Statistical analysis of competing risks in an unemployment study

Petr Volf[1]

Abstract. This study continues in the theme of contribution Volf (2010) and extends it considerably. While the previous paper was devoted mainly to the analysis of real incidence of competing events, the present one is much more concerned with the analysis of dependence of these events (more precisely, of random variables - latent times to events). To do it, we discuss first the problem of identifiability of marginal and joint distributions of competing random variables. Then, the copula models are utilized in order to express the dependence. Finally, the Gauss copula is used to solution of a real example with unemployment data.

Keywords: statistics, survival analysis, competing risks, copula, unemployment data.

JEL classification: C41, J64
AMS classification: 62N02, 62P25

1 Introduction

The problem of competing risks, except in the field of reliability, biostatistics and medical studies, is also often studied in demography, labour statistics, and in econometrics generally. In the insurance mathematics the setting of competing risks is sometimes called the multiple decrement model (c.f. Arnold and Brockett, 1983). The interest in the problem dates back to 70-ties of the last century. From the beginning it was revealed that in the competing risks setting the background model may not be identifiable. A proof and an example of this phenomenon is given in Tsiatis (1975), some instances of identifiable (or not) models are presented in Basu and Ghosh (1978). In these classical studies the notion of copula has not been used yet. Just later it was recognized that the use of copula for multi-dimensional continuous distribution can lead to a 'nice' closed form of the model. Therefore we are facing the problem of reasonable copula selection. Fortunately, it is known (cf. Zheng and Klein, 1995), that the selection of copula type is not crucial to a good fit of the model, that the finding proper value of its parameter (connected with correlation) is much more important.

The outline of the paper is the following: The next section introduces the scheme of competing risks, presents the method of analysis of competing events incidence, and points to the problem of possible non-identifiability of their marginal distributions. We shall mention also certain identifiability results in the framework of regression models. Then the notion of copula is recalled and used in competing risks model formulation. As a particular example, in Section 3 the Gauss copula is introduced and the procedure of simultaneous maximum likelihood estimation of marginal distributions and correlation is shown. This approach is applied in Section 4 containing a real example. We use the data on unemployed people from Han and Hausman (1990). There are two competing chances to re-gain an employment, we analyze their marginal and joint distribution (i.e. also their dependence), with the aid of Gauss copula model.

2 Competing risks and incidence

Let us recall the competing risks situation: Certain event (e. g. a failure of a device) can be caused by K reasons. It means that there are K (possibly dependent) random variables $T_j, j = 1, ..., K$, some-

[1] Institute of Information Theory and Automation, Czech Academy of Sciences, Pod vodárenskou věží 4, Praha 8, Czech Republic, volf@utia.cas.cz

times accompanied by a variable C of random right censoring (C is then independent of all T_j). Let $\overline{F}_K(t_1, ..., t_K) = P(T_1 > t_1, ..., T_K > t_K)$ be the joint survival function of $\{T_j\}$. However, instead the 'net' survivals T_j we standardly observe just 'crude' data (sometimes called also 'the identified minimum') $Z = \min(T_1, ..., T_K, C)$ and the indicator $\delta = j$ if $Z = T_j$, $\delta = 0$ if $Z = C$. Such data lead us to direct estimation of the distribution of $Z = \min(T_1, ..., T_K)$, for instance its survival function $S(t) = P(Z > t) = \overline{F}_K(t, ..., t)$. Further, we can estimate so called **incidence densities**

$$f_j^*(t) = dP(Z = t, \delta = j) = -\frac{\partial \overline{F}_K(t_1, ..., t_K)}{\partial t_j}|(t_1 = ... = t_K = t),$$

and also their integrals, **cumulative incidence functions**

$$F_j^*(t) = \int_0^t f_j^*(s)\, ds = P(Z \le t, \delta = j).$$

Notice that $\lim F_j^*(t) = P(\delta = j) < 1$ if $t \to \infty$, $S(t) = 1 - \sum_{j=1}^K F_j^*(t)$.

