*Bernoulli* **10**(6), 2004, 1011–1037

# Asymptotically optimal model selection method with right censored outcomes

SÜNDÜZ KELEŞ*, MARK VAN DER LAAN** and SANDRINE DUDOIT†

*Division of Biostatistics, University of California, Berkeley CA 94720, USA. E-mail: *sunduz@stat.berkeley.edu; **laan@stat.berkeley.edu; †sandrine@stat.berkeley.edu*

Over the last two decades, nonparametric and semi-parametric approaches that adapt well-known techniques such as regression methods to the analysis of right censored data, e.g. right censored survival data, have become popular in the statistics literature. However, the problem of choosing the best model (predictor) among a set of proposed models in the right censored data setting has received little attention. We develop a new cross-validation based-model selection method to select among predictors of right censored outcomes such as survival times. The proposed method considers the risk of a given predictor based on the training sample as a parameter of the full data distribution in a right censored data model. Then, the doubly robust locally efficient estimation method or an *ad hoc* inverse probability of censoring weighting method, as presented by Robins and Rotnitzky and later by van der Laan and Robins, is used to estimate this conditional risk parameter based on the validation sample. We prove that, under general conditions, the proposed cross-validated selector is asymptotically equivalent to an oracle benchmark selector based on the true data generating distribution. The method presented covers model selection with right censored data in prediction (univariate and multivariate) and density/hazard estimation problems.

*Keywords:* cross-validation; density/hazard estimator selection; model selection; multivariate prediction; nonparametric/semi-parametric regression; prediction of survival; right censored data; univariate prediction

## 1. Introduction

Recent developments in the field of functional genomics have attracted much attention to nonparametric and semi-parametric regression methods in statistics. The availability of thousands of genomes and the microarray technology that allow measurement of the gene expression of thousands of genes simultaneously generate different types of high-dimensional data (e.g. typically $\sim 30\,000$ gene expression profiles in humans). Modelling of such high-dimensional data with complex interactions among variables in combination with clinical outcomes is an important and challenging task. There has been a tremendous increase in the number of microarray gene expression studies that aim to link or correlate expression of genes with the survival of patients or time to recurrence of various types of cancers (Alizadeh *et al.* 2000; Garber *et al*. 2001; Sorliea *et al*. 2001; Beer *et al*. 2002; Wigle *et al*. 2002). In these studies the parameter of interest is typically mean or median survival of patients (or time to recurrence of cancer or some other event) given the expression of thousands of genes. This recent genomic literature indicates that analysis of

1350–7265 

such high-dimensional data structures by investigating the relationship between the gene expression profiles and other phenotypes such as survival times is very important in clinical applications.

Another type of high-dimensional data structure of functional genomics that requires the attention of nonparametric and semi-parametric regression methods is the single nucleotide polymorphism (SNP) data sets. Recently, Ruczinski *et al*. (2001) developed a method called *logic regression* that is entirely motivated by the binary structure of SNP data. This method essentially constructs predictive boolean expressions for SNPs (which can equivalently be represented as a tree structure) for a given phenotype outcome. In SNP data sets, one question of interest is, again, the linking of SNPs to quantitative phenotypes such as survival of patients. Besides being high-dimensional, another challenge of these data sets is that the outcome of interest such as survival is often right censored. Motivated by these recent developments in functional genomics, we turn our attention to model selection (and also predictor performance assessment) with right censored data that has applications in a variety of models such as the Cox proportional hazards model and survival trees. Given an algorithm searching through sets of variables (covariates) and corresponding models, e.g. regression models, we need a criterion for selecting among the corresponding predictors, and a measure of performance for any given predictor.

Nonparametric and semi-parametric regression methods are among the most popular methods for analysing right censored data. Recently, nonparametric alternatives to the Cox proportional hazards model have gained importance. These methods adapt well-known regression techniques to the analysis of censored survival data. In general, the available methods propose a sequence of regression models of increasing complexity, typically determined by a stepwise feature selection method. This sequence of models is referred to as a *sieve*. The final choice of model is determined with a particular model selection criterion. Although the actual regression methodology for analysing right censored data has been extensively studied, the problem of selecting the best model or predictor among a sieve has not gained much attention.

Our goal in this paper is to develop a model selection method to select among predictors of right censored outcomes in the context of prediction and density/hazard estimation problems. In particular, we propose a model selection criterion that generalizes the resampling-based cross-validated risk criterion of a given loss function used with uncensored data to censored data. This criterion uses an *inverse probability of censoring weighted* (IPCW) risk estimator. The consistency of this estimator relies on consistent estimation of the censoring mechanism and the condition that the uncensored data structure (where all data are observed) has a positive probability of being completely observed. This, in particular, rules out the type I censoring scheme which prevents the regions of the time axis being observable no matter what the sample size is, i.e. fixed (known) censoring times.

The method that we describe here is applicable to two different tasks: (i) selecting among a set of predictors (models); and (ii) assessings the performance of a given predictor. In the next two subsections, we first give a brief overview of the available nonparametric and semi-parametric regression methods for censored survival data (or, more broadly, for right censored outcomes that measure time to occurrence of an event) and the model selection methods used by them (task (i)). We then review some of the less technical literature on

assessing the performance of a given predictor in the context of the Cox proportional hazards model and survival trees (task (ii)).

## 1.1. Model selection in the context of nonparametric and semi-parametric regression with right censored data

Some of the most commonly used regression methods for censored survival data are based on splines or partitioning trees. In particular, Hastie and Tibshirani (1990a; 1990b) use additive Cox proportional hazards models that model covariate effects with smoothing splines. Kooperberg *et al*. (1995) follow a polynomial spline approach and propose a sieve of multiplicative intensity models for the hazard of survival which allows interaction effects between covariates and with time. In these approaches, model selection techniques such as Akaike's information criterion (AIC: Akaike 1973; Bozdogan 2000) and the Bayesian information criterion (BIC: Schwartz 1978) are used to adaptively select the best model. Several extensions of classification and regression trees (CART: Breiman *et al*. 1984) have also been proposed for censored survival data. These are sometimes referred to as *survival trees* and can roughly be divided into two categories. Methods in the first category use a *within-node homogeneity measure*. Examples of such approaches are provided in Gordon and Olshen (1985), Davis and Anderson (1989) and Leblanc and Crowley (1992). Davis and Anderson (1989), for example, use the negative log-likelihood of an exponential model for each node as a measure of split homogeneity and the squared difference of the parent node log-likelihood and a weighted sum of child node log-likelihoods as the split function. Methods in the second category, first proposed by Segal (1988), use a *between-node homogeneity measure* and a split function based on the two-sample log-rank test statistic.

In essence, these methods replace the least-squares split functions utilized by CART in the uncensored continuous outcome setting with suitable alternatives to deal with right censored outcomes, i.e. to evaluate risk of a given predictor based on censored data. In summary, the available spline-based regression methods for censored survival data use AIC or BIC or variants to model selection, whereas the tree-based methods replace the least-squares split criterion with an alternative split criterion that is entirely specific to censored data.

## 1.2. Predictor performance assessment with right censored survival data

In the prediction and model selection problems with uncensored data, resampling-based risk estimators that are obtained by $V$-fold cross-validation or Monte Carlo cross-validation (repeated sample splitting) are commonly used (Breiman *et al*. 1984; Burman 1989; Shao 1993; Zhang 1993). The performance of a given predictor with uncensored outcome is assessed by estimating its risk with respect to a user-supplied loss function with the empirical mean of the corresponding loss function over an (independent) validation sample. Contrary to the prediction and model selection literature for uncensored outcomes, we

observe that, in general, the literature on assessing the performance of predictors with right censored data is sparse.

There is a range of *ad hoc* approaches (Korn and Simon 1990; O'Quigley and Xu 2001; Schemper and Henderson 2000) that address the problem of quantifying the predictive accuracy of a survival model or assessing its performance (reviewed in Schemper and Stare 1996). In general, these approaches attempt to provide an analogue of the multiple correlation coefficient $R^2$ of linear regression models for survival models such as the Cox proportional hazards model and survival trees. Furthermore, they do not aim to measure the performance of a predictor on future data points, i.e. they do not intend to measure the *generalization error*.

In particular, Korn and Simon (1990) define a measure of explained variation as the proportional reduction in risk obtained by using the model for prediction over the null model in survival time models. The authors specifically consider the limited data setting where the explanatory variables (covariates) are fixed and propose to derive the expected loss from estimated model-based survival curves. This model-based approach gives inconsistent estimates of the expected loss if the survival model is misspecified; moreover, it is not suited to comparing different models. Schemper and Henderson (2000) suggest a measure of predictive accuracy and explained variation based on the mean absolute distance between the survival process (indicator function of observed survival time) and its predictor in the Cox proportional hazards model. In particular, an observed data loss function is used and its consistent estimation under non-informative censoring is considered. The proposed estimator is inconsistent under dependent censoring, and whether it can be used for model selection under independent censoring is not discussed. Similarly O'Quigley and Xu (2001) suggest a measure of explained variation specifically for the Cox proportional hazards model utilizing Schoenfeld residuals. As mentioned at the beginning of this paragraph, none of these approaches measures the performance of a given predictor based on independent observations (future observations) and they are not suited to model selection. They are mainly intended to provide additional summary information to complement a fitted survival model.

Our approach of assessing the performance of a predictor in a censored survival model differs from all of the above methods. We propose a method that is suitable for model selection and is capable of predictor accuracy assessment based on future observations. Moreover, this approach does not depend on the model assumed for survival times (i.e. the form of the regression model) and it handles dependent censoring. Graft *et al*. (1999) propose an interesting observed data loss function based on the so called *Brier score* to measure the performance of a given predictor. To accommodate independent censoring, a weighted version of the original loss function, which results in an IPCW loss function, is used. Their method is similar in spirit to ours since it can be used to measure the performance of a predictor based on observations that are not used to build the predictor and it relies on reweighting of the full data loss function. The approach that we present here is much more general since it treats the risk of a given predictor as a full data parameter and considers natural ways to estimate it based on the observed data. Furthermore, it handles dependent censoring and extends over model selection and predictor performance

assessment in general right censored data settings. More importantly, we provide theoretical results showing the asymptotic optimality of our method under general conditions.

