Inference in nonparametric current status models with covariates

Odile Pons

INRA, Mathématiques, 78352 Jouy en Josas cedex, France
e-mail: Odile.Pons@jouy.inra.fr

Abstract: In interval censored models with current status observations, the
variables are indicators of the presence of individuals on observation inter-
vals and covariates. When several individuals share the same observation
interval, a simple procedure provides new estimators for the distribution of
the observation times and their intensity, in a closed form. They are $n^{1/2}$
consistent for piece-wise constant covariates. Estimators of the sample-sizes
are deduced and asymptotic $\chi^2$ tests for independence of the observations
on consecutive intervals and for independence between consecutive classes
for the observed individuals are proposed.

AMS 2000 subject classifications: Primary 60J25, 62A10, G2G05.
Keywords and phrases: Markov process, interval censoring, current status.

Contents

1 Introduction ......................................................... 1
2 Models with independent observations .......................... 2
3 Identifiability and estimation of the parameters .................. 5
  3.1 Model without covariates .................................... 5
  3.2 Models with covariates ...................................... 6
  3.3 Estimation of the sample size ............................... 9
4 Models with dependent observations on consecutive intervals .... 10
  4.1 Nonparametric models .................................. 10
  4.2 Markov models ........................................ 11
References ......................................................... 12

1. Introduction

Statistical inference for sequential observations of individuals in a large popula-
tion differs according to the nature of the samples. The observation of presence
of individuals at specific locations is often restricted to a sequence of time inter-
vals. In capture-recapture models, the size of finite and closed populations has
been estimated under the assumptions of the same parametric model for the
consecutive samples and time-dependent intensities for the transitions of the
populations between several states, with individual covariates [1, 6, 7].
The discrete observation sampling leads to cumulative observations on fixed or random intervals, it is an interval censored model with only current status observations. With individual observation times for all the individuals, the monotonic nonparametric maximum likelihood estimator of the time-dependent cumulative hazard function relies on the greatest convex minorant algorithm, it weighs the random observation times and converges at the rate $n^{1/3}$ (see [2, 3] and [4] in a model with constant covariates). Here a nonparametric Markov model with piece-wise constant covariate processes is considered as in [5] for continuous observations, and the observations are current status data with common observation intervals. A simple reparametrization leads to easily calculated parametric estimators for the distribution functions of the observation times and the population sizes are estimated (section 3). The convergence rates of the estimators in several nonparametric models is $n^{1/2}$. In section 4, models with dependent observations on consecutive time intervals are considered and new estimators and tests for independence are proposed.

2. Models with independent observations

Consider a population of $L$ independent classes $C_1, \ldots, C_L$ of respective unknown sizes $\nu_l$, $l = 1, \ldots, L$ and $\nu = \nu_1 + \ldots + \nu_L$. In each class, a sample of the population is performed on a time interval $[0, \tau]$ with random sampling sizes $n_l$, $l = 1, \ldots, L$ and $n$. Let $\tau_{l,1} < \ldots < \tau_{l,K_l} \leq \tau$ be the end-point observation intervals for class $C_l$ and $(N_{l}(t))_{t \leq \tau}$ be the counting process of the observations of individual $i$ of $C_l$ restricted to the intervals $I_{l,k} = [\tau_{l,k-1}, \tau_{l,k}]$, $k = 1, \ldots, K_l$ up to time $t$,

$$N_{li}(t) = \sum_{k=1}^{K_l} \delta_{li,k} 1\{I_{l,k} \cap [0, t] \neq \emptyset\}, \quad \text{with } \delta_{li,k} = 1\{i \in C_l \text{ is observed on } I_{l,k}\},$$

with $N_{li}(\tau) \leq K_l$, $\sum_{l=1}^{n} 1\{N_{li}(\tau) > 0\} = n_l$. Only cumulated numbers $N_{li}(I_{l,k})$ are observed.

An individual $i$ of $C_l$ is supposed to be characterized by a $p$-dimensional random covariate vector process $Z_{li}$ having left-continuous sample-paths with right-hand limits. The individuals are sampled independently and for $l = 1, \ldots, L$, the processes $(N_{li}, Z_{li})$, $i = 1, \ldots, n_l$, are mutually independent and identically distributed. The distribution of $N_{li}$ conditionally on $Z_{li}$ is supposed to follow a Markov model with independent increments, where the probability of observing individuals only depends on their characteristics on the observation interval

$$\Pr(N_{li}(I_k) | (Z_{li}(s))_{s \leq \tau_{l,k}}) = \Pr(N_{li}(I_k) | Z_{li}(I_{l,k})), \quad (1)$$

only a countable set of values of the process $Z$ appears in the whole sample-path of $N_{li}$.

