Inference for a general semi-Markov model and a sub-model: Independent competing risks

Catherine Huber *, O. Pons †, N. Heutte ‡

May 5, 2004

Abstract

In the analysis of a multi-state process with a finite number of states, a semi-Markov model allows to weaken the often used Markov assumption. The behavior of the process is defined through the initial probabilities on the set of possible states, the direct transition probabilities from any state to any other state and the sojourn times distributions as functions of the actual state and the state reached from there at the end of the sojourn. The most usual model in this framework is the so-called independent competing risk model. Then the transition probabilities can be deduced from the distribution of the sojourn times. For both cases, this submodel and the general one, we propose estimators of the transition probabilities and the distribution functions of the sojourn times when n i.i.d. sample paths of the process are observed under right-censoring. A comparison of the estimators allows us to test for an ICR model against the general semi-Markov model and a simulation study is performed.

1 Introduction

The motivation for this paper is the analysis of a cohort of patients where not only the survival time of the patients but also a finite number of life states are under study. The behavior of the process is assumed to be semi-Markov in order to weaken the very often used, and often too restrictive, Markov assumption. The behavior of such a process is defined through the initial probabilities on the set of possible states, and the transition functions defined as the probabilities, starting from any specified state, to reach another state within a certain amount of time. In order to define this behavior, the set of the transition functions may be replaced by two sets. The first one is the set of direct transition probabilities $p_{jj'}$ from any state j to any other state j'. The second one is the set of the sojourn times distributions $F_{|jj'}$ as functions of the actual state j and the state j' reached from there at the end of the sojourn (section 2).

*MAP 5, FRE CNRS 2428, UFR Biomédicale, Université René Descartes, et U 472 INSERM, France, e-mail: catherine.huber@univ-paris5.fr.

†MAP 5, FRE CNRS 2428, Université René Descartes, Paris 5, and INRA, Jouy-en-Josas, France

‡Université de Caen, IUT, 14 000 Lisieux, France

The most usual model in this framework is the so-called competing risk model. This model may be viewed as one where, starting in a specific state j, all states that may be reached directly from j are in competition: the state j' with the smallest random time $W_{jj'}$ to reach it from j will be the one. It is well known that the joint distribution and the marginal distribution of the latent sojourn times $W_{jj'}$ is not identifiable in a general competing risk model (Tsiatis (1975)). In a semi-Markov model as well as in a competing risk model, only the sub-distribution functions $F_{j'|j} = p_{jj'}F_{|jj'}$ are identifiable and it is always possible to define an independent competing risk (ICR) model by assuming that the variables $W_{jj'}$, $j' = 1, \ldots, m$, are independent with distributions $F_{|jj'} = F_{j'|j}/F_{j'|j}(\infty)$. Without an assumption about their dependence, their joint distribution is not identifiable and a test of an ICR model against an alternative of a general competing risk model is not possible. Similarly, there is always a representation of any general semi-Markov model as a competing risk model with possibly dependent $W_{jj'}$ but it is not uniquely defined. When the random variables $W_{jj'},\ j' \in J(j)$, are assumed to be independent, the semi-Markov model simplifies : the transition probabilities can be deduced from the laws of the sojourn times $W_{jj'}$ (section 3). As the term "competing risk" is also used in case of dependence of the $W_{jj'}$, we shall sometimes emphasize the independence we always assume in a competing risk model, by calling it the Independent Competing Risk (ICR) model.

For a general right-censored semi-Markov process, Lagakos, Sommer and Zelen (1978) proposed a maximum likelihood estimator for the direct transition probabilities and the distribution functions of the sojourn times, under the assumption of a discrete function with a finite number of jumps. In non-parametric models for censored counting processes, Gill (1980), Voelkel and Crowley (1986) considered estimators of the sub-distribution functions $F_{j'|j} = p_{jj'}F_{|jj'}$ and they studied their asymptotic behavior. Here, we consider maximum likelihood estimation for the general semi-parametric model defined by the probabilities $p_{jj'}$ and the hazard functions related to the distribution functions $F_{|jj'}$ (section 4). If the mean number of transitions by an individual tends to infinity, then, the maximum likelihood estimators are asymptotically equivalent to those of the uncensored case. In section 5, we present new estimators defined for the case of a right-censored process with a bounded number of transitions. The difficulty comes from the fact that we do not observe the next state after a right-censored duration in a state.

Under the ICR assumption, specific estimators of the distribution functions $F_{|jj'}$ and of the direct transition probabilities $p_{jj'}$ are deduced from Gill's estimator of the transition functions $F_{j'|j}$. A comparison of those estimators to the estimators for a general semi-Markov process leads to tests for an ICR model against the semi-Markov alternative in section 6.

2 Framework

For each individual i, $i = 1, \cdots, n$, we observe, during a period of time t_i, his successive states $J(i) = (J_0(i), J_1(i), \cdots, J_{K(i)}(i))$, where $J_0(i)$ is the initial state, $J_{K(i)}(i)$ the final state after $K(i)$ transitions. The total number of possible states is assumed to be finite and equal to m. The successive observed sojourn times are denoted $X(i) = (X_1(i), X_2(i), \cdots, X_{K(i)}(i))$, where $X_k(i)$ is the sojourn time i spent in state $J_{k-1}(i)$ after $(k-1)$ transitions, and the cumulative sojourn times are $T_k = \Sigma_{\ell=1}^{k} X_\ell$.
One must notice that, if i changes state $K(i)$ times, the sojourn time i spent in his last state $J_{K(i)}$ is generally right censored by $t_i - T_{K(i)}(i)$, where t_i is the total period of observation for subject i. We simplify the rather heavy notation for this last duration, and the last state $J_{K(i)}(i)$ as

$$X^*(i) \equiv t_i - T_{K(i)}(i), \quad J^*(i) \equiv J_{K(i)}(i).$$

The subjects are assumed independent and the probability distribution of the sojourn times absolutely continuous. The two models we propose for the process describing the states of the patient are renewal semi-Markov processes. Their behavior is defined through the following quantities:

1. The initial law $\rho = (\rho_1, \rho_2, \cdots, \rho_m)$:

$$\begin{aligned} \rho_j &= P(J_0 = j), \quad j \in \{1, 2, \cdots, m\}, \\ \sum_{j \in \{1,2,\cdots,m\}} \rho_j &= 1. \end{aligned} \tag{1}$$

