INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2007-37

Duration Models and Point Processes

J.-P. FLORENS[1]
D. FOUGÈRE[2]
M. MOUCHART[3]

[1] Institut Universitaire de France, Toulouse School of Economics.
[2] CNRS, CREST-INSEE, Paris, France, CEPR, London, UK and IZA, Bonn, Germany.
[3] Université Catholique de Louvain.

Duration Models and Point Processes*

Jean-Pierre FLORENS, Denis FOUGERE, Michel MOUCHART

Abstract

This survey is devoted to the statistical analysis of duration models and point processes. The first section introduces specific concepts and definitions for single-spell duration models. Section two is devoted to the presentation of conditional duration models which incorporate the effects of explanatory variables. Competing risks models are presented in the third section. The fourth section is concerned with statistical inference, with a special emphasis on non- and semi- parametric estimation of single-spell duration models. Section 5 sets forth the main definitions for point and counting processes. Section 6 presents important elementary examples of point processes, namely Poisson, Markov and semi-Markov processes. The last section presents a general semi-parametric framework for studying point processes with explanatory variables.

JEL Classification : C41, C33, C44, C51

Keywords : duration models, hazard function, point processes, Markov chains, semi-Markovian processes

Corresponding author:
Denis Fougère
CREST-INSEE
15, Boulevard Gabriel Péri
92245 Malakoff Cedex
France
E-mail: fougere@ensae.fr

* This chapter summarizes and updates chapter 19 (on Duration Models) and chapter 20 (on Point Processes) in Mátyás and Sevestre (1996). Being a survey, this chapter by-passes many details, explanations and justifications, given in Mouchart (2004) in a textbook form, with a similar approach and notation

Many economic phenomena are characterized by the observation of a sequence of events on a continuous interval of time. Think, for instance, to observing the dates of a specific type of financial transactions, or to observing the dates of changes of the individual labour market situation (full-time employed, part-time employed, unemployed, etc.). The length of the interval between two successive events is called a duration. A duration is a positive random variable, denoted T, representing the length of a time period spent by an individual or a firm in a given state. For simplicity, we assume that the distribution of T is not defective, *i.e.* $\Pr(T = \infty) = 0$. This variable is also called a failure time when the date of change is interpreted as a breakdown or a failure.

The most elementary duration model is based on a "death process" $\{X_t, t \in \mathbb{R}_+\}$, for which X_t takes its values in the discrete state space $\{E_0, E_1\}$. At the time origin, called the birth date, the process is in state E_0, *i.e.* $X_0 = E_0$. Trajectories of the process X_t have at most a unique transition from state E_0 to state E_1, which occurs at time T, called the death date. Consequently, the duration T generated by a trajectory of the death process X_t is defined as follows:

$$T = \inf\{t \mid X_t = E_1\}.$$

In most structural models, T is a continuous random variable, *but* the empirical distribution function is a discrete time process and nonparametric methods are often based on (functional) transformations of the empirical distribution function, considered as the best estimator of the "true" distribution function. Therefore, in this chapter, we explicitly consider both continuous and discrete durations.

The first section of this survey concentrates on marginal models of durations, *i.e.* models without explanatory variables. It presents the main functions characterizing the distribution of a duration variable, the survivor and hazard functions among others. Section two is devoted to the presentation of conditional duration models, and more particularly, proportional hazards and accelerated life models, which incorporate the effects of explanatory variables in two different ways. In this section, a special emphasis is put on the problem of unobserved individual heterogeneity. The basic duration model treats a single spell (of unemployment, for example) ending with a given kind of transition (from unemployment to employment, for example).[1] But, in general, as a death could be due to various causes, an individual could exit from unemployment to enter one among different states: full-time employment, part-time employment, or training, for example. When a single-spell duration has many (at least two) outcomes, the duration model

[1] Recently, duration models have been used to analyze the determinants of time intervals between two successive changes in the price of a product sold in a given outlet (see, for instance, Fougère, Le Bihan and Sevestre [2007]).

may be modelled by means of a so-called competing risks model. Competing risks models are presented in the third section, which also contains a discussion on their identifiabiliaty. The right–censoring issue is presented here as a particular case of a competing risks duration model. The fourth section is concerned with statistical inference, with a special emphasis on non- and semi- parametric estimation of single-spell duration models.
The remaining part of this chapter is devoted to point processes, which can be viewed as a generalization of duration models. Such processes are a mathematical formalization which allows to examine individual mobilities or transitions between a finite number of discrete states through (continuous) time. They are particularly useful for the microeconometric analysis of labour market dynamics. Section 5 sets forth the main definitions for point and counting processes. Distribution, intensity and likelihood functions of such processes are also examined. Section 6 presents important elementary examples of point processes, namely Poisson, Markov and semi-Markov processes. Such processes are of great interest because they are well adapted to the case of observed censored or truncated realizations. The last section presents a general semiparametric framework for studying point processes with explanatory variables. It also focuses on the definition of martingale estimators, which are particularly useful in this framework.

1 Marginal duration models

1.1 Distribution, survivor and density functions

We first recall the general definition of the distribution function and of its complement, the survivor function. Next, we give more details for the continuous and the discrete cases, particularly from the point of view of the continuity of these functions.

Definition 1.1 ***: Distribution function***
The distribution function of the duration variable T is denoted F and is defined as

$$F(t) = \Pr(T < t), \quad t \geq 0.$$

■

The main properties of the distribution function F are: $F(t) \in [0,1]$, F is monotone non-decreasing, right continuous and $\lim_{t\to\infty} F(t) = 1$.

Definition 1.2 ***: Survivor function***
The survivor function of the duration variable T, denoted S, is defined as

$$S(t) = \Pr(T \geq t) = 1 - F(t) + \Pr(T = t).$$

■

Its main properties are: $S(t) \in [0,1], S$ is monotone non-increasing, left-continuous and $\lim_{t\to\infty} S(t) = 0$.

Definition 1.3 ***: Density function*** *If there exists a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that*

$$F(t) = \int_0^t f(u)du \quad or \quad f(t) = \frac{dF}{dt} = -\frac{dS}{dt},$$

f is called the density of T. ■

Thus, the density function may be interpreted as the "instantaneous probability" of a failure, a death or an exit (from unemployment, for instance). Remember that in the continuous case, there exists a value of t such that $F(t) = S(t) = 0.5$; that value is the median of the distribution.

Definition 1.4 ***: Discrete duration***

$$\exists(f_j, a_j), \quad j \in J \subseteq \mathbb{N},\ f_j > 0, \quad \sum\nolimits_{j\in J} f_j = 1,\ 0 \le a_j < a_{j+1}$$

such that

$$F(t) \quad = \sum\nolimits_{j\in J} f_j\ \mathbf{1}\{t \ge a_j\} \quad = \sum\nolimits_{\{j|a_j\le t\}} f_j$$

$$S(t) \quad = \sum\nolimits_{j\in J} f_j\ \mathbf{1}\{t \le a_j\} \quad = \sum\nolimits_{\{j|a_j\ge t\}} f_j$$

or equivalently

$$f_j \quad = F(a_j) - F\left(a_{j^-}\right) \quad = F(a_j) - F(a_{j-1})$$

$$= S(a_j) - S\left(a_{j^+}\right) \quad = S(a_j) - S(a_{j+1})$$

■

In the framework of a death process, the event $\{T = a_j\}$ means "alive up to age a_{j_-} and dead at age a_j" and that event has probability f_j.

1.2 Truncated distributions and hazard functions

The use of statistical duration models may be justified by several arguments:

(i) Problem of time dependence. Consider the following question. What is the "instantaneous" probability of dying at time t given you are still living at time t_-? More generally, this is the problem of the probability law of duration T, *conditional* on $T \ge t$ (remember that the event $\{T \ge t\}$ means "still alive at time t"). This problem is exactly that of analyzing the dynamic behavior of the process. Such conditional distributions are "truncated" distributions.

(ii) The preceding question is often so natural that modelling those truncated distributions may be *economically more meaningful* than modelling the untruncated distributions. For instance, in job search models, the reservation wage, at a given instant, is a function of the duration of unemployment up to that instant.

(iii) Right-censoring (see Section 3.4) makes truncated distributions particularly useful.

Definition 1.5 ***Integrated Hazard Function***

The integrated hazard function of the duration variable T is denoted Λ and is defined as

$$\begin{array}{rccl} \Lambda: & \mathbb{R}_+ & \to & \mathbb{R}_+ \\ & t & \mapsto & \Lambda(t) = \int_{[0,t[} \dfrac{1}{S(u)} dF(u) \end{array}$$

■

The function Λ is monotone non-decreasing, left-continuous and verifies $\Lambda(0) = 0$ and $\Lambda(\infty) = \infty$. As we will see later, the integrated hazard function is a useful tool for characterizing some duration distributions. Let us consider now the hazard function (or age–specific failure rate).

(i) Continuous case
In the continuous case, there is a density function $f(t)$ and

$$\Lambda(t) = \int_0^t \frac{f(u)}{S(u)} du = -\int_0^t \frac{1}{S(u)} dS(u) = -\ln\, S(t)\,.$$

Definition 1.6 ***Hazard function***

The hazard function of the duration variable T is denoted λ and is defined as

$$\lambda(t) = d\Lambda(t)/dt = f(t)/S(t) = -d\ln\, S(t)/dt\,.$$

■

The function $\lambda(t)$ may be viewed as the "instantaneous probability" of leaving the current state, indeed

$$\lambda(t) = \lim_{\Delta \to 0} \frac{\Pr[t \le T < t + \Delta \mid T \ge t]}{\Delta}\,.$$

Thus, $\lambda(t)$ is also called the "age–specific failure rate" or the "age–specific death rate". The function λ is non negative and $\int_0^t \lambda(u)\, du < \infty$, $\forall t \in R_+$, but $\int_0^\infty \lambda(u)\, du = \infty$ for non-defective distributions. Note that λ is *not* necessarily monotone.

Straightforward relationships between the distribution, survivor and hazard functions should be noticed:

$$\Lambda(t) = \int_0^t \lambda(u)\,du, \quad f(t) = \lambda(t)\exp\left(-\int_0^t \lambda(u)\,du\right)$$

$$S(t) = \exp\left(-\int_0^t \lambda(u)\,du\right), \quad F(t) = 1 - \exp\left(-\int_0^t \lambda(u)\,du\right)$$

showing that each of these functions completely characterizes the distribution of a duration.

Definition 1.7 ***Temporal independence***

The hazard function of the duration T has the property of *temporal independence* if and only if it is constant over time, *i.e.* $\lambda(t) = \lambda,\ \forall t \in \mathbb{R}\ (\lambda > 0)$■

(ii) Discrete case

Remember that, in the discrete case, for any (integrable) function $g(u)$ we have

$$\int_{[0,t[} g(u)dF(u) = \sum_{\{j|a_j<t\}} g(a_j)f_j = \sum_j g(a_j)f_j\mathbf{1}\{a_j < t\}.$$

Therefore

$$\Lambda(t) = \sum_{\{j|a_j<t\}} \frac{f_j}{S(a_j)} = \sum_{\{j|a_j<t\}} \frac{f_j}{f_j + f_{j+1} + \ldots}\,.$$

So, we obtain the discrete version of the (instantaneous) hazard function as

$$\lambda_j = \Lambda(a_{j^+}) - \Lambda(a_j) = \frac{f_j}{f_j + f_{j+1} + f_{j+2} + \ldots} = \frac{f_j}{S(a_j)}\,.$$

In particular, $\lambda_1 = f_1$. The last formula may also be interpreted as

$$\lambda_j = \Pr(T = a_j \mid T \geq a_j)\,.$$

To deduce relationships between survivor and hazard functions in the discrete case, let us write the survivor function as:

$$S(t) = \prod\nolimits_{\{j|a_j<t\}}(1 - \lambda_j)$$

based on the familiar identity

$$\begin{aligned}
a_0 + a_1 &= a_0\left(1 + \frac{a_1}{a_0}\right)\\
a_0 + a_1 + a_2 &= a_0\left(1 + \frac{a_1}{a_0}\right)\left(1 + \frac{a_2}{a_0 + a_1}\right)\\
&\ldots\\
\textstyle\sum_{0\leq j<k} a_j &= a_0 \textstyle\prod_{1\leq j<k}\left(1 + \frac{a_j}{\sum_{0\leq m<j-1} a_m}\right)
\end{aligned}$$

applied to:

$$S(t) = 1 - \sum\nolimits_{\{j|a_j<t\}} \lambda_j \,.$$

Thus we obtain the relationship

$$\ln S(t) \;\; = \sum\nolimits_{\{j|a_j<t\}} (1-\lambda_j) \approx - \sum\nolimits_{\{j|a_j<t\}} \lambda_j = -\Lambda\,(t)$$

if λ_j is "small", *i.e.* $-\ln\,(1-\lambda_j) \approx \lambda_j$. Thus, in the discrete case, $\Lambda(t)$ is *approximately* equal to $-\ln\,S(t)$ if all λ_j are small, while in the continuous case, $\Lambda(t)$ is exactly equal to $-\ln\,S(t)$. Moreover, in the discrete case:

$$f_j = \lambda_j \prod\nolimits_{1\leq i\leq j-1} (1-\lambda_i)\,.$$

Figure 1 presents the main distributions used for the statistical analysis of duration data.

2 Conditional models

2.1 General considerations

a) The two levels of analysis to be considered

(i) For a *descriptive (or exploratory) data analysis*, covariates may be used to control for observable factors of heterogeneity by performing separate analyses.
(ii) When the objective is to estimate a *structural model*, the parameter of interest may be such that the (marginal) process generating some covariates may be uninformative about the parameter of interest which, at the same time, is a function of a parameter sufficient to parametrize the process conditional on those covariates. Those covariates are then called "exogenous variables" and are generally denoted by Z whereas the other variables, denoted by Y (or T, in case of a duration variable), are called "endogenous", because the model describes the way they are generated conditionally on the exogenous variables.
In such a case, it is admissible to specify only the process conditional on those exogenous variables, leaving the marginal process generating those exogenous variables virtually unspecified. In other words, for the parameter of interest, $p(t \mid z, \theta)$ is as informative as $p(t, z \mid \theta)$. According to a general principle of parsimony, the conditional model is therefore preferred.

b) How to specify conditional models

(i) In general, a natural way of specifying conditional models is to make the parameters of a distribution dependent on the conditioning variable. Thus, in $F_T(t \mid \theta)$, one would transform θ into $g(z, \theta)$ where g would be a known function. For example, $Y \sim N(\mu, \sigma^2)$ could be transformed into

Distribution	Parameters	Survivor function $S(t)$	Density function $f(t)$	Hazard function $h(t)$	Time variation of the hazard function $dh(t)/dt$
Exponential	$\lambda > 0$	$\exp(-\lambda t)$	$\lambda \exp(-\lambda t)$	λ	constant $(dh(t)/dt=0)$
Weibull	$\alpha > 0, \lambda > 0$	$\exp(-\lambda t^{\alpha})$	$\alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha})$	$\alpha\lambda t^{\alpha-1}$	increasing if $\alpha>1$ decreasing if $\alpha<1$ constant if $\alpha=1$ $(h(t)=\lambda)$
Gamma	$\lambda>0,\ \gamma>0$	$\frac{1}{\Gamma(\gamma)}\int_t^{\infty}\lambda(\lambda s)^{\gamma-1}\exp(-\lambda s)ds$	$\frac{\lambda(\lambda t)^{\gamma-1}\exp(-\lambda t)}{\Gamma(\gamma)}$	$\frac{\lambda(\lambda t)^{\gamma-1}\exp(-\lambda t)}{\int_t^{\infty}\lambda(\lambda s)^{\gamma-1}\exp(-\lambda s)ds}$	increasing if $\gamma>1$, decreasing if $\gamma<1$, constant if $\gamma=1$ $(h(t)=\lambda)$
Generalized Gamma	$\lambda>0,\ \alpha>0,\ \gamma>0$	$\frac{1}{\Gamma(\gamma)}\int_t^{\infty}\lambda\alpha(\lambda s)^{\alpha\gamma-1}\exp(-(\lambda s)^{\alpha})ds$	$\frac{\lambda\alpha(\lambda t)^{\alpha\gamma-1}\exp(-(\lambda t)^{\alpha})}{\Gamma(\gamma)}$	$\frac{\lambda\alpha(\lambda t)^{\alpha\gamma-1}\exp(-(\lambda t)^{\alpha})}{\int_t^{\infty}\lambda\alpha(\lambda s)^{\alpha\gamma-1}\exp(-(\lambda s)^{\alpha})ds}$	Gamma distribution with parameter $\lambda>0$ *et* $\gamma>0$ *if* $\alpha=1$, exponential distribution with parameter λ if $\alpha=1$ and $\gamma=1$
Lognormal	$\mu,\ \sigma>0$	$1-\Phi\left[\frac{\ln t-\mu}{\sigma}\right]$	$\frac{1}{\sigma t}\phi\left[\frac{\ln t-\mu}{\sigma}\right]$	$\frac{\frac{1}{\sigma t}\phi\left[\frac{\ln t-\mu}{\sigma}\right]}{1-\Phi\left[\frac{\ln t-\mu}{\sigma}\right]}$	increasing, then decreasing
Log-logistic	$\lambda > 0, \gamma > 0$	$\left[1+(\lambda t)^{\gamma}\right]^{-1}$	$\gamma\lambda^{\gamma}t^{\gamma-1}\left[1+(\lambda t)^{\gamma}\right]^{-2}$	$\gamma\lambda^{\gamma}t^{\gamma-1}\left[1+(\lambda t)^{\gamma}\right]^{-1}$	increasing then decreasing if $\gamma>1$, decreasing if $\gamma\leq 1$
Singh-Maddala	$\lambda > 0, \alpha > 0, \gamma > 0$	$\left[1+(\lambda t)^{\gamma}\right]^{-\left(\frac{\alpha}{\lambda}\right)^{\gamma}}$	$\gamma\alpha^{\gamma}t^{\gamma-1}\left[1+(\lambda t)^{\gamma}\right]^{-2}$	$\gamma\alpha^{\gamma}t^{\gamma-1}\left[1+(\lambda t)^{\gamma}\right]^{-1}$	log-logistic $(\lambda, \gamma$ if $\lambda=\alpha$, Weibull $(\alpha^{\gamma}, \gamma$ if $\lambda=0$, exponential (α) if $\lambda=0$ and $\gamma=1$, increasing then decreasing if $\gamma>1$, decreasing if $\gamma<1$

Figure 1: **Examples of distributions for durations**

$(Y \mid Z) \sim N(\alpha + \beta Z, \sigma^2)$. Similarly, $T \sim Exp(\theta)$ could be transformed into $(T \mid Z) \sim Exp[g(Z, \theta)]$ where, e.g. $g(Z, \theta) = \exp(-Z'\theta)$.

