Nonparametric copula estimation under censoring

Svetlana Gribkova, **Olivier Lopez**

Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie Paris 6
ANR Project Lolita (Dynamic Models for human **Lo**ngevity with **Li**fes**t**yle **A**djustments)

CIRM, February 23th 2016

OUTLINE

OBSERVATIONS

BIVARIATE RIGHT-CENSORED AND LEFT-TRUNCATED DATA

We observe n i.i.d. copies $(Y_i, Z_i, \mu_i, \nu_i, \delta_i, \gamma_i)_{1\leq i\leq n}$, with

$$\begin{cases} Y_i &= \inf(T_i, C_i), \\ Z_i &= \inf(U_i, D_i), \end{cases}$$

where C_i and D_i are censoring variables, and

$$\begin{cases} \delta_i &= \mathbf{1}_{T_i\leq C_i}, \\ \gamma_i &= \mathbf{1}_{U_i\leq D_i}, \end{cases}$$

where $Y_i \geq \mu_i$ and $Z_i \geq \nu_i$.

EXAMPLES

- T = lifetime of a man, U = lifetime of his wife
- Examples of applications : pricing and/or reserving of pensions contracts with reversion clause.
- T and U are not independent.

- T = time between the occurrence of a claim and when its amount is settled, U = total amount paid by the insurer.
- Application : reserving in non-life insurance.
- Once again, T and U are not independent.

EXAMPLES

- T = lifetime of a man, U = lifetime of his wife
- Examples of applications : pricing and/or reserving of pensions contracts with reversion clause.
- T and U are not independent.

- T = time between the occurrence of a claim and when its amount is settled, U = total amount paid by the insurer.
- Application : reserving in non-life insurance.
- Once again, T and U are not independent.

SKLAR'S THEOREM

SKLAR'S THEOREM - DISTRIBUTION FUNCTIONS

Let (T, U) be absolutely continuous variables with d.f. F, $F_T(t) = \mathbb{P}(T \leq t)$, $F_U(u) = \mathbb{P}(U \leq u)$. There exists a unique copula function $\mathfrak{C}$ such that

$$F(t,u) = \mathfrak{C}(F_T(t), F_U(u)).$$

SKLAR'S THEOREM - SURVIVAL FUNCTIONS

Let (T, U) be absolutely continuous variables with survival function S_F, $S_T(t) = \mathbb{P}(T > t)$, $S_U(u) = \mathbb{P}(U > u)$. There exists a unique copula function $\mathfrak{C}_S$ such that

$$S_F(t,u) = \mathfrak{C}_S(S_T(t), S_U(u)).$$

Moreover,

$$\mathfrak{C}_S(u,v) = u + v - 1 + \mathfrak{C}(1-u, 1-v).$$

AIM OF THIS WORK

- Let $F(t,u)=\mathbb{P}(T\leq t, U\leq u)$ denote the bivariate distribution function of (T,U).
- Let $\hat{F}(t,u)$ denote an estimator of F of the type

$$\hat{F}(t,u)=\sum_{i=1}^{n} W_{i,n}\mathbf{1}_{Y_i\leq t, Z_i\leq u}.$$

- Questions :
  - How to estimate $\mathfrak{C}$ with at hand $\hat{F}$?
  - Asymptotic properties ?

OUTLINE

Generalizing the empirical copula

- Due to Sklar's theorem,

$$\mathfrak{C}(u,v) = F(F_T^{-1}(u), F_U^{-1}(v)).$$

- Let $\hat{F}_T(t) = \hat{F}(t,\infty)$ and $\hat{F}_U(u) = \hat{F}(\infty, u)$.
- Define

$$\hat{\mathfrak{C}}(u,v) = \hat{F}(\hat{F}_T^{-1}(t), \hat{F}_U^{-1}(u)),$$

same idea as in Deheuvels (1979) who defined the empirical copula.
- Works if $\hat{F}$ defines a true distribution function.

GENERALIZING THE EMPIRICAL COPULA

- Due to Sklar's theorem,

$$\mathfrak{C}(u,v) = F(F_T^{-1}(u), F_U^{-1}(v)).$$

- Let $\hat{F}_T(t) = \hat{F}(t,\infty)$ and $\hat{F}_U(u) = \hat{F}(\infty,u)$.
- Define

$$\hat{\mathfrak{C}}(u,v) = \hat{F}(\hat{F}_T^{-1}(t), \hat{F}_U^{-1}(u)),$$

  same idea as in Deheuvels (1979) who defined the empirical copula.
- Works if $\hat{F}$ defines a true distribution function.