2. The transition functions $F_{j'|j}(t)$:

$$F_{j'|j}(t) = P(J_k = j', X_k \leq t | J_{k-1} = j) \ , \ \ j, j' \in \{1, 2, \cdots, m\}. \tag{2}$$

Equivalent to the set of the transition functions $F_{j'|j}$, is the set of the transition probabilities, $p = \{p_{jj'} \ , \ j, j' \in \{1, 2, \cdots, m\}$, together with the set of the distribution functions $F_{|jj'}$ of the sojourn times in each state conditional on the final state as defined below:

1. The direct transition probabilities from a state j to another state j' :

$$p_{jj'} = P(J_k = j' | J_{k-1} = j), \tag{3}$$

2. The law of the sojourn time between two states j and j' defined by its distribution function:

$$F_{|jj'}(t) = P(X_k \leq t | J_{k-1} = j, J_k = j'), \tag{4}$$

$$\text{where} \ \sum_{j'=1}^{m} p_{jj'} = 1 \ , \ \ p_{jj'} \geq 0 \ , \ \ j, j' \in \{1, 2, \cdots, m\}. \tag{5}$$

We notice that the distribution functions $F_{|jj'}$ conditional on states (j, j') do not depend on the value of k, the rank of the state reached by the patient along the process, which is a characteristic of a renewal process. We can define the hazard rate conditional on the present state and the next one:

$$\lambda_{|jj'}(t) = \lim_{dt \longrightarrow 0} \frac{P(t \leq X_k \leq t + dt | X_k \geq t, J_{k-1} = j, J_k = j')}{dt}, \tag{6}$$

as well as the cumulative conditional hazard:

$$\Lambda_{|jj'}(t) = \int_0^t \lambda_{|jj'}(u) du. \tag{7}$$

Let W_j be a sojourn time in state j when no censoring is involved, F_j its distribution function, and $\overline{F}_j \equiv 1 - F_j$ its survival function, such that

$$\overline{F}_j(x) \equiv P(W_j > x) = \sum_{j'=1}^{m} p_{jj'} \overline{F}_{|jj'}(x). \tag{8}$$

The potential sojourn time in state j may be right censored by a random variable C_j having distribution function G_j, density g_j and survival function $\overline{G}_j$. The observed sojourn time in state j is $W_j \wedge C_j$.
A general notation will be $\overline{F}$ for the survival function corresponding to a distribution function F, so that, for example, $\overline{F}_{|jj'} = 1 - F_{|jj'}$ and similarly, for the transition functions, $\overline{F}_{j'|_j} = p_{jj'} - F_{j'|j}$.

3 Independent Competing Risks Model

We assume now that, starting from a state j, the potential sojourn times $W_{jj'}$ until reaching each of the states j' directly reachable from j are independent random variables having distribution functions defined through (4). The final state is the one for which the duration is the smallest. One can thus say that all other durations are right censored by this one. Without restriction of the generality, we assume that the subject is experiencing his k^{th} transition. The competing risks model is defined by

$$\begin{aligned} X_k &= \min_{j'=1,\ldots,m} W_{jj'}, \\ J_k &= j' \text{ such that } W_{jj'} < W_{jj"}, j" \neq j', \end{aligned} \tag{9}$$

where $W_{jj'}$ has the distribution function $F_{|jj'}$.

In this simple case, independence, both of the subjects and of the potential sojourn times in a given state, allows us to write down the likelihood as a product of factors dealing separately with the time elapsed between two specific states (j, j'). For the Independent Competing Risk model, one derives from (6), (8) and(9) that

$$\begin{aligned} F_{j'|j}(t) &= P(J_k = j', X_k \leq t | J_{k-1} = j) = \int_0^t \{ \prod_{j" \neq j'} \overline{F}_{|jj"}(u) \} \, dF_{|jj'}(u) \qquad (10) \\ &= \int_0^t \lambda_{|jj'}(u) e^{-\sum_{j"} \Lambda_{|jj"}(u)} du. \end{aligned}$$

A consequence is that the direct transition probabilities $p_{jj'}$ defined in (3) may be derived from the probabilities defined in (4),

$$p_{jj'} = P(J_{k+1} = j'|J_k = j) = \int_0^\infty \lambda_{|jj'}(u) e^{-\sum_{j"} \Lambda_{|jj"}(u)} du. \tag{11}$$

In this special case, the likelihood is fully determined by the initial ρ_j and the functions $\lambda_{|jj'}$ defined in (6). The likelihood $L_{rc,n}$ for the independent competing risks is proportional to

$$\begin{aligned} L_{rc,n} &= \prod_{i=1}^{n} \rho_{J_0(i)} \prod_{k=1}^{K(i)} \lambda_{|J_{k-1}(i),J_k(i)}(X_k(i)) \\ &\quad \times e^{-\sum_{j"} \Lambda_{J_{k-1}(i),j"}(X_k(i))} e^{-\sum_{j"} \Lambda_{|J^*(i),j"}(X^*(i))} . \end{aligned} \tag{12}$$

It can be decomposed into the product of terms each of which is relative to an initial state j and a final state j'. When gathering the terms in $L_{rc,n}$ that are relative to a same hazard rate $\lambda_{|jj'}$ or else $\Lambda_{|jj'}$, one observes that the hazard rates appear separately in the likelihood for each pair (j, j')

$$\begin{aligned} L_{rc,n} &= \{\prod_{i=1}^{n} \rho_{J_0(i)}\} \prod_{j} \prod_{j'=1}^{m} L_{rc,n}(j,j'), \\ L_{rc,n}(j,j') &= \prod_{i=1}^{n} \prod_{k=1}^{K(i)} [\lambda_{|jj'}(X_k(i)) e^{-\Lambda_{|jj'}(X_k(i))}]^{1\{J_{k-1}(i)=j,\ J_k(i)=j'\}} \\ &\quad \times \left[e^{-\Lambda_{|jj'}(X_k(i))}\right]^{1\{J_{k-1}(i)=j,\ J_{k(i)}\neq j'\}} \left[e^{-\Lambda_{|jj'}(X^*(i))}\right]^{1\{J^*(i)=j\}}. \end{aligned} \tag{13}$$

This problem may be treated as m parallel and independent problems of right censored survival analysis. The only link between them is the derivation of the direct transition probabilities using (11).

