K-Adaptive Partitioning for Survival Data: The kaps Add-on Package for R

Soo-Heang Eo
Korea University

Seung-Mo Hong
Asan Medical Center

HyungJun Cho
Korea University

Abstract

The partitioning of an ordered prognostic factor is important in order to obtain several groups having heterogeneous survivals in medical research. For this purpose, a binary split has often been used once or recursively. We propose the use of a multi-way split in order to afford an optimal set of cut-off points. In practice, the number of groups (K) may not be specified in advance. Thus, we also suggest finding an optimal K by a resampling technique. The algorithm was implemented into an R package that we called kaps, which can be used conveniently and freely. It was illustrated with a toy dataset, and was also applied to a real data set of colorectal cancer cases from the Surveillance Epidemiology and End Results.

Keywords: cutpoint model, adaptive partitioning, multi-way split, TNM staging, SEER.

1. Introduction

Clinicians are interested in obtaining several groups with heterogeneous survivals by partitioning an ordered prognostic factor. A staging system can be constructed by a kind of partitioning. The tumor node metastasis (TNM) staging system is the most widely used cancer staging system, and provides critical information about prognosis and about estimation for responsiveness to specific treatment for cancer patients (Edge, Byrd, Compton, Fritz, Greene, and Trotti 2010). The TNM staging system is composed of 3 classifications: T classification based on the extent or size of the primary tumor, N classification determined by the involvement of the regional lymph nodes (LNs), and M classification by distant metastasis. Each T, N, or M classification is decided by grouping cases with similar prognosis. When T classification, based solely on the size of the primary tumor such as breast cancer, or N classification, in several gastrointestinal tract cancers, was determined, increased tumor size or increased number of metastatic regional LNs is linked with worse prognosis of cancer patients.

Several studies have been conducted for partitioning an ordered prognostic factor or fineing cut-off points. Horhorn and Lausen (2003) utilized the maximally selected rank statistic and Abdolell, LeBlanc, Stephens, and Harrison (2002) used the likelihood ratio test statistics to obtain two subgroups with different survivals or cancer groups. However, these approaches only revealed two subgroups such as high-risk and low-risk groups. Negassa, Ciampi, Abrahamowicz, Shapiro, and Boivin (2005) and Hong, Cho, Moskaluk, and Yu (2007) utilized tree-structured methods to find an optimal set of cut-off points, so as to obtain several heterogeneous subgroups. Recursive partitioning selects the best point at the first split, but its
Figure 1: Tree diagram from the log-rank survival tree for the colorectal cancer data. Each oval including a split rule depicts an intermediate node and each rectangle with the node number (Node), the number of observations (n), and the median survival time (med) describes a terminal node. An observation goes to the left subnode if and only if the condition is satisfied. Information related to the intermediate node is presented on the right side of the ovals.
subsequent split points may not be optimal together. For illustration, we consider the data regarding colorectal cancer (Edge et al. 2010) from the Surveillance Epidemiology and End Results (SEER), which can be obtained from the SEER website (http://seer.cancer.gov). The number of metastatic LNs acts as a prognostic factor to obtain several heterogeneous subgroups with different levels of survival. For analysis, we selected 65,186 cases with more than 12 examined LNs because the examination of more than 12 LNs is accepted for proper evaluation of the prognosis of patients with colorectal cancers (Otchy, Hyman, Simmang, Anthony, Buie, Cataldo, Church, Cohen, Dentsman, Ellis, Kilkenny, Ko, Orsay, Moore, Place, Rafferty, Rakinic, Savoca, Tjandra, and Whiteford 2004). Figure 1 shows a tree-diagram for the colorectal cancer data by the tree-based method used in Hong et al. (2007). The Kaplan-Meier survival curves for the resulting subgroups are also displayed in Figure 2. This indicates that survivals of some subgroups differ insignificantly or their differences are not equal-spaced. Thus, we propose an algorithm for overcoming these limitations and introduce a convenient software program in this paper.

Our algorithm evaluates multi-way split points simultaneously and finds an optimal set of cut-off points. In addition, an optimal number of subgroups is selected by a resampling technique. The algorithm was implemented into an R package kaps, which can be used conveniently and freely.

