SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2016, Vol. 25(6) 2840–2857
© The Author(s) 2014
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280214533378
smm.sagepub.com
SAGE

Gene selection for survival data under dependent censoring: A copula-based approach

Takeshi Emura[1] and Yi-Hau Chen[2]

Abstract

Dependent censoring arises in biomedical studies when the survival outcome of interest is censored by competing risks. In survival data with microarray gene expressions, gene selection based on the univariate Cox regression analyses has been used extensively in medical research, which however, is only valid under the independent censoring assumption. In this paper, we first consider a copula-based framework to investigate the bias caused by dependent censoring on gene selection. Then, we utilize the copula-based dependence model to develop an alternative gene selection procedure. Simulations show that the proposed procedure adjusts for the effect of dependent censoring and thus outperforms the existing method when dependent censoring is indeed present. The non-small-cell lung cancer data are analyzed to demonstrate the usefulness of our proposal. We implemented the proposed method in an R "compound.Cox" package.

Keywords

bivariate survival distribution, competing risk, compound covariate prediction, Cox regression, cross-validation, Kendall's tau

1 Introduction

For survival data with microarrays, the primary task is selecting a small fraction of genes that are relevant to survival. To handle the censoring that is ubiquitous in survival data, most available approaches use Cox regression analysis[1] to select relevant genes. The simplest approach is to select subsets of genes by using univariate Cox regression analyses.[2–5] This approach is called univariate selection and used extensively in medical research. A predictor based on the linear combinations of the selected genes, often called the compound covariate predictor,[6,7] has been shown to be useful for survival prediction with high-dimensional settings of microarrays.[3–5,8,9]

The aforementioned univariate selection critically relies on the independent censoring assumption; survival time and censoring time need to be statistically independent at a given gene. As further elaborated in Section 2.2, such an independence assumption in univariate analysis is even more stringent than its counterpart in multivariate analysis.

[1]Graduate Institute of Statistics, National Central University, Jhongli, Taiwan
[2]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Corresponding author:
Yi-Hau Chen, Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan.
Email: yhchen@stat.sinica.edu.tw

If the independent censoring assumption is violated, univariate Cox regression analyses may not correctly identify the effect of each gene and thus may fail to select truly effective genes. In the presence of dependent censoring, the univariate Cox regression may instead identify the effective genes on the cause-specific hazard for the survival outcome.[10] However, the effect of a gene on the cause-specific hazard may not reflect well the effect of the same gene on the cumulative incidence, a typical phenomenon in the competing risks literature.[10–13] Therefore, the resultant predictor that uses the selected genes may have reduced ability to predict survival outcomes of interest.

In the presence of dependent censoring, a natural approach is to select genes that influence the cumulative incidence function. For low-dimensional settings, this approach is implemented by fitting the Cox proportional hazards model on the subdistribution hazard, the hazard for the cumulative incidence function.[11] Adapted to the high dimensionality of microarrays, Binder et al.[12] propose a boosting algorithm under the proportional subdistribution hazards model, which provides a short list of relevant genes. Another approach for high dimensionality is to perform a random forest algorithm for the competing risks data after imputing unknown event status for censored individuals.[14]

In this paper, we follow a different approach that selects relevant genes on the marginal survival function, where the nuisance aspects of dependent censoring are removed. For the marginal survival to be identifiable, it is necessary to specify either the dependence structure (i.e. copula[15]) between the survival and censoring times[16] or the marginal regression models[17] (e.g. proportional hazard models). Unfortunately, there are still no practical method for simultaneously estimating the dependence structure and the marginal regression models. So far, statistical inference for the marginal models relies on the sensitivity analysis under an assumed copula.[18,19] Despite the technical difficulty, the major attraction of the present approach is to offer a way for selecting genes that are predictive of a well-defined survival endpoint free of the nuisance aspects. The predictive values of the selected genes are simple to interpret within the framework of traditional survival analysis. In the following, we propose a gene selection method that fits a copula model for the dependence structure and fits the proportional hazards model on the marginal survival based on the method of Chen.[19] We also propose a novel approach to estimate the dependence parameter by using cross-validation, which is useful for both gene selection and survival prediction. We choose the copula-based approach since it not only gives a practical framework for both gene selection and prediction, but also an analytical tool to investigate the bias of univariate selection under dependent censoring.

In Section 2, we review univariate selection and study the potential bias of univariate selection due to dependent censoring. To study that bias, we specifically model the dependency between survival time and censoring time via copulas.[15] In Section 3, we consider a new gene selection method that adjusts for dependent censoring using the copula model. Section 4 compares the performance of univariate selection with the new method via simulations. Section 5 includes the analysis of the non-small-cell lung cancer data for illustration. Section 6 concludes the paper.

2 Univariate selection under dependent censoring

2.1 Univariate selection for censored survival data

The approach called *univariate selection* is performed using the following procedure. As the initial step, a univariate Cox regression is performed for each gene, one by one. Then a subset of genes that have low P-values is selected from the univariate analysis.

More specifically, let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ be a p-vector of genes from individual i. Also, let T_i and U_i be survival time and censoring time, respectively. We observe $(t_i, \delta_i, \mathbf{x}_i)$, where $t_i = \min\{T_i, U_i\}$

and $\delta_i = \mathbf{I}\{T_i \le U_i\}$, where $\mathbf{I}\{\cdot\}$ is the indicator function. In univariate selection, a Cox regression[1] based on the univariate models

$$h(t|x_{ij}) = h_{0j}(t)e^{\beta_j x_{ij}}, \quad j = 1, \ldots, p \tag{1}$$

is performed one by one for each j. The resultant estimator $\hat{\beta}_j$ is used to obtain the P-value for the Wald test for $H_{oj}: \ \beta_j = 0$. One selects genes that exhibit smaller P-values than a threshold value that can be determined by various different criteria, such as cross-validation[3,20] and false discovery rate,[21] which are often complemented by biological consideration.

The estimator $\hat{\beta}_j$ can correctly identify the true value of β_j under the so-called independent censoring assumption[22,23]:

Assumption I: The survival time T and censoring time U are conditionally independent given a gene x_j for all $j = 1, \ldots, p$.

Even when model (1) is incorrect, the univariate estimate $\hat{\beta}_j$ still possesses a valid meaning under Assumption I. To understand why, we consider dichotomous covariates with $x_{ij} = 0$ or 1. It follows that

$$\hat{\beta}_j = \log \frac{\int W_{\hat{\beta}_j}(t) d\bar{N}_1(t)/\bar{Y}_1(t)}{\int W_{\hat{\beta}_j}(t) d\bar{N}_0(t)/\bar{Y}_0(t)}$$

where $\bar{N}_\ell(t) = \sum_{i=1}^n \mathbf{I}(t_i \le t, \delta_i = 1, x_{ij} = \ell)$, $\bar{Y}_\ell(t) = \sum_{i=1}^n \mathbf{I}(t_i \ge t, x_{ij} = \ell)$ for $\ell = 0$, 1, and $W_{\beta_j}(u) = \bar{Y}_1(u)\bar{Y}_0(u)/\{e^{\beta_j}\bar{Y}_1(u) + \bar{Y}_0(u)\}$. This implies that $\hat{\beta}_j$ is the log of the cumulative observed hazard rate for those with $x_{ij} = 1$ relative to that for $x_{ij} = 0$. If Assumption I is valid, the underlying (net) hazard $h(t|x_j = \ell)dt = \Pr(t \le t_j < t + dt | t_j \ge t, x_j = \ell)$ can be correctly estimated by $d\bar{N}_\ell(t)/\bar{Y}_\ell(t)$. Hence, the statistic $\hat{\beta}_j$ still makes sense as the univariate effect of $x_j = 1$ over $x_j = 0$ under Assumption I. This interpretation is irrelevant to the correctness of the model assumption (1). For continuous covariates, we refer to the results of the misspecified Cox model.[9,24]

Thus, it should be stressed that Assumption I is even more important than the model assumption (1) in applying univariate selection.

2.2 Models for dependent censoring

Assumption I will be shown to be a fairly strong assumption. A more reasonable assumption is *the conditional independence* in which T and U are conditionally independent given all components of $\mathbf{x}$, which is routinely imposed for Cox regression models (e.g. Section VII.2 of Andersen et al.[22]; Section 8.4 of Fleming and Harrington[23]).

Figure 1 shows an example of how Assumption I fails to hold, yet the conditional independence still holds. Suppose that two genes, $x_1, x_2 \in \mathbf{x}$, relate to both T and U. Then, T and U are not conditionally independent given only x_1, since the variation in x_2 induces the observed dependence. Here, the variant x_2 can be interpreted as a frailty, a popular concept to construct bivariate survival models.[25] This example implies that Assumption I may not hold when $\mathbf{x}$ relates to both T and U.