4 General Model

The patients are assumed to be independent, while the potential times for a given subject are no longer assumed to be independent. We model separately the hazard rate and the transition functions ρ_j, $p_{jj'}$ and $\lambda_{|jj'}$ defined as in (1), (3) and (6). The direct transition probabilities $p_{jj'}$ can no longer be derived from the hazard rates.They are now free, except for the constraints (5). The distributions of the time elapsed between two successive states j and j' and those of the censoring are assumed to be absolutely continuous. The likelihood L_n is proportional to

$$\begin{aligned} L_n &= \prod_{i=1}^{n} \rho_{J_0(i)} \prod_{k=1}^{K(i)} \overline{G}_{J_{k-1}(i)}(X_k(i)) p_{J_{k-1}(i),J_k(i)} \lambda_{|J_{k-1}(i),J_k(i)}(X_k(i)) e^{-\Lambda_{|J_{k-1}(i),J_k(i)}(X_k(i))} \\ &\quad \times g_{J^*(i)}(X^*(i)) \left\{ \sum_{j'=1}^{m} p_{J^*(i),j'} e^{-\Lambda_{|J^*(i),j'}(X^*(i))} \right\} \\ &= \prod_{i=1}^{n} \prod_{j=1}^{m} \quad \rho_j^{1\{J_0(i)=j\}} \prod_{k=1}^{K(i)} \prod_{j'=1}^{m} [p_{jj'} \lambda_{|jj'}(X_k(i)) e^{-\Lambda_{|jj'}(X_k(i))}\ \overline{G}_j(X_k(i))]^{1\{J_{k-1}(i)=j, J_k(i)=j'\}} \\ &\quad \times \left\{g_j(X^*(i))\ \sum_{j'=1}^{m} p_{jj'} e^{-\Lambda_{|jj'}(X^*(i))}\right\}^{1\{J^*(i)=j\}}. \end{aligned}$$

This likelihood may be written as a product of terms each of which implies sojourn times exclusively in one specific state j, $L_n = \prod_{j=1}^{m} L_n(j)$.

For each subject i, and for each $k \in \{1, 2, \cdots, K(i)\}$, we denote $1 - \delta_k(i)$ the censoring indicator of its sojourn time in the k^{th} visited state, $J_{k-1}(i)$, with the convention that $\delta_0(i) \equiv 1$ for every i. If j' is an absorbing state, and if $J_k(i) = j'$, then j' is he last state observed for subject i, $k \equiv K(i)$, and we denote it $X^*(i) = 0$ and $\delta_{K(i)+1}(i) = 1$.
Another convention is that subject i is censored, when the last visited state $J^*(i)$ is not absorbing and the sojourn time in this state $X^*(i)$ is strictly positive and we denote $1 - \delta_i$ the censoring indicator. In all other cases, in particular if the last visited state is absorbing or if the sojourn time there is equal to 0, we say that the subject is not censored and we thus have $\delta_i = 1$. We can then write

$$\delta_k(i) = \prod_{k'=1}^{k} \delta_{k'}(i), \quad \delta_i = 1\{X^*(i) = 0\}.$$

For each state j of $\{1, 2, \cdots, m\}$, we define the following counts where k varies, for each subject i, between 1 and $K(i)$, $i \in \{1, 2, \cdots, n\}$, and $x \geq 0$,

$$\begin{aligned} N_{i,k}(x, j, j') &= 1\{J_{k-1}(i) = j, J_k(i) = j'\}1\{X_k(i) \leq x\}, \\ Y_{i,k}(x, j, j') &= 1\{J_{k-1}(i) = j, J_k(i) = j'\}1\{X_k(i) \geq x\}, \\ N_i^c(x, j) &= (1 - \delta_i)1\{J^*(i) = j\}1\{X^*(i) \leq x\}, \\ Y_i^c(x, j) &= (1 - \delta_i)1\{J^*(i) = j\}1\{X^*(i) \geq x\}. \end{aligned} \tag{14}$$

By summation of the counts thus defined on the indices j', i, or k, we get

$$\begin{aligned} N(x, j, j', n) &= \sum_{i=1}^{n} \sum_{k=1}^{K(i)} N_{i,k}(x, j, j'), \\ N^{nc}(x, j) &= \sum_{j'=1}^{m} N(x, j, j', n), \\ N(x, j, n) &= \sum_{i=1}^{n} N_i^{nc}(x, j) + N^{nc}(x, j), \\ Y^{nc}(x, j, j', n) &= \sum_{i=1}^{n} \sum_{k=1}^{K(i)} Y_{i,k}(x, j, j'), \\ Y(x, j, n) &= \sum_{j'=1}^{m} Y^{nc}(x, j, j', n) + \sum_{i=1}^{n} Y_i^c(x, j). \end{aligned} \tag{15}$$

By taking for x the limiting value ∞ we define $N_{i,k}(j, j') = N_{i,k}(\infty, j, j')$, $N_i^c(j) = N_i^c(\infty, j)$, $N(j, j', n) = N(\infty, j, j', n)$, $N^{nc}(j, n) = N^{nc}(\infty, j, n)$, so that $N(j, j', n)$ is the number of direct transitions from j to

j' that are fully observed,$N(j,n)$ is the number of sojourn times in state j, whose $N^{nc}(j,n)$ (nc for not censored) are fully observed and $N^c(j,n)$ (c for censored) are censored. For $x=0$, we denote $Y_i^c(j)=Y_i^c(0,j)$. The number of individuals initially in state j is $N^0(j,n)=\sum_{i=1}^n 1\{J_0(i)=j\}$.

The true parameter values are denoted ρ_j^0 and $p_{jj'}^0$, and the true functions of the model are $\overline{F}^0_{j'|j}$, $\overline{F}^0_{|jj'}$, $\overline{F}^0_j$, $\overline{G}^0_j$ and $\Lambda^0_{|jj'}$.

Let $l_n=\log(L_n)$ and $l_n(j)=\log(L_n(j))$. The log-likelihood relative to state j is proportional to

$$\begin{aligned} l_n(j) &= \rho_j N^0(j,n)+\sum_{j'=1}^m N(j,j',n)\log(p_{jj'}) \\ &+\sum_{i=1}^n\sum_{k=1}^{K(i)}\sum_{j'=1}^m N_{i,k}(j,j')[\log(\lambda_{|jj'}(X_k(i)))-\Lambda_{|jj'}(X_k(i))] \\ &+\sum_{i=1}^n N_i^c(j)[\log\{\sum_{j'=1}^m p_{jj'}e^{-\Lambda_{|jj'}(X^*(i))}\}] \qquad (16) \\ &= l_n^0(j)+l_n^{nc}(j)+l_n^c(j), \end{aligned}$$

Among the sum of four terms giving (16), let l_n^0 be the first term relative to the initial state, l_n^{nc} (nc for non censored) the sum of the second and third terms, which involve exclusively fully observed sojourn times in state j, and finally l_n^c (c for censored) the last term which deals with censored sojourn times in state j.

