Biometrika (2018), **105**, 4, *pp.* 891–903
Printed in Great Britain

doi: 10.1093/biomet/asy050
Advance Access publication 17 October 2018

The change-plane Cox model

BY SUSAN WEI

School of Mathematics and Statistics, University of Melbourne, Parkville, Victoria 3010, Australia

susan.wei@unimelb.edu.au

AND MICHAEL R. KOSOROK

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

kosorok@unc.edu

SUMMARY

We propose a projection pursuit technique in survival analysis for finding lower-dimensional projections that exhibit differentiated survival outcomes. This idea is formally introduced as the change-plane Cox model, a nonregular Cox model with a change-plane in the covariate space that divides the population into two subgroups whose hazards are proportional. The proposed technique offers a potential framework for principled subgroup discovery. Estimation of the change-plane is accomplished via likelihood maximization over a data-driven sieve constructed using sliced inverse regression. Consistency of the sieve procedure for the change-plane parameters is established. In simulations the sieve estimator demonstrates better classification performance for subgroup identification than alternatives.

Some key words: Latent supervised learning; Projection pursuit; Random projection; Sieve estimation; Sliced inverse regression; Subgroup discovery.

1. INTRODUCTION

Projection pursuit, the analysis of high-dimensional data via lower-dimensional projections, is a common tool in exploratory data analysis. The idea is to search for projections that reveal interesting structure in the data. In this paper we present a projection pursuit technique in survival analysis, where a projection is considered interesting if it leads to a separation of survival outcomes. The proposed technique is based on the change-plane Cox model, set forth below.

Let (X, Z, U) be a random vector of covariates, where $X \in \mathbb{R}^p$, $Z \in \mathbb{R}^{q_1}$ and $U \in \mathbb{R}^{q_2}$. Let $\mathbb{S}^p$ be the collection of unit vectors in $\mathbb{R}^p$. The following assumptions constitute the change-plane Cox model.

Assumption 1. The hazard function of the true survival time T° has the form

$$\lambda(t \mid X, Z, U) = \exp\{\beta_1^{\mathrm{T}} Z + \beta_2 1(\omega^{\mathrm{T}} X \geqslant \gamma) + \beta_3^{\mathrm{T}} Z 1(\omega^{\mathrm{T}} X \geqslant \gamma) + \beta_4^{\mathrm{T}} U\}\lambda(t), \qquad (1)$$

where ω is an element of $\mathbb{S}^p$, γ is in some known interval $[a, b]$, $\beta = (\beta_1, \ldots, \beta_4)$ is the vector of regression parameters, with at least one of β_2 or β_3 being nonzero for identifiability, and $\lambda(t)$ is an unknown baseline hazard function.

Assumption 2. The survival time T° with hazard function (1) may be subject to right-censoring at a censoring time C which, conditional on (X, Z, U), is independent of T°;

Assumption 3. X and (Z, U) are independent.

We observe the covariate vector (X, Z, U), the censored time $T = \min(T^\circ, C)$, and the censoring indicator δ, where $\delta = 1$ if $T^\circ \leqslant C$ and $\delta = 0$ otherwise. By seeking the change-plane, given by $\omega^{\mathrm{T}} X = \gamma$, we accomplish our goal of finding a lower-dimensional projection of X that reveals two subgroups with differentiated survival outcomes.

To fix ideas, imagine X to be a set of biomarkers potentially predictive of survival, Z a categorical treatment variable, and U a set of baseline covariates such as age or gender. In this case, the regression coefficient β_3 represents the interaction effect between treatment and the subgroup indicator $1(\omega^{\mathrm{T}} X \geqslant \gamma)$. A significant β_3 is of practical interest since it would suggest the presence of treatment heterogeneity.

Rigorous assessment of β's significance is likely to be challenging in light of the results of Pons (2003). In that paper it is shown that for a certain change-point Cox model, which may be viewed as a special case of (1), the maximum partial likelihood estimator for the change-point is n-consistent but root-n-consistent for the regression coefficients. Such nonregularity can be expected in the change-plane Cox model as well. Leaving distributional theory to future work, we propose a resampling procedure in the Supplementary Material that serves as a heuristic for assessing the significance of β. However, we have not yet ascertained the rigour of this bootstrap-based assessment of significance in light of established bootstrap consistency theory.

2. METHODOLOGY

2·1. *Overview*

In this section we propose an estimation scheme for the change-plane parameters in (1) based on a sample of n independent and identically distributed replicates of (R, T, δ), where $R = (X, Z, U)$ denotes the full covariate set. The maximum partial likelihood estimator of the change-plane parameters can incur overfitting even when the dimension of X is moderately high, such as $p = 25$. This leads us to employ a regularization technique known as Grenander's method of sieves (Grenander, 1981), in which maximization takes place over an approximating subset of the parameter space called a sieve. It is desired that the sieve be dense, in a sense made rigorous in Definition 2. Interestingly, as demonstrated by Geman & Hwang (1982) in the context of nonparametric density estimation, regularization of the likelihood via the method of sieves may produce consistent estimators even when the full maximum likelihood estimator is inconsistent.

A sieve maximization scheme for fitting (1) is as follows. Collect the parameters into $\theta = (\beta, \omega, \gamma)$. The sample log partial likelihood under (1) is

$$L_n(\theta) = n^{-1} \sum_{i=1}^{n} \left(\delta_i \eta(R_i, \theta) - \delta_i \log\left[\sum_{j: T_j \geqslant T_i} n^{-1} \exp\{\eta(R_j, \theta)\} \right] \right), \tag{2}$$

where $\eta(R, \theta) = \beta_1^{\mathrm{T}} Z + \beta_2 1(\omega^{\mathrm{T}} X \geqslant \gamma) + \beta_3^{\mathrm{T}} Z 1(\omega^{\mathrm{T}} X \geqslant \gamma) + \beta_4^{\mathrm{T}} U$. The factor n^{-1} is included for consistency with the empirical process notation in § 3. Let

$$M_n(\omega, \gamma) = L_n\{\hat{\beta}_n(\omega, \gamma), \omega, \gamma\}, \tag{3}$$

where the quantity $\hat{\beta}_n(\omega, \gamma) = \arg\max_{\beta} L_n(\beta, \omega, \gamma)$ is uniquely defined and can be found via Newton's method. We shall focus on the estimation of ω since, once it is determined, the other parameters in (1) can be estimated by profiling.