### 1.3. Outline of the paper

In this paper, we propose an asymptotically optimal method for model/predictor selection with right censored outcomes. This, in particular, includes prediction problems and density/hazard estimation problems with right censored outcomes. In this method, we treat the risk of a given predictor based on the training sample as a parameter of the full data distribution in a censored data model. Subsequently, we utilize doubly robust locally efficient estimation methods (Robins and Rotnitzky 1992; 2001; Robins *et al*. 2000; van der Laan and Robins 2003) for estimating this parameter based on the validation sample. The proposed model selection method also handles informative censoring, and the performance of a given predictor can be assessed consistently even when the censoring mechanism is informative. The organization of the paper is as follows. Section 2 introduces the new methodology and uses an IPCW risk estimator that is consistent under informative censoring provided that the censoring mechanism is estimated consistently. In Section 3 we present our main theorem, which states that under general conditions the proposed method is asymptotically optimal and the cross-validation selector performs asymptotically as well as the benchmark selector based on the true data-generating distribution. Section 4 presents a simulation study illustrating this result. In Section 5 we provide a detailed proof of our result, together with the lemmas required in the proof. Section 6 is devoted to a discussion and, in particular, focuses on a doubly robust risk estimator obtained by the general doubly robust locally efficient estimation methodology.

## 2. Model selection with right censored data

### 2.1. Data structure and model

Let $T$ denote the survival time, and define the random variable $\bar{X}(T) = \{X(s) : 0 \leqslant s \leqslant T\}$ where $X(s)$ is a multivariate stochastic process evolving in time. $X(s)$ includes, in particular, $R(s) = I(T \leqslant s)$ and the covariate process $L(s)$ evolving in time. The random variable $X \equiv \bar{X}(T)$, which we refer to as the *full data random variable*, stands for everything that can be observed on a randomly selected subject in the interval $(0, T]$ if the subject is not subject to censoring. We denote the distribution of the full data by $F$ and the general class of statistical models to which $F$ belongs by $\mathcal{M}^F$. Let $W = L(0)$ denote a $p$-dimensional vector of time-independent baseline covariates. One goal in this setting might be to build a predictor of log-survival $Z = \log T$ based on the time-independent covariates $W$, which could then be used to predict the survival time of a new incoming subject. Such a goal is not restricted to survival outcomes; in other words, $Z$ could be any function of the full data. Some examples of the parameter of interest $\psi_0$ are the conditional mean, $\mathrm{E}[Z|W]$,

or the conditional median, median$(Z|W)$, of the outcome $Z$ or the marginal or conditional density of the survival time $T$, i.e. $f(T)$ or $f(T|W)$.

In real-life applications, we often do not observe the full data but its censored version. Let $C$ denote the censoring time and let $A(t) = I(C < t)$ denote the censoring process. We will represent the *observed data random variable* with $Y = (C, \Delta = I(T \leqslant C), \overline{X}(C))$. The distribution of the observed data $Y$ is indexed by the full data distribution $F$ and the conditional distribution $G(\cdot|X)$ of the censoring variable $C$ given the full data $X$. We will denote this observed data distribution by $P$. We refer to $G(\cdot|X)$ as the censoring mechanism and sometimes simply denote it by $G$. By convention we have $C = \infty$ (hence $\Delta = 1$) if $T$ occurs before right censoring. The conditional hazard of the censoring mechanism $A(t)$ given the full data $X$ will be denoted by $\lambda_C(t|X) = \mathrm{E}(\mathrm{d}A(t)|\overline{A}(t-) = 0, X)$. We assume that $C$ is either discrete or continuous with respect to Lebesgue measure, that is $\lambda_C(\cdot|X)$ is either a probability or an intensity as in Andersen *et al*. (1993).

We assume *coarsening at random* (CAR) on the censoring mechanism. For right censored data structures, this assumption is

$$\text{CAR: } \lambda_C(t|X) = \lambda_C(t|\overline{X}(t)).$$

Coarsening at random was originally formulated by Heitjan and Rubin (1991) and further generalized by Jacobsen and Keiding (1995) and Gill *et al*. (1997). We refer to Robins and Rotnitzky (1992) and Robins (1993) for the introduction and discussion of this CAR definition for the right censored data structure.

## 2.2. The parameter of interest

We will keep our discussion of the parameter of interest quite general but will provide specific examples. In our general framework the parameter of interest is a parameter of the full data distribution. Formally, the parameter of interest, $\psi_0 = \psi(\cdot|F)$, is a mapping from the Euclidean space into the real line. For example, in the regression context the parameter of interest might be the conditional mean of the log of the survival time given the baseline covariates, $\mathrm{E}[Z|W]$, or the conditional median of the log of the survival time given the baseline covariates, median$[Z|W]$. In these two examples the Euclidean space captures the baseline covariates. Alternatively, in the context of density estimation the parameter of interest might be the conditional density $f(T|W)$ of the survival time given the baseline covariates (or the hazard given the baseline covariates), or just the marginal density $f(T)$ of the survival time. We will denote the parameter space of our parameter of interest by $\Psi = \{\psi(\cdot|F) : F \in \mathcal{M}^F\}$. Next, we define the parameter of interest $\psi_0$ in terms of a *loss function*, $L(X, \psi)$. The performance of a given predictor $\psi$ is usually quantified with a loss function and its risk is defined as the expected loss. Our parameter of interest is one of the minimizers of the risk or expected loss of a specified loss function. We have that

$$\psi_0 = \underset{\psi \in \Psi}{\operatorname{argmin}} \, \mathrm{E}_F[L(X, \psi)],$$

where the choice of the loss function defines a parameter of interest (possibly non-unique). Here, $\mathrm{E}_F[L(X, \psi)]$ is the risk of the candidate predictor $\psi$. Note that the parameter of

interest minimizes the expectation of a particular loss function of a candidate parameter value. When we plug in the true parameter value $\psi_0$, i.e. the optimal predictor, as the candidate predictor, we obtain the *optimal risk* $\theta_{\text{opt}}$:

$$\theta_{\text{opt}} = \int L(X, \psi_0) \mathrm{d}F(X).$$

Next we provide examples of parameters of interest.

## 2.3. Examples of full data loss functions and parameter of interests

We will consider parameters of interest from prediction and density/hazard estimation problems.

### *2.3.1. Univariate prediction*

In this setting, one is interested in predicting a univariate outcome $Z$, such as the log of the survival time, $Z = \log T$, or the indicator function of the survival time at a given time point $t$, $Z = I(T > t)$, based on a vector of baseline covariates $W$. The parameter of interest in nonparametric and semi-parametric regression approaches is typically the conditional mean, $\psi_0 = \mathrm{E}[Z|W]$, or the conditional median, $\psi_0 = \text{median}(Z|W)$. $\mathrm{E}[Z|W]$ corresponds to the $L_2$ quadratic (squared error) loss function $L(X, \psi) = (Z - \psi(W))^2$, and $\text{median}(Z|W)$ corresponds to the $L_1$ absolute error loss function $L(X, \psi) = |Z - \psi(W)|$.

### *2.3.2. Multivariate prediction*

In this setting, one is concerned with $l$ outcomes of interest, i.e. $Z$ is an $L$-dimensional vector. One parameter of interest in this data structure is the conditional mean vector $\psi_0 = \mathrm{E}[Z|W] = (\mathrm{E}[Z_1|W], \ldots, \mathrm{E}[Z_r|W])$. For a candidate multivariate predictor $\psi(W)$, the following loss function can be defined:

$$L(Z, \psi) = (Z - \psi(W))^{\mathrm{T}} \eta(W)(Z - \psi(W)).$$

Here, $\eta$ is a symmetric $l \times l$ matrix function of $W$. A natural choice for $\eta(W)$ is the inverse of the conditional covariance matrix of the outcome vector $Z$ given the baseline covariates $W$, i.e.

$$\eta(W) = [\mathrm{E}(\{Z - \mathrm{E}[Z|W]\}\{Z - \mathrm{E}[Z|W]\}^{\mathrm{T}})]^{-1}.$$

This type of loss function aims to take into account the dependence structure among the components of the response vector $Z$.

### *2.3.3. Density or hazard estimation*

In this setting, the parameter of interest might be the conditional density of survival time $T$ given the baseline covariates, $\psi_0 = f(T|W)$; the hazard at a time point $t_0$, $\psi_0 =$

$f(t_0|W, T \geqslant t_0) \equiv P(T > t_0|W, T \geqslant t_0)$; or just the marginal density of $T$, $\psi_0 = f(T)$. In this case, corresponding full data loss function is the negative log-likelihood loss function $L(X, \psi) = -\log \psi(T, W)$.

## 2.4. Risk estimation with the full data

As emphasized in Section 1, we may be interested in evaluating the risk of a predictor for at least two purposes: (i) model or predictor selection, where the best predictor is chosen to minimize risk over a given class of predictors; (ii) predictor performance assessment, where the generalization error of the selected predictor is assessed. In the literature it is common practice to use cross-validation for these purposes. There are various cross-validation schemes: in general terms the idea is to divide the data into a training and a validation set and estimate the predictor parameters on the training set while estimating the conditional risk of the predictor based on the validation set. We refer to this risk as the conditional risk since it belongs to a predictor based on a particular training set. Next, we will outline the steps that lead to performing cross-validation with censored outcomes.

## 2.5. Risk estimation with the observed data

Given an empirical distribution $P_n$ based on an independent and identically distributed sample $\{Y_i, i = 1, \ldots, n\}$ of size $n$, let $\psi_k(\cdot|P_n)$, $k \in \{1, \ldots, K(n)\}$, be well-defined estimators, e.g. predictors based on different models, of the parameter of interest $\psi_0$. We will not discuss the methods available to obtain such estimators in different full data models in this paper. We refer the reader to Chapter 3 of van der Laan and Robins (2003) for a presentation of doubly robust locally efficient estimation of the full data parameters in generalized linear models and multiplicative intensity models of the full data. Some examples of sequences of predictor estimators in such full data models are as follows. Consider a linear regression model for $\log T$. Then the $\psi_k$ can be a sequence of fitted models based on different subsets of $W$. Similarly, if a Cox proportional hazards model is used to model survival times, a sequence of models can be obtained through various subsets of $W$. Our interest here is in selecting a $\hat{k}$ among $\{1, \ldots, K(n)\}$ such that $\psi_{\hat{k}}(\cdot|P_n)$ converges to $\psi_0$ in an optimal manner.

