimensional or three-dimensional scatterplot of the response variable against the regressor is helpful in obtaining general ideas about the shape of the regression function, the pattern of heterogeneity and other valuable structural information. However, as the dimension increases, the total number of two-dimensional or three dimensional scatterplots escalates quickly. Very soon this task could turn into an extremely laborious exercise. Without proper guidance, it may not be easy for us to put together a clear global picture about the data from various plots. How to bypass the curse of dimensionality has been an important issue; see, for example, Huber (1985).

Li's framework for dimension reduction in regression begins with the following formulation:

$$Y = g(\beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x}, \varepsilon). \tag{1.1}$$

The main feature of (1.1) is that g is completely unknown and so is the distribution of ε, which is independent of the p-dimensional regressor $\mathbf{x}$. When k is smaller than p, (1.1) imposes a dimension reduction structure by claiming that the dependence of Y on the p-dimensional $\mathbf{x}$ only comes from the k variates, $\beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x}$, but *the functional form of the dependence structure is not specified*. The k-dimensional space spanned by the $k\,\beta$ vectors is called the e.d.r. (effective dimension reduction) space and any vector in this space is referred to as an e.d.r. direction. The primary goal of Li's approach is to estimate the e.d.r. directions so that we can plot y against the e.d.r. variates for visually exploring the structure of the regression and for more effectively applying various low-dimensional regression techniques to the reduced space. The notion of e.d.r. space and its role in regression graphics are further explored in Cook (1994) and Cook and Weisberg (1994).

To incorporate censoring into the dimension reduction framework, let

Y^o = the true (unobservable) lifetime,

C = the censoring time,

δ = the censoring indicator; $\delta = 1$, if $Y^o \le C$ and $\delta = 0$, otherwise,

$Y = \min\{Y^o, C\}$, the observed time.

We assume that

$$Y^o \text{ follows model (1.1)}; \tag{1.2}$$

$$\text{Conditional on } \mathbf{x}, C \text{ is independent of } Y^o. \tag{1.3}$$

The observed sample consists of n i.i.d. observations, (Y_i, x_i, δ_i), $i = 1, \ldots, n$ from the distribution of $(Y, \mathbf{x}, \delta)$. The continuous random variables, Y^o, C, are not observed. Condition (1.3) is the usual independence assumption to ensure identifiability under the random censoring scheme. If (1.3) is violated, then one needs more information on the censoring mechanism to build an appropriate model. This is not considered in this paper.

For $k = 1$, our formulation may include the generalized linear model [McCullagh and Nelder (1989)] and the linear transformation model [Doksum (1987)] as special cases. The latter also includes several survival analysis models such as the accelerated failure time model, the proportional hazard model, the proportional odds model and the logit and probit models [Doksum and Gasko (1990)].

Without censoring, sliced inverse regression (SIR) is a simple method for finding the e.d.r. space. Instead of directly estimating $E(Y \mid \mathbf{x})$, a p-dimensional surface, the roles of $\mathbf{x}$ and Y are reversed—the focus turns to the inverse regression $E(\mathbf{x} \mid Y)$, which is a curve in R^p. Under appropriate conditions [Lemma 3.1 of Li (1991)], the inverse regression curve is shown to fall into a k-dimensional subspace. In particular, when the regressor distribution has mean zero and with the identity covariance, this k-dimensional subspace coincides with the e.d.r. space. Exploring this connection, SIR begins with a simple estimate of the inverse regression curve by partitioning the data into several slices according to the Y values and computing the mean of $\mathbf{x}$ within each slice; this is the slicing step. It is then followed by an eigenvalue decomposition step—a principal component type of analysis intended to locate the subspace containing the inverse regression curve. See Li (1991) for further details. Properties of SIR have been studied in several places: Carroll and Li (1992, 1995), Chen and Li (1998), Cook and Weisberg (1991, 1994), Duan and Li (1991), Hsing and Carroll (1992), Schott (1994), Zhu and Ng (1995).

How does censoring affect SIR? This depends on the relationship between the censoring time C and the regressor $\mathbf{x}$. Section 2 considers the independence case in which

$$C \text{ is independent of } \mathbf{x} \text{ and } Y^o. \tag{1.4}$$

We show that the general theory of SIR is applicable without modification and the directions found by SIR are still consistent. Thus for the independence case, censoring does not introduce bias to the SIR estimates.

However, SIR will be affected by other censoring mechanisms that do not follow (1.4). In Section 3, we introduce a general strategy to overcome this difficulty. The proposed approach is to introduce a suitable weight function for the censored observations for offsetting bias in estimating the slice means. The weight function can be estimated by nonparametric estimation techniques for conditional survival functions. For simplicity, the kernel method is used and we establish the root-n consistency for the modified SIR.

In Section 4, we bring out a dimension reduction setting for studying the pattern of censoring when the independent censoring condition (1.4) is violated. We argue for the importance of visualizing the heavy censoring region, a nontrivial task in the high-dimensional situation. Data analysts have to recognize this region because heavy censoring sets severe limitations in finding the structure of regression. The dimension reduction assumption on C is a natural counterpart of (1.2),

$$C = h(\gamma_1'\mathbf{x}, \ldots, \gamma_c'\mathbf{x}, \varepsilon'). \tag{1.5}$$

Then (1.2) and (1.5) together allow us to treat the survival time and censoring time equivalently. But to avoid confusion, we shall refer to γ_i's and their linear combinations as e.d.r. *censoring* directions. In contrast, the e.d.r. directions for Y^o will be called e.d.r. *lifetime* directions. Both censoring and lifetime directions as well as their linear combinations will be called joint e.d.r. directions. We show how to estimate the joint e.d.r. directions through a double slicing procedure.

From the joint e.d.r. directions, we can recover the e.d.r. lifetime directions by further applying the modified SIR strategy of Section 3. This is illustrated in Section 5. The performance of the procedure is examined through two simulation studies. We apply our method to a data set about the study of primary biliary cirrhosis (PBC) at the Mayo Clinic. Section 6 concludes this article by summarizing our findings. Some questions are raised for further study.

2. SIR under the independence assumption (1.4). Denote the uncensored inverse regression curve by $\eta^o(y^o) = E(\mathbf{x} \mid Y^o = y^o)$. Without censoring (i.e., $Y = Y^o$), the population version of SIR is based on the following eigenvalue decomposition:

$$\begin{aligned} \Sigma_{\eta^o} b_i &= \lambda_i \Sigma_{\mathbf{x}} b_i, \\ \lambda_1 &\geq \cdots \geq \lambda_p, \end{aligned} \tag{2.1}$$

where

$$\Sigma_{\eta^o} = \operatorname{cov}(E(\mathbf{x} \mid Y^o)) \tag{2.2}$$

and

$$\Sigma_{\mathbf{x}} = \operatorname{cov}(\mathbf{x}).$$

The justification for using the first k eigenvectors b_i with nonzero eigenvalues to estimate the e.d.r. (lifetime) directions follows from Lemma 3.1 of Li (1991), which can be stated as follows.

LEMMA 2.1. *Assume that the dimension reduction assumption* (1.2) *holds. Then for any* y^o, $\Sigma_{\mathbf{x}}^{-1}(\eta^o(y) - E(\mathbf{x}))$ *falls into the e.d.r.* (*lifetime*) *space under the condition that*

$$\text{for any vector } b,\ E(b'\mathbf{x} \mid \beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x}) \text{ is linear}. \tag{2.3}$$

Design condition (2.3) has been discussed in several places. The performance of SIR is not very sensitive to this condition; see the discussion and the rejoinder of Li (1991), Cook (1994), Cook and Weisberg (1991, 1994). Carroll and Li (1995). In view of the fact that most low-dimensional projections of high-dimensional data often appear like normal distributions [Diaconis and Freedman (1984)], Hall and Li (1993) argue for the generality of this condition in high-dimensional situations. On the other hand, reweighting and subsampling methods can also be applied to achieve (2.3): Brillinger (1991), Cook and Nachtsheim (1994). Further discussion on this condition can be found in Li (1997).