A more practical form of the cumulative incidence function (more convenient for statistical estimation) uses so called **cause–specific hazard functions** for events $j = 1, 2, \ldots, K$:

$$h_j^*(t) = \lim_{d \to 0} \frac{P(t \le Z < t + d,\, \delta = j \,|\, Z \ge t)}{d}.$$

Overall hazard rate for $Z = \min(T_1, ..., T_K)$ is then:

$$h^*(t) = \lim_{d \to 0} \frac{P(t \le Z < t + d \,|\, Z \ge t)}{d} = \sum_{j=1}^K h_j^*(t),$$

by integration the cumulated hazard rates $H_j^*(t)$, $H^*(t)$ are obtained. Consequently, $S(t) = P(Z > t) = \exp(-H^*(t))$. Then $f_j^*(t) = h_j^*(t) \cdot S(t)$ and the cumulative incidence functions can be written as

$$F_j^*(t) = P(Z \le t, \delta = j) = \int_0^t S(s) \cdot h_j^*(s)\, \mathrm{d}s.$$

As both components, i.e. S and h_j^*, are estimable consistently by standard survival analysis methods, it follows that there also exist consistent estimates of F_j^*, see for instance Lin (1997), Scheike and Zheng (2008) in a regression context, also Volf (2010).

2.1 Problem of non-identifiability

However, in general, from data (Z_i, δ_i), $i = 1, \ldots, N$ it is not possible to identify neither marginal nor joint distribution of $\{T_j\}$. A. Tsiatis (1975) has shown that for arbitrary joint model we can find a model with independent components having the same incidences, i.e. we cannot distinguish the models. Namely, this 'independent' model is given by cause-specific hazard functions $h_j^*(t)$. In other words, even if the model is parametrized and the MLE yields consistent estimates, in general we do not know parameters of which model are estimated.

The situation can be better in the case of a regression model, because the covariates provide an additional information, especially when their structure is rich enough. There are numerous results showing conditions for full model identifiability, let us mention here Heckman and Honoré (1989) and their proof of identifiability in the Cox or the AFT model cases. Lee (2006) investigated more general transformation models of regression. Berg et al. (2007) have studied two competing transition rates from unemployment state. They have used a discrete-time multiplicative regression model with latent heterogeneities and their main identifying assumption is basically the same as that of Heckman and Honoré, namely that exit rates should not vary with the observed covariates in exactly the same way. However, all these studies rely on an assumption that the dependence structure (in the next section given by a copula parameter) does not change with covariates. If it is not the case, the problem of identifiability arises anew.

Further, as a consequence of the Tsiatis (1975) result, in competing risks models without regressors it is necessary to make certain functional assumptions about the form of both marginal and joint distribution in order to identify them. Several such cases are studied in Basu and Ghosh (1978) and in some other papers.

2.2 Competing risks and copula

In the sequel we shall consider just 2 competing events, i.e. random variables S, T, censoring variable C, and observed data – realizations of N i.i.d. random variables $Z_i = min(S_i, T_i, C_i), \delta_i = 1, 2, 0$, $i = 1, 2, ..., N$. The notion of copula offers a way how to model multivariate distributions, namely the joint distribution function $F_2(s, t)$ of S, T:

$$F_2(s, t) = C(F_S(s),\, F_T(t),\, \theta), \tag{1}$$

F_S, F_T are marginal distribution functions of S, T, $C(u, v, \theta)$ is a copula, i.e. a two-dimensional distribution function on $[0, 1]^2$, with uniformly on $[0, 1]$ distributed marginals U, V, θ is a copula parameter. The parameter is, as a rule, uniquely connected with correlation of U, V, hence also with correlation of S, T. It is seen that the use of copula allows to model the dependence structure separately from the analysis of marginal distributions. Hence, the identifiability of the copula (and its parameter) and marginals can be considered as two separate steps.