The fundamental idea behind our method is to consider the risk of a given predictor as a full data parameter of interest and apply the general methods developed by Robins and Rotnitzky (1992; 2001), Robins *et al*. (2000) and van der Laan and Robins (2003) for estimating it consistently and efficiently using the observed data. This is a crucial step since full data loss functions such as quadratic loss, $L(X, \psi) = (Z - \psi(W))^2$, cannot be evaluated for an observation with censored survival time ($\Delta = 0$), and hence risk estimators based on only uncensored observations, such as $n^{-1}\sum_i L(X_i, \psi(W_i))\Delta_i$, are biased for the expected loss $\mathrm{E}_F[L(X, \psi)]$ of the predictor $\psi$. We now consider an IPCW estimator of the risk of a given predictor. The IPCW estimating function was introduced by Robins and Rotnitzky (1992) and is widely used in censored data models (van der Laan and Robins 2003), for

handling missing covariates (Pugh *et al.* 1993) and when modelling survey data (Binder 1992).

Given a predictor $\psi$, we define the observed data function

$$IC_0[Y|G, L(\cdot, \psi)] = \frac{\Delta L(X, \psi)}{\bar{G}(T|X)}. \tag{1}$$

This is an IPCW estimating function, i.e. a full data loss function weighted by the inverse of the censoring probability times the censoring indicator. Note that this estimating function basically maps the full data (uncensored) loss function into an observed data loss function that has the same expected value, i.e.

$$\begin{aligned} \mathrm{E}_Y[IC_0[Y|G, L(\cdot, \psi)]] &= \mathrm{E}_X \mathrm{E}_Y[\Delta L(X, \psi)/\bar{G}(T|X)|X] \\ &= \mathrm{E}_X[L(X, \psi)]. \end{aligned} \tag{2}$$

We will now describe the formal notation that we use to define cross-validated risk estimator. Let $S_n \in \{0, 1\}^n$ be a random vector independent of $P_n$ with a finite number $V$ of support points. For technical reasons, let the probability on each support point of $S_n$ be bounded away from zero uniformly in $n$. A realization of $S_n$ defines a particular split of the sample of $n$ observations into a training sample $\{i \in \{1, \ldots, n\} : S_{n,i} = 0\}$ and a validation sample $\{i \in \{1, \ldots, n\} : S_{n,i} = 1\}$. In particular, $V$-fold cross-validation corresponds to a specific $S_n$. Let $P^0_{n,S_n}$, $P^1_{n,S_n}$ denote the empirical distributions of the training and the validation sample, respectively. Let the proportion $p(n) \equiv p = (1/n)\sum_{i=1}^n S_{n,i} \in (0, 1)$ of observations in the validation sample be constant.

Our proposed risk estimator utilizes the correspondence between the given full data loss function and its IPCW version as elucidated in equation (2). Let $\bar{G}(\cdot|X) = 1 - G(\cdot|X)$ be the survival of the censoring conditional on the full data $X$ and let $\bar{G}^0_{n,S_n}(\cdot|X)$ be an estimator of $\bar{G}(\cdot|X)$ based on the training sample. We define the following cross-validated risk estimate:

$$\begin{aligned} \hat{\theta}_{n(1-p)}(k) &= \mathrm{E}_{S_n}\int IC_0[Y|G^0_{n,S_n}, L(\cdot, \psi_k(\cdot|P^0_{n,S_n}))]\mathrm{d}P^1_{n,S_n}(Y), \\ &= \mathrm{E}_{S_n}\left[\frac{1}{np}\sum_{i=1}^n I(S_{n,i} = 1)\frac{\Delta_i}{\bar{G}^0_{n,S_n}(T_i|X)} L(X_i, \psi_k(\cdot|P^0_{n,S_n}))\right]. \end{aligned}$$

Here, the expectation indexed by $S_n$ corresponds to taking the empirical mean of the risk estimate $\int IC_0[Y|G^0_{n,S_n}, L(\cdot, \psi_k(\cdot|P^0_{n,S_n}))]$ of a given predictor $\psi_k(\cdot|P^0_{n,S_n})$ over validation samples corresponding to different $S_n$. This risk estimate defines an optimal choice $\hat{k}$ of $k$ given by

$$\hat{k} = \min^{-1}_{k\in\{1,\ldots,K(n)\}} \hat{\theta}_{n(1-p)}(k).$$

### 2.6. Benchmark for $\hat{k}$

A natural way to benchmark this choice of predictor index $\hat{k}$ is by defining the true conditional risk function

$$\tilde{\theta}_{n(1-p)}(k) = \mathrm{E}_{S_n}\int L(X, \psi_k(\cdot|P^0_{n,S_n}))\mathrm{d}F(X),$$

which equals the true conditional risk of the predictor in the full data model. We note that

$$\tilde{\theta}_{n(1-p)}(k) = \mathrm{E}_{S_n}\int IC_0[Y|G, L(\cdot, \psi_k(\cdot|P^0_{n,S_n}))]\mathrm{d}P(Y) = \mathrm{E}_{S_n}\int L(X, \psi_k(\cdot|P^0_{n,S_n}))\mathrm{d}F(X)$$

by the double expectation property. The minimizer

$$\tilde{k} = \min^{-1}_{k\in\{1,\ldots,K(n)\}} \tilde{\theta}_{n(1-p)}(k)$$

of the true conditional risk function for a given $P_n$ indexes the best predictor, among the predictors $\{\psi_k(\cdot|P_{n(1-p)}),\ k \in \{1, \ldots, K(n)\}\}$ based on $n(1-p)$ observations, that achieves the optimal conditional risk. Hence, $\tilde{k}$ defines a best choice or benchmark for $\hat{k}$. In practice, we do not observe the true conditional risk since it depends also on the true observed data distribution $P$. Consequently, we do not have $\tilde{k}$ available to us.

Finally, note that the optimal risk $\theta_{\mathrm{opt}} \equiv \mathrm{E}[L(X, \psi_0)]$ can also be represented as

$$\theta_{\mathrm{opt}} = \int IC_0[Y|G, L(\cdot, \psi_0)]\mathrm{d}P(Y).$$

It is crucial to establish how the performance of $\hat{k}$ obtained by the above resampling-based cross-validation method in estimating the optimal risk compares with the performance of the minimizer $\tilde{k}$ of the true conditional risk. In Theorem 1 of the next section, we prove that $\hat{k}$ performs asymptotically as well as the benchmark selector $\tilde{k}$ in the sense that

$$\frac{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\mathrm{opt}}}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\mathrm{opt}}} \to 1 \qquad \text{in probability as } n \to \infty,$$

under general conditions.

## 3. Asymptotic optimality result

Before we present our main theorem stating the optimality result, we wish to elaborate on the main assumptions of the theorem. Assumption (A.1) requires the supremum of the difference between the full data loss function evaluated at a candidate estimator $\psi_k$ and at the true parameter value $\psi_0$ to be bounded away from infinity over a support of the observed data distribution. This assumption is easily satisfied if the outcome variable, e.g. $Z = \log T$, is bounded provided that the predictor estimators are bounded in supremum norm by the same bound as the data (this can be ensured by truncation). Assumption (A.2) specifies loss functions whose expectations can be estimated at a quadratic rate. These loss

functions will be referred to here as *quadratic loss functions*. This assumption may not be satisfied for all kinds of loss functions, but many loss functions used in practice, such as the $L_2$ squared error loss function and the negative log-likelihood loss function, do satisfy it. Our theorem also addresses *general loss functions* where (A.2) does not hold, e.g. $L_1$ absolute error loss function, and establishes the same optimality result under slightly different conditions. Assumption (A.3) is an identifiability assumption required for the IPCW risk estimator used in our method. It requires possible realizations of the full data $X$ to have a positive probability of being observed, hence excludes the type I censoring mechanisms. Assumption (A.5) roughly requires the survival estimator of the censoring mechanism to converge to its true value at a rate faster than the rate at which the true conditional risk converges to the optimal risk.

Throughout the following theorem the following conventions apply. Firstly, let $Z_n$ and $Y_n$ be two random variables and define $Z_n = O_P(Y_n)$ as $\limsup_{n\to\infty} P_{Z_n,Y_n}(|Z_n| > mY_n) \leqslant \epsilon_m$, where $\epsilon_m \to 0$ as $m \to \infty$. Secondly, we define $O_{P^-}(Z_n)$ as a term that is equivalent to $O_P(C(n)Z_n)$ for any deterministic sequence $C(n)$ converging to infinity at an arbitrary slow rate. Moreover, the supremum in assumption (A.1) below is taken over a support of the observed data distribution $P$.

We now present our main theorem.

**Theorem 1.** *Suppose that the following assumptions are valid:*

(A.1) $\sup_X L(X, \psi_k(\cdot|P^0_{n,S_n})) - L(X, \psi_0) \leqslant M_1$, *for all* $k$, *almost surely, for some* $M_1 < \infty$.

(A.2) $\int [L(X, \psi_k(\cdot|P^0_{n,S_n})) - L(X, \psi_0)]^2 \mathrm{d}F(X) \leqslant M_2 \int [L(X, \psi_k(\cdot|P^0_{n,S_n})) - L(X, \psi_0)] \mathrm{d}F(X)$ *for all* $k$, *for some* $M_2 < \infty$.

(A.3) $\overline{G}(T|X) > \delta > 0$, $F$-*almost everywhere, for some* $\delta > 0$.

(A.4) $\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}} = O_P(1/R(n))$ *for some deterministic sequence* $R(n)$.

(A.5) $\sqrt{\mathrm{E}_{S_n} \int (\overline{G}^0_{n,S_n} - \overline{G})^2 (T|X) \mathrm{d}F(X)} = O_P(\max(\log K(n)/(np)^{0.5}, 1/R(n)^{0.5}(np)^{\gamma}))$, *for some* $\gamma > 0$.