Censoring alters the distribution of the observed time Y. Its effect on SIR can be studied by comparing the censored inverse regression curve $\eta(y) = E(\mathbf{x} \mid Y = y)$ with the uncensored one $\eta^o(y^o)$. By conditioning, we have

$$E(\mathbf{x} \mid Y = y) = E\big(E(\mathbf{x} \mid Y^o, C) \mid Y = y\big). \tag{2.4}$$

Under (1.4), $E(\mathbf{x} \mid Y^o, C)$ is equal to $E(\mathbf{x} \mid Y^o)$, implying that

$$E(\mathbf{x} \mid Y = y) = E\big(\eta^o(Y^o) \mid Y = y\big). \tag{2.5}$$

Since Lemma 2.1 applies to $\eta^o(y^o)$, the following result is obtained.

LEMMA 2.2. *Assume that* (1.2) *and* (1.4) *hold. Then* $\Sigma_{\mathbf{x}}^{-1}(\eta(y) - E(\mathbf{x}))$ *falls into the e.d.r lifetime space under* (2.3).

To implement SIR on the data $(Y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, we follow Li (1991). First we partition $y_i's$ into H intervals, I_h, $h = 1, \ldots, H$. Then for each interval, we compute the partition slice mean $\overline{\mathbf{x}}_h$ by averaging

$$\overline{\mathbf{x}}_h = \frac{1}{n_h} \sum_{\mathbf{x}_i \in I_h} \mathbf{x}_i,$$

where n_h is the number of cases falling into I_h. Then the covariance matrix

$$\sum_{h=1}^{H} \frac{n_h}{n} (\overline{\mathbf{x}}_h - \overline{\mathbf{x}})(\overline{\mathbf{x}}_h - \overline{\mathbf{x}})' \equiv \hat{\Sigma}_\eta$$

is formed. Finally we conduct the eigenvalue decomposition

$$\hat{\Sigma}_\eta \hat{b}_i = \hat{\lambda}_i \Sigma_{\mathbf{x}} \hat{b}_i,$$
$$\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p.$$

With Lemma 2.2 and following the argument in Li (1991), we obtain the root-n consistency of SIR estimates $\hat{b}_i$ for finding e.d.r. lifetime directions. Thus censoring does not introduce bias to SIR. However, this is true only when the censoring time is independent of the regressors and the true lifetime. Without (1.4), this appealing result vanishes and substantial bias may be induced by censoring under the more general condition (1.3).

3. A strategy for modifying SIR under (1.3). An ideal way of bypassing the difficulties caused by general censoring (1.3) is to slice the true survival time Y^o. At first sight, this does not appear feasible because under censoring, Y^o is unobservable. The promise comes from an identity derived in Section 3.1, which relates the conditional expectation of $\mathbf{x}$ in each slice to the observed time Y and the censored indicator. This leads to a modified slicing step by a suitable weighting scheme for offsetting the censoring bias in estimating the slice means. The consistency of this new procedure is discussed in Section 3.2.

3.1. *An identity.* Let $0 = t_1 < t_2 < \cdots < t_H < \infty = t_{H+1}$ be a partition on the survival time. The expected value of $\mathbf{x}$ in a slice, $\mathbf{m}_j = E\{\mathbf{x} \mid Y^o \in [t_j, t_{j+1})\}$, can be written as

$$\mathbf{m}_j = \frac{E\left\{\mathbf{x}1\left(Y^o \in \left[t_j, t_{j+1}\right)\right)\right\}}{P\left\{Y^o \in \left[t_j, t_{j+1}\right)\right\}} = \frac{E\left\{\mathbf{x}1\left(Y^o \geq t_j\right)\right\} - E\left\{\mathbf{x}1\left(Y^o \geq t_{j+1}\right)\right\}}{E\left\{1\left(Y^o \geq t_j\right)\right\} - E\left\{1\left(Y^o \geq t_{j+1}\right)\right\}}, \tag{3.1}$$

where $1(\cdot)$ is the indicator function. The two numerator terms take the same form, which involves the unobservable indicator $1(Y^o \geq t)$. They can be converted into terms with Y and δ via the identity,

$$E\{\mathbf{x}1(Y^o \geq t)\} = E\{\mathbf{x}1(Y \geq t)\} + E\{\mathbf{x}1(Y < t, \delta = 0)w(Y, t, \mathbf{x})\}, \tag{3.2}$$

where for $t' < t$,

$$w(t', t, \mathbf{x}) = \frac{S^o(t \mid \mathbf{x})}{S^o(t' \mid \mathbf{x})}, \tag{3.3}$$

$$\begin{aligned} S^o(t \mid \mathbf{x}) &= P\{Y^o \geq t \mid \mathbf{x}\} \\ &= \text{conditional survival function for } Y^o, \text{ given } \mathbf{x}. \end{aligned} \tag{3.4}$$

Consider the plane of variables Y^o and C in Figure 1. The integration region $Y^o \geq t$ is decomposed into two parts. The first region (area **I** in Figure 1) with $Y^o \geq t$, $C \geq t$ or equivalently, $Y \geq t$, contributes to the first term on the right side of (3.2). The second region with $Y^o \geq t$, $C < t$, falls into the censored area, $\delta = 0$. It is contained in the larger region (dashed area **II** in Figure 1) with $Y < t$ and $\delta = 0$. The second term on the right side of (3.2) comes from integration over this larger region with the weight adjustment

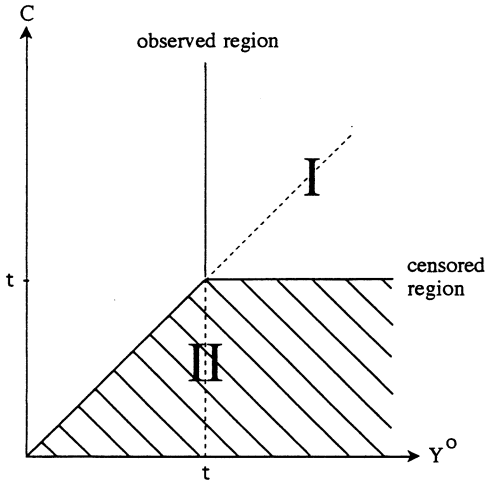

FIG. 1. *Integration regions.*

$w(\cdot,\cdot,\cdot)$. Conditioning is the key to justify this term:

$$\begin{aligned} E\{\mathbf{x}1(Y^o \geq t, C < t)\} &= E\{\mathbf{x}1(Y < t, \delta = 0)1(Y^o \geq t)\} \\ &= E\{\mathbf{x}1(Y < t, \delta = 0)E[1(Y^o \geq t) \mid Y, \delta = 0, \mathbf{x}]\} \\ &= E\{\mathbf{x}1(Y < t, \delta = 0)E[1(Y^o \geq t) \mid C, Y^o > C, \mathbf{x}]\} \\ &= E\{\mathbf{x}1(Y < t, \delta = 0)w(C, t, \mathbf{x})\} \\ &= E\{\mathbf{x}1(Y < t, \delta = 0)w(Y, t, \mathbf{x})\}. \end{aligned}$$

Here the next to the last equality is due to the conditional independence assumption (1.3), which assures that conditional on $\mathbf{x}$, the probability for the true survival time Y^o to exceed t given $C = t'$ and $Y^o \geq t'$ is equal to the conditional probability given by (3.3).

By a similar argument, the denominator terms can be converted via the identity

$$E\{1(Y^o \geq t)\} = E\{1(Y \geq t)\} + E\{1(Y < t, \delta = 0)w(Y, t, \mathbf{x})\}. \tag{3.5}$$

The weight function (3.3) can be further expressed as

$$w(t', t, \mathbf{x}) = \exp\{-\Lambda(t', t \mid \mathbf{x})\}, \tag{3.6}$$

where

$$\Lambda(t', t \mid \mathbf{x}) = E\left\{ \frac{1(t' < Y < t, \delta = 1)}{S_Y(Y \mid \mathbf{x})} \middle| \mathbf{x} \right\},$$

$S_Y(\cdot \mid \mathbf{x})$ = the conditional survival function of Y conditional on $\mathbf{x}$.

Then (3.6) follows from the well-known relationship between survival functions and cumulated hazards; for a proof, see the Appendix. The term $\Lambda(t', t \mid \mathbf{x})$ is simply the integrated conditional hazard (given $\mathbf{x}$) function over the interval $[t', t]$.