The process $Z_{li}$ is sometimes restricted to a piece-wise constant process with values $Z_{l,j}$ on a random sub-partition $I_{l,i,j} = [U_{l,i,j-1}, U_{l,i,j}]$, $j = 1, \ldots, J$ of
\((I_{l,k})_{l,k}\)

\[
Z_{li}(t) = \sum_{j=1}^{J} Z_{l,j} 1\{t \in I'_{l,j}\}. \tag{2}
\]

The probability of observation of \(i \in C_l\) on the partitions \((I_{l,k})_k\) is a discrete process defined according to the assumption (1) or (2). Let \(T_{li,k}\) be the unknown first presence time of \(i\) during the time interval \(I_{l,k}\), and we suppose that the model is defined by

\[
p_{l,k}(Z_{li}) = \Pr(\tau_{l,k} - 1 < T_{li,k} \leq \tau_{l,k} | Z_{li}) = \sum_{j} 1\{t \in I'_{l,j} \subset I_{l,k} \} \Pr(U_{li,j-1} < U_{li,j} | Z_{li}(U_{li,j-1}))
\]

\[
P_l(Z_{l,j}) = \Pr(Z_{li}(U_{li,j-1}) = Z_{lj}),
\]

\[
p_l = \Pr(N_{li}(\tau_{l,k}) > 0) = \int \Pr(N_{li}(\tau_{l,k}) > 0 | Z_{li}(\tau_{l,k})) dP_l(Z_{li})
\]

\[
= \sum_{j=1}^{J} \Pr(N_{li}(I'_{l,j}) > 0 | Z_{li}(U_{li,j-1}) = Z_{lj}) P_l(Z_{lj})
\]

\[
= \sum_{k=1}^{K_l} \sum_{j=1}^{J} p_{l,k}(Z_{lj}) P_l(Z_{lj}),
\]

\[
1 - p_l = \Pr(N_{li}(\tau_{l,k}) = 0).
\]

However individuals \(i\) with \(N_{li}(\tau_{l,k}) = 0\) are not observed. An underlying time-continuous model is defined by the intensities of observation of the individuals. The conditional intensity of observation of class \(C_l\) is supposed to depend only on the current value of the covariate, for individual \(i\) in \(C_l\) and \(t\) in \(I_{l,k}\), it is defined by

\[
\lambda_{l,k}(t, Z_{li}) = \lim_{h \to 0} \frac{1}{h} \Pr(N_{li}(t+h) - N_{li}(t) > 0 | Z_{li}(t) = z)
\]

More generally, the capture intensity for class \(l\) is defined as one of the intensity \(\lambda_{l,k}\) by

\[
\lambda_l(t, Z_{li}) = \lim_{h \to 0} \frac{1}{h} \Pr(N_{li}(t+h) - N_{li}(t) > 0 | Z_{li}(t))
\]

\[
= \lim_{h \to 0} \sum_{k=1}^{K_l} 1\{t \in I_{l,k}\} \sum_{j=1}^{J} 1\{t \in I'_{l,j}\} \lambda_{l,k}(t, Z_{lj})\text{ under (2)}.
\]

The variation of the cumulative intensities on each sub-interval are denoted

\[
\Delta \lambda_{l,k}(t, Z_{li}) = \int_{I_{l,k} \cap [0,t]} \lambda_{l,k}(s, Z_{li}(s)) ds
\]

\[
= \sum_{j=1}^{J} 1\{I'_{l,j} \subset I_{l,k}\} \int_{I'_{l,j} \cap [0,t]} \lambda_{l,k}(s, Z_{lj}) ds
\]
under \(\text{(2)}\) and the cumulative intensities from 0 is
\[
\Lambda_l(t, Z_{li}) = \sum_{k=1}^{K_l} 1\{t \in I_{l,k}\} \sum_{k'=1}^{k} \Delta \Lambda_{l,k'}(t, z).
\]
The unobserved apparition time \(T_{li,k}\) of \(i\) in \(C_l\) during the time interval \(I_{l,k}\) has a conditional distribution \(\Pr\{T_{li,k} \leq t|Z_{li}(\tau_{l,k}) = Z_{li,j}\} = 1 - S_l(t, Z_{li,j})\), for a covariate value \(Z_{li,j}\). The probability of observation in \(C_l\) is continuously defined as
\[
p_{l,i}(t, Z_{li}) = \Pr(N_{li}(t) - N_{li}(\tau_{l,k-1}) > 0|Z_{li}(t) = z) = S_l(\tau_{l,k-1}, z) - S_l(t, z) = \exp\{-\Delta \Lambda_{l,k}(t, z)\} - \exp\{-\Delta \Lambda_{l,k}(\tau_{l,k-1}, z)\}, \quad t \in I_{l,k},
\]
\[
p_{l}(t, Z_{li}) = \Pr(N_{li}(t) > 0|Z_{li}) = 1 - \exp\{-\Lambda_l(t, Z_{li})(t)\},
\]
where \(p_{l}(t, Z_{li})\) is the distribution function of observation for an individual of \(C_l\) before \(t\) conditionally on the covariate. For \(t\) in \(I_{l,k}\), it is written \(p_{l,i}(t, Z_{li}) = \sum_{k' < k} p_{l,i,k'}(Z_{li}) + p_{l,i,k}(t, Z_{li})\).

In a discrete nonparametric model, the hazard function of individual \(i\) in \(C_l\) with covariate value \(Z_{li,j}\) on an interval \(I_{l,i,j}\) is written \(\sum_k \lambda_{l,k}(t, Z_{li,j}) 1\{t \in I_{l,k} \cap I_{l,i,j}\}\).