Alternative procedure

- Another estimator :

$$\tilde{\mathfrak{C}}(u,v) = \sum_{i=1}^{n} W_{i,n} \mathbf{1}_{\hat{F}_T(Y_i) \leq u, \hat{F}_U(Y_i) \leq v}.$$

- $\tilde{\mathfrak{C}}$ is not a copula function in this case.
- $\tilde{\mathfrak{C}}$ is close to $\hat{\mathfrak{C}}$ (the difference is $O_P(n^{-1})$) if both are defined.

Bivariate distribution of (T, U)

- Many estimators of $F(t, u) = \mathbb{P}(T \leq t, U \leq u)$ exist (see e.g. Campbell et Földes, 1982, Dabrowska, 1988, van der Laan, 1994, Prentice, Moodie, Wu, 2004, Lopez, 2013...).
- Many of them do not define probability distributions (for example, they put negative masses to some observations).

PARTICULAR CASES

- Case 1 : $(C, D) \perp (T, U)$ and C and D are linked through a known copula $\mathbb{C}$.
- Lopez and Saint Pierre (2012) :

$$W_{i,n} = \frac{1}{n} \frac{\delta_i \gamma_i}{\mathbb{C}(\hat{S}_C(Y_i), \hat{S}_D(Z_i))},$$

  where $S_C(t) = \mathbb{P}(C \geq t)$, $S_D(t) = \mathbb{P}(D \geq t)$, and $\hat{S}_C$ and $\hat{S}_D$ their Kaplan-Meier estimates.
- Case 2 : $C = D + \varepsilon$, with ε an observed variable.
- Gribkova, Lopez, Saint Pierre (2013) :

$$W_{i,n} = \frac{1}{n} \frac{\delta_i \gamma_i}{\hat{S}_C(\max(Y_i, Z_i - \varepsilon_i))}.$$

ASSUMPTIONS

- Assume that $\mathbb{H}_n(t,u) := \sqrt{n}(\hat{F}(t,u) - F(t,u)) \rightsquigarrow \mathbb{G}_F(t,u)$ in $l^\infty(\mathbb{R}^2)$, where $\mathbb{G}_F(t,u)$ is a tight gaussian process and $\rightsquigarrow$ denotes the weak convergence.

- Let $(T^*, C^*, U^*, D^*) = (F_T(T), F_T(C), F_U(U), F_U(D))$, and let $W^*_{i,n}$ denote the weights of the estimator the distribution of (T^*, U^*) similar to $\hat{F}$, but based on $Y^* = \inf(T^*, C^*)$, $Z^* = \inf(U^*, D^*)$, $\delta^* = \mathbf{1}_{T^* \leq C^*}$ and $\gamma^* = \mathbf{1}_{U^* \leq D^*}$. Assume that $W_{i,n} = W^*_{i,n}$.

ASYMPTOTIC DISTRIBUTION

$n^{1/2}$−CONSISTENCY

Suppose that F has continuous marginal distribution functions and partial derivatives of its copula function exist and are continuous. Then the censored empirical copula process $\{\mathbb{Z}_n(u,v) = n^{1/2}(\hat{\mathfrak{C}}(u,v) - \mathfrak{C}(u,v)),\ 0 \leq u, v \leq 1\}$ converges weakly in $l^\infty([0,1]^2)$ to the tight Gaussian process,

$$\mathbb{Z}_{\mathfrak{C}}(u,v) = \mathbb{Z}^*_{\mathfrak{C}}(u,v) - \partial_1 \mathfrak{C}(u,v)\mathbb{Z}^*_{\mathfrak{C}}(u,1) - \partial_2 \mathfrak{C}(u,v)\mathbb{Z}^*_{\mathfrak{C}}(1,v),$$

where

$$\mathbb{Z}^*_{\mathfrak{C}}(u,v) = \mathbb{G}_F(F_T^{-1}(u), F_U^{-1}(v)).$$

- Tools : essentially Hadamard differentiability.
- Weaker versions : if $\sup_{t,u\in\mathcal{T}\times\mathcal{U}} |\hat{F}(t,u) - F(t,u)| = O_P(\eta_n)$, then $\sup_{u,v\in F_T^{-1}(\mathcal{T})\times F_U^{-1}(\mathcal{U})} |\hat{\mathfrak{C}}(u,v) - \mathfrak{C}(u,v)| = O_P(\eta_n)$.

