VARIABLE RANKING IN BIVARIATE COPULA SURVIVAL MODELS

Danilo Petti [1], Marcella Niglio[2] and Marialuisa Restaino[2]

[1] Department of Mathematical Sciences, University of Essex, (e-mail: d.petti@essex.ac.uk)

[2] Department of Economics and Statistics, University of Salerno, (e-mail: mniglio@unisa.it, mlrestaino@unisa.it)

ABSTRACT: We propose a variable ranking procedure based on copula bivariate time-to-event margins under a general censoring scheme. The procedure identifies the important variables influencing the two time-to-events in a high dimensional setting introducing a proper metric able to take into account the probabilistic copula structure. The proposal is the first attempt to apply a variable selection method to a copula bivariate time-to-event domain. The advantages of the proposed approach are illustrated in a case study based on AREDS dataset.

KEYWORDS: bivariate survival, copula, variable ranking, ultrahigh dimensionality.

1 Introduction

Technologies have had a deep impact on society and on data collection in a wide range of scientific areas. With a relatively low cost, we are able to collect massive amounts of information (and noise). This has led to the high dimensional data phenomenon where the variable selection plays a central role. This is even more true in the case of bivariate copula survival models under a censoring scheme (presence of two outcomes and missing information). Under this domain, we are interested in identifying two sets of relevant covariates for two random times to event (T_1 and T_2). This can be achieved by ranking the covariates in order of importance through a given metric ω to assess the contribution of each feature in the dataset. As far as the authors are aware, there is no valuable variable selection or variable ranking method nor implementation available in the literature for Bivariate Copula Survival models. In Sect. 2 we shortly present the model under analysis, and in Sect. 3 we sketch the algorithm of variable ranking. The application to AREDS data is presented in Sect. 4.

2 The model

Let us consider the pair of survival times (T_{1i}, T_{2i}), a vector of covariates $\mathbf{x}_i$, for $i = 1, 2, \ldots, n$, and an associated generic parameter vector $\delta \in \mathbb{R}^w$ of dimension w. We assume that T_{1i} and T_{2i} have marginal survival functions given by $S_v(t_{vi}|\mathbf{x}_{vi};\beta_v) = P(T_{vi} > t_{vi}|\mathbf{x}_{vi};\beta_v) \in (0,1)$, for $v = 1,2$, and a joint survival function expressed as follows $S(t_{1i}, t_{2i}|\mathbf{x}_i;\delta) = C(S_1(t_{1i}|\mathbf{x}_{1i};\beta_1), S_2(t_{2i}|\mathbf{x}_{2i};\beta_2)$ $;m\{\eta_{3i}(\mathbf{x}_{3i};\gamma)\})$, where $\delta^T = (\beta_1^T, \beta_2^T, \gamma^T)$, $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$ and $\mathbf{x}_{3i}$ are vectors of covariates, with associated coefficient vectors $\beta_1 \in \mathbb{R}^{w_1}$, $\beta_2 \in \mathbb{R}^{w_2}$ and $\gamma \in \mathbb{R}^{w_3}$ such that $w = w_1 + w_2 + w_3$, $C : (0,1)^2 \to (0,1)$ is a uniquely defined 2-dimensional copula function with coefficient $\theta_i = m\{\eta_{3i}(\mathbf{x}_{3i};\gamma)\}$ modelling the potentially varying dependence of (T_{1i}, T_{2i}) across observations, $\eta_{3i}(\mathbf{x}_{3i};\gamma) \in \mathbb{R}$ is a predictor which includes generic additive covariate effects, and m is a monotonic and differentiable one-to-one transformation function. The marginal survival functions can be written as

$$g_v\left[S(t_{vi}|\mathbf{x}_{vi};\beta_\nu)\right] = \eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \mathbf{f}_v(\beta_v)),\ \nu = 1,2 \tag{1}$$

