Chapter 5
Gene Selection and Survival Prediction Under Dependent Censoring

Abstract To select genes that are predictive of survival, univariate selection based on the Cox model has been routinely employed in biomedical research. However, this conventional approach relies on the independent censoring assumption, which is often an unrealistic assumption in many biomedical applications. We introduce an alternative approach to selecting genes by utilizing copulas to account for the effect of dependent censoring. We also introduce a method to construct a predictor based on the selected genes to predict patient survival. We use the non-small-cell lung cancer data to demonstrate the copula-based procedure for selecting genes, developing a predictor, and validating the predictor. We provide detailed instructions to implement the proposed statistical methods and to reproduce the real data analyses through the *compound.Cox* R package.

Keywords Clayton's copula · Competing risk · Compound covariate
Copula-graphic estimator · Cox regression · C-index · Gene expression
Overall survival · Univariate selection

5.1 Introduction

Recent years have witnessed a rapid increase in the use of genetic covariates to build survival prediction models in biomedical research. Accurate prediction of survival is often possible by incorporating genetic covariates into prediction models, as reported in breast cancer (Jenssen et al. 2002; Sabatier et al. 2011; Zhao et al. 2011), diffuse large-B-cell lymphoma (Lossos et al. 2004; Alizadeh et al. 2011), lung cancer (Beer et al. 2002; Chen et al. 2007; Shedden et al. 2008), ovarian cancer (Popple et al. 2012; Yoshihara et al. 2010, 2012; Waldron et al. 2014), and other cancers. Evaluating predictive accuracy of the survival prediction models has been a challenging area of research due to the high-dimensionality of genes (Michiels et al. 2005; Schumacher et al. 2007; Bøvelstad et al. 2007, 2009; Witten and Tibshirani 2010; Zhao et al. 2014; Emura et al. 2017).

T. Emura and Y.-H. Chen, *Analysis of Survival Data with Dependent Censoring*,
JSS Research Series in Statistics, https://doi.org/10.1007/978-981-10-7164-5_5

To overcome the difficulty of handling the high-dimensional genetic covariates, one often needs to obtain a small fraction of genes that are predictive of survival. The traditional approach, called *univariate selection*, is a forward variable selection method according to univariate association between each gene and survival, where the association is measured through univariate Cox regression. A predictor constructed from the selected genes has been shown to be useful for survival prediction (Beer et al. 2002; Wang et al. 2005; Matsui 2006; Chen et al. 2007; Matsui et al. 2012; Emura et al. 2017).

It is well known that Cox regression relies on the independent censoring assumption. From our discussions in Chap. 3, this assumption seems unrealistic in univariate Cox regression, where many covariates are omitted. If the independent censoring assumption is violated, univariate Cox regression may not correctly capture the effect of each gene and thus may fail to select useful genes. Accordingly, the resultant predictor based on the selected genes may have a reduced ability to predict survival.

Emura and Chen (2016) introduced a copula-based method for performing gene selection. With this method, dependence between survival and censoring times is modeled via a copula, whereby relaxing the independent censoring assumption. In the subsequent discussions, we revisit their method by providing more detailed developments than the original paper. We have made the lung cancer data publicly available in the *compound.Cox* R package (Emura et al. 2018) to enhance reproducibility.

The chapter is organized as follows. Section 5.2 reviews the conventional univariate selection. Sections 5.3–5.5 introduce the copula-based method of Emura and Chen (2016). Section 5.6 includes the analysis of the non-small-cell lung cancer data for illustration. Section 5.7 provides discussions.

5.2 Univariate Selection

Univariate selection is the traditional method for selecting a subset of genes that is predictive of survival. As the initial step, one fits the univariate Cox model for each gene, one-by-one. Then, one selects a subset of genes that are univariately associated with survival. Finally, one builds a multi-gene predictor using the subset of genes for purpose of survival prediction. The predictor is usually a weighted sum of gene expressions whose weights reflect the degree of association.

Let $\mathbf{x} = (\ x_1,\ \ldots,\ x_p\ )'$ be a p-dimensional vector of gene expressions, where the dimension p can be large. Let T be survival time having the hazard function $h(t|\mathbf{x}) = \Pr(\ t \le T < t + dt\ |T \ge t,\ \mathbf{x}\ )/dt$. It is well known that the multivariate Cox model $h(t|\mathbf{x}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x})$ does not yield proper estimates of $\boldsymbol{\beta}$ when p is very large (Witten and Tibshirani 2010).

In biomedical research, the univariate Cox regression analysis is the traditional strategy to deal with the large number of covariates (e.g., Beer et al. 2002; Chen et al. 2007). Let $h(t|x_j) = \Pr(\ t \le T < t + dt\ |T \ge t,\ x_j\ )/dt$ be the hazard function given the jth gene. The univariate Cox model is specified as $h_j(t|x_j) = h_{0j}(t)\exp(\beta_j x_j)$ for each gene $j = 1, \ldots, p$. The primary objective of using the univariate Cox model is to perform univariate selection as follows: For each $j = 1, \ldots, p$, the null hypothesis $H_0 : \beta_j = 0$ is examined by the Wald test (or score test) under the univariate Cox model. Then one picks out a subset of genes that have low P-values from the tests. The genes with low P-values are then selected for further analysis.