*Then*

$$\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}} \leqslant \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}} + O_{P^-}(H(n)), \tag{3}$$

*where*

$$H(n) = \max\left( \frac{\log^{1.5} K(n)}{R(n)^{0.25}(np)^{0.75}}, \frac{\log^2 K(n)}{(np)}, \frac{\log K(n)}{R(n)^{0.5}(np)^{0.5}}, \frac{\log^{0.5} K(n)}{R(n)^{0.75}(np)^{0.25}(np)^{\gamma}}, \frac{1}{R(n)(np)^{\gamma}} \right).$$

*If*

$$\frac{\log K(n)}{(np)^{0.5}} = O\left( \frac{1}{R(n)^{0.5}(np)^{\alpha}} \right), \tag{4}$$

*for some* $\alpha > 0$, *then*

$$H(n) = \frac{1}{R(n)(np)^{\min(\alpha,\gamma)}}.$$

*Thus, if also*

$$\frac{1}{R(n)(np)^{\min(\alpha,\gamma)}\,(\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}})} = o_P(1), \tag{5}$$

*then*

$$\frac{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}}}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}}} \to 1 \qquad \textit{in probability for } n \to \infty. \tag{6}$$

*If* (A.2) *does not hold, then*

$$H(n) = \max\left(\frac{\log K(n)}{(np)^{0.5}}, \frac{1}{R(n)^{0.5}(np)^{\gamma}}\right) \tag{7}$$

*and if*

$$\frac{\log K(n)}{(np)^{0.5}} = O\left(\frac{1}{R(n)^{0.5}(np)^{\alpha}}\right), \tag{8}$$

*for some* $\alpha > 0$*, then*

$$H(n) = \frac{1}{R(n)^{0.5}(np)^{\min(\alpha,\gamma)}}.$$

*Thus, if also*

$$\frac{1}{R(n)^{0.5}(np)^{\min(\alpha,\gamma)}(\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}})} = o_P(1), \tag{9}$$

*then* (6) *holds.*

For a fixed $p$, if $K(n)$ converges to infinity with $n$ at a polynomial rate and $R(n)/\sqrt{n} \to 0$ and the conditions (A.3) and (A.5) on $\overline{G}(\cdot|X)$ hold, then condition (4) of the theorem will hold for loss functions whose optimal risk $\theta_{\text{opt}}$ can be estimated at a quadratic rate, i.e. when (A.2) holds. Moreover, if also $R(n)$ is chosen to be the actual rate of convergence for $(\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}})$ then condition (5) is satisfied, and thus the optimality result (6) holds.

Theorem 1 establishes the optimality of $\hat{k}$ in selecting the best choice among the predictors $\{\psi_k(\cdot|P_{n(1-p)}) : k \in \{1, \ldots, K(n)\}\}$ that are based on $n(1-p)$ observations. Here, $P_{n(1-p)}$ denotes the empirical distribution of $Y$ based on $n(1-p)$ observations. The following corollary proves that if $p(n) \to 0$ and the result (6) of Theorem 1 holds then the proposed model choice of $\hat{k}$ is also optimal for selecting a predictor among the predictors $\{\psi_k(\cdot|P_n) : k \in \{1, \ldots, K(n)\}\}$ that are based on the whole sample. We use the notation $p(n) = p_n$ in this corollary.

**Corollary 1.** *Let*

$$\tilde{k}_{n(1-p)} = \min_{k\in\{1,\ldots,K(n)\}}{}^{-1} \mathrm{E}_{S_n}\int L(X, \psi_k(\cdot|P^0_{n(1-p),S_n}))\mathrm{d}F(X),$$

*denote the previously defined* $\tilde{k}$. *Let* $\hat{k}$ *be defined as previously. Define*

$$\tilde{k}_n = \min_{k\in\{1,\ldots,K(n)\}}{}^{-1}\int L(X, \psi_k(\cdot|P_n))\mathrm{d}F(X).$$

*If* $p_n \to 0$, *the assumptions of Theorem 1 hold so that* (6) *holds, and*

$$\frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{\mathrm{opt}}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{\mathrm{opt}}} \to 1 \qquad \textit{in probability}, \tag{10}$$

*then*

$$\frac{\tilde{\theta}_{n(1-p_n)}(\hat{k}) - \theta_{\mathrm{opt}}}{\tilde{\theta}_n(\tilde{k}_n) - \theta_{\mathrm{opt}}} \to 1 \qquad \textit{in probability}. \tag{11}$$

*A sufficient condition for* (10) *to hold is that*

$$(n^\gamma(\tilde{\theta}_n(\tilde{k}_n) - \theta_{\mathrm{opt}}), (n(1-p_n))^\gamma(\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{\mathrm{opt}})) \overset{D}{\Rightarrow} (Z, Z)$$

*for some* $\gamma > 0$ *and random variable* $Z$ *with* $\Pr(Z > a) = 1$ *for some* $a > 0$. *In particular, if* $\Pr(S_n = s) = 1$ *for some* $s \in \{0, 1\}^n$ *(i.e. single-split cross-validation), then it suffices to assume* $n^\gamma\left(\tilde{\theta}_n(\tilde{k}_n) - \theta_{\mathrm{opt}}\right) \overset{D}{\Rightarrow} Z$ *for some* $\gamma > 0$ *and* $\Pr(Z > a) = 1$ *for some* $a > 0$.

We expect the latter sufficient condition in this corollary to be sufficient for any general $S_n$ as well.

# 4. Simulation study

To illustrate the asymptotic optimality result of Theorem 1 we conducted a simulation study where $\psi_k(\cdot|P_n)$ are univariate histogram regressions indexed by the binwidth $1/k$. The full data structure in this simulation is $(T_i, W_i)$, $i = 1, \ldots, n$, where $W \sim U(0, 1)$ and the survival time $T$ is generated from the model $Z \equiv \log T = W^2 + \epsilon$. For example, if $k = 5$, then $\psi_k(\cdot|P_n)$ represents a histogram regression predictor with five bins of width 0.2. The distribution of $\epsilon$ is chosen to be a truncated normal distribution with a compact support in the interval $[-10, 10]$ and $\sigma^2 = 2$. We obtain $\theta_{\mathrm{opt}} = 2$ by using the numerical integration function `integrate` of the statistical software R. The censoring times are generated from a uniform distribution as follows: $\log C \sim \mathcal{U}(l_1, 12)$ and $l_1$ is set to $-2$ and $-5$ to generate 15% and 30% censoring. This censoring distribution ensures that $P(C > T|X) > \delta > 0$, $F$-a.e. Specifically, $\delta$ approximately equals 0.07 with $l_1 = -2$ and 0.06 with $l_1 = -5$.

Define

$$B^k_j(P^0_{n,S_n}) = \{i : S_{n,i} = 0,\ W_i \in j\text{th bin of the } k\text{-bin histogram regression}\},$$

as the set of observations in the $j$th bin of the $k$-bin histogram regression. We have the

following IPCW predictor $\psi_k(\cdot|P^0_{n,S_n})$ of log survival for an observation with a covariate value $w$ in the $j$th bin of the $k$-bin histogram regression:

$$\psi_k(w|P^0_{n,S_n}) = \frac{1}{|B^k_j(P^0_{n,S_n})|} \sum_{i \in B^k_j(P^0_{n,S_n})} \frac{(\log T_i)\Delta}{\bar{G}^0_{n,S_n}(T_i|W)},$$

where $\bar{G}^0_{n,S_n}(\cdot|W)$ is obtained by fitting a Cox proportional hazards model of the form $P(C = t|W = w, C \geqslant t) = \lambda_0(t)\exp(\beta_0 + \beta_1 w)$ with $(T_i, W_i, 1 - \Delta_i)$, $i \in \{1, \ldots, n\}$, ignoring independent censoring. Here, $\lambda_0(\cdot)$ represents the unspecified baseline hazard of censoring. We used 100 different bin sizes, resulting in 100 different predictors. The true conditional risk functions $\tilde{\theta}_{n(1-p)}(k)$ and $\tilde{\theta}_n(k)$ for each $k \in \{1, \ldots, 100\}$ are evaluated using an analytical formula. This formula is easy to derive and is given in Keleş (2003, Appendix D). We set the proportion of the validation sample $p(n)$ to 0.2 (fivefold cross-validation) and simulated from five different sample sizes $n \in \{100, 200, 500, 2000, 10\,000\}$. At each sample size and censoring proportion 50 data sets were generated. Simulation results illustrating Theorem 1 are summarized in Table 1.

This simulation study illustrates the convergence to unity of the ratio of the true conditional risk difference of the cross-validated selector to the true conditional risk difference of the benchmark selector. However, as observed in Table 1, this convergence might be slow in practice. One issue that we have not emphasized in these simulations is the choice of the proportion of the validation sample $p(n)$, i.e. the value of $V$. Keleş (2003) extends the simulations presented here to other choices of $p(n)$ and illustrates that the results are not sensitive to the choice of $p(n)$.

## 5. Proof of Theorem 1 and Corollary 1

In this section we present the proofs for our main theorem (Theorem 1) and its corollary (Corollary 1). We will specifically provide the proof of Theorem 1 for quadratic loss functions (assumption (A.2) holds). The proof of the theorem for general loss functions not satisfying (A.2) requires the analogue of the lemmas we establish in this section. When

**Table 1.** *Simulation study.* Each column reports the average ratio of $(\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}})/(\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}})$ over 50 data sets for censoring proportions 0%, 15%, and 30%, respectively

| $n$ | 0% | 15% | 30% |
|---|---|---|---|
| 100 | 4.479 037 | 3.336 332 | 3.209 060 |
| 200 | 2.377 004 | 2.449 200 | 2.724 465 |
| 500 | 1.829 994 | 1.989 088 | 2.319 114 |
| 2 000 | 1.593 553 | 1.812 296 | 1.553 712 |
| 10 000 | 1.553 950 | 1.470 549 | 1.467 048 |

(A.2) does not hold, the proof is more straightforward. We refer to our technical report (Keleş *et al*. 2003) for the details of the proof in the case of general loss functions.