The violation of Assumption I is seen by using a more formal argument based on copulas. Specifically, suppose that T and U are conditionally independent given $\mathbf{x}$ and their marginal

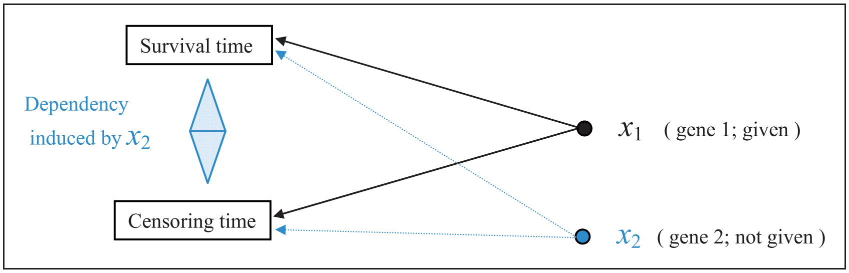

Figure 1. An example that T and U are not conditionally independent given only x_1. The dependency between T and U is induced by x_2, which affects both T and U.

cumulative hazard functions are $e^{\boldsymbol{\beta}'\mathbf{x}}\Lambda_T(t)$ and $e^{\boldsymbol{\gamma}'\mathbf{x}}\Lambda_U(u)$, respectively. It follows that, for a given x_j

$$\Pr(T > t, U > u|x_j) = \varphi_{\boldsymbol{\beta}(-j),\boldsymbol{\gamma}(-j)}\left[\varphi^{-1}_{\boldsymbol{\beta}(-j)}\{\Pr(T > t|x_j)\}, \varphi^{-1}_{\boldsymbol{\gamma}(-j)}\{\Pr(U > u|x_j)\}\right] \tag{2}$$

where $\varphi_{\boldsymbol{\beta}(-j),\boldsymbol{\gamma}(-j)}$, $\varphi_{\boldsymbol{\beta}(-j)}$, and $\varphi_{\boldsymbol{\gamma}(-j)}$ are Laplace transforms defined in Appendix A of the Supplementary Materials and $\boldsymbol{\beta}(-j)$ is $\boldsymbol{\beta}$ excluding β_j with $\boldsymbol{\gamma}(-j)$ similarly defined. For the special case where $\boldsymbol{\beta} = \boldsymbol{\gamma}$, we obtain an Archimedean copula model

$$\Pr(T > t, U > u|x_j) = \varphi_{\boldsymbol{\beta}(-j)}\left[\varphi^{-1}_{\boldsymbol{\beta}(-j)}\{\Pr(T > t|x_j)\} + \varphi^{-1}_{\boldsymbol{\beta}(-j)}\{\Pr(U > u|x_j)\}\right]. \tag{3}$$

The above analysis indicates that the conditional independence yields dependency between T and U given only x_j, and thus Assumption I does not hold.

In general, T and U may be dependent for any given x_j with an unknown dependence structure. Sklar's theorem (Nelsen,[15] p. 18) guarantees that the joint survival function is always written as

$$\Pr(T > t, U > u|x_j) = C_j\{\Pr(T > t|x_j), \Pr(U > u|x_j)\}$$

where C_j is called copula and describes the dependency between T and U. Assumption I corresponds to $C_j(u, v) = uv$. This is clearly a strong assumption in light of equations (2) and (3). Although the form of C_j is fairly difficult to specify, we consider applying certain parametric copulas to relax Assumption I.

2.3 Effect of dependent censoring

Performing univariate selection under dependent censoring may lead to a bias in estimation and hence the inability to select genes of interest. Here, we provide an analytic framework to study the bias when the dependence is modeled via copulas.

The cause-specific hazard

$$h^{\#}(t|x_j) = \Pr(t \leq T < t + dt, T \leq U|T \geq t, U \geq u, x_j)/dt$$

describes the "apparent" hazard rate for death in the presence of dependent censoring (Kalbfleisch and Prentice,[26] p. 251). If Assumption I holds, then

$$h^{\#}(t|x_j) = h(t|x_j) \equiv \Pr(t \leq T < t + dt, |T \geq t, x_j)/dt$$

Otherwise, $h^{\#}(t|x_j)$ and $h(t|x_j)$ are usually different. This implies that the data with dependent censoring give misleading information about the underlying (net) hazard $h(t|x_j)$.

We formulate the effect of dependent censoring under the copula models as

$$\Pr(T > t, U > u|x_j) = C_\alpha\{S_T(t|x_j), S_U(u|x_j)\}$$

where $S_T(t|x_j) = \Pr(T > t|x_j)$ and $S_U(u|x_j) = \Pr(U > u|x_j)$ are the marginal survival functions, and α is the dependence parameter. As indicated in Rivest and Wells,[18] the cause-specific hazard becomes $h^{\#}_\alpha(t|x_j) = r_\alpha(t|x_j)h(t|x_j)$, where

$$r_\alpha(t|x_j) = \frac{C^{[1,0]}_\alpha\{S_T(t|x_j), S_U(t|x_j)\}S_T(t|x_j)}{C_\alpha\{S_T(t|x_j), S_U(t|x_j)\}}$$

and $C^{[1,0]}_\alpha(u, v) = \partial C_\alpha(u, v)/\partial u$. We define the "apparent effect" of gene x_j as

$$\beta^{\#}_\alpha(t) \equiv \log\frac{h^{\#}_\alpha(t|x_j = 1)}{h^{\#}_\alpha(t|x_j = 0)} = \log\frac{h(t|x_j = 1)}{h(t|x_j = 0)} + \log\frac{r_\alpha(t|x_j = 1)}{r_\alpha(t|x_j = 0)}$$

This equation shows that the apparent effect can be partitioned into the true (net) effect and the bias due to dependent censoring. Here, the copula structure only enters in the bias term. The bias vanishes if $\alpha = 0$, the value leading to $C_\alpha(u, v) = uv$. If $\alpha \neq 0$, then the bias is usually nonzero except for some special copulas.

We conducted numerical analysis to gain insight into how dependent censoring affects the apparent effect $\beta^{\#}_\alpha(t)$ under the Clayton, Frank, and Gumbel copulas. We set marginal distributions as $S_T(t|x_j) = S_T(t|0)^{\exp(\beta_j)}$, $S_U(t|x_j) = S_U(t|0)^{\exp(\beta_j)}$, and $S_U(t\,|0) = S_T(t\,|0)^{p_C/(1-p_C)}$, where $p_C \times 100$ (%) is the censoring percentage, using 0, 40, 50, and 60 percentiles, and t is fixed by setting $S_T(t\,|0) = 0.5$.

Figure 2 displays the apparent effect $\beta^{\#}_\alpha(t)$ under the Clayton copula model. If the censoring probability is high (60%), the apparent effect differs significantly from the true (net) effect. Furthermore, the difference inflates as the association parameter α deviates from zero. The signs of the apparent and true effects are even different for $\alpha \geq 2$. For censoring probability 40 or 50%, the apparent effect is still different from the true effect, but the difference becomes more modest. The apparent effect is identical to the true effect when no censoring is present (0%).

Figure 3 displays the apparent effect $\beta^{\#}_\alpha(t)$ under the Frank copula model. Overall, the characteristics of $\beta^{\#}_\alpha(t)$ are similar to those under the Clayton copula. The difference between the apparent and true effects is remarkably large under high censoring probability (60%) but mild under moderate (50%) or low censoring (40%) probability.

The characteristics of the Gumbel copula are completely different from the Clayton and Frank copulas. There is no difference between the apparent and the true effects under any censoring probability or association parameter. Why this result occurs is not completely clear at this stage, but such a property seems to be a rather special circumstance.

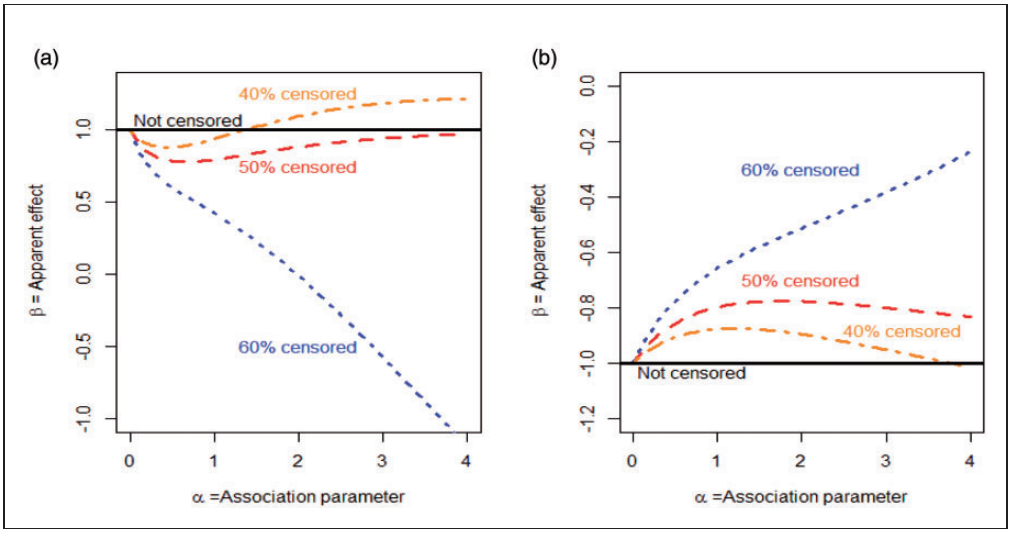

Figure 2. The plots for the apparent effect $\beta^{\#}_{\alpha}(t)$ against the association parameter α under the Clayton copula model with censored percentages 0, 40, 50, and 60%. (a) Net effect =1, (b) net effect = −1.