The proportional hazards model is defined by multiplicative intensities
\[
\lambda_{l,k}(t, Z_{li}) = \lambda_l(t) e^{\beta_l^\prime_i Z_{li}(t)} = \lambda_l(t) \sum_{j=1}^{J} e^{\beta_{l,j}^\prime Z_{li,j}} 1\{t \in I_{l,i,j}\},
\]
then
\[
\Delta \Lambda_{l,k}(t, Z_{li}(t)) = \sum_{j=1}^{J} e^{\beta_{l,j}^\prime Z_{li,j}} \int_{I_{l,k} \cap I_{l,i,j} \cap [0,t]} \lambda_l(s) \, ds
\]
\[
= \sum_{j=1}^{J} e^{\beta_{l,j}^\prime Z_{li,j}} \Lambda_l(I_{l,k} \cap I_{l,i,j} \cap [0,t]).
\]

Let \(S_l(t) = \exp\{-\int_0^t \lambda_l(s) \, ds\}\), for the \(\nu_l\) individuals, then the probability of being unobserved is \(\Pr(T_{li} > \tau_{li,K_l}) = 1 - p_{l}(\tau_{li,K_l})\), where \(T_{li}\) the first presence time of \(i\),
\[
1 - p_{l}(\tau_{li,K_l}, Z_{li}) = \exp\{-\Lambda_l(\tau_{li,K_l}, Z_{li})\} = \prod_{k=1}^{K_l} \exp\{-\Delta \Lambda_{l,k}(\tau_{li,K_l}, Z_{li})\}
\]
\[
= \prod_{k=1}^{K_l} \frac{S_l(\tau_{li,K_l}, Z_{li})}{S_l(\tau_{li,K_l-1}, Z_{li})},
\]
\[
= \prod_{k=1}^{K_l} \prod_{j=1}^{J} 1\{I_{l,i,j} \subset I_{l,k}\} \left(\frac{S_l(\tau_{li,k})}{S_l(\tau_{li,k-1})}\right)^{\exp\{\beta_{l,j}^\prime Z_{li,j}\}}
\]
and the conditional observation probability of $i$ on $I_{l,k}$ is

$$p_{l,k}(Z_{li}) = S_l(\tau_{l,k-1}, Z_{li}(\tau_{l,k})) - S_l(\tau_{l,k}, Z_{li}(\tau_{l,k})),
= \prod_{j=1}^{J} \{I_{li,j} \subset I_{l,k}\}\exp(\beta_{l,k} Z_{li}).$$

3. Identifiability and estimation of the parameters

3.1. Model without covariates

Without covariates the parameters are only the probabilities $p_{l,k}$ and $p_l(\tau_{l,K_l})$. Assuming that the observations on the different intervals are independent, the model is multinomial and the probabilities of independent observations on the $K_l + 1$ intervals are written with the differences $\Delta_{l,k} = \Delta \Lambda_{l,k}(I_{l,k}) > 0$, $1 \leq k \leq K_l$.

$$1 - p_l(\tau_{l,K_l}) = \sum_{k \leq K_l} p_{l,k},$$

$$\log(1 - p_l(\tau_{l,k})) = \sum_{k' \leq k} \log S_l(\tau_{l,k'-1}) - \log S_l(\tau_{l,k'}) = - \sum_{k' \leq k} \Delta_{l,k'}, \quad (3)$$

$$\log p_{nl,k} = \log S_l(\tau_{l,k-1}) - S_l(\tau_{l,k}) = \log(1 - \exp(-\sum_{k' \leq k} \Delta_{l,k'})).$$

The log-likelihood for class $C_l$ is

$$l_n(l) = \sum_{i=1}^{n_l} \sum_{k \leq K_l} \{\delta_{li,k} \log p_{l,k} + (1 - \delta_{li,k}) \log(1 - p_{l,k})\}$$

under (2) and the MLE of the parameters $p_{l,k}$ and the function $S_l$ are

$$\hat{p}_{nl,k} = n_l^{-1} \sum_{i=1}^{n_l} \delta_{li,k}, \quad \hat{p}_{nl}(\tau_{l,K_l}) = 1 - n_l^{-1} \sum_{i=1}^{n_l} \delta_{li,k},$$

$$\hat{S}_{nl}(\tau_{l,k}) = \hat{S}_{nl}(\tau_{l,k-1}) - \hat{p}_{nl,k} = 1 - n_l^{-1} \sum_{i=1}^{n_l} \delta_{li,k'}.$$

The estimator $\hat{S}_{nl}$ is decreasing with weights at the sampling times $\tau_{l,k}$. From (3), the differences $\Delta_{l,k}$ satisfy

$$\Delta_{l,k} = \log \frac{1 - \sum_{k' \leq k} \hat{p}_{l,k}}{1 - \sum_{k' \leq k} \hat{p}_{l,k}} > 0,$$

their estimators are deduced from the $\hat{p}_{nl,k}$’s and the cumulative hazard function for $C_l$ is estimated by

$$\hat{\Lambda}_{nl}(t) = \sum_{k=1}^{K_l} 1\{\tau_{l,k-1} < t \leq \tau_{l,k}\} \log \frac{1 - \sum_{k' \leq k} \hat{p}_{nl,k}}{1 - \sum_{k' \leq k} \hat{p}_{nl,k}}. \quad (4)$$
Let $p_{0l,k}$, $S_{0l}$ and $\Lambda_{0l}$ be the actual values of the model parameters, then

