Statistics Surveys
Vol. 3 (2009) 1–17
ISSN: 1935-7516
DOI: 10.1214/08-SS039

Statistical models: Conventional, penalized and hierarchical likelihood*

Daniel Commenges

Epidemiology and Biostatistics Research Center, INSERM
Université Victor Segalen Bordeaux 2
146 rue Léo Saignat, Bordeaux, 33076, France
e-mail: daniel.commenges@isped.u-bordeaux2.fr

Abstract: We give an overview of statistical models and likelihood, together with two of its variants: penalized and hierarchical likelihood. The Kullback-Leibler divergence is referred to repeatedly in the literature, for defining the misspecification risk of a model and for grounding the likelihood and the likelihood cross-validation, which can be used for choosing weights in penalized likelihood. Families of penalized likelihood and particular sieves estimators are shown to be equivalent. The similarity of these likelihoods with a posteriori distributions in a Bayesian approach is considered.

AMS 2000 subject classifications: Primary 62-02, 62C99; secondary 62A01.
Keywords and phrases: Bayes estimators, cross-validation, h-likelihood, incomplete data, Kullback-Leibler risk, likelihood, penalized likelihood, sieves, statistical models.

Received August 2008.

Contents

1 Introduction . . . 2
2 Definition of a density . . . 2
3 The Kullback-Leibler risk . . . 4
4 Statistical models and families . . . 5
4.1 Statistical families . . . 5
4.2 Statistical models . . . 5
4.3 Statistical models and true probability . . . 6
5 The likelihood . . . 7
5.1 Definition of the likelihood . . . 7
5.2 Computation of the likelihood . . . 8
5.3 Performance of the MLE in terms of Kullback-Leibler risk . . . 9
6 The penalized likelihood . . . 9
7 The hierarchical likelihood . . . 11
7.1 Random effects models . . . 11
7.2 Hierarchical likelihood . . . 11
8 Akaike and likelihood cross-validation criteria . . . 12

*This paper was accepted by Elja Arjas, Executive Editor for the Bernoulli.

9 Link with the MAP estimator 13
Conclusion 14
Acknowledgements 14
References 15

1. Introduction

Since its proposal by Fisher [16], likelihood inference has occupied a central position in statistical inference. In some situations, modified versions of the likelihood have been proposed. Marginal, conditional, profile and partial likelihoods have been proposed to get rid of nuisance parameters. Pseudo-likelihood and hierarchical likelihood may be used to circumvent numerical problems in the computation of the likelihood, that are mainly due to multiple integrals. Penalized likelihood has been proposed to introduce a smoothness a priori knowledge on functions, thus leading to smooth estimators. Several reviews have already been published, for instance [31], but it is nearly impossible in a single paper to describe with some details all the types of likelihoods that have been proposed. This paper aims at describing the conventional likelihood and two of its variants: penalized and hierarchical likelihoods. The aim of this paper is not to give the properties of the estimators obtained by maximizing these likelihoods, but rather to describe these three likelihoods together with their link to the Kullback-Leibler divergence. This interest more in the foundations rather than the properties, leads us to first develop some reflexions and definitions about statistical models and to give a slightly extended version of the Kullback-Leibler divergence.

In section 2, we recall the definition of a density and the relationship between a density in the sample space and for a random variable. In section 3, we give a slightly extended version of the Kullback-Leibler divergence (making it explicit that it also depends on a sigma-field). Section 4 gives an account of statistical models, distinguishing mere statistical families from statistical models and defining the misspecification risk. Section 5 presents the likelihood and discusses issues about its computation and the performance of the estimator of the maximum likelihood in terms of Kullback-Leibler risk. In section 6, we define the penalized likelihood and show that for a family of penalized likelihood estimators there is an identical family of sieves estimators. In section 7, we describe the hierarchical likelihood. In section 8, we briefly sketch the possible unification of these likelihoods through a Bayesian representation that allows us to consider the maximum (possibly penalized) likelihood estimators as maximum a posteriori (MAP) estimators; this question however cannot be easily settled due to the non-invariance of the MAP for reparameterization. Finally, there is a short conclusion.

2. Definition of a density

Consider a measurable space $(\mathcal{S}, \mathcal{A})$ and two measures μ and ν with μ absolutely continuous relatively to ν. For $\mathcal{G}$ a sub-σ-field of $\mathcal{A}$ the Radon-Nikodym

derivative of μ with respect to ν on $\mathcal{X}$, denoted by: $\frac{d\mu}{d\nu}_{|\mathcal{G}}$ is the $\mathcal{G}$-measurable random variable such that

$$\mu(G) = \int_G \frac{d\mu}{d\nu}_{|\mathcal{G}} d\nu,\ G \in \mathcal{G}.$$

The Radon-Nikodym derivative is also called the density. We are interested in the case where μ is a probability measure, which we will call P^1; ν may also be a probability measure, P^0. In that case we can speak of the likelihood ratio and denote it $\mathcal{L}_{\mathcal{G}}^{P^1/P^0}$. In order to speak of a likelihood function, we have to define a model (see section 4). Note that likelihood ratios (as Radon-Nikodym derivatives) are defined with respect to a sigma-field. The definitions and properties of these probabilistic concepts are very clearly presented in [45]. For the statistician, sigma-fields represent sets of events which may be (but are not always) observed. If $\mathcal{H}$ and $\mathcal{G}$ are different sigma-fields, $\frac{dP^1}{dP^0}_{|\mathcal{H}}$ and $\frac{dP^1}{dP^0}_{|\mathcal{G}}$ are different, but if $\mathcal{H} \subset \mathcal{G}$ the former can be expressed as a conditional expectation (given $\mathcal{H}$) of the latter and we have the fundamental formula:

$$\frac{dP^1}{dP^0}_{|\mathcal{H}} = \mathrm{E}_{P^0}\left[\frac{dP^1}{dP^0}_{|\mathcal{G}} | \mathcal{H}\right].$$

Consider now the case where the measurable space $(\Omega, \mathcal{F})$ is the sample space of an experiment. For the statistician $(\Omega, \mathcal{F})$ is not any measurable space: it is a space which enables us to represent real events. We shall write in bold character a probability on $(\Omega, \mathcal{F})$, for instance, $\boldsymbol{P}^1$. Let us define a random variable X, that is, a measurable function from $(\Omega, \mathcal{F})$ to $(\Re, \mathcal{B})$. The couple $(\boldsymbol{P}^1, X)$ induces a probability measure on $(\Re, \mathcal{B})$ defined by: $P_X^1(B) = \boldsymbol{P}^1\{X^{-1}(B)\}, B \in \mathcal{B}$. This probability measure is called the distribution of X. If this probability measure is absolutely continuous with respect to Lebesgue (resp. counting) measure, one speaks of continuous (resp. discrete) variable. For instance, for a continuous variable we define the density $f_X^1 = \frac{dP_X^1}{d\lambda}$, where λ is Lebesgue measure on $\Re$, which is the usual probability density function (p.d.f.). Note that the p.d.f. depends on both $\boldsymbol{P}^1$ and X, while $\frac{d\boldsymbol{P}^1}{d\boldsymbol{P}^0}_{|\mathcal{X}}$ depends on $\mathcal{X}$ but not on a specific random variable X. Often in applied statistics one works only with distributions, but this may leave some problems unsolved.