After genes are selected, they are used to build a prediction scheme for survival. In medical studies, it is a common practice to re-fit a multivariate Cox regression model based on the selected genes (e.g., Lossos et al. 2004). However, we have reservations about this commonly used strategy due to the poor predictive performance observed in many papers (e.g., Bøvelstad et al. 2007; van Wieringen et al. 2009). Alternatively, we suggest using Tukey's compound covariate predictor (Tukey 1993) that combines the results of univariate analyses without going through a multivariate analysis. The compound covariate has been successfully employed in many medical studies (e.g., Beer et al. 2002; Wang et al. 2005; Chen et al. 2007) and biostatistical studies (Matsui 2006; Matsui et al. 2012; Emura et al. 2012, 2017).

The two major assumptions of univariate selection are the correctness of the univariate Cox model and the independent censoring assumption. The violation of these assumptions yields bias in estimating the true effect of genes. Emura and Chen (2016) argued that the independence of censoring is a more crucial assumption than the correctness of the univariate Cox model. The bias due to dependent censoring gets large if either the degree of dependence or the percentage of censoring increases (see Sect. 3.5). In the following sections, we shall introduce a copula-based univariate selection method that copes with the problem of dependent censoring.

5.3 Copula-Based Univariate Cox Regression

Let T be survival time, U be censoring time, and $\mathbf{x} = (\, x_1, \ldots, x_p\,)'$ be gene expressions. The joint distribution of T and U can have an arbitrary dependence pattern for any given x_j. Sklar's theorem (Sklar 1959; Nelsen 2006) guarantees that the joint survival function is expressed as

$$\Pr(\, T > t\,,\ U > u|x_j\,) = C_j\{\, \Pr(\, T > t\,|x_j\,)\,,\ \Pr(\, U > u\,|x_j\,)\,\}, \quad j = 1, \ldots, p,$$

where C_j is a copula. The *independent censoring assumption* corresponds to $C_j(u, v) = uv$ for $j = 1, \ldots, p$, namely,

$$\Pr(\, T > t\,,\ U > u|x_j\,) = \Pr(\, T > t\,|x_j\,) \times \Pr(\, U > u\,|x_j\,), \quad j = 1, \ldots, p. \qquad (5.1)$$

This is clearly a strong assumption (Chap. 3).

To relax the independent censoring assumption, Emura and Chen (2016) suggested a one-parameter copula model

$$\Pr(\, T > t \,,\ U > u | x_j \,) = C_\alpha\{\, \Pr(\, T > t \,| x_j \,),\ \Pr(\, U > u | x_j \,) \,\},\quad j = 1,\ \ldots,\ p. \tag{5.2}$$

Since the same copula C is assumed for every j, this assumption may still be strong. Nevertheless, the copula relaxes the independent censoring assumption (5.1) by allowing a dependence parameter α to be flexibly chosen by users. One example is the Clayton copula

$$C_\alpha(\, u,\ v \,) = \ (\, u^{-\alpha} + v^{-\alpha} - 1 \,)^{-1/\alpha},\qquad \alpha > 0,$$

where the parameter α is related to Kendall's tau through $\tau = \alpha/(\alpha+2)$. The copula model (5.2) reduces to the independent censoring model (5.1) by letting $\alpha \to 0$.

For marginal distributions, Emura and Chen (2016) assumed the Cox models

$$\Pr(\, T > t \,| x_j \,) = \exp\{\, -\Lambda_{0j}(t) e^{\beta_j x_j} \,\},\quad \Pr(\, U > u \,| x_j \,) = \exp\{\, -\Gamma_{0j}(u) e^{\gamma_j x_j} \,\}, \tag{5.3}$$

where β_j and γ_j are regression coefficients and Λ_{0j} and Γ_{0j} are baseline cumulative hazard functions.

For purpose of gene selection, the target parameter is β_j that is the univariate effect of the jth gene on survival. Other parameters $(\, \gamma_j,\ \Lambda_{0j},\ \Gamma_{0j} \,)$ are nuisance. Under the independent censoring model (5.1), one can use the partial likelihood to estimate for β_j while ignoring the nuisance parameters. However, under the copula model (5.2), the partial likelihood estimator gives an inconsistent estimate of β_j (Chap. 3).

The full likelihood is necessary to consistently estimate $(\, \beta_j,\ \gamma_j,\ \Lambda_{0j},\ \Gamma_{0j} \,)$ under the copula model (5.2) and the Cox models (5.3). Define notations

$$D_{\alpha,1}(u,\ v) = \frac{\partial C_\alpha(u,\ v)/\partial u}{C_\alpha(u,\ v)} = -\frac{\partial \Phi_\alpha(u,\ v)}{\partial u},$$
$$D_{\alpha,2}(u,\ v) = \frac{\partial C_\alpha(u,\ v)/\partial v}{C_\alpha(u,\ v)} = -\frac{\partial \Phi_\alpha(u,\ v)}{\partial v},$$

where $\Phi_\alpha(u, v) = -\log C_\alpha(u, v)$. Observed data are denoted as $\{(t_i, \delta_i, x_{ij}), i = 1, \ldots, n\}$, where $t_i = \min(T_i, U_i)$ and $\delta_i = \mathbf{I}(T_i \le U_i)$, where $\mathbf{I}(\bullet)$ is the indicator function. As in Chen (2010), we treat Λ_{0j} and Γ_{0j} as increasing step functions that have jumps sizes $d\Lambda_{0j}(t_i) = \Lambda_{0j}(t_i) - \Lambda_{0j}(t_i - dt)$ for $\delta_i = 1$ and $d\Gamma_{0j}(t_i) = \Gamma_{0j}(t_i) - \Gamma_{0j}(t_i - dt)$ for $\delta_i = 0$. For any given α, the log-likelihood is defined as