(ii) When modelling individual data (and, in particular, duration data), a frequently used strategy consists of starting with a so–called "baseline" distribution for a reference individual, *i.e.* either an individual not belonging to the treatment group (e.g. an individual for which $Z = 0$) or a "representative" individual (e.g. an individual for which $Z = E(Z)$) and thereafter modelling, what makes the other individuals different from that individual of reference. Typical examples are the following:

- in the proportional hazard model, the global effect of all regressors Z is to multiply the baseline hazard function by a scale factor,
- in the accelerated life model, the global effect of all regressors Z is to rescale the duration variable. From now on, we shall only use the notation θ for the complete parameter characterizing the conditional distribution generating $(T \mid Z)$. This vector is decomposed into $\theta = (\alpha, \beta)$ where α parametrizes the baseline distribution and β represents the effect of the exogenous variables.

c) Time–varying and time–constant covariates must be distinguished

The covariates may represent:

- individual characteristics, such as gender, level of education, and so on, which are fixed over time,
- other individual characteristics, such as marital status, number of children, eligibility to social benefits or programs, which are typically varying through time,
- but also characteristics of the macroeconomic environment, such as the unemployment rate, the job vacancy rate, the employment structure, and so on, which are also time–varying but possibly common to several individuals.

Some variables may also represent interactions between several covariates. The dynamic properties of the model and the estimation procedures crucially depends on whether the covariates are time-dependent or not.

d) Interpretation of the parameters

Most models are typically nonlinear in the sense that partial derivatives (of interest) are not constant, but are functions of the values of the covariates and/or of the duration. This feature clearly makes the interpretation of the

coefficients more difficult. Furthermore, those partial derivatives are often not those of conditional expectations (as in regression analysis) but those of hazard functions (*i.e.* of "instantaneous probabilities").

2.2 The proportional hazard or Cox model

a) Definition

In the proportional hazard model, the effect of the exogenous variable is specified as multiplying a baseline hazard function by a function that depends on the exogenous variable. When Z is not time-dependent, this model is defined as

$$\lambda_T(t \mid z, \theta) = \lambda_0(t \mid \alpha) g(z, \beta), \quad \theta = (\alpha, \beta),$$

where $\lambda_0(t \mid \alpha)$ is the so-called baseline hazard function and g is a known function. The proportional hazard model is equivalently characterized as

$$\Lambda_T(t \mid z, \theta) = g(z, \beta) \int_0^t \lambda_0(u \mid \alpha) du = g(z, \beta)\ \Lambda_0(t \mid \alpha)\ ,$$

$$\begin{aligned} S_T(t \mid z, \theta) &= \exp\left\{-g(z, \beta) \textstyle\int_0^t \lambda_0(u \mid \alpha) du\right\} \\ &= \exp\left\{-g(z, \beta)\ \Lambda_0(t \mid \alpha)\right\} \\ &= [S_0(t \mid \alpha)]^{g(z,\beta)} \end{aligned}$$

where Λ_0 and S_0 are implicitly defined. Thus

$$\begin{aligned} f_T(t \mid z, \theta) &= \lambda_T(t \mid z, \theta)\ S_T(t \mid z, \theta) \\ &= g(z, \beta)\ \lambda_0(t \mid \alpha)\ [S_0(t \mid \alpha)]^{g(z,\beta)} \end{aligned}$$

b) Identification

The problem of identifying separately the functions g and λ_0 comes from the fact that for any $k > 0 : g \cdot \lambda_0 = gk \cdot k^{-1}\lambda_0$. A rather natural solution consists of defining a reference individual, *i.e.* a particular value z_0 of Z for which $g(z_0, \beta) = 1, \forall \beta$. Consequently, $\lambda_T(t \mid z_0, \theta) = \lambda_0(t \mid \alpha)$. When $Z = 0$ is meaningful, a typical normalization is $g(0, \beta) = 1$.
In the proportional hazard model with time-constant covariates, the first-order derivative

$$\frac{\partial}{\partial z} \ln\ \lambda_T(t \mid z, \theta) = \frac{\partial}{\partial z} \ln\ g(z, \beta),$$

depends on z and β only and is therefore independent of t.

c) Semi–parametric modelling

When interest is focused on the role of the exogenous variables, α is treated as a nuisance parameter and β is the sole parameter of interest. In such a case, modelling often relies on one of the following two extreme possibilities:

(i) $\lambda_0(t \mid \alpha)$ is specified in the most simplest way such as $\lambda_0(t \mid \alpha) = \lambda_0(t)$, *i.e.*. is completely known, or $\lambda_0(t \mid \alpha) = \alpha$, *i.e.* the baseline distribution is exponential and therefore depends on only one unknown parameter;

(ii) $\lambda_0(t \mid \alpha)$ is specified in the most general way: $\lambda_0(t \mid \alpha) = \alpha(t)$, *i.e.* a functional parameter (α is a non–negative function such that its integral on the positive real line diverges). This is a semiparametric model with parameter $\theta = (\alpha, \beta)$, where α takes its value in a functional space, whereas β takes its value in a (finite dimensional) Euclidean space. This approach is particularly attractive in situations where economic theory would not give much information on the structure of $\lambda_0(t \mid \alpha)$.

d) A particular case

The function $g(z, \beta)$ should clearly be non–negative. An easy way to obtain that property without restriction on β is the log–linear specification, viz.:

$$g(z, \beta) = \exp(z'\beta), \quad \beta \in \mathbb{R}^k.$$

In such a case $\Lambda_0(t|\alpha) = \Lambda_T(t|0, \theta)$. That specification has a number of interesting properties. First, let us remark that:

$$\frac{\partial}{\partial z} \ln \lambda_T(t \mid z, \theta) = \frac{\partial}{\partial z} \ln g(z, \beta) = \beta,$$

i.e. z has a constant proportional effect on the instantaneous conditional probability of leaving state E_0. As z is not time–dependent, one may also write

$$S_T(t \mid z, \theta) = \exp\{-\Lambda_0(t \mid \alpha) \exp(z'\beta)\} = [S_0(t \mid \alpha)]^{\exp(z'\beta)}$$

$$f_T(t \mid z, \theta) = \lambda_0(t \mid \alpha) \exp(z'\beta) \ [S_0(t \mid \alpha)]^{\exp(z'\beta)}$$

Let us define

$$\varepsilon_t = -\ln \Lambda_0(t \mid \alpha) - z'\beta \,.$$

where ε_t has a completely specified distribution, independent of α, z or β, namely a unit double–exponential distribution. Then we may write

$$-\ln \Lambda_0(t \mid \alpha) = z'\beta + \varepsilon_t \,.$$

This is a (non–normal) *nonlinear* regression but linear if α is known. This feature of the proportional hazard model was used by Han and Hausman [1990] for conducting a semiparametric estimation on grouped duration data.

2.3 The accelerated time model

a) The basic idea

In the accelerated time model, the effect of the exogenous variable is specified as modifying the time scale. For the ease of exposition, we assume that the exogenous variables are not time-dependent. The accelerated time model is accordingly defined as

$$T = [g(z,\beta)]^{-1} T_0 \quad \text{or} \quad T_0 = g(z,\beta) T$$

or, equivalently,

$$\begin{array}{ll} \lambda_T(t \mid z,\theta) & = g(z,\beta) \times \lambda_0 \left[t\ g(z,\beta) \mid \alpha\right] \\ \Lambda_T(t \mid z,\theta) & = \Lambda_0 \left[t\ g(z,\beta) \mid \alpha\right] \\ S_T(t \mid z,\theta) & = S_0 \left[t\ g(z,\beta) \mid \alpha\right] \\ f_T(t \mid z,\theta) & = g(z,\beta)\ f_0 \left[t\ g(z,\beta) \mid \alpha\right] \end{array}$$

with, as usual, $\theta = (\alpha, \beta)$. This specification may be particularly attractive when the baseline distribution admits a scale parameter.

b) Empirical test for the accelerated time model

Let us consider the quantile functions, *i.e.* the inverse of the survivor (rather than, as more usually, the distribution) function

$$q_T(p \mid z,\theta) = S_T^{-1}(p \mid z,\theta)\ , \quad 0 \leq p \leq 1,$$

$$q_0(p \mid \alpha) = S_0^{-1}(p \mid \alpha) \quad , \quad 0 \leq p \leq 1.$$

Because of the strict monotonicity (in the continuous case) of the survivor function, we have

$$q_0(p \mid \alpha) = g(z,\beta) \cdot q_T(p \mid z,\theta)\,.$$

In the $\{q_0(p \mid \alpha), q_T(p \mid z,\theta)\}$–space, this gives, for a fixed value of z, an homogenous straight line, the gradient of which is given by $g(z,\beta)$. This feature suggests that an easy empirical test for the accelerated time model may be obtained through an examination of the so–called "Q–Q–plot" (*i.e.* plot of the two quantiles) for a fixed value of Z and a fixed (typically, estimated) value of $\theta = (\alpha, \beta)$.

c) Regression representation of the accelerated time model

The accelerated time model may also be written, in logarithmic terms, as

$$\ln T = \ln T_0 - \ln g(z,\beta).$$

If we define $\mu_0 = E\left[\ln T_0\right]$ and $\varepsilon = \ln T_0 - E\left[\ln T_0\right]$, we may also write

$$\ln T = \mu_0 - \ln g(z,\beta) + \varepsilon.$$

In particular,

(i) if $\ln T_0 \sim N(\mu, \sigma^2)$, *i.e.* $T_0 \sim LN(\mu, \sigma^2)$, then $\varepsilon \sim N(0, \sigma^2)$. Thus we obtain a normal regression model (*if* there is no censoring);

(ii) if $g(z, \beta) = \exp(z'\beta)$, we obtain a linear regression model: $\ln T = \mu_0 - z'\beta + \varepsilon$.

d) Particular case: Weibull baseline

In the particular case of a Weibull baseline distribution, namely $\Lambda_0(t|x) = \lambda t^\tau$, where $\alpha = (\lambda, \tau)$, along with a log-linear effect of the exogenous variable, namely $g(z, \beta) = \exp(\beta' z)$, we obtain:

$$\Lambda_{PH}(t|z, \theta) = \exp\left(\beta'_{PH} z\right) \lambda t^\tau$$

$$\Lambda_{AT}(t|z, \theta) = \lambda[t \exp\left(\beta'_{AT} z\right)]^\tau$$

The two models, proportional hazards and accelerated time, become therefore identical under the reparametrization $\beta_{PH} = \tau \beta_{AT}$.

2.4 Aggregation and heterogeneity

Heterogeneity is the problem created by the non-observability or the omission of relevant exogenous variables. Aggregating over heterogenous individuals may create complicated structures of the hazard function. The analytical aspect is shown, for the general case, in the next lemma. An example illustrates a simple application of this lemma. Then it is shown that aggregation destroys the exponentiality of a duration.

a) A basic lemma

Let $T \mid Z \sim F_T^Z$ and $Z \sim F_Z$, *i.e.*

$$\Pr(T \leq t \mid Z = z) = F_T(t \mid z) \quad \text{and} \quad \Pr(Z \leq z) = F_Z(z)$$

Then

$$\begin{aligned}
f_T(t) &= \int f_T(t \mid z)\ dF_Z(z) \\
S_T(t) &= \int S_T(t \mid z)\ dF_Z(z) \\
\lambda_T(t) &= \frac{f_T(t)}{S_T(t)} = \frac{\int f_T(t \mid z)\ dF_Z(z)}{\int S_T(t \mid z)\ dF_Z(z)} \\
&= \int \lambda_T(t \mid z)\ \frac{S_T(t \mid z)}{\int S_T(t \mid z)\ dF_Z(z)} dF_Z(z) \\
&= \int \lambda_T(t \mid z)\ dF_Z(z \mid T \geq t)
\end{aligned}$$

■

This lemma may be interpreted as follows: aggregating over heterogenous individuals, characterized by z, produces a duration distribution for which the hazard function $\lambda_T(t)$ is a *weighted* average of the individual hazard functions $\lambda_T(t \mid z)$. This possibly complicated weighting scheme may eventually account for complex hazard functions when analyzing aggregate data. A simple example illustrates this point.

b) An example

Let $Z = 0$ for individuals with a low educational level, and $Z = 1$ for individuals with a high educational level. The distribution of this variable over the whole population is defined by $\Pr(Z = z) = \theta^z(1-\theta)^{1-z}$. Moreover, we suppose that:

$$(T \mid Z = j) \sim F_T^j, \quad j = 0, 1$$

Then we can deduce

$$\begin{aligned}
f_T(t) &= \theta f_T(t \mid z = 1) + (1-\theta) f_T(t \mid z = 0) \\
S_T(t) &= \theta S_T(t \mid z = 1) + (1-\theta) S_T(t \mid z = 0) \\
\lambda_T(t) &= \frac{f_T(t)}{S_T(t)} = \theta \frac{f_T^1(t)}{\theta S_T^1(t) + (1-\theta) S_T^0(t)} \\
&\qquad + (1-\theta) \frac{f_T^0(t)}{\theta S_T^1(t) + (1-\theta) S_T^0(t)} \\
&= \lambda_T^1(t) \frac{\theta S_T^1(t)}{\theta S_T^1(t) + (1-\theta) S_T^0(t)} + \lambda_T^0(t) \frac{(1-\theta) S_T^0(t)}{\theta S_T^1(t) + (1-\theta) S_T^0(t)}
\end{aligned}$$

c) The "mover–stayer" lemma

Lemma. If $(T \mid Z) \sim Exp\{\lambda_0(Z)\}$ and $Z \sim F_Z$ arbitrary, then $\lambda_T(t)$ is monotone decreasing. ■

Proof. Indeed, we successively obtain:

$$\begin{aligned}
S_T(t) &= \int_0^\infty S_T(t \mid z)\ dF_Z(z) = \int_0^\infty \exp[-t\ \lambda_0(z)]\ dF_Z(z) \\
f_T(t) &= -\frac{d}{dt} S_T(t) = \int_0^\infty \lambda_0(z) \exp[-t\ \lambda_0(z)]\ dF_Z(z) \\
\lambda_T(t) &= \frac{f_T(t)}{S_T(t)} = \frac{\int_0^\infty \lambda_0(z) \exp[-t\ \lambda_0(z)]\ dF_Z(z)}{\int_0^\infty \exp[-t\ \lambda_0(z)]\ dF_Z(z)}
\end{aligned}$$

It is then easy to check that

$$\frac{d}{dt} \lambda_T(t) < 0\ \forall t, \forall F_Z(Z),\ \forall \lambda_0(Z)$$

(see, for example, Fourgeaud, Gourieroux and Pradel [1990]).
This lemma may be interpreted as follows. Individuals are characterized by their value of z. Large values of $\lambda_0(z)$ represent so-called "movers": they will leave first, while individuals represented by small value of $\lambda_0(z)$, the so-called "stayers", will leave (in probability) later. This explains why $\lambda_T(t)$ will be decreasing because being determined at each t by the remaining individuals with smaller values of $\lambda_0(z)$. This lemma also shows that although each individual duration has exponential duration, the appropriate distribution not only is not exponential but has necessarily a decreasing hazard rate, whatever is the distribution of Z.

2.5 Endogeneity

In the previous section, we have considered models where the covariates are exogenous. In many cases, this assumption is not realistic. Consider, for example, a model constructed in the following way: T is a duration generated conditionally on $Z = (Z_1, Z_2)$, where Z_2 is an individual characteristic and Z_1 is the level of a treatment. The variable Z_2 is known by persons who assign the treatment but unknown by the statistician. If the parameters of interest are the parameters of the conditional distribution of T given (Z_1, Z_2) these parameters are in general not identified by the conditional distribution of T given Z_1 (after integration of Z_2). Using econometric terminology, Z_1 becomes an endogenous variable. Endogeneity of treatments in duration models has been studied by Abbring and Van den Berg [2003].

3 Competing risks and multivariate duration models

3.1 Multivariate durations

a) Introduction

Multivariate durations distributions are used in different situations. The first context is the analysis of multivariate elementary point processes, which occurs when we observe life lengths of several individuals belonging to the same family, or unemployment spells of couples. This is also the case when, for a given individual, we define a multivariate point process corresponding, for instance, to her labour market trajectories and to her marriage/divorce history. Another use is in point processes with more than one transition, as in the analysis of biographical data on unemployment. Yet another use is in situations where the vector of durations is latent and some sampling scheme allows one to observe only a part of this vector; this is the case in competing risks models to be presented later on.

