Chapter 1

Nonparametric estimation of linear functionals of a multivariate distribution under multivariate censoring with applications.

1.1. Introduction

The study of dependent durations arises in many fields. For example, in epidemiology, one can be interested in the lifetimes of twins or the times before healing after using two kinds of drugs on sick eyes. In actuarial sciences, provisioning of pension contracts with a reversion clauses requires to model the dependence between the two lifetimes involved in the contract. The reliability of several components of an engine could be studied in industry. Datasets used to perform statistical estimation in such problems often suffer from incompleteness of some observations due to censoring. Several approaches have been proposed to overcome this difficulty [HOU 00]. Among them, the estimation of the cumulative distribution function in a nonparametric way have been studied [HOU 00, DAB 88, WAN 97, LIN 93, PRE 92, TSA 86, PRU 91, LAA 96, AKR 03]. However, most of the estimators present some drawbacks, for instance, Tsai and al. [TSA 86] studied a method based on the conditional survival functions but the almost sure consistency is slow and the estimator is not symmetric in the sense that is not equivariant under reversal of coordinates. Dabrowska proposed a product limit estimator which can be seen as a bivariate Kaplan-Meier estimator [DAB 88]. The multivariate distribution function is rewritten with quantities which are easily estimated using their empirical counterparts, however, the estimator can assign negative

Chapter written by Olivier LOPEZ and Philippe SAINT-PIERRE.
mass to some observations [PRU 91]. Another approach based on slightly modify
the observations is proposed by van der Laan [LAA 96]. A nonparametric estimator
based on copula modeling is introduced by Wang et Wells [WAN 97] whereas Lin
and Ying [LIN 93] focus on univariate censoring. These estimators are designed to
estimate the cumulative distribution function but are not adapted to estimate the dis-
tribution. They do not define a proper joint distribution as they are not equivariant under
reversal coordinates or assigns negative mass to some points. Therefore, one can not
used them to estimate functionals of the multivariate distribution [GIL 92, FAN 00].
An estimator of the distribution based on weighted mean of marginal distribution
have been introduced by Akritas and Van Keilegom [AKR 03]. This estimator define
a proper distribution, however, the convergence rate does not achieve a \( \sqrt{n} \)-rate,
the computation and the generalization to multivariate case seems difficult. More
recently, Lopez and Saint-Pierre [LOP 12] introduced an estimator which does not
present these drawbacks. This estimator is asymptotically normal, can easily be
extended to multivariate censored lifetimes and defines a proper joint distribution.
Therefore it provides a valuable tool to study functionals of a bivariate distribution.
This paper deals with multivariate right censored survival data but the notations
can be quite complex. In order to facilitate the readability, a bivariate frame-
work is considered through the paper even if the proposed estimation method
can easily be adapted to deal with multivariate survival data. We are interested
in estimating integrals of the distribution of a bivariate right-censored variable
\((T^{(1)}, T^{(2)})\). Let us define \( F(t^{(1)}, t^{(2)}) = \mathbb{P}(T^{(1)} \leq t^{(1)}, T^{(2)} \leq t^{(2)}) \) the joint
distribution function and \( \phi(t^{(1)}, t^{(2)}) \) a measurable function from \( \mathbb{R}^2 \) to \( \mathbb{R} \) such as
\( \iint |\phi(t^{(1)}, t^{(2)})| dF(t^{(1)}, t^{(2)}) < \infty \). The aim is to estimate integrals of \( F \) of the
following form

\[
\xi(\phi) = \mathbb{E}(\phi(T^{(1)}, T^{(2)})) = \iint \phi(t^{(1)}, t^{(2)}) dF(t^{(1)}, t^{(2)}).
\] (1.1)