We denote $K_n=\max_{i=1,2\cdots,n}K(i)$ and $n\overline{K}_n=\sum_{i=1}^n K(i)$ respectively the maximum number of transitions and the total number of transitions for the n subjects. We consider two different designs of experiments, whether or not observations are stopped after a fixed amount K of direct transitions.

It is obvious that if the densities f_j of the sojourn times, without censoring, for every state j, are strictly positive on $]0;t_0[$ for some $t_0>0$, and if the distribution functions G_j of the censoring times are such that $G_j(t)<1$ for all $t>0$, the maximal number $K_n=\max_i K(i)$ of transitions experienced by a subject tends to infinity when n grows. If moreover the mean number of transitions $\overline{K}_n$ goes also to infinity, then the term relative to censored times $l_n^c(j)$ is the sum of terms of order n while the term $l_n^{nc}(j)$ is a sum of terms of order $n\overline{K}_n$. Therefore we have

Proposition 1 *Under the assumptions $\overline{K}_n\to\infty$, and*

$$\frac{N^{nc}(j,n)}{n\overline{K}_n} \longrightarrow q_j^0>0, \quad j\in\{1,2,\cdots,m\},$$

then

$$\lim_{n\longrightarrow\infty}\frac{l_n(j)}{n\overline{K}_n} = \lim_{n\longrightarrow\infty}\frac{l_n^{nc}(j)}{n\overline{K}_n}.$$

and the maximum likelihood estimators are asymptotically equivalent to

$$\begin{aligned}
\widehat{p}_{jj'} &= \frac{N(j,j',n)}{N^{nc}(j,n)}, \\
\widehat{\Lambda}_{|jj'}(x) &= \int_0^x \frac{dN(x,j,j',n)}{Y^{nc}(s,j,j',n)}, \\
\widehat{\overline{F}}_{|jj'}(x) &= \prod_{i=1}^{n}\prod_{k=1}^{K(i)}\Big\{1-\frac{N_{i,k}(x,j,j')}{Y^{nc}(X_k(i),j,j',n)}\Big\}.
\end{aligned}$$

5 Case of a bounded number of transitions

We now assume that the number of transitions is bounded by a finite number K fixed in advance. For each subject $i=1,\cdots,n$, the observation ends at time $t_i=\sum_{k=1}^{K(i)} X_k(i)$ if $K(i)=K$ or if $J_{K(i)}$ is an absorbing state, and at time t_i where there is a right censoring in the $K(i)^{\text{th}}$ visited state, $K(i)<K$.

Using notations in (15), the likelihood term relative to the initial state j may be written

$$l_n^0(j) \quad = \quad N^0(j,n)\log(\rho_j),$$

the terms relative to the fully observed sojourn times in state j is

$$\begin{aligned}
l_n^{nc}(j) &= \sum_{j'=1}^{m}\Big\{ \; N(j,j',n)\log(p_{jj'}) \\
&\quad +\sum_{i=1}^{n}\sum_{k=1}^{K} N_{i,k}(j,j')[\log(\lambda_{|jj'}(X_k(i)))-\Lambda_{|jj'}(X_k(i))] \;\Big\},
\end{aligned}$$

and the term relative to the censored sojourn times in state j is

$$l_n^c(j) \quad = \quad \sum_{i=1}^{n} N_i^c(j)[\log\{\sum_{j'=1}^{m} p_{jj'}e^{-\Lambda_{|jj'}(X^*(i))}\}].$$

The score equations for $p_{jj'}$ and $\Lambda_{jj'}$ do not lead to explicit solutions because they involve the survival function $\overline{F}_j$ and the transition function $\overline{F}_{j'|j}$. We define estimators $\widehat{p}_{n,jj'}$ and $\widehat{\Lambda}_{n,|jj'}$ by plugging in the score equations the Kaplan-Meier estimator of $\overline{F}_j$ and the estimator of $F_{j'|j}$ given by Gill (1980),

$$\widehat{\overline{F}}_{n,j}(x) \quad = \quad \prod_{i=1}^{n}\prod_{k=1}^{K(i)}\Big\{1-\frac{1}{Y(X_k(i),j,n)}\Big\}^{N_{i,k}^{nc}(x,j)} = \prod_{y\le x}\Big\{1-\frac{dN(y,j,n)}{Y(y,j,n)}\Big\}, \tag{17}$$

$$\widehat{F}_{n,j'|j}(x) \quad = \quad \sum_{i=1}^{n}\sum_{k=1}^{K(i)} \widehat{\overline{F}}_{n,j}(X_k^-(i))\frac{N_{i,k}(x,j,j')}{Y(X_k(i),j,n)} = \int_0^x \widehat{\overline{F}}_{n,j}(y^-)\frac{dN(y,j,j',n)}{Y(y,j,n)}. \tag{18}$$

We obtain the estimators

$$\begin{aligned}
\widehat{\rho}_{n,j} &= \frac{N^0(j,n)}{n}, \\
\widehat{p}_{n,jj'} &= \frac{N(j,j',n)+\widehat{N}^c(j,j',n)}{N^{nc}(j,n)+N^c(j,n)}, \qquad (19)\\
\widehat{\Lambda}_{n,|jj'}(x) &= \int_0^x \frac{dN(y,j,j',n)}{Y^{nc}(y,j,j',n)+\widehat{Y}^c(y,j,j',n)},
\end{aligned}$$

with

$$\widehat{Y}^c(y,j,j',n) = \sum_{i=1}^n Y_i^c(y,j)\frac{\widehat{\overline{F}}_{n,j'|j}(X^*(i))}{\widehat{\overline{F}}_{n,j}(X^*(i))},$$
$$\widehat{N}^c(j,j',n) = \sum_{i=1}^n N_i^c(j)\frac{\widehat{\overline{F}}_{n,j'|j}(X^*(i))}{\widehat{\overline{F}}_{n,j}(X^*(i))}.$$

The variable $(n^{1/2}(\widehat{p}_{n,jj'}-p^0_{jj'}))_{j'}$ and the process $\{n^{1/2}(\widehat{\Lambda}_{n,|jj'}-\Lambda^0_{|jj'}))_{j'}$ are asymptotically Gaussian, on every interval $[0,\tau]$ such that $\int_0^\tau(\overline{F}^0_{j'|j}\overline{G}^0_j)^{-1}\,d\Lambda^0_{j'|j}<\infty$ (Pons (2002)).