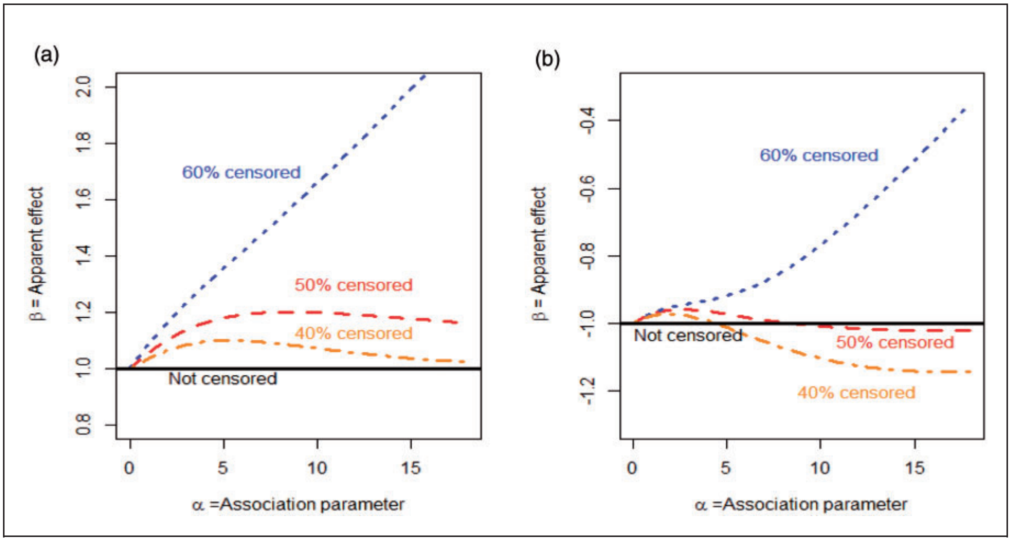

Figure 3. The plots for the apparent covariate effect $\beta^{\#}_{\alpha}(t)$ against the association parameter α under the Frank copula model with censored percentages 0, 40, 50, and 60%. (a) Net effect =1, (b) net effect = −1.

3 Gene selection method under dependent censoring

3.1 Copula-based adjustment for dependent censoring

We propose adjusting for the effect of dependent censoring by modeling the dependency with a given copula.[16,18,19,27,28] Specifically, we impose a copula model

$$\Pr\left(T_i > t, U_i > u|x_{ij}\right) = C_\alpha\left\{\Pr\left(T_i > t|x_{ij}\right), \Pr\left(U_i > u|x_{ij}\right)\right\}$$

where a copula C_α is assumed to be the same across all j and indexed by a single parameter α. The most convenient example is the Clayton copula

$$C_\alpha(u, v) = \left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}, \quad \alpha \geq 0$$

In this way, Assumption I is relaxed by free parameter α. For marginal distributions, we assume the proportional hazard models

$$\Pr\left(T_i > t|x_{ij}\right) = \exp\left\{-\Lambda_{0j}(t)e^{\beta_j x_{ij}}\right\}, \quad \Pr\left(U_i > u|x_{ij}\right) = \exp\left\{-\Gamma_{0j}(u)e^{\gamma_j x_{ij}}\right\}$$

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

For estimation, we apply the semiparametric maximum likelihood estimator in which Λ_{0j} and Γ_{0j} are unspecified.[19] For any given α, we maximize the full likelihood

$$\begin{aligned}\ell\left(\beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j}|\alpha\right) = {} & \sum_i \delta_i\left[\beta_j x_{ij} + \log\eta_{1ij}\left(t_i; \beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j}|\alpha\right) + \log d\Lambda_{0j}(t_i)\right] \\ & + \sum_i (1-\delta_i)\left[\gamma_j x_{ij} + \log\eta_{2ij}\left(t_i; \beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j}|\alpha\right) + \log d\Gamma_{0j}(t_i)\right] \\ & - \sum_i \Phi_\alpha\left[\exp\left\{-\Lambda_{0j}(t_i)e^{\beta_j x_{ij}}\right\}, \exp\left\{-\Gamma_{0j}(t_i)e^{\gamma_j x_{ij}}\right\}\right]\end{aligned} \tag{4}$$

where

$$\begin{aligned}\eta_{1ij}\left(t; \beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j}|\alpha\right) &= D_{\alpha,1}\left[\exp\left\{-\Lambda_{0j}(t)e^{\beta_j x_{ij}}\right\}, \exp\left\{-\Gamma_{0j}(t)e^{\gamma_j x_{ij}}\right\}\right]\exp\left\{-\Lambda_{0j}(t)e^{\beta_j x_{ij}}\right\} \\ \eta_{2ij}\left(t; \beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j}|\alpha\right) &= D_{\alpha,2}\left[\exp\left\{-\Lambda_{0j}(t)e^{\beta_j x_{ij}}\right\}, \exp\left\{-\Gamma_{0j}(t)e^{\gamma_j x_{ij}}\right\}\right]\exp\left\{-\Gamma_{0j}(t)e^{\gamma_j x_{ij}}\right\} \\ \Phi_\alpha &= -\log C_\alpha, \text{ and } D_{\alpha,k}(u_1, u_2) = -\partial\Phi_\alpha(u_1, u_2)/\partial u_k \quad \text{for } k = 1, 2\end{aligned}$$

The maximizer of equation (4) with respect to β_j is denoted by $\hat{\beta}_j(\alpha)$. The standard error $se\{\hat{\beta}_j(\alpha)\}$ can be computed from the observed information matrix.[19] We implement the computation in an R compound.Cox package.[29] Computational details under the Clayton copula are given in Appendix B of the Supplementary Materials.

The P-value for gene j is computed by the Wald test based on a Z-statistic $\hat{\beta}_j(\alpha)/sd\{\hat{\beta}_j(\alpha)\}$. If $\alpha = 0$, a value that leads to $C_\alpha(u, v) = uv$, the resultant P-value is the same as the P-value from univariate Cox analysis.

For a future subject with covariate vector $\mathbf{x} = (x_1, \ldots, x_p)'$, the survival prediction can be made by the prognostic index (PI) defined as $\hat{\boldsymbol{\beta}}(\alpha)'\mathbf{x}$, where $\hat{\boldsymbol{\beta}}(\alpha)' = (\hat{\beta}_1(\alpha), \ldots, \hat{\beta}_p(\alpha))$. If $\alpha = 0$, the resultant prediction method is equal to the compound covariate prediction.[3,9]

3.2 Choosing association parameters by cross-validation

Due to the nonidentifiability of competing risks data,[30] likelihood (4) may provide little information about the true parameter α. A more practical approach is to choose an α that maximizes prediction power. A widely used predictive measure for this purpose is cross-validated partial likelihood.[31] Unfortunately, the form of Cox's partial likelihood is derived under independent censoring, which renders it unsuitable for dependent censoring.

A more robust predictive measure under dependent censoring is Harrell's concordance measure known as c-index.[32,33] The interpretation of c-index does not depend on a specific model. To perform a K-fold cross-validation, we first divide n individuals into K groups of approximately equal sample sizes and label them as $\Im_k$ for $k = 1, \ldots, K$. The estimator based on all individuals not in $\Im_k$ is calculated and denoted by $\hat{\boldsymbol{\beta}}_{(-k)}(\alpha)$. For a subject $i \in \Im_k$, we consider $\mathrm{PI}_i(\alpha) = \hat{\boldsymbol{\beta}}'_{(-k)}(\alpha)x_i$, a predictor of the survival outcome (t_i, δ_i). We choose an α that maximizes the cross-validated c-index

$$CV(\alpha) = \frac{\sum_{i<j}\left\{\mathbf{I}\big(t_i < t_j\big)\mathbf{I}\big(\mathrm{PI}_i(\alpha) > \mathrm{PI}_j(\alpha)\big)\delta_i + \mathbf{I}\big(t_j < t_i\big)\mathbf{I}\big(\mathrm{PI}_j(\alpha) > \mathrm{PI}_i(\alpha)\big)\delta_j\right\}}{\sum_{i<j}\left\{\mathbf{I}\big(t_i < t_j\big)\delta_i + \mathbf{I}\big(t_j < t_i\big)\delta_j\right\}}$$

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

The cross-validation curve may be used as a heuristic way to test the presence of dependent censoring. The subsequent simulations will show that, if dependent censoring is present, then the curve may have an $\hat{\alpha}$ value that is far from $\alpha = 0$. We will also demonstrate this method using data analysis.