***Proof of Corollary 1.*** Firstly, note that

$$\frac{\tilde{\theta}_{n(1-p_n)}(\hat{k}) - \theta_{\text{opt}}}{\tilde{\theta}_n(\tilde{k}_n) - \theta_{\text{opt}}} \frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{\text{opt}}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{\text{opt}}} \to 1$$

by Theorem 1. This proves the first statement of the corollary (11). We now show that (10) holds under the given sufficient condition. Define

$$Z_{1,n} = n^{\gamma}(\tilde{\theta}_n(\tilde{k}_n) - \theta_{\text{opt}}), \tag{12}$$

$$Z_{2,n} = (n(1-p_n))^{\gamma}(\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{\text{opt}}). \tag{13}$$

If $(Z_{1,n}, Z_{2,n}) \overset{D}{\Rightarrow} (Z, Z)$ then, by the continuous mapping theorem, we have $Z_{1,n}/Z_{2,n} \to 1$. However, note that

$$\frac{Z_{1,n}}{Z_{2,n}} = \frac{1}{(1-p_n)^{\gamma}} \frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{\text{opt}}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{\text{opt}}}.$$

Thus, if $p_n \to 0$, then we have

$$\frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{\text{opt}}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{\text{opt}}} \to 1,$$

and thus (10) holds. If there is only one split, i.e. $P(S_n = s) = 1$ for some $s$, then $Z_{1,n} = Z_{2,n/(1-p_n)}$ (i.e., $n$ in the definition (13) is replaced by $n/(1-p_n)$), and hence $Z_{1,n} \overset{D}{\Rightarrow} Z$ implies $(Z_{1,n}, Z_{2,n}) \overset{D}{\Rightarrow} (Z, Z)$. This completes the proof. □

***Proof of Theorem 1.*** We will sometimes use the following shorthand notation in this proof:

$$IC_0[Y|G^0_{n,S_n} - G, L(\cdot, \psi_k(\cdot|P^0_{n,S_n}))] \equiv$$
$$IC_0[Y|G^0_{n,S_n}, L(\cdot, \psi_k(\cdot|P^0_{n,S_n}))] - IC_0[Y|G, L(\cdot, \psi_k(\cdot|P^0_{n,S_n}))].$$

Note that $IC_0[\cdot|G, L(\cdot, \psi_k(\cdot|P^0_{n,S_n}))]$ is linear in $L(\cdot, \psi_k(\cdot|P^0_{n,S_n}))$ and hence $IC_0[\cdot|G, L(\cdot, \psi_k(\cdot|P^0_{n,S_n}))] - IC_0[\cdot|G, L(\cdot, \psi_0)]$ equals $IC_0[\cdot|G, L(\cdot, \psi_k(\cdot|P^0_{n,S_n})) - L(\cdot, \psi_0)]$.

We have

$$
\begin{aligned}
0 &\leqslant \mathrm{E}_{S_n}\int IC_0[Y|G,\, L(\cdot,\, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot,\, \psi_{\tilde{k}}(\cdot|P^0_{n,Sn}))]\mathrm{d}P(Y) \\
&= \mathrm{E}_{S_n}\int IC_0[Y|G,\, L(\cdot,\, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot,\, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n}))]\mathrm{d}P^1_{n,S_n}(Y) \\
&\quad + \mathrm{E}_{S_n}\int IC_0[Y|G,\, L(\cdot,\, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot,\, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n}))]\mathrm{d}(P - P^1_{n,S_n})(Y) \\
&= \underbrace{\mathrm{E}_{S_n}\int IC_0[Y|G^0_{n,S_n}, L(\cdot, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n}))]\mathrm{d}P^1_{n,S_n}(Y)}_{\leqslant 0} \\
&\quad + \mathrm{E}_{S_n}\int IC_0[Y|G - G^0_{n,S_n},\, L(\cdot,\, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot,\, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n}))]\mathrm{d}P^1_{n,S_n}(Y) \\
&\quad + \mathrm{E}_{S_n}\int IC_0[Y|G,\, L(\cdot,\, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot,\, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n}))]\mathrm{d}(P - P^1_{n,S_n})(Y) \\
&\leqslant \mathrm{E}_{S_n}\int IC_0[Y|G - G^0_{n,S_n},\, L(\cdot,\, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot,\, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n}))]\mathrm{d}P^1_{n,S_n}(Y) \\
&\quad + \mathrm{E}_{S_n}\int IC_0[Y|G,\, L(\cdot,\, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot,\, \psi_{\tilde{k}})]\mathrm{d}(P - P^1_{n,S_n})(Y) \\
&= -\mathrm{E}_{S_n}\int IC_0[Y|G^0_{n,S_n},\, L(\cdot,\, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot,\, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n}))]\mathrm{d}P^1_{n,S_n}(Y) \\
&\quad + \mathrm{E}_{S_n}\int IC_0[Y|G,\, L(\cdot,\, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot,\, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n}))]\mathrm{d}P(Y) \\
&= -\mathrm{E}_{S_n}\underbrace{\int IC_0[Y|G^0_{n,S_n} - G, L(\cdot, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n}))]\mathrm{d}P(Y)}_{R_{1,n}(S_n)} \\
&\quad - \mathrm{E}_{S_n}\underbrace{\int IC_0[Y|G^0_{n,S_n}, L(\cdot, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n}))]\mathrm{d}(P^1_{n,S_n} - P)(Y)}_{R_{2,n}(S_n)} \\
&= -\mathrm{E}_{S_n}R_{1,n}(S_n) - \mathrm{E}_{S_n}R_{2,n}(S_n).
\end{aligned}
$$

Thus, we have

$$0 \leqslant \tilde{\theta}_{n(1-p)}(\hat{k}) - \tilde{\theta}_{n(1-p)}(\tilde{k}) \leqslant -\mathrm{E}_{S_n}R_{1,n}(S_n) - \mathrm{E}_{S_n}R_{2,n}(S_n),$$

which implies

$$0 \leqslant \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\mathrm{opt}} \leqslant \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\mathrm{opt}} - \mathrm{E}_{S_n}R_{1,n}(S_n) - \mathrm{E}_{S_n}R_{2,n}(S_n). \tag{14}$$

Using the linearity of $IC_0[Y|G, L(\cdot, \psi_k)]$ in $L(\cdot, \psi_k)$, $R_{1,n}$ and $R_{2,n}$ can be decomposed as:

$$\begin{aligned}
R_{1,n}(S_n) &= \underbrace{\int IC_0[Y|G^0_{n,S_n} - G, L(\cdot, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot, \psi_0)]\mathrm{d}P(Y)}_{R_{1,n}(S_n,\hat{k})} \\
&\quad - \underbrace{\int IC_0[Y|G^0_{n,S_n} - G, L(\cdot, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n})) - L(\cdot, \psi_0))\mathrm{d}P(Y)}_{R_{1,n}(S_n,\tilde{k})} \\
&\equiv R_{1,n}(S_n, \hat{k}) - R_{1,n}(S_n, \tilde{k}), \\
R_{2,n}(S_n) &= \underbrace{\int IC_0[Y|G^0_{n,S_n}, L(\cdot, \psi_{\hat{k}}(\cdot|P^0_{n,S_n})) - L(\cdot, \psi_0)]\mathrm{d}(P^1_{n,S_n} - P)(Y)}_{R_{2,n}(S_n,\hat{k})} \\
&\quad - \underbrace{\int IC_0[Y|G^0_{n,S_n}, L(\cdot, \psi_{\tilde{k}}(\cdot|P^0_{n,S_n})) - L(\cdot, \psi_0)]\mathrm{d}(P^1_{n,S_n} - P)(Y)}_{R_{2,n}(S_n,\tilde{k})} \\
&\equiv R_{2,n}(S_n, \hat{k}) - R_{2,n}(S_n, \tilde{k}).
\end{aligned}$$

We will analyse the terms of $R_{1,n}$ and $R_{2,n}$ separately. In particular, we use two general lemmas based on Bernstein's inequality (van der Vaart and Wellner 1996) to bound the $R_{2,n}$ terms in probability. The $R_{1,n}$ terms are bounded in probability using the Cauchy–Schwarz inequality. Specifically, using the lemmas given later in this proof, we obtain

$$\begin{aligned}
\mathrm{E}_{S_n} R_{1,n}(S_n, \tilde{k}) &= O_P\left(\sqrt{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\mathrm{opt}}}\sqrt{\mathrm{E}_{S_n}\int(\overline{G}^0_{n,S_n} - \overline{G})^2(T|X)\mathrm{d}F(X)}\right), \\
\mathrm{E}_{S_n} R_{1,n}(S_n, \hat{k}) &= O_P\left(\sqrt{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\mathrm{opt}}}\sqrt{\mathrm{E}_{S_n}\int(\overline{G}^0_{n,S_n} - \overline{G})^2(T|X)\mathrm{d}F(X)}\right), \\
\mathrm{E}_{S_n} R_{2,n}(S_n, \tilde{k}) &= O_P\left(\max\left(\frac{\log K(n)}{R(n)^{0.5^-}(np)^{0.5}}, \frac{\log K(n)}{np}\right)\right), \\
\mathrm{E}_{S_n} R_{2,n}(S_n, \hat{k}) &= O_P\left(\max\left(\frac{a(n)\log K(n)}{(np)^{0.5}}, \frac{\log K(n)}{np}\right)\right),
\end{aligned}$$

where $a(n)$ is any deterministic sequence such that $\Pr(\|\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\mathrm{opt}}\| \leqslant a(n)) \to 1$ and $R(n)^{0.5^-}$ refers to a deterministic sequences that is slower than $R(n)$, i.e. $R(n)^{0.5-\lambda}$ for an arbitrarily small $\lambda$.