Zheng and Klein (1995) proved that when the copula is known, the marginal distributions are estimable consistently (and then the joint distribution, too, from (1)), even in non-parametric (so that quite general) setting. However, in general, also value of θ is needed, because (again due to Tsiatis, 1975) without fully determined copula we are not able to distinguish between the 'true' model and corresponding independent one. On the other hand, Zheng and Klein also discussed importance of proper selection of copula form. As it has already been said, the knowledge (or a good estimate) of parameter θ is much more crucial for correct model of joint distribution. As a consequence, because the knowledge of copula type is still an unrealistic supposition, we can try to use certain sufficiently flexible class of copulas, as approximation, and concentrate to reliable estimation of its parameter.

3 Gauss copula

There exist a large number of different copula functions, among them for instance a set of Archimedean copulas. However, we concentrate here to one rather universal and flexible copula type, namely to Gauss copula (in our setting used for connecting just two random variables). Let X, Y be standard normal random variables $\sim N(0, 1)$ tied with (Pearson) correlation $\rho = \rho(X, Y)$. We denote ϕ, φ univariate standard normal distribution function and density and by $\phi_2(x, y)$, $\varphi_2(x, y)$ corresponding 2-dimensional functions. Then

$$\varphi_2(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big\{ -\frac{1}{2}\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}\Big\}$$

with $\boldsymbol{x} = (x, y)'$ and Σ the covariance matrix $[1, \rho; \rho, 1]$. If we define $U = \phi(X),\ V = \phi(Y)$, we obtain a 2-dimensional distribution on $(0, 1)^2$ with the copula

$$C(u, v) = \phi_2\big(\phi^{-1}(u),\, \phi^{-1}(v)\big). \tag{2}$$

Naturally, $\rho(U, V) \neq \rho(X, Y)$ (though they are rather close, as a rule), while Spearman's correlations coincide, namely $\rho_{\mathrm{SP}}(X, Y) = \rho_{\mathrm{SP}}(U, V) = \rho(U, V)$. We can connect also density functions. Let $c(u, v)$ be the joint density of (U, V), then

$$c(u, v) = \frac{\varphi_2(x, y)}{\varphi(x) \cdot \varphi(y)},$$

again with $u = \phi(x),\ v = \phi(y)$. As we are primary interested in the model for dependence of competing variables S, T, let us assume that their joint distribution function is given by Gauss copula (2),

$$F_2(s, t) = \phi_2\big(\phi^{-1}(F_S(s)),\, \phi^{-1}(F_T(t))\big), \tag{3}$$

and $S = F_S^{-1}(\phi(X)), T = F_T^{-1}(\phi(Y))$. Again $\rho_{\mathrm{SP}}(S, T) = \rho_{\mathrm{SP}}(U, V)$, and "initial" $\rho = \rho(X, Y)$ is the only parameter describing the dependence of S and T. It, naturally, differs from $\rho(S, T)$, however, all values $\rho(S, T)$ can be achieved by convenient choice of $\rho(X, Y)$. Let us remark here that the real dependence among S, T can be much more complicated, nevertheless the use of Gauss copula offers here certain rather simple and sufficiently flexible (as regards the correlation) set of distributions.

3.1 Estimation in Gauss copula model

When parameter ρ is known, copula (2) is fully defined and from Zheng, Klein (1995) it follows that the distribution of (S, T) can be estimated, in parametric and even non-parametric setting. However, without knowledge of ρ nonparametric model is not identifiable and in the parametric setting explicit proofs of identifiability are available for just certain types of marginal distributions. That is why in the following example we shall assume log-normal marginal distributions. Their identifiability in a framework of Gauss copula follows from the result of Basu and Ghosh (1978, Sect. 7), as log-normal variables are the same monotone transformation of normal variables. It also means that after log transformation of data we can work with Gauss marginal distributions. Naturally, the fit of chosen model to the data has to be tested.