4 Simulations

We compare the performance of the two gene selection strategies, namely univariate selection and the proposed method in Section 3, in the presence of dependent censoring.

4.1 Simulation set-up

We generate $n = 100$ random samples (T_i, U_i) either from the Clayton copula model

$$\Pr(T_i > t, U_i > u|\mathbf{x}_i) = \left(\exp\left\{-te^{\boldsymbol{\beta}'\mathbf{x}_i}\right\}^{-\alpha} + \exp\left\{-ue^{\boldsymbol{\gamma}'\mathbf{x}_i}\right\}^{-\alpha} - 1\right)^{-1/\alpha}, \quad \alpha \geq 0 \tag{5}$$

or the Frank copula model

$$\Pr(T_i > t, U_i > u|\mathbf{x}_i) = \log_\alpha\left\{1 + \left(\alpha^{\exp(-te^{\boldsymbol{\beta}'\mathbf{x}_i})} - 1\right)\left(\alpha^{\exp(-ue^{\boldsymbol{\gamma}'\mathbf{x}_i})} - 1\right)/(\alpha - 1)\right\}, \quad \alpha \geq 0 \tag{6}$$

We choose α so that Kendall's τ between T_i and U_i given $\mathbf{x}_i$ is $\tau = 0.5$. Let $\boldsymbol{\beta}' = \big(\beta_1, \ldots, \beta_q, \beta_{q+1}, \ldots, \beta_p\big)$, where the first q =5, 10, or 20 genes are related to survival among the $p = 100$ coefficients; the coefficients of the first q genes are nonzero (nonnull genes) and those of the remaining p–q genes are zero (null genes). We introduce the blocks of correlated genes $\mathbf{x}' = \big(x_1, \ldots, x_q, x_{q+1}, \ldots, x_p\big)$ by mimicking the microarray structure of "tag gene sequence" and "gene pathway" as in Binder et al.[12] and Emura et al.[9] More details for generating $\mathbf{x}$ are given in Appendix C of the Supplementary Materials. We consider the model that nonnull genes influence both T_i and U_i by setting $\boldsymbol{\beta} = \boldsymbol{\gamma}$. Under this setting, there are approximately 50% censored samples.

For each gene j, we obtained the univariate estimator $\hat{\beta}_j$ and the proposed estimator $\hat{\beta}_j(\hat{\alpha})$ calculated under the Clayton copula model. The fitted Clayton copula is misspecified since the true models in equations (5) and (6) involve multiple genes while the fitted model involves only a single gene at a time. Here, $\hat{\alpha}$ is calculated from the $K = 5$ cross-validation curve on the grid $\alpha \in$ (0, 0.5, 1.33, 3, 8), which corresponds to Kendall's tau (0, 0.2, 0.4, 0.6, 0.8). Then, the P-value for each gene is computed with the Wald test.

We compare the performance of gene selection in terms of sensitivity and specificity. Let $(P_1, \ldots, P_p)$ be a vector of P-values obtained by a gene selection method (univariate selection or proposed method) and let $P_{(c)}$ be the c^{th} smallest P-value. Then

$$\text{Sensitivity} = \frac{\sum_{j=1}^{p} \mathbf{I}\left(P_j \le P_{(q)}, \beta_j \neq 0\right)}{\sum_{j=1}^{p} \mathbf{I}\left(\beta_j \neq 0\right)} \times 100(\%)$$

is the percentage of selecting truly informative genes while

$$\text{Specificity} = \frac{\sum_{j=1}^{p} \mathbf{I}\left(P_j > P_{(q)}, \beta_j = 0\right)}{\sum_{j=1}^{p} \mathbf{I}\left(\beta_j = 0\right)} \times 100(\%)$$

is the percentage of not selecting uninformative genes. Larger values of sensitivity and specificity correspond to better gene selection ability. We report the results in terms of the average of 50 Monte Carlo replications.

We also compare the first component of the regression estimates, namely $\hat{\beta}_1$ and $\hat{\beta}_1(\hat{\alpha})$. In particular, we calculate their mean and standard deviation to assess the performance of the Wald test.

4.2 Simulation results

Table 1 summarizes gene selection performance under the tag gene sequence. In terms of sensitivity and specificity, the proposed method outperforms univariate selection in all configurations. The proposed method improves sensitivity by up to 12.8–13.2% when the number of nonzero coefficients is small ($q = 5$) and 4.7–7.9% when the number of nonzero coefficients is large ($q = 20$). The columns of $E[\hat{\beta}_1]$ show that the proposed estimates of the nonzero β_1 tend to be larger than the estimates of univariate selection. Also, the standard deviations of the proposed estimates are smaller than that of the univariate estimates. This explains the improved power of the Wald test in our proposal. The columns of $E[\hat{\alpha}]$ vary between 3.9 and 4.8, which corresponds to Kendall's tau between 0.66 and 0.71. Hence, on average, the proposed method fits the Clayton copula with strongly positive association.

Note that the fitted Clayton model is misspecified under the present simulation setups. As inferred from Section 2.2, the true copula structure is fairly complicated for a given x_{ij} only, which is difficult to specify in practice. In addition, the Cox proportional hazard models may not hold for a given x_{ij} only. Hence, one cannot expect that $E[\hat{\beta}_1]$ and $E[\hat{\alpha}]$ are close to the true value in equations (5) and (6). Nevertheless, the simulation results exhibit the good performance of the proposed method in terms of selection of true nonnull genes. This robustness to the model misspecification is important since the strong model assumptions are often controversial in the marginal approaches.

Table 2 summarizes gene selection performance under the gene pathway. Similar to Table 1, the proposed method produces higher sensitivity and specificity than those of univariate selection in all configurations. The improvement in sensitivity becomes more evident (up to 12.1–16.6%) when both

Table 1. Comparison of univariate selection and the proposed method based on $n = 100$ samples and 50 replications with tag gene sequences.

Case 1: $\boldsymbol{\beta} = (\underbrace{0.8, \ldots, 0.8}_{\times 5}, \underbrace{0, \ldots, 0}_{\times 95})$; $s = 4$, $\beta_1 = 0.8$

	Underlying model	Sensitivity % (Specificity %)	$E[\hat{\beta}_1]$ (±SD)	$E[\hat{\alpha}]$
Univariate selection	Clayton	47.60 (97.24)	0.36 (±0.15)	/
	Frank	49.60 (97.35)	0.38 (±0.17)	/
Proposed method	Clayton	60.80 (97.94)	0.41 (±0.13)	4.0
	Frank	62.40 (98.02)	0.43 (±0.14)	4.0

Case 2: $\boldsymbol{\beta} = (\underbrace{0.4, \ldots, 0.4}_{\times 5}, \underbrace{-0.4, \ldots, -0.4}_{\times 5}, \underbrace{0, \ldots, 0}_{\times 90})$; $s = 4$, $\beta_1 = 0.4$

	Underlying model	Sensitivity % (Specificity %)	$E[\hat{\beta}_1]$ (±SD)	$E[\hat{\alpha}]$
Univariate selection	Clayton	33.80 (92.64)	0.25 (±0.18)	/
	Frank	34.20 (92.69)	0.27 (±0.17)	/
Proposed method	Clayton	39.60 (93.29)	0.28 (±0.16)	4.0
	Frank	41.20 (93.47)	0.28 (±0.15)	4.3

Case 3: $\boldsymbol{\beta} = (\underbrace{0.4, \ldots, 0.4}_{\times 10}, \underbrace{0, \ldots, 0}_{\times 90})$; $s = 2$, $\beta_1 = 0.4$

	Underlying model	Sensitivity % (Specificity %)	$E[\hat{\beta}_1]$ (±SD)	$E[\hat{\alpha}]$
Univariate selection	Clayton	32.80 (92.53)	0.23 (±0.17)	/
	Frank	36.40 (92.93)	0.25 (±0.17)	/
Proposed method	Clayton	42.80 (93.64)	0.26 (±0.13)	4.5
	Frank	44.00 (93.78)	0.27 (±0.14)	4.5

Case 4: $\boldsymbol{\beta} = (\underbrace{0.2, \ldots, 0.2}_{\times 10}, \underbrace{-0.2, \ldots, -0.2}_{\times 10}, \underbrace{0, \ldots, 0}_{\times 80})$; $s = 2$, $\beta_1 = 0.2$

	Underlying model	Sensitivity % (Specificity %)	$E[\hat{\beta}_1]$ (±SD)	$E[\hat{\alpha}]$
Univariate selection	Clayton	30.60 (82.65)	0.11 (±0.17)	/
	Frank	30.20 (82.55)	0.12 (±0.17)	/
Proposed method	Clayton	35.30 (83.83)	0.13 (±0.15)	3.9
	Frank	38.10 (84.53)	0.13 (±0.16)	4.8

Higher sensitivity and specificity correspond to better gene selection performance.

the positive and negative coefficients exist. The proposed method enjoys the higher power of the Wald test, as implied from the large magnitude of $E[\hat{\beta}_1]$. Compared to the tag gene sequence of Table 1, sensitivity and specificity in Table 2 substantially increases under the gene pathway for both univariate selection and the proposed method. This is due to the presence of a positively correlated blocks of genes among nonzero coefficients, which enlarge the regression estimates $\hat{\beta}_1$ and hence the power.