The rest of the paper is organized as follows. In Section 2, we describe our K-adaptive partitioning called KAPS. In Section 3, the package is described and illustrated with a toy data set. In Section 4, the system requirements, availability and installation for the package kaps are summarized.
2. K-adaptive partitioning for survival data

We describe a new partitioning algorithm that combines the merits of multi-way splits and pairwise comparisons between subgroup levels for survival data. The data are divided into several heterogeneous subgroups to explain the status of progression and determine the number of subgroups by permutation test. In this paper, a novel adaptive partitioning algorithm based on the multi-way split approach is proposed to escape the limitation of tree-based models. We call our proposed algorithm K-Adaptive Partitioning for Survival data, or KAPS for short.

2.1. Selecting a set of cut-off points

For any standard set of data \(D\), there is a collection \(S\) of split sets \(s\), i.e., \(s \in S\). Such a split set consists of \((K - 1)\) cut-off points, which divide the cases into \(K\) subsets. The cut-off points are simply defined from the ordered unique values of a prognostic variable \(X\), so as to divide the data \(D\) into \(K\) disjoint subgroups. To compare the subgroups, we can utilize test statistics such as the Logrank test or Gehan-Wilcoxon test. Let \(\chi^2_k(g, h; s, K)\) be the \(\chi^2\) statistic with 1 degree of freedom (df) for comparing the \(g\)th and \(h\)th subgroups created by split \(s\) when \(K\) is given. For split \(s\) of \(D\) into \(D_1, D_2, \ldots, D_K\), the test statistic provides a measure of deviance defined as

\[
\chi^2_1(s, K) = \min\{\chi^2_k(g, h; s, K), 1 \leq g < h \leq K\}. \tag{1}
\]

Take \(s^*\) as the best split such that

\[
X^2(s^*, K) = \max_{s \in S} \chi^2_1(s, K). \tag{2}
\]

The best split \(s^*\) is a set of \((K - 1)\) cut-off points which well separate the data \(D\) into \(K\) subsets: \(D_1, D_2, \ldots, D_K\). The overall performance can be evaluated by the overall test statistic \(X^2_k(s^*, s, K) = \chi^2_k\) statistic with \(s^*\) and \(K\), where \(k = K - 1\).

Figure 3 illustrates our proposed algorithm. For instance, assume \(K = 3\). This indicates that the data \(D\) are divided into three subgroups, \(D_1, D_2,\) and \(D_3\), by an ordered prognostic factor \(X\). Therefore, we need to compare three pairs of subgroups, i.e., \(D_1\) vs. \(D_2, D_2\) vs. \(D_3,\) and \(D_1\) vs. \(D_3\) by an appropriate test such as the Log-rank or Gehan-Wilcoxon tests. Suppose there are three candidate split sets \(s_1, s_2,\) and \(s_3\) in \(S\), each of which consists of two cut-off points of \(X\), \(s_1 = \{0, 2\}, s_2 = \{2, 4\},\) and \(s_3 = \{0, 4\}\). For example, \(s_3 = \{0, 4\}\) means that \(D_1 = \{X = 0\}, D_2 = \{0 < X \leq 4\},\) and \(D_3 = \{X > 4\}\). Out of the three pairs of the subgroups, the smallest test statistic, equivalently the largest \(p\)-value, is selected as a representative statistic \(\chi^2_1(s, K = 3)\) for the partition generated by \(s\) since it is the worst case. The test statistic, \(\chi^2_1(g = D_1, h = D_2; s = s_1, K = 3)\), for \(D_1\) vs. \(D_2\) is 263. Likewise, the statistics for \(D_2\) vs. \(D_3\) and \(D_1\) vs. \(D_3\) are 6353 and 1039, respectively. We select the statistic for \(D_1\) vs. \(D_2\) as a test statistic because it has the smallest value among all pairs. Thus, the statistic \(\chi^2_1(s_1, 3)\) is 263. Then, we repeat the process for all elements in \(S\) to take the representative statistic, i.e., we compute test statistics for all possible splits of \(S = \{s_1, s_2, s_3\}\). As a result, the best split \(s^*\) is declared as \(\{0, 4\}\) because \(\chi^2_1(s_3 = \{0, 4\}, 3)\) is the largest. This algorithm is summarized below.