In this section we focus our attention on general issues, namely basic definitions and properties, and methods of construction. For expository purposes we limit the presentation to bivariate distributions; extensions to more than two dimensions are fairly obvious, although notations may become cumbersome.

b) Basic concepts

We start with the *multivariate survivor function* defined and denoted as

$$S_{T_1,T_2}(t_1,t_2) = \Pr(T_1 \geq t_1, T_2 \geq t_2)\,.$$

In what follows we assume that S_{T_1,T_2} is twice differentiable but in the last section we show how to treat a continuous but not everywhere differentiable survivor function as well. The *multivariate density* is defined as

$$f_{T_1,T_2}(t_1,t_2) = \frac{\partial^2}{\partial t_1 \partial t_2} S_{T_1,T_2}(t_1,t_2)\,.$$

The *marginal survivor and density functions* are defined asas

$$\begin{aligned} S_{T_1}(t_1) &= S_{T_1,T_2}(t_1,0) \\ f_{T_1}(t_1) &= -\frac{d}{dt_1} S_{T_1}(t_1) \end{aligned}$$

and similarly for T_2. Often we shall write, for simplicity, $S_{1,2}$, $f_{1,2}$ or $S_j (j = 1,2)$ instead of S_{T_1,T_2}, etc.
Conditional distributions occur in different contexts and should be carefully distinguished according to the relevant conditioning event. Thus we need both $S_{1|2}(t_1|T_2 = t_2)$, $f_{1|2}(t_1|T_2 = t_2)$ and $S_{1|2}(t_1|T_2 \geq t_2)$, $f_{1|2}(t_1|T_2 \geq t_2)$. They are defined and denoted as follows:

$$S^{\geq}_{1|2}(t_1|t_2) = \Pr(T_1 \geq t_1 \mid T_2 \geq t_2) = \frac{S_{1,2}(t_1,t_2)}{S_2(t_2)}$$

$$f^{\geq}_{1|2}(t_1|t_2) = -\frac{\partial}{\partial t_1} S_{1|2}(t_1 \mid T_2 \geq t_2) = \frac{-\dfrac{\partial}{\partial t_1} S_{1,2}(t_1,t_2)}{S_2(t_2)}\,.$$

Furthermore, as shown more precisely in next subsection,

$$S^{=}_{1|2}(t_1|t_2) = \Pr(T_1 \geq t_1 | T_2 = t_2) = -\frac{\dfrac{\partial}{\partial t_2} S_{1,2}(t_1,t_2)}{f_2(t_2)}$$

$$f^{=}_{1|2}(t_1|t_2) = -\frac{\partial}{\partial t_1} S^{=}_{1|2}(t_1|t_2) = \frac{f_{1,2}(t_1,t_2)}{f_2(t_2)}\,.$$

To each of these univariate conditional distributions, there corresponds a unique hazard function. For instance, marginal hazard functions are defined and denoted as:

$$\begin{aligned}\lambda_j(t_j) &= \lim_{\Delta\downarrow 0}\frac{1}{\Delta}\Pr\left[t_j \le T_j < t_j+\Delta \mid T_j \ge t_j\right] \\ &= -\frac{d\ln\ S_j(t_j)}{dt_j} = \frac{f_j(t_j)}{S_j(t_j)}\end{aligned}$$

Conditional hazard functions are respectively defined as

$$\begin{aligned}\lambda^{\geq}_{1|2}(t_1|t_2) &= \lim_{\Delta\downarrow 0}\frac{1}{\Delta}\Pr\left[t_1 \le T_1 < t_1+\Delta \mid T_1 \ge t_1, T_2 \ge t_2\right] \\ &= \frac{f^{\geq}_{1|2}(t_1|t_2)}{S^{\geq}_{1|2}(t_1|t_2)} = -\frac{\partial}{\partial t_1}\ln\ S_{1,2}(t_1,t_2) \\ \lambda^{=}_{1|2}(t_1|t_2) &= \lim_{\Delta\downarrow 0}\frac{1}{\Delta}\Pr\left[t_1 \le T_1 < t_1+\Delta \mid T_1 \ge t_1, T_2 = t_2\right] \\ &= \frac{f^{=}_{1|2}(t_1|t_2)}{S^{=}_{1|2}(t_1|t_2)} \\ &= -\frac{\partial}{\partial t_1}\left[\ln\ \left(-\frac{\partial}{\partial t_2}S_{1,2}(t_1,t_2)\right)\right]\end{aligned}$$

c) Construction of multivariate distributions

Several techniques for constructing multivariate distributions are worth mentioning. The most trivial one is the case of independent components in which case the joint survivor and density functions are the products of (arbitrary) corresponding marginal functions, and in which the conditional survivor, density and hazard functions coincide with the corresponding marginal functions.
For the dependent case, two general procedures are: (i) take two univariate distributions, choose one to be marginal and take the other one to be conditional to the first by making its parameters to be a function of the conditioning variable; (ii) take a joint distribution with survivor $S(t_1,t_2,y)$ where y is an auxiliary variable such that $S(t_1,t_2 \mid y)$ is meaningful, and marginalize it into $S_{1,2}(t_1,t_2)$.

3.2 Competing risks models: definitions

Competing risks duration models apply to situations where the state space E has more than two elements : $E = \{E_0, E_1, \ldots E_J\}$, $J > 2$. Such models

involve specifying not only the date at which the process leaves the initial state E_0, but also which state in $\{E_1, \dots E_J\}$ is entered.
Consider, for instance, a medical trial where a patient is submitted to a "treatment" for a supposedly known disease and where the survival time is observed. Typically, the cause of death is multiple; in particular, it may be different from the disease for which the treatment was originally designed, and the cause is possibly associated with the treatment itself. One says that several risks "compete" to cause the death of the patient. Similarly, in the labour market, when the initial state E_0 is unemployment, it may be relevant to distinguish several exit states, for example full-time employment, part–time employment or early retirement. The relevance of these distinctions is based on the fact that economic, social and institutional factors may be important to explain both durations and transitions of the individual trajectories; in other words, they are particularly important when analyzing biographical data.
Thus the data have the form (T, K) where T is the sojourn duration in the initial state and K is the destination state. Therefore the law of such a process is specified by the so-called sub-distribution

$$\Pr(T \geq t, K = k) = \Pr(T_j \geq T_k \geq t, \quad \forall j \neq k)$$

Competing risk models provide a specification of $\Pr(T \geq t, K = k)$ based on the following idea. T represents the duration of sojourn in the initial state E_0, whatever the destination state is. The latent random variable T_j would represent the duration of sojourn in the initial state if E_j were the only possible destination. In the competing risk models, if ties have zero probability, i.e. $\Pr(T_i = T_j) = 0, \ \forall i \neq j$, the T_j's are connected by the relationships:

$$T = \min_j \{T_j\}, \quad j = 1, \cdots J,$$

$$K = \ \operatorname{argmin}_j \{T_j = T\}.$$

Thus, the T_j's are latent duration variables because only their minimum is observed. This structure permits to write easily the marginal laws of T and K, which are given by:

$$S_T(t) = \Pr(T \geq t) = \Pr\left\{\cap_{j=1,\dots,K} (T_j \geq t)\right\}$$
$$\Pr[K = k] = \Pr\left\{\cap_{j \neq k} (T_k < T_j)\right\}$$

Intuitively, k is the index of the lowest latent duration (given an ascending order on the $j's$). In order to evaluate the likelihood function, we start by the joint survivor function, using * as an upper index in the notation of the joint distribution of the latent durations $(T_1, ..., T_J)$ to stress that those durations are latent:

$$S^*(t_1, \ldots t_J) = \Pr(T_1 \geq t_1, \ldots, T_J \geq t_J)$$

for any $(t_1, \ldots t_J) \in \mathbb{R}_+^J$. The survivor function of the observed duration $T = \min_j(T_j)$ satisfies

$$S_T(t) = S^*(t, \ldots t), \quad t \in \mathbb{R}^+$$

The marginal survivor function of the latent duration T_j, for $j = 1, ..., J$, is denoted S_j and defined as:

$$S_j(t_j) = S^*(0, \ldots 0, t_j, 0, \ldots 0)\,.$$

In the case where the T_j's are independent, we have

$$S^*(t_1, \ldots t_J) = \Pi_{j=1}^J S_j(t_j).$$

Now, let us suppose that the functions S^* and consequently S_T and S_j are continuously differentiable. The marginal and relevant conditional hazard functions of the latent duration T_j, for $j = 1, ..., J$, are denoted and defined as

$$\lambda_j(t) \quad = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr(t \leq T_j < t + \Delta \mid T_j \geq t) = -d\ln\, S_j(t)/dt, \quad t \in \mathbb{R}_+,$$

$$\begin{aligned} \lambda_{j|T}^{\geq}(t) &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr(t \leq T_j < t + \Delta \mid T \geq t) \\ &= -\frac{\partial}{\partial t_j} \ln\, S^*(t_1, \ldots t_J)\mid_{t_1 = t_2 = \cdots = t_J = t} \end{aligned}$$

where $\lambda_{j|T}^{\geq}(t)$ is a short cut for $\lambda_{T_j|T}^{\geq}(t|t)$. When the T_j's are mutually independent, it is obvious that:

$$\lambda_{j|T}^{\geq}(t) = \lambda_j(t), \qquad \text{for any } t \in \mathbb{R}_+.$$

The hazard function of the observed duration T is denoted and defined as

$$\begin{aligned} \lambda_T(t) &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr(t \leq T < t + \Delta \mid T \geq t) \\ &= -d\ln\, S_T(t)/dt\,, \quad t \in \mathbb{R}_+. \\ &= \sum_{j=1}^{J} \lambda_{j|T}^{\geq}(t) \end{aligned}$$

because, in the definition of $h_T(t)$, the derivative of $S_T(t)$ is a directional derivative (in the direction of the main diagonal $(1, 1, \cdots 1)$) of $S^*(t_1 \cdots t_J)$. In the continuously differentiable case, the likelihood function may be evaluated by differentiating the sub-distribution, namely:

$$\begin{aligned} l_{T,K}(t,k) &= -\frac{d}{dt}\Pr(T \geq t, K = k) \\ &= -\frac{d}{dt}\Pr\left\{\cap_{j\neq k}\left(T_j > T_k \geq t\right)\right\} \end{aligned}$$

Remember that a basic result of differential calculus gives :

$$S^*(t_1 \cdots t_k) = -\int_{t_k}^{\infty} \frac{\partial}{\partial t_k} S^*(t_1 \cdots t_k)\, dt_k$$

and, similarly, a basic result of conditional probability gives :

$$S^*(t_1 \cdots t_k) = \int_{t_k}^{\infty} S_{\bar{k}|k}(t_1, \cdots, t_{k-1}, t_{k+1}, \cdots, t_K \mid T_k = t_k)\; f_k(u)\; du$$

where $\bar{k} = \{1, 2, \cdots K\} \setminus \{K\}$ and

$$\begin{aligned} &S^{=}_{\bar{k}|k}(t_1 \cdots, t_{k-1}, t_{k+1}, \cdots, t_K \mid t_k) \\ &= \Pr\left(T_1 \geq t_1, \cdots, T_{k-1} \geq t_{k-1}, T_{k+1} \geq t_{k+1}, \cdots, T_K \geq t_K \mid T_k = t_k\right) \end{aligned}$$

Thus the likelihood function may be written as :

$$S^{=}_{\bar{j}|j}(t_1, \cdots, t_{j-1}, t_{j+1}, \cdots, t_J \mid t_j) = -\frac{\frac{\partial}{\partial t_j} S^*(t_1, \cdots, t_J)}{f_j(t_j)}\,.$$

In the sequel we use the following simplified notation

$$S^{=}_{\bar{j}|j}(t) = S^{=}_{\bar{j}|j}(t, t, \cdots, t \mid t)$$

Then, the sub-distribution may be written as :

$$\int_{t_k}^{\infty} S_{\bar{k}|k}(u) f_k(u) du$$

Therefore,

$$\begin{aligned} l_{T,K}(t,k) &= -\frac{d}{dt}\int_t^{\infty} S^{=}_{\bar{k}|k}(u) f_k(u) du \\ &= S^{=}_{\bar{k}|k}(t) \times f_k(t) \\ &= -\frac{\partial}{\partial t_k} S^*(t, \cdots, t) \\ &= -S^*(t, \cdots t) \times \frac{\partial}{\partial t_k} \ln\, S^*(t, \cdots t) \end{aligned}$$

Using a disjunctive coding for the exit state, namely

$$A = (A_1, \cdots, A_J), \quad A_j = \mathbb{1}\{K = j\}$$

we may also write

$$l_{T,A}(t,a) = \prod_{j=1}^{J} \left[f_j(t) S^{=}_{\bar{j}|j}(t) \right]^{a_j} = S_T(t) \prod_{j=1}^{J} \left[\lambda^{\geq}_{j|T}(t) \right]^{a_j} .$$

In case of independent latent durations, we have :

$$\begin{aligned} l_{T,K}(t,k) &= f_k(t) \prod_{j \neq k} S_j(t) \\ &= \lambda_k(t) S_T(t) . \end{aligned}$$

3.3 Identifiability of competing risks models

The basic idea of competing risks models is to interpret the data (T, K), representing the sojourn duration in the initial state and the label of the exit state, as the observation of the minimum component of a random vector along with the coordinate where the minimum is obtained. Intuition suggests that these observations give no information on the question whether the coordinate of the random vector, *i.e.* of the latent durations, are independent or not. This intuition is confirmed by next theorem

Theorem 3.1

Let us denote $\mathcal{S} = \{S^*(t_1 \cdots t_J)\}$ the set of J-dimensional survivor functions, $\mathcal{S}_I = \{S^* \in \mathcal{S} \mid S^*(t_1 \cdots t_J) = \prod_j S_j(t_j)\}$ the subset of J-dimensional survivor functions with independent components, $l^*(t,k)$ the likelihood function for a model in $\mathcal{S}$, and $l_I(t,k)$ the likelihood function for a model in $\mathcal{S}_I$. Then:

$$\forall S^* \in \mathcal{S}, \quad \exists!\ S_I \in \mathcal{S}_I \text{ such that } l^*(t,k) = l_I(t,k)$$

In particular,

$$\lambda^{*,\geq}_{j|T}(t) = \lambda_{j,I}(t)$$

■

In the continuous case, the proof of this theorem comes from the fact that, in the general case, $l^*(t,k) = \lambda^{*,\geq}_{k|(t)} S_T(t)$ and that $\lambda_T(t) = \sum_j \lambda^{*,\geq}_{j|T}(t)$, *i.e.* the distribution of the observed duration depends only on the sum of the conditional hazard functions. Therefore the equality $\lambda^{*,\geq}_{k|T}(t) = \lambda_{k,I}(t)$ ensures the equality of likelihood functions. Mouchart and Rolin [2002] gives a slightly more general statement and proof of this theorem.

This theorem means that to any competing risks model with dependent latent durations, one may associate an observationally equivalent model with independent latent durations. The rule of association is simply to build the joint latent distribution with marginal hazard functions of the independent model that are equal to the conditional hazard functions of the dependent model. To illustrate this point, we can consider the following bivariate example. Suppose that the joint survivor function of the two latent durations (T_1, T_2) is given by:

$$S^*(t_1, t_2) = \exp\{1 - \alpha_1 t_1 - \alpha_2 t_2 - \exp[\alpha_{12}(\alpha_1 t_1 + \alpha_2 t_2)]\}$$

where $\alpha_1, \alpha_2 > 0$ and $\alpha_{12} > -1$. Here the parameter α_{12} measures the dependence between the two latent durations T_1 and T_2 in the sense that T_1 and T_2 are independent once $\alpha_{12} = 0$. The conditional and marginal hazard functions of this model are respectively:

$$\lambda^{*,\geq}_{j|T}(t) = \alpha_j\{1 + \alpha_{12}\exp[\alpha_{12}(\alpha_1 + \alpha_2)t]\},\ j = 1, 2$$

and

$$\lambda^*_j(t) = \alpha_j[1 + \alpha_{12}\exp(\alpha_j\alpha_{12}t)], \quad j = 1, 2.$$

Marginal survivor functions are then

$$S^*_j(t_j) = \exp[1 - \alpha_j t_j - \exp(\alpha_{12}\alpha_j t_j)], \quad j = 1, 2,$$

from which it is obvious that

$$S^*(t_1, t_2) \neq S^*_1(t_1)S^*_2(t_2), \quad (t_1, t_2) \in \mathbb{R}^2_+$$

except if $\alpha_{12} = 0$. The likelihood element of an observation (t, k) may be written as

$$\begin{aligned} l_*(t, k) \quad &= \alpha_k\{1 + \alpha_{12}\exp[\alpha_{12}(\alpha_1 + \alpha_2)t]\} \\ &\quad \times \exp\{1 - (\alpha_1 + \alpha_2)t - \exp[\alpha_{12}(\alpha_1 + \alpha_2)t]\} \end{aligned}$$

The observationally equivalent model (*i.e.* having the same likelihood function) with independent latent durations has marginal hazard functions given by $\lambda^{*,\geq}_{j|T}(t)$ above and eventually marginal and joint survivor functions of latent durations given by:

$$S_{j,I}(t_j) = \exp\left\{\frac{\alpha_j}{\alpha_1 + \alpha_2} - \alpha_j t_j - \frac{\alpha_j}{\alpha_1 + \alpha_2}\exp\alpha_{12}(\alpha_1 + \alpha_2)t_j\right\}, \quad j = 1, 2$$

$$\begin{aligned} S_I(t_1, t_2) = \quad \exp\Bigg\{1 - \alpha_1 t_1 - \alpha_2 t_2 - \frac{1}{\alpha_1 + \alpha_2}\big[&\alpha_1\exp\alpha_{12}(\alpha_1 + \alpha_2)t_1 \\ &+\alpha_2\exp\alpha_{12}(\alpha_1 + \alpha_2)t_2\big]\Bigg\} \end{aligned}$$

Note that the latent models are clearly different unless $\alpha_{12} = 0$, *i.e.*

$$S^*(t_1, t_2) \neq S_I(t_1, t_2)$$

but the statistical models are observationally equivalent, *i.e.* $l_*(t, k) = l_I(t, k)$. Note also that both latent models have been identifiably parametrized, but the parameters have very different meaning in the two latent models. In particular, α_{12} measures the association among the latent variables in the case of dependence whereas α_{12} is a common parameter of the two marginal distributions in the case of independent latent variables. The identifiability of the competing-risks duration model with unobserved heterogeneity has been studied by Heckman and Honoré [1989]. Their results have been completed by those obtained by Honoré [1993] for duration models with multiple spells and with unobserved heterogeneity.