Definition 1. *For a sieve $\Omega_n \subset \mathbb{S}^p$, the corresponding sieve estimator for ω in* (1) *is $\hat{\omega}(\Omega_n) = \arg\max_{\omega \in \Omega_n} M_n\{\omega, \tilde{\gamma}(\omega)\}$ where*

$$\tilde{\gamma}(\omega) = \arg\max_{\gamma \in [a,b]} M_n(\omega, \gamma). \tag{4}$$

The success of the sieve estimator hinges on the specification of the sieve. The remainder of § 2 describes the construction of a data-driven sieve.

2·2. *Initialization of the sieve*

Algorithm 1 details the construction of an initial sieve consisting of vectors that represent possible change-planes in the X covariate space. Consideration of computation time leads to the particular choices in Algorithm 1, such as the number of clusters K, chosen deliberately so that $|\Omega_0|$ is linear in n. Similarly, the discarding of clusters with fewer than four elements and the downsampling of clusters with more than ten elements are merely for computational gain. To get a sense of the size of Ω_0, Algorithm 1 applied to the simulations in § 4 results in $|\Omega_0| \approx 3000$ for sample size $n = 100$. If computation time is not a factor, better empirical performance of the overall sieve procedure, Algorithm 2, has been observed for Ω_0 in Algorithm 1 with a larger number of elements. In particular, our claim is supported by various simulations we have conducted, including the ones presented in § 4.

Algorithm 1. Initial sieve Ω_0.

```
Input  : $\{X_1, \ldots, X_n\}$
Initialize $\Omega_0$ to the empty set.
Set $K$ to $n/10$.
Partition the data $\{X_1, \ldots, X_n\}$ into $K$ clusters using $K$-means clustering.
Discard clusters with fewer than four elements.
Retain ten elements at random for clusters with more than ten elements.
foreach remaining cluster do
    foreach non-overlapping partition of the cluster into two parts $P_1$ and $P_2$ do
        Add to $\Omega_0$ the unit-length vector that connects the centroids of $P_1$ and $P_2$.
Output: $\Omega_0$
```

2·3. *Updating the sieve using sliced inverse regression*

We next update Ω_0 by incorporating survival information using sliced inverse regression (Li, 1991). This is based on a model in which a response variable S and a covariate vector X in $\mathbb{R}^p$

satisfy

$$S = f(\kappa_1^{\mathrm{T}} X, \ldots, \kappa_k^{\mathrm{T}} X, \epsilon) \tag{5}$$

for unknown constant vectors κ_j of the same dimension as X, an unknown function f, and a noise term ϵ that is independent of X. Let $\Sigma = \mathrm{cov}(X)$. If for any $b \in \mathbb{R}^p$ the conditional expectation $E(b^{\mathrm{T}} X \mid \kappa_1^{\mathrm{T}} X, \ldots, \kappa_k^{\mathrm{T}} X)$ is linear in $\kappa_1^{\mathrm{T}} X, \ldots, \kappa_k^{\mathrm{T}} X$, then for every s the centred inverse regression curve, $E(X \mid S = s) - E(X)$, lies in the span of $\{\Sigma\kappa_1, \ldots, \Sigma\kappa_k\}$. This condition on the design distribution is satisfied by X with elliptically symmetric distribution. Under this linearity condition, the space spanned by the k eigenvectors of the covariance matrix of $E(X \mid S)$ associated with the k largest eigenvalues coincides with the span of $\{\Sigma\kappa_1, \ldots, \Sigma\kappa_k\}$. The span of $\{\kappa_1, \ldots, \kappa_k\}$ itself can be obtained through standardization by Σ^{-1}. The inverse regression curve is estimated empirically by slicing the range of S into H non-overlapping intervals $I_1, \ldots, I_H$ and computing the sample version of $E(X \mid S \in I_h)$.

The subscript zero will be used to denote the true parameter value under (1). The survival time T° with hazard function (1) satisfies (5) with ω_0. Consider the following condition on the distribution of X in the change-plane Cox model.

Condition 1. For any $b \in \mathbb{R}^p$, $E(b^{\mathrm{T}} X \mid \omega_0^{\mathrm{T}} X)$ is linear in $\omega_0^{\mathrm{T}} X$.

Under Condition 1, the recovery of ω_0 in the change-plane Cox model can be accomplished via eigendecomposition of the covariance matrix of $E(X \mid T^\circ)$, followed by standardization using Σ^{-1}. To avoid issues in estimating Σ and Σ^{-1} using their sample versions, we assume that $n > p$. However, rather than slicing on T°, we slice simultaneously on T° and on $1\{\omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\}$, where $\omega \in \Omega_0$. Let $0 = t_1 < \cdots < t_H < \infty = t_{H+1}$ be a partition of the positive real line into non-overlapping intervals $I_h = [t_h, t_{h+1})$. Let $\nu(\omega)$ denote the largest-eigenvalue eigenvector of the weighted covariance matrix

$$V(\omega) = \sum_{l=0}^{1} \sum_{h=1}^{H} p_{hl}(\omega) \{m_{hl}(\omega) - E(X)\}\{m_{hl}(\omega) - E(X)\}^{\mathrm{T}}, \tag{6}$$

where

$$m_{hl}(\omega) = E[X \mid T^\circ \in I_h, 1\{\omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\} = l],$$
$$p_{hl}(\omega) = \mathrm{pr}[T^\circ \in I_h, 1\{\omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\} = l].$$

Under Condition 1, the rescaled eigenvector $\Sigma^{-1}\nu(\omega)$ coincides with the desired ω_0.