Such functionals are involved in several quantities useful when analyzing the relation-
ship between bivariate data. For example, \( \mathbb{E}(T^{(1)}T^{(2)}) \) allows to obtain the covari-
ance and the correlation coefficient. The distribution function \( \mathbb{E}(1_{T^{(1)} \leq t^{(1)}, T^{(2)} \leq t^{(2)}}) \),
\( \mathbb{E}(T^{(1)} - T^{(2)}) \), \( \mathbb{E}(\min(T^{(1)}, T^{(2)})) \) or \( \mathbb{P}(T^{(1)} < T^{(2)}) \) are others desirable quan-
ties for the analysis. Kendall’s coefficient and others dependance measures [GIL 83,
FAN 00] can also be derived and will be discuss later in the paper.
In absence of censoring, the estimation of integrals \( \xi \) has been studied by many
authors. In the case of univariate survival data, among others, Stute has provided
several works on the estimation of Kaplan-Meier integrals [STU 95]. In a multivariate
framework, only few papers concern the estimation \( \xi \). Indeed, most of the estima-
tors of the bivariate distribution function are not adapted to this question. A class
of asymptotically normal estimators designed for univariate censoring has been pro-
posed by Lu and Burke [LU 08]. As regards bivariate censoring, Gill [GIL 92] and
Fan and al. [FAN 00] has studied quantities (1.1) by correcting the Dabrowska’s estimator [DAB 88] in view to obtain a regular distribution estimator. The nonparametric estimator proposed by Lopez and Saint-Pierre [LOP 12] is well adapted as it allows to obtain a proper joint distribution. Moreover, it present some desirable properties : classical empirical estimator is obtained in absence of censoring, estimates of marginal distribution lead to Kaplan-Meier estimates and it is equivariant under reversal of coordinates. Several applications useful in multivariate analysis can be derived from such estimator. Estimation of several dependence measures as the Kendall’s coefficient are discussed. A bootstrap procedure for multivariate survival data is derived. The estimator [LOP 12] which defines a proper distribution (as it do not assign negative mass to some points) can be used to simulate samples. The Efron’s resampling procedure [EFR 81] can then be adapted to the multivariate framework. Moreover, a regression model where the response and the covariate are both randomly right-censored is also studied.

The rest of the paper is organized as follows. Section 2 introduces the bivariate distribution estimator to estimate the quantities (1.1). Section 3, provides an asymptotic i.i.d. representation for the estimators. Consistency and asymptotic normality are derived. Section 4 is devoted to the estimation of dependence measures, a bootstrap procedure and regression modeling. Finally, these methods are applied to the lifetimes of twins in section 5.

1.2. nonparametric estimation of the distribution

The aim is to estimate the distribution of a random vector $T = (T^{(1)}, T^{(2)})$, subject to bivariate censoring. Therefore, $T = (T^{(1)}, T^{(2)})$ is not fully observed and the available data in case of right-censoring is an i.i.d. sample $(Y_j, \Delta_j)_{1 \leq j \leq n}$ where

$$Y_j = (Y_j^{(1)}, Y_j^{(2)}) = (\inf(T_j^{(1)}, C_j^{(1)}), \inf(T_j^{(2)}, C_j^{(2)})),$$

$$\Delta_j = (\Delta_j^{(1)}, \Delta_j^{(2)}) = (1_{T_j^{(1)} \leq C_j^{(1)}}, 1_{T_j^{(2)} \leq C_j^{(2)}}),$$

and $C = (C^{(1)}, C^{(2)})$ is the censoring vector. Some identifiability assumptions are needed to derive a consistent estimate of the joint distribution of $T = (T^{(1)}, T^{(2)})$. Multivariate counterpart of traditional assumptions used in univariate survival analysis are considered in the rest of the paper,

$T$ and $C$ are independent variables, \hspace{1cm} (1.2)

$\mathbb{P}(T^{(i)} = C^{(i)}) = 0$ for $i = 1, 2$. \hspace{1cm} (1.3)

In absence of censoring, the joint distribution function can easily be estimated using the empirical function $F_{emp}(t^{(1)}, t^{(2)}) = \frac{1}{n} \sum_{j=1}^{n} 1_{T_j^{(1)} \leq t^{(1)}, T_j^{(2)} \leq t^{(2)}}$. Unfortunately,
the empirical distribution is unavailable if $T = (T^{(1)}, T^{(2)})$ are censored and the empirical distribution of observed variables is a biased estimator. The well known Kaplan-Meier product limit estimate is often used in case of univariate right-censored data. This estimator can be rewritten as a weighted empirical sum [STU 95, SAT 01]. Each observation $Y = (Y^{(1)}, Y^{(2)})$ is weighted in order to compensate the censoring. Lopez and Saint-Pierre [LOP 12] have extended this approach to bivariate right-censored lifetimes. The nonparametric estimator of the distribution function is given by
\[ \hat{F}(y^{(1)}, y^{(2)}) = \frac{1}{n} \sum_{j=1}^{n} \hat{W}_j I_{Y^{(1)}_j \leq y^{(1)}}, Y^{(2)}_j \leq y^{(2)}}, \]
(1.4)
where the weights $\hat{W}_j$ are converging towards some limit weights $W_j$ which must satisfy the following equality
\[ \mathbb{E}[W_j \phi(Y^{(1)}, Y^{(2)})] = \mathbb{E}[\phi(T^{(1)}, T^{(2)})]. \]