**Proposition 3.1** The estimators $\hat{p}_{nl,k}$, $\hat{\Lambda}_{nl}$ and $\hat{S}_{nl}$ are a.s. consistent as $n \to \infty$, $n^{1/2}(\hat{p}_{nl,k} - p_{0l,k})$ converge to centered Gaussian variable with covariances $n^{-1}p_{0l,k}(1 - p_{0l,k})$ and zero otherwise, and the processes $n^{1/2}(\hat{S}_{nl} - S_{0l})$ and $n^{1/2}(\hat{\Lambda}_{nl} - \Lambda_{0l})$ converge to centered Gaussian process with independent increments and variances

$$n_l E[(\hat{S}_{nl} - S_{0l})^2(\tau_{l,k})] = \sum_{k' < k} p_{0l,k'}(1 - p_{0l,k'}),$$

$$n_l E[(\hat{\Lambda}_{nl} - \Lambda_{0l})^2(\tau_{l,k})] = \sum_{k' < k} p_{0l,k'}(1 - p_{0l,k'}) \left( \frac{p_{0l,k}}{\Pr(T_{li} > \tau_{l,k-1}) \Pr(T_{li} > \tau_{l,k})} \right)^2 + p_{0l,k}(1 - p_{0l,k}) \left( \frac{1}{\Pr(T_{li} > \tau_{l,k})} \right)^2.$$

### 3.2. Models with covariates

The parameters of the model are the probabilities $p_l$ and $p_{l,k} = p_l(I_{l,k})$, or the functions $p_l(z)$ and $p_{l,k}(z) = p_l(I_{l,k}, z)$ in regression model. The probabilities $p_l$ are expressions of the $p_{l,k}$'s and of the distribution of the covariates, their estimators satisfy

$$\hat{p}_{nl,k} = \sum_{j=1}^J \hat{p}_{nl,k}(Z_{l,j}) \hat{p}_{nl}(Z_{l,j}),$$

$$\hat{p}_l = \sum_{k=1}^{K_l} \sum_{j=1}^J \hat{p}_{nl,k}(Z_{l,j}) \hat{p}_{nl}(Z_{l,j})$$

but the distributions $p_l$ are not directly estimable since all the individuals are not observed. Only the probabilities $\Pr(Z_{li} \leq z | \delta_{li,k} = 1)$ are directly estimable as the proportion of the individuals observed in $I_{lk}$ such that $Z_{li} \leq z$. Then $P_l(z)$ is deduced from the equation

$$P_l(z) = \frac{\sum_{j=1}^J \Pr(Z_{li} \leq z | \delta_{li,k} = 1) \Pr(\delta_{li,k} = 1)}{\sum_{j=1}^J \Pr(\delta_{li,k} = 1 | Z_{li} \leq z)}, \forall i = 1, \ldots, n$$

which is easily estimated with the empirical probabilities.

The estimable parameters are always the values of the functions $S_l$ and $\Lambda_l$ at the observation times $\tau_{l,k}$ and model parameters when it is appropriate. Conditionally on the covariates, the log-likelihood for class $C_l$ is

$$l_n(l) = \sum_{i=1}^{n} \sum_{k \leq K_l} \{ \delta_{li,k} \log p_{l,k}(Z_{li}) + (1 - \delta_{li,k}) \log(1 - p_{l,k}(Z_{li})) \}.$$
\[\begin{align*}
&= \sum_{i=1}^{n_i} \sum_{k \leq K_i} \sum_{j=1}^{J} 1 \{I_{i,j} \cap I_{l,k}\} \{\delta_{i,k} \log p_{i,k}(Z_{l,j}) \\
&\quad + (1 - \delta_{i,k}) \log (1 - p_{i,k}(Z_{l,j}))).
\end{align*}\]

The MLEs are identical to the previous estimators if the covariates are on the intervals \(I_{l,k}\) and \(p_{i,k}(Z_{l,i}) \equiv p_{l,k}\). If \(J\) is finite, and the variations of the processes \(Z_{l,i}\) are observed though those of \(N_{l,i}\) are only observed on \(I_{l,k}, i = 1, \ldots, n\), they are modified

\[\widehat{p}_{nl,k}(Z_{l,j}) = n_i^{-1} \sum_{i=1}^{n_i} \delta_{i,k} 1 \{I_{i,j} \cap I_{l,k}\},\]

\[\widehat{S}_{nl}(\tau_{l,k}, Z_{l,j}) = 1 - n_i^{-1} \sum_{i=1}^{n_i} \sum_{k' = 1}^{k} \delta_{i,k'} 1 \{I_{i,j} \cap I_{l,k}\},\]

\[\widehat{S}_{nl}(\tau_{l,k}, z) = 1 - n_i^{-1} \sum_{i=1}^{n_i} \sum_{k' = 1}^{k} \delta_{i,k'} \sum_{j=1}^{J} 1 \{Z_{l,j} = z\} 1 \{I_{l,i} \cap I_{l,k}\},\]

\[\widehat{\Lambda}_{nl}(t, z) = \sum_{k=1}^{K} \sum_{j=1}^{J} 1 \{t \in I_{l,i} \cap I_{l,k}\} 1 \{Z_{l,j} = z\} \log \frac{1 - \sum_{k' < k} \hat{p}_{nl,k}(z)}{1 - \sum_{k' \leq k} \hat{p}_{nl,k}(z)}.\]