$$\begin{aligned} \ell(\beta_j,\ \gamma_j,\ \Lambda_{0j},\ \Gamma_{0j}|\alpha) = & \sum_i \delta_i[\, \beta_j x_{ij} + \log \eta_{1ij}(t_i;\ \beta_j,\ \gamma_j, \Lambda_{0j},\ \Gamma_{0j}|\alpha) + \log d\Lambda_{0j}(t_i) \,] \\ & + \sum_i (1-\delta_i)[\, \gamma_j x_{ij} + \log \eta_{2ij}(t_i;\ \beta_j,\ \gamma_j, \Lambda_{0j},\ \Gamma_{0j}|\alpha) + \log d\Gamma_{0j}(t_i) \,] \\ & - \sum_i \Phi_\alpha[\, \exp\{\, -\Lambda_{0j}(t_i) e^{\beta_j x_{ij}} \,\},\ \exp\{\, -\Gamma_{0j}(t_i) e^{\gamma_j x_{ij}} \,\}] \,, \end{aligned} \tag{5.4}$$

where,

$$\eta_{1ij}(\,t;\ \beta_j,\ \gamma_j,\ \Lambda_{0j},\ \Gamma_{0j}|\alpha\,) = \exp\{\,-\Lambda_{0j}(t)e^{\beta_j x_{ij}}\,\}D_{\alpha,1}[\,\exp\{\,-\Lambda_{0j}(t)e^{\beta_j x_{ij}}\,\},\ \exp\{\,-\Gamma_{0j}(t)e^{\gamma_j x_{ij}}\,\}],$$
$$\eta_{2ij}(\,t;\ \beta_j,\ \gamma_j,\ \Lambda_{0j},\ \Gamma_{0j}|\alpha\,) = \exp\{\,-\Gamma_{0j}(t)e^{\gamma_j x_{ij}}\,\}D_{\alpha,2}[\,\exp\{\,-\Lambda_{0j}(t)e^{\beta_j x_{ij}}\,\},\ \exp\{\,-\Gamma_{0j}(t)e^{\gamma_j x_{ij}}\,\}].$$

The maximizer of Eq. (5.4) given α is denoted as $(\,\hat{\beta}_j(\alpha),\ \hat{\gamma}_j(\alpha),\ \hat{\Lambda}_{0j}(\alpha),\ \hat{\Gamma}_{0j}(\alpha)\,)$. The standard error $SE\{\,\hat{\beta}_j(\alpha)\,\}$ is computed from the information matrix (Chen 2010).

The log-likelihood in Eq. (5.4) can be easily computed under the Clayton copula. It can be shown that $\Phi_\alpha(u,v) = \alpha^{-1}\log(u^{-\alpha}+v^{-\alpha}-1)$, $D_{\alpha,1}(u,\ v) = u^{-\alpha-1}(u^{-\alpha}+v^{-\alpha}-1)^{-1}$, and $D_{\alpha,2}(u,\ v) = u^{-\alpha-1}(u^{-\alpha}+v^{-\alpha}-1)^{-1}$. Hence,

$$\eta_{1ij}(\,t;\ \beta_j,\ \gamma_j,\ \Lambda_{0j},\ \Gamma_{0j}|\alpha\,) = \frac{[\exp\{\,-\Lambda_{0j}(t)e^{\beta_j x_{ij}}\,\}]^{-\alpha}}{[\exp\{\,-\Lambda_{0j}(t)e^{\beta_j x_{ij}}\,\}]^{-\alpha}+[\exp\{\,-\Gamma_{0j}(t)e^{\gamma_j x_{ij}}\,\}]^{-\alpha}-1},$$
$$\eta_{2ij}(\,t;\ \beta_j,\ \gamma_j,\ \Lambda_{0j},\ \Gamma_{0j}|\alpha\,) = \frac{[\exp\{\,-\Gamma_{0j}(t)e^{\gamma_j x_{ij}}\,\}]^{-\alpha}}{[\exp\{\,-\Lambda_{0j}(t)e^{\beta_j x_{ij}}\,\}]^{-\alpha}+[\exp\{\,-\Gamma_{0j}(t)e^{\gamma_j x_{ij}}\,\}]^{-\alpha}-1}.$$

One can apply these formulas to Eq. (5.4) to calculate the log-likelihood function and maximize it by optimization algorithms.

We implemented the computation of $\hat{\beta}_j(\alpha)$ and $SE\{\hat{\beta}_j(\alpha)\}$ in the *compound.Cox* R package (Emura et al. 2018). In the package, the maximization of Eq. (5.4) is performed by the *nlm* function after the log-transformations log $d\Lambda_{0j}(t_i)$ and log $d\Gamma_{0j}(t_i)$. The package uses the initial values $\beta_j = \gamma_j = 0$ and $d\Lambda_{0j}(t_i) = d\Gamma_{0j}(t_i) = 1/n$.

Technical remarks: Theoretically, if $\alpha \downarrow 0$, $\hat{\beta}_j(\alpha)$ approaches to the partial likelihood estimate of β_j. Numerically, however, the value α too close to zero makes the likelihood optimization unstable. Hence, we set $\hat{\beta}_j(\alpha) = \hat{\beta}_j(0.01)$ for $0 \le \alpha < 0.01$ in the package. The value of $\hat{\beta}_j(\alpha) = \hat{\beta}_j(0.01)$ is almost the same as the partial likelihood estimate.