Algorithm 1. Selecting the best split for given \(K\)
Step 1: Compute test statistics $\chi^2(g, h; s, K)$ for all possible pairs, $g$ and $h$, of a $K$-way partition generated by $s$, where $1 \leq g < h \leq K$.

Step 2: Obtain $\chi^2(s, K)$ by minimizing $\chi^2(g, h; s, K)$ for all $g, h$.

Step 3: Repeat Steps 1 and 2 for all possible partitions generated by $S$.

Step 4: Take the best split $s^*$ such that $X^2(s^*, K) = \max_{s \in S}\chi^2(s, K)$.

2.2. Finding an optimal $K$

One of the important issues in partitioning the data is to determine the number of subgroups, i.e., the selection of an optimal $K$. Test statistics should not be compared directly due to their different degrees of freedom. To solve this problem, we utilize the premutation procedure for finding the automatic selection of an optimal $K$.

Let $s^*_K$ and $X^2(s^*, K)$ be the best split set and the minimum pairwise statistic using the raw data for each $K$. The data can be reconstructed by matching their labels after permuting the labels of $X$ with retaining the labels of $(Y, \delta)$. The survival time $Y$ is independent of the covariate $X$ in the reconstructed data, which is called the permuted data. When the permuted data are allocated into each subgroup by $s^*_K$, there should be no significant differences in survival among the subgroups. The repetition of this procedure generates the null distribution of the test statistics. If we repeat this procedure many times ($R$ times), and then we obtain the permutation $p$-value $p_K$ for each $K$. This is the ratio where the minimum pairwise statistics of the permuted data are greater than or equal to that of the raw data.
i.e.,

\[ p_K = \frac{\sum_{r=1}^{R} I(X^2(r)(s^*, K) \geq X^2(s^*, K))/R, K = 2, 3, \ldots,} \]

where \( X^2(r)(s^*, K) \) is the \( r^{th} \) repeated minimum pairwise statistic for the permuted data. In addition, we correct the \( p \)-values for multiple comparison because there are \((K - 1)\) comparisons between two adjacent subgroups when there are \( K \) subgroups. For example, the corrected \( p \)-value can be obtained using Bonferroni correction, i.e., \( p^c_K = p_K/(K-1), K = 2, 3, \ldots \). Lastly, we choose the largest number to discover as many significantly different subgroups as possible, given that the corrected \( p \)-values are smaller than or equal to a predetermined significance level, e.g. \( \alpha = 0.05 \). Formally, 

\[ \hat{K} = \max\{K| p^c_K \leq \alpha, K = 2, 3, \ldots\}. \] 

Formally, the plots of \( p^c_K \) against \( K \) would be helpful in order to find an optimal \( K \). At abrupt change (elbow) point, \( K \) can be selected.

3. Package description

The algorithm in the previous section was implemented into an R package \texttt{kaps} (Eo and Cho 2014). In this section, we illustrate the use of the algorithm, KAPS, with a toy dataset.

3.1. Overview

The package \texttt{kaps} was written in the R (R Core Team 2013) programming environment which allow clean interface implementation and great extension. The R package \texttt{kaps} mainly depends on \texttt{methods} and \texttt{survival} (Therneau and Lumley 2011) packages. It also depends on other packages, \texttt{Formula} (Zeileis and Croissant 2010) and \texttt{coin} (Zeileis, Wiel, Hornik, and Hothorn 2008). The package \texttt{Formula} is utilized to handle multiple parts on the right-hand side of the \texttt{formula} object for convenient use. The package \texttt{coin} is used for the permutation test for the selection of optimal number of subgroups. In addition, the packages \texttt{locfit} (Loader 2010), \texttt{foreach} (Analytics and Weston 2013) and \texttt{doMC} (Analytics 2013) are suggested to give fancy visualization and minimize computational cost, respectively. The package \texttt{kaps} is available at the Comprehensive R Archive Network (CRAN, \url{http://cran.r-project.org/}).