With continuous covariate and under (1), kernel estimators of the functions conditionally on \(z\) are defined with a kernel \(K\), a bandwidth \(h\) and \(K_h(z) = h^{-1} K(h^{-1} x)\), by smoothing these estimators or the previous ones

\[\widehat{p}_{nl,k}(z) = \frac{\sum_{i=1}^{n_i} K_h(z - Z_{l,i}(\tau_{l,k})) \delta_{i,k}}{\sum_{i=1}^{n_i} K_h(z - Z_{l,i}(\tau_{l,k}))},\]

\[\widehat{S}_{nl}(\tau_{l,k}, z) = 1 - \sum_{k' = 1}^{k} \hat{p}_{nl,k'}(z),\]

\[\widehat{\Lambda}_{nl}(t, z) = \sum_{k=1}^{K} \sum_{i=1}^{n_i} \frac{K_h(z - Z_{l,i}(\tau_{l,k})) \delta_{i,k}}{\sum_{i=1}^{n_i} K_h(z - Z_{l,i}(\tau_{l,k}))} \times \sum_{j=1}^{J} 1 \{t \in I_{l,i} \cap I_{l,k}\} \log \frac{1 - \sum_{k' < k} \hat{p}_{nl,k}(z)}{1 - \sum_{k' \leq k} \hat{p}_{nl,k}(z)}\]

and they converge at the usual rate of the kernel estimators if the bandwidth tends to zero at the optimal rate \(n^{-\frac{1}{p+1}}\), for a \(p\)-dimensional covariate having a density with a \(s\)-order derivative.

For estimation in the proportional hazards model with constant covariates \(Z_{l,i,k}\) on \(I_{l,i,k}\), let \(\omega_{l,i,k} = \exp(\beta_{l,i,k} Z_{l,i,k})\), \(\Omega_l = \{\omega_{l,i,k}\}_{i \leq n, k \leq K_l}\),

\[\log \Delta S_l(I_{l,k}) = \log S_l(\tau_{l,k-1}) + \log \left\{ 1 - \frac{S_l(\tau_{l,k})}{S_l(\tau_{l,k-1})} \right\} .\]
If Proposition 3.2 same covariate value as $Z$ previously, $l$ and estimators are defined by

\[ \log p_{l,k}(Z_{l,k}) = \omega_{l,k} \log \Delta S_{l}(I_{l,k}) = -\omega_{l,k} \{ \sum_{k'<k} \Delta_{l,k'} - \log(1 - e^{-\Delta_{l,k}}) \}. \]

Denote $\mu_{l,k} = \log \Delta S_{l}(I_{l,k}) = \log p_{l}(I_{l,k})$, then the estimator of $p_{l,k}(Z_{l,k}) = \exp\{\omega_{l,k}h_{l,k}\}$ of proposition 3.1 has to be restricted to the individuals with the same covariate value as $Z_{l,k}$.

**Proposition 3.3** If $\Omega_l$ is a finite set $\{\omega_{l,j}\}_{j=1,...,J}$, then

\[ \omega_{l,j} = \log \frac{p_{l}(I_{l,k}, Z_{l,k})}{p_{l}(I_{l,k})}, \]

and estimators are defined by

\[ \hat{p}_{nl}(I_{l,k}, Z_{l,j}) = \frac{\sum_{i \leq n_l} 1 \{ \omega_{l,i,k} = \omega_{l,j} \} \delta_{l,i,k}}{\sum_{i \leq n_l} 1 \{ \omega_{l,i,k} = \omega_{l,j} \}}, \]

\[ \hat{\mu}_{nl,k} = \log \hat{p}_{nl,k} = \log \{ n_l^{-1} \sum_{i=1}^{n_l} \delta_{l,i,k} \}, \]

\[ \hat{\omega}_{nl,j} = \log \frac{n_l(\sum_{i \leq n_l} 1 \{ \omega_{l,i,k} = \omega_{l,j} \} \delta_{l,i,k})}{(\sum_{i \leq n_l} \delta_{l,i,k})(\sum_{i \leq n_l} 1 \{ \omega_{l,i,k} = \omega_{l,j} \})}, \]

\[ \hat{S}_{nl}(\tau_{l,k}, Z_{l,j}) = 1 - \frac{\sum_{i \leq n_l} \sum_{k'=1}^{k} 1 \{ \omega_{l,i,k} = \omega_{l,j} \} \delta_{l,i,k}}{\sum_{i \leq n_l} 1 \{ \omega_{l,i,k} = \omega_{l,j} \}}. \]