We now describe an estimate of $V(\omega)$ that accounts for censoring by employing the conditioning argument in Li et al. (1999). First, we have

$$m_{h1}(\omega) = \frac{E[X1\{T^\circ \geqslant t_h, \omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\}] - E[X1\{T^\circ \geqslant t_{h+1}, \omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\}]}{E[1\{T^\circ \geqslant t_h, \omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\}] - E[1\{T^\circ \geqslant t_{h+1}, \omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\}]},$$

which can be further decomposed as

$$\begin{aligned} &E[X1\{T^\circ \geqslant t, \omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\}] \\ &\quad = E[X1\{T \geqslant t, \omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\}] + E[X1\{T < t, \delta = 0, \omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\}\alpha(T, t, X)], \end{aligned}$$

where

$$\alpha(t', t, X) = \mathrm{pr}(T^\circ \geqslant t \mid X)/\mathrm{pr}(T^\circ \geqslant t' \mid X), \quad t' < t, \tag{7}$$

can be interpreted as a weight adjusting for the presence of censoring. This decomposition allows us to rewrite the numerator of $m_{h1}(\omega)$ as

$$\begin{aligned} &E[X1\{T^\circ \geqslant t_h, \omega^{\mathrm{T}}X \geqslant \tilde{\gamma}(\omega)\}] - E[X1\{T^\circ \geqslant t_{h+1}, \omega^{\mathrm{T}}X \geqslant \tilde{\gamma}(\omega)\}] \\ &\quad = E[X1\{t_h \leqslant T \leqslant t_{h+1}, \omega^{\mathrm{T}}X \geqslant \tilde{\gamma}(\omega)\}] + E[X1\{T < t_h, \delta = 0, \omega^{\mathrm{T}}X \geqslant \tilde{\gamma}(\omega)\}\alpha(T, t_h, X)] \\ &\qquad - E[X1\{T < t_{h+1}, \delta = 0, \omega^{\mathrm{T}}X \geqslant \tilde{\gamma}(\omega)\}\alpha(T, t_{h+1}, X)]. \end{aligned}$$

Thus we can slice on the observed survival time T rather than on T°. Let

$$\begin{aligned} \hat{c}_{i,h1}(\omega) = {} & 1\{t_h \leqslant T_i < t_{h+1}, \omega^{\mathrm{T}}X_i \geqslant \tilde{\gamma}(\omega)\} \\ & + 1\{T_i < t_h, \delta_i = 0, \omega^{\mathrm{T}}X_i \geqslant \tilde{\gamma}(\omega)\}\hat{\alpha}(T_i, t_h, X_i) \\ & - 1\{T_i < t_{h+1}, \delta_i = 0, \omega^{\mathrm{T}}X_i \geqslant \tilde{\gamma}(\omega)\}\hat{\alpha}(T_i, t_{h+1}, X_i), \end{aligned}$$

where $\hat{\alpha}(\cdot\,,\cdot\,,\cdot)$ denotes a nonparametric estimate of (7) to be discussed in § 2·4. To estimate m_{h1} and p_{h1}, we use the sample moments

$$\hat{m}_{h1}(\omega) = \sum_{i=1}^{n} X_i \hat{c}_{i,h1}(\omega) \Big/ \sum_{i=1}^{n} \hat{c}_{i,h1}(\omega)$$

and $\hat{p}_{h1}(\omega) = n^{-1}\sum_{i=1}^{n} \hat{c}_{i,h1}(\omega)$, respectively. The estimation of m_{h0} and p_{h0} is analogous. These components are incorporated into the data-driven sieve detailed in Algorithm 2. Let the resulting sieve be denoted by $\hat{\Omega}_n$. The sieve estimator associated with it will be written as $\hat{\omega}(\hat{\Omega}_n)$, following the notation introduced in Definition 1.

Algorithm 2. Data-driven sieve $\hat{\Omega}_n$ based on sliced inverse regression.

```
Input  : $(X_i, T_i, \delta_i)$ for $i = 1, \ldots, n$;
         $H$, the number of slices;
         $\Omega_0$, the initial sieve;
         $\hat{\alpha}(\cdot\,,\cdot\,,\cdot)$, censoring weight estimate.
Initialize $\hat{\Omega}_n \subset \mathbb{S}^p$ to the empty set.
Find $\hat{\Sigma}$, the empirical covariance matrix based on $X_1, \ldots, X_n$.
Set $\{t_h\}_{h=1}^{\infty}$ according to the observed range of the $T_i$ divided into $H$ equal intervals with
  $t_1 = 0$ and $t_{H+1} = \infty$.
Find $\hat{\alpha}(T_i, t_{h+1}, X_i)$ for $i = 1, \ldots, n$ and $h = 1, \ldots, H$.
foreach $\omega \in \Omega_0$ do
    Find $\hat{V}_n(\omega) = \sum_{l=0}^{1}\sum_{h=1}^{H} \hat{p}_{hl}(\omega)\{\hat{m}_{hl}(\omega) - \bar{X}\}\{\hat{m}_{hl}(\omega) - \bar{X}\}^{\mathrm{T}}$.
    Find the largest-eigenvalue eigenvector of $\hat{V}_n(\omega)$ and denote this by $\hat{v}_n(\omega)$.
    Add $\hat{\Sigma}^{-1}\hat{v}_n(\omega)$, normalized to unit length, to $\hat{\Omega}_n$.
Output: $\hat{\Omega}_n$
```

Algorithm 2 is rather insensitive to H, and we recommend setting $H = 10$. Far more critical for Algorithm 2 is the estimation of the censoring weight, the focus of the next subsection.

2·4. *Estimation of censoring weights*

The estimation of the censoring weight α in (7) reduces to that of $\text{pr}(T^\circ \geqslant t \mid X)$, the conditional survival function of T°. We shall consider two nonparametric estimates of the latter, and hence of (7) itself. The first is the nonparametric kernel estimator of Dabrowska (1987), which is described in equations (3.11)–(3.13) of Li et al. (1999) in notation similar to ours. The corresponding censoring weight estimate will also be referred to as Beran's kernel estimate.

The performance of Beran's kernel estimate quickly deteriorates as the dimension of X increases; this may be overcome by machine learning techniques. We shall employ the recursively imputed survival tree method proposed by Zhu & Kosorok (2012), a powerful, albeit complex, method for estimating the conditional survival function for censored data.

The recursively imputed survival tree combines imputation of censored observations with the idea of extremely randomized trees. Like the random forest, the extremely randomized tree selects a subset of candidate features at random. However, it does not search for the most discriminative cut-points as in the random forest, instead basing itself on random thresholds for each covariate. The imputation of censored observations enables more terminal nodes, and hence more complex trees, to be constructed. Full details of the recursively imputed survival tree algorithm are given in the Supplementary Material. We have found that the recursively imputed survival tree estimate of α leads to better performance of Algorithm 2 than does Beran's kernel estimate, as soon as the dimension of X increases beyond $p > 5$.

3. Consistency

Theorem 1 below establishes the consistency of the sieve estimator corresponding to a general sieve Ω_n under the following conditions.