Substituting these upper bounds into inequality (14) yields

$$\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}} \leqslant \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}} - \mathrm{E}_{S_n} R_{1,n}(S_n) - \mathrm{E}_{S_n} R_{2,n}(S_n)$$

$$\leqslant \tilde{\theta}_n(\tilde{k}) - \theta_{\text{opt}} + O_P\left(\frac{a(n)\log K(n)}{(np)^{0.5}}\right) + O_P\left(\frac{\log K(n)}{R(n)^{0.5^-}(np)^{0.5}}\right) + O_P\left(\frac{\log K(n)}{np}\right)$$

$$+ O_P\left(\sqrt{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}}}\sqrt{\mathrm{E}_{S_n}\int(\bar{G}^0_{n,S_n} - \bar{G})^2(T|X)\mathrm{d}F(X)}\right)$$

$$+ O_P\left(\sqrt{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}}}\sqrt{\mathrm{E}_{S_n}\int(\bar{G}^0_{n,S_n} - \bar{G})^2(T|X)\mathrm{d}F(X)}\right).$$

Because $\mathrm{E}_{S_n} R_{1,n}(S_n, \hat{k})$ dominates $\mathrm{E}_{S_n} R_{1,n}(S_n, \tilde{k})$, this inequality reduces to

$$\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}} \leqslant \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}} + O_P\left(\frac{a(n)\log K(n)}{(np)^{0.5}}\right) + O_P\left(\frac{\log K(n)}{R(n)^{0.5^-}(np)^{0.5}}\right)$$

$$+ O_P\left(\frac{\log K(n)}{np}\right) + O_P\left(\sqrt{\tilde{\theta}_n(\hat{k}) - \theta_{\text{opt}}}\sqrt{\mathrm{E}_{S_n}\int(\bar{G}^0_{n,S_n} - \bar{G})^2(T|X)\mathrm{d}F(X)}\right).$$

We now use assumption (A.5) to simplify the last term, giving

$$\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}} \leqslant \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}} + O_P\left(\frac{a(n)\log K(n)}{(np)^{0.5}}\right)$$

$$+ O_P\left(\frac{\log K(n)}{R(n)^{0.5^-}(np)^{0.5}}\right) + O_P\left(\frac{\log K(n)}{np}\right) + O_P\left(\frac{a(n)}{R(n)^{0.5}(np)^{\gamma}}\right). \quad (15)$$

Since $a(n)$ is the width of an interval containing $\tilde{\theta}_{n(1-p)}(k) - \theta_{\text{opt}}$ with probability tending to 1, this probabilistic inequality implies a convergence in probability result for $\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}}$. We solve this inequality iteratively, and the limit of this iterative process is given by

$$\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}} \leqslant \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}} + O_{P^-}(H(n)),$$

where $H(n)$ is defined in Theorem 1.

Specifically, the details of this iterative process are as follows. Define

$$c_1(n) = \max\left(\frac{1}{R(n)}, \frac{\log K(n)}{R(n)^{0.5^-}(np)^{0.5}}, \frac{\log K(n)}{np}\right),$$

$$c_2(n) = \max\left(\frac{\log K(n)}{(np)^{0.5}}, \frac{1}{R(n)^{0.5}(np)^{\gamma}}\right).$$

Then (15) implies that

$$\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}} = O_P(c_1(n)) + O_P(a(n)c_2(n)). \quad (16)$$

*Initialize*. Set $a(n) = 1$ so that (16) implies

$$\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}} = O_P(\max(c_1(n), c_2(n))).$$

A. If $\max(c_1(n), c_2(n)) = c_1(n)$, set $a(n) = c_1(n)^{0.5^-}$ and stop iterating. Then (16) becomes

$$\tilde{\theta}_n(\hat{k}) - \theta_{\text{opt}} = O_P(c_1(n)) + O_P(c_1(n)^{0.5^-}(n)c_2(n)). \tag{17}$$

B. If $\max(c_1(n), c_2(n)) = c_2(n)$, set $a(n) = c_2(n)^{0.5^-}$ and go to the iteration step.

*Iteration step*. This step updates $a(n)$ and iterates (16).

A. If $\max(c_1(n), a(n)c_2(n)) = c_1(n)$, set $a(n) = c_1(n)^{0.5^-}$ and stop iterating. As a result we obtain (17).

B. If $\max(c_1(n), a(n)c_2(n) = a(n)c_2(n)$, set $a_{\text{new}}(n) = a(n)^{0.5^-}c_2(n)^{0.5^-}$, and set $a(n) = a_{\text{new}}(n)$. Then repeat the iteration step with this $a(n)$.

We note that $a(n)$ converges to $c_2(n)^{1^-}$ in branch B, while it converges to $c_1(n)^{0.5^-}$ in branch A. Substituting $a(n) = \max(c_1(n)^{0.5^-}, c_2(n)^{1^-})$ in (15) and after some trivial simplification, we obtain

$$\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\text{opt}} \leqslant \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\text{opt}} + O_{P^-}(H(n)),$$

where $H(n)$ is defined in Theorem 1. This proves the first statement (3) in the theorem. We note that if

$$\frac{\log K(n)}{(np)^{0.5}} = O\left(\frac{1}{R(n)^{0.5}(np)^{\alpha}}\right),$$

for some $\alpha > 0$, then

$$H(n) = \frac{1}{R(n)(np)^{\min(\alpha,\gamma)}}.$$

Finally, if (5) holds, then we obtain the desired result (6). This completes the proof. □

We now present lemmas dealing with the analysis of the $R_{1,n}$ and $R_{2,n}$ terms. The following general lemma is obtained from Bernstein's inequality (van der Vaart and Wellner 1996) and is used in the analysis of $R_{2,n}(S_n, \tilde{k})$ term.

**Lemma 1.** *Let* $Z_{n,i}$, $i = 1, \ldots, n$, *be* $n$ *independent mean-zero random variables with variance* $\text{var}(Z_{n,i}) \leqslant \sigma_n^2$ *and* $\Pr(\max_i |Z_{n,i}| \leqslant W_n) = 1$ *for* $W_n < \infty$. *If* $W_n/\sqrt{n}\sigma_n = O(1)$, *then*

$$\frac{\sqrt{n}}{\sigma_n}\left|\frac{1}{n}\sum_{i=1}^{n} Z_{n,i}\right| = O_P(1).$$

*Specifically,*

$$\Pr\left(\frac{\sqrt{n}}{\sigma_n}\left|\frac{1}{n}\sum_{i=1}^{n} Z_{n,i}\right| > x\right) \leqslant 2\exp\left(-\frac{1}{2}\frac{x^2}{1 + W_n x/3\sqrt{n}\sigma_n}\right).$$

***Proof.*** By Bernstein's inequality we have

$$\Pr\left(\left|\sum_{i=1}^{n} Z_{n,i}\right| > x\right) \leqslant 2\exp\left(-\frac{1}{2}\frac{x^2}{n\sigma_n^2 + W_n x/3}\right).$$

Thus

$$\Pr\left(\frac{1}{\sqrt{n}\sigma_n}\left|\sum_{i=1}^{n} Z_{n,i}\right| > x\right) \leqslant 2\exp\left(-\frac{1}{2}\frac{n\sigma_n^2 x^2}{n\sigma_n^2 + W_n\sqrt{n}/(\sigma_n x/3)}\right).$$

If $\limsup_{n\to\infty} W_n/\sqrt{n}\sigma_n < \infty$, then the $\limsup_{n\to\infty}$ of the right-hand side converges to zero for $x \to \infty$. This completes the proof. □

The following lemma is used in the analysis of $R_{2,n}(S_n, \hat{k})$.

**Lemma 2.** *For each $n$ and $k \in \{1, \ldots, K(n)\}$, let $Z_{k,n,i}$, $i = 1, \ldots, n$, be $n$ independent mean-zero random variables with variance* $\mathrm{var}(Z_{k,n,i}) \leqslant \sigma_n^2$ *and* $\Pr(\max_{i,k}|Z_{k,n,i}| \leqslant W_n) = 1$ *for $W_n < \infty$. If $W_n/n^{0.5}\sigma_n = O(1)$, then*

$$\max_{k\in\{1,\ldots,K(n)\}}\left|\frac{1}{n}\sum_{i=1}^{n} Z_{k,n,i}\right| = O_P\left(\frac{\log K(n)}{n^{0.5}\sigma_n^*}\right),$$

*where $\sigma_n^* = \max(\sigma_n, n^{-0.5})$. Specifically,*

$$\Pr\left(\frac{n^{0.5}}{\sigma_n \log K(n)}\max_{k\in\{1,\ldots,K(n)\}}\left|\frac{1}{n}\sum_{i=1}^{n} Z_{k,n,i}\right| > x\right) \leqslant K(n)2\exp\left(-\frac{1}{2}\frac{x^2\log K(n)}{\{1/\log K(n) + W_n x/3\sigma_n n^{0.5}}\right).$$

***Proof.*** By the Bonferoni argument and Bernstein's inequality, we have

$$\begin{aligned}
\Pr\left(\frac{1}{\sigma_n n^{0.5}\log K(n)}\max_k\left|\sum_{i=1}^{n} Z_{k,n,i}\right| > x\right) &\leqslant K(n)\max_k \Pr\left(\left|\sum_{i=1}^{n} Z_{k,n,i}\right| > x\sigma_n n^{0.5}\log K(n)\right)\\
&\leqslant K(n)2\max_k \exp\left(\frac{-x^2\sigma_n^2 n\log^2 K(n)}{n\sigma_n^2 + W_n\sigma_n n^{0.5}\log K(n)x/3}\right)\\
&= K(n)2\max_k \exp\left(-\frac{1}{2}\frac{x^2\log K(n)}{1/\log K(n) + W_n x/3\sigma_n n^{0.5}}\right).\\
&= K(n)2\max_k\left(\frac{1}{K(n)}\right)^{x2/2\{1/\log K(n)+W_n x/3\sigma_n n^{0.5}\}}.
\end{aligned}$$

□

We now analyse $R_{2,n}(S_n, \hat{k})$ using Lemma 2.