3.2. Function \texttt{kaps}

The \( K \)-adaptive partitioning algorithm can be conducted by the function \texttt{kaps}. The function usage and input arguments are as follows. The type of the arguments is given in brackets.

\texttt{kaps(formula, data, K = 2:4, mindat, type = c("perm", "NULL"), \ldots)}

- \texttt{formula} [S4 class \texttt{Formula}]: a \texttt{Formula} object with a response variable on the left hand side of the \( \sim \) operator and covariate terms on the right side. The response has to be a survival object with survival time and censoring status in the \texttt{Surv} function.

- \texttt{data} [data.frame]: a data frame with variables used in the \texttt{formula}. It needs at least three variables including survival time, censoring status, and a covariate.
Table 1: The main slots for the kaps S4 class.

| Slot       | Type      | Description                                           |
|------------|-----------|-------------------------------------------------------|
| call       | language  | evaluated function call                               |
| formula    | formula   | formula to be used                                     |
| data       | data.frame| data to be used in the model fitting                  |
| groupID    | vector    | subgroup classified                                    |
| index      | vector    | an index for the selected K                           |
| split.pt   | vector    | cut-off points selected                                |
| results    | list      | results for each K                                    |
| Options    | kapsOptions| minor parameters to be used                           |
| X          | scalar    | test statistic with the worst pair of subgroups       |
| Z          | scalar    | overall test statistic                                 |
| pair       | numeric   | selected pair of subgroups                            |

- **K** [vector]: the number of subgroups. The default value is 2:4.
- **mindat** [scalar]: the minimum number of observations at each subgroup. The default value is 5% of data.
- **type** [character]: a type of optimal subgroup selection algorithm. At this stage, we offer two options. The option "perm" utilizes permutation test, while "NULL" passes a selection algorithm.
- **...** [S4 class kapsOptions]: a list of minor parameters.

The primary arguments used for analysis are **formula** and **data**. All of the information created by kaps is stored into an object from the kaps S4 class. The output structure is given in Table 3.2. Five generic functions are available for the kaps class: show-method, print-method, plot-method, predict-method and summary-method.

### 3.3. Illustrative example

To illustrate the function kaps with various options, we use an artificial data, toy, which consists of 150 artificial observations of the survival time (time), its censoring status (status) and 6 covariates: the patient’s race (Race), age at the initial time (Age), pathological grade (Grade), early onset depression (EOD), the number of metastasis LNs (meta) and the number of examined LNs (exam). The data can be called up from the package kaps:

```R
R> library("kaps")
R> data("toy", package = "kaps")
R> head(toy)
```

| ID | Race | Age | Grade | EOD | meta | exam | status | time |
|----|------|-----|-------|-----|------|------|--------|------|
| 1  | 27156581 | 1   | 86    | NA  | 1    | 12   | 0      | 0    |
| 2  | 16399774 | 1   | 81    | 2   | 45   | 14   | 1      | 26   |
| 3  | 8694870  | 1   | 77    | 2   | 20   | 16   | 1      | 22   |
| 4  | 649393   | 1   | 52    | 3   | 40   | 9    | 13     | 15   |
Here we utilize just 3 variables: *meta*, *status* and *time*. The number of metastasis LNs, *meta*, is used as an ordered prognostic factor for finding heterogeneous subgroups.

R> toy <- toy[, c("meta", "status", "time")]

The available data have the following structure:

R> str(toy)

'data.frame': 150 obs. of 3 variables:
$ meta : int 1 4 0 9 0 1 0 5 0 0 ... 
$ status: num 0 1 1 1 1 0 0 0 1 0 ... 
$ time : int 0 26 22 15 70 96 97 10 32 127 ... 