Example 1. Consider the case where concentrations of CD4 lymphocytes are measured. Ω represents the set of physical concentrations that may happen. Let the random variables X and Y express the concentration in number of CD4 by mm^3 and by ml respectively. Thus we have $Y = 10^3 X$. So X and Y are different, although they are informationally equivalent. For instance the events $\{\omega : X(\omega) = 400\}$ and $\{\omega : Y(\omega) = 400000\}$ are the same. The densities of X and Y, for the same $\boldsymbol{P}^1$ on $(\Omega, \mathcal{F})$, are obviously different. So, if we look only at distributions, we shall have difficulties to rigorously define what a model is.

3. The Kullback-Leibler risk

Many problems in statistical inference can be treated from the point of view of decision theory. That is, estimators for instance are chosen as minimizing some risk function. The most important risk function is based on the Kullback-Leibler divergence. Maximum likelihood estimators, use of Akaike criterion or likelihood cross-validation can be grounded on the Kullback-Leibler divergence. Given a probability $\boldsymbol{P}^2$ absolutely continuous with respect to a probability $\boldsymbol{P}^1$ and $\mathcal{X}$ a sub-σ-field of $\mathcal{F}$, the loss using $\boldsymbol{P}^2$ in place of $\boldsymbol{P}^1$ is the log-likelihood ratio $L_{\mathcal{X}}^{\boldsymbol{P}^1/\boldsymbol{P}^2} = \log \frac{d\boldsymbol{P}^1}{d\boldsymbol{P}^2}_{|\mathcal{X}}$. Its expectation is $\mathrm{E}_{\boldsymbol{P}^1}[L_{\mathcal{X}}^{\boldsymbol{P}^1/\boldsymbol{P}^2}]$. This is the Kullback-Leibler risk, also called divergence [28, 29], information deviation [4] or entropy [1]. The different names of this quantity reflects its central position in statistical theory, being connected to several fields of the theory. Several notations have been used by different authors. Here we choose the Cencov [4] notation:

$$\boldsymbol{I}(\boldsymbol{P}^2|\boldsymbol{P}^1;\mathcal{X}) = \mathrm{E}_{\boldsymbol{P}^1}[L_{\mathcal{X}}^{\boldsymbol{P}^1/\boldsymbol{P}^2}].$$

If $\mathcal{X}$ is the largest sigma-field defined on the space, then we omit it in the notation. Note that the Kullback-Leibler risk is asymmetric and hence does not define a distance between probabilities; we have to take on this fact. If X is a random variable with p.d.f. f_X^1 and f_X^2 under $\boldsymbol{P}^1$ and $\boldsymbol{P}^2$ respectively we have $\frac{d\boldsymbol{P}^1}{d\boldsymbol{P}^2}_{|\mathcal{X}} = \frac{f_X^1(X)}{f_X^2(X)}$ and the divergence of the distribution P_X^2 relative to P_X^1 can be written:

$$\mathcal{I}(P_X^2|P_X^1) = \int \log \frac{f_X^1(x)}{f_X^2(x)} f_X^1(x)dx. \tag{3.1}$$

We have that $\boldsymbol{I}(\boldsymbol{P}^2|\boldsymbol{P}^1;\mathcal{X}) = \mathcal{I}(P_X^2|P_X^1)$, if $\mathcal{X}$ is the σ-field generated by X on $(\Omega,\mathcal{F})$. Note that on $(\Omega,\mathcal{F})$ we have to specify that we assess the divergence on $\mathcal{X}$; we might assess it on a different sigma-field and would of course obtain a different result. This provides more flexibility. In particular, we shall use this in the case of incomplete data. The observation is represented by a sigma-field $\mathcal{O}$. Suppose we are interested in making inference about the true probability on $\mathcal{X}$. We have complete data if our observation is $\mathcal{O} = \mathcal{X}$. With incomplete data, in the case where the mechanism leading to incomplete data is deterministic, we have $\mathcal{O} \subset \mathcal{X}$. In that case it will be very difficult to estimate $\boldsymbol{I}(\boldsymbol{P}^2|\boldsymbol{P}^1;\mathcal{X})$ and it will be more realistic to use $\boldsymbol{I}(\boldsymbol{P}^2|\boldsymbol{P}^1;\mathcal{O}) = \mathrm{E}_{\boldsymbol{P}^1}[L_{\mathcal{O}}^{\boldsymbol{P}^1/\boldsymbol{P}^2}]$. We need this flexibility to extend Akaike's argument for the likelihood and for developing model choice criteria to situations with incomplete data. This will become important in section 5, where $\boldsymbol{P}^1$ will be the true unknown probability (denoted $\boldsymbol{P}^*$) and the problem will be to estimate this divergence rather than to compute it.

Example 2. Suppose we are interested in modeling the time to an event, X, and we wish to evaluate the divergence of $\boldsymbol{P}^2$ with respect to $\boldsymbol{P}^1$. It is natural to compute the divergence on the sigma-field $\mathcal{X}$ generated by X, $\boldsymbol{I}(\boldsymbol{P}^2|\boldsymbol{P}^1;\mathcal{X}) = \mathcal{I}(P_X^2|P_X^1)$ given by formula (3.1). Suppose that we have an observation of X under $\boldsymbol{P}^1$ which is right-censored at a fixed time C. We observe $(\tilde{X},\delta)$ where

$\tilde{X} = \min(X, C)$ and $\delta = 1_{\{X \leq C\}}$. Thus on $\{X \leq C\}$ we observe all the events of $\mathcal{X}$ but on $\{X > C\}$ we observe no more events. If we represent the observation by the sigma-field $\mathcal{O}$, we can say that $\mathcal{O}$ is generated by $(\tilde{X}, \delta)$. It is clear that we have $\mathcal{O} \subset \mathcal{X}$. Although in theory it is still interesting to compute the divergence of $\boldsymbol{P}^2$ with respect to $\boldsymbol{P}^1$ on the sigma-field $\mathcal{X}$ it is also interesting to compute it on the observed sigma-field, which is $\boldsymbol{I}(\boldsymbol{P}^2|\boldsymbol{P}^1; \mathcal{O})$. It can be proved by simple probabilistic arguments that on $\{X \leq C\}$ we have $\frac{d\boldsymbol{P}^1}{d\boldsymbol{P}^2}_{|\mathcal{O}} = \frac{f_X^1(X)}{f_X^2(X)}$ and on $\{X > C\}$ we have $\frac{d\boldsymbol{P}^1}{d\boldsymbol{P}^2}_{|\mathcal{O}} = \frac{S_X^1(C)}{S_X^2(C)}$ and thus

$$\boldsymbol{I}(\boldsymbol{P}^2|\boldsymbol{P}^1; \mathcal{O}) = \int_0^C \log \frac{f_X^1(x)}{f_X^2(x)} \, f_X^1(x)dx + \log \frac{S_X^1(C)}{S_X^2(C)} \, S^1(C),$$

where $S_X^1(.)$ and $S_X^2(.)$ are the survival functions of X under $\boldsymbol{P}^1$ and $\boldsymbol{P}^2$ respectively.