**Lemma 3.** *Define*

$$K(a(n)) \equiv \{k \in \{1, \ldots, K(n)\} : \|\tilde{\theta}_{n(1-p)}(k) - \theta_{\text{opt}}\|_F \leqslant a(n)\}.$$

*Let* $a(n)$ *be such that* $P(\hat{k} \in K(a(n))) \to 1$. *We have*

$$R_{n,2}(S_n, \hat{k}) = O_P\left(\max\left(\frac{a(n)\log K(n)}{(np)^{0.5}}, \frac{\log K(n)}{np}\right)\right). \tag{18}$$

***Proof.*** Define $B(n, k) = I(k \in K(a(n)), \overline{G}_{n,S_n}(T|X) > \delta/2)$ and decompose $R_{2,n}(S_n, \hat{k})$ as

$$R_{2,n}(S_n, \hat{k}) = R_{2,n}(S_n, \hat{k})B(n, \hat{k}) + R_{2,n}(S_n, \hat{k})(1 - B(n, \hat{k})).$$

Since $P(\hat{k} \in K(a(n))) \to 1$ and $\overline{G}(T|X) > \delta > 0$, $F$-a.e. (by (A.3)) we have that the probability that the second term equals 0 converges to 1 as $n \to \infty$. Thus this term converges to 0 in probability at an arbitrary rate. We note that

$$\Pr(R_{2,n}(S_n, \hat{k})B(n, \hat{k}) > x|S_n, P^0_{n,S_n}) \leqslant \Pr\left(\max_{k \in K(a(n))} R_{2,n}(S_n, k)B(n, k) > x \middle| S_n, P^0_{n,S_n}\right).$$

Let $Z_{k,n} = I(\overline{G}^0_{n,S_n}(T|X) > \delta/2)IC_0[Y|G^0_{n,S_n}, L(\cdot, \psi_k(\cdot|P^0_{n,S_n})) - L(\cdot, \psi_0)]$. For $k \in K(a(n))$ we have that

$$R_{2,n}(S_n, k)B(n, k) = \int Z_{k,n} \mathrm{d}(P^1_{n,S_n} - P)(Y).$$

We now apply Lemma 2 with $Z_{k,n}$. It is straightforward to show that $W_n = M_1/\delta$ so that $\Pr(\max|Z_{k,n}| \leqslant W_n) = 1$. Next, we note that the conditional variance of $Z_n$ is bounded by

$$\begin{aligned}\mathrm{E}[Z^2_{k,n}|S_n, P^0_{n,S_n}] &\leqslant \frac{1}{\delta}\int \Big(L(Z, \psi_k(\cdot|P^0_{n,S_n})) - L(X, \psi_0)\Big)^2 \mathrm{d}F(X) \\ &\leqslant \frac{M_2}{\delta}\int \Big(L(X, \psi_k(\cdot|P^0_{n,S_n})) - L(X, \psi_0)\Big)\mathrm{d}F(X),\end{aligned}$$

where we have used the double expectation theorem ($\mathrm{E}_Y[Z] = \mathrm{E}_X\mathrm{E}_{Y|X}[Z|X]$) and the last inequality follows by assumption (A.2). Define

$$a(n)^* = \max\left(a(n), \frac{1}{(np)^{0.5}}\right).$$

We now note that $\mathrm{var}(Z_{k,n}|S_n = s_n, P^0_{n,S_n})$ is bounded above by $O(a(n)^*)$, up to a constant. A direct application of Lemma 2 with $Z_{k,n} - \int Z_{k,n}\mathrm{d}P$ gives, for each $x > 0$,

$$\Pr\left(\frac{(np)^{0.5}}{a(n)^* \log K(n)} I(\hat{k} \in K(a(n))) R_{n,2}(S_n, \hat{k}) > x \middle| S_n, P^0_{n,S_n}\right) \leqslant K(n) \exp\left(-\frac{1}{2} \frac{x^2 \log K(n)}{1/\log K(n) + W_n x/3a(n)^*(np)^{0.5}}\right), \tag{19}$$

where $W_n = M_1/\delta$. The definition of $a(n)^*$ implies that $W_n/a(n)^*(np)^{0.5} = O(1)$. We note that this bound in particular holds marginally, and hence we obtain (18). This completes the proof. □

**Lemma 4.**

$$R_{2,n}(S_n, \tilde{k}) = O_P\left(\max\left(\frac{\log K(n)}{R(n)^{0.5^-}(np)^{0.5}}, \frac{\log K(n)}{np}\right)\right). \tag{20}$$

***Proof.*** This result can be proved in a similar way to Lemma 3 with $a(n) = R(n)^{-0.5^-}$ and $a(n)^* = \max(R(n)^{-0.5^-}, (np)^{-0.5})$, where the notation $R(n)^{-0.5^-}$ implies a rate slower than $R(n)^{-0.5}$, as mentioned previously. □

Since the number of support points of $S_n$ is bounded by some $V < \infty$ and the probability on each support point is bounded away from zero uniformly in $n$, Lemmas 3 and 4 readily provide the required results for $\mathrm{E}_{S_n} R_{2,n}(S_n, \tilde{k})$ and $\mathrm{E}_{S_n} R_{2,n}(S_n, \hat{k})$. The corollary given below states these results.

**Corollary 2.** *Redefine*

$$K(a(n)) \equiv \left\{k \in \{1, \ldots, K(n)\} : \sqrt{\mathrm{E}_{s_n} \int \left(L(X, \psi_k(\cdot|P^0_{n,S_n})) - L(X, \psi_0)\right)^2 \mathrm{d}F(X)} \leqslant a(n)\right\}.$$

*Let $a(n)$ be such that $P(\hat{k} \in K(a(n))) \to 1$. Then we have*

$$\mathrm{E}_{S_n} R_{2,n}(S_n, \hat{k}) = O_P\left(\max\left(\frac{a(n) \log K(n)}{(np)^{0.5}}, \frac{\log K(n)}{np}\right)\right). \tag{21}$$

*Similarly, with $a(n) = R(n)^{-0.5^-}$ so that $P(\tilde{k} \in K(a(n))) \to 1$, we have*

$$\mathrm{E}_{S_n} R_{2,n}(S_n, \tilde{k}) = O_P\left(\max\left(\frac{\log K(n)}{R(n)^{0.5^-}(np)^{0.5}}, \frac{\log K(n)}{np}\right)\right). \tag{22}$$

We now present the lemma used in the analysis of $R_{1,n}$ terms.

**Lemma 5.** *We have*

$$\mathrm{E}_{S_n} \int IC_0\left[Y|G^0_{n,S_n} - G, L(\cdot, \psi_k(\cdot|P^0_{n,S_n})) - L(\cdot, \psi_0)\right] \mathrm{d}P(Y)$$

$$= O_P\left(\sqrt{\tilde{\theta}_{n(1-p)}(k) - \theta_{\text{opt}}}\sqrt{\mathrm{E}_{S_n}\int(\overline{G}^0_{n,S_n} - \overline{G})^2(T|X)\mathrm{d}F(X)}\right).$$

***Proof.***

$$\begin{aligned}
&\mathrm{E}_{S_n}\left[\int IC_0[Y|G^0_{n,S_n} - G,\, L(\cdot,\, \psi_k(\cdot|P^0_{n,S_n})) - L(\cdot,\, \psi_0)]\mathrm{d}P(Y)\right] \\
&\quad = \mathrm{E}_{S_n}\mathrm{E}_Y\left[\left(L(X,\, \psi_k(W|P^0_{n,S_n})) - L(X,\, \psi_0)\right)\frac{\Delta(\overline{G} - \overline{G}^0_{n,S_n})(T|X)}{\overline{G}(T|X)\overline{G}^0_{n,S_n}(T|X)}\right] \\
&\quad = \mathrm{E}_{S_n}\mathrm{E}_X\left[\left(L(X,\, \psi_k(W|P^0_{n,S_n})) - L(X,\, \psi_0)\right)\frac{(\overline{G} - \overline{G}^0_{n,S_n})(T|X))}{\overline{G}^0_{n,S_n}(T|X)}\right] \\
&\quad \leqslant \sup_X\left\{\frac{(|\overline{G} - \overline{G}^0_{n,S_n})(T|X)|}{\overline{G}^0_{n,S_n}(T|X)}\right\} \times \sqrt{\mathrm{E}_{S_n}\int(\overline{G}^0_{n,S_n} - \overline{G})^2(T|X)\mathrm{d}F(X)} \\
&\qquad \times \sqrt{\mathrm{E}_{S_n}\int(L(X,\, \psi_k(W|P^0_{n,S_n})) - L(X,\, \psi_0))^2\mathrm{d}F(X)} \\
&\quad \leqslant \sup_X\left\{\frac{(|\overline{G} - \overline{G}^0_{n,S_n})(T|X)|}{\overline{G}^0_{n,S_n}(T|X)}\right\} \times \sqrt{\mathrm{E}_{S_n}\int(\overline{G}^0_{n,S_n} - \overline{G})^2(T|X)\mathrm{d}F(X)} \\
&\qquad \times \sqrt{\mathrm{E}_{S_n}\int(L(X,\, \psi_k(W|P^0_{n,S_n})) - L(X,\, \psi_0))^2\mathrm{d}F(X)} \\
&\quad \leqslant M_2\sup_X\left\{\frac{(|\overline{G} - \overline{G}^0_{n,S_n})(T|X)|}{\overline{G}^0_{n,S_n}(T|X)}\right\} \times \sqrt{\mathrm{E}_{S_n}\int(\overline{G}^0_{n,S_n} - \overline{G})^2(T|X)\mathrm{d}F(X)} \\
&\qquad \times \sqrt{\mathrm{E}_{S_n}\int(L(X,\, \psi_k(W|P^0_{n,S_n})) - L(X,\, \psi_0))\mathrm{d}F(X)} \\
&\quad = O_P\left(\sqrt{\tilde{\theta}_{n(1-p)}(k) - \theta_{\text{opt}}}\sqrt{\mathrm{E}_{S_n}\int(\overline{G}^0_{n,S_n} - \overline{G})^2(T|X)\mathrm{d}F(X)}\right),
\end{aligned}$$

where we have used the fact that $\Pr(\overline{G}^0_{n,S_n}(T|X) > \delta/2) \to 1$. (The first inequality above follows by Cauchy–Schwarz, the third by (A.2).) This completes the proof. □

Using the above lemma with $k = \tilde{k}$ and $k = \hat{k}$, we obtain

$$\mathrm{E}_{S_n} R_{1,n}(S_n, \tilde{k}) = O_P\left(\sqrt{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{\mathrm{opt}}} \sqrt{\mathrm{E}_{S_n} \int (\overline{G}^0_{n,S_n} - \overline{G})^2(T|X)) \mathrm{d}F(X)}\right),$$

$$\mathrm{E}_{S_n} R_{1,n}(S_n, \hat{k}) = O_P\left(\sqrt{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{\mathrm{opt}}} \sqrt{\mathrm{E}_{S_n} \int (\overline{G}^0_{n,S_n} - \overline{G})^2(T|X)) \mathrm{d}F(X)}\right).$$