Selecting a set of cut-off points for given $K$

Suppose we specify the number of subgroups in advance. For instance, $K = 3$. To select an optimal set of two cut-off points when $K = 3$, the function *kaps* is called via the following statements

R> fit1 <- kaps(Surv(time, status) ~ meta, data = toy, K = 3)
R> fit1

Call:
kaps(formula = f, data = toy, K = 3)

K-Adaptive Partitioning for Survival Data

Samples= 150 Optimal K=3

Selecting a set of cut-off points:

Xk df Pr(>|Xk|) X1 df Pr(>|X1|) adj.Pr(|X1|) cut-off points
K=3 36.8 2 0 7.2 1 0.0073 0.026603 0, 10 *

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

P-values of pairwise comparisons

0<=meta<=0 0<meta<=10
0<meta<=10 1e-04 -
10<meta<=38 <.0000 0.0073

On the R command, we first create an object *fit1* by the function *kaps* with the three input arguments formula, data, and K. The object *fit1* has the S4 class *kaps*. The function
show returns the outputs of the object, consisting of three parts: Call, Selecting a set of cut-off points, and P-values of pairwise comparisons.

The first part, Call, displays the model formula with a dataset and a number for \( K \). In this example, the prognostic factor, \( meta \), is used to find three heterogeneous subgroups since \( K = 3 \). Next, the information regarding the selection of an optimal set of cut-off points is provided for given \( K \) in a table. In this part, the \( X_k \) and \( X_1 \) mean the overall and worst-pair test statistics and the \( \Pr(>|X_k|) \) and \( \Pr(>|X_1|) \) denote their corresponding \( p \)-values. The \( \text{adj.} \Pr(|X_1|) \) indicates permuted \( p \) value for the worst-pair with the Bonferroni correction.

When \( K = 3 \), an optimal set of two cut-off points selected by the algorithm is \( s^* = \{0, 10\} \).

The two cut-off points are used to partition the data into three groups: \( meta = 0, 0 < meta \leq 10 \), and \( 10 < meta \leq 38 \). For the three subgroups, the log-rank test statistic \( S_k^2(s^*, K) \left( X_k \right) \), the degree of freedom (df), and the \( p \)-value (\( \Pr(|X_k|) \)) are given in Section 2.1. Note that if \( K \) is not significant, the output part is changed from "Optimal K=3" to "Optimal K<3". It means the value of the argument \( K \) may be less than the present input value.

In fact, the cut-off points were selected by maximizing the test statistic \( X^2(s^*, K) \) of the worst pair of subgroups (\( X \)) in Eq (2) in Section 2.1. Lastly, the \( p \)-values of pairwise two-sample test comparisons among all the pairs of subgroups are provided. The \( p \)-values can be adjusted for multiple comparison, as shown below.

```r
R> fit2 <- kaps(Surv(time, status) ~ meta, data = toy, K=3, +    p.adjust.methods = "holm")
R> fit2
```

It is based on the internal function \( \text{p.adjust} \). The default value of \( \text{p.adjust.methods} \) is "none". The only difference between the objects \( \text{fit1} \) and \( \text{fit2} \) is the \( p \)-values of pairwise comparisons. For more information, refer to the help page of the function \( \text{p.adjust} \). The Kaplan-Meier survival curves can be obtained by

```r
R> plot(fit1)
```

It provides Kaplan-Meier survival curves for the selected subgroups as seen in Figure 4. The method summary shows the tabloid information for the subgroups. It consists of the number of observations (\( N \)), the survival median time (\( \text{Med} \)), and the 1-year (\( \text{yrs.1} \)), 3-year (\( \text{yrs.3} \)), and 5-year (\( \text{yrs.5} \)) survival times. The rows mean orderly for all the data (\( \text{All} \)) and each subgroup.

```r
R> summary(fit1)
```

|     | N  | Med | yrs.1 | yrs.3 | yrs.5 |
|-----|----|-----|-------|-------|-------|
| All | 150| 59.0| 0.812 | 0.622 | 0.495 |
| Group=1 | 94 | 101.0| 0.879 | 0.730 | 0.590 |
| Group=2 | 42 | 35.0 | 0.772 | 0.495 | 0.000 |
| Group=3 | 14 | 19.5 | 0.500 | 0.214 | 0.000 |
**Finding an optimal K**

The number \( K \) of subgroups is usually unknown and may not therefore be specified in advance. Rather, an optimal \( K \) can be selected by the algorithm for a given range of \( K \) as follows:

```r
R> fit3 <- kaps(Surv(time, status) ~ meta, data = toy, K = 2:4)
R> fit3
```