Condition 2. The parameter $\theta_0 = (\beta_0, \omega_0, \gamma_0)$ lies in a compact subset $\Theta = \Theta_1 \times \Theta_2$ of $\mathbb{R}^{2q_1+q_2+1} \times \mathbb{S}^p \times [a, b]$, where Θ_1 and Θ_2 are compact subsets of $\mathbb{R}^{2q_1+q_2+1}$ and $\mathbb{S}^p \times [a, b]$, respectively.

Condition 3. The covariate X has a continuous distribution, and the projection $\omega_0^{\mathrm{T}} X$ has a strictly bounded and positive density f over $[a, b]$.

Condition 4. The probability $\text{pr}(C = 0) = 0$. There exists a $\tau \in (0, \infty)$ such that $\text{pr}(C \geqslant \tau \mid X) = \text{pr}(C = \tau \mid X) > 0$ almost surely.

Condition 5. The variables Z and U lie in bounded sets.

Conditions 2 and 3 are rather technical and simplify the proof. Condition 4 is common in survival analysis, though it is not precisely true in practice, for example in a clinical trial with staggered entry. Condition 5 is needed for an application of the dominated convergence theorem. The statement of Theorem 1 requires a definition first.

Definition 2. *A sieve $\Omega_n \subset \mathbb{S}^p$ is said to be dense for* (1) *if there exists a sequence $\omega_n \in \Omega_n$ such that $\{\omega_n, \tilde{\gamma}(\omega_n)\}$ converges to (ω_0, γ_0) as $n \to \infty$.*

Theorem 1 (Consistency of general sieve estimator). *Assume Conditions 2–5 and let $\Omega_n \subset \mathbb{S}^p$ be a dense sieve for* (1). *If $\hat{\omega}_n = \hat{\omega}(\Omega_n)$ denotes the sieve estimator, then $\{\hat{\omega}_n, \tilde{\gamma}(\hat{\omega}_n)\}$ is consistent for (ω_0, γ_0) as $n \to \infty$.*

The proof of Theorem 1 can be found in the Appendix. Corollary 1 establishes the consistency of the sieve estimator corresponding to Algorithm 2 under Condition 1 and the following meta-condition.

Condition 6. The censoring weight estimate $\hat{\alpha}$ is such that for every $\omega \in \Omega_0$, $\hat{m}_{hl}(\omega)$ is consistent for $m_{hl}(\omega)$ as $n \to \infty$ for $h = 1, \ldots, H$ and $l = 0, 1$.

Although we will limit our discussion of Condition 6 to the two estimators considered in § 2·4, its specification is left broad to allow for other possible censoring weight estimators.

For Beran's kernel estimate, the arguments in the proof of Lemma 3.1 in Li et al. (1999) can be used to verify Condition 6. The application of Lemma 3.1 requires regularity conditions labelled therein as (B.1), (B.3), (B.5) and (B.8), which mostly pertain to the relationship between the bandwidth rate and the bias and variance terms of the kernel estimate.

As for the recursively imputed survival tree estimate of α, Theorem 1 of Cui et al. (2017) addresses the consistency of estimating the underlying hazard function using a similar survival tree-based method. In both cases, a single tree is partitioned enough that the failure and censoring observations in the terminal nodes are approximately independent while maintaining a sufficient number of observations. In Theorem 1 of Cui et al. (2017), this is used to establish consistency of the resulting local Nelson–Aalen estimators for the conditional hazard estimators. For the recursively imputed survival tree, the Kaplan–Meier estimator is used instead of the Nelson–Aalen estimator.

For both Lemma 3.1 in Li et al. (1999) and Theorem 1 in Cui et al. (2017), suitable smoothness of the conditional survival function is most convenient for ascertaining the key conditions. Under Condition 3, the region where the smoothness is not met by the change-plane Cox model, i.e., the change-plane, can be bounded by a region with arbitrarily small probability.

Corollary 1 (Consistency of sieve estimator corresponding to Algorithm 2). *Let $\hat{\Omega}_n$ denote the sieve produced by Algorithm 2 for some nonempty initial sieve Ω_0. Suppose that Conditions 1–6 hold. If $\hat{\omega}_n = \hat{\omega}(\hat{\Omega}_n)$ denotes the sieve estimator, then $\{\hat{\omega}_n, \tilde{\gamma}(\hat{\omega}_n)\}$ is consistent for (ω_0, γ_0) as $n \to \infty$.*

Proof. Let $\omega \in \Omega_0$. Through conditioning,

$$\begin{aligned} m_{h1}(\omega) &= E\{X \mid T^\circ \in [t_h, t_{h+1}), \omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\} \\ &= E\{E(X \mid T^\circ) \mid T^\circ \in [t_h, t_{h+1}), \omega^{\mathrm{T}} X \geqslant \tilde{\gamma}(\omega)\}. \end{aligned}$$

A similar identity holds for m_{h0}. By Condition 1, $\nu(\omega)$, the largest-eigenvalue eigenvector of (6), is a scalar multiple of $\Sigma\omega_0$. By Condition 6, the individual components in $\hat{V}_n(\omega)$ are consistent for their theoretical counterparts. Thus, as $n \to \infty$, $\hat{V}_n(\omega)$ is consistent for $V(\omega)$, and hence the eigenvector $\hat{\nu}_n(\omega)$ is consistent for $\nu(\omega)$. Therefore the sieve $\hat{\Omega}_n$ is dense and Theorem 1 yields the desired result. □

4. Simulation study

In this section we use simulation to compare the sieve estimator with two alternatives. To focus on subgroup identification in the change-plane Cox model, we set $Z = 1$ and $U = 0$ in (1). This yields the reduced change-plane Cox model, with hazard function

$$\lambda(t \mid X) = \exp\{\beta 1(\omega^{\mathrm{T}} X \geqslant \gamma)\}\lambda(t).$$

Table 1. *Censoring mechanisms: the independent setting is so-called because censoring is independent of X; in the linear setting censoring is dependent on X only through the change-plane, whereas in the nonlinear setting censoring depends nonlinearly on X*

Name	Distribution
Independent	$C \sim \text{Un}(0, 10)$
Linear	$C \sim \min\{\text{Un}(0, 31{\cdot}97), 20\}1(\omega^{\mathrm{T}}X \geqslant \gamma) + \min\{\text{Un}(0, 3{\cdot}2), 2\}1(\omega^{\mathrm{T}}X < \gamma)$
Nonlinear	$C \sim \text{Ex}\{10^{-1}\exp(X_1 + X_2^2 + \log\lvert X_3\rvert)\}$

Un(a, b), the uniform distribution with parameters a and b; Ex(μ), the exponential distribution with mean μ.