where $g_v : (0,1) \to \mathbb{R}$ is a monotone and twice continuously differentiable link function with bounded derivatives, $\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \mathbf{f}_v(\beta_v)) \in \mathbb{R}$ is an additive predictor which models the baseline hazard and several types of covariate effects, and $\mathbf{f}_v(\beta_v)$ has the role of imposing a monotonicity constraint. Equation 1 can be written as $S(t_{\nu i}|\mathbf{x}_{\nu i};\beta_\nu) = G_\nu(\eta_{\nu i}(t_{\nu i}, \mathbf{x}_{\nu i}, \mathbf{f}(\beta_\nu)))$ where G_ν is an inverse link function. The key difference between $\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \mathbf{f}_v(\beta_v))$, for $v = 1,2$, and $\eta_{3i}(\mathbf{x}_{3i};\gamma)$ is that the two former predictors must include smooth functions of times t_{vi} which can be treated as regressors. We, therefore, consider a generic $\eta_{\nu i}(\nu = 1,2,3)$, where the dependence on the covariates and parameters is momentarily dropped, an overall covariate vector $\mathbf{z}_{\nu i}$ containing $\mathbf{x}_{\nu i}$ and $t_{\nu i}$ when $\nu = 1,2$, and $\mathbf{z}_{3i} = \mathbf{x}_{3i}$. For simplicity, the dimensions of $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$ are assumed to be W_1 and W_2. A generic additive predictor is specified as follows

$$\eta_{\nu i} = \beta_{\nu 0} + \sum_{k_\nu=1}^{K_\nu} s_{\nu k_\nu}(\mathbf{z}_{\nu k_\nu i}),\ \nu = 1,2,3 \tag{2}$$

where $\beta_{\nu 0} \in \mathbb{R}$ is an overall intercept, $\mathbf{z}_{\nu k_\nu i}$ denotes the k_ν^{th} sub-vector of the complete vector $\mathbf{z}_{\nu i}$ and the K_ν functions $s_{\nu k_\nu}(\mathbf{z}_{\nu k_\nu i})$ represent generic effects which are chosen according to the type of covariate(s) considered (Wood, 2017). The above formulation allows for many types of flexible covariate effects. For more details see Marra, 2020.

3 The Variable Selection Algorithm

We extend the variable selection procedure proposed by Baranowski, 2020 to the bivariate survival data. Given the set of w covariates, the variables with higher influence on $\eta_{\nu i}(x_{\nu i}, \beta_{\nu i})$, $(\nu = 1,2)$ are those that even in presence of randomly selected sub-samples exhibit consistent relationship to explain the dependence of the two survival functions.

Let $Z_i = \{T_{1i}, T_{2i}, X_{i1}, X_{i2}, \ldots, X_{iw}\}$, for $i = 1,2,\ldots,n$ and with w that grows with n, be the observed dataset used to select the subset of covariates $\{X_1, \ldots, X_w\}$. Further, let $\mathcal{A}^\nu \subset (1, \ldots, w_\nu)$ for $\nu = 1,2$ be the indices that identify a subset of covariates for the ν-th margin and let $|\mathcal{A}^\nu| = k$ be the cardinality of $\mathcal{A}^\nu$, for $k = 0,1,\ldots,w_\nu$. Let $R^s_{nj}(Z_1,\ldots,Z_n)$ be the ranking of the j-th covariate, based on a metric $\hat{\omega}^\nu_j = \hat{\omega}^\nu_j(Z_1,\ldots,Z_n)$ assessing the importance of each covariate of each margin, such that $\omega^\nu_{R^s_{n1}} \geq \cdots \geq \omega^\nu_{R^s_{n|\mathcal{A}^\nu|}}$. The probability of the set of $|\mathcal{A}^\nu|$ top-ranked variables in $\mathcal{A}^\nu$ is:

$$\pi_{n,m}(\mathcal{A}^\nu) = \mathbb{P}\left(\left\{R^\nu_{n1}(Z_1,\ldots,Z_m),\ldots,R^\nu_{n|\mathcal{A}^\nu|}(Z_1,\ldots,Z_m)\right\} = \mathcal{A}^\nu\right), \nu = 1,2 \tag{3}$$

that is obtained from a subset of m observations, with $1 \leq m \leq n$. To estimate 3 a bootstrap approach is proposed in Baranowski, 2020. It follows that $\pi_{n,m}(\mathcal{A}^\nu)$ is the probability that the covariates in $\mathcal{A}^\nu$ are ranked at the top, using a subset of m observations. The selection can be then performed on the set of top-ranked variables $\mathcal{A}^\nu$ from which the number of terms $\hat{s}^\nu$ can be determined using equation (2.5) in Baranowski, 2020. In practice, given the estimated probabilities of $\hat{\pi}_{n,m}\left(\hat{\mathcal{A}}^\nu{}_{k,m}\right)$, for $k = 0,\ldots,k_{\max}-1$, with $k_{\max}$ a fixed large integer, the number of relevant variables is related to the evaluation of the magnitude of the estimated probability.