3.4 Right–censoring

One usual feature of duration data is that the sampling scheme often produces right–censored observations, *i.e.* observations which have not yet left the initial state E_0 at the end of the sampling period. For example, in the case of single-spell unemployment duration data, the sampling scheme is often the following. Individual observations are sampled from the inflow of individuals entering unemployment at time t_0 and followed up until date C, which is possibly determined by the researcher. Now let us assume that C is greater than t_0. Some observations correspond to individuals leaving the unemployment status before C, in which case they generate complete unemployment durations. Other sampled individuals have not left the unemployment state at date C and so they generate right–censored unemployment durations. Rather than sampling from the inflow into unemployment at a given date t_0, the analyst may sample from inflows considered at several staggered dates $t_0^1, t_0^2 \ldots$ and follow up observations once again up to a censoring time C. Right–censoring can be modelled using the framework of the competing risks models with state space $\{E_0, E_1, \ldots, E_K\}, K > 1$, where the last state E_K denotes the right-censored situation. To illustrate this kind of formalization, let us consider a bivariate competing risks model (T_1, T_2) with state space $\{E_0, E_1, E_2\}$, E_0 labelling unemployment, E_1 employment and E_2 right–censoring. Thus $T_2 = C$. In other words, censoring is often associated with a residual state in a model with multiple states. Suppose first that all individual observations are sampled at the same date t_0. Without loss of generality, one may write $t_0 = 0$ (after some relevant time translation). Within the framework presented in the previous section, this model may be viewed as resulting from a latent survivor function $S_{1,2}(t_1, t_2 \mid \theta)$ with parameter θ, and a generic element of the likelihood function may be

written as:

$$l_{T,D}(t,d) = \left[f_1(t \mid \theta) S^{=}_{2|1}(t \mid \theta)\right]^d \left[f_2(t \mid \theta) S^{=}_{1|2}(t \mid \theta)\right]^{1-d}$$

where $D = \mathbb{1}_{\{T_1 \leq T_2\}}$. In view of the identification problem, and because in many cases censoring mechanisms are independent of the unemployment process, it is often assumed that T_1 and T_2 are independent. Then,

$$l_{T,D}(t,d) = [f_1(t \mid \theta) S_2(t \mid \theta)]^d \, [f_2(t \mid \theta) S_1(t \mid \theta)]^{1-d}$$

If moreover θ may be factorized into $\theta = (\theta_1, \theta_2)$, such that θ_1 characterizes the distribution of T_1 and θ_2 the distribution of T_2, the likelihood reduces to

$$l_{T,D}(t,a) = L_1(\theta_1) L_2(\theta_2)$$

where

$$L_1(\theta_1) = f_1(t \mid \theta_1)^d S_1(t \mid \theta_1)^{1-d} .$$

The parameters of interest are in general those of the distribution of duration T_1, and their estimation could be deduced from $L_1(\theta_1)$ only. Then the generic element of the relevant factor of the likelihood function is $f_1(t \mid \theta_1)$ (resp. $S_1(t \mid \theta_1)$) for an uncensored (resp. a right-censored) observation.
Another model generating censored data may be the following one. Let T_0 be the age of an individual entering unemployment. This age is randomly generated by the individual previous labour market history. The duration of the unemployment spell is T_1 and the age at the end of the unemployment spell is then $T_0 + T_1$. The econometric model specifies the joint distribution of (T_0, T_1) and these two random variables are not, in general, assumed to be independent. A natural specification could be a sequential one: the (marginal) distribution of T_0 is first specified and a conditional distribution of T_1 given T_0 completes the model.
Let us now assume that all the individuals are observed at a given date T_*. In general this date is also random but, for simplicity, we consider T_* as fixed (the model is conditional to T_*). Let us also assume that the sample is constructed in such a way that $T_0 \leq T_*$ (all the individuals have entered unemployment). Then T_0 is always observed but T_1 is not censored if $T_0 + T_1 \leq T_*$. Otherwise, the unemployment spell duration is censored.
Let us define $T_2 = T_* - T_0$. From the distribution of (T_0, T_1) we obtain the distribution of (T_1, T_2), and we may consider the observations as generated by a censored duration model: T_1 is observed only if $T_1 \leq T_2$. But the following specification of a likelihood based on the generic element:

$$l_{T,D}(t,d) = f_1(t)^d S_1(t)^{1-d}$$

where $T = \min(T_1, T_2)$, $D = \mathbb{1}(T_1 \leq T_2)$, f_1 and S_1 are the density and the survivor functions of T_1, is incorrect for two reasons:
i) First if T_0 and T_1 are dependent, T_1 and T_2 are also dependent and the likelihood function must be based on their joint distribution.
ii) The censoring mechanism is different from the usual competing risks model because T_0 or T_2 is always observed and the likelihood of the actual data must be the density of (T_2, T, D). The generic element of this likelihood is then

$$l_{T_2,T,D}(t_2, t, d) = f_2(t_2)\ f^{=}_{1|2}(t \mid t_2)^a\ S^{=}_{1|2}(t \mid t_2)^{1-d}$$

using our previous notations. Finally, note that the identification result of subsection 3.3 does not apply to this case since the censoring mechanism is different from the competing risks model.

4 Inference in duration models

4.1 Introduction

Models actually used in econometrics for dealing with duration data are characterized by two noteworthy features: durations are non-negative random variables and most data sets involve right-censored data. In this section, we focus our attention on the implications of censoring, both for adapting the inference procedure and for evaluating the consequences of misspecification. We first review the inference in parametric models, both in the marginal and in the conditional case, with a particular attention on a rigorous specification of the likelihood function; next we consider non- and semi-parametric models. In each case, we first specify the structure of the model and next give some illustrations with significantly relevant particular cases.

4.2 Parametric models

4.2.1 Inference in marginal models

a) The basic model

The basic model considers a random censoring process that is indepen-

dent of the duration variable. Let us introduce the following notations:

$$\begin{array}{rcl}
\eta & = & (\eta_1 \dots \eta_n)' \quad \text{denote } \textit{latent durations}, \\
\zeta & = & (\zeta_1 \dots \zeta_n)' \quad \text{denote } \textit{latent censoring indicators}, \\
T & = & (T_1 \dots T_n)', \text{ with } T_i = \eta_i \wedge \zeta_i, \text{ are } \textit{observed durations} \\
D & = & (D_1 \dots D_n)', \text{ with } D_i = \mathbb{1}_{\{\eta_i \leq \zeta_i\}} = \mathbb{1}_{\{T_i = \eta_i\}}, \\
X & = & (X_1 \dots X_n)', \text{ with } X_i = (T_i, D_i)', \text{ denote } \textit{complete data} \\
X & = & (T, D) \text{ with } \dim(X) = (n \times 2, 1) \\
\phi & \text{is a} & \text{sufficient parametrisation for the process generating } (\eta, \zeta)
\end{array}$$

Assumptions

A.1 (independent sampling): $\perp\!\!\!\perp_i(\eta_i, \zeta_i) \mid \phi$
A.2 (independent censoring): $\eta_i \perp\!\!\!\perp \zeta_i \mid \phi$
A.3 (definition of θ as a sufficient parametrization for η): $\eta_i \perp\!\!\!\perp \phi \mid \theta$
A.4 (definition of ω as a sufficient parametrization for ζ): $\zeta_i \perp\!\!\!\perp \phi \mid \omega$
A.5 (variation-free parameters) : $(\theta, \omega) \in \Theta_\theta \times \Theta_\omega$
A.6: θ is the only parameter of interest

Latent likelihood

Under $(A.1)$ to $(A.5)$, the complete latent likelihood is therefore:

$$L^{**}(\phi) = \prod_i f_\eta(\eta_i \mid \theta) \cdot \prod_i f_\zeta(\zeta_i \mid \omega) = L_1^*(\theta) \cdot L_2^*(\omega)$$

Under $(A.6)$, the relevant latent likelihood is

$$L_1^*(\theta) = \prod_i f_\eta(\eta_i \mid \theta) = f_\eta(\eta \mid \theta)$$

Actual likelihood

Considering the actually available data, namely (T, D), the complete actual likelihood is

$$\begin{array}{rcl}
L(\phi) & = & \displaystyle\prod_i f_\eta(T_i \mid \theta)^{D_i} S_\eta(T_i \mid \theta)^{1-D_i} \prod_i f_\zeta(T_i \mid \omega)^{1-D_i} S_\zeta(T_i \mid \omega)^{D_i} \\
& = & L_1(\theta) L_2(\omega)
\end{array}$$

Under $(A.6)$, the relevant actual likelihood is:

$$L_1(\theta) = \prod_i f_\eta(T_i \mid \theta)^{D_i} S_\eta(T_i \mid \theta)^{1-D_i} = \prod_i \lambda_\eta(T_i \mid \theta)^{D_i} S_\eta(T_i \mid \theta)$$

Thus the logarithm of the relevant actual likelihood is:

$$\begin{aligned} L(\theta) &= \ln L_1(\theta) = \sum_i D_i \ln f_\eta(T_i \mid \theta) + \sum_i (1 - D_i) \ln S_\eta(T_i \mid \theta) \\ &= \sum_i [D_i \ln \lambda_\eta(T_i \mid \theta) + \ln S_\eta(T_i \mid \theta)] \\ &= \sum_i [D_i \ln \lambda_\eta(T_i \mid \theta) - \Lambda_\eta(T_i \mid \theta)] \end{aligned}$$

b) The exponential case

The consequences of censoring are best understood by considering with some detail the case where the duration of interest is exponentially distributed, which means that $f_\eta(\eta_i \mid \theta) = \theta e^{-\theta \eta_i}$ while $f_\zeta(\zeta_i \mid \omega)$ is left unspecified. Thus, the latent process generating η is a member of the exponential family, $\sum_i \eta_i = \eta_+$ is a minimal sufficient complete statistic of the latent process and, for a sample of size n, the Fisher information is $n\theta^{-2}$. With censoring, the relevant actual likelihood is written as:

$$L(\theta) = \sum_i D_i \ln \theta - \sum_i T_i \theta = (\ln \theta) D_+ - \theta T_+$$

where $D_+ = \sum_i D_i$ and $T_+ = \sum_i T_i$. The score and the statistical information are accordingly:

$$\begin{aligned} S(\theta) &= \frac{d}{d\theta} L(\theta) = \frac{D_+}{\theta} - T_+ \\ J(\theta) &= -\frac{d^2}{d\theta^2} L(\theta) = \frac{D_+}{\theta^2} \end{aligned}$$

taking into account that $J(\theta)$ and therefore $I(\theta)$ are block diagonal. Therefore the maximum likelihood estimator of θ is:

$$\hat{\theta}_{ML} = \frac{D_+}{T_+}$$

Let us recall that:

$$\sqrt{n}(\hat{\theta}_{ML,n} - \theta) \xrightarrow{\mathcal{L}} N\left\{0, [I(\theta)]^{-1}\right\}$$

where

$$I(\theta) = V\left\{\frac{d}{d\theta} L(\theta) \mid \theta\right\} = \mathbb{E}[J(\theta)|\theta] = \frac{E[D_+ \mid \theta]}{\theta^2}$$

Note that:

$$E[D_i \mid \phi] = \Pr[\eta_i \leq \zeta_i \mid \phi] = E[F_\eta(\zeta_i \mid \theta) \mid \phi] = 1 - E[e^{-\theta \zeta_i} \mid \phi]$$

Therefore:

$$E[D_i \mid \theta] = 1 - E[e^{-\theta\zeta_i}|\theta] = 1 - \int e^{-\theta\zeta_i} dF_\zeta(\zeta_i)$$

In practice, $I(\theta)$ is estimated as:

$$I(\hat{\theta}_{MV,n}) = \frac{D_+}{\hat{\theta}^2_{ML,n}}$$

Let us turn now to the uncensored case. In the model with censoring, there is only one parameter, $\theta \in I\!R_+$, and the bivariate statistic (D_+, T_+) is minimal sufficient but not complete. This is an example of a curved exponential family with canonical parameter $(\theta, \ln\,\theta)$. Also, let us notice the differences in the maximum likelihood estimations:

$$\begin{aligned} D_+ \quad &\longrightarrow n > D_+ \\ L(\theta) = (\ln\,\theta)D_+ - \theta T_+ \quad &\longrightarrow n\ln\,\theta - \theta T_+ \\ \hat{\theta}^c_{ML} = \tfrac{D_+}{T_+} \quad &\longrightarrow \hat{\theta}^{nc}_{ML} = \tfrac{n}{T_+} > \tfrac{D_+}{T_+} \end{aligned}$$

In other words, the cost of overlooking censoring may be appreciated by considering the difference between the (true) Fisher information, and the numerical value of the maximum likelihood estimator:

$$\frac{\hat{\theta}^c_{ML}}{\hat{\theta}^{nc}_{ML}} = \frac{D_+}{n} \leq 1 \quad \text{and} \ = 1 \Longleftrightarrow D_+ = n$$

4.2.2 Inference in conditional models

a) The general statistical model

Let us introduce the following definitions and assumptions:

- $\theta = (\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta \subset I\!R^{k_\alpha} \times I\!R^{k_\beta}$, $\qquad k_\alpha$ and k_β finite.
- *Data*:

$$Y_i = (T_i, D_i), \quad Y = (Y_1 \ldots Y_n)$$

$$X_i = (Y_i, Z_i), \quad X = (X_1 \ldots X_n)$$

- *Definition of* κ *and* θ : $Z \perp\!\!\!\perp \theta \mid \kappa$ and $Y \perp\!\!\!\perp \kappa \mid Z, \theta$
- *Assumptions*
  variation-free parameters: $(\kappa, \theta) \in \Theta_\kappa \times \Theta_\theta$
  conditional independence: $\perp\!\!\!\perp_i Y_i \mid Z, \theta$ and $Y_i \perp\!\!\!\perp Z \mid Z_i, \theta$
  θ is the only parameter of interest

Therefore, the relevant actual loglikelihood takes the form:

$$\begin{aligned} L(\theta) &= \sum_i D_i \ln\, f_\eta(T_i \mid z_i, \theta) + \sum_i (1 - D_i) \ln\, S_\eta(T_i \mid z_i, \theta) \\ &= \sum_i D_i \ln\, \lambda_\eta(T_i \mid z_i, \theta) - \sum_i \Lambda_\eta(T_i \mid z_i, \theta) \end{aligned}$$

The score and the statistical information are equal to:

$$S(\theta) = \tfrac{d}{d\theta} L(\theta) = \sum_i \frac{D_i}{\lambda_\eta(T_i \mid z_i, \theta)} \frac{d}{d\theta} \lambda_\eta(T_i \mid z_i, \theta) - \sum_i \frac{d}{d\theta} \Lambda_\eta(T_i \mid z_i, \theta)$$

$$J(\theta) = \tfrac{-d^2}{d\theta d\theta'} L(\theta) = \textstyle\sum_i D_i [\lambda_\eta(T_i|z_i, \theta)]^{-2} \tfrac{d}{d\theta} \lambda_\eta(T_i|z_i, \theta) \tfrac{d}{d\theta'} \lambda_\eta(T_i|z_i, \theta)$$

$$- \textstyle\sum_i D_i \lambda_\eta(T_i|z_i, \theta)^{-1} \tfrac{d^2}{d\theta d\theta'} \lambda_\eta(T_i|z_i, \theta) + \sum_i \tfrac{d^2}{d\theta d\theta'} \Lambda_\eta(T_i|z_i, \theta)$$

Notice once more that the expectation of $I(\theta)$ depends both on θ and w, and thus, on the parameter of the censoring variable.

b) The proportional hazard model

When

$$\lambda_\eta(t \mid z, \theta) = g(z, \beta) \lambda_0(t \mid \alpha),$$

the log-likelihood function may be written as:

$$\begin{aligned} L(\theta) &= \sum D_i \ln\ \lambda_\eta(T_i \mid z, \theta) - \sum_i \Lambda_\eta(T_i \mid z, \beta) \\ &= \sum_i D_i \ln\ g(z_i, \beta) + \sum_i D_i \ln\ \lambda_0(T_i|\alpha) - \sum_i g(z_i, \beta) \Lambda_0(T_i|\alpha) \end{aligned}$$

and, under the log-linear specification $g(z, \beta) = \exp\left(z'\beta\right)$:

$$L(\theta) = \beta' \sum_i D_i z_i + \sum_i D_i\ \ln\ \lambda_0(T_i|\alpha) - \sum_i e^{z_i'\beta} \Lambda_0(T_i|\alpha)$$

c) The mixed proportional hazard model and its identifiability

The mixed proportional hazard (MPH) model is characterized by the following hazard function:

$$\lambda_T(t \mid z) = \lambda_0(t)\, g(z)\, \upsilon$$

where $\lambda_0(t)$ is a baseline hazard function, $g(z)$ is the function measuring the proportional effect of observable covariates z on the hazard function, and υ is an individual-specific random term representing unobserved individual

heterogeneity. The cumulative density function of υ is denoted H. This model is supposed to verify the following assumptions:

Assumption 1 : The covariate vector z is a finite-dimensional vector of dimension k $(1 \leq k \leq \infty)$. The function $g(z)$ is positive for every $z \in \mathcal{Z} \subset \mathbb{R}^k$.

Assumption 2 : The function $\lambda_0(t)$ is positive and continuous on $[0, \infty)$, except that $\lim_{t \to 0} \lambda_0(t)$ may be infinite. For every $t \geq 0$,

$$\int_0^t \lambda_0(u)du < \infty \quad \text{while} \quad \lim_{t \to \infty} \int_0^t \lambda_0(u)du = \infty$$

Assumption 3 : The distribution H of the random term υ in the inflow (i.e. when $t = 0$) satisfies $\Pr\{\upsilon \in]0, \infty)\} = 1$.

Assumption 4 : The individual value of υ is time-invariant.

Assumption 5 : In the inflow (i.e. when $t = 0$), υ is independent of z.

This model is nonparametrically identified if there is a unique set of functions λ_0, g and H that generates the observable distribution of the data, namely $F(t \mid z)$. Conditions for identification are the following (see Van den Berg [2001], for a very clear exposition):

Assumption 6 (variation in observed covariates): The set $\mathcal{Z}$ of possible values of z contains at least two values, and $g(z)$ is not constant on $\mathcal{Z}$.

Assumption 7 (normalizations): For some a priori chosen t_0 and z_0, there holds:

$$\int_0^{t_0} \lambda_0(u)\, du = 1 \quad \text{and} \quad g(z_0) = 1$$

Assumption 8 (tail of the unobserved heterogeneity distribution): $E(\upsilon) < \infty$.