5.4 Copula-Based Univariate Selection

One can use the copula-based method in Sect. 5.3 to perform univariate selection adjusted for the effect of dependent censoring. The P-value for testing the null hypothesis $H_0 : \beta_j = 0$ is computed by the Wald test based on a Z-statistic $\hat{\beta}_j(\alpha)/SE\{\hat{\beta}_j(\alpha)\}$. One can select a subset of genes according to the P-values. With $\alpha \approx 0$ in the Clayton copula, one has $C(u,\ v) \approx uv$. Hence, the resultant test is approximately equal to the Wald test under univariate Cox regression. In this sense, the copula-based test is a generalization of the conventional univariate selection.

For a future subject with a covariate vector $\mathbf{x} = (x_1, \ldots, x_p)'$, survival prediction can be made by the prognostic index (PI) defined as $\hat{\boldsymbol{\beta}}(\alpha)'\mathbf{x}$, where $\hat{\boldsymbol{\beta}}(\alpha)' = (\ \hat{\beta}_1(\alpha),\ \cdots,\ \hat{\beta}_p(\alpha)\ )$. The PI is a weighted sum of genes whose weights reflect the degree of univariate association. If $\alpha = 0$, one obtains PI = $\hat{\boldsymbol{\beta}}(0)'\mathbf{x}$ which is equal to the *compound covariate* based on univariate Cox regression under the independent censoring assumption (Matsui 2006; Emura et al. 2012).

5.5 Choosing the Copula Parameter by the *C*-Index

Estimation of the copula parameter α is inherently difficult due to the non-identifiability of competing risks data (Tsiatis 1975). An estimator maximizing the profile log-likelihood for α based on Eq. (5.4) typically shows very large sampling variation (Chen 2010). In our experience, the profile likelihood often has a peak at extreme values; for instance, either $\alpha \approx 0$ or $\alpha \approx \infty$ under the Clayton copula. These undesirable properties make the likelihood-based strategy less useful.

Following Emura and Chen (2016), we introduce a prediction-based strategy for choosing α. A widely used predictive measure is a cross-validated partial likelihood (Verveij and van Houwelingen 1993). Unfortunately, the partial likelihood is not a valid likelihood under dependent censoring.

A more plausible predictive measure under dependent censoring is Harrell's *c*-index (Harrell et al. 1982). The interpretation of the *c*-index does not depend on a specific model. We adopt a cross-validated version of the *c*-index defined as follows.

We calculate the *c*-index based on a *K*-fold cross-validation. We first divide *n* patients into *K* groups of approximately equal sample sizes. This process can be specified by a function $\kappa : \{1, \ldots, n\} \mapsto \{1, \ldots K\}$ indicating the group to which each patient is allocated (Hastie et al. 2009). For each patient *i*, define the PI:

$$\mathrm{PI}_i(\alpha) = \hat{\boldsymbol{\beta}}'_{-\kappa(i)}(\alpha)\mathbf{x}_i = \ \hat{\beta}_{1,-\kappa(i)}(\alpha)x_{i1} + \ \cdots + \hat{\beta}_{p,-\kappa(i)}(\alpha)x_{ip},$$

where $\hat{\beta}_{j,-\kappa(i)}(\alpha)$ is obtained based on Eq. (5.4) with the $\kappa(i)$th group of patients removed. In this way, $\mathrm{PI}_i(\alpha)$ is a predictor of the survival outcome (t_i, δ_j) for the patient *i*. We define the cross-validated *c*-index:

$$CV(\alpha) = \frac{\sum\limits_{i<j} \{\ \mathbf{I}(\ t_i < t_j\ )\mathbf{I}(\ \mathrm{PI}_i(\alpha) > \mathrm{PI}_j(\alpha)\ )\delta_i + \mathbf{I}(\ t_j < t_i\ )\mathbf{I}(\ \mathrm{PI}_j(\alpha) > \mathrm{PI}_i(\alpha)\ )\delta_j\ \}}{\sum\limits_{i<j} \{\ \mathbf{I}(\ t_i < t_j\ )\delta_i + \mathbf{I}(\ t_j < t_i\ )\delta_j\ \}}.$$

Finally, we define $\hat{\alpha}$ that maximizes $CV(\alpha)$. We recommend $K = 5$ that is often used when *n* or *p* is large.

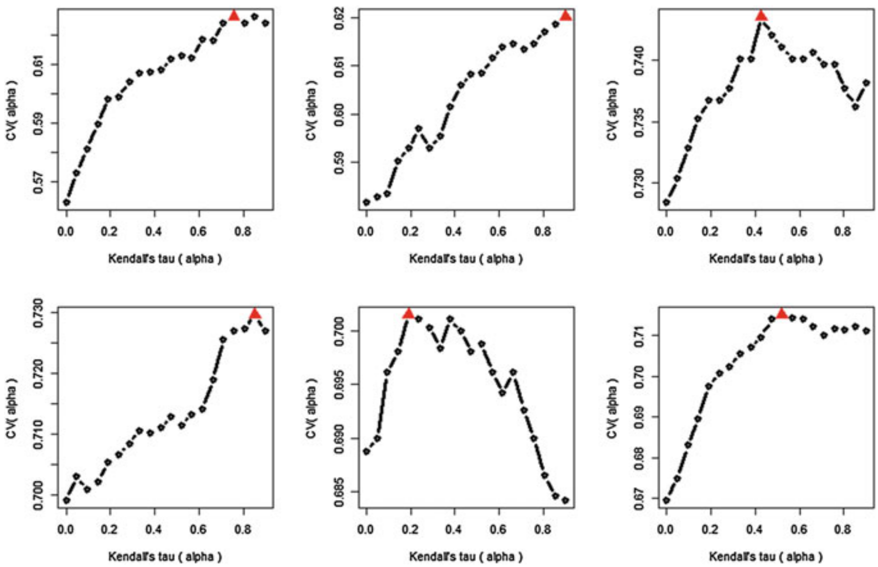


Fig. 5.1 Six replications of the cross-validated *c*-index $CV(\alpha)$. The maximum of $CV(\alpha)$ is signified as a triangle (in red color)

