Choice between Semi-parametric Estimators of Markov and Non-Markov Multi-state Models from Coarsened Observations

Author(s): DANIEL COMMENGES, PIERRE JOLY, ANNE GÉGOUT-PETIT and BENOIT LIQUET

Source: *Scandinavian Journal of Statistics*, March 2007, Vol. 34, No. 1 (March 2007), pp. 33-52

Published by: Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics

Stable URL: https://www.jstor.org/stable/41548537

doi: 10.1111/j.1467-9469.2006.00536.x
 Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA Vol 34: 33–52, 2007

Choice between Semi-parametric Estimators of Markov and Non-Markov Multi-state Models from Coarsened Observations

DANIEL COMMENGES, PIERRE JOLY

Unité INSERM de Biostatistique, Université Victor Segalen Bordeaux 2

ANNE GÉGOUT-PETIT

Département Science et Modélisation, Université Victor Segalen Bordeaux 2

BENOIT LIQUET

Laboratoire Statistique et Analyse de données, Université de Grenoble

ABSTRACT. We consider models based on multivariate counting processes, including multi-state models. These models are specified semi-parametrically by a set of functions and real parameters. We consider inference for these models based on coarsened observations, focusing on families of smooth estimators such as produced by penalized likelihood. An important issue is the choice of model structure, for instance, the choice between a Markov and some non-Markov models. We define in a general context the expected Kullback–Leibler criterion and we show that the likelihood-based cross-validation (LCV) is a nearly unbiased estimator of it. We give a general form of an approximate of the leave-one-out LCV. The approach is studied by simulations, and it is illustrated by estimating a Markov and two semi-Markov illness–death models with application on dementia using data of a large cohort study.

Key words: counting processes, cross-validation, dementia, interval-censoring, Kullback–Leibler loss, Markov models, multi-state models, penalized likelihood, semi-Markov models

1. Introduction

Multi-state models, and more generally models based on multivariate counting processes, are well adapted for modelling complex event histories (Andersen *et al.*, 1993; Hougaard, 2000). Assumptions have to be made about the law of the processes involved. In particular, the Markov assumption has been made in many applications (Aalen & Johansen, 1978; Joly *et al.*, 2002) while semi-Markov models have been considered in other applications (Joly & Commenges, 1999). Subject-matter knowledge can be a guide for making these assumptions [for instance, risk of AIDS essentially depends on time since infection, leading Joly & Commenges (1999) to choose a semi-Markov model]; however, in many cases the choice is not obvious. Other assumptions have to be made relative to the influence of explanatory variables: multiplicative or additive structures, for instance, may be considered. The problem is generally not to assert whether the 'true' model is Markov or has a multiplicative structure but to choose the best model relative to the data at hand.

The semi-parametric approaches offer the greatest flexibility. Aalen (1978) has studied non-parametric inference for counting processes. If we wish to estimate smooth intensities, we have to consider families of estimators such as kernel estimators (Ramlau-Hansen, 1983), sieve-estimators (Kooperberg & Clarkson, 1997) or penalized likelihood estimators (Good & Gaskin, 1971; O'Sullivan, 1988; Joly *et al.*, 2002). These families are indexed by a parameter

that we may call 'smoothing coefficient'. A practical way for choosing the smoothing coefficient is by optimizing a cross-validation criterion. In particular, likelihood cross-validation (LCV) has been shown to have good properties in simulation (Liquet *et al.*, 2003; Liquet & Commenges, 2004), while it has been shown that in some cases it could be considered as a proxy for the expected Kullback–Leibler loss and had the optimal property of being asymptotically as efficient as this theoretical criterion (Hall, 1987; van der Laan *et al.*, 2004). Liquet *et al.* (2006) argued that LCV could be used not only for choosing the smoothing coefficient but also for choosing between different semi-parametric models, such as stratified and non-stratified proportional hazard survival models.

Additional complexity comes from the fact that the model must be estimated from incomplete data. Coarsening mechanisms have been studied in a general context by Gill *et al.* (1997). Commenges & Gégout-Petit (2005) have studied a general time-coarsening model for processes which they called GCMP; we will use this coarsening process under the name TCMP for 'time-coarsening model for processes'; in effect it is not completely general because it assumes that there are times where the process is exactly observed; this is most often the case for counting processes but not for more general processes. Even for counting processes, the TCMP does not include the filtering process of Andersen *et al.* (1993) in a natural way. Writing the likelihood for observations of multi-state models through the TCMP has been done by Commenges & Gégout-Petit (2006).

The aim of this paper is to advocate the use of the expected Kullback–Leibler risk, EKL, based on the observation, for the choice between semi-parametric estimators for coarsened observations. We also advocate the use of LCV as an estimator of EKL. Thus, LCV can be used in particular for choosing between estimators of Markov and non-Markov multi-state models or between multiplicative and additive models in the presence of generally coarsened observations. It is worth noting that the LCV choice fits well with using families of smooth estimators, such as produced by penalized likelihood, because non-smooth estimators are strongly rejected by this criterion.

In section 2, we recall the description of multi-state models as multivariate counting processes and suggest possible Markov and non-Markov structures for the illness-death model. In section 3, we recall the construction of the likelihood ratio for counting processes and its extension to penalized likelihood and we unify the problem of choice of smoothing coefficients and model structure. In section 4, we tackle the problem of likelihood ratio and penalized likelihood in the TCMP framework. In section 5, we define the expected Kullback–Leibler loss as a general criterion for choosing an estimator in a family of estimators based on generally coarsened observations; we also study the case where there are observed explanatory variables. In section 6, it is proposed to use LCV as a proxy for this theoretical criterion and we give a general approximation of the leave-one-out LCV. In section 7, we present a simulation study in which we study in particular the variability of LCV and we give insight into the interpretation of a difference of LCV. This approach is then applied in section 8 for choosing and estimating an illness–death model for dementia, based on the data of a large cohort study; section 9 concludes.

2. Multi-state and counting process models: illness–death model

2.1. Multi-state and counting process models

A multi-state process $X=(X_t)$ is a right-continuous process which can take a finite number of values $\{0,1,\ldots,K\}$. If the model is Markov it can be specified by the transition intensities $\alpha_{hj}(\cdot), h, j=0,\ldots,K$. The correspondence between multi-state processes and multivariate counting processes was studied in Commenges & Gégout-Petit (2006) where the advantage

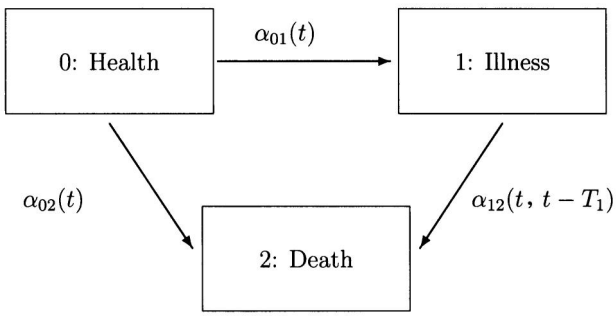

Fig. 1. The illness–death model, t: age; T_1: age at onset of illness.

of representing multi-state processes as multivariate 'basic' counting processes is highlighted. Multi-state models are often generated by p types of events, each type occurring just once. For instance, the three-state illness–death model (see Fig. 1) is generated by considering the events 'illness' and 'death'; the five-state model considered by Commenges & Joly (2004) is generated by 'dementia', 'institutionalization' and 'death'. So these multi-state models can be represented by a p-variate counting process $N=(N_j, j=1,\ldots,p)$, each N_j making at most one jump, and we will denote T_j the jump time of N_j.

2.2. *Possible semi-parametric models*

A model for a multivariate counting process $N=(N_j, j=1,\ldots,p)$ is specified for a given filtration $\{\mathcal{F}_t\}$ by its intensity $\lambda^\theta(t)=(\lambda_j^\theta(t), j=1,\ldots,p)$ under P_θ.

For efficient inference one has to make assumptions, often the Markov assumption is made: the process is Markov if and only if $\lambda^\theta(t)$ is a function only of time t and the indicator functions $1_{\{T_j<t\}}, j=1,\ldots,p$. An interesting non-Markov model occurs if $\lambda^\theta(t)$ depends on the time elapsed since the last jump of one of the components of N. If the intensities do not depend on time t itself this defines a particular semi-Markov model (used, for instance, by Lagakos *et al.*, 1978) that we will call 'current-state' model because the transition intensities depend only on the time spent in the current state.

Completely parametric models are often too rigid but parametric assumptions may be made for some parts of the model: thus a semi-parametric approach, in which a great flexibility is preserved on some part of the model while some parametric assumptions are made for simplicity and easier interpretation, is often attractive. In such an approach, $\lambda^\theta(t)$ will depend on a certain number of functions on which no parametric assumptions are made, some of them representing baseline transition intensity functions, and parameters which appear in modelling how these baseline intensities will be changed as a function of events that have happened.

Let us consider some possible three-state irreversible illness–death models; these models are Markov or semi-Markov. Any model of this type is described by a bivariate counting process, N_1 counting illness and N_2 counting death. The intensity of N_1 necessarily takes the form

$$\lambda_1^\theta(t)=1_{\{T_1\geq t\}}1_{\{T_2\geq t\}}\alpha_{01}(t),$$

where $\alpha_{01}(t)$ has the interpretation of the transition intensity toward illness. The intensity of N_2 can generally be written

$$\lambda_2^\theta(t)=1_{\{T_2\geq t\}}[1_{\{T_1\geq t\}}\alpha_{02}(t)+1_{\{T_1<t\}}\alpha_{12}(t, t-T_1)].$$

The function $\alpha_{02}(t)$ has the interpretation of the transition intensity from health toward death (the mortality rate of healthy subjects). To avoid having to estimate non-parametrically a bivariate function, we may consider models in which $\alpha_{12}(t, t-T_1)$ depends on two univariate functions $h(t)$ and $g(t-T_1)$; for instance, we may consider an additive model $\alpha_{12}(t, t-T_1) = h(t)+g(t-T_1)$ as in Scheike (2001).

Particular cases of this model are:

$\mathcal{M}_1$: $g(t)=0$: non-homogeneous Markov model; here $h(t)$ has the interpretation of $\alpha_{12}(t)$, the transition intensity from illness toward death;

$\mathcal{M}_2$: $h(t)=0$: current-state model; here $g(t)$ has the interpretation of a random transition intensity from illness toward death;

$\mathcal{M}_3$: $h(t)=\alpha_{02}(t)$: excess mortality model; here $g(t-T_1)$ has the interpretation of an excess mortality because of illness as a function of time passed in the illness state.