Subgroup identification in this model can be viewed as a type of latent supervised learning (Wei & Kosorok, 2013) where the right-censored survival time plays the role of a surrogate training label.

The first alternative we consider is the double-slicing procedure proposed in Li et al. (1999), which simultaneously slices on the censored survival time and on the censoring indicator. A critical assumption is that the censoring time also satisfies a sliced inverse regression representation; that is,

$$C = g(\kappa_1^{\mathrm{T}}X, \ldots, \kappa_c^{\mathrm{T}}X, \epsilon'), \tag{8}$$

where g and ϵ' are unspecified and ϵ' is independent of X. As Li's double-slicing method does not automatically produce an estimate of γ, we obtain one by applying $\tilde{\gamma}$ in (4) to the estimated ω. A complete description of Li's double-slicing method can be found in the Supplementary Material.

The second alternative we consider is the standard survival tree implemented by means of the R (R Development Core Team, 2018) package rpart (Therneau & Atkinson, 2018). We use the rpart tree to produce a direct estimate of subgroup membership, since one cannot be obtained for the change-plane itself. This is done by thresholding the hazard rate at unity to divide the terminal nodes of the rpart tree into two subgroups. The rpart survival tree should not be confused with the recursively imputed survival tree. The latter is used in this paper solely for the estimation of α. rpart was implemented using default rather than carefully tuned parameters.

The sieve estimator corresponding to Algorithm 2 is implemented as follows. The initial sieve Ω_0 is produced using Algorithm 1 with $K = n/10$. The recursively imputed survival tree is used to estimate the conditional survival function of T° and, in turn, the censoring weight α.

The simulation set-up is as follows. We draw $n = 100$ independent and identically distributed observations (X, T, δ) from the reduced change-plane Cox model with parameters

$$\beta = \log 10, \quad \lambda(t) = 1, \quad X \sim N(0, I_p),$$
$$\omega = (\underbrace{p^{-1/2}, \ldots, p^{-1/2}}_{[p/2]}, \underbrace{-p^{-1/2}, \ldots, -p^{-1/2}}_{p-[p/2]}), \quad \gamma = 1/4$$

and one of three censoring mechanisms in Table 1. As this set-up results in exponential survival times on either side of the change-plane with all components of ω nonzero, we call it the abundant exponential simulation.

The average misclassification rate over 100 Monte Carlo simulations on a large independent test set, with sample size 10 000, of the covariate X will serve as the measure of performance.

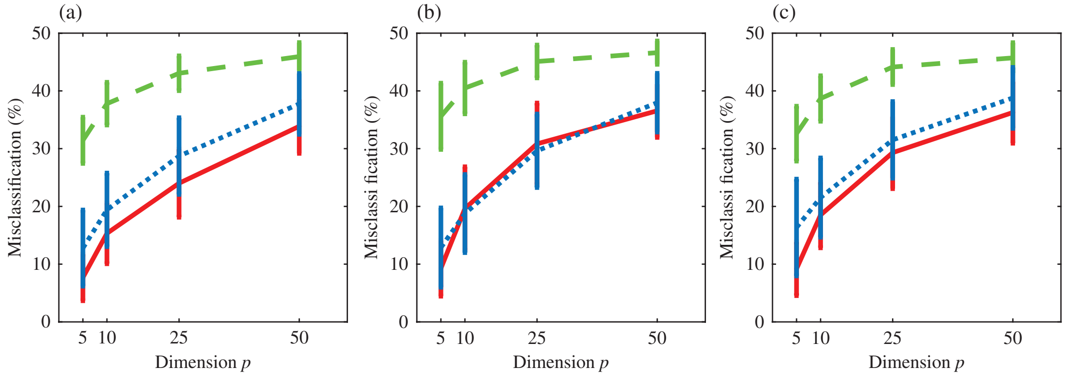

Fig. 1. Results for abundant exponential simulation in the (a) independent, (b) linear, and (c) nonlinear settings: misclassification rate over 100 Monte Carlo simulations for the sieve (solid), Li's double-slicing method (dotted), and the rpart tree (dashed), as a function of the dimension p; vertical bars indicate Monte Carlo simulation error.

Figure 1 summarizes the classification performance of the three methods as a function of the dimension p for each of the three censoring mechanisms in Table 1.

The sieve estimator performs better than Li's double-slicing procedure under the independent censoring mechanism, since there is no benefit to slicing on the censoring variable. In the linear censoring case, the two methods have similar performance, as the sieve estimator is unlikely to provide a substantial improvement when C satisfies (8). In contrast, under the nonlinear censoring mechanism, C cannot be written as a function of a linear combination of the covariates, which violates (8) in Li's double-slicing model. The sieve estimator slightly outperforms it in this case.

Figure 1 reveals that the rpart tree has difficulty across all censoring mechanisms and dimensions, probably because the geometry of the change-plane is far from that assumed in the method. When the geometry is favourable to the rpart survival tree, it can be expected to perform substantially better; see the sparse exponential simulation presented in the Supplementary Material. The rpart approach is outperformed by both the sieve estimator and Li's double-slicing for dimensions $p = 5, 10, 25$ but shows its advantages when $p = 50$. Nonetheless, survival tree methods for subgroup identification cannot produce subgroups that are contiguous in the covariate space, which may hamper interpretability in certain settings.

The abundant exponential simulation in this section and the sparse exponential simulation in the Supplementary Material both consider an idealized setting where the data are generated according to the reduced change-plane Cox model. The sieve estimator offers generally better classification performance than both Li's double-slicing and the rpart tree across a range of dimensions p and censoring mechanisms.

5. Future work

We originally envisioned the change-plane Cox model as a tool for performing subgroup discovery, which aims to identify subgroups with heterogeneous treatment responses from a very large pool of candidate subgroups (Lipkovich et al., 2017). Given its post hoc nature, subgroup discovery, and more generally subgroup analysis, is controversial (Wang et al., 2007). The change-plane Cox model may provide a principled, data-driven framework for subgroup discovery when the outcome of interest is survival. However, as the data examples in the Supplementary Material highlight, several issues must be addressed before the potential can be realized.