3.2. *Estimation*. To construct an estimate for $\mathbf{m}_j$, we replace each expectation term in (3.2) and (3.5) by the corresponding first sample moment,

$$\hat{\mathbf{m}}_j = \frac{\hat{E}\{\mathbf{x}1(Y^o \geq t_j)\} - \hat{E}\{\mathbf{x}1(Y^o \geq t_{j+1})\}}{\hat{P}\{Y^o \geq t_j\} - \hat{P}\{Y^o \geq t_{j+1}\}}, \tag{3.7}$$

$$\hat{E}\{\mathbf{x}1(Y^o \geq t)\} = n^{-1} \sum_{i:\, Y_i \geq t}^{n} \mathbf{x}_i + n^{-1} \sum_{i:\, Y_i < t,\, \delta_i = 0}^{n} \mathbf{x}_i \hat{w}(Y_i, t, \mathbf{x}_i), \tag{3.8}$$

$$\hat{P}\{Y^o \geq t\} = \#\{i:\, Y_i \geq t\}/n + n^{-1} \sum_{i:\, Y_i < t,\, \delta_i = 0}^{n} \hat{w}(Y_i, t, \mathbf{x}_i), \tag{3.9}$$

where $\hat{w}(\cdot,\cdot,\cdot)$ denotes an estimate of the weight function (3.3) to be discussed later.

After estimating each slice mean by (3.7), we can form the covariance matrix of the slice means in the usual way;

$$\hat{\Sigma}_{\eta_o} = \sum_j \big(\hat{\mathbf{m}}_j - \bar{\mathbf{x}}\big)\big(\hat{\mathbf{m}}_j - \bar{\mathbf{x}}\big)' \hat{p}_j,$$

$$\hat{p}_j = \hat{P}\big\{Y^o \geq t_j\big\} - \hat{P}\big\{Y^o \geq t_{j+1}\big\}.$$

Finally, we may conduct the eigenvalue decomposition as before to find the SIR directions

$$\begin{aligned} \hat{\Sigma}_{\eta^o} \hat{b}_i^o &= \hat{\lambda}_i \hat{\Sigma}_{\mathbf{x}} \hat{b}_i^o, \\ \hat{\lambda}_1 &\geq \cdots \geq \hat{\lambda}_p. \end{aligned} \tag{3.10}$$

Smoothing is needed in estimating $w(t', t, \mathbf{x})$. There are several ways to proceed. For example, we can apply Beran's estimates for conditional survival functions and their variants. Under appropriate conditions, Beran (1981) and Dabrowska (1987, 1992) established the consistency of their estimates at convergence rates (slower than the root n rate) similar to those commonly found in nonparametric regression. These consistency results lead to the consistency of $\hat{\mathbf{m}}_h$ as an estimate of $\mathbf{m}_h$. It is easy to see that $\hat{\Sigma}_{\eta^o}$ is also consistent for the covariance matrix of the slice means $\mathbf{m}_h$s. As in Li (1991), we can apply Lemma 2.1 to establish the consistency of $\hat{b}_i$ as estimates of e.d.r. lifetime directions.

Despite the slow rate of convergence in estimating conditional survival functions [hence the weight (3.3)], it is still possible to establish the root n convergence for $\hat{\mathbf{m}}_h$. We only consider the kernel smoothing method here for simplicity. Let $K_p(\cdot)$ be a kernel function on R^p and h_n be the bandwidth in each coordinate. We shall assume that $h_n = o(1)$ and nh_n^p tends to infinity. Further constraints will be imposed later. It is common for $K_p(\cdot)$ to take a product form, $K_p(x_1, \ldots, x_p) = K(x_1) \cdots K(x_p)$, for some one-dimensional kernel function $K(\cdot)$. Our kernel estimate of (3.6) is defined by setting

$$\hat{\Lambda}(t', t \mid \mathbf{x}) = \frac{n^{-1} \sum_{i:\, t' < Y_i < t,\, \delta_i = 1} \big(\hat{S}_Y(Y_i \mid \mathbf{x}_i)\big)^{-1} h_n^{-p} K_p\big(h_n^{-1}(\mathbf{x}_i - \mathbf{x})\big)}{\hat{f}(\mathbf{x})}, \tag{3.11}$$

$$\hat{S}_Y(Y_i \mid \mathbf{x}_i) = \frac{n^{-1} \sum_{j:\, Y_j > Y_i}^n h_n^{-p} K_p\big(h_n^{-1}(\mathbf{x}_j - \mathbf{x}_i)\big)}{\hat{f}(\mathbf{x}_i)}, \tag{3.12}$$

$$\hat{f}(\mathbf{x}) = n^{-1} \sum_i^n h_n^{-p} K_p\big(h_n^{-1}(\mathbf{x}_i - \mathbf{x})\big). \tag{3.13}$$

A sketch proof of the following claim together with the regularity conditions needed is given in the Appendix.

LEMMA 3.1. *Under the regularity conditions* (B.1), (B.3), (B.5) *and* (B.8), *given in the Appendix*, $\hat{\mathbf{m}}_h$ *is a root-n consistent estimate for* $\mathbf{m}_h$, $h = 1, \ldots, H$.

We can use this lemma and follow the same argument as in Li (1991) to show that the modified SIR is root-n consistent. This is stated in the following theorem. The proof is omitted.

THEOREM 3.2. *Under the assumption of Lemma* 3.1, *each* $\hat{b}_h$ *is a root-n consistent estimate for an e.d.r. direction.*

Assume that $f(\cdot)$ and $S(\cdot \mid \mathbf{x})$ are d-times continuously differentiable and that the kernel function satisfied the moment conditions $\int x^i K_p(x)\,dx = 0$ for $i = 1, \ldots, d-1$, and $\int x^d K_p(x)\,dx$ is nonzero. Then the regularity assumptions in Lemma 3.1 can be satisfied with bandwidth $H \propto n^{-1/2d}$ provided $p \leq d$.

What we have presented so far in this section is a general strategy for offsetting the bias due to censoring. The theoretical result of Theorem 3.2, however, may not help much in practice. The problem is that kernel smoothing only works well in the low-dimensional case. Thus, before applying the kernel method in estimating the weight function, we may want to reduce the dimensionality first. This is to be discussed in the next two sections.

4. Dimension reduction model for censoring time. Analyzing the censoring pattern is an important step in studying the censored data. It helps the recognition of the information-poor region in $\mathbf{x}$, the region where censoring is heavy and the regression structure is thus harder to explore. Sometimes such an analysis may even become a primary part of the study. In some industrial applications, Y^o may be the potential yield of a production process and censoring C may occur because of machine malfunctioning, for example. In addition to learning how various input variables $\mathbf{x}$ may affect the potential yield, quality control engineers may equally be interested in how they affect the censoring rate; they need such knowledge to prevent machine malfunctioning as much as possible.

Like its counterpart Y^o, we now assume that the censoring time C also has a dimension reduction structure given by (1.5). Again, the functional form of h and the distributional form of ε' are both unspecified. This model suggests only that the dimension of the regressor can be reduced from p to c. The relationship between the e.d.r. space for the censoring time and the e.d.r. space for the true lifetime is arbitrary. They can be either identical, partly overlapped, or disjoint. Linear combinations of their elements form a space which will be called the *joint* e.d.r. space. If Y^o and C were used for slicing, then by the same argument used in deriving Lemma 2.1, it is easy to see that

(4.1) $\quad \Sigma_{\mathbf{x}}^{-1}(E(\mathbf{x} \mid Y^o, C) - E(\mathbf{x}))$ falls into the joint e.d.r. space.

However, instead of (Y^o, C), we can only observe Y and δ. This suggests that Y and δ can be used simultaneously for slicing. Let $\eta_d(Y, 0) = E(\mathbf{x} \mid Y = y, \delta = 0)$, and $\eta_d(Y, 1) = E(\mathbf{x} \mid Y = y, \delta = 1)$. We may replace (2.1) with

$$\text{(4.2)} \qquad \begin{aligned} \operatorname{Cov}(\eta_d(Y, \delta))b_i &= \lambda_i \Sigma_{\mathbf{x}} b_i, \\ \lambda_1 &\geq \cdots \geq \lambda_p. \end{aligned}$$

By conditioning, $E(\mathbf{x} \mid Y, \delta) = E(E(\mathbf{x} \mid Y^o, C) \mid Y, \delta))$. Thus from (4.1), we see that

$$\Sigma_{\mathbf{x}}^{-1}\big(\eta_d(y, \delta) - E(\mathbf{x})\big) \text{ falls into the joint e.d.r. space.}$$

This justifies the use of eigenvectors from (4.2) to estimate the joint e.d.r. space.