If explanatory variables $Z_i(t)$ for subject i are available, we may consider different models for the dependence of the intensities on the $Z_i(t)$ [the variables are either external or internal and in the latter case the filtration must be rich enough so that the processes $(Z_i(t))$ are adapted]; in particular a multiplicative structure (in the spirit of the proportional hazard model) or an additive structure (in the spirit of the Aalen additive model: Aalen *et al.*, 2001) could be considered. For instance, a multiplicative structure for the explanatory variables could be:

$$\alpha^i_{01}(t) = \alpha^0_{01}(t)\exp(\beta_{01}Z_i(t)),$$

$$\alpha^i_{02}(t) = \alpha^0_{02}(t)\exp(\beta_{02}Z_i(t)),$$

$$\alpha^i_{12}(t, t-T_1) = \alpha^0_{12}(t, t-T_1)\exp(\beta_{12}Z_i(t)),$$

where $\alpha^0_{01}(t)$, $\alpha^0_{02}(t)$ and $\alpha^0_{12}(t, t-T_1)$ are baseline transition intensities (the last one being generally random and in that case defined only on $\{t>T_1\}$).

3. Likelihood and penalized likelihood for counting processes

3.1. Likelihood ratios

The model specifies a family of probability measures $\{P_\theta\}_{\theta\in\Theta}$; consider a reference probability measure P_0 such that each P_θ is absolutely continuous relative to P_0 (P_0 may or may not belong to $\{P_\theta\}_{\theta\in\Theta}$). The likelihood ratio on a σ-field $\mathcal{X}$ is defined by:

$$\mathcal{L}^{P_\theta/P_0}_{\mathcal{X}} = \frac{dP_\theta}{dP_{0\,|\mathcal{X}}} \quad \text{a.s.},$$

where

$$\frac{dP_\theta}{dP_{0\,|\mathcal{X}}}$$

is the Radon–Nikodym derivative of P_θ relatively to P_0 on $\mathcal{X}$.

Remark 1. All equalities involving likelihood ratios or conditional expectations are a.s. equalities; this may not be recalled every time.

Remark 2. Often likelihoods are computed using a reference measure that is not a probability measure and is not even specified. Here we will make it explicit and take a probability measure, in which case the term 'likelihood ratio' is warranted. If the reference probability P_0 belongs to $(P_\theta)_{\theta\in\Theta}$ then there exist θ_0 such that $P_0 = P_{\theta_0}$ and we may write $\mathcal{L}^{\theta/\theta_0}_{\mathcal{X}} = \mathcal{L}^{P_\theta/P_0}_{\mathcal{X}}$.

One of the advantages of representing multi-state models in the framework of counting processes (such as in section 2.1) is the availability of Jacod's formula for the likelihood ratio based on observation on $[0, C]$ in the filtration $\{\mathcal{G}_t\}$ where $\mathcal{G}_t = \mathcal{G}_0 \vee \mathcal{N}_t$, where $\mathcal{N}_t = \sigma(N_{ju}, j = 1, \dots p,\ 0 \le u \le t)$. The model is specified by the intensities $\lambda_j^\theta(t)$ of the N_js under P_θ. It is advantageous to take as reference probability, a probability P_0 under which the N_js are independent with intensities $\lambda_j^0(t) = 1_{\{N_{jt-}=0\}}$; equivalently, the T_js are independent with exponential distributions with unit parameter. Using Jacod's formula (Jacod, 1975) the likelihood ratio for this reference probability is:

$$\mathcal{L}_{\mathcal{G}_C}^{P_\theta/P_0} = \mathcal{L}_{\mathcal{G}_0}^{P_\theta/P_0} \prod_{r=1}^{N_{.C}} \lambda_{J_r}^\theta(T_{(r)}) \exp(-\Lambda_.^\theta(C)) \prod_{j=1}^{p} \mathrm{e}^{T_j \wedge C}, \tag{1}$$

where for each $r \in \{1, \dots, N_{.C}\}$, J_r is the unique j such that $\Delta N_{jT_{(r)}} = 1$; $N_{.t} = \sum_{j=1}^p N_{jt}$, $\Lambda_.^\theta(t) = \sum_{j=1}^p \Lambda_j^\theta(t)$, $\Lambda_j^\theta(t) = \int_0^t \lambda_j^\theta(u)\,\mathrm{d}u$. This formula allows us to compute the likelihood for any multi-state model once we write it as a multivariate counting process.

3.2. Families of penalized likelihood estimators

Consider models specified by a set of parameters $\theta = (g, \beta)$ where $g(\cdot) = (g_j(\cdot), j = 1, \dots, K)$ is a vector of functions from $\Re$ to $\Re$ and β a vector of real parameters. For instance, for the Markov illness–death model $\theta = (\alpha_{01}(\cdot), \alpha_{02}(\cdot), \alpha_{12}(\cdot), \beta)$, where the α_{hj} are transition intensities and β is a vector of regression coefficients. If no parametric assumptions are made about the functions to be estimated and if smooth estimators are favoured, the two main approaches are sieve estimators, extending the so-called hazard regression of Kooperberg & Clarkson (1997) or using orthogonal expansions such as in Müller & Stadtmüller (2005), and penalized likelihood (Gu, 1996; Joly *et al.*, 2002).

Suppose that the sample consists of n independent observations of multivariate counting processes $N^i = (N_j^i, j = 1, \dots, p), i = 1, \dots, n$, represented by $\mathcal{G}_{iC_i}$; the likelihood ratio $\mathcal{L}_{\bar{\mathcal{O}}_n}^{P_\theta/P_0}$, where $\bar{\mathcal{O}}_n = \vee \mathcal{G}_{iC_i}$, is the product of contributions computed with formula (1). A penalized log-likelihood is defined as:

$$pl_{\bar{\mathcal{O}}_n}^\kappa(\theta) = \log\, \mathcal{L}_{\bar{\mathcal{O}}_n}^{P_\theta/P_0} - J(\theta, \kappa), \tag{2}$$

where $\kappa = (\kappa_j, j = 1, \dots, K)$ is a set of smoothing coefficients. It is common to use a penalty based on the L_2-norms of the second derivatives of the unknown functions:

$$J(g(\cdot), \kappa) = \sum_{j=1}^{K} \kappa_j \int (g_j'')^2(u)\,\mathrm{d}u.$$

The penalized likelihood defines a family of estimators of θ, $(\hat{\theta}_\kappa)_{\kappa \in \Re^K{}_+}$, and thus a family of estimators of the probability P_θ, $(P_{\hat{\theta}_\kappa})_{\kappa \in \Re^K{}_+}$. Asymptotic results have been given for particular cases (Cox & O'Sullivan, 1990; Gu, 1996; Eggermont & LaRiccia, 1999, 2001).

Consider now the situation where we can choose between different basic assumptions for our model (such as Markov or semi-Markov assumptions), indexed by $\eta = 1, \dots, m$. Formally, we could include η in θ. However, we prefer to formalize the problem in a way that is closer to intuition and practice: for each value of η and each κ we have a maximum penalized likelihood estimator $\hat{\theta}_\kappa^\eta$; thus, we have a family of estimators of the probability specified by $(P_{\hat{\theta}_\kappa^\eta})_{\eta = 1, \dots, m; \kappa \in \Re^K{}_+}$. The problem is to choose one estimator in this family. In the following, we will include η in κ considering that κ indexes a family of estimators, thus unifying the problem of smoothing coefficient and model structure.

4. Penalized likelihood for coarsened at random counting processes

4.1. Coarsening at random in the TCMP

Here we consider a general case of incomplete data: we recall the TCMP model proposed in Commenges & Gégout-Petit (2005) and we give a version of the coarsening at random condition CAR(TCMP) and the factorization theorem it implies, adapted to the case where the reference probability is outside of the model; also we exhibit a 'reduced' model that we will use in the sequel. The TCMP can be considered for any stochastic process X; when X is a counting process, the TCMP includes in particular extensions of the concepts of right-, left- and interval-censoring that have been defined for survival data, as well as a combination of these different types of censoring.

Definition 1 (The time coarsening model for processes, TCMP)
A TCMP is a scheme of observation for a multivariate process $X=(X_1,\ldots,X_p)$ *specified by a multivariate response process* $R=(R_1,\ldots,R_p)$, *where the* R_{jt}*s take values* 0 *or* 1 *for all* j *and* t, *such that* X_{jt} *is observed at time* t *if and only if* $R_{jt}=1$, *for* $j=1,\ldots,p$; *that is, the observed* σ*-field is* $\mathcal{O}=\sigma(R_t, R.X_t, t\geq 0)$.

In the definition we denote $R.X_t=(R_{1t}X_{1t},\ldots,R_{pt}X_{pt})$. A model for (X,R) is a family of measures $\{P_{\theta\psi}\}_{(\theta,\psi)\in\Theta\times\Psi}$ on a measurable space $(\Omega,\mathcal{F})$. X (resp. R) takes values in a measurable space (Ξ,ξ) (resp. (Γ,ρ)). For us X and R will be p-dimensional càdlàg stochastic processes, so (Ξ,ξ) and (Γ,ρ) are Skorohod spaces endowed with their Borel σ-fields. The parameter spaces Θ and Ψ need not be finite dimensional. We will assume that the measures in the family are equivalent. The processes X and R generate σ-fields $\mathcal{X}$ and $\mathcal{R}$, and we shall take $\mathcal{F}=\mathcal{X}\vee\mathcal{R}$. $P_{\theta\mathcal{X}}$ is the restriction of $P_{\theta\psi}$ to $\mathcal{X}$: that is, the marginal probability of X does not depend on ψ. The additional parameter ψ will be considered as a nuisance parameter. We assume a 'non-informativeness' assumption in the coarsening mechanism, which, writing $P^{\mathcal{X}}_{\theta\psi}=P_{\theta\psi}(\cdot|\mathcal{X})$ (a conventional notation for conditional probabilities: Kallenberg, 2001), is:

$$P^{\mathcal{X}}_{\theta_1\psi}=P^{\mathcal{X}}_{\theta_2\psi}, \quad \text{a.s., for all } \theta_1,\theta_2,\psi. \tag{3}$$

This is a conventional assumption (although it has not been expressed in this form) and means that the coarsening mechanism (conditionally on X) depends on a distinct (from the parameter of interest θ), 'variation independent' parameter ψ. Now we consider a family of equivalent probabilities $\mathcal{Q}$ including in addition to $\{P_{\theta\psi}\}_{(\theta,\psi)\in\Theta\times\Psi}$ a family of possible reference probabilities. Let P_0 be one such probability; we denote by $P_{0\mathcal{X}}$ its restriction to $\mathcal{X}$ and by $P^{\mathcal{X}}_0$ the associated conditional probability given $\mathcal{X}$. The likelihood ratio is then $\mathcal{L}^{P_{\theta\psi}/P_0}_{\mathcal{F}}=\mathcal{L}^{P_{\theta\psi}/P_0}_{\mathcal{R}|\mathcal{X}}\mathcal{L}^{P_{\theta\mathcal{X}}/P_{0\mathcal{X}}}_{\mathcal{X}}$, where $\mathcal{L}^{P_{\theta\psi}/P_0}_{\mathcal{R}|\mathcal{X}}$ is the conditional likelihood of $\mathcal{R}$ given $\mathcal{X}$. Note that with the non-informativeness assumption $\mathcal{L}^{P_{\theta\psi}/P_0}_{\mathcal{R}|\mathcal{X}}$ does not depend on θ; we will note it, $\mathcal{L}^{P_{,\psi}/P_0}_{\mathcal{R}|\mathcal{X}}$, to emphasize this fact. $\mathcal{L}^{P_{\theta\psi}/P_0}_{\mathcal{F}}$ is the full likelihood and $\mathcal{L}^{P_{\theta\psi}/P_0}_{\mathcal{O}}$ the observed likelihood.