Table 2. Comparison of univariate selection and the proposed method based on $n = 100$ samples and 50 replications with gene pathways.

Case 1: $\boldsymbol{\beta} = (\underbrace{0.4,\ldots,0.4}_{\times 5}, \underbrace{0,\ldots,0}_{\times 95}), \beta_1 = 0.4$

	Underlying model	Sensitivity % (Specificity %)	$E[\hat{\beta}_1]$ (±SD)	$E[\hat{\alpha}]$
Univariate selection	Clayton	98.40 (99.92)	0.75 (±0.15)	/
	Frank	98.80 (99.94)	0.81 (±0.15)	/
Proposed method	Clayton	99.60 (99.98)	0.83 (±0.13)	4.8
	Frank	100.00 (100.00)	0.91 (±0.15)	6.2

Case 2: $\boldsymbol{\beta} = (\underbrace{0.2,\ldots,0.2}_{\times 5}, \underbrace{-0.2,\ldots,-0.2}_{\times 5}, \underbrace{0,\ldots,0}_{\times 90}), \beta_1 = 0.2$

	Underlying model	Sensitivity % (Specificity %)	$E[\hat{\beta}_1]$ (±SD)	$E[\hat{\alpha}]$
Univariate selection	Clayton	64.40 (96.04)	0.34 (±0.13)	/
	Frank	69.40 (96.60)	0.38 (±0.14)	/
Proposed method	Clayton	81.00 (97.89)	0.41 (±0.13)	4.2
	Frank	85.60 (98.40)	0.43 (±0.13)	4.7

Case 3: $\boldsymbol{\beta} = (\underbrace{0.2,\ldots,0.2}_{\times 10}, \underbrace{0,\ldots,0}_{\times 90}), \beta_1 = 0.2$

	Underlying model	Sensitivity % (Specificity %)	$E[\hat{\beta}_1]$ (±SD)	$E[\hat{\alpha}]$
Univariate selection	Clayton	95.20 (99.47)	0.67 (±0.15)	/
	Frank	95.80 (99.53)	0.72 (±0.16)	/
Proposed method	Clayton	98.80 (99.87)	0.75 (±0.14)	3.5
	Frank	99.80 (99.98)	0.81 (±0.13)	4.3

Case 4: $\boldsymbol{\beta} = (\underbrace{0.1,\ldots,0.1}_{\times 10}, \underbrace{-0.1,\ldots,-0.1}_{\times 10}, \underbrace{0,\ldots,0}_{\times 80}), \beta_1 = 0.1$

	Underlying model	Sensitivity % (Specificity %)	$E[\hat{\beta}_1]$ (±SD)	$E[\hat{\alpha}]$
Univariate selection	Clayton	71.50 (92.88)	0.36 (±0.15)	/
	Frank	74.10 (93.53)	0.38 (±0.15)	/
Proposed method	Clayton	83.80 (95.95)	0.41 (±0.13)	4.4
	Frank	86.20 (96.55)	0.45 (±0.14)	3.9

Higher sensitivity and specificity correspond to better gene selection performance.

5 Data analysis

We revisit the 128 non-small-cell lung cancer patients of Chen et al.[4] available at http://www.ncbi.nlm.nih.gov/projects/geo/ with accession number GSE4882. In this study, the primary endpoint is overall survival, i.e. death from any cause. During the follow-up, 38 patients died (35 patients due to recurrence of cancer and 3 patients due to other causes). The remaining 87 patients are censored, i.e. survived at the end of their follow-up times. Dependent censoring may arise in univariate selection due to the unadjusted effect of genes as demonstrated in Section 2.2 (see Figure 1). In addition, it is suspicious that some early dropouts were related to patients' health status.

We split the 125 lung cancer patients into 63 training and 62 testing samples. Chen et al.[4] used univariate selection on the 63 training patients to identify a 16-gene signature, which led to a highly accurate separation of the patients with a good prognosis from those with a poor prognosis among the 62 testing patients. This univariate analysis relies on the independent censoring assumption (Assumption I).

Figure 4 plots the cross-validated c-index $CV(\alpha)$ using the training samples. The c-index is maximized at the association parameter $\hat{\alpha} = 18$ (Kendall's tau $= 0.90$). This implies the possible gain in prediction for testing samples by considering dependent censoring models.

We compare univariate selection with our proposed method in terms of selecting the top $\tau = 16$ genes among the 485 genes. The two gene selection methods used on the training samples resulted in two different lists of the top 16 genes as given in Table 3. We find that the gene list from univariate selection is the same as Chen et al.[4] Among the 16 genes, six genes are selected by both methods, while the other 10 genes differ between the two methods.

We compare the predictive value of the two methods based on the PI. The PI of univariate selection is

$$\begin{aligned}\text{PI(univariate selection)} = {} & (-1.09 * \text{ANXA5}) + (1.32 * \text{DLG2}) + (0.55 * \text{ZNF264}) \\ & + (0.75 * \text{DUSP6}) + (0.59 * \text{CPEB4}) \\ & + (-0.84 * \text{LCK}) + (-0.58 * \text{STAT1}) + (0.65 * \text{RNF4}) + (0.52 * \text{IRF4}) \\ & + (0.58 * \text{STAT2}) + (0.51 * \text{HGF}) + (0.55 * \text{ERBB3}) + (0.47 * \text{NF1}) \\ & + (-0.77 * \text{FRAP1}) + (0.92 * \text{MMD}) + (0.52 * \text{HMMR})\end{aligned}$$

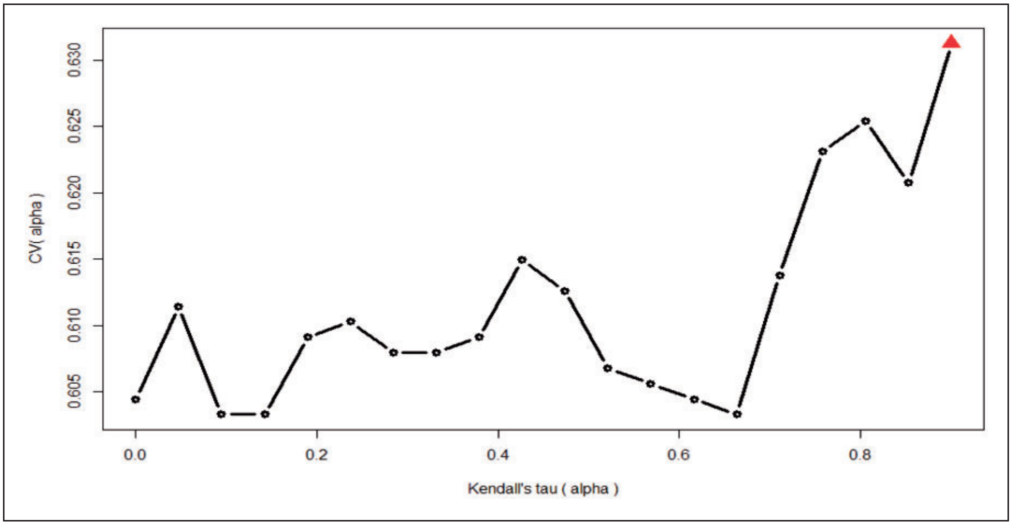

Figure 4. The cross-validated c-index for the 63 training set from the lung cancer data. The cross-validated c-index is maximized at $\alpha = 18$, which corresponds to Kendall's tau $= 0.90$.