Assumptions 6 and 8 can be alternatively stated:

Assumption 6b (variation in observed covariates): The vector z includes an element z^a such that the set $\mathcal{Z}^a$ of its possible values contains a non-empty open interval. For given values of the other elements of z, the value of z^a varies over this interval. Moreover, $g(z)$ as a function of z^a is differentiable and not constant on this interval.

Assumption 8b (tail of the unobserved heterogeneity distribution): The random variable υ is continuous, and the probability density function $h(\upsilon)$ of υ verifies the following property:

$$\lim_{\upsilon \to \infty} \frac{h(\upsilon)}{\upsilon^{-1-\epsilon} V(\upsilon)} = 1$$

where $\epsilon \in [0,1]$ is fixed in advance, and where $V(v)$ is a function such that:

$$\lim_{s\to\infty} \frac{V(sv)}{V(s)} = 1.$$

Identification of the MPH model has been analyzed successively by Elbers and Ridder [1982], Heckman and Singer [1984b], Ridder [1990], Melino and Sueyoshi [1990], and Kortram, Lenstra, Ridder and Van Rooij [1995].

c) The accelerated life model
When $\lambda_\eta(t \mid z,\theta) = g(z,\beta)\ \lambda_0(t\, g(z,\beta) \mid \alpha)$, the log-likelihood function, for an arbitrary family of baseline distributions, may be written as:

$$L(\theta) = \sum_i D_i[\ln\ g(z_i,\beta) + \ln\ \lambda_0(T_i\, g(z_i,\beta)|\alpha)] - \sum_i \Lambda(T_i\, g(z_i,\beta)|\alpha)$$

When the baseline distribution is exponential, namely when $\lambda_0(t_i|\alpha) = \alpha$, we obtain:

$$L(\theta) = \sum_i D_i[\ln\ g(z_i,\beta) + \ln\ \alpha] - \alpha \sum_i T_i g(z_i,\beta)$$

In the particular case where $g(z_i,\beta) = \exp\left(z_i'\beta\right)$, we obtain a proportional hazard model. More generally, this is also the case for a Weibull baseline distribution:

$$L(\theta) = \ln\ \alpha \sum_i D_i + \beta' \sum_i D_i z_i - \alpha \sum_i T_i e^{z_i'\beta}$$

4.3 Nonparametric and semiparametric models

4.3.1 Marginal models : the Kaplan-Meier estimator of the survivor factor

If we want to estimate $S_T(t)$ in presence of right-censoring, a simple idea is to adjust the hazard rates of the product form of the (discrete) empirical survivor function. With the same data as for the parametric models :

$$Y_i = (T_i, D_i)$$

$$T_i = \min(\eta_i, \zeta_i)$$

$$D_i = \mathbb{1}_{\{T_i=\eta_i\}}$$

we now evaluate:

$$\begin{aligned} T_i &\rightarrow T_{(1)} < T_{(2)} \ldots T_{(n)} \quad \text{(order statistics)} \\ D_i &\rightarrow D_1^{'}, D_2^{'}, \ldots, D_n^{'}\text{: (censoring indicators corresponding to the } T_{(i)}) \\ R(t) &= \sum_i \mathbb{1}_{\{T_{(i)} \geq t\}} \\ B(T_{(i)}) &= \sum_j D_j^{'} \mathbb{1}_{\{T_j = T_{(i)}\}} \end{aligned}$$

Thus $R(t)$ represents the number of individuals at risk at time t, *i.e.* those who are neither "dead" nor censored at time t^-, and $B(T_{(i)})$ represents the number of deaths (*i.e.* exiting without being censored) at the observed time $T_{(i)}$. A natural way of taking censoring into account is to consider that at the time $T_{(i)}$, $B(T_{(i)})$ is the realization of a binomial variable with parameter $\left(R\left(T_{(i)}\right), \lambda(T_{(i)})\right)$. Then the hazard function at (observed) time $T_{(i)}$ and the survivor functions are estimated as:

$$\begin{aligned} \hat{\lambda}(T_{(i)}) &= \frac{B(T_{(i)})}{R(T_{(i)})} \\ \hat{S}_{KM}(t) &= \prod_{\{T_{(i)} < t\}} [1 - \hat{\lambda}(T_{(i)})] \end{aligned}$$

Remarks

1. If at $T_{(i)}$ there are only censored data, we have $B(T_{(i)}) = 0$ and therefore $\hat{S}_{KM}(T_{(i)})$ is continuous at $T_{(i)}$.

2. If the largest observation is a censored one, $\hat{S}_{KM}(t)$ is strictly positive and continuous, at $T_{(n)}$:

$$\hat{S}_{KM}(t) = \hat{S}_{KM}(T_{(n)}) > 0, \qquad \forall t > T_{(n)}$$

If furthermore $T_{(n-1)}$ is not censored, $\lim_{t \to \infty} \widehat{F}_{KM}(t) > 0$, which means that F could be defective. A natural interpretation of this occurrence, in the case of a life duration, is the following: if the largest observation does not correspond to an exit (or a death), there is no empirical reason not to believe that such a life could possibly be infinite. If one is willing to avoid defective distributions, one may modify the Kaplan-Meier estimator as follows:

$$\widehat{S}_{KM}^{m}(t) = \prod_{\{T_{(i)} \leq t\}} [1 - \hat{h}(T_{(i)})] \mathbb{1}_{\{t \leq \max\{D_i T_i\}\}} = \widehat{F}_{KM}(t) \mathbb{1}_{\{t \leq \max\{D_i T_i\}\}}$$

where $\max\{D_i T_i\}$ represents the largest uncensored duration.

3. If there are no ties at $T_{(i)}$, then:

$$\begin{aligned} B(T_{(i)}) &= D_i', \quad R(t_{(i)}) = n - i + 1, \\ \widehat{S}_{KM}(t) &= \prod_{\{T_{(i)} \leq t\}} (1 - \frac{D_i}{n - i + 1}) \end{aligned}$$

In many data sets, ties are observed, as a matter of fact. They call for two remarks: *(i)* even if F_η and F_ζ are continuous, $\Pr(\eta = \zeta) > 0$ is possible when η is not independent of ζ (see, for instance, Marshall and Olkin[1967]); *(ii)* the rounding problem : although theoretical models assume the time is continuous, actual measurements are discrete in nature. We have just seen that the Kaplan-Meier estimator accommodates for ties. When the rounding problem is too severe because spells are actually observed through intervals, truncated survivor functions may be used for an explicit modelling.

4. If, at the largest observation, some censored *and* uncensored data are tied, the estimated distribution, $\hat{S}_{KM}(T_{(i)})$, is again defective and discontinuous at $T_{(n)}$, with:

$$\hat{S}_{KM}(T_{(n)}) > \hat{S}_{KM}(T_{(\infty-)}) > 0$$

4.3.2 Conditional models : the semiparametric proportional hazard model (the Cox model)

Remember that in $\theta = (\alpha, \beta)$, α is a sufficient parameter for the baseline distribution, whereas β is introduced for describing the action of the exogenous variables. The semiparametric version of the proportional hazard model takes the form:

$$\lambda_T(t \mid z, \theta) = \alpha(t) \exp\left(z'\beta\right)$$

where $\alpha(t) = \lambda_0(t|z, \theta)$, which is the baseline hazard function, is now a functional parameter. Thus the parameter space takes the following form:

$$\begin{aligned} \theta &= (\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta \\ \Theta_\alpha &= \{\alpha : I\!R_+ \to I\!R_+ \mid \alpha \text{ is continuous and } \int_0^\infty \alpha(t)dt = \infty\} \\ \Theta_\beta &\subset I\!R^k \end{aligned}$$

The functional parameter α is often a nuisance parameter, whereas the Euclidean parameter β is the parameter of interest. It is therefore important to try to separate inferences on α and β. A natural idea is to construct a

statistic $W = f(Y)$ such that the likelihood function $L_{Y|Z}(\alpha, \beta)$ factorizes as follows:

$$L_{Y|Z}(\alpha, \beta) = L_{W|Z}(\beta) \times L_{Y|W,Z}(\alpha, \beta)$$

In such a case, the inference on β would be made simpler by considering only the partial likelihood $L_{W|Z}(\beta)$ instead of $L_{Y|Z}(\alpha, \beta)$. A heuristic argument in favour of this simplification is that the information on β contained in $L_{Y|W,Z}(\alpha, \beta)$ is likely to be "eaten up" by the functional parameter α. This simplified estimator may now be build as follows. Similarly to the Kaplan-Meier estimator, let us reorder the sample according to the observed durations:

$$\begin{aligned} T_i &\longrightarrow T_{(1)} < T_{(2)} < \ldots < T_{(n)} \\ D_i &\longrightarrow D_1^{'}, D_2^{'}, ..., D_n^{'} \end{aligned}$$

and let us also define:

$$\begin{aligned} R(t) &= \sum_{1 \leq i \leq n} \mathbb{1}_{\{T_{(i)} \geq t\}} \\ \mathcal{R}(t) &= \{k | T(k) \geq t\} = \{i | T_i \geq t\} \end{aligned}$$

Thus $R(t)$ represents the number of individuals at risk at time t and $\mathcal{R}(t)$ represents the set of such individuals. Notation will be usefully simplified as follows:

$$R_{(i)} = R(T_{(i)}), \qquad \mathcal{R}_{(i)} = \mathcal{R}(T_{(i)})$$

Let us now represent the sample $(T_1 \ldots T_n)$ by its order statistics $(T_{(1)} \ldots T_{(n)})$ and its rank statistics $(R_1 \ldots R_n)$ where R_i is the rank of the i-th observation in the vector of order statistics. Giving the rank statistics, which plays the role of W in the previous expression, we may write the likelihood function of the rank statistics as follows:

$$L(\beta) = \prod_{1 \leq i \leq n} \left[\frac{e^{z_i' \beta}}{\sum_{k \in \mathcal{R}(i)} e^{z_k' \beta}} \right]^{D_i} = \prod_{1 \leq i \leq D_+} \left[\frac{e^{z_i' \beta}}{\sum_{k \in \mathcal{R}(i)} e^{z_k' \beta}} \right]^{D_i}$$

where $D_+ = \sum_i D_i$. The (partial) likelihood estimator of β is then defined as

$$\hat{\beta} = \arg\max_{\beta} L(\beta)$$

This estimator is consistent and its asymptotic properties have been studied e.g. by Tsiatis [1981] and by Andersen *et alii* [1993].

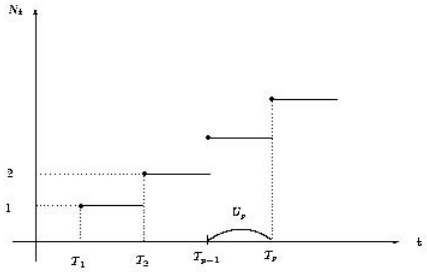

Figure 2: **A realization of a univariate counting process**

5 Counting processes and point processes

Point processes provide the framework for modelling trajectories with more than one transition and more than two states (such trajectories are sometimes called duration models with multiple spells and multiple states). Formally a point process is a continuous time process with a finite state space and right continuous with left limit (*cadlag*) trajectories. A point process is conveniently represented by means of a multivariate counting process that counts, as time increases, the number of possible transitions. Consequently, we will first present counting processes.

5.1 Definitions

Let us consider a (finite or infinite) sequence $(T_p)_{p\geq 1}$ of increasing random durations ($0 < T_1 < T_2 < ...$). This sequence characterizes a univariate counting process:

$$N_t = \sum_{p\geq 1} \mathbb{1}(T_p \leq t)$$

The trajectories of N_t are right continuous, and such that $N_0 = 0$ and N_t only increases by jumps of size 1. A typical realization is shown in Figure 2.

A duration model defines a process with a single jump ($N_t = \mathbb{1}(T \geq t)$). From the definition of N_t, we can deduce easily the definition of the date T_p of the j-th jump of the process:

$$T_p = \inf\{t | N_t = p\}, \quad p \geq 1$$

The distribution of N_t may be characterized by the distribution of the sequence $(T_p)_{p\geq 1}$. Equivalently, that sequence may be replaced by the sequence of positive numbers:

$$U_p = T_p - T_{p-1} \quad (T_0 = 0)$$

The random variable is now the duration between the $(p-1)$-th and the p-th jumps. If the random variables $(U_p)_{p\geq 1}$ are *i.i.d.*, the process is called a *renewal process*. The information denoted by $\mathcal{F}_t^N$ and carried by N_t, observed from 0 to t (included), is equivalent to the knowledge of $T_1, ..., T_p$ $(T_p \leq t < T_{p+1})$ and the event $T_{p+1} > t$. Equivalently this information may be described by the random variables $U_1, ..., U_p$ and by the event $U_{p+1} > t - \sum_{q=1}^{p} U_q$.

A *multivariate counting process* is a vector $N_t = (N_t^1, ..., N_t^J)$ of counting processes. This vectorial process is characterized by J sequences $(T_p^j)_{p\geq 1}$ $(j = 1, ..., J)$ of increasing durations and by:

$$N_t^j = \sum_{p\geq 1} \mathbb{1}(T_p^j \leq t)$$

The information content of the observation of this multivariate counting process up to time t is described by the family of random variables T_p^j such that $T_{p_j}^j \leq t$, and by the J events $T_{p_j+1}^j > t$.

A multivariate counting process may also be represented by a unique sequence $(T_r)_{r\geq 1}$ of the jump times of any element of the vector N_t, and by e_r $(r \geq 1)$ which is a discrete-time process valued in $(1, ..., J)$. In this sequence $(T_r, e_r)_{r\geq 1}$, e_r indicates the component j that jumps at date T_r. Note that the sequence (T_r) has the property:

$$\bar{N}_t = \sum_{j=1}^{J} N_t^j = \sum_{r\geq 1} \mathbb{1}(T_r \leq t)$$

The distribution of N_t may then be described by the sequence of conditional distributions:

$$(T_r, e_r) \mid (T_s, e_s)_{s=1,...,r-1}$$

Consider for example a bivariate duration (T^1, T^2), where $\Pr(T^1 = T^2) = 0$. This pair defines two single jump counting processes:

$$N_t^1 = \mathbb{1}(T^1 \leq t) \quad \text{and} \quad N_t^2 = \mathbb{1}(T^2 \leq t)$$

Then the $(T_r)_r$ sequence becomes:

$$T_1 = min(T^1, T^2), \quad T_2 = max(T^1, T^2)$$

and

$$\begin{aligned} e_1 &= \mathbb{1}(T^1 < T^2) + 2\,\mathbb{1}(T^2 \leq T^1) \\ e_2 &= 3 - e_1 \end{aligned}$$

A point process is a continuous-time process valued in a finite (or more generally discrete) state space $\{1, ..., K\}$. Such a process X_t may represent, for example, the labour market situation of an individual at time t. In such a case, the set $\{1, ..., K\}$ describes the different possible labour market states (full-time employed, part-time employed, unemployed, retired,...) and X_t is characterized by the dates of the transitions between two different states. Indeed, a point process defines a multivariate counting process. Consequently, we denote by $j = (k, k')$ the pair of states such that a transition from k to k' is possible and $\{1, ..., J\}$ is the set of all these ordered pairs. Then $(T_p^j)_{p\geq 1}$ is the sequence of jump times from k to k' if $j = (k, k')$ and

$$N_t^j = \sum_{p\geq 1} \mathbb{1}(T_p^j \leq t)$$

This multivariate counting process satisfies the following constraint by construction: after a jump of the component N_t^j, $j = (k, k')$, the next process which may jump is necessarily an element of the subfamily $(N_t^\ell)_\ell$ where $\ell = (k', k")$ and $k^" \neq k'$.

5.2 Stochastic intensity, compensator and likelihood of a counting process

The stochastic intensity of a univariate counting process is defined as follows:

$$h_N(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \Pr(N_{t+\Delta t} - N_t = 1 \mid \mathcal{F}_{t-}^N)$$

If for instance $N_t = \mathbb{1}(T \leq t)$, this definition implies, where T is a continuous variable, that $h(t) = \lambda(t)$, which is the hazard function of T if $T > t$ and $h(t) = 0$ after the jump T. Equivalently:

$$h_N(t) = \lambda_T(t)(1 - N_{t-}),$$

where $N_{t-} = \mathbb{1}(T < t)$.
If N_t is a general univariate counting process $N_t = \sum_{p\geq 1} \mathbb{1}(T_p \geq t)$, the stochastic intensity is obtained by the following rule:

- If $t > \max_p(T_p)$ then $h(t) = 0$
- If t verifies $T_{p-1} < t \leq T_p$ (where $p = N_t + 1$) then

  $$h_N(t) = \lambda_p(t \mid T_1, ..., T_{p-1})$$

  where λ_p is the hazard function of the duration T_p conditional on $T_1, ..., T_{p-1}$.