An estimator of $\Lambda_{l}(\tau_{l,k}, Z_{l,j})$ is deduced from the $\hat{p}_{nl}(I_{l,k}, Z_{l,j})$'s and (3) as previously,

\[ \hat{\Lambda}_{nl}(t, Z_{l,j}) = \sum_{k=1}^{K} 1 \{ \tau_{l,k-1} < t \leq \tau_{l,k} \} \log \frac{1 - \sum_{k'<k} \hat{p}_{nl}(I_{l,k}, Z_{l,j})}{1 - \sum_{k'<k} \hat{p}_{nl}(I_{l,k}, Z_{l,j})} \]

and the results of Proposition 3.1 extend to these estimators.

Let $p_{0l,k}$, $S_{0l}$ and $\Lambda_{0l}$ be the actual values of the model parameters, then

**Proposition 3.3** The estimators $\hat{p}_{nl,k}$, $\hat{\Lambda}_{nl,k}$ and $\hat{S}_{nl}$ are a.s. consistent as $n \to \infty$, $n_l^{1/2}(\hat{p}_{nl,k} - p_{0l,k})$ converge to centered Gaussian variable with covariances $n_l^{-1}p_{0l,k}(1 - p_{0l,k})$ and zero otherwise, and the processes $n_l^{1/2}(\hat{S}_{nl} - S_{0l})$ and $n_l^{1/2}(\hat{\Lambda}_{nl} - \Lambda_{0l})$ converge to centered Gaussian process with independent increments and variances

\[ n_l E(\hat{S}_{nl} - S_{0l})^2(\tau_{l,k}) = \sum_{k'<k} p_{0l,k'}(1 - p_{0l,k'}), \]

\[ n_l E(\hat{\Lambda}_{nl} - \Lambda_{0l})^2(\tau_{l,k}) = \sum_{k'<k} p_{0l,k'}(1 - p_{0l,k'}), \]

\[ n_l E(\hat{p}_{nl,k} - p_{0l,k})^2(\tau_{l,k}) = \sum_{k'<k} p_{0l,k'}(1 - p_{0l,k'}), \]

\[ n_l E(\hat{\omega}_{nl,j} - \omega_{l,j})^2(\tau_{l,k}) = \sum_{k'<k} p_{0l,k'}(1 - p_{0l,k'}), \]

\[ n_l E(\hat{\omega}_{nl,j} - \omega_{l,j})^2(\tau_{l,k}) = \sum_{k'<k} p_{0l,k'}(1 - p_{0l,k'}), \]

\[ n_l E(\hat{S}_{nl} - S_{0l})^2(\tau_{l,k}) = \sum_{k'<k} p_{0l,k'}(1 - p_{0l,k'}), \]

\[ n_l E(\hat{\Lambda}_{nl} - \Lambda_{0l})^2(\tau_{l,k}) = \sum_{k'<k} p_{0l,k'}(1 - p_{0l,k'}), \]

\[ n_l E(\hat{p}_{nl,k} - p_{0l,k})^2(\tau_{l,k}) = \sum_{k'<k} p_{0l,k'}(1 - p_{0l,k'}), \]
\[ n_l E(\hat{\Delta}_{nl} - \Lambda_0 l)^2 (\tau_{l,k}) = \sum_{k' < k} p_{0l,k'} (1 - p_{0l,k'}) \left( \frac{p_{0l,k}}{\Pr(T_{il} > \tau_{l,k-1}) \Pr(T_{il} > \tau_{l,k})} \right)^2 \\
+ p_{0l,k} (1 - p_{0l,k}) \left( \frac{1}{\Pr(T_{il} > \tau_{l,k})} \right)^2. \]

The proportional hazards model without finite \( \Omega \) is still parametric but maximum likelihood estimators are not written in closed form. Denoting \( \Delta_{li,j} = \Lambda_l(U_{li,j}) - \Lambda_l(U_{li,j-1}) \), the probabilities are now

\[
\log(1 - p_l(\tau_{l,K_l}, \beta_{k,l}, Z_{l,j})) \\
= \sum_{k=1}^{K_l} \sum_{j=1}^{J} \mathbb{1}\{I'_{li,j} \subset I_{l,k}\} \exp\{\beta'_{l,k} Z_{l,j}\} \{\log S_l(U_{li,j}) - \log S_l(U_{li,j-1})\} \\
= - \sum_{k=1}^{K_l} \sum_{j=1}^{J} \mathbb{1}\{I'_{li,j} \subset I_{l,k}\} \exp\{\beta'_{l,k} Z_{l,j}\} \Delta_{li,j},
\]

\[
\log p_{l,k}(Z_{li}) = \sum_{j=1}^{J} \mathbb{1}\{I'_{li,j} \subset I_{l,k}\} \exp\{\beta'_{l,k} Z_{l,j}\} \log S_l(I'_{li,j}) \\
= \sum_{j=1}^{J} \mathbb{1}\{I'_{li,j} \subset I_{l,k}\} \exp\{\beta'_{l,k} Z_{l,j}\} \{\log S_l(U_{li,j}) - \log S_l(U_{li,j-1})\} + \log \left(1 - \frac{S_l(U_{li,j})}{S_l(U_{li,j-1})}\right) \\
= - \sum_{j=1}^{J} \mathbb{1}\{I'_{li,j} \subset I_{l,k}\} \exp\{\beta'_{l,k} Z_{l,j}\} \{\sum_{j' < j} \Delta_{li,j'} \}
+ \log \left(1 - \exp(- \exp\{\beta'_{l,k} Z_{l,j}\} \Delta_{li,j})\right).\]

