Covariate Significance Testing in the Conditional Bivariate Distribution Function

Gustavo Soutinho[1,a)], Luís Meira-Machado[2,b)] and Artur Araujo[3,c)]

[1]*Research on Economics, Management and Information Technologies (REMIT) - Portucalense University, Rua Dr. António Bernardino de Almeida, 541 4200-072 Porto, Portugal*
[2]*Centre of Mathematics, University of Minho, Campus de Azurém, 4800 - 058 Guimarães, Portugal*

[a)]Corresponding author: gustavo.soutinho@upt.pt
[3]*University of Vigo, Spain*
[b)]lmachado@math.uminho.pt
[c)]artur.stat@gmail.com

Abstract. One significant goal in recurrent events analysis is the estimation of the bivariate distribution function. Estimating this function for censored gap times is crucial across various fields and applications, as it helps elucidate recurring events and their underlying patterns. 'Gap time' refers to the duration between successive occurrences of an event, while the bivariate distribution function describes the joint probability distribution of two such gap times. Despite considerable progress in this area, most approaches do not account for the influence of covariates. In this paper, we introduce a viable nonparametric method for estimating the bivariate distribution function conditioned on current or past covariate measures. In addition to this, the primary aim of this paper, however, is to introduce ideas of possible methods for testing the significance of covariates in the estimation of the conditional bivariate distribution function.

INTRODUCTION

In this paper, we revisit the problem of estimating the bivariate distribution function. Unlike previous efforts, our focus is on a regression framework, aiming to estimate these probabilities in the presence of a covariate. Discrete covariates can be included using available methods, such as those introduced by Lin et al. (1999) or de Uña-Álvarez and Meira-Machado (2008), by splitting the sample for each level of the covariate and applying the procedures to each subsample. Surprisingly, there has been scarce research on estimating these probabilities conditional on continuous covariates. A common approach involves parametric specifications of covariate effects, often employing estimators based on Cox's model, with the baseline hazard function estimated using Breslow's method (Breslow, 1972). However, these estimates are prone to model misspecification errors, making standard techniques insufficiently flexible and robust for estimating covariate effects on the target probabilities. The estimators proposed in this paper utilize local smoothing techniques through the introduction of kernel weights based on local constants (Nadaraya-Watson). Right censoring is addressed by appropriately reweighting selected components.

The study is structured as follows. The next section presents the fundamental mathematical background. Subsequent sections detail the application of these methods to bladder cancer data and discuss the main conclusions derived from the analysis.

NOTATION AND ESTIMATORS

Consider n independent and identically distributed pairs of successive failure (gap) times (T_{1i}, T_{2i}), where $1 \leq i \leq n$, with a joint distribution function $F_{12}(x, y)$. These pairs of gap times are subject to univariate right-censoring at times C_i with a distribution function $G(t) = P(C \leq t)$, which we assume to be independent of (T_{1i}, T_{2i}). Consequently, we

observe only $(\widetilde{T}_{1i}, \widetilde{T}_{2i}, \Delta_{1i}, \Delta_{2i})$, where $\widetilde{T}_{1i} = \min(T_{1i}, C_i)$, $\Delta_{1i} = I(T_{1i} \leq C_i)$, $\widetilde{T}_{2i} = \min(T_{2i}, C_{2i})$, $\Delta_{2i} = I(T_{2i} \leq C_{2i})$, and $C_{2i} = (C_i - T_{1i})I(T_{1i} \leq C_i)$. Let $Y = T_1 + T_2$ be the total time, and denote $\widetilde{Y} = \min(Y, C)$.

Since the censoring time is assumed to be independent of the process, the marginal distribution of the first gap time T_1, denoted as F_1, can be consistently estimated by the Kaplan-Meier estimator based on $(\widetilde{T}_1, \Delta_{1i})$. Similarly, the distribution of the total time can be consistently estimated by the Kaplan-Meier estimator based on $(\widetilde{Y}_i, \Delta_{2i})$.

With this notation, the bivariate distribution is expressed as $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$. An estimator for the bivariate distribution function $F_{12}(x, y)$ was introduced by de Uña-Álvarez and Meira-Machado (2008). This estimator is based on inverse probability of censoring weights and is defined as follows:

$$\widehat{F}_{12}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{I\left(\widetilde{T}_{1i} \leq x, \widetilde{T}_{2i} \leq y\right) \Delta_{2i}}{1 - \widehat{G}(\widetilde{Y}_i^-)}.$$

In this manuscript, we are also interested in estimating the conditional distribution function: $F_{12}(x, y \mid Z)$, which can be computed for any times x and y, but conditional on a covariate value denoted by Z. Again, following the notation introduced above, the conditional bivariate distribution is expressed as $F_{12}(x, y \mid Z) = P(T_1 \leq x, T_2 \leq y \mid Z)$.