The weights $W_j$ represent the mass assigned to the $jth$ observation to correct the effect of bivariate censoring. Assuming the independence between $T$ and $C$, one can easily show that
\[ W_j = \frac{\Delta_j^{(1)} \Delta_j^{(2)}}{\mathbb{E}[\Delta_j^{(1)} \Delta_j^{(2)} | T^{(1)} = y^{(1)}, T^{(2)} = y^{(2)}]} = \frac{\Delta_j^{(1)} \Delta_j^{(2)}}{\mathbb{P}(C^{(1)} > y^{(1)}, C^{(2)} > y^{(2)})}. \]

One can observe that weights only affect mass to doubly uncensored observations. In case of independence between $C^{(1)}$ and $C^{(2)}$, the estimation of weights only consists in estimating the marginal distribution functions. In case of dependence between censoring times, a copula approach can be used to model the cumulative distribution function. In this paper, we suppose that
\[ \mathbb{P}(C^{(1)} > y^{(1)}, C^{(2)} > y^{(2)}) = C(G_1(y^{(1)}), G_2(y^{(2)})) \]
(1.7)
where $C$ is known copula. Indeed, the copula estimation for censored data has already been considered in the literature ([SHI 95]). Moreover, the theoretical results presented in the next section are still valid if the known copula is replaced by a consistent estimator of $C$ [SHI 95]. In this case it is sufficient to estimate the marginal distribution functions $G_1$ and $G_2$ to obtain an estimator of the weights
\[ \hat{W}_j = \frac{\Delta_j^{(1)} \Delta_j^{(2)}}{C(\hat{G}_1(Y^{(1)}_j), \hat{G}_2(Y^{(2)}_j))}. \]
(1.8)
where \( \hat{G}_i \) is the Kaplan-Meier estimator of the survival function of censoring. Indeed, by reversing the labels of \( T^{(i)} \) and \( C^{(i)} \), the censoring \( C^{(i)} \) can be considered as right-censored variable. Therefore, the survival function of the censoring times can be estimated by the Kaplan-Meier estimator, for \( i = 1, 2 \),

\[
\hat{G}_i(t) = \prod_{k: Y_k^{(i)} \leq t} \left( 1 - \frac{\hat{H}_i(Y_k^{(i)})}{\hat{H}_i(Y_k^{(i)})} \right),
\]

where \( \hat{H}_i(t) = \sum_{j=1}^n 1_{Y_j^{(i)} \geq t} \) and \( \hat{H}_0(t) = \sum_{j=1}^n (1 - \Delta_j^{(i)}) 1_{Y_j^{(i)} \geq t} \).

The estimations of weights are used to define estimators of the cumulative distribution function and integrals of the distribution

\[
\hat{F}(y^{(1)}, y^{(2)}) = \frac{1}{n} \sum_{j=1}^n \hat{W}_j 1_{Y_j^{(1)} \leq y^{(1)}, Y_j^{(2)} \leq y^{(2)}},
\]

\[
\hat{\xi}(\phi) = \int \int \phi(y^{(1)}, y^{(2)}) d \hat{F}(y^{(1)}, y^{(2)}) = \frac{1}{n} \sum_{j=1}^n \hat{W}_j \phi(Y_j^{(1)}, Y_j^{(2)}),
\]

These estimates can be generalized to multivariate observations using a multivariate copula function. Left-censored data or left and right censored data can also be considered. Indeed, the left-censored counterpart of the Kaplan-Meier estimator proposed by Gomez and al. [GÓM 92] can be used to estimate the marginal survival function in case of left censoring.

### 1.3. Asymptotic properties

The weights used in equations (1.10) and (1.11) depend on the whole sample. Hence, they are not independent and the central limit theorem does not apply. An asymptotic i.i.d. representation for integrals (1.11) can be obtained under some assumptions related to the copula function and to the belonging of the function \( \phi \) to a Donsker class of functions. Some conditions on regularity of the copula function \( C \) are needed.