The sample version of (4.2) is easy to carry out. Denote the number of slices for the uncensored ($\delta = 1$) observations by H_1. Let I_{1j}, $j = 1, \ldots, H_1$ be a partition of the positive real line into nonoverlapping intervals. Similarly, denote the number of slices for the censored ($\delta = 0$) observations by H_0, and let I_{0j}', $j = 1, \ldots, H_0$ be another partition of the positive real line. We first form the individual slice means by taking

$$\bar{\mathbf{x}}_{lj} = \big(n\hat{p}_{lj}\big)^{-1} \sum_{i=1}^{n} \mathbf{x}_i 1\big(\delta_i = l, Y_i \in I_{lj}\big),$$

where $\hat{p}_{lj}$ is the proportion of cases with $\delta_i = l$ falling into interval I_{lj}. Then we compute the covariance matrix for the slice means, $\hat{\Sigma}_d = \sum_l \sum_j \hat{p}_{lj}(\bar{\mathbf{x}}_{lj} - \mathbf{x})(\bar{\mathbf{x}}_{lj} - \mathbf{x})$. Finally we conduct the eigenvalue decomposition

$$\begin{aligned} \hat{\Sigma}_d \hat{b}_{di} &= \hat{\lambda}_{di} \hat{\Sigma}_{\mathbf{x}} \hat{b}_{di}, \\ \hat{\lambda}_{d1} &\geq \cdots \geq \hat{\lambda}_{dp}. \end{aligned} \tag{4.3}$$

Li (1991) proposed a chi-squared test for determining the number of significant e.d.r. directions obtained by SIR. It should be clear that we can use the same test for the double slicing case.

So far we have only located the joint e.d.r. directions. We shall show in the next section how to use the procedure in Section 3 to recover the e.d.r. lifetime directions. Likewise, we can also recover the e.d.r. directions for censoring time by exchanging the roles of censoring time and lifetime. Before we proceed, an example is given below to illustrate the double slicing procedure discussed in this section.

EXAMPLE 4.1. Take $p = 6$ and let $\mathbf{x} = (x_1, \ldots, x_6)'$ be generated from the standard normal distribution. Suppose

$$\begin{aligned} Y^o &= 4 - \big(|x_1 - 1|\big) + \sigma_1 \varepsilon_1, \\ C &= 3 + \sigma_2 \varepsilon_2 \quad \text{for } x_1 > 0,\ x_2 + x_3 > 0, \\ &= 10 \quad \text{otherwise,} \end{aligned}$$

where $\sigma_1 = \sigma_2 = 0.1$. Here ε_1, ε_2 are normal random variables. Generate 300 cases. Sixty-six observations in the data set are censored. Now apply double slicing with the number of slices equal to 5 and 10, respectively, for the censored and the uncensored groups. The eigenvalues of SIR are found to be 0.76, 0.35, 0.08, 0.06, ..., indicating that the first two eigenvectors, $\hat{b}_{d1} = (1.14, 0.05, -0.03, -0.00, -0.04, 0.04)'$ and $\hat{b}_{d2} = (-0.06, 0.69, 0.74, -0.02, -0.10, -0.05)'$ are important. This is confirmed by the chi-squared test in Li (1991).

The joint e.d.r. directions $(1, 0, 0, 0, 0, 0)'$ and $(1/\sqrt{2})(0, 1, 1, 0, 0, 0)'$ are captured successfully by $\hat{b}_{d1}$ and $\hat{b}_{d2}$. The censored cases are found to cluster in the first quadrant in the plot of the first two SIR variates; see Figure 2(c). Statistical information about the behavior of the true lifetime in that region is very sparse.

5. Implementation of modified SIR. The directions found by double slicing can be used to relieve the difficulties encountered in Section 3 when kernel smoothing is to be applied for estimating the weight function (3.3). Under the dimension reduction assumptions for both the true lifetime and the censoring time, (1.2) and (1.5), it is easy to see that the dependence of the weight function (3.3) on $\mathbf{x}$ is only through joint e.d.r. variates. This suggests the following two-stage procedure:

1. Apply double slicing on (Y, δ) and find the joint e.d.r. directions, $\hat{b}_{di}$. Let $\hat{B}_r = (\hat{b}_{d1}, \ldots, \hat{b}_{dr})$ be the matrix formed by the first r significant directions.
2. Apply r-dimensional kernel smoothing on $\hat{B}_r'\mathbf{x}$, to obtain the weight function $\hat{w}$,

$$\hat{w}(t', t, \mathbf{x}) = \exp\{-\hat{\Lambda}(t', t \mid \mathbf{x})\}, \tag{5.1}$$

where

$$\hat{\Lambda}(t', t \mid \mathbf{x}) = \frac{n^{-1}\sum_{i:\, t' < Y_i < t,\, \delta_i = 1}\left(\hat{S}_Y(Y_i \mid \mathbf{x}_i)\right)^{-1} h_n^{-r} K_r\left(h_n^{-1}\left(\hat{B}_r(\mathbf{x}_i - \mathbf{x})\right)\right)}{\hat{f}(\mathbf{x})}, \tag{5.2}$$

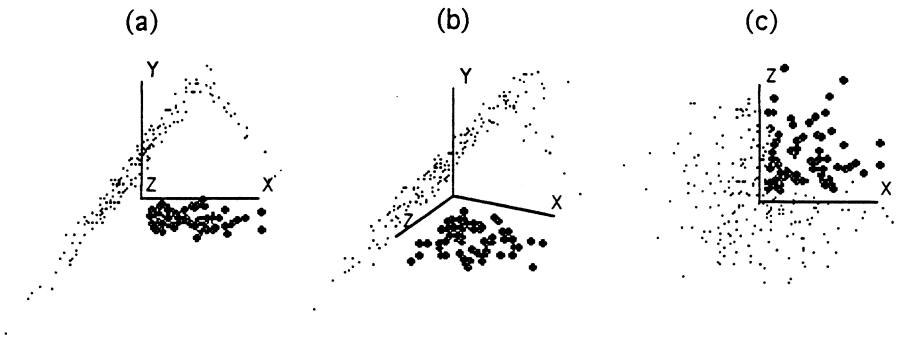

FIG. 2. *Three-dimensional scatterplot of Y against the first SIR variate (x-axis) and the second SIR variate (z-axis) found by double slicing. The highlighted points are censored.*

$$\hat{S}_Y(Y_i \mid \mathbf{x}_i) = \max\left\{ \frac{n^{-1}\sum_{j:\, Y_j > Y_i}^{n} h_n^{-r} K_r\left(h_n^{-1}\left(\hat{B}_r(\mathbf{x}_j - \mathbf{x}_i)\right)\right)}{\hat{f}(\mathbf{x}_i)}, c \right\}, \tag{5.3}$$

$$\hat{f}(\mathbf{x}) = n^{-1}\sum_{i}^{n} h_n^{-r} K_r\left(h_n^{-1}\hat{B}_r((\mathbf{x}_i - \mathbf{x}))\right). \tag{5.4}$$

Note that a small positive number c (set to 0.05 in our examples) is used to bound $\hat{S}_Y(Y_i \mid \mathbf{x}_i)$ away from zero. This is needed in order to increase the stability of the factor $\hat{S}_Y(Y_i \mid \mathbf{x}_i)^{-1}$ in (5.2). After estimating the weight function, we can apply (3.7) ~ (3.9) and then carry out the eigenvalue decomposition (3.10) to obtain estimates of e.d.r. lifetime directions.

We first report two simulation studies to illustrate how this strategy works. Then we apply our method to a data set concerning a study of primary biliary cirrhosis in the liver (PBC).