4. Statistical models and families

4.1. Statistical families

We consider a subset $\mathcal{P}$ of the probabilities on a measurable space $(\mathcal{S}, \mathcal{A})$. We shall call such a subset a family of probabilities, and we may parameterize this family. Following [22], a parameterization can be represented by a function from a set Θ with values in $\mathcal{P}$: $\theta \to P^\theta$. It is desirable that this function be one-to-one, a property linked to the identifiability issue which will be discussed later in this section. The parameterization associated with the family of probabilities $\mathcal{P}$ can be denoted $\Pi = (P^\theta; \theta \in \Theta)$ and we have $\mathcal{P} = \{P^\theta; \theta \in \Theta\}$. We may denote $\Pi \sim \mathcal{P}$. If $\Pi_1 \sim \mathcal{P}$ and $\Pi_2 \sim \mathcal{P}$, Π_1 and Π_2 are two parameterizations of the same family of probabilities and we may note $\Pi_1 \sim \Pi_2$.

$\mathcal{P}$ is really a family of probabilities and Π a parametrized family of probabilities. We may call them statistical families if the aim of considering such families is to make statistical inference. However, a family of probability on $(\Re, \mathcal{B})$ is not sufficient to specify a statistical model (here, we do not follow [22]). A statistical model depends on the random variables chosen, as exemplified in section 2.

4.2. Statistical models

A family of probabilities on the sample space of an experiment $(\Omega, \mathcal{F})$ will be called a statistical model and a parameterization of this family will be called a parameterized statistical model.

Definition 1. Two parameterized statistical models $\boldsymbol{\Pi} = (\boldsymbol{P}^\theta, \theta \in \Theta)$ on $\mathcal{X}$ and $\boldsymbol{\Pi}' = (\boldsymbol{P}^\gamma, \gamma \in \Gamma)$ on $\mathcal{Y}$ are equivalent (in the sense that they specify the same statistical model) if $\mathcal{X} = \mathcal{Y}$ and they specify the same family of probability on $(\Omega, \mathcal{X})$.

The pair $(X, \mathbf{\Pi})$ of a random variable and a parameterized statistical model induces the parameterized family (of distributions) on $(\Re, \mathcal{B})$: $\Pi_X = (P_X^\theta; \theta \in \Theta)$. Conversely, the pair (X, Π_X) induces $\mathbf{\Pi}$ if $\mathcal{X} = \mathcal{F}$. In that case, we may describe the statistical model by (X, Π_X). Two different random variables X and Y induce two (generally different) parameterized families on $(\Re, \mathcal{B})$, Π_X and Π_Y. Conversely, one may ask whether the pairs (X, Π_X) and (Y, Π_Y) define equal or equivalent parameterized statistical models. We need the definition of "informationally equivalent" random variables (or more generally random elements).

Definition 2. X and Y are informationally equivalent if the sigma-fields $\mathcal{X}$ and $\mathcal{Y}$ generated by X and Y are equal.

Each pair (X, P_X^θ) induces a probability on $(\Omega, \mathcal{X})$ $\boldsymbol{P}^{X,\theta} = P_X^\theta o X$ and thus the pair (X, Π_X) induces the parameterized statistical model $(\boldsymbol{P}^{X,\theta}, \theta \in \Theta)$. Similarly, each pair (Y, P_Y^γ) induces a probability on $(\Omega, \mathcal{Y})$ $\boldsymbol{P}^{Y,\gamma} = P_Y^\gamma o Y$ and the pair (Y, Π_Y) induces the parameterized statistical model $(\boldsymbol{P}^{Y,\gamma}, \gamma \in \Gamma)$. Tautologically, we will say that (X, Π_X) and (Y, Π_Y) define the same statistical models if $(\boldsymbol{P}^{X,\theta}, \theta \in \Theta)$ and $(\boldsymbol{P}^{Y,\gamma}, \gamma \in \Gamma)$ are equivalent.

Example 1 (continued). (i) $\Pi_X = (\mathcal{N}(10^3; \sigma^2), \sigma^2 > 0)$ and $\Pi_Y = (\mathcal{N}(10^3; \sigma^2), \sigma^2 > 0)$ are the same parameterized families on $(\Re, \mathcal{B})$. However, since X and Y are measurements of the same quantity in different units, these parameterized families correspond to different statistical models.

(ii) $\Pi_X = (\mathcal{N}(\mu, \sigma^2); \mu \in \Re, \sigma^2 > 0)$ and $\Pi_Y = (\mathcal{N}(\mu, \sigma^2); \mu \in \Re, \sigma^2 > 0)$ are the same parameterized family on $(\Re, \mathcal{B})$. (X, Π_X) and (Y, Π_Y) specify the same statistical model but not the same parameterized statistical model.

(iii) $\Pi_X = (\mathcal{N}(10^3; \sigma^2), \sigma^2 > 0)$ and $\Pi_Y = (\mathcal{N}(10^6; 10^6\sigma^2), \sigma^2 > 0)$ are different families on $(\Re, \mathcal{B})$. However (X, Π_X) and (Y, Π_Y) specify the same statistical model (with the same parameterization).

For sake of simplicity we have considered distributions of real random variables. The same can be said about random variables with values in $\Re^d$ or stochastic processes that are random elements with values in a Skorohod space. Commenges and Gégout-Petit [6] gave an instance of two informationally equivalent processes. The events described by an irreversible three-state process $X = (X_t)$, where X_t takes values $0, 1, 2$, can be described by a bivariate counting process $N = (N_1, N_2)$. The law of the three-state process is specified by the transition intensities $\alpha_{01}, \alpha_{02}, \alpha_{12}$. There is a way of expressing the intensities λ_1 and λ_2 of N_1 and N_2 such that the laws of X and N correspond to the same probability on $(\Omega, \mathcal{F})$. Thus the same statistical model can be described with X or with N.