In the Supplementary Material, we apply the full change-plane Cox model to two datasets. The significance of β is assessed by repeatedly partitioning the data into training and test sets. Each time, only the training dataset is used to obtain an estimate of the change-plane parameters ω and γ. The significance of the regression coefficient β is then assessed in the test set, ignoring the fact that the change-plane was learned from the data. For both datasets, the resampling strategy reveals that significant β coefficients in the training data may not remain so in the test set.

Distributional theory for the parameters in the change-plane Cox model, which is currently lacking, could help identify these instances of over-optimism. For now, we recommend that any application of the proposed technique always be accompanied by the resampling strategy, which seems adequate for detecting whether the subgroups discovered are real or not. A deeper issue is the challenge that data-driven approaches pose to the standard paradigm of the scientific method. When hypotheses are generated from the data, care is needed to avoid confirmation bias.

Acknowledgement

The authors are grateful to all the reviewers, especially the associate editor, for their meticulous reading of the manuscript and their invaluable input. The authors would also like to thank Ruoqing Zhu for helpful comments on an earlier version. The second author was supported by the U.S. National Cancer Institute and the U.S. National Science Foundation.

Supplementary material

Supplementary material available at *Biometrika* online includes descriptions of the recursively imputed survival tree and Li's double-slicing method, implementation details of all methods used in the simulations and data analyses, results of the sparse exponential simulation, and analysis of two survival datasets.

Appendix

Proof of Theorem 1. Let P denote the probability measure of $W = (R, T, \delta)$ under (1). Define the empirical measure to be $\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{W_i}$, where δ_w is the measure that assigns mass 1 at w and zero elsewhere. For a measurable function f, we write $\mathbb{P}_n f = n^{-1} \sum_{i=1}^{n} f(W_i)$ and $Pf = \int f \,\mathrm{d}P$. Let $\tilde{W} = (\tilde{R}, \tilde{T}, \tilde{\delta})$ be a realization from P, independent of W. Let $\tilde{P}$ and $\tilde{\mathbb{P}}_n$ be defined analogously for $\tilde{W}$. Next, let $Y(t) = 1(T \geqslant t)$ be the at-risk process. Using empirical process notation, we can write (2) and (3) as $L_n(\theta) = \tilde{\mathbb{P}}_n \tilde{\delta}\{\eta(\tilde{R}, \theta) - \log F_n(\tilde{T}, \theta)\}$, where $F_n(t, \theta) = \mathbb{P}_n Y(t) \exp\{\eta(R, \theta)\}$, and $M_n(\omega, \gamma) = \tilde{\mathbb{P}}_n \tilde{\delta}[\eta\{\tilde{R}, \hat{\beta}_n(\omega, \gamma), \omega, \gamma\} - \log F_n\{\tilde{T}, \hat{\beta}_n(\omega, \gamma), \omega, \gamma\}]$. In the expressions for L_n and M_n, the random variables $(\tilde{R}, \tilde{T}, \tilde{\delta})$ in the first term on the right-hand side have their expectations taken with respect to $\tilde{\mathbb{P}}_n$. In the second term on the right-hand side, two successive integrations take place: first the expectation of (R, T, δ) in F_n with respect to $\mathbb{P}_n$, and then the expectation of $\tilde{T}$ with respect to $\tilde{\mathbb{P}}_n$. Let $F_0(t, \theta) = PY(t) \exp\{\eta(R, \theta)\}$. The corresponding population versions of L_n and M_n are

$$L_{\mathrm{p}}(\theta) = \tilde{P} \tilde{\delta}\{\eta(\tilde{R}, \theta) - \log F_0(\tilde{T}, \theta)\} \tag{A1}$$

and

$$M(\omega, \gamma) = \tilde{P} \tilde{\delta}\big[\eta\{\tilde{R}, \beta(\omega, \gamma), \omega, \gamma\} - \log F_0\{\tilde{T}, \beta(\omega, \gamma), \omega, \gamma\}\big]$$

where $\beta(\omega,\gamma)=\arg\max_\beta L_\mathrm{p}(\beta,\omega,\gamma)$. The subscript in L_p refers to the fact that this is a partial likelihood. Later we will use L to denote the full likelihood.

Following the argmax theorem in M-estimation theory (Kosorok, 2008, Theorem 14.1), the following conditions are sufficient to obtain consistency: (i) the sequence $\{\hat\omega_n,\tilde\gamma(\hat\omega_n)\}$ is uniformly tight; (ii) the map $(\omega,\gamma)\mapsto M(\omega,\gamma)$ is upper semicontinuous and (iii) has a unique maximum at (ω_0,γ_0); (iv) M_n converges to M uniformly over every compact set K in Θ_2; and (v) the sieve estimator nearly maximizes the objective function, i.e., $M_n\{\hat\omega_n,\tilde\gamma(\hat\omega_n)\}\geqslant M_n(\omega_0,\gamma_0)-o_P(1)$. We now check these conditions.

The first condition of the argmax theorem holds since $\|\hat\omega_n\|=1$ and $\tilde\gamma(\hat\omega_n)$ must lie in the interval $[a,b]$. For condition (ii), we will show that $M(\omega,\gamma)$ is continuous. Let (ω_n,γ_n) be a sequence converging to (ω,γ) and β_n a sequence converging to β. Then $\theta_n=(\beta_n,\omega_n,\gamma_n)$ is a sequence converging to $\theta=(\beta,\omega,\gamma)$. We first show that $\tilde P\tilde\delta\eta(\tilde R,\theta_n)\to\tilde P\tilde\delta\eta(\tilde R,\theta)$ if $\theta_n\to\theta$. This can be seen to hold componentwise for η in light of Conditions 3 and 5. We will show it explicitly for one of the components. Since X is continuous by Condition 3, we have

$$\begin{aligned}&\left|P\delta 1(\omega_n^{\mathrm T}X\geqslant\gamma_n)-\delta 1(\omega^{\mathrm T}X\geqslant\gamma)\right|\\&\leqslant P\delta\left|1(\omega_n^{\mathrm T}X\geqslant\gamma_n)-1(\omega^{\mathrm T}X\geqslant\gamma)\right|1\big(|\omega_n^{\mathrm T}X-\gamma_n-\omega^{\mathrm T}X-\gamma_0|\leqslant\epsilon\big)\\&\quad+P\delta\left|1(\omega_n^{\mathrm T}X\geqslant\gamma_n)-1(\omega^{\mathrm T}X\geqslant\gamma)\right|1\big(|\omega_n^{\mathrm T}X-\gamma_n-\omega^{\mathrm T}X-\gamma_0|>\epsilon\big)\to 0.\end{aligned}$$