Definition 2 [CAR(TCMP)]
We will say that CAR(TCMP) holds for the couple (X,R) *in* $\mathcal{Q}$ *if* $\mathcal{L}^{P_1/P_0}_{\mathcal{R}|\mathcal{X}}$ *is* $\mathcal{O}$*-measurable for all* $P_1,P_0\in\mathcal{Q}$.

We will use the following result (which is an adaptation of Theorem 2 in Commenges & Gégout-Petit, 2005):

Theorem 1 (Factorization)
If the couple (R, X) *satisfies CAR(TCMP) then we have* $\mathcal{L}_{\mathcal{O}}^{P_{\theta\psi}/P_0} = \mathcal{L}_{\mathcal{R}|\mathcal{X}}^{P_{\psi}/P_0} E_{P_0}[\mathcal{L}_{\mathcal{X}}^{P_{\theta\mathcal{X}}/P_{0\mathcal{X}}}|\mathcal{O}]$ *and* $E_{P_0}[\mathcal{L}_{\mathcal{X}}^{P_{\theta\mathcal{X}}/P_{0\mathcal{X}}}|\mathcal{O}]$ *does not depend on* $P_0^{\mathcal{X}}$.

This factorization is the first step toward ignorability because, for instance, it is the same value of θ which maximizes $\mathrm{E}_{P_0}[\mathcal{L}_{\mathcal{X}}^{P_{\theta\mathcal{X}}/P_{0\mathcal{X}}}|\mathcal{O}]$ and which maximizes $\mathcal{L}_{\mathcal{O}}^{P_{\theta\psi}/P_0}$; a slightly stronger condition is necessary for ignorability (see Commenges & Gégout-Petit, 2005).

We can achieve a nicer result which will help simplifying notations in section 5. If we knew the true conditional probability given $\mathcal{X}$, $P_*^{\mathcal{X}}$, we would use the following 'reduced' model: $\{P_\theta\}_{\theta\in\Theta}$ such that $P_\theta^{\mathcal{X}} = P_*^{\mathcal{X}}$.

Definition 3 (Reduced model)
Given a model $\{P_{\theta,\psi}\}_{(\theta,\psi)\in\Theta\times\Psi}$ *such that the restriction of* $P_{\theta,\psi}$ *to* $\mathcal{X}$ *is* $P_{\theta\mathcal{X}}$ *for all* θ *and* ψ, *we call 'reduced model' the model* $\{P_\theta\}_{\theta\in\Theta}$ *such that the restriction of* P_θ *to* $\mathcal{X}$ *is* $P_{\theta\mathcal{X}}$ *and* $P_\theta^{\mathcal{X}} = P_*^{\mathcal{X}}$ *for all* θ, *where* $P_*^{\mathcal{X}}$ *is the true conditional probability given* $\mathcal{X}$.

Note that this reduced model is 'reduced' in the sense that it is a smaller family than the original one; however, it is a submodel only if the original model was well specified ($P_* \in \{P_{\theta,\psi}\}_{(\theta,\psi)\in\Theta\times\Psi}$). Taking a reference probability P_0 such that $P_0^{\mathcal{X}} = P_*^{\mathcal{X}}$ we have that $\mathcal{L}_{\mathcal{R}|\mathcal{X}}^{P_\theta/P_0} = 1$ a.s. Thus we have, without additional assumption, $\mathcal{L}_{\mathcal{O}}^{P_\theta/P_0} = \mathrm{E}_{P_0}[\mathcal{L}_{\mathcal{X}}^{P_{\theta\mathcal{X}}/P_{0\mathcal{X}}}|\mathcal{O}]$. If CAR(TCMP) holds, $\mathrm{E}_{P_0}[\mathcal{L}_{\mathcal{X}}^{P_{\theta\mathcal{X}0}/P_{0\mathcal{X}}}|\mathcal{O}]$ does not depend on $P_0^{\mathcal{X}} = P_*^{\mathcal{X}}$. That is, we can compute the exact likelihood that we would like to compute if we knew $P_*^{\mathcal{X}}$, without in fact knowing it.

There remains in practice to compute $\mathrm{E}_{P_0}[\mathcal{L}_{\mathcal{X}}^{P_{\theta\mathcal{X}}/P_{0\mathcal{X}}}|\mathcal{O}]$ for the observed value of R and this is explained for the case of multi-state processes in section 4.2.

4.2. Likelihood and penalized likelihood for counting processes

The likelihood for the deterministic TCMP (that is, in which the Rs are deterministic functions) and when X is a multivariate counting process has been given in Commenges & Gégout-Petit (2006). The observation in this scheme is denoted by the σ-field $\mathcal{O}$ and we have $\mathcal{O} \subset \mathcal{G}_C$, so that the observed likelihood can be expressed as: $\mathcal{L}_{\mathcal{O}}^{P_\theta/P_0} = \mathrm{E}_{P_0}[\mathcal{L}_{G_C}^{P_\theta/P_0}|\mathcal{O}]$. The formula for the likelihood follows from the computation of this conditional expectation using the disintegration theorem (Kallenberg, 2001).

In the case of the stochastic TCMP, when CAR(TCMP) holds, we use the reduced model $\{P_\theta\}_{\theta\in\Theta}$ so that the likelihood ratio is $\mathcal{L}_{\mathcal{O}}^{P_\theta/P_0} = \mathrm{E}_{P_0}(\mathcal{L}_{G_C}^{P_\theta/P_0}|\mathcal{O})$ and $\mathrm{E}_{P_0}(\mathcal{L}_{G_C}^{P_\theta/P_0}|\mathcal{O})$ does not depend on $P_0^{\mathcal{X}}$; with a slightly stronger condition it can be computed in practice as if the TCMP was deterministic, using the formulae given in Commenges & Gégout-Petit (2006), with values of the responses functions r equal to what has been observed.

As for the penalized likelihood, it can also be extended to the case where the observations are CAR(TCMP): the formula is the same as (2).

5. Choice between semi-parametric models: expected Kullback–Leibler loss

5.1. General theory

The problem of choice of an estimator among a family of estimators using the expected Kullback–Leibler risk has been studied in particular by Hall (1987), van der Laan & Dudoit (2003) and van der Laan *et al.* (2004). As our aim here is to choose an estimator of a probability measure among a family of estimators, this theoretical criterion is particularly relevant. We formalize this criterion in this general context and for incomplete observations.

Although the focus of the paper is on counting processes, the formalism developed in this section applies to more general processes, so we will call the process of interest X as in section 4.1.

Let us now model n independent and identically distributed (i.i.d.) random elements $(X_i, R_i), i=1,\ldots,n$. We consider a measurable space $(\bar{\Omega}_n, \bar{\mathcal{X}}_n, \bar{\mathcal{R}}_n)$ where $\bar{\Omega}_n = \times\Omega_i$, $\bar{\mathcal{X}}_n = \otimes\mathcal{X}_i$, $\bar{\mathcal{R}}_n = \otimes\mathcal{R}_i$ where the σ-fields for different i are independent. The probability measures on this space are the product measures. Finally, we define full and observed σ-fields $\mathcal{F}_i = \mathcal{R}_i \vee \mathcal{X}_i$ and $\mathcal{O}_i = \sigma(R_{it}, R_t.X_{it}, t \geq 0)$ respectively.

The problem is to estimate θ. When Θ is a functional space, a conventional strategy is to define estimators (that is $\bar{\mathcal{O}}_n$-measurable functions) depending on a meta-parameter κ: $\hat{\theta}(\kappa, \bar{\mathcal{O}}_n)$. κ may index nested models such as in sieve estimators or may be a smoothing parameter such as in penalized likelihood or kernel estimators. According to statistical decision theory (Le Cam & Yang, 1990), we should choose an estimator which minimizes a risk function, the expectation of a loss function. In statistics, we assume that there is a true probability P_*. We do not make the assumption that P_* belongs to the model; this assumption would significantly reduce the scope of the theory. Making the assumption that P_* is equivalent to the probabilities of the model, the most natural 'all-purpose' loss function relative to P_* is $-\log \mathcal{L}_{\mathcal{X}_{n+1}}^{P_{\hat{\theta}\mathcal{X}}/P_{*\mathcal{X}}}$ [we write $\hat{\theta} = \hat{\theta}(\kappa, \bar{\mathcal{O}}_n)$], where $P_{*\mathcal{X}}$ is the restriction of P_* to $\mathcal{X}$. However, the problem will be to 'estimate' the risk, that is to find a statistic ($\bar{\mathcal{O}}_n$-measurable) which takes values close to that risk (this is not exactly an estimation problem because the target moves with n but we will use the word 'estimate' for simplicity). It may be considered as intuitive that a risk based on $-\log \mathcal{L}_{\mathcal{X}_{n+1}}^{P_{\hat{\theta}\mathcal{X}}/P_{*\mathcal{X}}}$ will be very difficult to estimate; this is why Liquet & Commenges (2004) suggested using the expectation of the observed loglikelihood of the sample, a criterion they denoted ELL. The use of the stochastic TCMP and the CAR(TCMP) assumption allows us to work in the more comfortable i.i.d. framework for incomplete data leading to more elegant and general results.