4.3. Statistical models and true probability

So-called objectivist approaches to statistical inference assume that there is a true, generally unknown, probability $\boldsymbol{P}^*$. Frequentists as well as objectivist Bayesians adopt this paradigm while subjectivist Bayesians, following De Finetti [11], reject it. We adopt the objectivist paradigm, which in our view is more

suited to answer scientific issues. Statistical inference aims to approach $\boldsymbol{P}^*$ or functionals of $\boldsymbol{P}^*$. Model $\boldsymbol{\Pi}$ is well specified if $\boldsymbol{P}^* \in \boldsymbol{\Pi}$ and is mis-specified otherwise. If it is well specified, then there is a $\theta^* \in \Theta$ such that $\boldsymbol{P}^{\theta^*} = \boldsymbol{P}^*$. If we consider a probability $\boldsymbol{P}^\theta$, we may measure its divergence with respect to $\boldsymbol{P}^*$ on a given sigma-field $\mathcal{O}$ by $\boldsymbol{I}(\boldsymbol{P}^\theta|\boldsymbol{P}^*;\mathcal{O})$, and we may choose θ that minimizes this divergence. We assume that there exists a value θ_{opt} that minimizes $\boldsymbol{I}(\boldsymbol{P}^\theta|\boldsymbol{P}^*;\mathcal{O})$. We call $\boldsymbol{I}(\boldsymbol{P}^{\theta_{opt}}|\boldsymbol{P}^*;\mathcal{O})$ the misspecification risk of model Π. Of course, if the model is well specified, then $\boldsymbol{I}(\boldsymbol{P}^\theta|\boldsymbol{P}^*;\mathcal{O})$ is minimized at θ^*, and the misspecification risk is null.

5. The likelihood

5.1. Definition of the likelihood

Conventionally, most statistical models assume that independently identically distributed (i.i.d.) random variables, say $X_i, i = 1, \ldots, n$, are observed. However, in case of complex observation schemes, the observed random variables become complicated. Moreover the same statistical model can be described by different random variables. For instance, in Example 2 the observed random variables are the pairs $(\tilde{X}_i, \delta_i)$. However, we may also describe the observation by $(\delta_i X_i, \delta_i)$, or in terms of counting processes by $(N_u^i, 0 \leq u \leq C)$, where $(N_u^i = 1_{\{X_i \leq u\}})$. These three descriptions are observationally equivalent, in the sense that they correspond to the same sigma-field, say $\mathcal{O}_i = \sigma(\tilde{X}_i, \delta_i) = \sigma(\delta_i X_i, \delta_i) = \sigma(N_u^i, 0 \leq u \leq C)$.

We shall adopt the description of observations in terms of sigma-fields because it is more intrinsic. We shall work with a measure space $(\Omega, \mathcal{F})$ containing all events of interest. For instance the observation of subject i, $\mathcal{O}_i$, belongs to $\mathcal{F}$. Saying that observations are i.i.d. means that the $\mathcal{O}_i$ are independent, that there is a one-to-one correspondence between $\mathcal{O}_i$ and $\mathcal{O}_{i'}$ and that the restrictions of $\boldsymbol{P}^*$ to $\mathcal{O}_i$ (denoted $\boldsymbol{P}^*_{\mathcal{O}_i}$) are the same. We call $\bar{\mathcal{O}}_n$ the global observation: $\bar{\mathcal{O}}_n = \vee_{i=1}^n \mathcal{O}_i$. Since we do not know $\boldsymbol{P}^*$, we may in the first place reduce the search by restricting our attention to a statistical model $\boldsymbol{\Pi}$ and find a $\boldsymbol{P}^\theta \in \boldsymbol{\Pi}$ close to $\boldsymbol{P}^*$, that is, one which minimizes $\boldsymbol{I}(\boldsymbol{P}^\theta|\boldsymbol{P}^*;\mathcal{O}_i)$. We have already given a name to it, $\boldsymbol{P}^{\theta_{opt}}$, but we cannot compute it directly because we do not know $\boldsymbol{P}^*$. The problem is that $\boldsymbol{I}(\boldsymbol{P}^\theta|\boldsymbol{P}^*;\mathcal{O}_i)$ doubly depends on the unknown $\boldsymbol{P}^*$: (i) through the Radon-Nikodym derivative and (ii) through the expectation. Problem (i) can be eliminated by noting that $L_{\mathcal{O}_i}^{\boldsymbol{P}^*/\boldsymbol{P}^\theta} = L_{\mathcal{O}_i}^{\boldsymbol{P}^*/\boldsymbol{P}^0} + L_{\mathcal{O}_i}^{\boldsymbol{P}^0/\boldsymbol{P}^\theta}$. Thus, by taking expectation under $\boldsymbol{P}^*$:

$$\boldsymbol{I}(\boldsymbol{P}^\theta|\boldsymbol{P}^*;\mathcal{O}_i) = \boldsymbol{I}(\boldsymbol{P}^0|\boldsymbol{P}^*;\mathcal{O}_i) - \mathrm{E}_{\boldsymbol{P}^*}\big(L_{\mathcal{O}_i}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0}\big).$$

Minimizing $\boldsymbol{I}(\boldsymbol{P}^\theta|\boldsymbol{P}^*;\mathcal{O}_i)$ is equivalent to maximizing $\mathrm{E}_{\boldsymbol{P}^*}(L_{\mathcal{O}_i}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0})$. We cannot compute $\mathrm{E}_{\boldsymbol{P}^*}(L_{\mathcal{O}_i}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0})$, but we can estimate it. The law of large numbers tells

us that, when $n \to \infty$:

$$n^{-1} \sum_{i=1}^{n} L_{\mathcal{O}_i}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0} \to \mathrm{E}_{\boldsymbol{P}^*}\left(L_{\mathcal{O}_i}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0}\right).$$

Thus, we may maximize the estimator on the left hand, which is the loglikelihood $L_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0}$ divided by n. Maximizing the loglikelihood is equivalent to maximizing the likelihood function $\mathcal{L}_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0}$. The likelihood function is the function $\theta \to \mathcal{L}_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0}$. In conclusion, the maximum likelihood estimator (MLE) can be considered as an estimator that minimizes a natural estimator of the Kullback-Leibler risk.

5.2. Computation of the likelihood

Computation of the likelihood is simple in terms of the probability on the observed σ-field. The conventional way of specifying a model is in terms of a random variable and a family of distributions $(X, (f_X^\theta(.))_{\theta\in\Theta})$. Then the likelihood for observation X is simply $f_X^\theta(X)$. When the events of interest are represented by stochastic processes in continuous time, it is also possible to define a density and hence a likelihood function. See [15] for diffusion processes and [23] for counting processes.