The estimation procedure will be based on the maximum likelihood method. The data are (Z_i, δ_i), $i = 1, \ldots, N$, the likelihood function then has the form

$$L = \prod_{i=1}^{N} \left\{ -\frac{\partial}{\partial s} \overline{F}_2(s,t) \right\}^{I[\delta_i=1]} \cdot \left\{ -\frac{\partial}{\partial t} \overline{F}_2(s,t) \right\}^{I[\delta_i=2]} \cdot \overline{F}_2(s,t)^{I[\delta_i=0]},$$

evaluated at $s = t = Z_i$, with $\overline{F}_2(s,t) = P(S > s, T > t) = 1 - F_S(s) - F_T(t) + F_2(s,t)$. From transformation (3) it follows that $F_2(s,t) = \phi_2(x,y)$ with $x = \phi^{-1}(F_S(s))$, $y = \phi^{-1}(F_T(t))$. Hence, when we put $X_i = \phi^{-1}(F_S(Z_i))$, $Y_i = \phi^{-1}(F_T(Z_i))$, we obtain after some computation – integration of 2-dimensional Gauss density $\varphi_2(x,y)$, that

$$L = \prod_{i=1}^{N} \left\{ f_S(Z_i) \left[1 - \phi_1(Y_i; \rho X_i, 1-\rho^2) \right] \right\}^{I[\delta_i=1]} \cdot$$

$$\cdot \left\{ f_T(Z_i) \left[1 - \phi_1(X_i; \rho Y_i, 1-\rho^2) \right] \right\}^{I[\delta_i=2]} \cdot \left\{ 1 - F_S(Z_i) - F_T(Z_i) + \phi_2(X_i, Y_i \right\}^{I[\delta_i=0]}, \qquad (4)$$

where $\phi_1(x; \mu, \sigma^2)$ denotes the distribution function of normal distribution $(N(\mu, \sigma^2)$, evaluated at x. Parameter ρ is hidden in ϕ_1 and in ϕ_2. Distributions of S and T are present both explicitly and also implicitly, in transformed X_i, Y_i. It is seen that the problem of maximization is not an easy task and has to be solved by a convenient search procedure.

4 Application

Han and Hausman (1990) have analyzed the data on unemployment duration, with two competing chances to leave the unemployment state, either by obtaining a new job (our variable S) or by a recall to the former employer (variable T). Some histories of unemployment were censored (by a variable C), independently, i.e. terminated from other reasons. The data on together 1051 people are collected in Table III of Han and Hausman (with several insignificant misprints which we have corrected), the time is in fact discrete, data are aggregated to weeks. The data show some (rare, however) non-regularities, visible also in the graph of cause-specific hazard rates in Figure 1, for instance significantly larger numbers of events in 26-th week which may be related to a change of support after the first half-year of unemployment. We analyzed first the incidence of both competing variables separately. It was assumed that cause-specific hazards were constant during each week and that each person could experience just one (potential) event in a week (i.e. that for each person i the values S_i, T_i, C_i – though two of them 'not-realized' – must be different). Figure 1 shows estimated cause-specific hazard rates, their cumulated sums, and finally estimated cumulated incidence functions, together with estimated distribution function of $\min(S, T)$.

Han and Hausman had also an information on several covariates which was not available to us. They therefore used a discrete version of Cox regression model. It has to be said that they also used certain not fully correct approximations, for instance substituting Gumbel distribution by the Gauss one. In fact their estimate of correlation was not significant, i.e. they actually have shown that in the framework of their model (with their approximations) the risks are conditionally independent, given the covariates.