When covariate only depend on the observation intervals, the parameters are all identifiable by maximization of the likelihood, as it is the case with continuously observed individuals. The parameters are not identifiable when the covariates vary individually.

### 3.3. Estimation of the sample size

The unknown population size \( \nu \) has to be estimated. For a population of \( L \) observed classes \( C_1, \ldots, C_L \) of respective sizes \( \nu_l \), estimators of the catching or observation probabilities \( p_{l,k} \) would be \( n_{l,k} \nu_l^{-1} \) if \( \nu_l \) was known, \( k = 1, \ldots, K_l \). By inverting this expression after an estimator \( \hat{\nu}_{nl} \) has been defined, the sizes are usually estimated by

\[
\hat{\nu}_{nl} = \frac{n_l}{\hat{p}_{nl}}, \quad l = 1, \ldots, L, \quad \hat{\nu} = \sum_{l=1}^{L} \hat{\nu}_{nl} = \sum_{l=1}^{L} \frac{n_l}{\hat{p}_{nl}}.
\]

With consecutive intervals under the same conditions and with varying catching or observation probabilities \( p_{l,k} \), define a moving average estimator of \( p_{l,k} \) and
mean estimators of classes and population sizes for \( k > a \geq 1 \) by

\[
\hat{p}_{nl,k} = \sum_{k=a}^{b+1} \hat{p}_{nl,k'} \quad \hat{\nu}_{nl} = \sum_{k>a} n_{l,k} \hat{p}_{nl,k}, \quad \hat{\nu}_n = \sum_{l=1}^L \hat{\nu}_{nl}.
\]

The same method applies for covariate dependent probabilities, using the estimators of section 3.2 and (5)-(6).

4. Models with dependent observations on consecutive intervals

4.1. Nonparametric models

When the probability of observing individuals in \( I_{l,k} \) depends on their observation in \( I_{l,k-1} \), several nonparametric models may be considered. Let

\[
\pi_{l,k} = \Pr \{ \tau_{l,k-1} < T_i \leq \tau_{l,k} \},
\]

\[
\pi_{l,k}(Z_{l,i}) = \Pr \{ \tau_{l,k-1} < T_i \leq \tau_{l,k} \mid \tau_{l,k-1} < T_i \leq \tau_{l,k}, Z_{l,i} \},
\]

then

\[
p_{l,k,k+1} = \Pr \{ \tau_{l,k-1} < T_i \leq \tau_{l,k+1} \} = \pi_{l,k} p_{l,k}
\]

and conditionally on \( Z_{l,i} \),

\[
p_{l,k,k+1}(Z_{l,i}) = \pi_{l,k}(Z_{l,i}) p_{l,k}(Z_{l,i}).
\]

The estimators are now defined for joint intervals,

\[
\hat{\pi}_{nl,k} = \sum_{i=1}^{n_l} \delta_{l_i,k} \delta_{l_i,k+1},
\]

\[
\hat{p}_{nl,k,k+1} = \frac{\sum_{i=1}^{n_l} \delta_{l_i,k} \delta_{l_i,k+1}}{\sum_{i=1}^{n_l} \delta_{l_i,k}},
\]

\[
\hat{p}_{nl,k,k+1}(Z_{l,j}) = \sum_{i=1}^{n_l} \delta_{l_i,k} \delta_{l_i,k+1} I_{l_i,j} \subset I_{l,k} \cup I_{l,k+1},
\]

\[
\hat{\pi}_{nl,k}(Z_{l,j}) = \sum_{i=1}^{n_l} \delta_{l_i,k} \delta_{l_i,k+1} I_{l_i,j} \subset I_{l,k} \cup I_{l,k+1}.
\]

All the other models and estimators of section 4.1 are generalized by the same method. In the model without covariates, a test for the hypothesis \( H_0 \) of independence between intervals \( I_{l,k} \) and \( I_{l,k+1} \) is a test for \( p_{l,k,k+1} = p_{l,k} p_{l,k+1} \) or \( \pi_{l,k} = \pi_{l,k+1} \).