**Assumption 1** The function \( (x_1, x_2) \rightarrow C(x_1, x_2) \) is twice continuously differentiable. Moreover, we will denote \( \partial_1 C(x_1, x_2) \) (resp. \( \partial_2 C(x_1, x_2) \)) the partial derivative of \( C \) with respect to \( x_1 \) (resp. \( x_2 \)), and we assume that

\[
\sup_{x_1, x_2} |\partial_1 C(x_1, x_2)| + \sup_{x_1, x_2} |\partial_2 C(x_1, x_2)| + \sum_{i,j \in \{1,2\}} \sup_{x_1, x_2} |\partial_i \partial_j C(x_1, x_2)| < \infty,
\]

and that \( C(x_1, x_2) \neq 0 \) for \( x_1 \neq 0 \) and \( x_2 \neq 0 \).
Assumption 2 Assume that
\[ C(x_1, x_2) \geq x_1^{\alpha_1} x_2^{\alpha_2}, \]
with \( 0 \leq \alpha_i \leq 1 \) for \( i = 1, 2 \).

For instance, this assumption holds in the case of independence copula or copula from Frank’s family with \( \alpha_1 = \alpha_2 = 1 \). Consistency of alternative estimators in multivariate analysis have only been studied on a compact set strictly include in the support of the distribution of \( Y = (Y^{(1)}, Y^{(2)}) \). Such result is restrictive, for instance in a regression framework, since only functions which vanish at the tail of the distribution are considered. Tightness arguments similar to ones used in [GIL 83] to study the consistency of Kaplan-Meier estimator on the whole line can be used to extend the consistency to functions taking non-zero values on the whole support of \( Y \).

Assumption 3 \( \mathcal{F} \) is a Donsker class of locally bounded functions with positive locally bounded envelope function \( \Phi \) such that
\[
\int \frac{\Phi^2(y) dF(y)}{C(G_1(y^{(1)}), G_2(y^{(2)}))} < \infty,
\]
and, for some \( \varepsilon > 0 \) arbitrary small,
\[
\int \Phi(y) \left[ \frac{G_1^{1-\alpha_1}(y^{(1)}) K_1^{1/2+\varepsilon}(y^{(1)})}{G_1^{\alpha_1}(y^{(1)})} + \frac{G_2^{1-\alpha_2}(y^{(2)}) K_2^{1/2+\varepsilon}(y^{(2)})}{G_2^{\alpha_2}(y^{(2)})} \right] dF(y) < \infty
\]
where for \( i = 1, 2 \),
\[
K_i(u) = -\int_0^u \frac{dG_i(t)}{G_i(t) F_i(t)}.\]

The functions \( H_i, G_i \) and \( F_i \) represent the survival function of the variables \( Y^{(i)} \), \( C^{(i)} \) and \( T^{(i)} \) respectively. The moment condition (1.12) corresponds to a finite asymptotic variance and condition (1.13) is useful to obtain the tightness of the process. This assumption is a bivariate adaptation of assumption (1.6) described in [STU 95]. It is shown in [STU 95] that such assumption holds for a large class of distributions, for instance, when the lifetime distributions are sub-exponential the function \( \Phi \) is polynomial. Donsker classes of functions [VAA 96] satisfy a uniform Central Limit Theorem property. Their asymptotic properties is a valuable tool to derive our main theorem. For any function \( \phi \in L^1(\xi) \), it is natural to compare \( \xi(\phi) \) with the following quantity,
\[
\hat{\xi}(\phi) = \frac{1}{n} \sum_{j=1}^{n} W_j \phi(Y_j^{(1)}, Y_j^{(2)}).
\]
Indeed, $\hat{\xi}(\phi)$ is a sum of i.i.d. quantities which tends to $\xi(\phi)$ by the strong law of large numbers. The following i.i.d. representation can be obtained