It is computationally demanding to obtain a high-dimensional vector $\hat{\boldsymbol{\beta}}_{-\kappa(i)}(\alpha)$ for every group $\kappa(i)$. To release the computational cost, we suggest reducing the number p by using the initial univariate selection under $\alpha = 0$, e.g., based on P-value <0.2. The technique shall be applied to the subsequent data analysis.

A graphical diagnostic plot for $CV(\alpha)$ is informative to see how the proposed method of choosing $\hat{\alpha}$ works. We suggest using a grid search to find the approximate value of $\hat{\alpha}$ and plot the values of $CV(\alpha)$ against the grids. Figure 5.1 shows the plots of $CV(\alpha)$ with simulated data under our previously considered setting (Case 2 of Table 2 in Emura and Chen 2016). The figure shows that $CV(\hat{\alpha})$ is noticeably larger than $CV(0)$. This suggest that $\mathrm{PI}_i(\hat{\alpha})$ has better ability to predict survival than $\mathrm{PI}_i(0)$ does.

5.6 Lung Cancer Data Analysis

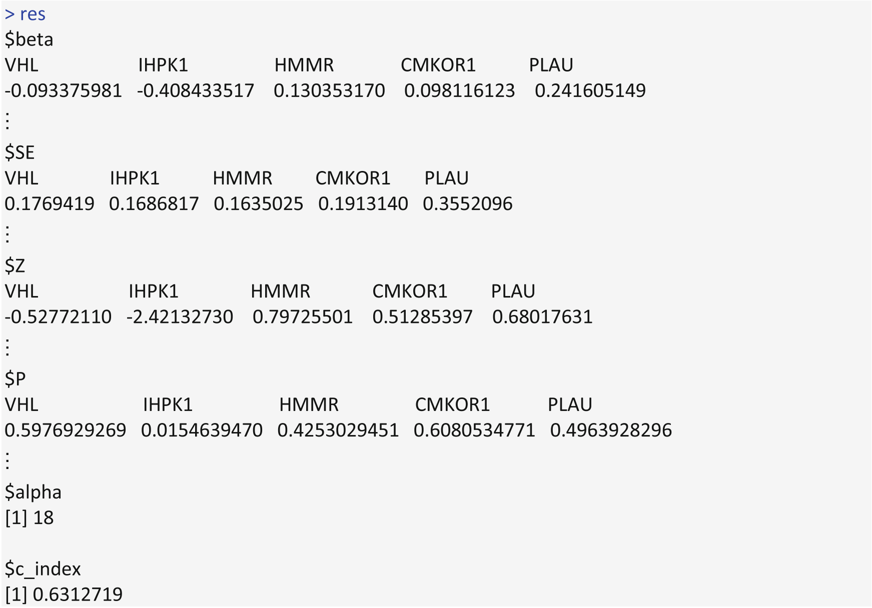
We analyze the survival data on the non-small-cell lung cancer patients of Chen et al. (2007). The data analysis was performed previously by Emura and Chen (2016) using the copula-based methods. Here, we update the analysis based on the data available in the *compound.Cox* R package, providing more detailed explanations than the previous one. In addition, this demonstration allows researchers to reproduce all the results easily through R.

In the lung cancer data, the primary endpoint is overall survival, i.e., time-to-death. During the follow-up, 38 patients died and the remaining 87 patients were censored. The 125 patients were split into either a training set (63 patients) or a testing set (62 patients) in the same manner as Chen et al. (2007).

The *Lung* object in the *compound.Cox* R package contains censored survival times t, censoring indicators δ_i, training/testing indicators, and gene expressions $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ for the 125 patients. Available are $p = 97$ gene expressions that satisfy P-value <0.20 under the usual univariate selection performed on the training set. All the gene expressions were coded as 1, 2, 3, or 4 according to Chen et al. (2007). In the original analysis of Chen et al. (2007), univariate selection yielded 16 genes with P-value <0.05. In our analysis, we shall apply the copula-based univariate selection to select 16 genes.

5.6.1 Gene Selection and Prediction

We applied the copula-based univariate Cox regression to the 63 patients (training set) by using the R codes available in Appendix B. Here, we used $K = 5$ cross-validation for examining the diagnostic plot of $CV(\alpha)$. The outputs are shown below:



```
> res
$beta
VHL           IHPK1         HMMR          CMKOR1        PLAU
-0.093375981  -0.408433517  0.130353170   0.098116123   0.241605149
⋮
$SE
VHL        IHPK1      HMMR       CMKOR1     PLAU
0.1769419  0.1686817  0.1635025  0.1913140  0.3552096
⋮
$Z
VHL          IHPK1        HMMR         CMKOR1       PLAU
-0.52772110  -2.42132730  0.79725501   0.51285397   0.68017631
⋮
$P
VHL           IHPK1         HMMR          CMKOR1        PLAU
0.5976929269  0.0154639470  0.4253029451  0.6080534771  0.4963928296
⋮
$alpha
[1] 18

$c_index
[1] 0.6312719
```