Two situations make the computation of the likelihood more complex. The first is when there is incomplete observation of the events of interest. If the mechanism leading to incomplete data is random we should in principle model it. The theory of ignorable missing observation of Rubin [38] has been extended to more general mechanisms leading to incomplete data in [20]. This has been developed in the stochastic process framework by Commenges and Gégout-Petit [5] (who also give some general formulas for likelihood calculus). The second situation occurs when the law is described through a conditional probability and the conditioning events are not observed. This is the framework of random effects models (see section 7.1). Although conceptually different these two situations lead to the same problem: the likelihood for subject i can be relatively easily computed for a "complete" observation $\mathcal{G}_i$ and the likelihood for the observation $\mathcal{O}_i \subset \mathcal{G}_i$ is the conditional expectation (which derives from the fundamental formula):

$$\mathcal{L}_{\mathcal{O}_i}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0} = \mathrm{E}_{\boldsymbol{P}^0}\left[\mathcal{L}_{\mathcal{G}_i}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0} | \mathcal{O}_i\right]. \tag{5.1}$$

The conditional expectation is expressed as an integral which must be computed numerically in most cases. The only notable exception is the linear mixed effects model where the integral can be analytically computed. For examples of algorithms for non-linear mixed effects see [12] and [19]. For general formulas for the likelihood of interval-censored observations of counting processes see [6].

5.3. Performance of the MLE in terms of Kullback-Leibler risk

We expect good behavior of the MLE $\hat{\theta}$ when the law of large numbers can be applied and when the number of parameters is not too large. Some cases of unsatisfactory behavior of the MLE are reported for instance in [30]. The properties of the MLE may not be satisfactory when the number of parameters is too large, and especially when it increases with n as in an example given by Neymann and Scott [36]. In this example $(X_i, Y_i), i = 1, \ldots, n$ are all independent random variables with X_i and Y_i both normal $N(\xi_i, \sigma^2)$. It is readily seen that not only the MLE of $\xi_i, i = 1 \ldots, n$, but also the MLE of σ^2 are inconsistent. This example is typical of problems where there are individual parameters (a ξ_i for each i), so that in fact the statistical model changes with n. Such situations are better approached by random effects models.

To assess the performance of the MLE we can use a risk which is an extended version of the Kullback-Leibler risk with respect to $\boldsymbol{P}^*$:

$$\mathrm{EKL}(\boldsymbol{P}^{\hat{\theta}}, \mathcal{O}_i) = \mathrm{E}_{\boldsymbol{P}^*}\left(L_{\mathcal{O}_i}^{\boldsymbol{P}^*/\boldsymbol{P}^{\hat{\theta}}}\right).$$

The difference with the classical Kullback-Leibler risk is that here $\boldsymbol{P}^{\hat{\theta}}$ is random: so $\mathrm{EKL}(\boldsymbol{P}^{\hat{\theta}}, \mathcal{O}_i)$ is the expectation of the Kullkack-Leibler divergence between $\boldsymbol{P}^{\hat{\theta}}$ and $\boldsymbol{P}^*$. In parametric models (that is, Θ is a subset of $\Re^p$), it can be shown [9, 35] that

$$\mathrm{EKL}(\boldsymbol{P}^{\hat{\theta}}, \mathcal{O}_i) = \mathrm{E}_{\boldsymbol{P}^*}[L_{\mathcal{O}_i}^{\boldsymbol{P}^*/\boldsymbol{P}^{\theta_{opt}}}] + \frac{1}{2}n^{-1}\mathrm{Tr}(I^{-1}J) + o(n^{-1}), \tag{5.2}$$

where I is the information matrix and J is the variance of the score, both computed in θ_{opt}; here the symbol Tr means the trace. This can be nicely interpreted by saying that the risk $\mathrm{EKL}(\boldsymbol{P}^{\hat{\theta}}, \mathcal{O}_i)$ is the sum of the misspecification risk $\mathrm{E}_{\boldsymbol{P}^*}[L_{\mathcal{X}}^{\boldsymbol{P}^*/\boldsymbol{P}^{\theta_{opt}}}]$ and the statistical risk $\frac{1}{2}n^{-1}\mathrm{Tr}(I^{-1}J)$. Note in passing that if $\boldsymbol{\Pi}$ is well specified we have $\mathrm{E}_{\boldsymbol{P}^*}[L_{\mathcal{O}_i}^{\boldsymbol{P}^*/\boldsymbol{P}^{\theta_{opt}}}] = 0$ and $I = J$, and thus $\mathrm{EKL}(\boldsymbol{P}^{\hat{\theta}}, \mathcal{O}_i) = \frac{p}{2n} + o(n^{-1})$.

6. The penalized likelihood

There is a large literature on the topic: see [13, 14, 17, 18, 21, 25, 37, 44] among others. Penalized likelihood is useful when the statistical model is too large to obtain good estimators, while conventional parametric models appear too rigid. A simple form of the penalized log-likelihood is

$$pl_\kappa(\theta) = L_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0} - \kappa J(\theta),$$

where $J(\theta)$ is a measure of dislike of θ and κ weights the influence of this measure on the objective function. A classical example is when $\theta = (\alpha(.), \beta)$, where $\alpha(.)$

is a function and β is a real parameter. $J(\theta)$ can be chosen as

$$J(\theta) = \int_0^\infty \alpha''(u)^2 du.$$

In this case $J(\theta)$ measures the irregularity of the function $\alpha(.)$. The maximum penalized likelihood estimator (MpLE) θ_κ^{pl} is the value of θ which maximizes $pl_\kappa(\theta)$. κ is often called a smoothing coefficient in the cases where $J(\theta)$ is a measure of the irregularity of a function. More generally, we will call it a meta-parameter. We may generalize the penalized log-likelihood by replacing $\kappa J(\theta)$ by $J(\theta, \kappa)$, where κ could be multidimensional. When κ varies, this defines a family of estimators,$(\theta_\kappa^{pl}; \kappa \geq 0)$. κ may be chosen by cross-validation (see section 8).