**Call:**

```
kaps(formula = f, data = toy, K = 2:4)
```

---

**K-Adaptive Partitioning for Survival Data**

Samples= 150

Optimal K=3

Selecting a set of cut-off points:

| \( K \) | \( \text{Xk} \) | \( \text{df} \) | \( \text{Pr}(>|X_k|)\) | \( \text{X1} \) | \( \text{df} \) | \( \text{Pr}(>|X_1|)\) | \( \text{adj.Pr}(|X_1|)\) | \text{cut-off points} |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | 26.4 | 1 | 0.0000 | 26.37 | 1 | 0.0001 | 0.0001 | 8 *** |
| 3 | 36.8 | 2 | 0.0073 | 7.20 | 1 | 0.0246 | 0.0246 | 0, 10 * |
| 4 | 38.0 | 3 | 0.1692 | 1.89 | 1 | 0.4887 | 0.4887 | 0, 3, 6 |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

P-values of pairwise comparisons

Figure 4: Kaplan-Meier survival curves for the toy dataset with three subgroups: \( G_1 = \{\text{meta} = 0\} \), \( G_2 = \{0 < \text{meta} \leq 10\} \), and \( G_3 = \{10 < \text{meta} \leq 38\} \).
Optimal sets of cut-off points are selected for each $K$, as seen in the output with the title "Selecting a set of cut-off points". The explanation for the output is the same as that of the previous subsection. Then an optimal $K$ is selected by the algorithm with permutation test as described in Section 2.2, respectively. In the output, $X_k$ and $X_1$ indicate the overall and worst-pair test statistics. Their degrees of freedom and $p$-values are followed in the output. The "adj. Pr(|$X_1|)" is the Bonferroni corrected permuted $p$ value for the worst pair by which we make a decision for the optimal $K$. In this example, an optimal $K$ is 3 because the worst pairs of comparisons were significant with significance level $\alpha = 0.05$ when $K = 2$ and 3, and the worst-pair $p$-value for $K = 4$ is rapidly increased.

The test statistic for determining an optimal $K$ can be displayed by

\[ R> \text{plot(fit3)} \]

It generates the four plots shown in Figure 5. The top left panel is the scatterplot of survival times against the prognostic factor $meta$ with the line fitted by local censored regression (Loader 1999). The top right panel is the Kaplan-Meier survival curves for the subgroups selected with the optimal $K$. At the bottom are displayed the plots of the overall and worst-pair $p$ values against $K$. The dotted lines indicate thresholds for significance ($\alpha = 0.05$).

The outputs for $K$s can also be printed out. For instance, when $K$ is 4, the output is printed out as follows.

\[ R> \text{print(fit3, K= 4)} \]

P-values of pairwise comparisons when $K = 4$

| $0<=$meta<=$0$ | $0<=$meta<=$10$ |
|----------------|----------------|
| $0<=$meta<=$3$ | $1e-04$       |
| $3<=$meta<=$6$ | $0.3307$      |
| $6<=$meta<=$38$| $<.0000$      |

It gives information about pairwise comparisons for a specific $K$.

4. System requirements, availability and installation

The **kaps** is an R package developed newly by employing the following R packages: **methods**, **survival**, **Formula** and **coin**. It requires R (>3.0.0) and runs under Windows and Unix like operating systems. The source code of development version and detailed installation guide for **kaps** are freely available under the terms of GNU license from STATLAB (http://statlab.korea.ac.kr/kaps). The stable version of **kaps** is also available at CRAN (http://cran.r-project.org).
Figure 5: The top left panel is the scatter plot of survival times against the prognostic factor with the line fitted by local censored regression. The top right panel is the Kaplan-Meier survival curves for the selected subgroups. The panels at the bottom are the plots of the overall and worst-pair $p$-values against $K$ with significance level $\alpha = 0.05$. 
5. Conclusion

In this paper, we propose a novel new algorithm for obtaining heterogeneous subgroups by partitioning an ordered prognostic factor. The algorithm finds an optimal set of cut-off points in order to obtain heterogeneous subgroups by evaluating possible multi-way splits. In case the number \((K)\) of subgroups cannot be specified in advance, \(K\) is selected by permutation test. We call the algorithm \(K\)-adaptive partitioning for survival data, short for KAPS. The KAPS algorithm is implemented into an R package called kaps for convenient and free use. The package is build on the S4 formulation. Its use was illustrated with a toy dataset.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2010-0007936). It was also supported by the Asan Institute for Life Sciences, Seoul, Korea (2013-554).