We concentrated to the analysis of competing risks in the framework of the Gauss copula model. Further, we assumed log-normal marginal distributions of S and T in order to assure the identifiability. Reasonability of such an assumption was checked in the following graphical way: $M = 500$ samples of new competing-risks data, each of extent $N = 100$, were randomly generated from estimated model. From each sample, the cumulated incidence function was estimated. They are plotted on Figure 2 and

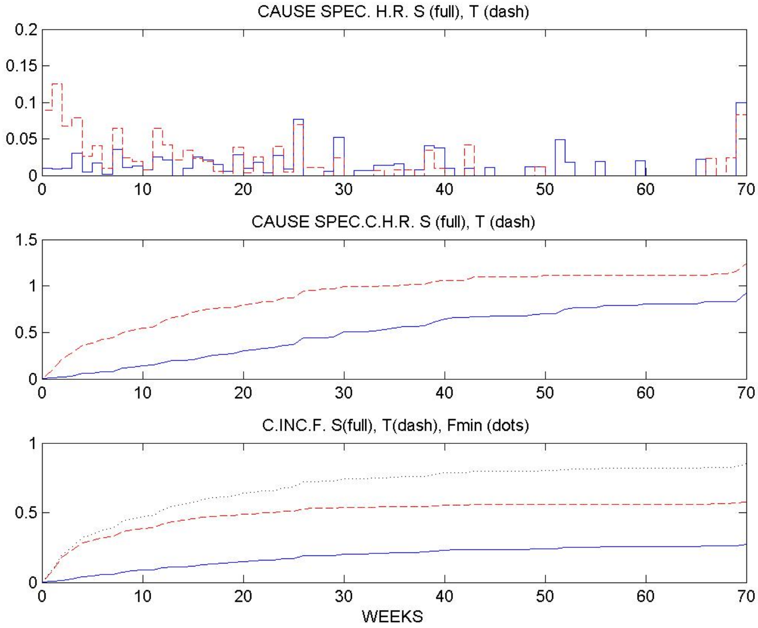

Figure 1 Estimated cause-specific hazard functions (above), cumulated cause-specific hazards (middle), cumulated incidence functions (below), for S (full) and for T (dashed curves). Dotted is estimate of F_{min}.

compared with cumulated incidence functions obtained from real data (thick curves). It is seen that the 'clouds' of generated curves are around real ones, in both cases. It could be taken as a graphical goodness-of-fit test supporting our idea of log-normal distributions.

A random search procedure for maximum of likelihood (4) brought us to the following final results:

$$\mu_S = 2.7966,\ \sigma_S = 1.2299,\ \mu_T = 2.4749,\ \sigma_T = 1.5852,\ \rho = 0.8620.$$

Here $\rho = \rho(X, Y)$ of corresponding standard Gauss variables (see definition of Gauss copula in preceding section), while numerically computed $\rho(U, V) = 0.849$ and, finally, $\rho(S, T) = 0.519$. It is large positive, indicating strong dependence between both competing variables. As the solution of the MLE was based on a numerical optimization procedure, we were not able to assess confidence intervals of involved parameters. In fact, graphs on Figure 2 provide at least partial information on the whole model reliability. As an alternative we have considered also Weibull marginal distributions, also connected by Gauss copula. While estimated correlation was comparable, achieved maximum of likelihood was smaller and the graphical comparison as in Figure 2 indicated significantly worse fit of this model.

5 Conclusion

We have studied the problem of competing risks with the focus on assessing the dependence of competing random variables and identifying their marginal as well as joint distributions. The joint distribution was expressed with the aid of a copula, the case of Gauss copula was investigated in more details. Proposed model was then utilized in an example with real unemployment data. Statistical analysis revealed positive correlation between times to both competing events. On the other hand, even the experience with artificial data indicates that in the framework of chosen Gauss copula competing risks model the log-likelihood is flat and the convergence of computations to its (hardly detectable) maximum is rather slow.

Acknowledgements

The research was supported by the grant No. 13-14445S of the Grant Agency of the Czech Republic.