This 16-gene signature leads to the same risk score reported in the original work of Chen et al.[4] (their supplemental, p. 4). On the other hand, the proposed method yields a different PI

$$\begin{aligned}\text{PI (proposed method)} = {} & (0.51 * \text{ZNF264}) + (0.50 * \text{MMP16}) \\ & + (0.50 * \text{HGF}) + (-0.49 * \text{HCK}) + (0.47 * \text{NF1}) \\ & + (0.46 * \text{ERBB3}) + (0.57 * \text{NR2F6}) + (0.77 * \text{AXL}) \\ & + (0.51 * \text{CDC23}) + (0.92 * \text{DLG2}) \\ & + (-0.34 * \text{IGF2}) + (0.54 * \text{RBBP6}) + (0.51 * \text{COX11}) \\ & + (0.40 * \text{DUSP6}) + (-0.37 * \text{CKMT1A}) + (-0.41 * \text{ENG})\end{aligned}$$

We begin our analysis by comparing the cumulative incidence curves for the good (or poor) prognosis groups separated by the low (or high) values of the PIs on the testing samples (Figure 5). Gray's two-sample test is used to measure the separation between the two curves. The proposed method leads to a slightly poorer separation of the good and poor prognoses (P-value = 0.256) compared to that of the univariate selection (P-value = 0.247). Although prediction in terms of the cumulative incidence probability is the standard in the presence of dependent censoring, the goal of Chen et al.[4] is to identify genes that are predictive for overall survival.

To validate the predictive ability of the top 16 genes on overall survival, we draw the survival curves for the good (or poor) prognosis groups separated by the low (or high) values of the PIs (Figure 6). Since the Kaplan–Meier survival curve is biased under dependent censoring, we apply the

Table 3. The 16 most strongly associated genes based on two methods: univariate selection and the proposed method. The genes are ordered according to the P-values.

	Univariate selection			Proposed method		
No.	Gene symbol	Coefficient	P-value	Gene symbol	Coefficient	P-value
1	ANXA5	−1.09	0.0039	ZNF264	0.51	0.0004
2	DLG2	1.32	0.0041	MMP16	0.50	0.0005
3	ZNF264	0.55	0.0079	HGF	0.50	0.0010
4	DUSP6	0.75	0.0086	HCK	−0.49	0.0012
5	CPEB4	0.59	0.0162	NF1	0.47	0.0016
6	LCK	−0.84	0.0171	ERBB3	0.46	0.0016
7	STAT1	−0.58	0.0198	NR2F6	0.57	0.0030
8	RNF4	0.65	0.0220	AXL	0.77	0.0035
9	IRF4	0.52	0.0299	CDC23	0.51	0.0050
10	STAT2	0.58	0.0311	DLG2	0.92	0.0055
11	HGF	0.51	0.0334	IGF2	−0.34	0.0081
12	ERBB3	0.55	0.0335	RBBP6	0.54	0.0082
13	NF1	0.47	0.0380	COX11	0.51	0.0118
14	FRAP1	−0.77	0.0408	DUSP6	0.40	0.0121
15	MMD	0.92	0.0419	ENG	−0.37	0.0139
16	HMMR	0.52	0.0481	CKMT1A	−0.41	0.0155

Gray shading signifies genes that appear in both univariate selection and the proposed method.

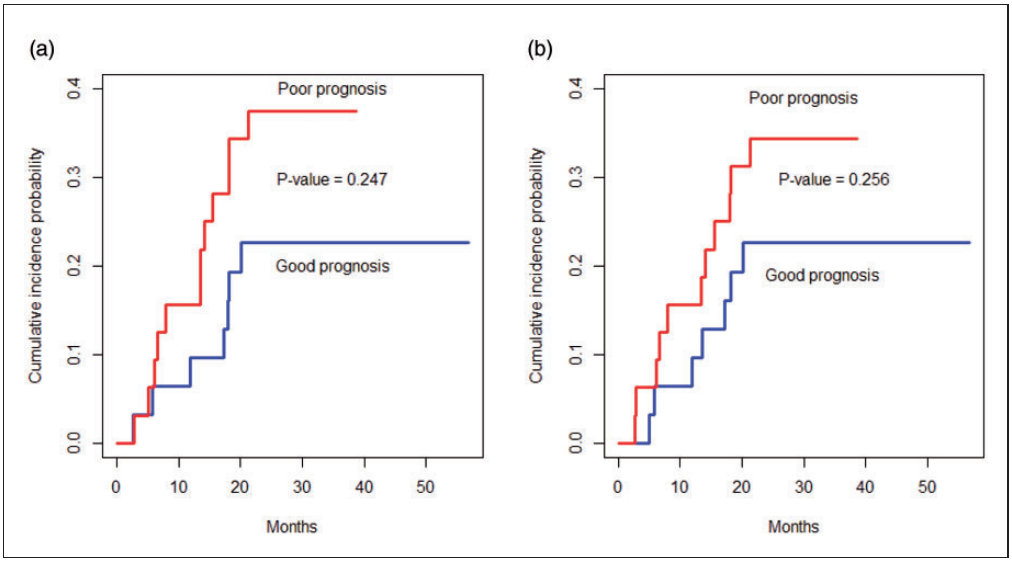

Figure 5. The cumulative incidence curves for the good (or poor) prognosis group separated by the top 16 genes. The good (or poor) group is determined by the low (or high) values of the 16-gene PI with equal sample sizes. (a) Univariate selection, (b) proposed method.

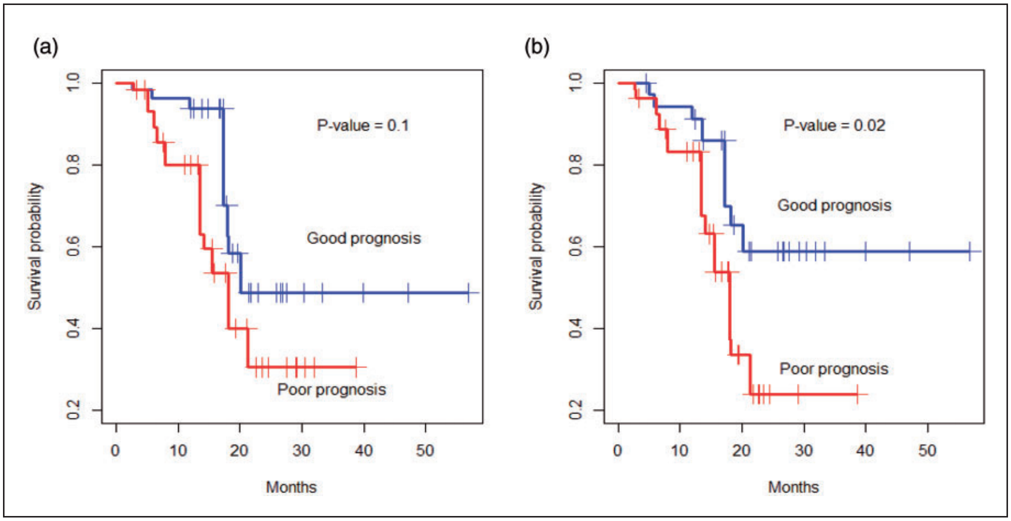

Figure 6. The marginal survival curves for the good (or poor) prognosis group separated by the top 16 genes. The good (or poor) group is determined by the low (or high) values of the 16-gene PI with equal sample sizes. (a) Univariate selection, (b) proposed method.

Table 4. The average vertical difference in the survival curves between good and poor prognosis groups in the test dataset, where the good or poor prognosis is determined by the prognostic index based on the top $\tau = 10, 20, \ldots, 90$ genes. The corresponding P-value is obtained using the permutation test for the weighted Kaplan–Meier statistics.

	Univariate method	Proposed method
τ	Difference in survival (P-value)	Difference in survival (P-value)
10	0.072 (0.453)	0.185 (0.062)
20	0.114 (0.249)	0.154 (0.120)
30	0.025 (0.806)	0.251 (0.012)
40	0.102 (0.301)	0.230 (0.021)
50	0.131 (0.186)	0.210 (0.035)
60	0.136 (0.162)	0.252 (0.012)
70	0.112 (0.253)	0.322 (0.002)
80	0.226 (0.023)	0.335 (0.001)
90	0.177 (0.072)	0.335 (0.001)

Smaller P-values correspond to a better separation of the patients with a good prognosis from those with a poor prognosis.

copula-graphic estimator[16,18] of the survival curves adjusted under the Clayton copula at $\hat{\alpha} = 18$ (Kendall's tau $= 0.90$). Figure 6 shows that the proposed method appears to give a clearer separation between the good and poor groups than univariate selection does. The separation of the two curves is measured by the average vertical difference in the survival curves over the study period, and the corresponding P-value is obtained using the permutation test.[34,35] The proposed method leads to a significantly better separation of the good and poor prognoses (average difference=0.230; P-value = 0.02) compared to that for the univariate selection (average difference = 0.162; P-value = 0.10).