6 A Test of the Hypothesis of Independent Competing Risks.

In the ICR case, the initial probabilities jointly with the survival functions $\overline{F}_{|jj'}$ of the sojourn times conditional on states on both ends, are sufficient to determine completely the law of the process. In the general case, however, the two sets of parameters $p_{jj'}$ and $\overline{F}_{|jj'}$ are independent and may be modeled separately. Our aim is to derive a test of the hypothesis of Independent Competing Risks (ICR):

$$H_0 \;:\; \text{The process is ICR}$$
$$H_1 \;:\; \text{The process is not ICR}$$

The Kaplan-Meier estimator $\widehat{\bar{F}}_{n,j}$ of $\bar{F}_j$, given in (17), and the estimator $\widehat{F}_{n,j'|j}$ of $F_{j'|j}$, given in (18), are consistent and asymptotically Gaussian both under H_0 and under H_1. It is also true for the straightforward estimator $\widehat{\rho}_{n,j}$ of the initial probabilities. From those estimators, one may derive general estimators of the transition probability $p_{jj'}$ and of the survival function $\overline{F}_{|jj'}$ of the time elapsed between two successive jumps in states j and j'. For these estimators, we shall use the same notations as the estimators of $p_{jj'}$ and $\overline{F}_{|jj'}$ defined in section 5, though they are now given by

$$\widehat{p}_{n,jj'} = \max_t \widehat{F}_{n,j'|j}(t) \tag{20}$$

$$\widehat{\overline{F}}_{n,|jj'}(t) = 1-\frac{\widehat{F}_{n,j'|j}(t)}{\widehat{p}_{n,jj'}}. \tag{21}$$

In the independent competing risk model, the transition probability $F_{j'|j}$ satisfies (10) and thus may be estimated as

$$\begin{aligned}\widehat{F}^{RC}_{n,j'|j}(t) &= -\int_0^t \prod_{j"\neq j'} \widehat{\overline{F}}_{n,|jj"}(s)\ d\widehat{\overline{F}}_{n,|jj'}(s) \\ &= \frac{1}{\prod_{j"}\widehat{p}_{n,jj"}}\int_0^t \prod_{j"\neq j'} \widehat{\overline{F}}_{n,j"|j}(s)\ \widehat{\overline{F}}_{n,j}(s^-)\ d\widehat{\Lambda}_{n,j'|j}(s),\end{aligned} \tag{22}$$

where

$$\widehat{\Lambda}_{n,j'|j}(t) = \int_0^t 1\{Y(s,j,n)>0\}\frac{dN(s,j,j',n)}{Y(s,j,n)} \tag{23}$$

is the estimator of the cumulative hazard function $\Lambda_{n,j'|j}$ in the general model. A competitor to $\widehat{p}_{n,jj'}$ is deduced as

$$\widehat{p}^{RC}_{n,jj'} = \max_t \widehat{F}^{RC}_{n,j'|j}(t). \tag{24}$$

Proposition 2 *If $p^0_{jj'} > 0$, $\sqrt{n}(\widehat{p}_{n,jj'} - p^0_{jj'})$ is asymptotically distributed as a normal random vector with mean 0, variances and covariances*

$$\begin{aligned}
\sigma^2_{jj'} &= \frac{1}{\pi^0_j}\int_0^\infty \frac{1}{\overline{G}^0_j(s)\overline{F}^0_j(s)}\left\{(\overline{F}^0_{j'|j}(s) - p^0_{j'|j})^2\frac{dF^0_j(s)}{\overline{F}^0_j(s)} + \{\overline{F}^0_j(s) + 2(\overline{F}^0_{j'|j}(s) - p^0_{j'|j})\}\, d\overline{F}^0_{j'|j}(s)\right\},\\
\sigma^2_{jj'j"} &= \frac{1}{\pi^0_j}\int_0^\infty \frac{1}{\overline{G}^0_j(s)\overline{F}^0_j(s)}\left\{(\overline{F}^0_{j'|j}(s) - p^0_{j'|j})(\overline{F}^0_{j"|j}(s) - p^0_{j"|j})\frac{dF^0_j(s)}{\overline{F}^0_j(s)}\right.\\
&\quad \left. +(\overline{F}^0_{j'|j}(s) - p^0_{j'|j})\}\, d\overline{F}^0_{j"|j}(s) + (\overline{F}^0_{j"|j}(s) - p^0_{j"|j})\}\, d\overline{F}^0_{j'|j}(s)\right\}.
\end{aligned}$$

Moreover, $\sqrt{n}(\widehat{p}^{RC}_{n,jj'} - p^0_{jj'})$ is asymptotically distributed as a centered Gaussian variable.

Estimators of the asymptotic variance and covariances of $(\widehat{p}_{n,jj'})_{j'\in J(j)}$ may be obtained by replacing the functions $\overline{F}^0_j$, $F^0_{j'|j}$ and $\Lambda^0_{j'|j}$ by their estimators in the general model, (17, (18) and (23). Due to their intricate formulas, it seems difficult to use an empirical estimator of the asymptotic variance of $\widehat{p}^{RC}_{n,jj'}$ and a bootstrap estimator should be preferred. Asymptotic confidence intervals for $p^0_{jj'}$ at the level α are deduced from the $(1-\alpha/2)$-quantile c_α of their boostrap distributions, $I_{n,jj'}(\alpha)$ in the general case and $I^{RC}_{n,jj'}(\alpha)$ under the null hypothesis of Independent Competing Risks.

A test of the Independent Competing Risks hypothesis may be defined by rejecting H_0 if $I_{n,jj'}(\alpha)$ and $I^{RC}_{n,jj'}(\alpha)$ are not overlapping for some j'. As the estimators of the parameters $p^0_{jj'}$ are not independent, the level α^* of this test with critical region

$$R_{nj}(\alpha) = \cap_{j'=1}^m R_{njj'}(\alpha), \text{ where } R_{njj'}(\alpha) = \{I_{n,jj'}(\alpha) \cap I^{RC}_{n,jj'}(\alpha) \neq \emptyset\},$$

satisfies $\alpha^* \geq 1 - (1-\alpha)^m$.