The straightforward loss function on $\mathcal{O}_{n+1}$ is $-\log \mathcal{L}_{\mathcal{O}_{n+1}}^{P_{\hat{\theta},\hat{\psi}}/P_*}$. A difficulty arises in that this loss function requires an estimator of ψ. Hopefully, the CAR(TCMP) assumption allows us to construct a reduced model as in section 4.1 in which we can construct a loss function that does not depend on the conditional probability given $\mathcal{X}$, $P_*^{\mathcal{X}}$. The first step is to construct the reduced model associated with the reference probability P_*. Then we have, by applying the result of section 4.1 to that case and to σ-fields $\mathcal{X}_{n+1}, \mathcal{R}_{n+1}, \mathcal{O}_{n+1}$:

$$\mathcal{L}_{\mathcal{O}_{n+1}}^{P_\theta/P_*} = \mathrm{E}_{P_*}[\mathcal{L}_{\mathcal{X}_{n+1}}^{P_{\theta\mathcal{X}}/P_{*\mathcal{X}}} | \mathcal{O}_{n+1}]$$

and this does not depend on $P_*^{\mathcal{X}}$. We can now construct the loss function as

$$-\log \mathcal{L}_{\mathcal{O}_{n+1}}^{P_{\hat{\theta}}/P_*} = -\log \mathrm{E}_{P_*}[\mathcal{L}_{\mathcal{X}_{n+1}}^{P_{\hat{\theta}\mathcal{X}}/P_{*\mathcal{X}}} | \mathcal{O}_{n+1}, \bar{\mathcal{O}}_n];$$

note that we must add the conditioning on $\bar{\mathcal{O}}_n$ because $\hat{\theta}$ is an $\bar{\mathcal{O}}_n$-measurable random variable.

The conditional expectation of this loss, or conditional risk, is

$$\mathrm{CKL}_n = \mathrm{E}_{P_*}[-\log \mathcal{L}_{\mathcal{O}_{n+1}}^{P_{\hat{\theta}}/P_*} | \bar{\mathcal{O}}_n]$$

and can be interpreted as the Kullback–Leibler divergence between $P_{\hat{\theta}}$ and P_*, as the Kullback–Leibler divergence of a probability P_1 relatively to P_* over the σ-field $\mathcal{O}_{n+1}$ is

$$\mathrm{KL}(P_1, P_*) = \mathrm{E}_{P_*}[-\log \mathcal{L}_{\mathcal{O}_{n+1}}^{P_1/P_*}].$$

Its expectation, or risk,

$$\mathrm{EKL}_n = \mathrm{E}_{P_*}[-\log \mathcal{L}_{\mathcal{O}_{n+1}}^{P_{\hat{\theta}}/P_*}],$$

can be interpreted as the expected Kullback–Leibler divergence over the σ-field $\mathcal{O}_{n+1}$ of interest.

5.2. Case of observed explanatory variables

Explanatory variables are considered as stochastic processes $Z=(Z_t)_{t\geq 0}$. We can then consider the process $W=(X,Z)$. In the TCMP framework, we associate the response process $R=(R_X,R_Z)$. The observed σ-field is $\mathcal{O}=\sigma(R_t, R.W_t, t\geq 0)$. With such a formulation, there is no need of a special theory for explanatory variables. However, it is often the case that:

(i) the marginal law of Z is not of interest, but only the conditional law of X given Z is of interest;
(ii) Z is completely observed, that is, all the components of R_Z are identically equal to one (we will refer to this by writing $R_Z=1$).

Because of (i), we consider the following parametrization of the model: the model is defined by the family of probability measures $(P_{\theta\gamma\psi})_{\theta\in\Theta,\,\gamma\in\Gamma,\,\psi\in\Psi}$, where $P_{\theta\gamma\psi}$ is specified by $P_{\gamma\mathcal{Z}}$, $P_{\psi}^{\mathcal{X},\mathcal{Z}}$ and $P_{\theta\mathcal{X}}^{\mathcal{Z}}$, i.e., γ indexes the marginal probability of Z, ψ the conditional probability given $\mathcal{X}$ and $\mathcal{Z}$, and θ the conditional probability of X given $\mathcal{Z}$ on which the interest focuses. We take a reference probability P_0 and we assume that CAR(TCMP) holds for (W,R), that is $\mathcal{L}_{\mathcal{R}|\mathcal{W}}^{P_{\theta\gamma\psi}/P_0}$ is $\mathcal{O}$-measurable; this is equivalent to $\mathcal{L}_{\mathcal{R}_X|\mathcal{W}}^{P_{\theta\gamma\psi}/P_0}\mathcal{O}$-measurable (where $\mathcal{R}_X$ is the σ-field generated by R_X) because $R_Z=1$ (this can be seen, for instance, by applying property (vi) of Commenges & Gégout-Petit, 2005). In that case, if (ii) holds, we can get rid of both the nuisance parameters ψ and γ for the likelihood inference on θ. This is a consequence of the double-factorization theorem.

Theorem 2 (Double-factorization)
Consider the process $W=(X,Z)$, where X is the process of interest, Z is a process of explanatory variables; $R=(R_X,R_Z)$ is the associated response process and we have $R_Z=1$. Consider the family of equivalent probability measures $\mathcal{Q}=\{\{P_{\theta\gamma\psi}\}_{\theta\in\Theta,\,\gamma\in\Gamma,\,\psi\in\Psi},\mathcal{Q}_0\}$, where $P_{\theta\gamma\psi}$ is specified by $P_{\gamma\mathcal{Z}}$, $P_{\theta\mathcal{X}}^{\mathcal{Z}}$ and $P_{\psi}^{\mathcal{X},\mathcal{Z}}$, and where $\mathcal{Q}_0$ is a family of possible reference probabilities; the restriction of $P_{\theta\gamma\psi}$ on $\mathcal{W}$ is denoted $P_{\theta\gamma\mathcal{W}}$. Consider a reference probability $P_0\in\mathcal{Q}_0$. If the couple (W,X) satisfies CAR(TCMP) in $\mathcal{Q}$, then we have:

$$\mathcal{L}_{\mathcal{O}}^{P_{\theta\gamma\psi}/P_0}=\mathcal{L}_{\mathcal{R}|\mathcal{W}}^{P_{\theta\gamma\psi}/P_0}\mathcal{L}_{\mathcal{Z}}^{P_{\gamma\mathcal{Z}}/P_{0\mathcal{Z}}}\mathrm{E}_{P_0}[\mathcal{L}_{\mathcal{X}|\mathcal{Z}}^{P_{\theta\gamma\mathcal{W}}/P_{0\mathcal{W}}}|\mathcal{O}] \tag{4}$$

and

(1) $\mathcal{L}_{\mathcal{R}|\mathcal{W}}^{P_{\theta\gamma\psi}/P_0}$ *depends neither on θ nor on γ and can be denoted $\mathcal{L}_{\mathcal{R}|\mathcal{W}}^{P_{\cdot\psi}/P_0}$;* $\mathcal{L}_{\mathcal{Z}}^{P_{\gamma\mathcal{Z}}/P_{0\mathcal{Z}}}$ *depends neither on θ nor on ψ;* $\mathcal{L}_{\mathcal{X}|\mathcal{Z}}^{P_{\theta\gamma\mathcal{W}}/P_{0\mathcal{W}}}$ *depends neither on ψ nor on γ and can be denoted $\mathcal{L}_{\mathcal{X}|\mathcal{Z}}^{P_{\theta,\mathcal{W}}/P_{0\mathcal{W}}}$;*
(2) $\mathrm{E}_{P_0}[\mathcal{L}_{\mathcal{X}|\mathcal{Z}}^{P_{\theta,\mathcal{W}}/P_{0\mathcal{W}}}|\mathcal{O}]$ *depends neither on* $P_0^{\mathcal{X},\mathcal{Z}}$ *nor on* $P_{0\mathcal{Z}}$.

Proof. As CAR(TCMP) holds for (W,R) we can apply the (simple) factorization theorem which gives:

$$\mathcal{L}_{\mathcal{O}}^{P_{\theta\gamma\psi}/P_0}=\mathcal{L}_{\mathcal{R}|\mathcal{W}}^{P_{\theta\gamma\psi}/P_0}\mathrm{E}_{P_0}[\mathcal{L}_{\mathcal{W}}^{P_{\theta\gamma\mathcal{W}}/P_{0\mathcal{W}}}|\mathcal{O}].$$

We next use the decomposition

$$\mathcal{L}_{\mathcal{W}}^{P_{\theta\gamma\mathcal{W}}/P_{0\mathcal{W}}}=\mathcal{L}_{\mathcal{Z}}^{P_{\gamma\mathcal{Z}}/P_{0\mathcal{Z}}}\mathcal{L}_{\mathcal{X}|\mathcal{Z}}^{P_{\theta\gamma\mathcal{W}}/P_{0\mathcal{W}}}$$

and as $\mathcal{L}_{\mathcal{Z}}^{P_{\gamma\mathcal{Z}}/P_{0\mathcal{Z}}}$ is $\mathcal{O}$-measurable (because $\mathcal{L}_{\mathcal{Z}}^{P_{\gamma\mathcal{Z}}/P_{0\mathcal{Z}}}$ is $\mathcal{Z}$-measurable and $\mathcal{Z}\subset\mathcal{O}$) we obtain (4). Point (1) is straightforward. Another way to express Point (2) is that if we consider two probabilities P_1 and P_0 such that $P_{0\mathcal{X}}^{\mathcal{Z}}=P_{1\mathcal{X}}^{\mathcal{Z}}$ we have

$$\mathrm{E}_{P_0}[\mathcal{L}_{\mathcal{X}|\mathcal{Z}}^{P_{\theta,\mathcal{W}}/P_{0\mathcal{W}}}|\mathcal{O}]=\mathrm{E}_{P_1}[\mathcal{L}_{\mathcal{X}|\mathcal{Z}}^{P_{\theta,\mathcal{W}}/P_{1\mathcal{W}}}|\mathcal{O}].$$

The proof is similar to that of Theorem 2 in Commenges & Gégout-Petit (2005).