If the model is specified in terms of $U_p = T_p - T_{p-1}$, we have

$$h_N(t) = \lambda_U(t - T_{p-1} \mid U_1, ..., U_{p-1})$$

where λ_p^U is the hazard function of U_p given $U_1, ...U_{p-1}$. This definition is easily extended to multivariate counting processes. The stochastic intensity is then multivariate and for each $j \in \{1, ..., J\}$:

$$h_N^j(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \Pr(N_{t+\Delta t}^j - N_t^j = 1 \mid \mathcal{F}_{t-}^N)$$

where $\mathcal{F}_{t-}^N$ represents the information carried by all the coordinates of the process observed before t.
If $N_t^j = \sum_{p \geq 1} \mathbb{1}(T_p^j \leq t)$, $h_N^j(t)$ is null if $t > \max_p(T_p^j)$. For each coordinate ℓ, we can choose $p_\ell = N_t^\ell + 1$ such that

$$T_{p_\ell - 1}^\ell \leq t < T_{p_\ell}^\ell$$

(where $T_{p_\ell}^\ell = +\infty$ if N_t^ℓ never jumps after $T_{p_\ell - 1}^\ell$) and $\lambda_T^i(t)$ is equal to the hazard function of $T_{p_j}^j$ at the point t, given all the T_q^ℓ, $\ell \neq j$ and $q < p_\ell$, and the family of events $T_{p_\ell}^\ell \geq t$. Let us take as an example the bivariate counting process $N_t^1 = \mathbb{1}(T^1 \leq t)$, $N_t^2 = \mathbb{1}(T^2 \leq t)$. The stochastic intensity $h_N^1(t)$ is equal to the hazard function of T^1 conditional on $T^2 = t_2$ if $T^2 < t$ or conditional on $T^2 \geq t$ if $T^2 \geq t$. The compensator of univariate counting process N_t with stochastic intensity $h_N(t)$ is defined by

$$H_N(t) = \int_0^t h(s) d_s$$

For a duration model $N_t = \mathbb{1}(T \leq t)$, $H_N(t)$ is equal to the integrated hazard $\Lambda(t)$ if $T > t$ and equal to $\Lambda(T)$ if $T \leq t$.
For a multivariate counting process N_t^j, we define a vector of compensators by:

$$H_N^j(t) = \int_0^t h_N^j(s) ds.$$

From now on, we simplify the notation $H_N^j(t)$ into H_t^j similarly to N_t instead of $N(t)$. The compensators are positive and non-decreasing predictable processes satisfying $H_0 = 0$. The main property of the compensator is that the difference:

$$M_t = N_t - H_t$$

is a zero mean $\mathcal{F}_t^N$- martingale (*i.e.* $E(M_t \mid \mathcal{F}_s^N) = M_s$). The decomposition $N_t = H_t + M_t$ is called the Doob-Meyer decomposition of the process N_t. The same decomposition may be constructed for a multivariate counting

process. In that case, M_t^j is a martingale with respect to the information sets generated by the whole process $(N_t^1, ..., N_t^J)$.
The stochastic intensity and the compensator both determine an elegant expression of the likelihood of a counting process. Consider first a univariate process $N_t = \sum_{p\geq 1} \mathbb{1}(T_p \leq t)$. If the process is observed between 0 and t such that $T_{p-1} < t < T_p$, the likelihood of this observation is:

$$\ell(t) = \left\{ \prod_{q=1}^{p-1} f_q(T_q \mid T_1, ..., T_{q-1}) \right\} \times S_p(t \mid T_1, ..., T_{p-1})$$

where f_q and S_q are respectively the density and the survivor functions of T_q given $T_1, ..., T_{q-1}$. One can easily check that:

$$\ell(t) = \prod_{T_q \leq t} h(T_q) e^{-H_t}$$

or

$$\ln \ell(t) = \int_0^t \ln h(s) dN_s - H_t$$

In this expression, we use the stochastic integral notation:

$$\int_0^t g(s) dN_s = \sum_{T_p \leq t} g(T_p)$$

The stochastic intensity notation can be generalized to multivariate processes for which the likelihood corresponding to the observation of all the coordinates of the process up to time t is equal to:

$$\ln l(t) = \sum_{j=1}^{J} \left\{ \int_0^t \ln h^j(s) dN_s^j - H_t^j \right\}$$

This way of writing the likelihood function is the basis for Cox's estimation and martingale estimations, to be presented in the last section of this chapter.

6 Poisson, Markov and semi-Markov Processes

In this section, we give first the example of a well-known single counting process, namely the Poisson process. Then we examine point processes displaying Markovian or semi-Markovian properties.

6.1 Poisson processes

We consider the familiar Poisson process as an example of a univariate counting process. Let M be a positive measure on $\mathbb{R}^+$ with density m with respect to the Lebesgue measure, *i.e.* , $M([a,b]) = \int_a^b m(x)\ d(x)$.
A stochastic process N_t is a *Poisson process* associated with the measure M if its distribution satisfies the following requirements:

i) $N_0 = 0$,

ii) N_t is a process with independent increments: $\forall t_1, \cdots, t_n$, the random variables $(N_{t_i} - N_{t_{i-1}})_{i=1,\cdots,n}$ are independent random variables,

iii) the distribution of $(N_t - N_s)$ is a Poisson distribution for any $s < t$, which means that:

$$\Pr(N_t - N_s = k) = \frac{M([s,t])^k}{k!} e^{-M([s,t])}$$

These three properties imply that a Poisson process is a counting process with unit jumps. If $m(x)$ is equal to some positive constant λ, then the process is said to be homogeneous and we may verify that sojourn times $U_p = T_n - T_{n-1}$ are *i.i.d.* random variables with an exponential distribution with parameter $\lambda > 0$. The *homogeneous Poisson process* is then the renewal process characterized by the exponential distribution.
The compensator and the intensity of a Poisson process, with respect to its canonical filtration, are equal to $H = M([0,t])$ and to $m(t)$, respectively. This result follows from the equalities :

$$\begin{aligned} h(t) &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \Pr(N_{t+\Delta t} - N_t = 1 \mid \mathcal{F}_{t^-}^N) \\ &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \Pr(N_{t+\Delta t} - N_t = 1 \mid N_t) \\ &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} [M(t, t+\Delta t]) e^{-M([t,t+\Delta t])}] \\ &= m(t) \end{aligned}$$

In particular, if the process is homogeneous, $h(t)$ is constant. The likelihood $\ell(t)$ relative to the observation of the process N_t between 0 and t is derived from the intensity and the compensator, *i.e.*

$$\begin{aligned} \ln \ell(t) &= \int_0^t [\ln m(s)]\ dN_s - M([0,t]) \\ &= \sum_{\tau_n \leq t} \ln m(\tau_n) - M([0,t])\,. \end{aligned}$$

If N_t is an homogeneous Poisson process with parameter λ, its likelihood satisfies:

$$\ln \ell(t) = N_t \ln \lambda - \lambda t \,.$$

6.2 Markov processes

a) Definitions

We consider a point process $X = (X_t)_{t\in\mathbb{R}^+}$ valued in the finite state space $E = \{1, \cdots, K\}$. The distribution of X_t is totally defined by a projective system:

$$\Pr(X_{t_1} = j_1, \cdots, X_{t_p} = j_p)$$

for any finite subset $(t_1, \cdots, t_p)$ of $\mathbb{R}^+$ satisfying $t_1 < t_2 < \cdots < t_p$. From these probabilities, one can compute:

$$\Pr(X_{t_p} = j_p \mid X_{t_1} = j_1, \cdots, X_{t_{p-1}} = j_{p-1})$$

and the process X_t is a *Markov process* if:

$$\Pr(X_{t_p} = j_p \mid X_{t_1} = j_1, \cdots, X_{t_{p-1}} = j_{p-1}) = \Pr(X_{t_p} = j_p \mid X_{t_{p-1}} = j_{p-1})$$

It follows that a Markov process is characterized by the distribution of the initial condition, *i.e.* by the distribution of X_0, and by the *transition probabilities*:

$$p_{jk}(s, s+t) = \Pr(X_{s+t} = k \mid X_s = j)$$

defined for any s and $t \in \mathbb{R}^+$, and for any j and $k \in E$. The Markov process is said to be *time–homogeneous* if:

$$p_{jk}(s, s+t) = p_{jk}(0, t), \quad \forall (s,t) \in \mathbb{R}^+ \times \mathbb{R}^+, \forall (j,k) \in E^2,$$

i.e. if the transition probability does not depend on the origin of the time set, but only on the difference between the two dates s and $(s+t)$. For a time–homogeneous Markov process, we denote the transition probability $p_{jk}(0,t)$ by $p_{jk}(t)$ and the matrix with elements $p_{jk}(t)$ by $P(t)$. So, $P(t)$ is a $K \times K$ matrix of non–negative numbers such that the sum of each row is equal to one, *i.e.*

$$\sum_{k=1}^{K} p_{jk}(t) = 1$$

Moreover, decomposing the trajectory on $[0,t]$ into two sub-trajectories on $[0,s]$ and $[s,t]$, we obtain the following properties of the matrices $P(t)$:

$$p_{jk}(t) = \textstyle\sum_{\ell=1}^{K} p_{j\ell}(s)\; p_{\ell k}(t-s)\,, \quad \forall\; 0 \le s \le t,\; \forall\; (j,k) \in E \times E$$

or equivalently:

$$P(t) \;=\; P(s)\; P(t-s)\,,\; 0 \le s \le t\,.$$

We will now restrict our attention to processes satisfying some regularity conditions.

Definition 6.1

A time–homogeneous Markov process X_t is said to be *standard* if:

i) $\forall j \in E,\ \lim_{t\downarrow 0} p_{jj}(t) = 1$, and then, $\quad \forall k \neq j,\ \lim_{t\downarrow 0} p_{jk}(t) = 0,$

ii) $\exists q_{jk} \in \mathbb{R}^+, \forall (j,k) \in (E \times E)$, with $k \neq j$,
$q_{jk} = \lim_{t\downarrow 0} \frac{1}{t} p_{jk}(t) = \frac{d}{dt} p_{jk}(t)\ |_{t=0}$,
$q_{jj} = -\lim_{t\downarrow 0} \frac{1}{t}(1 - p_{jj}(t)) = -\sum_{k\neq j} q_{jk}$ ■

As a direct consequence, quantities q_{jk} satisfy the following properties:

$$i) \qquad \sum_{k=1}^{K} q_{jk} = 0, \qquad j \in E,$$

$$ii) \qquad q_{jk} \geq 0, \quad k \neq j, \quad \text{and} \quad q_{jj} \leq 0, \qquad j \in E\,.$$

If $j \neq k$, q_{jk} is called the *intensity of transition* from state j to state k. The matrix Q is called the *intensity matrix* or the *generator* of the process X_t. Writing P_{jj} in the form $P_{jj}(t) = 1 - \sum_{k\neq j} P_{kj}(t)$, the previous definition implies that $Q = \dfrac{d}{dt} P(t)|_{t=0}$

Theorem 6.1

The transition matrix $P(t)$ of the time–homogeneous standard Markov process X_t satisfies the forward matrix equation

$$\frac{d}{dt} P(t) = P(t) \cdot Q$$

and the backward matrix equation

$$\frac{d}{dt} P(t) = Q \cdot P(t)\,.$$

Proof: see Doob [1953], p. 240–241, or Bhattacharya and Waymire [1990], p. 263–267.
These two equations are known as the *Kolmogorov forward and backward differential equations*, respectively. In general, these equations do not have a unique solution; however, if X_t is regular, the solution, subject to the border condition $P(0) = I$, is unique and has the exponential form given in the following theorem (where I is the identity matrix).

Theorem 6.2

If the time–homogeneous Markov process X_t with generator Q is regular, then the matrix

$$P(t) = \exp(Qt) = \sum_{n=0}^{\infty} t^n Q^n / n! \tag{1}$$

exists for any t, and is the unique solution to the Kolmogorov differential equations subject to the border condition $P(0) = I$.

Proof: See Doob[1953], p. 240–241, or Bhattacharya and Waymire [1990], p. 267–275.

b) Distributions related to a time–homogeneous standard Markov process

Since the state space E is finite, the Markov process X_t moves by jumping from one state to another. Let $0 = T_0 < T_1 < T_2 < \cdots$, be the times of these transitions. As the sample paths of the process X_t are right–continuous step functions, we can define $Y_n = X_{T_n}$ as the state entered at T_n. Moreover, we set:

$$U_n = T_n - T_{n-1}\,,\ n \in \mathbb{N}\,,\ \text{and } U_0 = 0$$

The random variable U_n represents the sojourn duration of the process in state $Y_{n-1} = X_{T_{n-1}}$ entered at time T_{n-1}. A Markov point process X_t can be represented by a multivariate counting process characterized by the sequence $(T_n, e_n)_{n\geq 0}$. In this representation, e_n is the transition at time T_n, *i.e.* :

$$e_n = (Y_{n-1}, Y_n) \text{ with } Y_{n-1} \neq Y_n.$$

Thus e_n takes its value in a finite set with $K(K-1)$ elements. Yet, the representation of X_t as a point process is easier to formalize. So, we are interested in the distribution of the sequences $(T_n, Y_n)_{n\geq 0}$ or $(U_n, Y_n)_{n\geq 0}$, rather than of the sequence $(T_n, E_n)_{n\geq 0}$.
For that purpose, we firstly set $\lambda_j = -q_{jj}$ for any $j \in E$, and we define quantities ρ_{jk} as follows :

- If $\lambda_j \neq 0$, $j \in E$, $\qquad \rho_{jj} = 0$ and $\rho_{jk} = q_{jk}/\lambda_j\,,\ k \neq j$
- If $\lambda_j = 0\,,\ j \in E \qquad \rho_{jj} = 1$ and $\rho_{jk} = 0\,,\ k \neq j$

Theorem 6.3

If X_t is a time–homogeneous standard Markov process, then

i) $(U_n, Y_n)_{n\geq 0}$ is a Markov sequence and (U_n, Y_n) is independent of U_{n-1} given Y_{n-1}. Moreover U_n and Y_n are conditionally independent given Y_{n-1}.

ii) U_n given $Y_{n-1} = j$ has an exponential distribution with parameter λ_j if $\lambda_j \neq 0$. If $\lambda_j = 0$, the state j is absorbing and $U_n = \infty$ with probability 1.

iii) $Y = (Y_n)_{n \geq 0}$ is a Markov chain with transition matrix:

$$\Pr(Y_n = k \mid Y_{n-1} = j) = \rho_{jk}, \ (j,k) \in E \times E$$

(see Bhattacharya and Waymire, [1990], p. 275-279).

Theorem 6.4

If X_t is irreducible ($\forall j, \forall k, \exists m$ such that $p_{jk}(m) > 0$) and recurrent ($\Pr(\inf\{m \mid Y_{n+m} = j\} < \infty \mid Y_n = j) = 1$) then:

(i) the limits of transition probabilities $p_{jk}(t)$ exist and are independent of the initial state, *i.e.* .

$$\lim_{t \uparrow \infty} p_{jk}(t) = \Pi_k$$

(ii) either $\Pi = (\Pi_1, \cdots, \Pi_K) = (0, \cdots, 0)$, in which case all states are said to be null recurrent, or $\sum_{k=1}^{K} \Pi_k = 1$, in which case all states are said to be non–null recurrent (or positive recurrent if $\Pi_k > 0, \forall k \in E$).

Proof: see Cox and Miller [1966], p. 106–117.

The *limiting distribution* Π is also invariant or stationary, because:

$$\Pi = \Pi \cdot P(t) \, , \, \forall t \in \mathbb{R}^+$$

In the case of an irreducible, recurrent non–null Markov process with generator Q, calculation of the vector Π is made easier by noting that Π is the unique invariant distribution probability satisfying the linear equation:

$$\Pi \cdot Q = 0$$

Moreover, if the embedded Markov chain Y is also irreducible and recurrent non–null, Y has a limit distribution v satisfying:

$$v = v \cdot R$$

where R is the transition matrix of the embedded Markov chain. The relationship between the two limit distributions Π and v is:

$$\Pi_j = \left[\frac{v_j}{\lambda_j} \right] \left[\sum_{k=1}^{K} \frac{v_k}{\lambda_k} \right], \qquad j \in E$$

or equivalently:

$$v_j = \frac{\Pi_j \lambda_j}{\left[\sum_{k=1}^{K} \Pi_k \lambda_k\right]}, \qquad j \in E$$

Generally, v and Π are different. The last equation has a very interesting interpretation: since v_j is the long–run frequency of visits of the chain Y to state j, and since $(1/\lambda_j)$ is the mean duration of a sojourn of the process X_t in state j, then Π_j, which is the long–run proportion of occupation of state j for the process x_t, is calculated as the long–run global duration of sojourn in state j for the process X_t (calculated as the product of v_j and $1/\lambda_j$), divided by the sum of the long–run global durations of sojourn in the different states.

c) Statistical inference for time–homogeneous Markov models

Now we shall discuss the problem of estimating the generator Q of a time–homogeneous Markov process X_t from the observation of N independent sample paths over a fixed time interval $[0, T]$. Firstly, we consider the non-parametric case of N *i.i.d.* realizations of X_t over $[0, T]$. Here the likelihood function $L_Q^{(N)}$ is given by

$$L_Q^{(N)} = \prod_{i=1}^{N} \left\{ \Pr(x_0^{(i)} = Y_0^{(i)}) \times e^{-(T-\tau_{n_i}(i))\lambda_{y_{n_i}}(i)} \right.$$
$$\left. \times \prod_{j=0}^{n_i-1} q_{Y_j^{(i)}, Y_{j+1}^{(i)}} \; e^{-u_{j+1}^{(i)} \lambda_{Y_j^{(i)}}} \right\}$$

where n_i is the number of transitions observed for the i–th sample path over $[0, T]$, $0 < \tau_1^{(i)} < \tau_2^{(i)} < \cdots < \tau_{n_i}^{(i)}$ being the ordered sequence of transition times for this sample, and $\{(u_0^{(i)}, Y_0^{(i)}), (u_1^{(i)}, Y_1^{(i)}), \cdots, (u_{n_i}^{(i)}, Y_{n_i}^{(i)})\}$ being the sequence of successive sojourn durations and visited states for the i–th sample path, with the conventions: $u_0^{(i)} = 0$ and $u_{n_i+1}^{(i)} = T - \tau_{n_i}^{(i)}$.
If we denote $\mathcal{N}_{T,N}(j, k)$ the total number of transitions from state j to state k observed over the N realizations and $D_{T,N}(j)$ the total length of time that state j is occupied during these N sample paths, then it is easy to show that the maximum likelihood estimator for $q_{j,k}$ is given by:

$$\widehat{q}_{j,k}(T, N) = \frac{\mathcal{N}_{T,N}(j, k)}{D_{T,N}(j)}$$

if $j \neq k$ and $D_{T,N}(j) \neq 0$. If $D_{T,N}(j) = 0$, the MLE of $q_{j,k}$ does not exist and we adopt the convention that :

$$\widehat{q}_{j,k}(T, N) = 0 \quad \text{if} \quad j \neq k \quad \text{and} \quad D_{T,N}(j) = 0\,.$$

Asymptotic properties of the MLE estimates $\widehat{q}_{j,k}(T, N)$ when $T \to \infty$ and N is fixed (typically, $N = 1$), or when $N \to \infty$ and T is fixed, are given by the following theorems (see Basawa and Prakasa Rao [1980], p. 195–197).