4 Application to AREDS dataset

The performance of the algorithm in Sect. 3 is assessed using the AREDS data (available in the `R` package `CopulaCenR`). The dataset includes 629 Caucasian participants. The event of interest is late-AMD progression, which is a disease affecting both eyes. Less than 50% of the subjects had late-AMD in both eyes (bivariate interval-censored). Around 20% had late-AMD in one eye but not the other by the study end (mixed interval- and right-censored). More than 33% did not develop late-AMD in either eye (bivariate right-censored). The variables are Severity score, values that reflect the progression of the disease in the eyes (`SevScale1E` `SevScale2E`), enrollment age (`Age`), and a

genetic variant (`rs2284665`), factor variable with levels 0 (`GG`), 1 (`GT`) and 2 (`TT`), respectively). Furthermore, the AREDS dataset has been perturbed by adding 100 independent realizations of a standard Gaussian distribution. For sake of completeness, the algorithm has also been evaluated through a Monte Carlo study (not included in the paper), which confirmed the effectiveness of the method returning false positives and negatives close to zero.

We carried out some preliminary fitting from which emerged that {`C0`, `POPO`} is the combination with the lowest BIC (4330.08) considering a full model specification (all features included in all three margins). The procedure has been applied on a standardized version of the dataset, where `rs2284665` has been encoded as $0/1$, resulting in three new covariates. The tuning parameters has been specified as follows: $k_{max} = 10$, $m = \lfloor n/2 \rfloor$, $\tau = 0.5$, 50 bootstrap replicates, Clayton copula (`C0`) and Proportional odds (`PO`, `PO`). We have considered two metrics: $\omega_\nu = \beta_j^2 i(\beta_j)$ (with $i(\beta_j)$ be the associated element of the Fisher information matrix) and $\psi_\nu = |\beta_j|$. In pseudo code (ignoring the smooth functions of times t_ν) $\eta_\nu = \beta_{\nu 0} + \beta_{\nu j} x_j$ for $j = 1, \ldots, w$. The former metric is proposed specifically for the class of Bivariate Copula Survival models while the latter is the absolute value of the coefficient.

Comparing the variable selected with the two metrics, the selection with $\beta_j^2 i(\beta)$ has greater cardinality and is able to select those characteristics considered relevant for the event of interest in the literature (see Sun, 2021), giving empirical evidence of its goodness.

Table 1. *Results of the algorithm in Sect. 3 using Clayton copula, proportional hazard margins and using* $\omega_\nu = \beta_j^2 i(\beta)$ *and* $\psi_\nu = |\beta_j|$*, for* $j = 1, \ldots, w$*, as metrics. The covariates are ordered according to their importance. The BIC and AIC are obtained by applying* `gjrm()` *function to a non-standardized AREDS.*

	$\hat{\mathfrak{A}}^1$	$\hat{\mathfrak{A}}^2$	BIC	AIC
$\mathcal{M}_\omega$	{`SevScale1E`, `SevScale2E`, `GG`, `TT`}	{`GG`, `SevScale2E`, `TT`, `SevScale1E`, `GT`}	4325.849	4225.385
$\mathcal{M}_\psi$	{`SevScale1E`, `SevScale2E`}	{`SevScale2E`, `SevScale1E`, `GG`, `TT`, `GT`}	4324.518	4220.659

References

BARANOWSKI, R., CHEN Y. FRYZLEWICZ P. 2020. Ranking-based variable selection for high-dimensional data. *Stat. Sinica*, **30**(3), 1485–1516.

MARRA, G., RADICE R. 2020. Copula link-based additive models for right-censored event time data. *J. of the Am. Stat. Ass.*, **115**(530), 886–895.

SUN, T., DING Y. 2021. Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostat.*, **22**(2), 315–330.

WOOD, S. N. 2017. *Generalized additive models: an introduction with R*. CRC Press.