We can now apply the trick of the reduced model of section 4.1 to this context. If we knew the true conditional probability given $(\mathcal{X}, \mathcal{Z})$, $P_*^{\mathcal{X},\mathcal{Z}}$, and the true marginal probability $P_{*\mathcal{Z}}$, we would use the reduced model $\{P_\theta\}_{\theta\in\Theta}$ such that $P_\theta^{\mathcal{X},\mathcal{Z}} = P_*^{\mathcal{X},\mathcal{Z}}$ and $P_{\theta\mathcal{Z}} = P_{*\mathcal{Z}}$. Taking a reference probability P_0 such that $P_0^{\mathcal{X},\mathcal{Z}} = P_*^{\mathcal{X},\mathcal{Z}}$ and $P_{0\mathcal{Z}} = P_{*\mathcal{Z}}$ we have that $\mathcal{L}_{\mathcal{R}|\mathcal{X},\mathcal{Z}}^{P_\theta/P_0} = 1$ a.s. Thus, we have, without additional assumption,

$$\mathcal{L}_{\mathcal{O}}^{P_\theta/P_0} = \mathrm{E}_{P_0}[\mathcal{L}_{\mathcal{X}|\mathcal{Z}}^{P_{\theta,\mathcal{W}}/P_0\mathcal{W}} | \mathcal{O}].$$

If CAR(TCMP) holds $\mathrm{E}_{P_0}[\mathcal{L}_{\mathcal{X}|\mathcal{Z}}^{P_{\theta,\mathcal{W}}/P_0\mathcal{W}} | \mathcal{O}]$ does not depend on $P_0^{\mathcal{W}} = P_*^{\mathcal{W}}$ nor on $P_{0\mathcal{Z}} = P_{*\mathcal{Z}}$. That is, we can compute the exact likelihood that we would like to compute if we knew $P_*^{\mathcal{X},\mathcal{Z}}$ and $P_{*\mathcal{Z}}$, without in fact knowing them.

We can adapt now the same reasoning as in section 5.1 for defining the risk function for choosing estimators in the case where there are explanatory variables. We assume that we have n i.i.d. triples (X_i, Z_i, R_i), where the stochastic processes Z_i represent time-dependent explanatory variables. Assuming that CAR(TCMP) holds for $(W_i, R_i), i = 1, \ldots, n$, with $W_i = (X_i, Z_i)$ and using the reduced model we have that

$$\mathcal{L}_{\mathcal{O}_{n+1}}^{P_\theta/P_*} = \mathrm{E}_{P_*}[\mathcal{L}_{\mathcal{X}_{n+1}|\mathcal{Z}_{n+1}}^{P_{\theta,\mathcal{W}}/P_*\mathcal{W}} | \mathcal{O}_{n+1}]$$

depends neither on $P_*^{\mathcal{X},\mathcal{Z}}$ nor on $P_{*\mathcal{Z}}$. We define the loss function as before as

$$-\log \mathcal{L}_{\mathcal{O}_{n+1}}^{P_{\hat\theta}/P_*} = -\log \mathrm{E}_{P_*}[\mathcal{L}_{\mathcal{X}_{n+1}|\mathcal{Z}_{n+1}}^{P_{\hat\theta,\mathcal{W}}/P_*\mathcal{W}} | \mathcal{O}_{n+1}, \bar{\mathcal{O}}_n]$$

and the risk function as $\mathrm{EKL}_n = \mathrm{E}_{P_*}[-\log \mathcal{L}_{\mathcal{O}_{n+1}}^{P_{\hat\theta}/P_*}]$.

6. Choice between semi-parametric models: likelihood cross-validation

6.1. Estimating EKL$_n$ by likelihood cross-validation

Making the assumption that CAR(TCMP) holds for $(X_i, R_i), i = 1, \ldots, n$ we use a reduced model based on a reference probability P_0; then the observed likelihood ratio is $\mathcal{L}_{\bar{\mathcal{O}}_n}^{P_\theta/P_0}$. If there are explanatory variables, we assume that CAR(TCMP) holds for $(W_i, R_i), i = 1, \ldots, n$, we use the reduced model and we still denote the observed likelihood $\mathcal{L}_{\bar{\mathcal{O}}_n}^{P_\theta/P_0}$.

Now, we are seeking an 'estimator' for our criterion EKL_n. We consider the leave-one-out LCV criterion as a possible 'estimator'. It is defined as:

$$\mathrm{LCV}_n(\hat\theta(\cdot;\cdot), \kappa, \bar{\mathcal{O}}_n) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathcal{L}_{\mathcal{O}_i}^{P_{\hat\theta(\kappa,\, \bar{\mathcal{O}}_{n|i})}/P_0},$$

where $\bar{\mathcal{O}}_{n|i} = \vee_{j\neq i}\mathcal{O}_j$.

A first property, bearing on expectation of LCV is:

Lemma 1
$\mathrm{E}_{P_*}[\mathrm{LCV}_n(\hat\theta(\cdot;\cdot), \kappa, \bar{\mathcal{O}}_n)] = \mathrm{EKL}_{n-1}(\hat\theta(\cdot;\cdot), \kappa) - \mathrm{KL}(P_0, P_*)$.

Proof. We have

$$\begin{aligned}
\mathrm{E}_{P_*}[\mathrm{LCV}_n(\hat\theta(\cdot;\cdot), \kappa, \bar{\mathcal{O}}_n)] &= -\mathrm{E}_{P_*}[\log \mathcal{L}_{\mathcal{O}_i}^{P_{\hat\theta(\kappa,\, \bar{\mathcal{O}}_{n|i})}/P_0}] \\
&= -\mathrm{E}_{P_*}[\log \mathcal{L}_{\mathcal{O}_i}^{P_{\hat\theta(\kappa,\, \bar{\mathcal{O}}_{n|i})}/P_*} - \log \mathcal{L}_{\mathcal{O}_i}^{P_*/P_0}] \\
&= \mathrm{EKL}_{n-1}(\hat\theta(\cdot;\cdot), \kappa) - \mathrm{KL}(P_0, P_*).
\end{aligned}$$

So, using LCV_n for the choice of κ, we are using an unbiased estimator of EKL_{n-1}: indeed, as $\mathrm{KL}(P_0, P_*)$ does not depend on κ, we have

$$\begin{aligned}&\mathrm{E}_{P_*}[\mathrm{LCV}_n(\hat{\theta}(\cdot;\cdot), \kappa_2, \bar{\mathcal{O}}_n)] - \mathrm{E}_{P_*}[\mathrm{LCV}_n(\hat{\theta}(\cdot;\cdot), \kappa_1, \bar{\mathcal{O}}_n)]\\ &\quad = \mathrm{EKL}_{n-1}(\hat{\theta}(\cdot;\cdot), \kappa_2) - \mathrm{EKL}_{n-1}(\hat{\theta}(\cdot;\cdot), \kappa_1).\end{aligned}$$

Thus, LCV estimates a difference in EKL without the assumption that the true probability belongs to the model. Moreover, we conjecture that the optimal properties obtained by Hall (1987) and van der Laan *et al.* (2004) extend under certain assumptions to the general context considered here and that cross-validation will effectively be able to choose between semi-parametric multi-state models.

6.2. *Computational algorithm*

6.2.1. *Approximation of the solution of the penalized likelihood*

The penalized likelihood estimator $\hat{\theta}_\kappa$ is the set of functions and parameters which maximize $pl_\kappa(\theta)$. In general, it is not possible to compute $\hat{\theta}_\kappa$ analytically so the $\hat{g}_k^\kappa$ are approximated, for instance, by splines. With this approximation, the optimization problem becomes a standard maximization on a finite number of parameters. (Note that the number of knots in the spline representation is limited only by computational issues: the smoothness of the final estimator for the $g_k(\cdot)$ is controlled by κ in the penalized likelihood, not by the number of knots.) Calling γ the set of new parameters (including β and the set of spline parameters) we are led to maximizing:

$$pl_{\bar{\mathcal{O}}_n} = pl^\kappa_{\bar{\mathcal{O}}_n}(\gamma) = L^\gamma_{\bar{\mathcal{O}}_n} - J(\gamma, \kappa), \tag{5}$$

where $L^\gamma_{\bar{\mathcal{O}}_n} = \log \mathcal{L}^{\gamma/P_0}_{\bar{\mathcal{O}}_n}$. We note $\hat{\gamma} = \hat{\gamma}(\bar{\mathcal{O}}_n, \kappa) = \mathrm{argmax}_\gamma(pl^\kappa_{\bar{\mathcal{O}}_n}(\gamma))$.

6.2.2. *Approximation of LCV*

As LCV_n (that we will note simply LCV from now on) is particularly computationally demanding when n is large an approximate version LCV_a has been proposed by O'Sullivan (1988) for the estimation of the hazard function in a survival case and adapted by Joly *et al.* (2002) to the case of interval-censored data in an illness–death model. We may still extend it to a general framework valid for any penalized likelihood depending on a vector of real parameters γ. We note $\mathrm{LCV} = -n^{-1}\sum_{i=1}^n L^{\hat{\gamma}_{-i}}_{\mathcal{O}_i}$, where $\hat{\gamma}_{-i} = \hat{\gamma}(\bar{\mathcal{O}}_{n|i}, \kappa)$. The first-order development of $L^{\hat{\gamma}_{-i}}_{\mathcal{O}_i}$ around $\hat{\gamma}$ yields:

$$L^{\hat{\gamma}_{-i}}_{\mathcal{O}_i} \approx L^{\hat{\gamma}}_{\mathcal{O}_i} + (\hat{\gamma}_{-i} - \hat{\gamma})^T \hat{d}_i, \tag{6}$$

where

$$\hat{d}_i = \frac{\partial L^\gamma_{\mathcal{O}_i}}{\partial \gamma} | \hat{\gamma}.$$