**Proposition 4.1** Under \( H_0 \), the statistic

\[
Z_l = \sum_{k=1}^{K_l-1} \left( \frac{\hat{p}_{nl,k} \hat{p}_{nl,k+1} - \hat{p}_{nl,k,k+1}}{\hat{p}_{nl,k} \hat{p}_{nl,k+1}} \right)^2
\]

converges to a \( \chi^2_{(K_l-2)} \) as \( n_l \to \infty \).
Proof. Let \( N_{i,k} = \sum_{i=1}^{n_i} N_{ii}(I_{i,k}) \), \( N_{i,k,k+1} = \sum_{i=1}^{n_i} N_{ii}(I_{i,k} \cup I_{i,k+1}) \) and

\[
Z_l = \sum_{k=1}^{K_l-1} \frac{(N_{i,k,k+1} - n_l^{-1} N_{i,k} N_{i,k+1})^2}{N_{i,k} N_{i,k+1}}
\]

is the test statistic for independent marginals in a two-dimensional array.

### 4.2. Markov models

As the individual classes change during the observation period, a second class index may be incorporated in the model to take into account the evolution. Let \( C_{i,T_i} \) denote the class at \( T_i \) for some observation time \( T_i \) of individual \( i \),

\[
\eta_{l',i} = 1\{C_{i,T_i} = C_l, C_{i,T_i} = C_{l'}\},
\]

\[
p_{l'|l,} = p_{l'|l}(I_{i,k}) = \Pr\{T_i \in I_{i,k}, C_{i,T_i} = C_l'\},
\]

\[
S_{l'|l,} = \Pr\{T_i \in I_{i,k}, T_i \geq t, C_{i,T_i} = C_l'\},
\]

\[
A_{l'|l,} = h^{-1}\lim_{h \to 0} \Pr\{T_i \in [t, t+h), C_{i,T_i} = C_l'\}.
\]

The likelihood is proportional to

\[
\prod_{l=1}^{K_l} \prod_{k=1}^{n_l} \prod_{i=1}^{L} \prod_{l'=1}^{L} \{p_{l'|l,}(1 - p_{l'|l,})^{1 - \eta_{l',i}}\}^{\eta_{l',i}}
\]

and the estimators become

\[
\hat{p}_{nl'|l,} = \frac{\sum_{j=1}^{n_l} \delta_{l',k}\eta_{l',i}}{\sum_{j=1}^{n_l} \eta_{l',i}},
\]

\[
\hat{S}_{nl'|l,}(\tau_{l,k}) = 1 - \frac{\sum_{j=1}^{n_l} \sum_{k'=1}^{k} \delta_{l',k'}\eta_{l',i}}{\sum_{j=1}^{n_l} \eta_{l',i}},
\]

\[
\hat{A}_{nl'|l,}(t) = \sum_{k=1}^{K_l}\sum_{j=1}^{n_l} 1\{\tau_{l,k-1} < t \leq \tau_{l,k}\} \log \frac{1 - \sum_{k' < k} \hat{p}_{nl'|l,k}}{1 - \sum_{k' < k} \hat{p}_{nl'|l,k}}.
\]

The extension to models and estimators with covariates follows easily from section 3.2. A test for the hypothesis \( H_0 \) of independence between observation and the variation between classes is a test for \( p_{l'|l,} = p_{i,k} \Pr\{C_{i,T_i} = C_l|C_{i,T_i} = C_{l'}\} \) for every \( l, l' = 1, \ldots, L \) and \( k = 1, \ldots, K_l \).

Let \( q_{l'} = \Pr\{C_{i,T_i} = C_l, C_{i,T_i} = C_{l'}\} \), then the estimators

\[
\hat{q}_{nl'|l,} = \frac{\sum_{j=1}^{n_l} \eta_{l',i}}{n_l},
\]

\[
\hat{p}_{nl'|l,} = \frac{\hat{p}_{nl'|l,} \hat{q}_{nl'|l,}}{\hat{q}_{nl'|l,}},
\]

provide a test statistic.
Proposition 4.2 Under $H_0$, the statistic

$$X_l = \sum_{k=1}^{K_l} \sum_{l=1}^{L} \frac{(\hat{p}_{nl,k} - \hat{p}_{nl,k} \tilde{q}_{nl})^2}{\hat{p}_{nl,k} \tilde{q}_{nl}}$$

converges to a $\chi^2_{(K_l-1)(L-1)}$ as $n_l \to \infty$.

References

[1] Chao, A. and Lee, S.M. (1992). Estimating population size for capture-recapture data when capture probability vary by time and individual. *Biometrics*, 48, 201–216.

[2] Groeneboom, P. (1985). Estimating a monotone density. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol.II, L.M. LeCam and R.A. Olsen eds.

[3] Groeneboom, P. and Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Basel.

[4] Huang, J. (1996). Efficient estimation for the Cox model with interval censoring. *Ann.Statist.*, 24, 540–568.

[5] Pons, O. (2005) Semi-Parametric Estimation for a Semi-Markov Process with Left-Truncated and Right-Censored Observations. *Statist. Probab. Lett.*, 76, 952–958.

[6] Hwang, W-H., Chao, A. and Yip, P. (2002). Continuous-time capture-recapture models with time variation and behavioural response. *Aust.N.Z.Stat.*, 44, 41–54.

[7] Yip, P.S.F., Huggins, R.M. and Lin, D.Y. (1996). Inference for capture-recapture experiments in continuous time with variable capture rates. *Biometrika*, 83, 477–483.