EXAMPLE 5.1. We take $p = 6$ and generate $\mathbf{x} = (x_1, \ldots, x_6)'$ from the standard normal distribution. The true survival time Y^o and the censoring time C are generated from

$$Y^o = -\frac{\log \varepsilon_1}{\varepsilon^{x_1}}; \qquad C = -\frac{\log \varepsilon_2}{\varepsilon^{x_2}},$$

where ε_1, ε_2 are independent uniform random variables from [0, 1]. Conditional on $\mathbf{x}$, Y^o and C are seen to follow the exponential distributions with the natural parameters λ_1, λ_2 linking to $\mathbf{x}$ via $\lambda_1 = \varepsilon^{x_1}$; $\lambda_2 = \varepsilon^{x_2}$, respectively.

We obtain 300 independent observations of (Y, δ); among them, 138 cases are censored. We proceed with the SIR analysis. First, the method of double slicing on Y and δ as described by (4.3) gives eigenvalues 0.34, 0.27, 0.05, ... The first two eigenvectors, $\hat{b}_{d1} = (-0.67, -0.70, -0.08, 0.06, 0.11, 0.15)'$ and $\hat{b}_{d2} = (0.69, -0.73, 0.12, -0.04, -0.10, -0.12)'$, are close to the joint e.d.r. space for Y^o and C. We use these two directions to reduce the $\mathbf{x}$ dimension before estimating the weight function $w(\cdot, \cdot, \cdot)$. With the weight adjustment given by (5.1), we perform SIR as described by (3.7) ~ (3.10) to find the e.d.r. lifetime directions. The eigenvalues are 0.40, 0.10, 0.03, ..., and the leading eigenvector is $\hat{b}_1^o = (-0.92, -0.12, -0.21, 0.08, 0.25, 0.11)'$. We see that $\hat{b}_1^o$ is quite close to the true e.d.r. lifetime direction $(1, 0, \ldots, 0)'$.

For comparison, we also carry out the SIR analysis on Y without weight adjustment as if the censoring were independent of $\mathbf{x}$. The first direction $(-0.68, -0.69, -0.058, 0.07, 0.13, 0.08)'$ does have a substantial bias. Therefore the weight adjustment is crucial in this example.

We used the bivariate normal kernel function here and the bandwidth is set at 0.18. The sensitivity to the bandwidth choice seems mild.

EXAMPLE 5.2. Important prognostic variables affecting the hazard rate may be different at different survival stages. In this example, we assume that the true survival time Y^o follows an exponential distribution with the

TABLE 1
The first three eigenvectors and eigenvalues of SIR for Example 5.2 *with the double slicing procedure*

First vector	$(-0.93, -0.11, 0.03, 0.03, 0.04, -0.06)$
Second vector	$(0.09, -0.76, -0.60, -0.01, 0.03, -0.13)$
Third vector	$(-0.10, 0.55, -0.73, -0.03, -0.02, 0.27)$
Eigenvalues	$(0.52, 0.21, 0.15, 0.03, 0.01, 0.01)$

natural parameter equal to ε^{2x_1} until time $\tau = \log 2$. From time τ on, the additional survival time follows the exponential distribution with the natural parameter ε^{3x_2}. More specifically, we assume

$$Y^* \sim \text{exponential with parameter } \lambda = \varepsilon^{2x_1},$$
$$Y^{**} \sim \text{exponential with parameter } \lambda = \varepsilon^{3x_2},$$
$$Y^o = Y^* 1(Y^* < \tau) + (\tau + Y^{**}) 1(Y^* > \tau).$$

The censoring time C follows an exponential distribution with parameter equal to $\varepsilon^{x_3 - 1}$.

Again 300 independent observations of (Y, δ) are obtained. Among them, 98 cases are censored. The output of the double slicing procedure is given in Table 1. The first three eigenvectors, which have relatively larger eigenvalues compared to the rest, are then used in estimating the weight function for finding the true e.d.r. lifetime directions. After the weight adjustment, the final output of SIR is given in Table 2. Now we see that only the first two eigenvectors stand out and the important variables x_1 and x_2 can be identified.

EXAMPLE 5.3. The PBC data set collected at the Mayo Clinic between 1974 and 1986 has been analyzed in the literature. The data set and a detailed description can be found in Fleming and Harrington (1991). There are originally seventeen regressors. Fleming and Harrington selected five of them in their final equation for fitting a Cox proportional model. These five regressors plus another variable, the platelet count (x_5 below), will be used in this illustration:

Y = number of days between registration and the earlier of death or censoring;
δ = 1 if Y is due to death; 0 otherwise;
x_1 = age in years;
x_2 = presence of edema;
x_3 = serum bilirubin, in mg/dl;
x_4 = albumin, in gm/dl;

TABLE 2
The first two eigenvectors and eigenvalues of SIR for final result of Example 5.2 with weight adjustment

First vector	(−0.97, −0.15, 0.10, −0.04, 0.10, −0.15)
Second vector	(0.16, −0.95, −0.18, −0.02, −0.06, −0.20)
Eigenvalues	(0.66, 0.34, 0.05, 0.04, 0.02, 0.02)

x_5 = platelet count;

x_6 = prothrombin time.

Cases with missing values are ignored and there are 308 cases remaining. We first apply double censoring with slice numbers $H_1 = H_0 = 10$. The first two directions are significant, as judged from the sequence of output eigenvalues 0.55, 0.15, 0.05, 0.0, 0.0, 0.0. Figure 3 shows the scatterplot of the first two SIR variates. Two outliers labeled as 104 and 276 are found from the three-dimensional plot (not shown here) of Y against the first two SIR variates. They are removed. We apply double slicing again to the remaining 306 cases. The SIR output essentially remains the same. This suggests that the dimension of the joint e.d.r. space is two.

We proceed to find the true e.d.r. lifetime directions. We take $r = 2$ and use the two SIR directions reported in Table 3 to reduce the $\mathbf{x}$ dimension before estimating the weight function. The kernel function and the bandwidth are the same as in Example 5.1. The output of the weighted SIR is given in Table 4. Judging from the eigenvalue sequence, the first direction $\hat{b}_1^o$ is

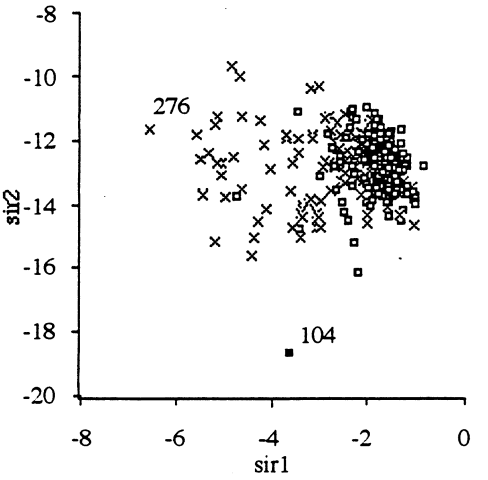

FIG. 3. *Scatterplot of the first two SIR variates found by double slicing.* × = *observed, square = censored cases.*

TABLE 3
The first two eigenvectors and eigenvalues of SIR for the PBC data in Example 5.3

First vector	(0.02, 1.04, 0.10, −0.50, −0.00, 0.39)
Second vector	(0.02, −1.62, 0.17, −0.97, −0.00, −0.87)
Eigenvalues	(0.54, 0.16, 0.05, 0.01, 0.00, 0.00)

clearly important. The second direction is also worth further examination. Figure 4(a) and (b) show the scatterplots of Y against $\hat{b}_1^{o\prime}\mathbf{x}$ and against $\hat{b}_2^{o\prime}\mathbf{x}$.

Earlier analysis in Fleming and Harrington (1991) yields that the true lifetime depends on $\mathbf{x}$ through the variate $Q = 0.0333x_1 + 0.7847x_2 + 0.8792 \log x_3 - 3.0553 \log x_4 + 3.0157 \log x_6$. This variate turns out highly correlated with the first SIR variate $\hat{b}_1^{o\prime}\mathbf{x}$; the correlation coefficient is $\sqrt{0.858}$. The correlation between Q and $1.3\hat{b}_1^{o\prime}\mathbf{x} - 0.25\hat{b}_2^{o\prime}\mathbf{x}$ is equal to $\sqrt{0.89}$. Variable x_5 makes very little contribution to the first two SIR variates, with a squared multiple correlation of only 0.11. This is consistent with Fleming and Harrington's finding that platelet count is not important.