Theorem 6.5

If there is a positive probability that the j–th state will be occupied at some date $t \geq 0$, then

$$\operatorname*{plim}_{T\uparrow\infty} \widehat{q}_{j,k}(T,N) = q_{j,k} \qquad a.s.$$

and if each state has a positive probability of being occupied, then the random variables

$$\left\{ N^{1/2}(\widehat{q}_{j,k}(T,N) - q_{j,k}) \right\}_{k\neq j}$$

are asymptotically normal and independent with zero mean and variance

$$\frac{q_{j,k}}{\int_0^T \Pr[X_t = j]\ dt}$$

Moreover, $\widehat{q}_{j,k}(T,N)$ is asymptotically efficient when N tends to infinity. ■

Theorem 6.6

If the time–homogeneous standard Markov process X_t is regular and recurrent positive, then

$$\operatorname*{plim}_{T\uparrow\infty} \widehat{q}_{j,k}(T,1) = q_{j,k} \qquad a.s.$$

and the random variables

$$\{T^{1/2}(\widehat{q}_{j,k}(T,1) - q_{j,k}\}_{j,k=1,\cdots,K,\ j\neq k}$$

are asymptotically normal and independent with zero mean and variance $q_{j,k}\rho \ / \ \overline{Q}(j,j)$ where ρ is the product of the non–zero eigenvalues of Q and $\overline{Q}(j,j)$ is the (j,j)–th cofactor of Q. ■

In the last case (where $N = 1, T \uparrow \infty$), it is shown that

$$\int_0^T \Pr[x_t = j]dt = \overline{Q}(j,j)T\rho^{-1} + o(T) \quad \text{for } T \uparrow \infty$$

Since

$$\frac{1}{T}E\left\{ \frac{\partial Log L_Q^{(1)}}{\partial q_{j,k}} \right\}^2 = \frac{\int_0^T \Pr[x_t = j]\ dt}{T q_{j,k}}$$

it follows from the previous theorem that

$$T^{1/2}\left\{\widehat{q}_{j,k}(T,1) - q_{j,k}\right\} \xrightarrow{d} N\left[0, q_{j,k}\rho/\overline{Q}(j,j)\right]$$

and so $\widehat{q}_{j,k}(T,1)$ is asymptotically efficient for $T \uparrow \infty$.

Now let us suppose that transition intensities are functions of a set $\theta = (\theta_1, \cdots, \theta_p)$ of unknown parameters, *i.e.* they have the form $q_{j,k}(\theta)$. The problem is then to obtain a MLE of θ from N independent observations of

the process $x_t(\theta)$ over the period $[0,T]$. In this case, the likelihood function is:

$$\begin{aligned} L_Q(N,\theta) &= \prod_{i=1}^{N} \left\{ \Pr\left(x_0(i,\theta) = Y_0^{(i)}\right) \times e^{-(T-\tau_{n_i}^{(i)})\,\lambda^{(\theta)}_{Y_{n_i}^{(i)}}} \right. \\ &\qquad \left. \times \prod_{j=0}^{n_i-1} q^{(\theta)}_{Y_j^{(i)},Y_{j+1}^{(i)}}\, e^{-u_{j+1}^{(i)}\lambda^{(\theta)}_{Y_j^{(i)}}} \right\} \\ &= \left\{ \prod_{i=1}^{N} \Pr(x_0(i,\theta) = Y_0^{(i)}) \right\} \times \left\{ \prod_{j,k=1;j\neq k}^{K} (q_{j,k}^{(\theta)})^{\mathcal{N}_{T,N}(j,k)} \right\} \\ &\qquad \times \left\{ \prod_{j=1}^{K} e^{-\lambda_j^{(\theta)} D_{T,N}(j)} \right\} \end{aligned}$$

where:

- $D_{T,N}(j) = \sum_{i=1}^{N} \sum_{\ell=0}^{n_i} u_{\ell+1}^{(i)} \mathbb{1}(Y_\ell^{(i)} = j)$ is the total sojourn duration in state j, which is observed over the N sample paths (with the convention $u_{n_i+1}^{(i)} = T - \tau_{n_i}^{(i)}$),
- $\mathcal{N}_{T,N}(j,k) = \sum_{i=1}^{N} \sum_{\ell=0}^{n_i-1} \mathbb{1}\left(Y_\ell^{(i)} = j,\, Y_{\ell+1}^{(i)} = k\right)$ is the total number of transitions from j to k, observed over the N sample paths.

With the assumption that the initial state $Y_0^{(i)}$ does not depend on θ, the ML equations for estimating θ are :

$$\frac{\partial LogL_Q(N,\theta)}{\partial \theta_m} = \prod_{\substack{j,k=1 \\ j\neq k}}^{K} \left[\frac{\mathcal{N}_{T,N}(j,k) - D_{T,N}(j) q_{jk}^{(\theta)}}{q_{jk}^{(\theta)}} \right] \frac{dq_{jk}^{(\theta)}}{d\theta_m} = 0,\ m = 1, \cdots, p$$

In the case where $N = 1$, Billingsley [1961], p. 46, has shown that these equations yield a consistent solution $\widehat{\theta} = (\widehat{\theta}_1, \cdots, \widehat{\theta}_p)$ such that

$$\mathcal{N}_T^{1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, i(\theta)^{-1}) \text{ as } T \uparrow \infty$$

where $\mathcal{N}_T$ is the total number of transitions during the interval $[0,T]$ and

$$\begin{aligned} i(\theta) &= -\frac{1}{\mathcal{N}_T} \left[E\left(\frac{\partial^2 LogL_Q(\theta)}{\partial\theta_m\, \partial\theta'_m} \right) \right]_{m=1,\cdots,p} \\ &= \left[\prod_{\substack{j,k=1 \\ j\neq k}}^{K} \frac{v_j(\theta)}{\lambda_j(\theta)\, q_{jk}^{(\theta)}} \left(\frac{dq_{jk}^{(\theta)}}{d\theta_m} \right) \left(\frac{dq_{jk}^{(\theta)}}{d\theta_{m'}} \right) \right]_{m,\ m'=1,\cdots,p} \end{aligned}$$

$v(\theta) = [v_j(\theta)]_{j=1,\cdots,K}$ being the limit distribution of the embedded Markov chain associated to the process X_t.

6.3 Semi–Markov processes

Semi–Markov processes can be viewed as direct extensions of Markov processes. Using notations and concepts introduced for the characterization of a Markov process, we will say that a stochastic process $\{X_t\}_{t\geq 0}$ taking its value in the discrete state space $E = \{1, \cdots, K\}$ is semi–markovian if the sequence $\{Y_n\}_{n\geq 0}$ of states visited remains a Markov chain, but time u_n spent in the state Y_{n-1} need not be exponentially distributed and may depend on the next state entered, namely Y_n.

Definition 6.2

If $(Y_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ denote respectively the sequences of visited states and sojourn durations of a continuous–time process $\{X_t\}_{t\geq 0}$ with a discrete state space $E = \{1, \cdots, K\}$, then $\{X_t\}_{t\geq 0}$ is a *semi–Markov process* if:

$$\Pr\left\{Y_\ell = j,\, u_\ell \leq t \mid (Y_n)_0^{\ell-1},\, (u_n)_0^{\ell-1}\right\}$$

$$= \Pr\{Y_\ell = j, u_\ell \leq t \mid Y_{\ell-1}\} \qquad \ell \in \mathbb{N},\, j \in E,\, t \in \mathbb{R}^+$$

with the convention $u_0 = 0$. Moreover, a semi–Markov process $\{X_t\}_{t\geq 0}$ is said to be *time–homogeneous* if transition probabilities

$$\Pr\{Y_\ell = j, u_\ell \leq t \mid Y_{\ell-1} = i\} = \mathcal{P}(i, j, t)\,,\ (i, j) \in E \times E$$

do not depend on ℓ. The function $\mathcal{P}$ is called the *kernel* of the semi–Markov process $\{x_t\}_{t\geq 0}$. Then the sequence $Y = (Y_n)_{n\in\mathbb{N}}$ is a Markov chain with transition matrix:

$$R(i, j) = \mathcal{P}(i, j, \infty) = \lim_{t\uparrow\infty} \mathcal{P}(i, j, t)\,,\ (i, j) \in E \times E$$

and $u_1, u_2, \cdots$ are conditionally independent given Y. ■

If the kernel $\mathcal{P}$ is defined as

$$\mathcal{P}(i, j, t) = \rho_{ij}(1 - e^{-\lambda_i t})\ ,\ (i, j) \in E \times E$$

where $\lambda_i \in]0, \infty[$, $\rho_{ii} = 0$ and $\sum_{j\in E} \rho_{ij} = 1\ , \forall i \in E$, then $\{X_t\}_{t\geq 0}$ is a time–homogeneous Markov process with generator $Q(i, j) = q_{ij} = \lambda_i \rho_{ij},\ j \neq i$. On the other hand, if $E = \{i\}$ is a singleton, then $(u_n)_{n\in\mathbb{N}}$ is a time–homogeneous *renewal process* with an inter–arrival time distribution of the form $F(t) = \mathcal{P}(i, i, t)$.

The law of a semi–Markov process $\{X_t\}_{t\geq 0}$ is jointly characterized by the transition probability $R(i,j)$ of the embedded Markov chain $(Y_n)_{n\in\mathbb{N}}$ and the conditional sojourn distributions:

$$G(i,j,t) = \Pr\{u_\ell \leq t \mid Y_{\ell-1} = i, Y_\ell = j\}, \ell \in \mathbb{N}, (i,j) \in E \times E$$

The kernel function of this semi–Markov process is then defined as:

$$\mathcal{P}(i,j,t) = R(i,j)\ G(i,j,t)$$

from which are deduced unconditional sojourn distributions:

$$F(i,t) = \Pr\{u_\ell \leq t \mid Y_{\ell-1} = i\} = \sum_{j\in E} \mathcal{P}(i,j,t)$$

Let us recall that if the Markov chain $Y = (Y_n)_{n\in\mathbb{N}}$ is irreducible and recurrent non–null, there exists a limiting probability distribution v on E of the form:

$$v_j = \sum_{i\in E} v_i\, R(i,j), \quad j \in E$$

or in matrix notation:

$$v = vR$$

Moreover, if $\overline{u}_1(i) = E[u_1 \mid Y_0 = i] < \infty,\ \forall i \in E$, then the limit distribution of the semi–Markov process $\{X_t\}_{t\geq 0}$ is given by:

$$\Pi_j = \lim_{t\uparrow\infty} \Pr\{x_t = j\} = \frac{v_j \overline{u}_1(j)}{\sum_{i\in E} v_i \overline{u}_1(i)}$$

Notice that this relation between Π and v is more general than the one for Markov processes, for which $\overline{u}_1(i) = \lambda_i^{-1}$.
The main statistical problem is to estimate the semi–Markov kernel $\mathcal{P}$. Here we concentrate on a fully nonparametric estimation procedure for a semi–Markov process$\{X_t\}_{t\geq 0}$, where the distribution of a sojourn in state i does not depend on the next state to be entered, *i.e.* :

$$G(i,j,t) = F(i,t), \quad \forall (i,j) \in E \times E, \forall t \in \mathbb{R}^+$$

Then $R(i,j)$ and $F(i,t)$ can be estimated from N *i.i.d.* realizations of $\{X_t\}_{t\geq 0}$ over a fixed time interval $[0,T]$. In that case, let us denote $\mathcal{N}_{T,N}(i,j)$ and $\mathcal{N}^*_{T,N}(i) = \sum_{j\in E} \mathcal{N}_{T,N}(i,j)$ the number of transitions from i to j in $[0,T]$ and the number of sojourns in state i completed before time T, respectively. Then nonparametric maximum–likelihood estimators of the unconditional sojourn distributions and of the transition matrix of the embedded Markov chain are respectively given by:

$$\widehat{F}(i,t) = \mathcal{N}^*_{T,N}(i)^{-1} \sum_{\ell=1}^{\mathcal{N}^*_{T,N}(i)} \mathbb{1}(Y_{\ell-1} = i, u_\ell \leq t)$$

and

$$\widehat{R}(i,j) = \mathcal{N}_{T,N}(i,j)/\mathcal{N}^*_{T,N}(i)\,.$$

Consequently, one obtains:

$$\widehat{\mathcal{P}}(i,j,t) = \widehat{R}(i,j) \cdot \widehat{F}(i,t)\,.$$

Asymptotic properties (convergence, normality) of these nonparametric estimators are reviewed by Karr [1986], Theorem 8.33. Non–parametric estimation of the kernel $\mathcal{P}$ of partially observed renewal processes has been considered by Gill [1980] and surveyed by Karr [1986], p. 347–351.

7 Statistical analysis of counting processes

In this section, we present both the statistical analysis of counting processes based on martingale estimators and the extension to these processes of the semiparametric inference initially proposed by Cox [1972], [1975], for duration models. For that purpose, we consider a multivariate counting process with covariates, but our presentation is restricted to the case of a non–censored independent sampling scheme for a counting process derived from a point process.

Let us denote n the number of individuals and i an element of $\{1, \cdots, n\}$. For any individual i, we observe both the path X^i_t of a point process valued in a discrete state space E with K elements and the path of a (multivariate) covariate process $Z^i = (Z^i_t)_t$. These two processes are observed over an interval $[0, T]$ for any i. Given $(Z^i_t)_t$, the $(X^i_t)_t$'s are assumed to be independent. The distribution of $(X^i_t)_t$ is also assumed to be independent of the $(Z^j_t)_t$ for any $j \neq i$, *i.e.* it is independent of the covariate processes of other individuals.

Now we have to describe the distribution of $(X^i_t)_t$ given $(Z^i_t)_t$. This goal is achieved by representing the point process $(X^i_t)_t$ through a multivariate counting process $(N^{i,j}_t)_t$, where $j = (k,\ell), k, \ell \in E,\ k \neq \ell,\ j \in \{1, \cdots, J = K(K-1)\}$. This counting process increases by jumps of size one when the individual i moves from state k to state ℓ. The distribution of $(N^{i,j}_t)_t$, given $(Z^i_t)_t$, is characterized by its stochastic intensity with respect to the filtration generated by both the past of *all* the $N^{i,j}_t$ processes, for $i \in \{1, \cdots, n\}$ and $j \in \{1, \cdots, J\}$, and by *the whole trajectories* of all the (Z^i_t) processes. These stochastic intensities are assumed to take the following form:

$$h^{i,j}_t = \psi^j\big((Z^i_t)_t, \theta\big)\, \lambda^j_t\ Y^{i,j}_t\ , i = 1, \cdots, n\,, \quad j = 1, \cdots, J$$

where:
i) $\psi^j\big((Z_t^i)_t, \theta\big)$ is a *known positive function* depending on an unknown parameter $\theta \in \Theta \subset I\!R^p$; in practice, each ψ^j may depend on a subvector of θ only, and then one has to check if the vector θ is identified by the vector of the ψ^j functions;
ii) λ_t^j is the *baseline intensity function* of $N_t^{i,j}$; it does not depend on individual i; the model is then a proportional hazard type model in which covariates act multiplicatively through the ψ^j functions on the baseline intensity; moreover, λ_t^j is assumed to be non–stochastic and hence a function valued in $I\!R^+$; in the semiparametric approach, the function λ_t^j is assumed to be totally unknown and the distribution of the X_t^i processes are then characterized by parameters θ and by functions $(\lambda_t^j),\ j = 1, \cdots, J$;
iii) finally, $Y_t^{i,j}$ is a family of *observable predictable stochastic processes* valued in $\{0, 1\}$; we restrict our attention to the case where $Y_t^{i,j}$ characterize the fact that the individual i is "at risk" at time t for jumping from state k to state ℓ, if $j = (k, \ell)$, or equivalently:

$$Y_t^{i,j} = 1 \qquad \text{if } x_{t^-}^i = k$$

$$Y_t^{i,j} = 0 \qquad \text{elsewhere}$$

As $h_t^{i,j}$ is the stochastic intensity with respect to all the trajectories of the covariate processes, it would be dependent of all the $(Z_t^i)_t, i = 1, \cdots, n$. However, we have assumed that $(X_t^i)_t$ is independent of $(Z_t^{i'})_t$ for any $i' \neq i$ given $(Z_t^i)_t$, and this assumption is expressed by the fact that ψ^j depends only on $(Z_t^i)_t$. In fact, this requirement is not an assumption but is a condition on the definition of the (Z_t^i) processes which may have some elements in common. Moreover, ψ^j may be a function of the whole trajectory of $(Z_t^i)_t$ or of the current value Z_t^i only. The first case requires the continuous–time observation of covariates, which is unrealistic, or some approximation procedures such as discretization of stochastic integrals. The more common case is the one where the instantaneous probability of a jump from state k to state ℓ for the individual i depends only on the current value of the process Z^i, which implies that $\psi^j\big((Z_t^i)_t, \theta\big)$ may be written as $\psi^j(Z_t^i, \theta)$. For example, if (Z_t^i) is a q–dimensional process, a usual specification is the following:

$$\psi^j(Z_t^i, \theta) = \exp(Z_t^{i'}\theta^j)$$

where $\theta^j \in I\!R^q$ and $\theta = (\theta^j)_{j=1,\cdots,J}$. More generally, such a specification may be constrained by imposing that some components of Z_t^i in the ψ^j function are eliminated.

Up to an additive constant, the log–likelihood of the model is equal to

$$\begin{aligned} & L_T(\theta, \lambda^1, \cdots, \lambda^J) \\ = & \sum_{i=1}^{n} \sum_{j=1}^{J} \int_0^T \ln\, \psi^j(Z_t^i, \theta)\, dN_t^{i,j} + \sum_{j=1}^{J} \int_0^T \ln\, \lambda_t^j\; d\overline{N}_t^j \\ & - \sum_{i=1}^{n} \sum_{j=1}^{J} \int_0^T \psi^j(Z_t^i, \theta) \lambda_t^j Y_t^{i,j}\; dt \end{aligned}$$

where $\overline{N}_t^j = \sum_{i=1}^n\; N_t^{i,j}$. The maximum likelihood estimator of θ can be derived from this last equation if the λ_t^j are known functions of unknown parameters. However, the log-likelihood is unbounded if the functions λ_t^j are taken as arguments: λ_t^j may be chosen arbitrarily large at observed jump times (and then the second element in the right hand side of the log-likelihood equation may be as large as desired) and null at other times (and then the third element in the r.h.s. of the log-likelihood equation becomes equal to zero). Then it appears clearly that estimation must be based on a different procedure: parameters θ can be estimated by maximizing Cox's partial likelihood, and integrals of the λ_t^j's are then estimated through martingale techniques.