The first-order development of

$$\frac{\partial pl^\gamma_{\bar{\mathcal{O}}_{n|i}}}{\partial \gamma} | \hat{\gamma}_{-i}$$

gives:

$$\hat{\gamma}_{-i} - \hat{\gamma} \approx -H^{-1}_{pl_{\bar{\mathcal{O}}_{n|i}}} \frac{\partial pl^\gamma_{\bar{\mathcal{O}}_{n|i}}}{\partial \gamma} | \hat{\gamma},$$

where

$$H_{pl_{\bar{\mathcal{O}}_{n|i}}} = \frac{\partial^2 pl_{\bar{\mathcal{O}}_{n|i}}}{\partial \gamma^2} | \hat{\gamma},$$

and more generally

$$H_g = \frac{\partial^2 g}{\partial \gamma^2} | \hat{\gamma}.$$

At first-order $H_{pl_{\bar{\mathcal{O}}_{n|i}}} \approx H_{pl_{\bar{\mathcal{O}}_n}}$. From the equality, $pl_{\bar{\mathcal{O}}_n}(\gamma) = pl_{\bar{\mathcal{O}}_{n|i}}(\gamma) + L^{\gamma}_{\mathcal{O}_i}$, we deduce by taking derivatives:

$$0 = \frac{\partial pl^{\gamma}_{\bar{\mathcal{O}}_{n|i}}}{\partial \gamma} | \hat{\gamma} + \hat{d}_i$$

which finally yields:

$$\hat{\gamma}_{-i} - \hat{\gamma} \approx H^{-1}_{pl_{\bar{\mathcal{O}}_n}} \hat{d}_i,$$

which inserted in (6) gives:

$$L^{\hat{\gamma}_{-i}}_{\bar{\mathcal{O}}_i} \approx L^{\hat{\gamma}}_{\bar{\mathcal{O}}_i} + \hat{d}_i^T H^{-1}_{pl_{\bar{\mathcal{O}}_n}} \hat{d}_i.$$

Substituting this expression in the expression of LCV we obtain:

$$\mathrm{LCV} \approx \mathrm{LCV}_{a_1} = -n^{-1} \left[L^{\hat{\gamma}}_{\bar{\mathcal{O}}_n} + \sum_{i=1}^{n} \hat{d}_i^T H^{-1}_{pl_{\bar{\mathcal{O}}_n}} \hat{d}_i \right].$$

Using the fact that both $n^{-1} \sum_{i=1}^{n} \hat{d}_i \hat{d}_i^T$ and $-n^{-1} H_{L_{\bar{\mathcal{O}}_n}}$ tend towards the individual information matrix $I = -\mathrm{E}_{P_*}(H_{L_{\mathcal{O}_i}})$, we get another approximation:

$$\mathrm{LCV} \approx \mathrm{LCV}_a = -n^{-1} [L^{\hat{\gamma}}_{\bar{\mathcal{O}}_n} - \mathrm{Tr}(H^{-1}_{pl_{\bar{\mathcal{O}}_n}} H_{L_{\bar{\mathcal{O}}_n}})].$$

This expression looks like an Akaike information criterion (AIC) and there are arguments to interpret $\mathrm{Tr}[H^{-1}_{pl_{\bar{\mathcal{O}}_n}} H_{L_{\bar{\mathcal{O}}_n}}]$ as the model degree of freedom. For instance, if there is no penalty ($J = 0$), $H_{pl_{\bar{\mathcal{O}}_n}} = H_{L_{\bar{\mathcal{O}}_n}}$ so that the correction term in LCV_a reduces to $\dim(\gamma)$, that is LCV_a reduces to AIC.

If κ is a scalar, minimization of LCV_a can be done by standard line-search algorithms; if it is a vector, a grid algorithm can be used (Joly *et al.*, 2002).

7. Simulation study

7.1. Description and main result

We did a simulation study to illustrate the ability of LCV to choose the right model structure; the possibilities were the Markov structure and the semi-Markov current-state structure, respectively, models $\mathcal{M}_1$ and $\mathcal{M}_2$ in section 2.2. The *best* model is not always the *right* model, especially in the case where the right model is larger than alternative models. Here, the model 1 and model 2 structures are of similar complexities, so the right model should be the best model.

We considered two particular models (or probability measures, $M_1 \in \mathcal{M}_1$ and $M_2 \in \mathcal{M}_2$). For both M_1 and M_2, the transition intensities toward illness, $\alpha_{01}(t)$, and death, $\alpha_{02}(t)$, were taken equal to the hazard function of Weibull distributions, namely $p\gamma^p t^{p-1}$ with parameters ($p = 2.4; \gamma = 0.05$) and ($p = 2.5; \gamma = 0.06$) respectively. Models M_1 and M_2 differed by the intensity of N_2 for $t > T_1$, that is the mortality rate of diseased. For M_1 the mortality rate was defined by $h(t)$ and was a Weibull hazard function with parameters ($p = 2.6; \gamma = 0.08$); for M_2

it was defined by $g(t-T_1)$ which was equal to a Weibull hazard function with parameters $(p=1.5;\gamma=0.2)$. For each subject, we generated an ignorable TCMP observation scheme by generating R_1 and R_2 independently from N_1 and N_2. In intuitive terms, R_1 and R_2 were constructed to represent a situation where N_1 was observed at discrete times and N_2 was observed in continuous time and possibly right-censored (the same as in the application). For each subject, we generated visit times V_j at which N_1 was observed as $V_j=V_{j-1}+2+3U_j$, where the U_js were independent uniform [0, 1] random variables, and observation of both N_1 and N_2 was right-censored by a variable C which had a uniform distribution on [2, 52]. We had $R_1(t)=1_{\{t<C\}}\sum_{j\geq 1}1_{\{t=V_j\}}$ and $R_2(t)=1_{\{t<C\}}$. To take into account the discrete time observation scheme on one component we used formula (13) of Commenges & Gégout-Petit (2006).

We generated 100 replicas of samples N_1 and N_2 from M_1 and M_2; each sample had $n=1000$ subjects. For each replica, the three functions determining the model $[\alpha_{01}(\cdot), \alpha_{02}(\cdot)$ for both models and $h(\cdot)$ for M_1 and $g(\cdot)$ for $M_2]$ were estimated by penalized likelihood while the parameter $\hat{\eta}$ (determining model structure) and the smoothing coefficients $\hat{\kappa}=(\hat{\kappa}_1,\hat{\kappa}_2,\hat{\kappa}_3)$ which minimized LCV were determined.

The first result is that when the Markov model was generated, it was chosen in 99 cases out of 100; when the semi-Markov model was generated, it was chosen in 93 cases out of 100. This shows that LCV does a good job in picking the right model structure. Table 1 shows the distance in terms of the risk EKL (the average Kullback–Leibler loss is an estimate of EKL) between the estimated models and the true model: choosing the model structure by LCV incurs a very slight additional risk (of order 10^{-4}) when compared with knowing the true model but a lower risk when compared with choosing the wrong model; in the latter case, the additional risk is of order 10^{-2}.

This result must not be falsely interpreted. First, the discrimination properties of LCV depends on many parameters and particularly on the quantity of information available in the samples. Second and even more important, the aim of estimator choice is not to choose the right model but to choose the best estimator. The choice between the two structures depends on how 'far' the two models are. If the models are 'close', it is, of course, more difficult to discriminate between them, but at the same time it becomes less important to choose the right one. For instance, the homogeneous Markov model belongs to both structures so it is possible by small perturbations of this model to construct two models, one Markov and one semi-Markov, which are very near in terms, for instance, of Kullback–Leibler divergence.

7.2. Study of the variability of LCV

In this section, we exploit the above simulation study to explore the variability of LCV. We have estimated from the 100 replicas from M_1 the density of LCV($\hat{\kappa}$) assuming $\mathcal{M}_1$ and assuming $\mathcal{M}_2$. The upper-left panel of Fig. 2 displays these estimated densities: there seems to be little difference between the two, so one may wonder whether LCV can be of any use

Table 1. *Average Kullback–Leibler loss* $\overline{KL}$ *and the corresponding standard errors (numbers in the parentheses) for estimators chosen by LCV*

	$\overline{\text{KL}}$		
	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_1$ or $\mathcal{M}_2$
True model $\mathcal{M}_1$	0.00489 (0.00024)	0.02918(0.00024)	0.00515 (0.00035)
True model $\mathcal{M}_2$	0.10335 (0.00018)	0.09675(0.00019)	0.09719 (0.00025)

Replication = 100; sample size = 1000.

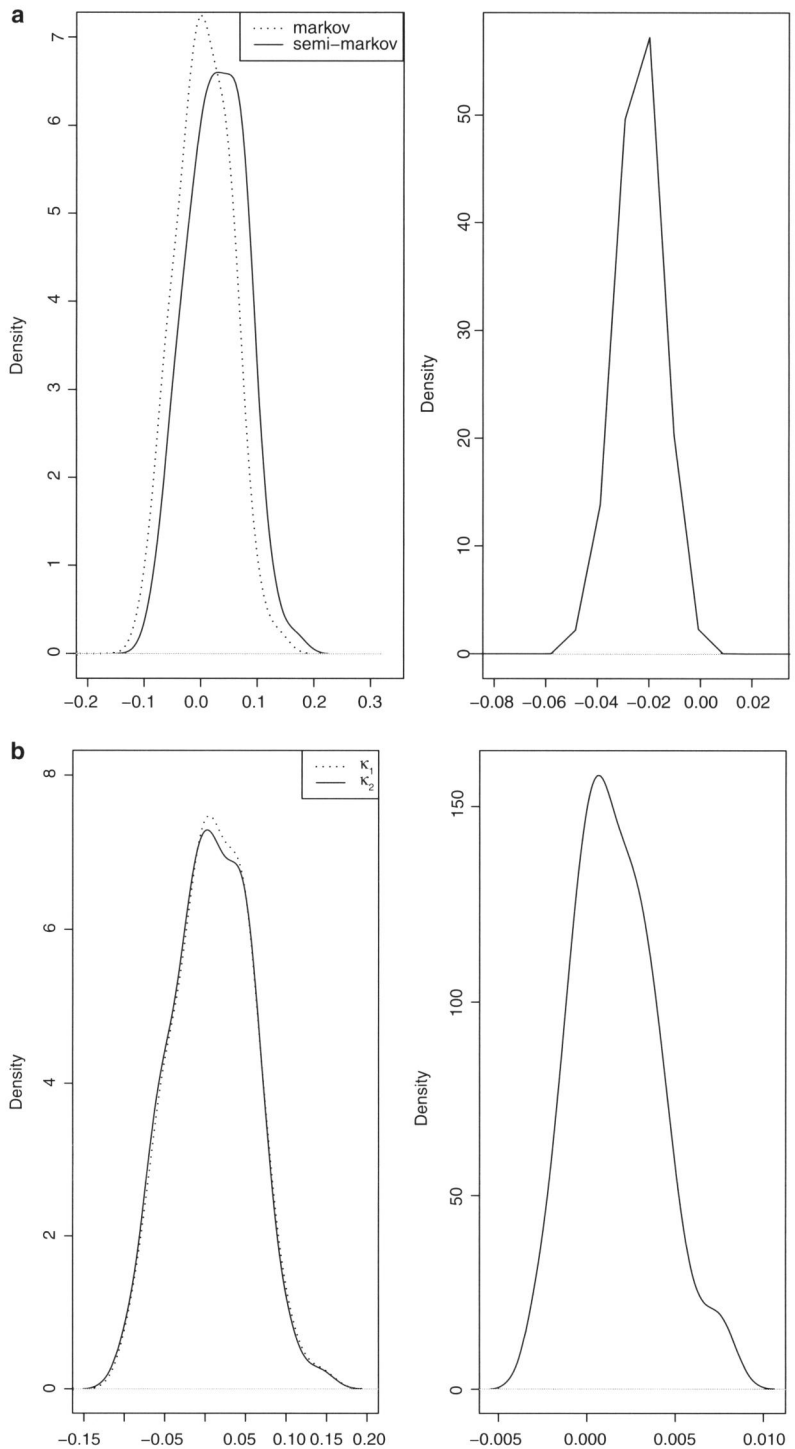

Fig. 2. Kernel density estimation of LCV (left) and of differences of LCV (right) for (a) Markov and semi-Markov choices, (b) for two different values of the smoothing parameter; in all cases the true model is Markov.