Finally, we estimate the censoring e.d.r. directions by reversing the roles of censoring time and the true lifetime. This amounts to replacing δ with $1 - \delta$ throughout our estimation procedure. The output is given by Table 5 and Figure 5. The assumption of independent censoring (1.4) is seen to be invalid for this data set. We further notice that the first censoring time direction is quite close to the first lifetime direction. The correlation coefficient between the first lifetime SIR variate and the first censoring SIR variate turns out to be $\sqrt{0.93}$.

Some caution needs to be taken regarding the design condition. Of special concern is the second regressor (presence of edema) which is discrete and takes only three values (0, 0.5, 1). Nevertheless, the corresponding regression coefficient from Table 4 is 0.90, which is quite close to the coefficient 0.7847 based on the Cox proportional hazard model. A further study would be to carry out another SIR analysis by focusing on the group with $x_2 = 0$. The other groups have only 29 and 19 cases and thus it is not feasible to carry out separate analyses for them.

TABLE 4
The first two eigenvectors and eigenvalues of the lifetime SIR directions for the PBC data in Example 5.3

First vector	(0.02, 0.90, 0.09, −0.62, −0.00, 0.38)
Second vector	(0.03, −2.3, 0.20, −0.28, −0.00, −0.68)
Eigenvalues	(0.54, 0.16, 0.05, 0.02, 0.01, 0.00)

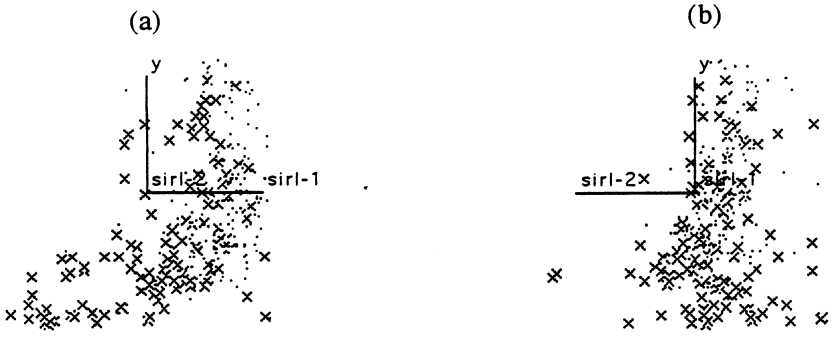

FIG. 4. *Scatterplot of Y against the first two lifetime SIR variates.* × = *observed*, *dot* = *censored cases.*

REMARK 5.1. In both of our simulation examples, we take $p = 6$. As the regressor dimension p gets larger, the problem certainly gets harder and one might expect the performance of our procedure to deteriorate as well. To study this effect, we vary p from 6 to 10, 15 and 20. The sample size is kept the same, $n = 300$. For each simulation run, we compute an R-squared term for evaluating how close to the true e.d.r. lifetime directions the estimated directions are. For the set-up of Example 5.1, which has only one true e.d.r. lifetime direction, the R-squared term is simply the squared correlation coefficient between $\hat{b}_1^{o\prime}\mathbf{x}$ and $\beta_1'\mathbf{x}$. Since $\beta_1'\mathbf{x} = x_1$, the R-squared term is equal to the square of the first coordinate of $\hat{b}_1^o$. Table 6 (left side panel) gives a summary of the R-squared values for 100 simulation runs in each case. For comparison, the R-squared values for the SIR estimate without the weight adjustment are given in the right side panel. We can see that the improvement for the modified SIR procedure is still substantial for p as large as 20.

The set-up of Example 5.2 has two true e.d.r. lifetime directions. For the first modified SIR direction, the R-squared term is just the R-squared value for regressing $\hat{b}_1^{o\prime}\mathbf{x}$ against $\beta_1'\mathbf{x}$ and $\beta_2'\mathbf{x}$ linearly. This is equal to the sum of the square of the first two coefficients in $\hat{b}_o$. The R-squared value for the second modified SIR direction is defined similarly. A summary for 100 simulation runs is given in Table 7.

TABLE 5
The first two eigenvectors and eigenvalues of the censoring time SIR variates for the PBC data in Example 5.3

First vector	(0.01, 1.43, 0.05, −0.42, −0.00, 0.55)
Second vector	(0.02, 0.38, −0.15, 1.22, 0.00, 0.85)
Eigenvalues	(0.39, 0.22, 0.05, 0.03, 0.01, 0.00)

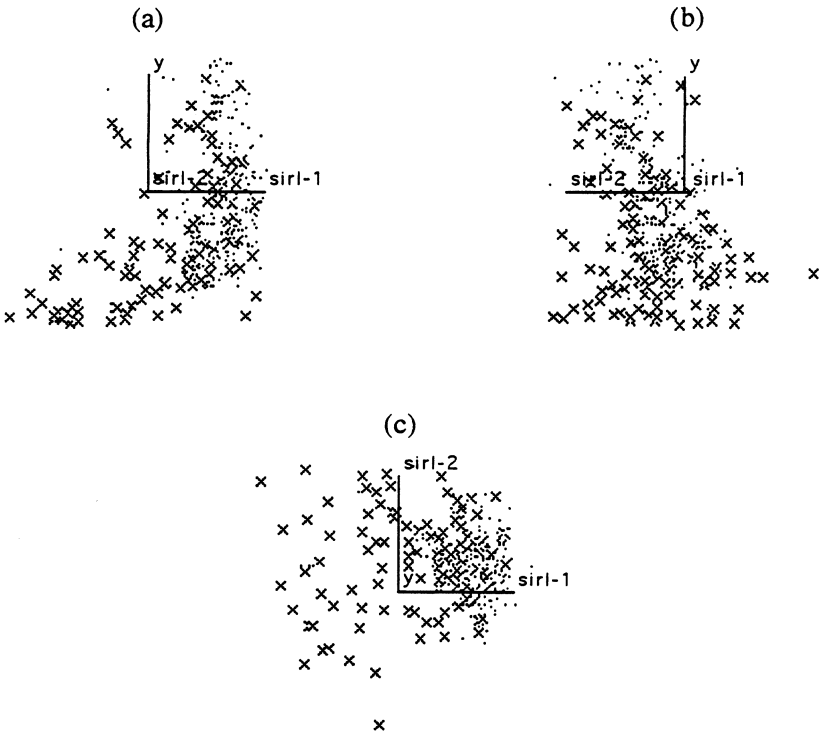

FIG. 5. *Scatterplot of Y against the first two censoring time SIR variates.* × = *observed, dot = censored observations.*

6. Conclusion. We have demonstrated how to extend the dimension reduction method of sliced inverse regression (SIR) to censored data. The extension is straightforward if censoring time is independent of the regressor. SIR can be applied to the observed data $(Y_i, \mathbf{x}_i)$ directly. However, if censoring time depends on the regressor, then SIR needs to be modified. We

TABLE 6
Performance of modified SIR as the number of regressors(p) increases under the setting of Example 5.1 *with* 100 *runs*

	Mean (standard deviation) for R^2	
p	**Modified SIR**	**Original SIR**
6	0.9172(0.0599)	0.4751(0.1100)
10	0.8630(0.0632)	0.4736(0.0937)
15	0.7963(0.0899)	0.4322(0.0915)
20	0.7576(0.0815)	0.4152(0.0881)

TABLE 7
Performance of modified SIR as the number of regressors(p) increases under the setting of Example 5.2 *with* 100 *runs*

	Mean (standard deviation) for R^2	
p	First modified SIR direction	Second modified SIR direction
6	0.9730(0.0237)	0.9132(0.0689)
10	0.9434(0.0270)	0.8455(0.0739)
15	0.9239(0.0267)	0.7911(0.0755)
20	0.8933(0.0340)	0.7149(0.1017)

introduce a weight function in estimating the slice means. The estimation of the weight function requires nonparametric smoothing. There are two options. The first one is to apply the kernel smoothing method of Section 3. This is feasible only if the number of regressors is small (e.g., $p \leq 3$) or if the sample size is substantially large. The other option, which seems more realistic, is the two-stage procedure of Section 5. We conduct a double slicing SIR first to reduce the dimension of $\mathbf{x}$ before applying kernel smoothing. This two-stage procedure relies on condition (1.5), which assumes that the censoring variable also has a dimension reduction structure with respect to the regressor. This assumption appears reasonable and it offers the possibility of examining the censoring pattern visually.