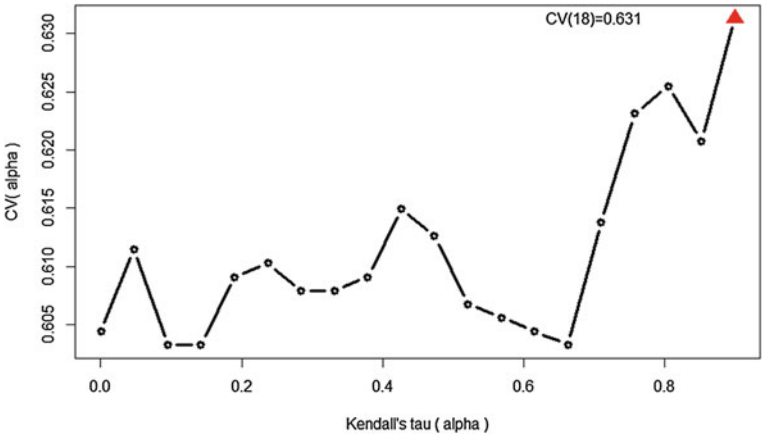


Fig. 5.2 Plot of $CV(\alpha)$ (the cross-validated c-index) based on the lung cancer data. The value of $CV(\alpha)$ is maximized at $\alpha = 18$ (Kendall's tau = 0.90)

Here, \$beta $= \hat{\beta}_j(\hat{\alpha})$, \$SE $= SE\{\ \hat{\beta}_j(\hat{\alpha})\ \}$, \$Z $= \hat{\beta}_j(\hat{\alpha})/SE\{\ \hat{\beta}_j(\hat{\alpha})\ \}$, and \$P is the P-value for each $j = 1,\ \ldots,\ 97$. Also, \$alpha $= \hat{\alpha}$ and \$c_index $= CV(\hat{\alpha})$.

Figure 5.2 displays the diagnostic plot of the cross-validated c-index $CV(\alpha)$ calculated on the 63 patients (training set). The c-index is maximized at the copula parameter $\hat{\alpha} = 18$ (Kendall's tau = 0.90). This implies a possible gain in prediction accuracy by using the Clayton copula for dependent censoring.

We selected the 16 genes among the 97 genes according to the P-values. The outputs are shown below:

```
          Coef    P.value
MMP16     0.51    0.0003
ZNF264    0.51    0.0004
HGF       0.50    0.0010
HCK       -0.49   0.0012
NF1        0.47    0.0016
ERBB3     0.46    0.0016
NR2F6     0.57    0.0030
AXL       0.77    0.0034
CDC23     0.51    0.0051
DLG2      0.92    0.0054
IGF2      -0.34   0.0081
RBBP6     0.54    0.0082
COX11     0.51    0.0116
DUSP6     0.40    0.0122
ENG       -0.37   0.0140
IHPK1     -0.41   0.0155
⋮
```

The resultant PI is defined as PI $= \hat{\beta}_j(\hat{\alpha})x_1 + \cdots + \hat{\beta}_{16}(\hat{\alpha})x_{16}$, where $(x_1, \ldots, x_{16})$ are gene expressions of the 16 genes. Accordingly,

$$\begin{aligned}\mathrm{PI} = {} & (0.51 \times \mathrm{MMP16}) + (0.51 \times \mathrm{ZNF264}) + (0.50 \times \mathrm{HGF}) + (-0.49 \times \mathrm{HCK}) + (0.47 \times \mathrm{NF1}) \\ & + (0.46 \times \mathrm{ERBB3}) + (0.57 \times \mathrm{NR2F6}) + (0.77 \times \mathrm{AXL}) + (0.51 \times \mathrm{CDC23}) + (0.92 \times \mathrm{DLG2}) \\ & + (-0.34 \times \mathrm{IGF2}) + (0.54 \times \mathrm{RBBP6}) + (0.51 \times \mathrm{COX11}) + (0.40 \times \mathrm{DUSP6}) + (-0.37 \times \mathrm{ENG}) \\ & + (-0.41 \times \mathrm{IHPK1}).\end{aligned}$$

5.6.2 Assessing Prediction Performance

To validate the ability of the PI for predicting overall survival, we separate the 62 testing patients into two groups of equal sizes: 31 good prognosis patients with low PIs and 31 poor prognosis patients with high PIs. We then calculate the two survival curves for each group (Fig. 5.3).

The prediction performance of the PI can be measured by the difference between the two survival curves in Fig. 5.3. The two survival curves were calculated by the *copula-graphic estimator* (Rivest and Wells 2001) that adjusts for the effect of dependent censoring with the Clayton copula at $\hat{\alpha} = 18$ (Kendall's tau = 0.90). This approach may be better than the conventional log-rank test to measure the difference between two Kaplan–Meier estimators that are biased under dependent censoring.