There is another way of dealing with the problem of statistical models that might be too large. This is by using the so-called sieve estimators [40]. Sieves are based on a sequence of approximating spaces. For instance rather than working with a functional parameter we may restrict to spaces where the function is represented on a basis (e.g. a splines basis). Here we consider a special sieves approach where the approximating spaces may be functional spaces. Consider a family of models $(\mathcal{P}_\nu)_{\nu \geq 0}$ where:

$$\mathcal{P}_\nu = (\boldsymbol{P}^\theta; \theta \in \Theta : J(\theta) \leq \nu).$$

For fixed ν, the MLE $\hat{\theta}_\nu$ solves the constrained maximization problem:

$$\max L_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0}; \text{ subject to } J(\theta) \leq \nu. \tag{6.1}$$

When ν varies this defines a family of sieve estimators: $(\hat{\theta}_\nu; \nu \geq 0)$. $\hat{\theta}_\nu$ maximizes the Lagrangian $L_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0} - \lambda[J(\theta) - \nu]$ for some value of λ. The Lagrangian superficially looks like the penalized log-likelihood function, but an important difference is that here the Lagrange multiplier λ is not fixed and is a part of the solution. If the problem is convex the Karush-Kuhn-Tucker conditions are necessary and sufficient. Here these conditions are

$$J(\theta) \leq \nu; \lambda \geq 0; \ \frac{\partial L_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0}}{\partial \theta} - \lambda \frac{\partial J(\theta)}{\partial \theta} = 0. \tag{6.2}$$

It is clear that when the observation $\bar{\mathcal{O}}_n$ is fixed, the function $\kappa \to J(\theta_\kappa^{pl})$ is a monotone decreasing function. Consider the case where this function is continuous and unbounded (when $\kappa \to 0$). Then for each fixed ν there exists a value, say κ_ν, such that $J(\theta_{\kappa\nu}^{pl}) = \nu$. Note that this value depends on $\bar{\mathcal{O}}_n$. Now, it is easy to see that $\theta_{\kappa\nu}^{pl}$ satisfies the Karush-Kuhn-Tucker conditions (6.2), with $\lambda = \kappa_\nu$. Thus, if we can find the correct κ_ν we can solve the constrained maximization problem by maximizing the corresponding penalized likelihood. However, the search for κ_ν is not simple, and we must remember that the relationship between ν and κ_ν depends on $\bar{\mathcal{O}}_n$. A simpler result, deriving from the previous considerations, is:

Lemma 6.1 (Penalized and sieves estimators). *The families* $(\boldsymbol{P}^{\theta^{pl}_{\kappa}}; \kappa \geq 0)$ *and* $(\boldsymbol{P}^{\hat{\theta}_{\nu}}; \nu \geq 0)$ *are identical families of estimators.*

The consequence is that since it is easier to solve the unconstrained maximization problem involved in the penalized likelihood approach, one should apply this approach in applications. On the other hand, it may be easier to develop asymptotic results for sieve estimators (because $\hat{\theta}_{\nu}$ is a MLE) than for penalized likelihood estimators. One should be able to derive properties of penalized likelihood estimators from those of sieve estimators.

7. The hierarchical likelihood

7.1. Random effects models

An important class of models arises when we define a potentially observable variable Y_i for each subject, and its distribution is given conditionally on unobserved quantities. This is the classical framework of random effects models, which we have already mentioned in subsections 5.2 and 5.3. Specifically, let us consider the following model: conditionally on b^i, Y_i has a density $f_{Y|b}(.;\theta, b^i)$, where θ is a vector of parameters of dimension m and b^i are random effects (or parameters) of dimension K. The (Y_i, b^i) are i.i.d. Typically Y_i is multivariate of dimension n_i. We assume that the b^i have density $f_b(.;\tau)$, where τ is a parameter. Typically Y_i is observed, while b^i is not. This can be made more general for including the case of censored observation of Y_i.

The conventional approach for estimating θ is to compute the maximum likelihood estimators. Empirical Bayes estimators of the b^i can be computed in a second stage. The likelihood (for observation i) is computed by taking the conditional expectation given $\mathcal{O}_i$ of the complete likelihood on the sigma-field including the random effect $\mathcal{G}_i = \mathcal{O}_i \vee \sigma(b_i)$. This is an application of formula (5.1). Practically the computation of this conditional expectation involves the integrals $\int f^{\theta}_{Y|b}(Y_i|b) f_b(b) db$. Random effects models have been thoroughly studied in both linear [43] and non-linear [10] cases. While in the linear case computation of the above integrals is analytical, in the non-linear case it is not. The numerical computation of these multiple integrals of dimension K is a daunting task if K is larger than 2 or 3, especially if the likelihood given the random effects is not itself very easy to compute; this is the curse of dimensionality.

7.2. Hierarchical likelihood

For hierarchical generalized linear models, the hierarchical likelihood (or h-likelihood), was proposed by Lee and Nelder [32]; see also [33, 34]. The h-likelihood is the joint (or complete) likelihood of the observations and the (unobserved) random effects, but where the random effects are treated as parameters. The complete loglikelihood is $L_{\bar{\mathcal{G}}_n \boldsymbol{P}^{\theta}/\boldsymbol{P}^{0}}$. It can be decomposed into

$L_{\bar{\mathcal{G}}_n}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0} = L_{\bar{\mathcal{G}}_n|b}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0} + L_b^{\boldsymbol{P}^\theta/\boldsymbol{P}^0}$; the last term can be written $\sum_{i=1}^n \log f_b(b^i;\tau)$. None of these likelihoods can be computed (is measurable for) $\bar{\mathcal{O}}_n$. The h-loglikelihood function is the function $\gamma \to L_{\bar{\mathcal{G}}_n}^{\boldsymbol{P}^\theta/\boldsymbol{P}^0}$ where $\gamma = (\theta, b)$ is the set of all the "parameters". Thus, estimators (here denoted MHLE) of both θ and b can be obtained by maximizing the h-loglikelihood:

$$hl_\tau(\gamma) = L_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^\gamma/\boldsymbol{P}^0} - \sum_{i=1}^n \log f_b(b^i;\tau).$$

Often the loglikelihood can be written $L_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^\gamma/\boldsymbol{P}^0} = \sum_i^n \log f(Y_i;\theta,b^i)$. However, this formulation is not completely general, because there are interesting cases where observations of the Y_i are censored. So, we prefer writing the loglikelihood as $L_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^\gamma/\boldsymbol{P}^0}$. We note $\hat{\gamma}_\tau = (\hat{\theta}_\tau, \hat{b}_\tau)$ the maximum h-likelihood estimators of the parameters for given τ; the latter (meta) parameter can be estimated by profile likelihood. The main interest of this approach is that there is no need to compute multiple integrals. This problem is replaced by that of maximizing $hl_\tau(\gamma)$ over γ. That is, the problem is now a large number of parameters that must be estimated, which this is equal to $m+nK$. This may be large, but special algorithms can be used for generalized linear models.

Therneau and Grambsch [41] used the same approach for fitting frailty models, calling it a penalized likelihood. It may superficially look like the penalized quasi likelihood of Breslow and Clayton [2], but this is not the same thing. There is a link with the more conventional penalized likelihood for estimating smooth functions discussed in section 6. The h-likelihood can be considered as a penalized likelihood but with two important differences relative to the conventional one: (i) the problem is parametric; (ii) the number of parameters grows with n. Commenges et al. [9] have proved that the maximum h-likelihood estimators for the fixed parameters are M-estimators [42]. Thus, under some regularity conditions they have an asymptotic normal distribution. However, this asymptotic distribution is not in general centered on the true parameter values, so that the estimators are biased. In practice the bias can be negligible so that this approach can be interesting in some situations due to its relative numerical simplicity.