References

[1] P.K. Andersen and N. Keiding, "Event history analysis in Continuous Time," *Technical report, Department of Biostatistics, University of Copenhagen*,2000.

[2] V. Bagdonavicius and M.N. Nikulin, "Accelerated Life Models, Modeling and Statistical Analysis," Chapman & Hall/CRC ed., 334 p., 2002.

[3] R. Gill, " Nonparametric estimation based on censored observations of a markov renewal process," *Z. Wahrsch. verw. Gebiete* vol. 53 pp. 97-116,1980.

[4] R. Gill, "Large sample behaviour of the product-limit estimator on the whole line," *Ann. Statist.* vol. 11 pp. 49-58, 1983.

[5] N. Heutte and C. Huber-Carol, "Semi-Markov Models for Quality of Life Data with Censoring," in *Statistical Methods for Quality of Life Studies*, Kluwer Academic Publishers, pp. 207-218, 2002.

[6] N. Limnios and G. Oprisan, "Semi-Markov Processes and Reliability," Birkhaser ed.,222 p., 2001.

[7] O. Pons, "Estimation of semi-Markov models with right-censored data," In *Handbook of Statistics 23, "Survival Analysis"*, Elsevier *(to appear)*, 2003.

[8] S.W. Lagakos, C. Sommer and M. Zelen, "Semi-markov models for partially censored data.," *Biometrika* vol. 65 pp. 311-317, 1978.

[9] R. Pyke, "Markov renewal processes: definitions and preliminary properties," *Ann. Math. Statist.* vol. 32 pp. 1231-1342, 1961.

[10] A. Tsiatis, "A nonidentifiability aspect of the problem of competing risks," *Proc. Nat. Acad. Sci. USA* vol. 37 pp. 20-22, 1975.

[11] J. Voelkel and , J. Crowley, "Nonparametric inference for a class of semi-markov processes with censored observations," *Ann. Statist.* vol. 12 pp. 142-160, 1984.

7 Appendix

Proof of Proposition 2.

Let $\tau_{n,j} = \arg\max_t \widehat{F}_{n,j}(t)$. The asymptotic behavior of $\widehat{p}_{n,jj'}$ is derived from theorem 3 in Gill (1980) which states the weak convergence of the process

$$(\sqrt{n}(\widehat{F}_{n,j'|j}(t \wedge \tau_{n,j}) - F^0_{j'|j}(t \wedge \tau_{n,j}))_{j' \in J(j)}, \sqrt{n}(\widehat{\overline{F}}_{n,j}(t \wedge \tau_{n,j}) - \overline{F}^0_j(t \wedge \tau_{n,j}))_{t \geq 0}$$

to a Gaussian process defined, for continuous transition functions $F^0_{j'|j}$, as

$$\left\{\int_0^t \frac{\overline{F}^0_{j'|j}(s)\,dV_{jj'}(s)}{EY_i(s,j)} - \overline{F}^0_{j'|j}(t)\int_0^t \frac{dV_j(s)}{EY_i(s,j)} + \int_0^t \frac{\overline{F}^0_{j'|j}(s)\,dV_j(s)}{EY_i(s,j)}, \overline{F}^0_j(t)\int_0^t \frac{dV_j(s)}{EY_i(s,j)}\right\}$$

where $V_{jj'}, j, j' \in \{1, 2, \cdots, m\}$ is a multivariate Gaussian process with independent increments, having mean 0 and covariances

$$var(V_{jj'}(t)) = \int_0^t EY_i(s,j)\frac{d\overline{F}^0_{j'|j}(s)}{\overline{F}^0_j(s)},$$

$cov(V_{jj'}(t), V_{jj"}(t)) = 0$ if $j' \neq j"$ and $cov(V_{jj'}(t), V_{j_1 j_1'}(t_1)) = 0$ if $j_1 \neq j$ or $t_1 \neq t$, and $V_j = \sum_{j'} V_{jj'}$.

As $EY_i(s,j) = \pi^0_j \overline{G}^0_j(s)\overline{F}^0_j(s)$, it follows that $\sqrt{n}(\widehat{p}_{n,jj'} - p^0_{jj'})$ is asymptotically distributed as

$$\int_0^\infty \frac{dV_{jj'}(s)}{\pi^0_j \overline{G}^0_j(s)} \quad - \quad p^0_{jj'}\int_0^\infty \frac{dV_j(s)}{\pi^0_j \overline{G}^0_j(s)\overline{F}^0_j(s)} \quad + \quad \int_0^\infty \overline{F}^0_{j'|j}(s)\frac{dV_j(s)}{\pi^0_j \overline{G}^0_j(s)\overline{F}^0_j(s)}.$$

Denoting this limit as $A - B + C$, we have

$$\begin{aligned}
var(A) &= \frac{1}{\pi^0_j}\int_0^\infty \frac{1}{\overline{G}^0_j(s)}\,d\overline{F}^0_{j'|j}(s) \\
var(B) &= \frac{p^2_{jj'}}{\pi^0_j}\int_0^\infty \frac{1}{\overline{G}^0_j(s)\overline{F}^0_j(s)^2}\,dF^0_j(s) \\
var(C) &= \frac{1}{\pi^0_j}\int_0^\infty \frac{\overline{F}^0_{j'|j}(s)^2}{\overline{G}^0_j(s)\overline{F}^0_j(s)^2}\,dF^0_j(s) \\
cov(A,B) &= \frac{p^0_{jj'}}{\pi^0_j}\int_0^\infty \frac{1}{\overline{G}^0_j(s)\overline{F}^0_j(s)}\,d\overline{F}^0_{j'|j}(s) \\
cov(A,C) &= \frac{1}{\pi^0_j}\int_0^\infty \frac{\overline{F}^0_{j'|j}(s)}{\overline{G}^0_j(s)\overline{F}^0_j(s)}\,d\overline{F}^0_{j'|j}(s) \\
cov(B,C) &= \frac{p^0_{jj'}}{\pi^0_j}\int_0^\infty \frac{\overline{F}^0_{j'|j}(s)}{\overline{G}^0_j(s)\overline{F}^0_j(s)^2}dF^0_j(s),
\end{aligned}$$

and $\sigma^2_{jj'}$ is the variance of $A - B + C$. The covariance $\sigma^2_{jj'j"}$ is obtained by similar calculations, but the covariance between the corresponding terms $A(jj')$ and $A(jj")$ is zero.