Under the Clayton copula model, the copula-graphic (CG) estimator (Chap. 4) is defined as

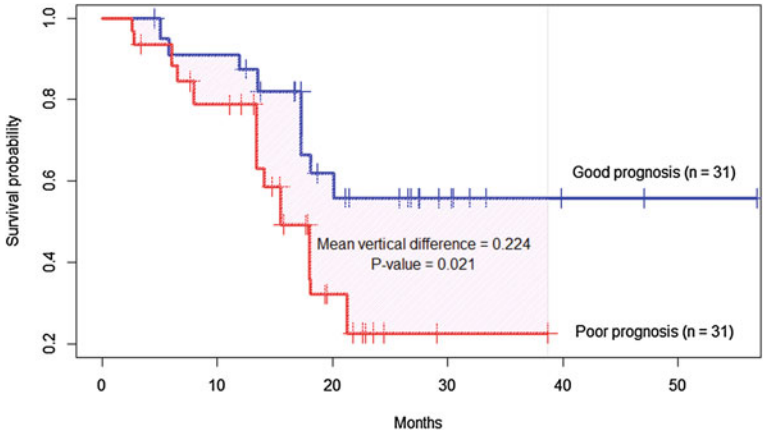


Fig. 5.3 Survival curves for the good and poor prognosis groups. The good (or poor) group is determined by the low (or high) values of the PI. Censored patients are indicated as the mark "+"

$$\hat{S}^{CG}(\,t\,) = \left[1 + \sum_{t_i \le t,\, \delta_i = 1} \left\{ \left(\frac{n_i - 1}{n}\right)^{-\hat{\alpha}} - \left(\frac{n_i}{n}\right)^{-\hat{\alpha}} \right\}\right]^{-1/\hat{\alpha}},$$

where $n_i = \sum_{j=1}^{n} \mathbf{I}(t_j \ge t_i)$ is the number at-risk at time t_i. We computed the CG estimator by using the *compound.Cox* R package (Emura et al. 2018).

The separation of the two curves in Fig. 5.3 is measured by the average vertical difference between the survival curves over the *study period*. This statistic is considered as a scaled version of the area between the two survival curves. It is also equivalent to a special case of the weighted Kaplan–Meier statistics (Pepe and Fleming 1989). When using this statistic, the choice of the study period strongly influences the test results. The common choice is the period where at least one survivor exists in both groups (Chap. 2; Klein and Moeschberger 2003). The study period is depicted in Fig. 5.3.

The P-value for testing the difference between the two groups is obtained using the permutation test (Frankel et al. 2007). In each permutation, good prognosis group (n = 31) and poor prognosis group (n = 31) are randomly allocated from the 62 testing samples, and then, the CG estimator is computed for each group. For each permutation, the study period is determined and the average vertical difference between the two CG estimators is calculated. The P-value is computed as the proportion of 10,000 permuted test statistics exceeding the original test statistic.

The two curves are significantly separated between the good and poor prognoses (Average difference = 0.224; P-value = 0.021). This result justifies the predictive ability of the PI derived by using the copula-based approach.

5.7 Discussions

We have introduced copula-based approaches for selecting genes and making survival prediction in the presence of dependent censoring. The method can be flexibly applied to accommodate different copulas, such as the Clayton, Gumbel, and FGM copulas. Due to its mathematical simplicity, we prefer the Clayton copula to other copulas in modeling dependence structure between survival time and censoring time. However, the effect of dependent censoring on estimates can be remarkably different between different copulas (Chap. 3). Rivest and Wells (2001) theoretically explored the sensitivity of using different copulas on estimating a marginal survival function.

Due to the inherent problem of the non-identifiability of competing risks data (Tsiatis 1975), it is not easy to identify the degree of dependence (i.e., the true copula parameter) between survival and censoring times. The problem is due to the fact that the likelihood function contains little information to identify the true copula parameter. Alternatively, we choose the copula parameter by using a cross-validated c-index, a predictive measure free from the likelihood criterion. This method exhibited sound numerical performances in our numerical analyses. Unfortunately, we do not have a theoretical justification of the method, such as consistency. Recently, Emura and Michimae (2017) proposed a goodness-of-fit procedure to test the assumption of the correct copula under competing risks. According to their simulation results, their approaches have certain ability to identify the correct copula under a large number of samples. However, their approaches have not been extended to include covariates.

After relevant genes are selected, researchers often use them to stratify patients between good and poor prognosis groups in validation samples. This is a common strategy to assess prediction performance of the selected genes. Researchers typically use the log-rank test to see how well the Kaplan–Meier survival curves are separated between the good and poor groups. Note that these commonly used validation strategies may give biased results if dependent censoring exists in validation samples. Copulas are used to adjust for this bias by replacing the Kaplan–Meier estimator by the copula-graphic estimator. Since the log-rank test is no longer valid in the presence of dependent censoring, we apply the permutation test based on the average vertical difference between the copula-graphic estimators. For purpose of constructing survival forests, Moradian et al. (2017) also suggested the copula-graphic estimator to measure the difference between two groups under dependent censoring.

One potential drawback of the proposed gene selection method is that it needs to impose a proportional hazards model for the censoring distribution in Eq. (5.3). On the other hand, the traditional univariate Cox regression does not require any model assumption on the censoring distribution. This elimination of the model assumption is the consequence of the independent censoring assumption. Once the independent censoring assumption is relaxed, certain model specifications for the censoring distribution appear to be mandatory (e.g., Siannis et al. 2005; Chen 2010). If the research interest lies in the effect of genes on both survival time and censoring time, the proportional hazards model for the censoring distribution may provide useful information. For instance, researchers may be interested in selecting genes associated with both disease-specific survival and time-to-death due to other causes as in the competing risks setting (Escarela and Carrière 2003).