7.1 The Cox likelihood

Following an argument given by Karr [1986], Chapter 5, the Cox's likelihood can be derived as the difference between the log–likelihood function of the observations and the log–likelihood function of the $\overline{N}_t^j$ processes. This difference is a function of θ only, and can be heuristically interpreted as the logarithm of the density of the $N_t^{i,j}$ given $\overline{N}_t^j$ and the covariates. Given the same filtration, intensities of the $\overline{N}_t^j$ processes are the sum over i of $h_t^{i,j}$, *i.e.*

$$\overline{h}_t^j = \lambda_t^j \left(\sum_{i=1}^{n} \psi^j(Z_t^i, \theta)\, Y_t^{i,j} \right)$$

and the log–likelihood of the statistic $\overline{N}_t^j$ is equal to

$$\begin{aligned} & L_T^*(\theta, \lambda^1, \cdots, \lambda^J) \\ & = \textstyle\sum_{j=1}^{J} \int_0^T \ln\; \sum_{i=1}^{n} \big(\psi^j(Z_t^i, \theta) Y_t^{i,j}\big) d\overline{N}_t^j + \sum_{j=1}^{J} \int_0^T \ln\, \lambda_t^j\, d\overline{N}_t^j \\ & - \textstyle\sum_{j=1}^{J} \int_0^T \lambda_t^j \big(\sum_{i=1}^{n} \psi^j(Z_t^i, \theta) Y_t^{i,j} \big)\, dt \end{aligned}$$

The Cox likelihood is then defined as:

$$\begin{aligned} C_T(\theta) &= L_T(\theta, \lambda^1, \cdots, \lambda^J) - L_T^*(\theta, \lambda^1, \cdots, \lambda^J) \\ &= \sum_{j=1}^{J} C_T^j(\theta) \end{aligned}$$

where

$$C_T^j(\theta) = \sum_{i=1}^{n} \int_0^T \ln\, \psi^j(Z_t^i, \theta)\, dN_t^{i,j} - \int_0^T \ln\, \big(\sum_{i=1}^{n} \psi^j(Z_t^i, \theta) Y_t^{i,j}\big)\, d\overline{N}_t^j$$

or equivalently:

$$\exp[C_T^j(\theta)] = \frac{\prod_{i=1}^{n} \prod_{\tau_u^{i,j} \leq T} \psi^j(Z^i_{\tau_u^{i,j}}, \theta)}{\prod_{\overline{\tau}_u^j \leq T} \sum_{i=1}^{n} \psi^j(Z^i_{\overline{\tau}_u^j}, \theta) Y^{i,j}_{\overline{\tau}_u^j}}$$

In this last expression, the second product of the numerator is computed for all the observed jump times $\tau_u^{i,j}$ of the process $N_t^{i,j}$ and the product in the denominator is computed for all the jump times $\overline{\tau}_u^j$ of the process $\overline{N}_t^j$, *i.e.* for all the transitions from state k to state ℓ (if $j = (k, \ell)$) observed over all the individuals. Parameters θ are estimated via the maximization of $C_T(\theta)$. Moreover, if $\psi^j(Z_t^i, \theta)$ depends on a subvector θ^j such that all the θ^j's are variation free, the estimator of θ^j may be obtained through a maximization of $C_T^j(\theta) = C_T^j(\theta^j)$ only. In this case, observations of $N_t^{i,j}$ for any i and t are sufficient for the estimation of θ^j.
Asymptotic properties of the maximand of $C_T(\theta)$, denoted $\widehat{\theta}$, have been studied initially by Andersen and Gill [1982] and surveyed, for example, by Karr [1986], Chapter 5. Under usual regularity conditions, it could be shown that $\widehat{\theta}$ is a consistent estimator of θ when n tends to ∞ and that $\sqrt{n}(\widehat{\theta} - \theta)$ is asymptotically normal with variance explicitly given, for example, by Karr [1986], Chapter 5, formulas (5.90a) to (5.91)).

7.2 The martingale estimation of the integrated baseline intensity

For simplicity, let us first present martingale estimators for *i.i.d.* counting processes, *i.e.* without presence of covariates. The likelihood of such a model is obtained by setting $\psi^j(Z_t^i, \theta)$ equal to 1 in the log-likelihood function:

$$\begin{aligned} & L_T(\theta, \lambda^1, \cdots, \lambda^J) \\ = & \sum_{i=1}^{n} \sum_{j=1}^{J} \int_0^T \ln\, \psi^j(Z_t^i, \theta)\, dN_t^{i,j} + \sum_{j=1}^{J} \int_0^T \ln\, \lambda_t^j\, d\overline{N}_t^j \\ & - \sum_{i=1}^{n} \sum_{j=1}^{J} \int_0^T \psi^j(Z_t^i, \theta) \lambda_t^j Y_t^{i,j}\, dt \end{aligned}$$

In this case, one can easily verify that the log–likelihood is a function of $\overline{N}_t^j$ only, up to an additive constant. This means that these processes constitute a sufficient statistic. Indeed, in this case, the log–likelihood function becomes:

$$L_T(\lambda^1, \cdots, \lambda^J) = \sum_{j=1}^{J} \big[\int_0^T \ln\, \lambda_t^j \, d\overline{N}_t^j - \int_0^T \lambda_t^j \overline{Y}_t^j \, dt \big]$$

and the processes $\overline{N}_t^j$ have the following stochastic intensities:

$$\overline{h}_t^j = \lambda_t^j \cdot \overline{Y}_t^j$$

where $\overline{Y}_t^j = \sum_{i=1}^n Y_t^{i,j}$ is the number of individuals at risk for the transition of type j (from state k to state ℓ) at time t.
We want to estimate the integrals of λ_t^j for any j. However, in practice, information is only available for the time interval in which there exists some individuals from the sample who are at risk for the analyzed transition. Functions of interest are then:

$$\Lambda_t^j = \int_0^t \lambda_s^j \, \mathbb{1}(\overline{Y}_s^j > 0) \, ds$$

where $\mathbb{1}(\overline{Y}_s^j > 0) = 1$ if $\overline{Y}_s^j > 0$ and 0 elsewhere.
The *martingale estimator* of Λ_t^j is defined by:

$$\widehat{\Lambda}_t^j = \int_0^t (\overline{Y}_s^j)^{-1} \, \mathbb{1}(\overline{Y}_s^j > 0) \; d\overline{N}_s^j$$

This estimator may be heuristically justified by the following argument. Let us start with the differential representation of a counting process:

$$d\overline{N}_s^j = h_s^j \; ds + dM_s^j$$

where M_s^j is a martingale. In our model, this expression becomes

$$d\overline{N}_s^j = \lambda_s^j \cdot \overline{Y}_s^j ds + dM_s^j$$

which can be pre–multiplied by $(\overline{Y}_s^j)^{-1} \; \mathbb{1}(\overline{Y}_s^j > 0)$ to give:

$$(\overline{Y}_s^j)^{-1} \, \mathbb{1}(\overline{Y}_s^j > 0) \, d\overline{N}_s^j$$

$$= \lambda_s^j \, \mathbb{1}(\overline{Y}_s^j > 0) ds + (\overline{Y}_s^j)^{-1} \, \mathbb{1}(\overline{Y}_s^j > 0) \; dM_s^j$$

Integrating the two sides of this relation yields:

$$\widehat{\Lambda}_t^j = \Lambda_t^j + \int_0^t (\overline{Y}_s^j)^{-1} \; \mathbb{1}(\overline{Y}_s^j > 0) \; dM_s^j$$

The difference between $\widehat{\Lambda}_t^j$ and Λ_t^j is then a stochastic integral of a predictable process with respect to a martingale; so it is a martingale (see Dellacherie and Meyer [1980], Chapter 7, Theorem 3). Moreover, it can be verified that

$$E(\widehat{\Lambda}_t^j - \Lambda_t^j) = 0$$

and

$$< \widehat{\Lambda}_t^j - \Lambda_t^j >= \int_0^t \lambda_s^j (\overline{Y}_s^j)^{-1} \mathbb{1}(\overline{Y}_s^j > 0)\, ds$$

Let us recall that the predictable variation $E_t^j =< \widehat{\Lambda}_t^j - \Lambda_t^j >$ plays the role of an instantaneous variance. In particular:

$$V(\widehat{\Lambda}_t^j - \Lambda_t^j) = E(< \widehat{\Lambda}_t^j - \Lambda_t^j >)$$

Using a martingale estimation approach, E_t^j may be estimated by

$$\widehat{E}_t^j = \int_0^t (\overline{Y}_s^j)^{-2}\ \mathbb{1}(\overline{Y}_s^j > 0)\, ds$$

Under standard regularity conditions, estimators $\widehat{\Lambda}_t^j$ are asymptotically well–behaved. They are consistent in a strong sense

$$E\left[\sup_t (\widehat{\Lambda}_t^j - \Lambda_t^j)^2\right] \to 0 \text{ when } n \to \infty$$

and $n(\widehat{\Lambda}_t^j - \Lambda_t^j)$ is asymptotically distributed as a centered Gaussian martingale with continuous trajectories and whose predictable variation may be estimated by $n\widehat{E}_t^j$.
Let us now return to the general model with covariates. The differential representation of a process $N_t^{i,j}$ is then:

$$d\ N_s^{i,j} = \psi^j(Z_s^i, \theta) \cdot \lambda_s^j \cdot Y_s^{i,j} ds + dE_s^{i,j}$$

where $E_s^{i,j}$ is a zero–mean martingale. From the definition of $d\overline{N}_s^j$, we obtain:

$$d\overline{N}_s^j = \left[\sum_{i=1}^n \psi^j(Z_s^i, \theta) \cdot Y_s^{i,j}\right] \lambda_s^j ds + d\overline{E}_s^j$$

in which $\overline{E}_s^j = \sum_{i=1}^n E_s^{i,j}$ is still a zero–mean martingale.
Now let us define:

$$w_t^j(\theta) = \sum_{i=1}^n \psi^j(Z_s^i, \theta) \cdot Y_s^{i,j}$$

Assuming that θ is known, the problem of inference on the integral of the λ_t^j's is identical to the previous case without covariates. The function parameters are now:

$$\Lambda_t^j(\theta) = \int_0^t \mathbb{1}(w_s^j(\theta) > 0)\, \lambda_s^j\, ds$$

and their estimators are given by:

$$\widehat{\Lambda}_t^j(\theta) = \int_0^t [w_s^j(\theta)]^{-1} \mathbb{1}(w_s^j(\theta) > 0)\, d\overline{N}_s^j$$

If a Cox procedure is initially used and provides an estimator $\widehat{\theta}$ of θ regardless of the λ_t^j's, an estimator of $\Lambda_t^j(\theta)$ is obtained by substituting $\widehat{\theta}$ for θ in this last expression. It can be proved (see Andersen and Gill [1982]) that asymptotic properties of $\widehat{\Lambda}_t^j(\widehat{\theta})$ are identical to those of $\widehat{\Lambda}_t^j(\theta)$ and that estimators $\widehat{\Lambda}_t^j(\widehat{\theta})$ are independent of each other and independent of $\widehat{\theta}$ asymptotically.

8 Conclusions

This chapter focused on definitions and statistical analysis of duration models and point processes. More extensive presentations are contained in textbooks by Kalbfleisch and Prentice [1980], Lawless [1982], Jacobsen [1982], Cox and Oakes [1984], Karr [1986], Daley and Vere–Jones [1988], Lancaster [1990], Andersen *et alii* [1993],or in detailed surveys by Heckman and Singer [1984a], Kiefer [1988], Serfozo [1990], Van den Berg [2001]. Markov chains have been completely studied by Chung [1967], Freedman [1971], Revuz [1975] and by Ethier and Kurtz [1986].

REFERENCES

Abbring, J. and Van den Berg, G. [2003]: The nonparametric identification of treatment effects in duration models, *Econometrica*, **71**, 1491-1518.

Andersen, P.K. and Gill, R. [1982]: Cox's regression model for counting processes: a large sample study, *Annals of Statistics*, **10**, 1100–1120.

Andersen , P. K., Borgan, Ø., Gill, R. D. and Keiding, N. [1993]: *Statistical models based on counting processes*. New-York: Springer.

Basawa, I.V. and Prakasa Rao, B.L.S. [1980]: *Statistical inference of stochastic processes*. New–York: Academic Press.

Bhattacharya, R.N. and Waymire, E.C. [1990]: *Stochastic processes with applications*. New-York: Wiley and Sons.

Billingsley, P. [1961]: *Statistical inference for Markov processes*. The University of Chicago Press.

Chung, K.L. [1967]: *Markov chains with stationary transition probabilities*. New-York: Springer–Verlag.

Cox, D.R. [1972]: Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, series B,* **34**, 187–220.

Cox, D.R. [1975]: Partial likelihood, *Biometrika,* **62**, 269-276.

Cox, D.R. and Miller, H.D. [1966]: *The theory of stochastic processes*. London: Methuen.

Cox, D.R. and Oakes, D. [1984]: *Analysis of survival data*. London: Chapman and Hall.

Daley, D.J. and Vere-Jones, D. [1988]: *An introduction to the theory of point processes*. New-York: Springer-Verlag.

Dellacherie, C. and Meyer, P.A. [1980]: *Probabilité et potentiel* (Chapitres V à VIII: Théorie des martingales). Paris: Hermann.

Doob, J.L. [1953]: *Stochastic processes*. New-York: Wiley and Sons.

Elbers, C. and Ridder, G. [1982]: True and spurious dependence: the identifiability of the proportional hazard model, *Review of Economic Studies,* **XLIX**, 403-409.

Ethier, S.N. and Kurtz, T.G. [1986]: *Markov processes: characterization and convergence*. New-York: Wiley and Sons

Fougère, D., Le Bihan, H. and Sevestre P. [2007]: Heterogenity in consumer price stickiness: a microeconometric approach, *Journal of Business and Economic Statistics*, **25**, 247-264.

Fourgeaud, C., Gouriéroux, C. and Pradel, J. [1990]: Heterogeneity and hazard dominance in duration data models (in french), *Annales d'Economie et de Statistique,* **18**, 1-24.

Freedman, D. [1971]: *Markov chains*. San Francisco: Holden-Day.

Gill, R.D. [1980]: Non-parametric estimation based on censored observations of a Markov renewal process, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **53**, 97-116.

Han, A. and Hausman, J.A. [1990]: Specification and semiparametric estimation of duration models, *Journal of Applied Econometrics,* **5**, 1-28.

Heckman, J.J. and Honoré, B. [1989]: The identifiability of the competing risks model, *Biometrika,* **76**, 325-330.

Heckman, J.J. and Singer, B. [1984a]: Econometric duration analysis, *Journal of Econometrics,* **24**, 63-132.

Heckman, J.J. and Singer, B. [1984b]: The identifiability of the proportional hazards model, *Review of Economic Studies,* **51**, 231-243.

Honoré B. [1993]: Identification results for duration models with multiple spells, *Review of Economic Studies,* **60**, 241-246.

Jacobsen, M. [1982]: *Statistical analysis of counting processes.* Berlin: Springer-Verlag.

Kalbfleisch, J. D., and Prentice, R. L. [1980]: *The statistical analysis of failure time data.* New York: Wiley.

Karr, A.F. [1986]: *Point processes and their statistical inference.* New-York: Marcel Dekker.

Kiefer N. [1988]: Economic duration data and hazard functions, *Journal of Economic Literature,* **XXVI**, 646-679.

Kortram R. A., Lenstra A. J. , Ridder G. and Van Rooij A. C. M. [1995]: Constructive identification of the mixed proportional hazards model, *Statistica Neerlandica*, **49**, 269-281.

Lancaster T. [1990]: *The Econometric analysis of transition data*, Econometric Society Monographs, Cambridge University Press: New York.

Lawless J.F. [1982]: *Statistical models and methods for lifetime data.* New York: Wiley.

Marshall A.W. and Olkin I. [1967]: A multivariate exponential distribution, *Journal of the American Statistical Association,* **62**, 30-44.

Mátyás L. and P. Sevestre (editors)[1996]: *The Econometrics of Panel Data*, Second Revised Edition, Dordrecht: Kluwer Academic Publishers.

Melino A. and G. T. Sueyoshi [1990]: A simple approach to the identifiability of the proportional hazard model, *Economics Letters*, **33**, 63-68.

Mouchart, M. [2004], The econometrics of duration data and of point processes, Lecture Notes, *http://www.stat.ucl.ac.be/ISpersonnel/mouchart/DuDa040908$_b$.pdf*

Mouchart, M. and J.-M. Rolin [2002]: Competing risks models : problems of modelling and of identification, in *Life tables, modelling survival and death,* edited by G. Wunsch, M. Mouchart and J. Duchêne, Dordrecht: Kluwer Academic Publishers, 245-267.

Ridder G. [1990]: The nonparametric identification of generalized accelerated failure-time models, *Review of Economic Studies,* **57**, 167-181.

Revuz, D. [1975]: *Markov chains.* New-York: North Holland / American Elsevier.

Serfozo, R.F. [1990]: Point processes, in *Handbook in operations research and management science*, vol. 2, edited by D.P. Heyman and M.J. Sobel, Amsterdam: North-Holland, 1-94.

Tsiatis A.A. [1981]: A large sample study of Cox's regression model, *Annals of Statistics,* **9**, 93-108.

Van den Berg G. [2001]: Duration models: specification, identification and multiple durations, in *Handbook of econometrics*, vol. 5, edited by J. Heckman and E.Leamer, Amsterdam: North-Holland, 3381-3460.