**Theorem.** Let $\mathcal{F}$ be a class of functions satisfying Assumption 3. Then, under identifiability assumptions 1.2 and 1.7 and under regularity assumptions 1 and 2,

$$\hat{\xi}(\phi) - \tilde{\xi}(\phi) = \frac{1}{n} \sum_{j=1}^{n} \eta(Y_j, \Delta_j, t) \phi(t) C(G_1(t^{(1)}), G_2(t^{(2)})) dF(t) + R_n(\phi),$$

with $\sup_{\phi \in \mathcal{F}} |R_n(\phi)| = o_P(n^{-1/2})$, and where $\eta$ is

$$\eta(Y_j, \Delta_j, t; y) = -\sum_{i=1,2} \frac{\partial_i C(G_1(y^{(i)}), G_2(y^{(2)}))}{C(G_1(y^{(i)}), G_2(y^{(2)}))^2} \left[ \begin{array}{l} \Delta_j \mathbb{1}_{Y_j^{(i)} > y} - G_i(y) \\ \int \mathbb{1}_{Y_j^{(i)} > u} G_i(u \vee y) dF_i(u) - H_i(u) \end{array} \right].$$

The proof of this theorem is obtained and detailed in [LOP 12]. Note that the asymptotic normality of integrals (1.11) can easily be deduced from the i.i.d. representation and the central limit theorem. In some applications, the function $\phi$ belongs to a Donsker class of functions with bounded envelope $\Phi$ satisfying $\Phi(y^{(1)}, y^{(2)}) \equiv 0$ if $y^{(1)} > \tau^{(1)}$ or $y^{(2)} > \tau^{(2)}$ for $\tau^{(i)} < \inf \{y : \mathbb{P}(Y^{(i)} > y) = 0\}, i = 1, 2$. In such cases, the function $\phi$ are bounded with compact support strictly included in the support of the distribution, the assumption 3 is satisfied and then the assumption 2 and 3 are not needed to prove the asymptotic representation.

### 1.4. Statistical applications of functionals

#### 1.4.1. Dependence measures

A popular dependence measure between two variables is the Kendall’s coefficient of concordance,

$$\tau = 2\mathbb{P}[(T_1^{(1)} - T_2^{(1)})(T_1^{(2)} - T_2^{(2)}) > 0] - 1.$$ 

Alternatively, Kendall’s coefficient $\tau$ can be evaluated by integration of the bivariate distribution function,

$$\tau = 4 \iint F(t^{(1)}, t^{(2)}) dF(t^{(1)}, t^{(2)}) - 1.$$ 

This expression is an integrals of the form (1.1). A nonparametric estimate $\hat{\tau}$ can be obtained using the estimator $\hat{F}$. Theorem 1.3 can be used to obtain an asymptotic
representation for \( \hat{\tau} \) ([GRI 13]). The asymptotic normality of \( \hat{\tau} \) can be derived from the Central Limit Theorem.

**Theorem.**— Under identifiability assumptions 1.2 and 1.7 and under regularity assumptions 1 and 2

\[
n^{-1/2}(\hat{\tau} - \tau) \Rightarrow \mathcal{N}(0, Q).
\]

The analytic expression of the limit covariance matrix \( Q \) can be derived from the asymptotic representation and then be estimated from the data by replacing each unknown quantity by its empirical counterpart. Nevertheless, the bootstrap can be recommended to evaluate the law of \( \hat{\tau} \). Kendall’s tau estimation helps to evaluate the dependence between the variables. It can also be useful when studying survival copula models under bivariate censoring. Indeed, Kendall’s coefficient is related to parameter of association for some copula families, for example, Frank or Clayton copulas. The association parameter can be estimated using maximum likelihood estimation [SHI 95]. However, due to censored data, the optimization is very sensitive to the choice of initial value. Our estimate of Kendall’s coefficient can be used as initial value of the association parameter in the maximization algorithm. The integrated hazard correlation, the median concordance [HOU 00] or the dependence measures defined in [GIL 92, FAN 00] can also be estimated using (1.11).

### 1.4.2. Bootstrap

Estimator of integrals (1.1) is proven to be asymptotically normal [LOP 12]. The expression of the asymptotic variance can be derived even if the explicit expression is very complex. The variance can be estimated by replacing the unknown functions by their empirical estimates and asymptotic confidence intervals can be obtained. However, such intervals are often larger than those obtain with bootstrap methods. In case of univariate survival data, a bootstrap procedure introduced by Efron [EFR 81] is consistent and seems to give the best results [AKR 86]. Instead of resampling directly the pairs of observed variables \((Y, \Delta)\), Efron’s methodology consist in simulating lifetimes and censoring times using nonparametric estimators of their distributions. The aim is to adapt this approach to a multivariate framework using the estimator (1.10) (which defines a proper distribution). The lifetimes \((T^{(1)}, T^{(2)})\) are simulated using the estimated distribution \( \hat{F} \). The assumption (1.7) can be used to estimate the distribution function \( G \) of the censoring times \((C^{(1)}, C^{(2)})\),

\[
\hat{G}(y^{(1)}, y^{(2)}) = \mathcal{C}(\hat{G}_1(y^{(1)}), \hat{G}_2(y^{(2)})),
\]