As previously mentioned, the methods introduced by de Uña-Álvarez and Meira-Machado (2008) can be used to estimate the bivariate distribution function conditional on discrete covariates by splitting the sample for each level of the covariate and applying the procedures to each subsample. The nonparametric estimation of these quantities conditional on continuous covariates can be achieved by estimating the functions $E\left[\varphi_{x,y}(T_1, T_2) \mid Z = z\right]$. To do so, inverse probability of censoring weighting (Satten et al., 2001) can be used to handle right censoring together with kernel smoothing techniques. The proposed conditional estimator:

$$\widehat{F}_{12}(z; x, y) = \sum_{i=1}^{n} W_i(z, b_n) \frac{I(\widetilde{T}_{1i} \leq x, \widetilde{T}_{2i} \leq y)\Delta_{2i}}{1 - \widehat{G}_{Z_i}(\widetilde{Y}_i^-)}.$$

To estimate these quantities, we need to estimate the distribution function of C given Z, denoted as G_Z. This can be accomplished using the Kaplan-Meier estimator as introduced by Beran (1981):

$$\widehat{G}_z(t) = \prod_{Y_i \leq t, \Delta_i = 0} \left[1 - \frac{W_i(z, a_n)}{\sum_{j=1}^{n} I(Y_j \geq Y_i) W_j(z, a_n)}\right]$$

with $W_i(z, a_n) = K\left(\frac{z - Z_i}{a_n}\right) / \sum_{j=1}^{n} K\left(\frac{z - Z_j}{a_n}\right)$. Here, $W_i(z, a_n)$ represents the Nadaraya-Watson weights, K denotes a known probability density function (kernel), and a_n signifies a sequence of bandwidths.

BLADDER CANCER RECURRENCE DATA

In this section, we focus on analyzing the gap times corresponding to the first and second recurrence in a bladder cancer study (Byar et al., 1980). The dataset consists of 118 patients with superficial bladder tumors who underwent transurethral tumor removal. We specifically examine the data of 85 individuals in the placebo and thiotepa treatment groups, comprising 47 and 38 patients, respectively.

The plot depicted in Figure 1 shows the curves of the bivariate distribution function for a fixed grid of values of x, while varying y. It displays the curves obtained by splitting the sample for each level of the treatment group (covariate *rx*). Additionally, the plot includes the curve without the treatment covariate, based on all individuals. This plot serves as a graphical test for checking the relevance of the discrete covariate *rx* in the estimation of the bivariate distribution function.

The results indicate that for lower values of x, specifically $x = 2$ and $x = 6$, the estimates for the two types of treatment are consistent with those obtained using the same method without covariates, based on the entire sample. However, as x increases, a noticeable divergence in the estimates between the treatments emerges. This suggests a growing treatment effect as x increases. Thus, based on graphical inspection, it can be concluded that the treatment effect may be significant at specific higher values of x.

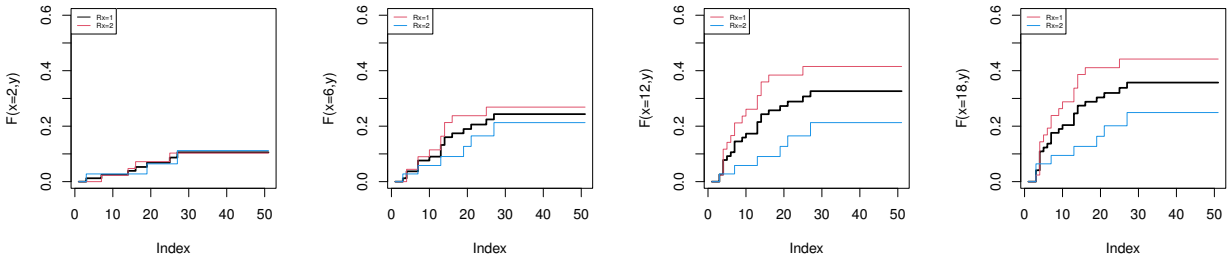

FIGURE 1. Estimates of the bivariate distribution for $rx = \{1, 2\}$ with $x = \{2, 6, 12, 18\}$; black solid line-without covariate, and coloured lines—with the covariate.

Our objective is to build a formal hypothesis test proposed for this context and determine if there is a statistically significant effect of the treatment on the estimation of the bivariate distribution function.

Next, in Figure 2, the plots of the bivariate distribution function for a fixed grid of values of x are shown. The plot shows the curves along y, both without any covariate and conditional on the covariate *size*—largest initial tumor ranging between 1 and 7 cm. Again, our aim is to adapt our proposal to this context. In this case, although the estimates shown in Figure 2 reveal major differences for all displayed values of x, the effect of the covariate *size* is dubious, as we cannot observe a direct influence of the covariate *size*, i.e., it does not show consistent results as the covariate increases.