The main feature that distinguishes our approach from most other methods in survival analysis is that it does not require the estimation of g at the dimension reduction stage of data analysis. Instead, after the dimension is reduced, the estimation of g can be pursued by applying any low-dimensional smoothing methods. Furthermore, our approach can be used to check if a popular survival model is appropriate by examining the eigenvalues and the low-dimensional plots generated by SIR. These plots provide valuable information about the general pattern of censoring, possible presence of outliers and the shape of the regression surface.

Imputation is a powerful way of dealing with the incomplete censored observation. We can impute the censored Y observation first and then apply the SIR method in Li (1991) directly to the imputed data. One possible imputation method is given in Fan and Gijbels (1994). While their method is effective for one or two regressors, it is not appropriate in the higher-dimensional situation. A feasible alternative is first to apply the dimension reduction method as outlined in this article and then apply imputation to the reduced variables. This prospect merits further study.

The proof of root n consistency as outlined in the Appendix can perhaps be improved with less strenuous assumptions. While this requires further theoretical investigation, it should not affect the applicability of the procedure proposed here.

APPENDIX

A. Derivation of (3.6). It suffices to show that

$$S^o(t \mid \mathbf{x}) = \exp\left\{E\left[\frac{1(t < Y, \delta = 1)}{S_Y(Y \mid \mathbf{x})}\,\middle|\,\mathbf{x}\right]\right\}. \tag{A.1}$$

First, the conditional independence assumption (1.3) implies that $S_Y(y \mid \mathbf{x}) = S^o(y \mid \mathbf{x})S_C(y \mid \mathbf{x})$, where $S_C(y \mid \mathbf{x}) = P\{C > y \mid \mathbf{x}\}$. Using this relationship, the expectation term in (A.1) can be written as

$$\begin{aligned} E\left\{\frac{1(t < Y, \delta = 1)}{S_Y(Y \mid \mathbf{x})}\,\middle|\,\mathbf{x}\right\} &= E\left\{\frac{1(t < Y^o)1(Y^o < C)}{S^o(Y^o \mid \mathbf{x})S_C(Y^o \mid \mathbf{x})}\,\middle|\,\mathbf{x}\right\} \\ &= E\left\{\frac{1(t < Y^o)}{S^o(Y^o \mid \mathbf{x})S_C(Y^o \mid \mathbf{x})}E(1(Y^o < C) \mid \mathbf{x}, Y^o) \mid \mathbf{x}\right\} \end{aligned}$$

By (1.3) again, we have $E(1(Y^o < C) \mid \mathbf{x}, Y^o) = S_C(Y^o \mid \mathbf{x})$. The last expression is seen to become

$$E\left\{\frac{1(t < Y^o)}{S^o(Y^o \mid \mathbf{x})}\,\middle|\,\mathbf{x}\right\}$$

The rest of the derivation is straightforward from the relationship between the hazard and the survival functions.

B. Proof of Lemma 3.1. To obtain the root n consistency for $\hat{\mathbf{m}}_h$ given by (3.7) using the kernel estimates (3.11)–(3.13), some regularity conditions will be imposed. Let

$$\begin{aligned} w_{ij} &= h_n^{-p}K_p\big(h^{-1}(\mathbf{x}_i - \mathbf{x}_j)\big), \\ u_{ij} &= w_{ij} - E\{w_{ij} \mid \mathbf{x}_j\}. \end{aligned}$$

We first require that the bias term of $\hat{f}(\mathbf{x}_i)$ is of the root n,

$$E\{w_{ij} \mid \mathbf{x}_j\} - f(\mathbf{x}_j) = O_p(n^{-1/2}). \tag{B.1}$$

The trade-off for imposing a smaller bias is the increasing of the variance, but by averaging out many point estimates over an interval, the variance will eventually remain small. The bias term, on the other hand, is harder to cancel out. To ensure (B1), we need to use a bandwidth smaller than the usual optimal rate. With (B.1), we can write

$$\hat{f}(\mathbf{x}_i) = f(\mathbf{x}_i) + n^{-1}\sum_k u_{ki} + O_p(n^{-1/2}). \tag{B.2}$$

The rate of convergence for the term contributing to the variance is more flexible. We need only assume that

$$n^{-1}\sum_k u_{ki} = O_p(n^{-1/4}). \tag{B.3}$$

Next we also assume that the bias for the kernel estimate of $S_Y(t \mid \mathbf{x})f(x)$ also has the root n rate

$$E\big[1(Y_k > Y_j)w_{kj} \mid \mathbf{x}_j, Y_j\big] - S_Y(Y_j \mid \mathbf{x}_j)f(\mathbf{x}_j) = O_p(n^{-1/2}). \tag{B.4}$$

Typically with suitable smoothness conditions on $S_Y(t \mid \mathbf{x})$, the same bandwidth used to achieve (B.1) may also imply (B.4). Denote

$$v_{kj} = 1(Y_k > Y_j)w_{kj} - E\big[1(Y_k > Y_j)w_{kj} \mid \mathbf{x}_j, Y_j\big].$$

Similarly to (B.3), we assume

$$n^{-1}\sum_k v_{kj} = O_p(n^{-1/4}). \tag{B.5}$$

By Taylor's expansion, we can find the leading terms for the term $\hat{S}_Y(Y_j \mid \mathbf{x}_j)^{-1}$,

$$\begin{aligned}
\hat{S}_Y(Y_j \mid \mathbf{x}_j)^{-1} &= \hat{f}(\mathbf{x}_j)\Big(S_Y(Y_j \mid \mathbf{x}_j)f(\mathbf{x}_j) + n^{-1}\sum_k v_{kj} + O_p(n^{-1/2})\Big)^{-1} \\
&= \hat{f}(\mathbf{x}_j)\Big(f(\mathbf{x}_j)^{-1}S_Y(Y_j \mid \mathbf{x}_j)^{-1} \\
&\qquad - f(\mathbf{x}_j)^{-2}S_Y(Y_j \mid \mathbf{x}_j)^{-2}n^{-1}\sum_k v_{kj} + O_p(n^{-1/2})\Big) \\
&= S_Y(Y_j \mid \mathbf{x}_j)^{-1} - f(\mathbf{x}_j)^{-1}S_Y(Y_j \mid \mathbf{x}_j)^{-2}n^{-1}\sum_k v_{kj} \\
&\quad + f(\mathbf{x}_j)^{-1}S_Y(Y_j \mid \mathbf{x}_j)^{-1}n^{-1}\sum_k u_{kj} + O_p(n^{-1/2}).
\end{aligned} \tag{B.6}$$

The last expression is obtained from (B.2) and the assumptions (B.3) and (B.5).

To proceed, let us simplify the notation by taking

$$\begin{aligned}
&f_i = f(\mathbf{x}_i), S_Y(Y_j \mid \mathbf{x}_j) = S_j, 1_{ij} = 1(Y_i < Y_j < t, \delta_j = 1), \\
&\Lambda_i = \Lambda(Y_i, t \mid \mathbf{x}_i), \hat{\Lambda}_i = \hat{\Lambda}(Y_i, t \mid \mathbf{x}_i).
\end{aligned}$$

Now apply (B.2) and (B.6) and expand the term $\hat{\Lambda}_i$:

$$\begin{aligned}
\hat{\Lambda}_i &= \hat{f}(\mathbf{x}_i)^{-1}n^{-1}\sum_j 1_{ij}\hat{S}_Y(Y_j \mid \mathbf{x}_j)^{-1}w_{ji} \\
&= f_i^{-1}n^{-1}\sum_j 1_{ij}S_j^{-1}w_{ji} \\
&\quad - f_i^{-1}n^{-1}\sum_j 1_{ij}w_{ji}f_j^{-1}S_j^{-2}n^{-1}\sum_k v_{kj}
\end{aligned} \tag{B.7}$$

$$+ f_i^{-1} n^{-1} \sum_j 1_{ij} w_{ji} f_j^{-1} S_j^{-1} n^{-1} \sum_k u_{kj}$$
$$- f_i^{-2} \left(n^{-1} \sum_k u_{ki} \right) \left(n^{-1} \sum_j 1_{ij} S_j^{-1} w_{ji} \right) + O_p(n^{-1/2}).$$