OUTLINE

Two strategies

- First procedure : consider a smooth estimator F, i.e.

$$\hat{F}_1(t,u) = \sum_{i=1}^{n} W_{i,n} K\left(\frac{t-Y_i}{h}\right) K\left(\frac{u-Z_i}{h}\right),$$

where $K(u) = \int_{-\infty}^{u} k(x)dx$, with k positive function with integral equal to 1 and $h \to 0$, and deduce an estimator $\tilde{\mathfrak{C}}_1$.

- Second procedure : Omelka, Gijbels and Veraverbeke (2009) proposed to transform the observations to make the procedure less sensitive to the marginal distributions, defining $\tilde{\mathfrak{C}}_2(u,v)$ as :

$$\sum_{i=1}^{n} W_{i,n} K\left(\frac{\Phi^{-1}(u) - \Phi^{-1}(\hat{F}_T(Y_i))}{h}\right) K\left(\frac{\Phi^{-1}(v) - \Phi^{-1}(\hat{F}_U(Z_i))}{h}\right),$$

with Φ a distribution function.

Two strategies

- First procedure : consider a smooth estimator F, i.e.

$$\hat{F}_1(t,u) = \sum_{i=1}^n W_{i,n} K\left(\frac{t-Y_i}{h}\right) K\left(\frac{u-Z_i}{h}\right),$$

  where $K(u) = \int_{-\infty}^u k(x)dx$, with k positive function with integral equal to 1 and $h \to 0$, and deduce an estimator $\tilde{\mathfrak{C}}_1$.

- Second procedure : Omelka, Gijbels and Veraverbeke (2009) proposed to transform the observations to make the procedure less sensitive to the marginal distributions, defining $\tilde{\mathfrak{C}}_2(u,v)$ as :

$$\sum_{i=1}^n W_{i,n} K\left(\frac{\Phi^{-1}(u) - \Phi^{-1}(\hat{F}_T(Y_i))}{h}\right) K\left(\frac{\Phi^{-1}(v) - \Phi^{-1}(\hat{F}_U(Z_i))}{h}\right),$$

  with Φ a distribution function.

THEORETICAL RESULTS

- Let $\mathbb{Z}_n^i(u,v) = n^{1/2}(\tilde{\mathfrak{C}}_i(u,v) - \mathfrak{C}(u,v))$ for $i = 1, 2$.

$n^{1/2}$–CONSISTENCY

Under some assumptions,

$$\sup_{u,v} |\mathbb{Z}_n^i(u,v) - \mathbb{Z}_n(u,v)| = o_P(1),$$

and the asymptotic distribution can then be deduced from the previous theorem.

ASSUMPTIONS ON THE COPULA FUNCTION

BEHAVIOR CLOSE TO THE BOUNDARIES

Assume that $\mathfrak{C}$ is twice continuously differentiable on $]0,1[^2$, and that

$$\frac{\partial^2 \mathfrak{C}(u,v)}{\partial u^2} = O\left(\frac{1}{u(1-u)}\right), \quad \frac{\partial^2 \mathfrak{C}(u,v)}{\partial v^2} = O\left(\frac{1}{v(1-v)}\right),$$
$$\frac{\partial^2 \mathfrak{C}(u,v)}{\partial u \partial v} = O\left(\frac{1}{\sqrt{uv(1-u)(1-v)}}\right).$$

- See Omelka et al. (2009)

ASSUMPTIONS ON THE CENSORING

- Essentially, we require asymptotic i.i.d. representations of the type

$$\sum_{i=1}^{n}[W_{in} - W_i]\psi(Y_i, Z_i) = \frac{1}{n}\sum_{i=1}^{n}\eta^{\psi}(Y_i, Z_i, \delta_i, \gamma_i) + R_n(\psi),$$

where $\sup_{\psi \in \mathcal{F}} |R_n(\psi)| = o_P(n^{-1/2})$, $E[\eta^{\psi}(Y_i, Z_i, \delta_i, \gamma_i)] = 0$, and $nW_i = \lim_{n\to\infty} nW_{i,n}$.