(1.14)

where \( \hat{G}_1 \) and \( \hat{G}_2 \) are the Kaplan-Meier estimators defined in (1.9). Samples of censoring times can be simulated according to the estimator of the distribution \( \hat{G} \). One can then construct samples of observed variables and estimated the integral (1.1) on each sample. The bootstrap procedure is as follow:
Step 0: Compute estimates \( \hat{F} \) and \( \hat{G} \) of the distribution functions of lifetimes \( F \) and censoring times \( G \) using estimators (1.10) and (1.14).

Step 1: For \( b = 1, \ldots, B \),
1) Simulate independent variables \( (T_{j,b}^{(1)}, T_{j,b}^{(2)})_{1 \leq j \leq n} \) according to the distribution \( \hat{F} \),
2) Simulate independent variables \( (C_{j,b}^{(1)}, C_{j,b}^{(2)})_{1 \leq j \leq n} \) according to the distribution \( \hat{G} \),
3) Compute the \( b \)th bootstrap sample \( (Y_{j,b}^{(1)}, Y_{j,b}^{(2)}, \Delta_{j,b}^{(1)}, \Delta_{j,b}^{(2)})_{1 \leq j \leq n} \) where for \( k = 1, 2 \),
   \[ Y_{j,b}^{(k)} = \inf(T_{j,b}^{(k)}, C_{j,b}^{(k)}) \]
   \[ \Delta_{j,b}^{(k)} = 1_{T_{j,b}^{(k)} \leq C_{j,b}^{(k)}} \]
4) Compute the integral (1.11) for the \( b \)th bootstrap sample \( \xi_b \).

Step 2: Use \( \hat{\xi}_1, \ldots, \hat{\xi}_B \) to estimate the distribution of \( \hat{\xi} \).

The sample \( \hat{\xi}_1, \ldots, \hat{\xi}_B \) can be used to estimate the distribution, the mean \( \hat{\mu}(\hat{\xi}) \) or the variance \( \hat{\sigma}(\hat{\xi}) \) of \( \hat{\xi} \) using the empirical estimate. A large sample confidence interval of the mean \( \mu(\xi) \) can be obtained using a normal approximation,

\[
CI_1(\alpha) = [\hat{\mu}(\hat{\xi}) \pm z_{\alpha/2} \hat{\sigma}(\hat{\xi}) / \sqrt{n}].
\]

Another confidence interval can be obtained by estimating the quantiles \( \hat{q}_{\alpha/2} \) and \( \hat{q}_{1-\alpha/2} \) using the empirical distribution. A confidence band is then given by

\[
CI_2(\alpha) = [\hat{q}_{\alpha/2}; \hat{q}_{1-\alpha/2}]\]

In practice, the total mass of the distributions defined by \( \hat{F} \) and \( \hat{G} \) can be lower than 1. A normalization by dividing the distribution by the total mass can be required. Such bootstrap procedure can be used to estimate the distribution of test statistics [GRI 13]. Bias corrections and weighted bootstrap approaches could also be adapted to our framework.
1.4.3. Linear regression

The linear regression model has been studied in the particular case where the covariate is observed and the response is censored [STU 93, DEL 08]. We consider the more general case where the response and the covariate are both randomly censored. Assume that

\[ T^{(1)} = f(\beta, T^{(2)}) + \epsilon \]

where \( \mathbb{E}[\epsilon | T^{(2)}] = 0 \) and \( f \) is a known family of functions depending on a finite dimensional parameter \( \beta \). As the two variables are censored, the nonparametric estimator (1.11) can be used to obtain a consistent least square estimator of \( \beta \),

\[
\hat{\beta} = \arg\min_{\beta} \iint \phi(t^{(1)}, t^{(2)}) d\hat{F}(t^{(1)}, t^{(2)})
\]

with \( \phi(t^{(1)}, t^{(2)}) = (t^{(1)} - f(\beta, t^{(2)}))^2 \). In the case of linear regression, the function \( \phi \) satisfy the assumptions required to obtain the asymptotic representation of the integrals. One can then deduced an asymptotic representation and the asymptotic normality of \( \hat{\beta} \) ([LOP 12]).