We perform the same prediction analysis as above for the different cut-off numbers for the top τ genes. The results for $\tau = 10, 20, 30, \ldots, 80, 90$ are summarized in Table 4. In most cases, the proposed method produces significant separation between the patients with a good prognosis and those with a poor prognosis (P-value < 0.05) while univariate selection seldom reaches 5% significance level for that. Both methods produce the best separation at $\tau = 80$. In this case, the proposed method provides an extremely clear separation between the good and poor prognosis patients (P-value $= 0.001$; Figure 7). In terms of cumulative incidence, the proposed method leads to a clearly better separation of the good and poor prognoses (P-value $= 0.082$; Figure 8) compared to that for univariate selection (P-value $= 0.230$). Hence, the list of genes selected by the proposed methods is consistently more predictive than those by univariate selection.

6 Conclusion

In this paper, we consider gene selection procedures for survival data with dependent censoring. We first develop a copula-based analytic tool to investigate the effect of dependent censoring on univariate gene selection. This tool facilitates the understanding of the bias due to dependent censoring under various types of dependence structures. Our analysis reveals that the bias grows when the censoring percentage and the degree of dependence increase. In addition, we find that the qualitative natures of the bias due to different copulas, namely Clayton, Frank, and Gumbel copulas, are remarkably different.

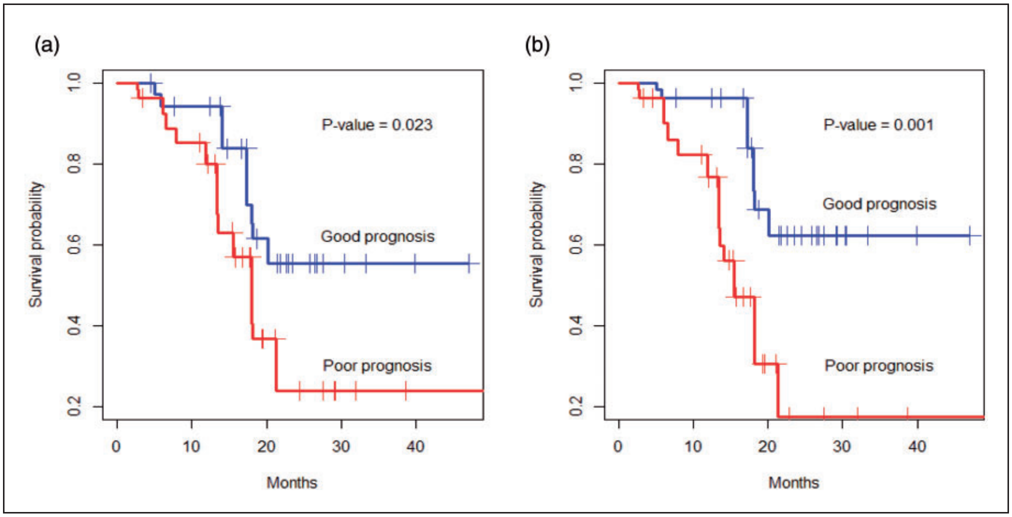

Figure 7. The marginal survival curves for the good (or poor) prognosis group separated by the top 80 genes. The good (or poor) group is determined by the low (or high) values of the 80-gene PI with equal sample sizes. (a) Univariate selection, (b) proposed method.

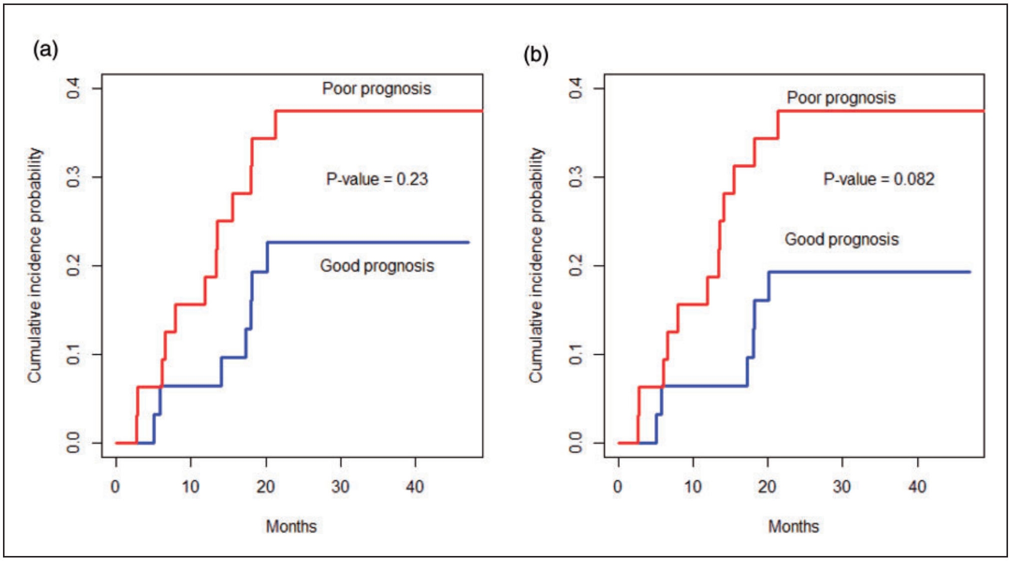

Figure 8. The cumulative incidence curves for the good (or poor) prognosis group separated by the top 80 genes. The good (or poor) group is determined by the low (or high) values of the 80-gene PI with equal sample sizes. (a) Univariate selection, (b) proposed method.

We also utilize the copula models to develop a new gene selection method for dependent censoring. This method, in contrast to univariate selection, does not require the strong independence assumption (Assumption I). Although the paper focuses on fitting the Clayton copula for dependence structure, the method can be applicable to other copulas and hence provides a very flexible framework for various different types of dependent censoring. The simulations show that the proposed method offers a higher percentage of selecting nonnull genes of interest than the traditional univariate selection when dependent censoring is indeed present. In addition, the method exhibits good performance even if the fitted model is misspecified. This robustness is particularly important for the copula-based approach, in which it is fairly difficult to specify the correct form of copulas. When applied to the aforementioned lung cancer data, the genes selected by using the proposed method have better predictive performance than the ones using univariate selection. The proposed method would be generally useful in datasets where dependent censoring is suspected.

The objective of the proposed approach is to select genes that are relevant to marginal survival, where the effect of dependent censoring is removed. While the simulations and data analysis provide firm evidence that the method selects effective genes on the marginal survival, it does not necessarily select relevant genes on the cumulative incidence function for the survival. As demonstrated in the data analysis, the selected genes that are highly predictive on marginal survival become somewhat offset in terms of the cumulative incidence. To investigate the predictive power of the selected genes, plotting both cumulative incidence function and the estimator of survival function for the validation samples would be informative, as demonstrated in Section 5. For estimating survival functions, the copula-graphic estimator adjusted for dependent censoring is suitable to reduce the bias caused by the traditional Kaplan–Meier estimator. The plots of these different curves are useful to clarify the benefit of the selected genes on survival prediction and to build risk prediction models involving high-dimensional microarrays. We note that in a true competing risks situation, e.g. when studying the risk of relapse, with death without prior relapse acting as a competing risk, the marginal survival is usually not of interest.

Supplementary material

Supplementary materials include Appendix A (Laplace transforms), Appendix B (Implementation of Chen[19] under the Clayton model), and Appendix C (Data generation for the covariates **x**).

Acknowledgements

The authors thank the reviewer and the editor for their helpful comments that substantially improve the manuscript. The authors also thank Dr Chen Hsuan-Yu for permission to use the lung cancer data of Chen et al.[4]

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Cox DR. Regression models and life-tables (with discussion). *J R Stat Soc Ser B* 1972; **34**: 187–220.
2. Jenssen TK, Kuo WP, Stokke T, et al. Association between gene expressions in breast cancer and patient survival. *Hum Genet* 2002; **111**: 411–420.
3. Matsui S. Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. *BMC Bioinform* 2006; **7**: 156.
4. Chen HY, Yu SL, Chen CH, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007; **356**: 11–20.
5. Matsui S, Simon R, Qu P, et al. Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. *Clin Cancer Res* 2012; **18**: 21.
6. Tukey JW. Tightening the clinical trial. *Control Clin Trials* 1993; **14**: 266–285.