- In the particular case $C = D$, and under the assumption $C \perp (T, U)$, such representations go back to Stute (1996), since they are derived from the Kaplan-Meier estimator.

Assumptions on the censoring

- To obtain $n^{1/2}-$consistency on $[0,1]^2$, we require assumptions on the tails of the distribution of (T,U) and (C,D).
- Example in the case $C=D$:

$$\begin{aligned} \int \frac{dF(t,u)}{S_C(\max(t,u))} &< \infty, \\ \int \frac{\mathcal{C}^{1/2+\varepsilon}(\max(t,u))dF(t,u)}{[S_C(\max(t,u))]} &< \infty, \end{aligned}$$

where $\mathcal{C}$ is a function that tends to infinity when $t \to \infty$.

Real data example

- 11 947 contracts from a Canadian insurer, observed between December 29th, 1988 and December 31st, 1993.
- $98,2\%$ observations are censored.
- Copula models have been proposed to study this population (Frees et al., 1996, Carriere, 2000, Luciano et al., 2008...)

Goodness-of-fit

- $$H_0 : \mathfrak{C} \in \{\mathfrak{C}_\theta : \theta \in \Theta\},$$
  against
  $$H_1 : \mathfrak{C} \notin \{\mathfrak{C}_\theta : \theta \in \Theta\}.$$
- Idea : evaluate $T_n = d(\hat{\mathfrak{C}}(u, v), \mathfrak{C}_{\hat{\theta}}(v))$, and reject H_0 if $T_n > s_\alpha$.
- Critical value computed by bootstrap to ensure $\mathbb{P}(T_n > s_\alpha) \approx \alpha$.

GOODNESS-OF-FIT

- $$H_0 : \mathfrak{C} \in \{\mathfrak{C}_\theta : \theta \in \Theta\},$$
  against
  $$H_1 : \mathfrak{C} \notin \{\mathfrak{C}_\theta : \theta \in \Theta\}.$$

- Idea : evaluate $T_n = d(\hat{\mathfrak{C}}(u,v), \mathfrak{C}_{\hat{\theta}}(v))$, and reject H_0 if $T_n > s_\alpha$.
- Critical value computed by bootstrap to ensure $\mathbb{P}(T_n > s_\alpha) \approx \alpha$.

GOODNESS-OF-FIT

- $$H_0 : \mathfrak{C} \in \{\mathfrak{C}_\theta : \theta \in \Theta\},$$

  against

  $$H_1 : \mathfrak{C} \notin \{\mathfrak{C}_\theta : \theta \in \Theta\}.$$

- Idea : evaluate $T_n = d(\hat{\mathfrak{C}}(u, v), \mathfrak{C}_{\hat{\theta}}(v))$, and reject H_0 if $T_n > s_\alpha$.
- Critical value computed by bootstrap to ensure $\mathbb{P}(T_n > s_\alpha) \approx \alpha$.

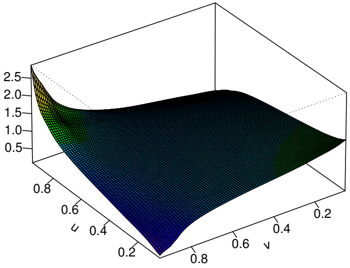
GOODNESS-OF-FIT BASED ON $\hat{\mathfrak{C}}$ - CANADIAN DATASET

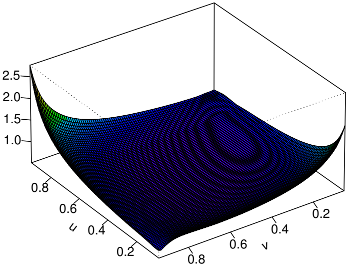
- Compare (using $\|\cdot\|_\infty$) the nonparametric copula with the one from the parametric model.

Model	Test statistic	95% quantile	p-value
Clayton	$7.6e^{-4}$	$2.08e^{-3}$	0.391
Frank	$3.6e^{-4}$	$9.2e^{-4}$	0.416
Nelsen 4.2.20	$1.16e^{-4}$	$1.37e^{-3}$	0.103

TABLE: Goodness-of-fit procedure based on the empirical copula for three copula models (Clayton, Frank, Nelsen 4.2.20), $p-$values obtained by bootstrap.