**Theorem.** Under identifiability assumptions 1.2 and 1.7 and under regularity assumptions 1 and 2

\[ n^{-1/2}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, R). \]

The complex expression of the asymptotic covariance matrix \( R \) can be derived but can easily be estimated using bootstrap procedure. This approach can be seen as a generalization of the results obtained by [STU 93] when only one variables is censored. Note that prediction measures as the coefficient of determination \( R^2 \) can also be estimate using (1.11). A generalized linear model with censored variables can be considered using the same approach. The extension to a regression model with two or more covariates is possible as the estimators can be adapted to multivariate censored data.

1.5. Illustration

Danish Twin Registry are used to illustrate the analysis of bivariate survival data using integrals of the distribution. The registry include more than 28000 pairs born from 1870-1952 in Denmark. The ascertainment of twins has taken place using different methods leading to several birth cohorts (1870-1910; 1911-1930; 1931-1952). An accurate information is the zygosity: monozygotic (MZ), dizygotic (DZ), unknown zygosity (UZ). We refer to [SKY 02] for detailed descriptions of the registry and the way it has been collected. The survival function can be estimated using estimator (1.10). In this application the censoring times are supposed to be independent. The estimation of survival function are obtained for each birth cohort and represented using contour plot (Fig. 1.1).
Kendall’s tau coefficient between the twin lifetimes is estimated for each birth cohort as well as the regression of $T_1$ on $T_2$. The estimations of Kendall’s tau and regression coefficient displayed in the following tables lead to the same conclusion. The dependence between lifetime is stronger for monozygotic and seems to grow with time.

<table>
<thead>
<tr>
<th>Kendall tau</th>
<th>1870-1910</th>
<th>1911-1930</th>
<th>1931-1950</th>
</tr>
</thead>
<tbody>
<tr>
<td>MZ</td>
<td>0.184</td>
<td>0.289</td>
<td>0.303</td>
</tr>
<tr>
<td>DZ</td>
<td>0.097</td>
<td>0.135</td>
<td>0.057</td>
</tr>
<tr>
<td>UZ</td>
<td>0.162</td>
<td>0.188</td>
<td>0.369</td>
</tr>
</tbody>
</table>

**Table 1.1. Estimation of Kendall tau for each birth cohort.**

<table>
<thead>
<tr>
<th>Regression coefficient</th>
<th>1870-1910</th>
<th>1911-1930</th>
<th>1931-1950</th>
</tr>
</thead>
<tbody>
<tr>
<td>MZ</td>
<td>0.149</td>
<td>0.280</td>
<td>0.354</td>
</tr>
<tr>
<td>DZ</td>
<td>0.088</td>
<td>0.115</td>
<td>0.037</td>
</tr>
<tr>
<td>UZ</td>
<td>0.161</td>
<td>0.212</td>
<td>0.279</td>
</tr>
</tbody>
</table>

**Table 1.2. Estimation of regression coefficients for each birth cohort.**

1.6. Conclusion

We introduce a nonparametric estimator of the distribution of multivariate failure times. This estimator has several interesting properties, it is asymptotically normal, it
generalize the univariate Kaplan-Meier estimator and the empirical function, it define a proper distribution and it is fast to compute. Using this estimator, integrals according to the distribution of failure times can be estimated consistently. We proposed several practical applications of the integrals of a distribution. Dependence measures as the Kendall’s coefficient can be estimated. This measure can be used for studying the dependence but also to estimate the association parameter of a copula. A bootstrap procedure according to Efron’s methodology is provided to derive confidence intervals. The regression model with censored response and variables is also considered. The obtained results can be generalized to the multivariate case and can be adapted to left censoring.

Acknowledgment This work is supported by French Agence Nationale de la Recherche (ANR) ANR Grant ANR-09-JCJC-0101-01. The authors wish to thank the Danish Twin Register, University of Southern Denmark for providing the data.

1.7. Bibliography


Estimation of multivariate distribution under censoring with applications. 13


Index

bivariate censoring, 3
bootstrap, 8
copula, 4, 8
Danish Twin Registry, 10
dependence measure, 7
Kaplan-Meier, 5
Kendall’s coefficient, 7
linear regression, 10
multivariate survival analysis, 2