References

Abdolell M, LeBlanc M, Stephens D, Harrison R (2002). “Binary Partitioning for continuous longitudinal data: categorizing a prognostic variable.” Statistics in Medicine, 21, 2295–2309.

Analytics R (2013). doMC: *Foreach parallel adaptor for the multicore package*. R package version 1.3.0, URL http://CRAN.R-project.org/package=doMC.

Analytics R, Weston S (2013). foreach: *Foreach looping construct for R*. R package version 1.4.1, URL http://CRAN.R-project.org/package=foreach.

Edge S, Byrd D, Compton C, Fritz A, Greene F, Trotti Ar (2010). AJCC Cancer staging manual. Springer, New York.

Eo SH, Cho H (2014). kaps: *K Adaptive Partitioning for Survival data*. R package version 1.0.0, URL http://CRAN.R-project.org/package=kaps.

Hong S, Cho H, Moskaluk C, Yu E (2007). “Measurement of the Invasion Depth of Extrahepatic Bile Duct Carcinoma: An Alternative Method Overcoming the Current T Classification Problems of the AJCC Staging System.” American Journal of Surgical Pathology, 31, 199–206.
Horhorn T, Lausen B (2003). “On the exact distribution of maximally selected rank statistics.” *Computational Statistics and Data Analysis*, **43**, 121–137.

Loader C (1999). *Local Regression and Likelihood*. Springer, New York.

Loader C (2010). *locfit: Local Regression, Likelihood and Density Estimation*. R package version 1.5-6, URL [http://CRAN.R-project.org/package=locfit](http://CRAN.R-project.org/package=locfit).

Negassa A, Ciampi A, Abrahamowicz M, Shapiro S, Boivin JF (2005). “Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria.” *Statistics and Computing*, **15**, 231–239.

Otchy D, Hyman N, Simmang C, Anthony T, Buie W, Cataldo P, Church J, Cohen J, Dentsman F, Ellis CN, Kilkenny JWr, Ko C, Orsay C, Moore R, Place R, Raffety J, Rakinic J, Savoca P, Tjandra J, Whiteford M (2004). “Practice parameters for colon cancer.” *Diseases of the colon and rectum*, **47**, 1269 – 1284.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL [http://www.R-project.org/](http://www.R-project.org/).

Ternheau T, Lumley T (2011). *survival: Survival analysis, including penalised likelihood*. R package version 2.36-10, URL [http://CRAN.R-project.org/package=survival](http://CRAN.R-project.org/package=survival).

Zeileis A, Croissant Y (2010). “Extended Model Formulas in R: Multiple Parts and Multiple Responses.” *Journal of Statistical Software*, **34**(XYZ), 1–12. URL [http://www.jstatsoft.org/v34/iXYZ/](http://www.jstatsoft.org/v34/iXYZ/).

Zeileis A, Wiel MA, Hornik K, Hothorn T (2008). “Implementing a class of permutation tests: The coin package.” *Journal of Statistical Software*, **28**(8), 1–23.
Affiliation:

HyungJun Cho (correspondence for statistical issues)
Department of Statistics
Korea University
Anam-Dong, Seoul, Korea
E-mail: hj4cho@korea.ac.kr
URL: http://statlab.korea.ac.kr/

Seung-Mo Hong (correspondence for clinical issues)
Department of Pathology
Asan Medical Center
University of Ulsan College of Medicine
Seoul, Korea
E-mail: smhong28@gmail.com

Soo-Heang Eo
Department of Statistics
Korea University
Anam-Dong, Seoul, Korea
E-mail: hanansh@korea.ac.kr