From (22), the asymptotic Gaussian distribution of $\sqrt{n}(\widehat{p}^{RC}_{n,jj'} - p^0_{jj'})$ is a consequence of the asymptotic behavior of the estimators $\widehat{\overline{F}}_{n,j}$ and $\widehat{\overline{F}}_{n,j'|j}$ and of the estimator $\widehat{\Lambda}_{n,j'|j}$ given by (23), using again theorem 3 in Gill (1980). □

Limiting covariance of $\sqrt{n}(\widehat{p}^{RC}_{n,jj'} - p^0_{jj'})$.

The limiting covariance of $\sqrt{n}(\widehat{p}^{RC}_{n,jj'} - p^0_{jj'})$ may be calculated using the following expressions

$$\begin{aligned}
\sigma^2_{jj'}(t) &= \frac{1}{\pi^0_j}\int_0^t \frac{1}{\overline{G}^0_j(s)\overline{F}^0_j(s)}\left\{(\overline{F}^0_{j'|j}(s) - \overline{F}^0_{j'|j}(t))^2\frac{dF^0_j(s)}{\overline{F}^0_j(s)} + \{\overline{F}^0_j(s) + 2(\overline{F}^0_{j'|j}(s) - \overline{F}^0_{j'|j}(t))\}\, d\overline{F}^0_{j'|j}(s)\right\}, \\
\sigma^2_{jj'j"}(t) &= \frac{1}{\pi^0_j}\int_0^t \frac{1}{\overline{G}^0_j(s)\overline{F}^0_j(s)}\left\{(\overline{F}^0_{j'|j}(s) - \overline{F}^0_{j'|j}(t))(\overline{F}^0_{j"|j}(s) - \overline{F}^0_{j"|j}(t))\frac{dF^0_j(s)}{\overline{F}^0_j(s)}\right. \\
&\quad \left. +(\overline{F}^0_{j'|j}(s) - \overline{F}^0_{j'|j}(t))\}\, d\overline{F}^0_{j"|j}(s) + (\overline{F}^0_{j"|j}(s) - \overline{F}^0_{j"|j}(t))\}\, d\overline{F}^0_{j'|j}(s)\right\},
\end{aligned}$$

$$\begin{aligned}
c^{(1)}_{jj'}(t) &= \lim_n \ \mathrm{Cov}\{\sqrt{n}(\widehat{\overline{F}}_{n,j}(t) - \overline{F}^0_j(t)), \sqrt{n}(\widehat{\overline{F}}_{n,j'|j}(t) - \overline{F}^0_{j'|j}(t))\} \\
&= \overline{F}^0_j(t)\{\int_0^t \frac{\overline{F}^0_{j'|j}}{\overline{G}^0_j(\overline{F}^0_j)^2}\,(dF^0_{j'|j} + dF^0_j) - \overline{F}^0_{j'|j}(t)\int_0^t \frac{dF^0_j}{\overline{G}^0_j(\overline{F}^0_j)^2},
\end{aligned}$$

$$\begin{aligned}
v^{(1)}_{jj'}(t) &\equiv \lim_n \ \mathrm{Var}\sqrt{n}\{\widehat{\overline{F}}_{n,j}(t^-)\prod_{j_1\neq j'}\widehat{\overline{F}}_{n,j_1|j}(t) - \overline{F}^0_j(t^-)\prod_{j_1\neq j'}\overline{F}^0_{j_1|j}(t)\} \\
&= \{\prod_{j_1\neq j'}\widehat{\overline{F}}_{n,j_1|j}(t)\}^2\lim_n[\sum_{j_2\neq j'}\mathrm{Var}\sqrt{n}\{\widehat{\overline{F}}_{n,j_2|j}(t) - \overline{F}^0_{j_2|j}(t)\}\{\frac{\overline{F}^0_j(t)}{\overline{F}^0_{j_2|j}(t)}\}^2 + \mathrm{Var}\sqrt{n}\{\widehat{\overline{F}}_{n,j}(t^-) - \overline{F}^0_j(t)\} \\
&+ \sum_{j_2\neq j'}\sum_{j_3\neq j',j_2}\frac{(\overline{F}^0_j(t))^2}{\overline{F}^0_{j_2|j}(t)\overline{F}^0_{j_3|j}(t)}\mathrm{Cov}\{\sqrt{n}(\widehat{\overline{F}}_{n,j_2|j}(t) - \overline{F}^0_{j_2|j}(t)), \sqrt{n}(\widehat{\overline{F}}_{n,j_3|j}(t) - \overline{F}^0_{j_3|j}(t))\} \\
&+ \sum_{j_2\neq j'}\frac{\overline{F}^0_j(t)}{\overline{F}^0_{j_2|j}(t)}\mathrm{Cov}\{\sqrt{n}(\widehat{\overline{F}}_{n,j}(t^-) - \overline{F}^0_j(t)), \sqrt{n}(\widehat{\overline{F}}_{n,j_2|j}(t) - \overline{F}^0_{j_2|j}(t))\}] \\
&= \{\overline{F}^0_j(t)\prod_{j_1\neq j'}\widehat{\overline{F}}_{n,j_1|j}(t)\}^2[\sum_{j_2\neq j'}\frac{\sigma^2_{jj_2}(t)}{(\overline{F}^0_{j_2|j}(t))^2} + \int_0^\infty \frac{dF^0_j(s)}{\pi^0_j\overline{G}^0_j(s)(\overline{F}^0_j)^2(s)} \\
&\qquad + \sum_{j_2\neq j'}\sum_{j_3\neq j',j_2}\frac{\sigma^2_{jj_2j_3}}{\overline{F}^0_{j_2|j}(t)\overline{F}^0_{j_3|j}(t)} + \sum_{j_2\neq j'}\frac{c^{(1)}_{jj_2}(t)}{\overline{F}^0_{j_2|j}(t)}].
\end{aligned}$$

and, for any sequence A_{nj} converging to A_j,

$$\begin{aligned}
&\lim_n \ \mathrm{Var}\sqrt{n}(\prod_j A_{nj} - \prod_j A_j) = \sum_j\prod_{j'\neq j}A^2_{j'}\lim_n n\mathrm{Var}(A_{nj} - A_j) \\
&+\sum_j\sum_{j'\neq j}A_jA_{j'}\prod_{j',j"\neq j}A^2_{j"}\lim_n n\mathrm{Cov}(A_{nj} - A_j, A_{nj'} - A_j).
\end{aligned}$$