for choosing between the two model structures. However, when we look at the density of the difference between LCV($\hat{\kappa}$) for $\mathcal{M}_1$ and $\mathcal{M}_2$, we see that most of the mass is in the negative values, so that most of the time the true model $\mathcal{M}_1$ will be chosen. Similarly, the lower panels show the estimated densities of LCV, assuming $\mathcal{M}_1$, for two different values κ_1 and κ_2 of the smoothing coefficient. Here, the two densities are nearly undistinguishable while the density of the difference is clearly shifted toward positive values. Another way to examine this issue is to look at the standard deviations of LCV(κ_1), LCV(κ_2) and LCV(κ_2) − LCV(κ_1): we estimated these values (under $\mathcal{M}_1$) to be 0.048, 0.048 and 0.0024 respectively. Thus the standard deviation of the difference is about 20 times less than that of LCV for κ_1 or κ_2. This explains why LCV does a good job in model choice in spite of its large variability.

7.3. Quantitative interpretation of Kullback–Leibler divergences

For the practical use of the method proposed in this paper it is important to have an idea of whether a particular EKL value, or a difference of EKL values or their LCV estimators are large or not. As in more conventional situations, we must distinguish the interpretational issue from statistical issues. For instance, in the conventional situation of a regression parameter, the statistical issues beyond the point estimation of the parameter are to test the hypothesis of a null value of the parameter and to give a confidence interval for this parameter; the interpretational issue is to be able to assess the importance of the effect on the variable of interest. For instance, in an epidemiological application using a Cox model, we would consider the exponential of the parameter and interpret it as a relative risk, considering that a value of 1.1, 2 and 5 would correspond to a small, moderate and large increase of risk respectively. We would like to have a guide toward such an interpretation when manipulating EKL values.

As EKL is an expected Kullback–Leibler divergence, it can be interpreted as a Kullback–Leibler divergence. So let us try to interpret KL($\tilde{P}, P_*$). If we consider that P_* is the true probability this means that we will make errors by evaluating the probability of an event A by $\tilde{P}(A)$ rather than by $P_*(A)$. For instance, we may evaluate the relative error

$$r_e(\tilde{P}(A), P_*(A)) = \frac{P_*(A) - \tilde{P}(A)}{P_*(A)}.$$

Consider the typical event on which $\tilde{P}(A)$ will be under-evaluated and defined as: $A = \{\omega : \mathcal{L}^{\tilde{P}/P_*} < 1\}$. To obtain a simple formula relating KL($\tilde{P}, P_*$) to the error on $P_*(A)$, we consider the particular case $P_*(A) = 1/2$ and $\mathcal{L}^{\tilde{P}/P_*}$ constant on A and A^C [or equivalently we compute KL for a likelihood defined on $\sigma(A)$]. In that case we easily find:

$$r_e(\tilde{P}(A), P_*(A)) = \sqrt{1 - e^{-2\mathrm{KL}(\tilde{P}, P_*)}} \approx \sqrt{2\mathrm{KL}(\tilde{P}, P_*)},$$

the approximation being valid for a small KL value. For KL values of 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, we find that $r_e(\tilde{P}(A), P_*(A))$ is equal to 0.44, 0.14, 0.045 and 0.014, errors that we may qualify as 'large', 'moderate', 'small' and 'negligible' respectively.

As an example, the KL divergence of a double exponential relative to a normal distribution with same mean and variance is of order 10^{-1} leading to a 'large' $r_e(\tilde{P}(A), P_*(A))$. In the previous simulation study, we have found that choosing the wrong model leads to an increase of the risk of order 10^{-2} which is 'moderate', while choosing the model by LCV leads to and increase of order 10^{-4} which may be qualified as 'negligible'.

8. Application on dementia

We illustrate the use of this general approach using the data of the Paquid study (Letenneur *et al.*, 1999), a prospective cohort study of mental and physical aging that evaluates social environment and health status. The target population consists of subjects aged 65 years and older living at home in south-western France. The diagnosis of dementia was made according to a two-stage procedure: the psychologist who filled the questionnaire screened the subjects as possibly demented according to DSM-III-R or not; subjects classified as positive were later seen by a neurologist who confirmed (or not) the diagnosis of dementia and made a more specific diagnosis, assessing in particular the NINCDS-ADRDA criteria for Alzheimer's disease. Subjects were re-evaluated 1, 3, 5, 8, 10 and 13 years after the initial visit. Subjects already demented at the initial visit were removed from the sample, a selection condition which is easily taken into account by using a conditional likelihood as mentioned in Commenges & Gégout-Petit (2006). The sample consisted of 3673 subjects, 1540 men and 2133 women. Previous work (Commenges *et al.*, 2004) has shown that the effect of gender on the risk of dementia is neither multiplicative nor additive; in fact, the dynamics of ageing is so different between men and women that it is safer to perform completely separate analyses. For the purpose of this illustration, we analysed only women. During the 13 years of follow-up, 396 incident cases of dementia and 835 deaths were observed. We wish to jointly model dementia and death, an approach conventionally referred to as the illness–death model; the model can be graphically described as in Fig. 1, where the mortality rate of demented is noted $\alpha_{12}(t, t - T_1)$ to emphasize the fact that it may depend on both age t and time since the onset of dementia $t - T_1$; we assume that the transition intensities do not depend in addition on universal (or calendar) time. Note that dementia is observed in discrete time while death is observed in continuous time. One effect of the observation scheme is that we miss a certain number of dementia cases: we do not observe a dementia case which has happened when the subject develops dementia and dies between two planned visits. This scheme of observation and the likelihood for it are explained heuristically in Commenges *et al.* (2004) and rigorously in Commenges & Gégout-Petit (2006).

We tried the three model structures depicted in section 2.2. We took as reference probability the homogeneous Markov model fitted to the data. Thus LCV estimated the change in EKL when going from the homogeneous Markov maximum likelihood estimator to another estimator. The values of the best LCV criteria for the different model structures were: non-homogeneous Markov model ($\mathcal{M}_1$): -0.2182; current-state model ($\mathcal{M}_2$): -0.2100; excess mortality model ($\mathcal{M}_3$): -0.2180. This means for instance that the best penalized likelihood estimator in the non-homogeneous model has an estimated expected Kullback–Leibler divergence (EKL) relative to the true model which is smaller by 0.2182 than the homogeneous Markov estimator. The best LCV was found for the non-homogeneous Markov model. However, the best 'excess mortality' estimator is not far from the best non-homogeneous estimator, while the current-state estimator seems to be farther. In terms of the interpretation of section 7.3, the difference between $\mathcal{M}_2$ and $\mathcal{M}_1$ is moderate while that between $\mathcal{M}_3$ and $\mathcal{M}_1$ is negligible.

We compared graphically the best estimators found for the three model structures considered. Figure 3 shows the three estimators for the age-specific incidence of dementia (α_{01}) and the mortality rates of non-demented respectively: the three estimators are very close for incidence of dementia; there is a certain difference between the Markov model and the two semi-Markov models for mortality rates of non-demented above 90 years. Figure 4 displays the three estimators of the age-specific mortality rates of demented for different ages at onset of dementia, respectively 70, 80 and 90 years. Here the patterns are different although the

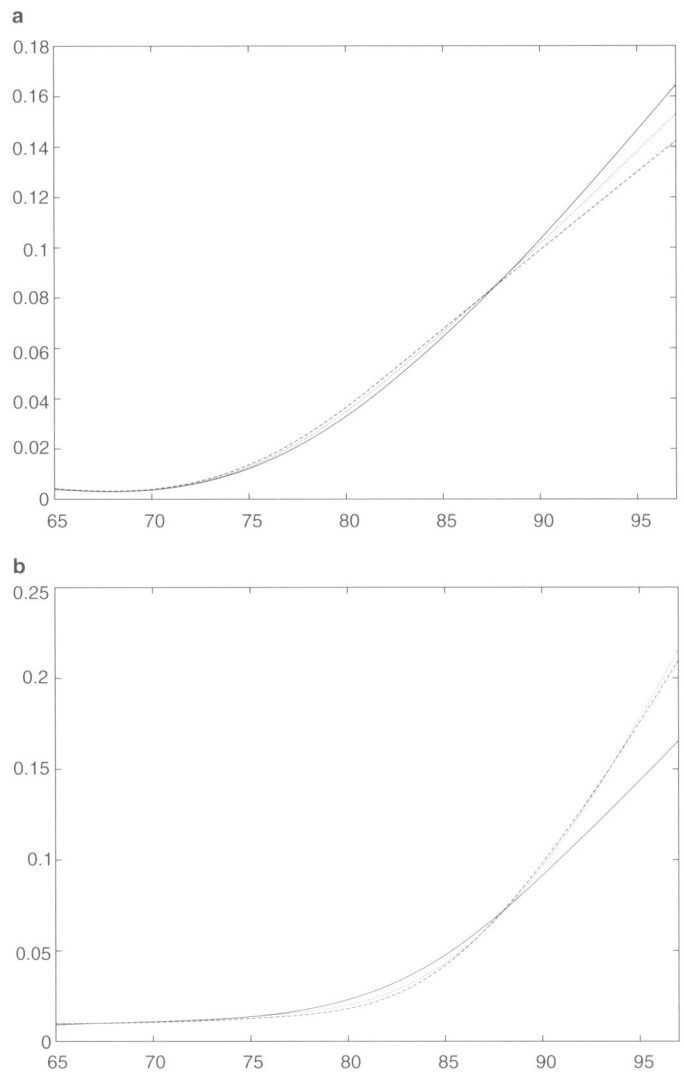

Fig. 3. Incidence of dementia (a) and mortality (b) for non-demented women for the three models. Continuous line: non-homogeneous Markov model; dashed line: current-state model; dotted line: excess mortality model.

magnitude of these estimators are similar. In particular, the current-state estimator is the same for the three ages at onset (by assumption) while we see a marked increase of mortality for age at onset of 90 years in the non-homogeneous Markov estimator. From a qualitative point of view we may say that the three estimators agree for ages at onset of 70 and 80 years: the mortality rate for demented women does not vary much either with time since the onset of dementia or with age at onset, and is around 0.2.