References

Alizadeh AA, Gentles AJ, Alencar AJ, Liu CL, Kohrt HE et al (2011) Prediction of survival in diffuse large B-cell lymphoma based on the expression of 2 genes reflecting tumor and microenvironment. Blood 118(5):1350–1358

Beer DG, Kardia SLR, Huang CC, Giordano TJ, Levin AM et al (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 8:816–824

Bøvelstad HM, Nygård S, Storvold HL, Aldrin M, Borgan Ø et al (2007) Predicting survival from microarray data—a comparative study. Bioinformatics 23:2080–2087

Bøvelstad HM, Nygård S, Borgan Ø (2009) Survival prediction from clinico-genomic models-a comparative study. BMC Bioinf 10(1):1

Chen YH (2010) Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. J R Stat Soc Ser B Stat Methodol 72:235–251

Chen HY, Yu SL, Chen CH, Chang GC, Chen CY et al (2007) A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 356:11–20

Emura T, Chen YH, Chen HY (2012). Survival prediction based on compound covariate under Cox proportional hazard models. PLoS One 7(10): e47627, https://doi.org/10.1371/journal.pone.0047627

Emura T, Chen HY, Matsui S, Chen YH (2018). compound.Cox: univariate feature selection and compound covariate for predicting survival, CRAN

Emura T, Chen YH (2016) Gene selection for survival data under dependent censoring, a copula-based approach. Stat Methods Med Res 25(6):2840–2857

Emura T, Michimae H (2017) A copula-based inference to piecewise exponential models under dependent censoring, with application to time to metamorphosis of salamander larvae. Environ Ecol Stat 24(1):151–173

Emura T, Nakatochi M, Matsui S, Michimae H, Rondeau V (2017) Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: meta-analysis with a joint model. Stat Methods Med Res, https://doi.org/10.1177/0962280216688032

Escarela G, Carrière JF (2003) Fitting competing risks with an assumed copula. Stat Methods Med Res 12(4):333–349

Frankel PH, Reid ME, Marshall JR (2007) A permutation test for a weighted Kaplan-Meier estimator with application to the nutritional prevention of cancer trial. Contemp Clin Trial 28:343–347

Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. JAMA 247:2543–2546

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York

Jenssen TK, Kuo WP, Stokke T, Hovig E (2002) Association between gene expressions in breast cancer and patient survival. Hum Genet 111:411–420

Klein JP, Moeschberger ML (2003) Survival analysis techniques for censored and truncated data. Springer, New York

Lossos IS, Czerwinski DK, Alizadeh AA, Wechser MA, Tibshirani R, Botstein D, Levy R (2004) Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. N Engl J Med 350(18):1828–1837

Matsui S (2006) Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. BMC Bioinf 7:156

Matsui S, Simon RM, Qu P, Shaughnessy JD, Barlogie B, Crowley J (2012) Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. Clin Cancer Res 18(21):6065–6073

Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365(9458):488–492

Moradian H, Denis Larocque D, Bellavance F (2017). Survival forests for data with dependent censoring. Stat Methods Med Res, https://doi.org/10.1177/0962280217727314

Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York

Pepe MS, Fleming TR (1989). Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. Biometrics: 497–507

Popple A, Durrant LG, Spendlove I, Scott PRI, Deen S, Ramage JM (2012) The chemokine, CXCL12, is an independent predictor of poor survival in ovarian cancer. Br J Cancer 106:1306–1313

Rivest LP, Wells MT (2001) A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. J Multivar Anal 79:138–155

Sabatier R, Finetti P, Adelaide J, Guille A, Borg JP, Chaffanet M, Bertucci F (2011) Down-regulation of ECRG4, a candidate tumor suppressor gene, in human breast cancer. PLoS One 6(11):e27656

Schumacher M, Binder H, Gerds T (2007) Assessment of survival prediction models based on microarray data. Bioinformatics 23(14):1768–1774

Shedden K, Taylor JMG, Enkemann SA, Tsao MS, Yeatman TJ et al (2008) Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 14:822–827

Siannis F, Copas J, Lu G (2005) Sensitivity analysis for informative censoring in parametric survival models. Biostatistics 6(1):77–91

Sklar A (1959) Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de L'Université de Paris. 8:229–31

Tsiatis A (1975) A nonidentifiability aspect of the problem of competing risks. Proc Natl Acad Sci 72(1):20–22

Tukey JW (1993) Tightening the clinical trial. Control Clin Trials 14:266–285

Yoshihara K, Tajima A, Yahata T, Kodama S, Fujiwara H et al (2010) Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. PLoS One 5(3):e9615

Yoshihara K, Tsunoda T, Shigemizu D, Fujiwara H, Hatae M et al (2012) High-risk ovarian cancer based on 126-gene expression signature is uniquely characterized by downregulation of antigen presentation pathway. Clin Cancer Res 18(5):1374–1385

van Wieringen WN, Kun D, Hampel R, Boulesteix AL (2009) Survival prediction using gene expression data: a review and comparison. Comput Stat Data Anal 53(5):1590–1603

Verveij PJM, van Houwelingen HC (1993) Crossvalidation in survival analysis. Stat Med 12:2305–2314

Waldron L, Haibe-Kains B, Culhane AC, Riester M, Ding J et al. (2014) Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. J Natl Cancer Inst 106(5): dju049

Wang Y, Klijn JG, Zhang Y, Sieuwerts AM et al (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365(9460):671–679

Witten DM, Tibshirani R (2010) Survival analysis with high-dimensional covariates. Stat Methods Med Res 19(1):29–51

Zhao X, Rødland EA, Sørlie T, Naume B, Langerød A et al (2011) Combining gene signatures improves prediction of breast cancer survival. PLoS One 6(3):e17845

Zhao SD, Parmigiani G, Huttenhower C, Waldron L (2014) Más-o-menos: a simple sign averaging method for discrimination in genomic data analysis. Bioinformatics 30(21):3062–3069