Thus

$$(\sigma^{RC}_{jj'})^2 = \frac{1}{\{\prod_{j"}p^0_{jj"}\}^2}\{v^{(2)}_{jj'} + v^{(3)}_{jj'} - 2c^{(2)}_{jj'}\}$$

with

$$
\begin{aligned}
v^{(2)}_{jj'} &\equiv \lim_n \ \mathrm{Var}\sqrt{n}\{\int_0^\infty \widehat{\overline{F}}_{n,j}(s^-) \prod_{j"\neq j'} \widehat{\overline{F}}_{n,j"|j}(s)\, d\widehat{\Lambda}_{n,j'|j}(s) - p^0_{jj'} \prod_{j"} p^0_{jj"}\} \\
&= \int_0^\infty \lim_n \ \mathrm{Var}\sqrt{n}\{\widehat{\overline{F}}_{n,j}(s^-) \prod_{j"\neq j'} \widehat{\overline{F}}_{n,j"|j}(s) - \overline{F}^0_j(s) \prod_{j"\neq j'} \overline{F}^0_{j"|j}(s)\}\, d\Lambda^0_{j'|j}(s) \\
&\quad + \int_0^\infty \{\overline{F}^0_j(s) \prod_{j"\neq j'} \overline{F}^0_{j"|j}(s)\}^2 \lim_n \ \mathrm{Var}\sqrt{n}(d\widehat{\Lambda}_{n,j'|j}(s) - d\Lambda^0_{j'|j}(s)) \\
&= \int_0^\infty v^{(1)}_{jj'}(s)\, d\Lambda^0_{j'|j}(s) + \int_0^\infty \{\overline{F}^0_j(s) \prod_{j"\neq j'} \overline{F}^0_{j"|j}(s)\}^2 \frac{dF^0_j(s)}{\pi^0_j \overline{G}^0_j(s)(\overline{F}^0_{j'|j})^2(s)},
\end{aligned}
$$

$$
\begin{aligned}
v^{(1)}_{jj'}(t) &\equiv \lim_n \ \mathrm{Var}\sqrt{n}\{\widehat{\overline{F}}_{n,j}(t^-) \prod_{j_1\neq j'} \widehat{\overline{F}}_{n,j_1|j}(t) - \overline{F}^0_j(t^-) \prod_{j_1\neq j'} \overline{F}^0_{j_1|j}(t)\} \\
&= \{\prod_{j_1\neq j'} \widehat{\overline{F}}_{n,j_1|j}(t)\}^2 \lim_n [\sum_{j_2\neq j'} \mathrm{Var}\sqrt{n}\{\widehat{\overline{F}}_{n,j_2|j}(t) - \overline{F}^0_{j_2|j}(t)\}\{\frac{\overline{F}^0_j(t)}{\overline{F}^0_{j_2|j}(t)}\}^2 + \mathrm{Var}\sqrt{n}\{\widehat{\overline{F}}_{n,j}(t^-) - \overline{F}^0_j(t)\} \\
&+ \sum_{j_2\neq j'} \sum_{j_3\neq j', j_2} \frac{(\overline{F}^0_j(t))^2}{\overline{F}^0_{j_2|j}(t)\overline{F}^0_{j_3|j}(t)} \mathrm{Cov}\{\sqrt{n}(\widehat{\overline{F}}_{n,j_2|j}(t) - \overline{F}^0_{j_2|j}(t)), \sqrt{n}(\widehat{\overline{F}}_{n,j_3|j}(t) - \overline{F}^0_{j_3|j}(t))\} \\
&+ \sum_{j_2\neq j'} \frac{\overline{F}^0_j(t)}{\overline{F}^0_{j_2|j}(t)} \mathrm{Cov}\{\sqrt{n}(\widehat{\overline{F}}_{n,j}(t^-) - \overline{F}^0_j(t)), \sqrt{n}(\widehat{\overline{F}}_{n,j_2|j}(t) - \overline{F}^0_{j_2|j}(t))\}] \\
&= \{\overline{F}^0_j(t) \prod_{j_1\neq j'} \widehat{\overline{F}}_{n,j_1|j}(t)\}^2 [\sum_{j_2\neq j'} \frac{\sigma^2_{jj_2}(t)}{(\overline{F}^0_{j_2|j}(t))^2} + \int_0^\infty \frac{dF^0_j(s)}{\pi^0_j \overline{G}^0_j(s)(\overline{F}^0_j)^2(s)} \\
&\qquad + \sum_{j_2\neq j'} \sum_{j_3\neq j', j_2} \frac{\sigma^2_{jj_2j_3}}{\overline{F}^0_{j_2|j}(t)\overline{F}^0_{j_3|j}(t)} + \sum_{j_2\neq j'} \frac{c^{(1)}_{jj_2}(t)}{\overline{F}^0_{j_2|j}(t)}], \\
v^{(3)}_{jj'} &= \lim_n \ \mathrm{Var}\sqrt{n}\{\prod_{j'} \widehat{p}_{n,jj'} - \prod_{j'} p^0_{jj'}\} = \sum_{j_1} \sigma^2_{jj_1}\{\prod_{j_2\neq j_1} p^0_{jj_2}\}^2 + \sum_{j_1} \sum_{j_2\neq j_1} p^0_{jj_1} p^0_{jj_2} (\prod_{j_3\neq j_1, j_2} p^0_{jj_3})^2 \sigma^2_{jj_1j_2},
\end{aligned}
$$

and similar calculations give the expression of

$$
c^{(2)}_{jj'} \equiv \lim_n \ \mathrm{Cov}[\sqrt{n}\{\prod_{j'"} \widehat{p}_{n,jj"} - \prod_{j"} p^0_{jj"}\}, \sqrt{n}\{\int_0^\infty \widehat{\overline{F}}_{n,j}(s^-) \prod_{j"\neq j'} \widehat{\overline{F}}_{n,j"|j}(s)\, d\widehat{\Lambda}_{n,j'|j}(s) - p^0_{jj'} \prod_{j"} p^0_{jj"}\}.
$$