The first term will converge to the cumulative hazard Λ_i. Again, under suitable smoothness and boundedness conditions on the hazard function, the same bandwidth used before should give a bias term at the root n rate. We shall assume that

$$E\left[1_{ij} S_j^{-1} w_{ji} \mid \mathbf{x}_i, Y_i\right] - \Lambda_i f_i = O_p(n^{-1/2}). \tag{B.8}$$

Let $\varepsilon_{ij} = 1_{ij} S_j^{-1} w_{ji} - E(1_{ij} S_j^{-1} w_{ji} \mid \mathbf{x}_i, Y_i)$, and denote the second, third and fourth terms on the right side of (B.7) as $(\mathrm{I})_i$, $(\mathrm{II})_i$, $(\mathrm{III})_i$, respectively. By (B.8), we can rewrite (B.7) as

$$\hat{\Lambda}_i = \Lambda_i + f_i^{-1} n^{-1} \sum_j \varepsilon_{ij} - (\mathrm{I})_i + (\mathrm{II})_i - (\mathrm{III})_i + O_p(n^{-1/2}).$$

Denote $1_i = 1(Y_i < t, \delta_i = 0)$, $w_i = \varepsilon^{-\Lambda_i}$. We can expand the second term on the right side of equation (3.8) to

$$\begin{aligned} n^{-1} \sum_i 1_i \mathbf{x}_i \hat{w}(Y_i, t, \mathbf{x}_i) &= n^{-1} \sum_i 1_i \mathbf{x}_i \exp\left\{-\hat{\Lambda}_i\right\} \\ &= n^{-1} \sum_i \mathbf{x}_i 1_i w_i + n^{-2} \sum_{i,j} \mathbf{x}_i 1_i w_i f_i^{-1} \varepsilon_{ij} \\ &\quad - n^{-1} \sum_i \mathbf{x}_i 1_i w_i \left[(\mathrm{I})_i - (\mathrm{II})_i + (\mathrm{III})_i\right]. \end{aligned} \tag{B.9}$$

It remains to show that the second and third terms are $O_p(n^{-1/2})$.

Abbreviate the conditional expectation $E[\cdot \mid \mathbf{x}_i, Y_i]$ by $E[\cdot \mid i]$. Note that we have $E(\varepsilon_{ij}) = 0$, $E[\varepsilon_{ij} \mid i] = 0$, $\mathrm{var}(\varepsilon_{ij}) = O(h_n^{-p})$. The second term takes the form of $n^{-2} \sum_{i,j} a_i \varepsilon_{ij}$ with $a_i = \mathbf{x}_i 1_i w_i f_i^{-1}$. To evaluate its variance, we first observe that

$$\begin{aligned} E(a_i \varepsilon_{ij} a_{i'} \varepsilon_{i'j'}) &= 0 \quad \text{if } j \neq j' \\ &= O(1) \quad \text{if } j = j', i \neq i' \\ &= O(h_n^{-p}) \quad \text{if } i = i', j = j'. \end{aligned}$$

From this, a straightforward calculation leads to

$$\mathrm{var}\left(n^{-2} \sum_{i,j} a_i \varepsilon_{ij} \right) = n^{-4} \left[n^3 O(1) + n^2 O(h_n^{-p}) \right] = O(n^{-1}). \tag{B.10}$$

The variance for the third term in (B.9) can also be evaluated similarly. We can rewrite $n^{-1} \sum_i \mathbf{x}_i 1_i w_i (I)_i$ as $n^{-2} \sum_{j,k} \tilde{a}_j v_{kj}$, with

$$\tilde{a}_j = n^{-1} \sum_i \mathbf{x}_i 1_i w_i f_i^{-1} 1_{ij} f_j^{-1} S_j^{-2} w_{ji},$$

where

$$E(v_{kj} \mid j) = 0, \mathrm{var}(v_{kj}) = O(h_n^{-p}).$$

From this expression, we can calculate its variance and obtain a result similar to (B.10): $\mathrm{var}(n^{-2} \sum_{j,k} \tilde{a}_j v_{kj}) = O(n^{-1})$. The calculation for the vari-

ance of $n^{-1}\sum_i \mathbf{x}_i 1_i w_i(\mathrm{II})_i$ can be carried out in exactly the same way. Finally, to deal with the term $n^{-1}\sum_i \mathbf{x}_i 1_i w_i(\mathrm{III})_i$, we express it as $n^{-2}\sum_{i,k} \bar{a}_i u_{ki}$ with $\bar{a}_i = \mathbf{x}_i 1_i w_i f_i^{-2} n^{-1}\sum_j 1_{ij} S_j^{-1} w_{ji}$. Then again, by the same argument, the variance is shown to have the order of n^{-1}.

We have now completed the proof for the root n convergence for (3.8). The proof for (3.9) is the same. Therefore, $\hat{\mathbf{m}}_h$ is root n consistent, as claimed in Lemma 3.1. □

REFERENCES

BERAN, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, Univ. California, Berkeley.

BRILLINGER. (1991). Comment on "Sliced inverse regression for dimension reduction," by K. C. Li. *J. Amer. Statist. Assoc.* **86** 333.

CARROLL, R. J. and LI, K. C. (1992). Measurement error regression with unknown link: dimension reduction and data visualization. *J. Amer. Statist. Assoc.* **87** 1040–1050.

CARROLL, R. J. and LI, K. C. (1995). Binary regressors in dimension reduction models: a new look at treatment comparisons. *Statist. Sinica* **5** 667–688.

CHEN, C. H. and LI, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica* **8** 289–316.

COOK, R. D. (1994). On the interpretation of regression plots. *J. Amer. Statist. Assoc.* **89** 177–189.

COOK, R. D. and NACHTSHEIM, C. J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *J. Amer. Statist. Assoc.* **89** 592–599.

COOK, R. D. and WEISBERG, S. (1991). Comment on "Sliced inverse regression for dimension reduction," by K. C. Li. *J. Amer. Statist. Assoc.* **86** 328–332.

COOK, R. D. and WEISBERG, S. (1994). *An Introduction to Regression Graphics*. Wiley, New York.

DABROWSKA, D. M. (1987). Non-parametric regression with censored survival time data. *Scand. J. Statist.* **14** 181–197.

DABROWSKA, D. M. (1992). Variable bandwidth conditional Kaplan–Meier estimate. *Scand. J. Statist.* **19** 351–361.

DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** 793–815.

DOKSUM, K. A. (1987). An extension of partial likelihood methods for proportional hazard models to general transformation models. *Ann. Statist.* **15** 325–345.

DOKSUM, K. A. and GASKO, M. (1990). On a correspondence between models in binary regression analysis and in survival analysis. *Internat. Statist. Rev.* **58** 243–252.

DUAN, N. and LI, K. C. (1991). Slicing regression: a link-free regression method. *Ann. Statist.* **19** 505–530.

FAN, J. and GIJBELS, I. (1994). Censored regression: local linear approximations and their applications, *J. Amer. Statist. Assoc.* **89** 560–570.

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

HALL, P. and LI, K. C. (1993). On almost linearity of low dimensional projection from high dimensional data. *Ann. Statist.* **21** 867–889.

HSING, T. and CARROLL, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20** 1040–1061.

HUBER, P. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13** 435–526.

LI, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–342.

LI, K. C. (1992). Uncertainty analysis for mathematical models with SIR. In *Probability and Statistics* (Z. P. Jiang, S. H. Yan, P. Cheng and R. Wu, eds.) 138–162. World Scientific Press, Singapore.

LI, K. C. (1997). Nonlinear confounding in high dimensional regression. *Ann. Statist.* **25** 577–612.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd. ed. Chapman and Hall, London.

SCHOTT, J. R. (1994). Determining the dimensionality in sliced inverse regression. *J. Amer. Statist. Assoc.* **89** 141–148.

ZHU, L. X. and NG, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5** 727–736.

K.-C. LI
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA
E-MAIL: kcli@math.ucla.edu

J.-L. WANG
DIVISION OF STATISTICS
UNIVERSITY OF CALIFORNIA
DAVIS, CALIFORNIA

C.-H. CHEN
INSTITUTE OF STATISTICAL SCIENCE
ACADEMIA SINICA
TAIWAN