This completes the proof of Theorem 1. □

# 6. Discussion

We have presented an asymptotically optimal model selection method for choosing among predictors of right censored outcomes such as survival times. This method deals with the censored data by replacing the full data loss function by an observed data loss function that has the same expectation as the full data loss function. Specifically, it treats the risk of a given predictor based on training sample as a full data parameter in a censored data model. We have discussed the estimation of this conditional risk with an inverse probability of censoring weighted estimator based on the validation sample. The consistency of this estimator relies on consistent estimation of the censoring mechanism and the condition that $\overline{G}(T|X) > \delta > 0$, $F$-a.e. One can improve on this estimator by constructing a doubly robust risk estimator. Applying the general doubly robust estimation methodology (Robins and Rotnitzky 1992; van der Laan and Robins 2003), we replace the observed data IPCW loss function $IC_0[Y|G, L(\cdot, \psi)]$ in our risk estimate by

$$IC[Y|G, Q(F, G), L(\cdot, \psi)] = IC_0[Y|G, L(\cdot, \psi)] + \int Q(F, G)(u, \overline{X}(u)) \mathrm{d}M_G(u),$$

where $\mathrm{d}M_G(u) = I(\tilde{T} \in \mathrm{d}u, \Delta = 0) - I(\tilde{T} \geqslant u)\lambda_C(u|X)\mathrm{d}u$ is the martingale of the censoring process $A(\cdot) = I(C \leqslant \cdot)$ and $Q(F, G)(u, \overline{X}(u))$ denotes the conditional expectation of $IC_0(Y|G, L(\cdot, \psi))$ given $\overline{X}(u)$, $C > u$ under the observed data generating distribution $P_{F,G}$. This new observed data loss function equals the observed data IPCW loss function $IC_0(\cdot|G, L(\cdot, \psi))$ minus its projection onto the nuisance tangent space for the censoring mechanism $G$ in the nonparametric observed data model only assuming CAR (Robins and Rotnitzky 1992). We refer to van der Laan and Robins (2003) for a detailed treatment of such projections and the general methodology of obtaining doubly robust estimators. $IC[Y|G, Q(F, G), L(\cdot, \psi)]$ has the so-called *double robust property*. Let $G^1$ and $F^1$ be the guessed models of the censoring mechanism $G$ and the full data distribution $F$, respectively. Then, we have

$$\mathrm{E}_{P_{F,G}} IC[Y|G_1, Q(F^1, G_1), L(\cdot, \psi)] = \mathrm{E}_F[L(X, \psi]$$

if either $G_1 = G$ and $\overline{G}(T|X) > \delta > 0$, $F$-a.e., or $Q(F^1, G^1) = Q(F, G^1)$. This implies that the double robustness property allows misspecification of either the censoring mechanism or

the part of the full data distribution that is utilized in the projections. We then define the doubly robust cross-validated risk estimator for a given predictor $\psi_k(\cdot|P^0_{n,S_n})$ as

$$\hat{\theta}_{n(1-p)}(k) = \mathrm{E}_{S_n}\int IC[Y|G^0_{n,S_n}, Q^0_{n,S_n}, L(\cdot, \psi_k(\cdot|P^0_{n,S_n}))]\mathrm{d}P^1_{n,S_n}(Y), \tag{23}$$

where $\bar{G}^0_{n,S_n}$ is an estimate of $\bar{G}(\cdot|X)$ and $Q^0_{n,S_n}$ is an estimate of $Q_{F,G}$ based on the training sample defined by the split $S_n$. $\hat{k}$ is again chosen such that $\hat{\theta}_{n(1-p)}(\hat{k})$ is minimized over the index set $k \in \{1, \ldots, K(n)\}$. This double robust risk estimator is also generally more efficient than the IPCW risk estimator. We refer to van der Laan and Robins (2003) for the estimation methods of the nuisance parameters $G$ and $Q(F, G)$ and the general properties of these double robust estimators. The properties of this doubly robust cross-validated model selection criteria are investigated in van der Laan and Dudoit (2003) and both finite-sample and asymptotic optimality results are established.

# References


Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csáki (eds), *Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akadémiai Kiadó.

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Hudson, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, J., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P. and Staudt, L., (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, **403**(6769), 503–511.

Andersen, P., Borgan, Ø., Gill, R. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Beer, D., Kardia, S., Huang, C.-C., Giordano, T., Levin, A., Misek, D., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G., Lizyness, M.L., Kuick, R., Hayasaka, S., Taylor, J., Iannettoni, M., Obringer, M. and Hanash, S. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med.*, **8**, 816–824.

Binder, D.A. (1992) Fitting Cox proportional hazards models from survey data. *Biometricka*, **79**, 139–147.

Bozdogan, H. (2000) Akaike's information criterion and recent developments in information complexity. *J. Math. Psych.*, **44**, 62–91.

Breiman, L., Friedman, J., Olshan, R. and Stone, C. (1984) *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole.

Burman, P. (1989) A comparative study of ordinary cross-validation, $v$-fold cross-validation and the repeated learning testing methods. *Biometrika*, **76**, 503–514.

Davis, R. and Anderson, J. (1989) Exponential survival trees. *Statist. Med.*, **8**, 947–961.

Garber, M.E., Troyanskaya, O.G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G.D., Perou, C.M., Whyte, R.I., Altman, R.B., Brown, P.O., Botstein, D. and Petersen, I. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Nat. Acad. Sci. USA*, **98**, 13 784–13 789.

Gill, R., van der Laan, M. and Robins, J. (1997) Coarsening at random: characterizations, conjectures

and counter-examples. In D. Lin and T. Fleming (eds), *Proceedings of the First Seattle Symposium in Biostatistics*, pp. 255–294. New York: Springer-Verlag.

Gordon, L. and Olshen, R. (1985) Tree-structured survival analysis. *Cancer Treatment Rep.*, **69**, 1062–1069.

Graft, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999) Assessment and comparison of prognostic classification schemes for survival data. *Statist. Med.*, **18**, 2529–2545.

Hastie, T. and Tibshirani, R. (1990a) Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, **46**, 1005–1016.

Hastie, T. and Tibshirani, R. (1990b) *Generalized Additive Models*. London: Chapman & Hall.

Heitjan, D. and Rubin, D. (1991) Ignorability and coarse data. *Ann. Statist.*, **19**, 2244–2253.

Jacobsen, M. and Keiding, N. (1995) Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Statist.*, **23**, 774–786.

Keleş S. (2003) Statistical methods for *cis*-regulatory motif detection in DNA sequences and two censored data problems. PhD thesis, University of California, Berkeley.

Keleş S., van der Laan, M. and Dudoit, S. (2003) Asymptotically optimal model selection method with right censored outcomes. Techinical Report 124, Division of Biostatistics, University of California: Berkely. http://www.bepress.com/ucbbiostat/paper124/

Kooperberg, C., Stone, C. and Truong, Y. (1995) Hazard regression. *J. Amer. Statist. Assoc.*, **90**, 78–94.

Korn, E. and Simon, R. (1990) Measures of explained variation for survival data. *Statist. Med.*, **9**, 487–503.

Leblanc, M. and Crowley, J. (1992) Relative risk trees for censored data. *Biometrics*, **48**, 411–425.

O'Quigley, J. and Xu, R. (2001) Explained variation in proportional hazards regression. In J. Crowely (ed.), *Handbook of Statistics in Clinical Oncology*, pp. 397–409. New York: Marcel Dekker.

Pugh, M., Robins, J., Lipsitz, S. and Harrington, D. (1993) Inference in the Cox proportional hazards model with missing covariate. Technical report, Department of Biostatistics, Harvard University.

Robins, J. (1993) Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. *Proc. Biopharm. Sect. Amer. Statist. Assoc.*, 24–33.

Robins, J. and Rotnitzky, A. (1992) Recovery of information and adjustment for dependent censoring using surrogate markers. In N.P. Jewell, K. Dietz and V.T. Farewell (eds), *AIDS Epidemiology: Methodoligical Issues*. Boston: Birkhäuser.

Robins, J. and Rotnitzky, A. (2001) Comment on the Bickel and Kwon article, 'Inference for semiparametric models: Some questions and an answer'. *Statist. Sinica*, **11**, 920–936.

Robins, J., Rotnitzky, A. and van der Laan, M. (2000) Comment on 'On Profile Likelihood' by S.A. Murphy and A.W. van der Vaart. *J. Amer. Statist. Assoc.*, **95**, 477–482.

Ruczinski, I., Kooperberg, C. and Leblanc, M.L. (2001) Logic regression. Manuscript.

Schemper, M. and Henderson, R. (2000) Predictive accuracy and explained variation in Cox regression. *Biometrics*, **56**, 249–255.

Schemper, M. and Stare, J. (1996) Explained variation in survival analysis. *Statist. medicine*, **15**, 1999–2012.

Schwartz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Segal, M. (1988) Regression trees for censored data. *Biometrics*, **44**, 35–47.

Shao, J. (1993) Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, **88**, 486–494.

Sorliea, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Lonning, P.E. and Borresen-Dale, A.-L. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Nat. Acad. Sci. USA*, **98**, 10 869–10 874.

van der Laan, M.J. and Dudoit, S. (2003) Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. Technical Report 130, Division of Biostatistics, University of California, Berkeley. http://www.bepress.com/ucbbiostat/paper130/

van der Laan, M. and Robins, J. (2003) *Unified Methods for Censored Longitudinal Data and Causality.* New York: Springer-Verlag.

van der Vaart, A. and Wellner, J. (1996) *Weak Convergence and Empirical Processes.* New York: Springer-Verlag.

Wigle, D.A., Jurisica, I., Radulovich, N., Pintilie, M., Rossant, J., Ni Liu, C.L., Woodgett, J., Seiden, I., Johnston, M., Shaf Keshavjee, G.D., Winton, T., Breitkreutz, B.J., Jorgenson, P., Mike Tyers, F.A.S. and Tsao, M.S. (2002) Molecular profiling of non-small cell lung cancer and correlation with disease-free survival, *Cancer Res.*, **62**, 3005–3008.

Zhang, P. (1993) Model selection via multifold cross-validation, *Ann. Statist.*, **21**, 299–313.

Received January 2003 and revised December 2003