9. Conclusion

We have extended the expected Kullback–Leibler risk function (EKL) for estimator choice from generally coarsened observations of a stochastic process, including in the case of explanatory variables. We have suggested that this could be used for choosing both smoothing

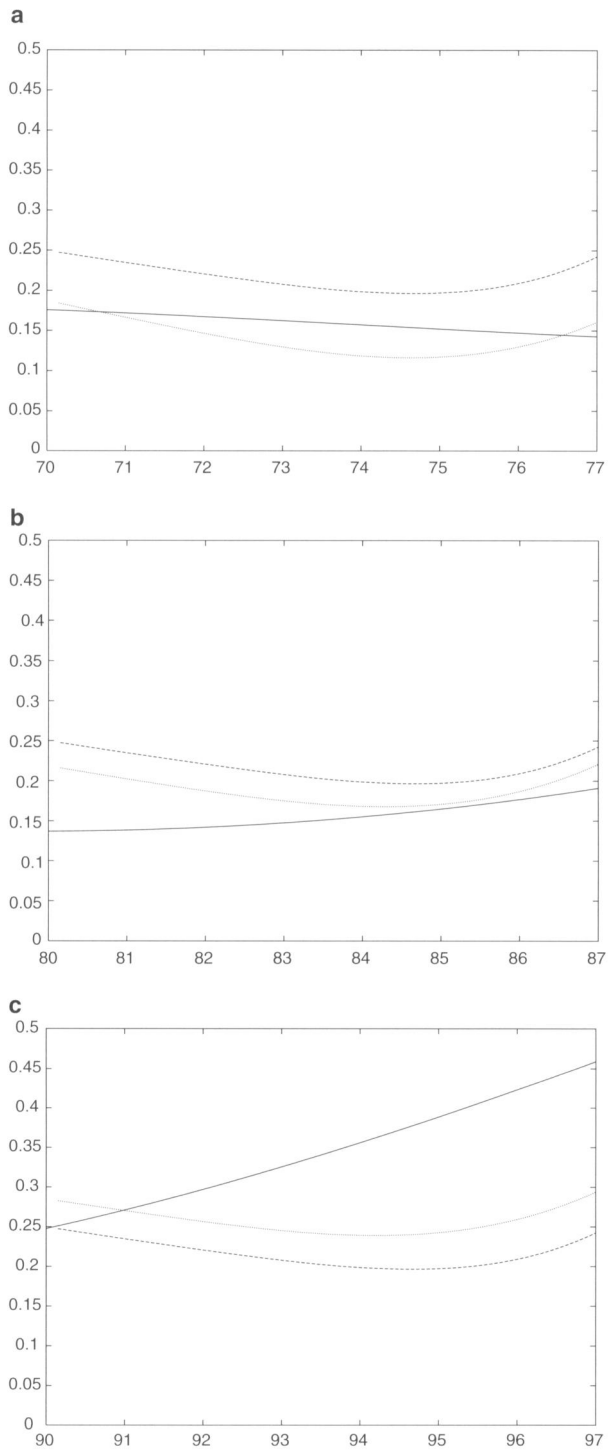

Fig. 4. Mortality of demented women for the three models: continuous line: non-homogeneous Markov model; dashed line: current-state model; dotted line: excess mortality model. Age (years) at onset of dementia: (a) 70; (b) 80; (c) 90.

coefficients and model structure; we have suggested that EKL could be approached by LCV and we have given a general approximation formula for the leave-one-out LCV. The simulation presented showed that the LCV did a good job in discriminating between model structures. The approach was illustrated in the problem of choosing between different additive illness–death models. The approach is in fact quite general and could be applied, for instance, to the choice between additive and multiplicative models.

Other choices might have been done: other loss functions, families of estimators and ways of estimating the risk function might have been chosen. However, the choices we have done for the different components of the approach are adapted to the problem and fit well together. For instance, the CAR(TCMP) assumption allows us to eliminate the nuisance parameters from the chosen loss function; LCV is a natural estimator of EKL; penalized likelihood yields a flexible family of smooth estimators for which an approximation of LCV can easily be computed. The approach yields an operational tool for exploring complex event histories, for instance, in the domain of ageing.

There are many open problems and useful developments would be: finding a better algorithm for minimizing LCV over multiple smoothing parameters; studying the variance of LCV (see Bengio & Grandvalet, 2004); finding asymptotic properties of the estimators chosen by minimizing LCV.

References

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6**, 701–726.

Aalen, O. O. & Johansen, S. (1978). An empirical transition matrix for non-homogenous Markov chains based on censored observations. *Scand. J. Statist.* **5**, 141–150.

Aalen, O. O., Borgan, Ø. & Fekjær, H. (2001). Covariate adjustment of event histories estimated from Markov chains: the additive approach. *Biometrics* **57**, 993–1001.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag, New York.

Bengio, Y. & Grandvalet, Y. (2004). No unbiased estimator of the variance of the K-fold cross-validation. *J. Mach. Learn. Res.* **5**, 1089–1105.

Commenges, D. & Gégout-Petit, A. (2005). Likelihood inference for incompletely observed stochastic processes: ignorability conditions. *arXiv:math.ST/0507151*. Available at: http://arxiv.org/abs/math/0705151.

Commenges, D. & Gégout-Petit, A. (2006). Likelihood for generally coarsened observations from multistate or counting process models. *Scand. J. Statist.*, in press, doi: 10.1111/j.1467.9469.2006.00518.

Commenges, D. & Joly, P. (2004). Multi-state model for dementia, institutionalization and death. *Commun. Statist. A* **33**, 1315–1326.

Commenges, D., Joly, P., Letenneur, L. & Dartigues, J. F. (2004). Incidence and prevalence of Alzheimer's disease or dementia using an illness-death model. *Statist. Med.* **23**, 199–210.

Cox, D. & O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18**, 1676–1695.

Eggermont, P. & LaRiccia, V. (1999). Optimal convergence rates for Good's nonparametric likelihood density estimator. *Ann. Statist.* **27**, 1600–1615.

Eggermont, P. & LaRiccia, V. (2001). *Maximum penalized likelihood estimation*. Springer-Verlag, New York.

Gill, R. D., van der Laan, M. J. & Robins, J. M. (1997). Coarsening at random: characterizations, conjectures and counter-examples. In *State of the art in survival analysis*, Springer Lecture Notes in Statistics 123 (eds D.-Y. Lin & T. R. Fleming), 255–294. Springer-Verlag, New York.

Good, I. J. & Gaskins, R. A. (1971). Nonparametric roughness penalty for probability densities. *Biometrika* **58**, 255–277.

Gu, C. (1996). Penalized likelihood hazard estimation: a general procedure. *Statist. Sinica* **6**, 861–876.

Hall, P. (1987). On Kullback–Leibler loss and density estimation. *Ann. Statist.* **15**, 1491–1519.

Hougaard, P. (2000). *Analysis of multivariate survival data*. Springer, New York.

Jacod, J. (1975). Multivariate point processes: predictable projection; Radon-Nikodym derivative, representation of martingales. *Z. Wahrsch. verw Geb.* **31**, 235–253.

Joly, P. & Commenges, D. (1999). A penalized likelihood approach for a progressive three-state model with censored and truncated data: application to AIDS. *Biometrics* **55**, 887–890.

Joly, P., Commenges, D., Helmer, C. & Letenneur, L. (2002). A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* **3**, 433–443.

Kallenberg, O. (2001). *Foundations of modern probabilities*. Springer-Verlag, New York.

Kooperberg, C. & Clarkson, D. B. (1997). Hazard regression with interval-censored data. *Biometrics* **53**, 1485–1494.

van der Laan, M. & Dudoit, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. UC Berkeley Division of Biostatistics working paper series, paper 130. Available at: http://www.bepress.com/ucbiostat/paper130.

van der Laan, M., Dudoit, S. & Keles, S. (2004). Asymptotic optimality of likelihood-based cross-validation. *Statist. Appl. Genet. Mol. Biol.* **3**.

Lagakos, S. W., Sommer, C. J. & Zelen, M. (1978). Semi-Markov models for partially censored data. *Biometrika* **65**, 311–317.

Le Cam, L. & Yang, G. (2000). *Asymptotics in Statistics*. Springer-Verlag, New-York.

Letenneur, L., Gilleron, V., Commenges, D., Helmer, C., Orgogozo, J. M. & Dartigues, J. F. (1999). Are sex and educational level independent predictors of dementia and Alzheimer's disease? Incidence data from the PAQUID project. *J. Neurol. Neurosurg. Psychiatr.* **66**, 177–183.

Liquet, B. & Commenges, D. (2004). Estimating the expectation of the log-likelihood with censored data for estimator selection. *Lifetime Data Anal.* **10**, 351–367.

Liquet, B., Sakarovitch, C. & Commenges, D. (2003). Bootstrap choice of estimators in parametric and semi-parametric families: an extension of EIC. *Biometrics* **59**, 172–178.

Liquet, B., Saracco, J. & Commenges, D. (2006). Selection between proportional and stratified hazards models based on expected log-likelihood. *Comput. Statist.*, in press.

Müller, H. G. & Stadtmüller, U. (2005). Generalized functional linear models. *Ann. Statist.* **33**, 774–805.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Comput.* **9**, 363–379.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11**, 453–466.

Scheike, T. (2001). A generalized additive regression model for survival analysis. *Ann. Statist.* **29**, 1344–1380.

Received November 2005, in final form July 2006

Daniel Commenges, INSERM E0338, Université Victor Segalen Bordeaux 2, 146 rue Léo Saignat, Bordeaux, 33076, France.
E-mail: daniel.commenges@isped.u-bordeaux2.fr