Nonparametric copula density estimation





- Copula density estimation (right-hand side : Omelka, Gijbels, Veraverke transformation)

CHOICE OF THE BANDWIDTH FOR THE SMOOTH VERSION

- Among the parametric models we considered, Frank's copula seems the most appropriate.
- The estimated copula density also "looks like" the density of a Frank copula.
- Let $\mathfrak{C}_{\hat{\theta}}$ the copula function estimated assuming that Frank's model holds.
- We consider a finite set of bandwidth $\mathcal{H}$.
- We select $\hat{h}$ as the minimizer of a distance $d(\tilde{\mathfrak{C}}_h, \mathfrak{C}_{\hat{\theta}})$.

CHOICE OF THE BANDWIDTH FOR THE SMOOTH VERSION

- Among the parametric models we considered, Frank's copula seems the most appropriate.
- The estimated copula density also "looks like" the density of a Frank copula.
- Let $\mathfrak{C}_{\hat{\theta}}$ the copula function estimated assuming that Frank's model holds.
- We consider a finite set of bandwidth $\mathcal{H}$.
- We select $\hat{h}$ as the minimizer of a distance $d(\tilde{\mathfrak{C}}_h, \mathfrak{C}_{\hat{\theta}})$.

CONCLUSION

- Results for estimating a copula function based on an estimator $\hat{F}$: if one wishes to consider another estimate, one has simply to check the conditions on the weights $W_{i,n}$.
- More details :
  S. Gribkova, O. Lopez (2015) *Nonparametric copula estimation under bivariate censoring*, to appear in Scand. Journ. of Stat.

- Extensions, further work :
  - taking covariates into account ;
  - for longevity issues in insurance, take into account the fact that the marginal distributions and the dependence structure evolve from one generation to another ;
  - for non-life insurance applications, considering the heterogeneity of the individuals (clustering).

CONCLUSION

- Results for estimating a copula function based on an estimator $\hat{F}$: if one wishes to consider another estimate, one has simply to check the conditions on the weights $W_{i,n}$.
- More details :
  S. Gribkova, O. Lopez (2015) *Nonparametric copula estimation under bivariate censoring*, to appear in Scand. Journ. of Stat.
- Extensions, further work :
  - taking covariates into account ;
  - for longevity issues in insurance, take into account the fact that the marginal distributions and the dependence structure evolve from one generation to another ;
  - for non-life insurance applications, considering the heterogeneity of the individuals (clustering).

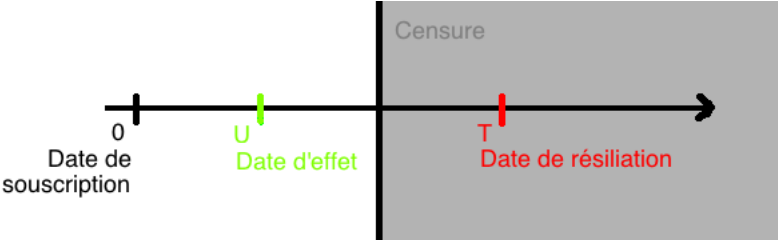
CONCLUSION

- Results for estimating a copula function based on an estimator $\hat{F}$: if one wishes to consider another estimate, one has simply to check the conditions on the weights $W_{i,n}$.
- More details :
  S. Gribkova, O. Lopez (2015) *Nonparametric copula estimation under bivariate censoring*, to appear in Scand. Journ. of Stat.

- Extensions, further work :
  - taking covariates into account ;
  - for longevity issues in insurance, take into account the fact that the marginal distributions and the dependence structure evolve from one generation to another ;
  - for non-life insurance applications, considering the heterogeneity of the individuals (clustering).

EXAMPLE OF APPLICATION IN NON-LIFE INSURANCE

- Online insurance subscription.
- Two duration variables :
  - $T =$ lifetime of the subscribed contract
  - $U =$ time at which the contract will be effective



- Specific form of the censoring.
- Presence of covariates that have influence on the dependence structure (conditional copulas)

Thank you for your attention !