If $\beta(\omega_n,\gamma_n)\to\beta(\omega,\gamma)$, then $F_0\{\tilde T,\beta(\omega_n,\gamma_n),\omega_n,\gamma_n\}\to F_0\{\tilde T,\beta(\omega,\gamma),\omega,\gamma\}$ almost surely. Note that $F_0\{\tilde T,\beta(\omega_n,\gamma_n),\omega_n,\gamma_n\}$ is bounded by an integrable function under Conditions 4 and 5. This gives $\tilde P\tilde\delta\log F_0\{\tilde T,\beta(\omega_n,\gamma_n),\omega_n,\gamma_n\}\to\tilde P\tilde\delta\log F_0\{\tilde T,\beta(\omega,\gamma),\omega,\gamma\}$. Hence, to show that $M(\omega,\gamma)$ is continuous, it suffices to establish continuity of $\beta(\omega,\gamma)$. To see this, first note that $L_\mathrm{p}(\theta)$ is continuous by the arguments above. Next, we establish that $L_\mathrm{p}(\theta)$ has a unique maximum in β for every pair (ω,γ). Consider

$$\frac{\partial}{\partial\beta}L_\mathrm{p}(\theta)=\tilde P\tilde\delta\left[\frac{\partial}{\partial\beta}\eta(\tilde R,\theta)-\frac{PY(\tilde T)\exp\{\eta(R,\theta)\}\frac{\partial}{\partial\beta}\eta(R,\theta)}{PY(\tilde T)\exp\{\eta(R,\theta)\}}\right],$$

where

$$\frac{\partial}{\partial\beta}\eta(R,\theta)=\{Z,1(\omega^{\mathrm T}X\geqslant\gamma),Z1(\omega^{\mathrm T}X\geqslant\gamma),U\}.$$

A straightforward calculation shows that the second partial derivative with respect to β is strictly negative definite. Thus $\beta(\omega_n,\gamma_n)\to\beta(\omega,\gamma)$.

We now verify condition (iii). Under (1), write the integrated hazard function of T° given X as $\exp\{\eta(R,\theta)\}\Lambda(t)$, where Λ is continuous and monotone increasing with $\Lambda(0)=0$. The joint likelihood in θ and the nuisance parameter Λ for a single observation (R,T,δ) is proportional to $L(\theta,\Lambda)\equiv\{b(R,\theta)\lambda(T)\}^\delta\exp\{-b(R,\theta)\Lambda(T)\}$ where $b(R,\theta)=\exp\{\eta(R,\theta)\}$. Next, we check (iii) by showing that the profile of L over Λ equals $L_\mathrm{p}(\theta)$ in (A1) up to a constant, which will then enable us to use the standard Kullback–Leibler argument for identifiability to show that θ_0 is a unique maximizer of (A4) and hence that (ω_0,γ_0) is a unique maximizer of $M(\omega,\gamma)$.

In L, replace $\lambda(t)$ with $\lambda_s(t)=\{1+sf(t)\}\lambda(t)$, where f is for now an unspecified bounded function, and take the Gateaux derivative of L with respect to s at $s=0$. Letting $N(t)=1(T\leqslant t,\delta=1)$ be the counting process and using the fact that $P\,\mathrm dN(t)=PY(t)b(R,\theta_0)\,\mathrm d\Lambda_0(t)$, we obtain that the expectation of the resulting derivative is

$$\int_0^\tau f(t)P\{Y(t)b(R,\theta_0)\}\,\mathrm d\Lambda_0(t)-\int_0^\tau f(t)P\{Y(t)b(R,\theta)\}\,\mathrm d\Lambda(t).\tag{A2}$$

Now, if we replace Λ in (A2) with $\Lambda_s(t)=\int_0^t\{1+sg(u)\}\,\mathrm d\Lambda(u)$ for some other function g and differentiate again with respect to s at $s=0$, we obtain that the second Gateaux derivative is

$-\int_0^\tau f(t)g(t)P\{Y(t)b(R,\theta)\}\,\mathrm{d}\Lambda(t)$, which is strictly negative when $f = g$, implying that for fixed θ, any Λ which is a zero of (A2) for a rich enough collection of functions f is a maximizer over all Λ for fixed θ. Insert $f(t) = 1(t \leqslant u)$ into (A2), and allow u to range over $[0,\tau]$; then we obtain that the profile maximizer of L over Λ satisfies $\int_0^u P\{Y(t)b(R,\theta_0)\}\,\mathrm{d}\Lambda_0(t) - \int_0^u P\{Y(t)b(R,\theta)\}\,\mathrm{d}\Lambda(t) = 0$ for all $u \in [0,\tau]$. Hence

$$\frac{\mathrm{d}\Lambda(t)}{\mathrm{d}\Lambda_0(t)} = \frac{P\{Y(t)b(R,\theta_0)\}}{P\{Y(t)b(R,\theta)\}}. \tag{A3}$$

Inserting (A3) back into L and removing additive terms which are constants with respect to θ, we obtain that the profile of L over the parameter Λ is

$$P\left(\int_0^\tau \log b(R,\theta)\,\mathrm{d}N(t) - \int_0^\tau \log[P\{Y(t)b(R,\theta)\}]\,\mathrm{d}N(t)\right), \tag{A4}$$

which equals $L_{\mathrm{p}}(\theta)$ in (A1). Now let θ_1 maximize (A4). Then, as (A4) is the profile of L over the parameter Λ, there exists a Λ_1 such that the joint parameter (θ_1, Λ_1) maximizes L. By the property of the Kullback–Leibler discrepancy and model identifiability, this implies that $\theta_1 = \theta_0$. Hence (A4) has a unique maximizer at θ_0 and we have shown that $M(\omega,\gamma)$ is uniquely maximized at (ω_0,γ_0).