7. Radamacher MD, Mcshane LM and Simon R. A paradigm for class prediction using gene expression profiles. *J Comput Biol* 2002; **9**: 505–511.
8. Beer DG, Kardia SLR, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002; **8**: 816–824.
9. Emura T, Chen YH and Chen HY. Survival prediction based on compound covariate under Cox proportional hazard models. *PLoS One* 2012; **7**: e47627.
10. Beyersmann J, Allignol A and Schumacher M. *Competing risks and multistate models with R*. New York: Springer-Verlag, 2012.
11. Fine JP and Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc* 1999; **94**: 548–560.
12. Binder H, Allignol A, Schumacher M, et al. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* 2009; **25**: 890–896.
13. Bakoyannis G and Touloumi G. Practical methods for competing risks data: A review. *Stat Method Med Res* 2012; **21**: 257–272.
14. Mogensen UB and Gerds TA. A random forest approach for competing risks based on pseudo-values. *Stat Med* 2011; **32**: 3102–3114.
15. Nelsen RB. An introduction to copulas. *Springer Series in Statistics*. 2nd ed. New York: Springer-Verlag, 2006.
16. Zheng M and Klein JP. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* 1995; **82**: 127–138.
17. Heckman JJ and Honore BE. The identifiability of the competing risks models. *Biometrika* 1989; **76**: 325–330.
18. Rivest LP and Wells MT. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *J Mult Anal* 2001; **79**: 138–155.
19. Chen YH. Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *J R Stat Soc Ser B* 2010; **72**: 235–251.
20. Wessels LFA, Reinders MJT, Hart AAM, et al. A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics* 2002; **21**: 3755–3762.
21. Witten DM and Tibshirani R. Survival analysis with high-dimensional covariates. *Stat Method Med Res* 2010; **19**: 29–51.
22. Andersen PK, Borgan O, Gill RD, et al. *Statistical models based on counting processes*. New York: Springer-Verlag, 1993.
23. Fleming TR and Harrington DP. *Counting process and survival analysis*. New York: John Wiley and Sons, 1991.
24. Struthers CA and Kalbfleish JD. Misspecified proportional hazard models. *Biometrika* 1986; **73**: 363–369.
25. Oakes D. Bivariate survival models induced by frailties. *J Am Stat Assoc* 1989; **84**: 487–493.
26. Kalbfleisch JD and Prentice RL. *The statistical analysis of failure time data*. 2nd ed. New York: John Wiley and Sons, 2002.
27. Escarela G and Carriere JF. Fitting competing risks with an assumed copula. *Stat Method Med Res* 2003; **12**: 333–349.
28. Braekers R and Veraverbeke N. A copula-graphic estimator for the conditional survival function under dependent censoring. *Can J Stat* 2005; **33**: 429–447.
29. Emura T and Chen YH. Regression estimation based on the compound shrinkage method under the Cox proportional hazard model. *R compound.Cox package, version 1.4*, 2013.
30. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proc Natl Acad Sci USA* 1975; **72**: 20–22.
31. Verveij PJM and van Houwelingen HC. Cross validation in survival analysis. *Stat Med* 1993; **12**: 2305–2314.
32. Harrell FE, Califf RM, Pryor DB, et al. Evaluating the yield of medical tests. *J Am Med Assoc* 1982; **247**: 2543–2546.
33. Harrell FE, Lee KL and Mark DB. Multivariate prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; **15**: 361–387.
34. Pepe MS and Fleming TR. Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics* 1989; **45**: 497–507.
35. Frankel PH, Reid ME and Marshall JR. A permutation test for a weighted Kaplan-Meier estimator with application to the nutritional prevention of cancer trial. *Contemp Clin Trial* 2007; **28**: 343–347.

Gene selection for survival data under dependent censoring: a copula-based approach

Supplementary Materials

Takeshi Emura Graduate Institute of Statistics, National Central University, Jhongda Road, Jhongli City Taoyuan 32001, Taiwan and

Yi-Hau Chen Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan

The supplementary materials includes Appendix A (Laplace transforms), Appendix B (Implementation of Chen (2010) under the Clayton model), and Appendix C (Data generation for the covariates $\mathbf{x}$).

Appendix A: Laplace transforms

The Laplace transform of a random vector $\mathbf{X}$ is defined as $\varphi(\mathbf{u}) = E\{\exp(-\mathbf{u}'\mathbf{X})\}$. Hence,

$$\varphi_{\boldsymbol{\beta}(-j),\boldsymbol{\gamma}(-j)}(u,v) = E\{\exp(-ue^{\boldsymbol{\beta}'_{(-j)}\mathbf{x}_{i(-j)}})\exp(-ve^{\boldsymbol{\gamma}'_{(-j)}\mathbf{x}_{i(-j)}}) \mid x_{ij}\},$$

$$\varphi_{\boldsymbol{\beta}(-j)}(u) = E\{\exp(-ue^{\boldsymbol{\beta}'_{(-j)}\mathbf{x}_{i(-j)}}) \mid x_{ij}\} = \varphi_{\boldsymbol{\beta}(-j),\boldsymbol{\gamma}(-j)}(u,0),$$

$$\varphi_{\boldsymbol{\gamma}(-j)}(u) = E\{\exp(-ue^{\boldsymbol{\gamma}'_{(-j)}\mathbf{x}_{i(-j)}}) \mid x_{ij}\} = \varphi_{\boldsymbol{\beta}(-j),\boldsymbol{\gamma}(-j)}(0,u),$$

are the Laplace transforms for some random vectors.

Appendix B: Implementation of Chen (2010) under the Clayton copula

The Clayton family of Archimedean copulas

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

is obtained under a copula generator $\phi_\alpha(t) = (t^{-\alpha} - 1)/\alpha$ with its inverse $\phi_\alpha^{-1}(t) = (1 + \alpha t)^{-1/\alpha}$. The degree of dependence is measure by Kendall's tau $\tau = \alpha/(\alpha+2)$. One has

$$\Phi_\alpha(u,v) = \alpha^{-1}\log(u^{-\alpha} + v^{-\alpha} - 1),$$

$$D_{\alpha,1}(u,v) = u^{-\alpha-1}(u^{-\alpha} + v^{-\alpha} - 1),\ D_{\alpha,2}(u,v) = v^{-\alpha-1}(u^{-\alpha} + v^{-\alpha} - 1).$$

Therefore, the likelihood given α is

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

where

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

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

The maximizer of the likelihood function is denoted by $(\, \hat{\beta}_j(\alpha), \hat{\gamma}_j(\alpha), \hat{\Lambda}_{0j}(\alpha), \hat{\Gamma}_{0j}(\alpha)\,)$, which is computed by nonlinear maximization routines. We implement the computation of $\hat{\beta}_j(\alpha)$ and $se\{\, \hat{\beta}_j(\alpha)\,\}$ in an R compound.Cox package (Emura & Chen, 2012).

Appendix C: Data generation for the covariate $\mathbf{x}$

Let $\mathbf{x}' = (\, x_1, ..., x_q, x_{q+1}, ..., x_p\,)$ be a vector of genes. The following algorithms for generationg $\mathbf{x}'$ are due to Emura et al. (2012):

1. Tag gene sequence: Each of the *q* covariates is positively correlated to *s* genes that have zero coefficients. Specifically, we set

$$x_j = \begin{cases} A_j + u_j & \text{if} \quad j \le q\,; \\ A_k + u_j & \text{if} \quad j = q + (k-1)s + 1, ..., q + ks, \;\; k = 1, ..., q\,; \\ U_j & \text{if} \quad j \ge q + qs + 1 \end{cases}$$

where $A_j \sim U(-0.75, 0.75)$, $u_j \sim U(-0.75, 0.75)$, $U_j \sim U(-1.5, 1.5)$, and they are independent of one another. This scenario represents the setting that *q* independent sets of genes are associated with survival; the (*s* + 1) genes in each set are correlated, and after accounting for one “tag gene” in each set of genes, the other genes have no net effects on survival.

2. Gene pathway: The q significant covariates are positively correlated. We set

$$x_j = \begin{cases} A_1 + u_j & \text{if} \quad 1 \leq j \leq q\,; \\ U_j & \text{if} \quad q < j \leq p\,, \end{cases}$$

or

$$x_j = \begin{cases} A_1 + u_j & \text{if} \quad 1 \leq j \leq q/2\,; \\ A_2 + u_j & \text{if} \quad q/2 < j \leq q\,; \\ U_j & \text{if} \quad q < j \leq p\,, \end{cases}$$

where $A_j \sim U(-0.75, 0.75)$, $u_j \sim U(-0.75, 0.75)$, $U_j \sim U(-1.5, 1.5)$, and they are independent of one another. The former represents the setting that there exists a "gene pathway" of *q* correlated genes that jointly affect survival, and the latter does for two gene pathways of *q*/2 correlated genes. Hence, scenario 2 represents a setting where the genes informative for survival are correlated while scenario 1 represents a setting where the informative genes are independent of each other.

For both scenarios, the covariates are standardized so that they have standard deviation 1.

References

Chen, Y.-H. (2010). Semiparametric marginal regression analysis for dependent competing risks under an assumed copula, Journal of the Royal Statistical Society, Ser B **72**, 235-251.

Emura T., Chen Y.-H., and Chen H.-Y. (2012). Survival prediction based on compound covariate under Cox proportional hazard models,. PLoS ONE **7**(10): e47627. doi:10.1371/journal.pone.0047627.

Emura, T. and Chen, Y.-H. (2014) Regression estimation based on the compound shrinkage method under the Cox proportional hazard model, Version 1.4. 2014.