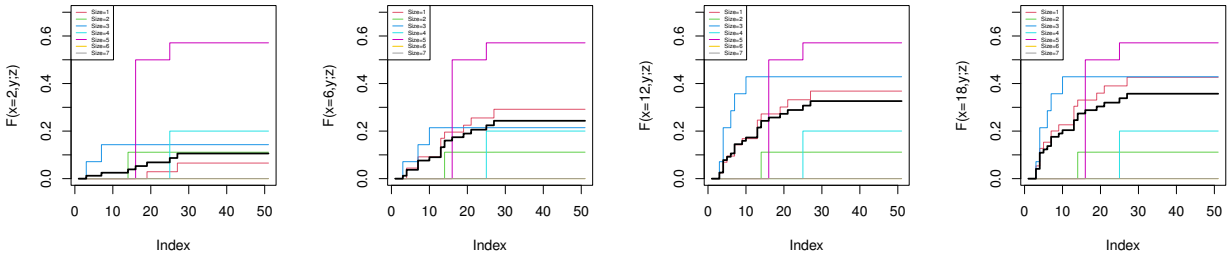

FIGURE 2. Estimates of the bivariate distribution function. The black solid line represents the estimates without covariates, while the coloured lines represent the estimator conditional on the covariate *size* for $x = \{2, 6, 12, 18\}$.

Finally, Figure 3, shows the estimated conditional bivariate distribution $\widehat{F}_{12}(z; x, y)$ for fixed values of (x, y), along size. Additionally, a red straight line displays the estimated bivariate distribution without covariates. We can observe that the effect of the covariate size is more pronounced for values around 5, showing deviations from those obtained without the covariate. However, for these values, the estimates can diverge in opposite directions, resulting in a dubious effect likely due to the small sample size.

COVARIATE SIGNIFICANCE TESTING

In this section, our goal is to introduce methods for assessing the influence of covariates on the estimation of the bivariate distribution function for censored gap times. We provide guidance on the implementation of these methods and propose ideas for evaluating the statistical significance of specific covariates in this context.

For discrete covariates, a localized version of the log-rank test can be utilized. This test compares the estimated bivariate distribution function $F_{12}(x, y)$ across different levels of the discrete covariate, holding x constant. This approach allows for a local assessment of the covariate effects at specific values of x. To derive a comprehensive, global assessment of the covariate effects, the results from the localized tests conducted at various values of x can be aggregated.

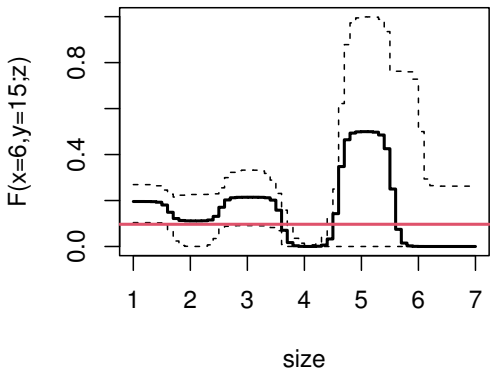

FIGURE 3. Comparison of the estimates of the bivariate distribution given *size* ranging from 1 to 7 cm in increments of 0.1 (black color), without covariates (red color), and the 95% bootstrap confidence bands (dashed lines).

For continuous covariates, a possible approach involves the following steps: First, categorize the continuous covariate z into a small number of levels j. Next, implement a local test, with a fixed x, to compare the j conditional distribution functions, such as the bootstrap test introduced by Villanueva et al. (2019). Finally, construct a global test by aggregating the results from these multiple local tests.

Alternatively, for fixed values of x and y, a Cramér-von Mises type statistic $U = \int_{\tau_1}^{\tau_2} \left(\widehat{F}_{12}(x, y|z) - \widehat{F}_{12}(x, y)\right)^2 dz$ where τ_1 and τ_2 are the bounds of the support of Z, could be used. A near-zero value of U indicates that the covariate has no influence on the estimation of the bivariate distribution function. These methods, adapted from Soutinho and Meira-Machado (2022), may be particularly valuable in practical applications as well.

Although we believe these methods will be well suited for the purpose, details on this issue, such as simulation results, are beyond the scope of the present work and will be reported elsewhere.

ACKNOWLEDGMENTS

This work was suported by Portuguese funds through the Fundação para a Ciência e a Tecnologia, Grant/Award Numbers: UIDB/00013/2020, UIDP/00013/2020 and UIDB/05105/2020.