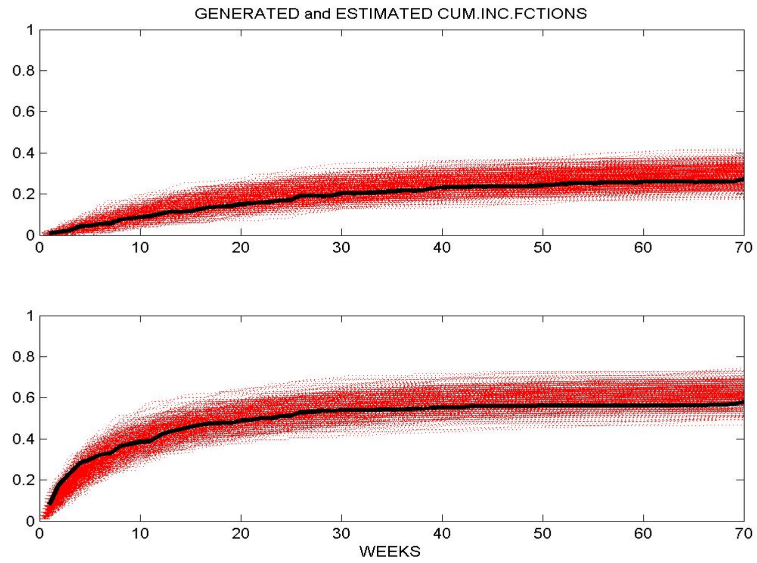

Figure 2 Set of cumulated incidence functions estimated from generated data, above for S, below for T, thick curves – cumulated incidence functions from real data (the same as in Figure 1 below)

References

[1] Arnold, B.C. and Brockett, P.L.: Identifiability for dependent multiple decrement/competing risk model, *Scand. Actuarial J.* 1983, 117–127.

[2] Basu, A.P. and Ghosh, J.K.: Identifiability of the Multinormal and Other Distributions under Competing Risks Model, *Journal of Multivariate Analysis* **8** (1978), 413–429.

[3] Van den Berg, G.J., van Lomwel, A.G.C., and van Ours, J.C.: Nonparametric estimation of a dependent competing risks model for unemployment durations, *Empirical Economics* **34** (2008), 477-491

[4] Han, A. and Hausman J.A.: Flexible parametric estimation of duration and competing risk models, *J. of Applied Econometrics* **5** (1990), 1–28.

[5] Heckman, J.J. and Honoré, B.E.: The identifiability of the competing risks model, *Biometrika* **76** (1989), 325–330.

[6] Lee, S.: Identifcation of a competing risks model with unknown transformations of latent failure times, *Biometrika* **93** (2006), 996–1002.

[7] Lin, D.Y.: Non-parametric inference for cumulative incidence functions in competing risks studies, *Statistics in Medicine* **16** (1997), 901–910.

[8] Scheike, T.H. and Zhang, M.: Flexible competing risks regression modelling and goodness-of-fit, *Lifetime Data Analysis* **14** (2008), 464–483.

[9] Tsiatis, A.: A nonidentifiability aspects of the problem of competing risks, *Proc. Nat. Acad. Sci. USA* **72** (1975), 20–22.

[10] Volf, P.: On statistical modeling of incidence of competing events, with application to labor mobility analysis. In: *Proceedings of the MME 2010*, JCU Ceske Budejovice, 2010, 670–675.

[11] Zheng, M. and Klein, J.P.: Estimates of marginal survival for dependent competing risks based on an assumed copula, *Biometrika* **82** (1995), 127–138.

33rd International Conference

Mathematical Methods in Economics

MME 2015

Conference Proceedings

Cheb, Czech Republic
September 9 – 11, 2015

Faculty of Economics
University of West Bohemia, Plzeň

Editors
David Martinčík
Jarmila Ircingová
Petr Janeček

Technical editor
David Martinčík

Cover photo
Hana Ovesleová

First edition
ISBN 978-80-261-0539-8