To verify condition (iv) of the argmax theorem, fix a compact set $K = K_1 \times K_2 \subset \Theta$, where K_1 is compact in Θ_1 and K_2 is compact in Θ_2. Let $m_\theta(v,t,\delta) = \delta\{\eta(v,\theta) - \log F_n(t,\theta)\}$ and consider the class of functions $\{m_\theta(v,t,\delta) : \theta \in K\}$. First we consider the component $\{\eta(v,\theta) : \theta \in K\}$. The classes $\{\beta_i\}$ for $i = 1,\ldots,4$ are each Donsker, as are the classes $\{Z\}$ and $\{U\}$. The class $\{1(\omega^{\mathrm{T}}x \geqslant \gamma) : (\omega,\gamma) \in K_2\}$ is also Donsker by the example in Kosorok (2008, § 4.1.1). Since products of bounded Donsker classes are Donsker, $\{\eta(v,\theta) : \theta \in K\}$ is Donsker. Next, we examine the component $\{\log F_n(t,\theta) : t \in [0,\tau], \theta \in K\}$. The class $[\exp\{\beta_2 1(\omega^{\mathrm{T}}x \geqslant \gamma)\}]$ is Donsker, since exponentiation is Lipschitz continuous on compact sets. The at-risk process $Y(t)$ is Donsker by Lemma 4.1 in Kosorok (2008). Thus $\{\log F_n(t,\theta)\}$ is Donsker. Repeating arguments for sums of Donsker classes and products of bounded Donsker classes shows that $\{m_\theta(v,t,\delta) : \theta \in K\}$ is a Donsker class of functions and therefore also a Glivenko–Cantelli class of functions.

Now let $m_{\omega,\gamma}(v,t,\delta) = \delta[\eta\{v,\hat\beta_n(\omega,\gamma),\omega,\gamma\} - \log F_n(t,\hat\beta_n(\omega,\gamma),\omega,\gamma)]$; then we can write $M_n(\omega,\gamma) = \tilde{\mathbb{P}}_n m_{\omega,\gamma}(\tilde R,\tilde T,\tilde\delta)$. Since the estimated log ratio hazard $\hat\beta_n(\omega,\gamma)$ lies in a compact set in Θ_1 for all $(\omega,\gamma) \in K_2$, the class $\{m_{\omega,\gamma}(v,t,\delta) : (\omega,\gamma) \in K_2\}$ is contained in a Donsker class, which implies that it is a Glivenko–Cantelli class. Thus

$$\sup_{(\omega,\gamma)\in K_2} \left|M_n(\omega,\gamma) - \tilde P m_{\omega,\gamma}(\tilde R,\tilde T,\tilde\delta)\right| \to 0$$

in probability as $n \to \infty$. Next we show that $\tilde P m_{\omega,\gamma}(\tilde R,\tilde T,\tilde\delta)$ converges uniformly to $M(\omega,\gamma)$. The uniform convergence of $\hat\beta_n(\omega,\gamma)$ to $\beta(\omega,\gamma)$ can be shown by adapting the arguments of Theorem 1 in Pons (2003). We then show that $F_n\{t,\hat\beta_n(\omega,\gamma),\omega,\gamma\} \to F_0\{t,\beta(\omega,\gamma),\omega,\gamma\}$ uniformly over $(\omega,\gamma) \in K_2$. We may write $F_n\{t,\hat\beta_n(\omega,\gamma),\omega,\gamma\} = \mathbb{P}_n Y(t)\exp[\eta\{R,\hat\beta_n(\omega,\gamma),\omega,\gamma\}]$ and $F_0\{t,\beta(\omega,\gamma),\omega,\gamma\} = PY(t)\exp[\eta\{R,\beta(\omega,\gamma),\omega,\gamma\}]$. We have already argued for the Donsker property of the classes $\{1(t \geqslant r) : r \in [0,\tau]\}$ and $\{\exp\{\eta(v,\theta)\} : \theta \in K\}$. Thus we conclude that $\{1(t \geqslant r)\exp\{\eta(v,\theta)\} : r \in [0,\tau], \theta \in K\}$ is Donsker and hence Glivenko–Cantelli. Therefore, $M_n(\omega,\gamma)$ converges uniformly to $M(\omega,\gamma)$ over compact $K_2 \subset \Theta_2$.

Finally, we check condition (v) of the argmax theorem. If the sieve Ω_n is dense, there is a sequence $\{\omega_n, \tilde\gamma(\omega_n)\} \in \Omega_n \times [a,b]$ that converges to (ω_0,γ_0). By definition, $M_n\{\hat\omega_n, \tilde\gamma(\hat\omega_n)\} \geqslant M_n\{\omega_n,\tilde\gamma(\omega_n)\}$. By the continuity of $M(\omega,\gamma)$, $M_n(\omega_0,\gamma_0) - M_n\{\omega_n,\tilde\gamma(\omega_n)\} = o_P(1)$ and hence $M_n\{\hat\omega_n,\tilde\gamma(\hat\omega_n)\} \geqslant M_n(\omega_0,\gamma_0) - o_P(1)$. All the conditions of the argmax theorem are met, and so consistency follows. □

References

CUI, Y., ZHU, R. & KOSOROK, M. (2017). Tree based weighted learning for estimating individualized treatment rules with censored data. *Electron. J. Statist.* **11**, 3927–53.

DABROWSKA, D. M. (1987). Non-parametric regression with censored survival time data. *Scand. J. Statist.* **14**, 181–97.

GEMAN, S. & HWANG, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10**, 401–14.

GRENANDER, U. (1981). *Abstract Inference*. New York: Wiley.

KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

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

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

LIPKOVICH, I., DMITRIENKO, A. & D'AGOSTINO, R. B. (2017). Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Statist. Med.* **36**, 136–96.

PONS, O. (2003). Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Ann. Statist.* **31**, 442–63.

R DEVELOPMENT CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

THERNEAU, T. M. & ATKINSON, E. J. (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13.

WANG, R., LAGAKOS, S. W., WARE, J. H., HUNTER, D. J. & DRAZEN, J. M. (2007). Statistics in medicine–reporting of subgroup analyses in clinical trials. *New Engl. J. Med.* **357**, 2189–94.

WEI, S. & KOSOROK, M. R. (2013). Latent supervised learning. *J. Am. Statist. Assoc.* **108**, 957–70.

ZHU, R. & KOSOROK, M. R. (2012). Recursively imputed survival trees. *J. Am. Statist. Assoc.* **107**, 331–40.

[*Received on* 10 *August* 2014. *Editorial decision on* 23 *May* 2018]