8. Akaike and likelihood cross-validation criteria

An important issue is the choice between different estimators. Two typical situations are: (i) choice of MLE's in different models; (ii) choice of MpLE's with different penalties. If we consider two models $\boldsymbol{\Pi}$ and $\boldsymbol{\Pi}'$ we get two estimators $\boldsymbol{P}^{\hat{\theta}}$ and $\boldsymbol{P}^{\hat{\gamma}}$ of the probability $\boldsymbol{P}^*$, and we may wish to assess which is better. This is the "model choice" issue. A penalized likelihood function produces a family of estimators $(\boldsymbol{P}^{\theta_\kappa^{pl}}; \kappa \geq 0)$, and we may wish to choose the best. Here, what we call "the best" estimator is the estimator that minimizes some risk function; in both cases we can use the extended version of the Kullback-Leibler

risk already used in section 5:

$$\mathrm{EKL}(\boldsymbol{P}^{\hat{\theta}};\mathcal{O}_i) = \mathrm{E}_{\boldsymbol{P}^*}(L_{\mathcal{O}_i}^{\boldsymbol{P}^*/\boldsymbol{P}^{\hat{\theta}}}).$$

Since $\boldsymbol{P}^*$ is unknown we can first work with $\mathrm{E}_{\boldsymbol{P}^*}(L_{\mathcal{O}_i}^{\boldsymbol{P}^0/\boldsymbol{P}^{\hat{\theta}}})$, which is equal, up to a constant, to $\mathrm{EKL}(\boldsymbol{P}^{\hat{\theta}};\mathcal{O}_i)$. Second we can, as usual, replace the expectation under $\boldsymbol{P}^*$ by expectation under the empirical distribution. For parametric models, Akaike [1] has shown that an estimator of $\mathrm{E}_{\boldsymbol{P}^*}(L_{\mathcal{O}_i}^{\boldsymbol{P}^0/\boldsymbol{P}^{\hat{\theta}}})$ was $-n^{-1}(L_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^{\hat{\theta}}/\boldsymbol{P}^0} - p)$, and Akaike criterion (AIC) can be deduced by multiplying this quantity by $2n$: $\mathrm{AIC} = -2L_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^{\hat{\theta}}/\boldsymbol{P}^0} + 2p$. Other criteria have been proposed for model choice, and for more detail about Akaike and other criteria we refer to [3, 27, 35]. Here, we pursue Akaike's idea of estimating the Kullback-Leibler risk. It is clear that the absolute risk itself can not in general be estimated. However, the difference of risks between two estimators in parametric models $\Delta(\boldsymbol{P}^{\hat{\theta}},\boldsymbol{P}^{\hat{\gamma}}) = \mathrm{EKL}(\boldsymbol{P}^{\hat{\theta}};\mathcal{O}_i) - \mathrm{EKL}(\boldsymbol{P}^{\hat{\gamma}};\mathcal{O}_i)$ can be estimated by the statistic $D(\boldsymbol{P}^{\hat{\theta}},\boldsymbol{P}^{\hat{\gamma}}) = (1/2n)(AIC(\boldsymbol{P}^{\hat{\theta}}) - AIC(\boldsymbol{P}^{\hat{\gamma}}))$ and a more refined analysis of the difference of risks can be developed, as in [9].

The leave-one-out likelihood cross-validation criterion can also be considered as a possible "estimator" up to a constant of EKL [7]. It is defined as:

$$LCV(\boldsymbol{P}^{\hat{\theta}_n};\mathcal{O}_{n+1}) = -\frac{1}{n}\sum_{i=1}^{n} L_{\mathcal{O}_i}^{\boldsymbol{P}^{\hat{\theta}(\bar{\mathcal{O}}_{n|i})}/\boldsymbol{P}^0},$$

where $\bar{\mathcal{O}}_{n|i} = \vee_{j\neq i}\mathcal{O}_j$ and $\mathcal{O}_{n+1}$ is another i.i.d. replicate of $\mathcal{O}_i$. Then, we define an estimator of the difference of risks between two estimators:

$$\Delta(\boldsymbol{P}^{\hat{\theta}},\boldsymbol{P}^{\hat{\gamma}}) = LCV(\boldsymbol{P}^{\hat{\theta}_n};\mathcal{O}_{n+1}) - LCV(\boldsymbol{P}^{\hat{\gamma}_n};\mathcal{O}_{n+1}) \tag{8.1}$$

The advantage of LCV is that it can be used for comparing smooth estimators in nonparametric models, and in particular it can be used for choosing the penalty weight in penalized likelihood. A disadvantage is the computational burden, but a general approximation formula has been given ([7, 37]):

$$LCV \approx -n^{-1}[L_{\bar{\mathcal{O}}_n}^{\boldsymbol{P}^{\hat{\theta}}/\boldsymbol{P}^0} - \mathrm{Tr}(H_{pl_\kappa}^{-1}H_{L_{\bar{\mathcal{O}}_n}})],$$

where $H_{L_{\bar{\mathcal{O}}_n}}$ and H_{pl_κ} are the Hessian of the loglikelihood and penalized log-likelihood respectively. This expression looks like an AIC criterion and there are arguments to interpret $\mathrm{Tr}[H_{pl_\kappa}^{-1}H_{L_{\bar{\mathcal{O}}_n}}]$ as the model degree of freedom.

9. Link with the MAP estimator

One important issue is the relationship between the three likelihoods considered here and the Bayesian approach. The question arises because it seems that these

three likelihoods can be identified with the numerator of *a posteriori* distributions with particular priors. Thus MLE, MpLE and MHLE could be identified with the maximum a posteriori (MAP) estimators with the corresponding priors. However, this relationship depends on the parameterization. Thus the MLE is identical to the MAP using a uniform prior for the parameters. If we change the parameterization, the uniform prior on the new parameters does not correspond in general to the uniform prior on the original parameters, as was already noticed by Fisher [16]. This apparent paradox led Jeffreys to propose a prior invariant for parameterization [24], known as Jeffrey's prior. However the MAP with Jeffreys's prior is no longer identical to the MLE when Jeffreys's prior is not uniform. For instance, for the parameter of a binomial trial, Jeffreys's prior is $1/\sqrt{p(1-p)}$. Adding the logarithm of this term to the loglikelihood shifts the maximum away from 0.5. Moreover it is questionable whether this invariance property can be identified with a non-informativeness character of this prior (for a review on the choice of priors, see [26]).

In the Bayesian paradigm, rather than considering estimators based on maximization of some expression such as the likelihood or posterior density, it is common to attempt to summarize the statistical inferences by using quantiles of the posterior distribution, such as the median, or expectations with respect to the posterior. While such estimators may be more satisfactory, they typically involve multiple integrals that are hard to compute: computations are mostly being done with the MCMC algorithm. Maximization methods have the advantage of being potentially easier in the case where multiple integrals can be avoided. There are also approximate Bayesian methods, which yield the a posteriori marginal distribution by approximating some of the multiple integrals by Laplace approximation, which in turn involves a maximization problem. Rue et al. [39] claim that this approach is much faster than the MCMC algorithm.

Conclusion

The Kolmogorov representation of a statistical experiment has to be taken seriously if we want to have a deep understanding of what a statistical model is. The Kullback-Leibler risk is underlying most of the reflexions about likelihood, as was clearly seen by Akaike [1]. Finally, the link with the Bayesian approach should be explored more thoroughly than could done in this paper. The MLE and MAP estimators are the same if, in a given paramterization, the prior used for the MAP is uniform. However, this identity is not stable with respect to reparameterizations. Similar remarks hold for the link between penalized likelihood and MAP.

Acknowledgements

I would like to thank Anne Gégout-Petit for helpful comments on the manuscript.

References

[1] Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. *Second International Symposium on Information Theory, Akademia Kiado.* 267–281. MR0483125

[2] Breslow, N.E. and Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. *J. Amer. Statist. Assoc.* **88**, 9–25.

[3] Burnham, K.P. and Anderson, D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Methods Res.* **33**, 261–304. MR2086350

[4] Cencov, N.N. (1982). *Statistical decisions rules and optimal inference.* American Mathematical Society. MR0645898

[5] Commenges, D. and Gégout-Petit, A. (2005). Likelihood inference for incompletely observed stochastic processes: general ignorability conditions. arXiv:math.ST/0507151.

[6] Commenges, D. and Gégout-Petit, A. (2007). Likelihood for generally coarsened observations from multi-state or counting process models. *Scand. J. Statist.* **34**, 432–450. MR2346649

[7] Commenges, D., Joly, P., Gégout-Petit, A. and Liquet, B. (2007). Choice between semi-parametric estimators of Markov and non-Markov multi-state models from generally coarsened observations. *Scand. J. Statist.* **34**, 33–52. MR2325241

[8] Commenges, D., Jolly, D., Putter, H. and Thiébaut, R. (2009). Inference in HIV dynamics models via hierarchical likelihood. *Submitted.*

[9] Commenges, D., Sayyareh, A., Letenneur, L., Guedj, J. and Bar-Hen, A. (2008). Estimating a difference of Kullback-Leibler risks using a normalized difference of AIC. *Ann. Appl. Statist.* **2**, 1123–1142.

[10] Davidian, M. and Giltinan, D.M. (2003). Nonlinear models for repeated measurement data: an overview and update, *J. Agric. Biol. Environ. Statist.* **8**, 387–419.

[11] De Finetti, B. (1974). *Theory of Probability.* Chichester: Wyley.

[12] Delyon B., Lavielle, M. and Moulines, E. (1999). Convergence of a Stochastic Approximation Version of the EM Algorithm. *Ann. Statist.* **27**, 94–128. MR1701103

[13] Eggermont, P. and Lariccia, V. (1999). Optimal convergence rates for Good's nonparametric likelihood density estimator. *Ann. Statist.* **27**, 1600–1615. MR1742501

[14] Eggermont, P. and Lariccia, V. (2001). *Maximum penalized likelihood estimation.* New-York: Springer-Verlag. MR1837879

[15] Feigin, P.D. (1976). Maximum likelihood estimation for continuous-time stochastic processes. *Adv. Appl. Prob.* **8**, 712–736. MR0426342

[16] Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Phil. Trans. Roy. Soc. A* **222**, 309–368.

[17] Good, I.J. and Gaskin, R.A. (1971). Nonparametric roughness penalty for probability densities. *Biometrika* **58**, 255–277. MR0319314

[18] Gu, C. and Kim, Y. J. (2002). Penalized likelihood regression.: general formulation and efficient approximation. *Can. J. Stat.* **30**, 619–628. MR1964431

[19] Guedj, J., Thiébaut, R. and Commenges, D. (2007). Maximum likelihood estimation in dynamical models of HIV. *Biometrics* **63**, 1198–1206. MR2414598

[20] Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and coarse data. *Ann. Statist.* **19**, 2244–2253. MR1135174

[21] Hastie, T. and Tibshirani, R. (1990). *Generalized additive models.* London: Chapman and Hall. MR1082147

[22] Hoffmann-Jorgensen, J. (1994). *Probability with a view toward statistics.* London: Chapman and Hall.

[23] Jacod, J. (1975). Multivariate point processes: predictable projection; Radon-Nikodym derivative, representation of martingales. *Z. Wahrsch. Verw. Geb.* **31**, 235–253. MR0380978

[24] Jeffreys, H. (1961). *Theory of probability.* Oxford University Press. MR0187257

[25] Joly, P. and 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.

[26] Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules *J. Amer. Statist. Assoc.* **91**, 1343–1370.

[27] Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling.* New-York: Springer Series in Statistics. MR2367855

[28] Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86. MR0039968

[29] Kullback, S. (1959). *Information Theory.* New-York: Wiley. MR0103557

[30] Le Cam, L. (1990). Maximum Likelihood: An Introduction. *Int. Statist. Rev.* **58**, 153–171.

[31] Lee, Y. and Nelder, J.A. (1992) Likelihood, Quasi-Likelihood and Pseudolikelihood: Some Comparisons. *J. Roy. Statist. Soc. B* **54**, 273–284. MR1157725

[32] Lee, Y. and Nelder, J.A. (1996). Hierarchical Generalized Linear Models. *J. Roy. Statist. Soc. B* **58**, 619–678. MR1410182

[33] Lee, Y. and Nelder, J.A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* **88**, 987–1006. MR1872215

[34] Lee, Y., Nelder, J.A. and Pawitan, Y. (2006). *Generalized linear models with random effects.* Chapman and Hall. MR2259540

[35] Linhart, H. and Zucchini, W. (1986). *Model Selection*, New York: Wiley. MR0866144

[36] Neymann, J. and Scott, E.L. (1988).Consistent estimates based on partially consistent observations. *Econometrika* **16**, 1–32. MR0025113

[37] O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Scient. Statist. Comput.* **9**, 363–379. MR0930052

[38] Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581–592. MR0455196

[39] Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *J. Roy. Statist. Soc. B* **71**, 1–35.

[40] Shen, X. (1997). On methods of sieves and penalization. *Ann. Statist.* **25**, 2555–2591. MR1604416

[41] Therneau, T.M. and Grambsch, P.M. (2000). *Modeling survival data: extending the Cox model.* Springer. MR1774977

[42] van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge.

[43] Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New-York: Springer. MR1880596

[44] Wahba, G. (1983). Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline *J. Roy. Statist. Soc. B* **45**, 133–150. MR0701084

[45] Williams, D. (1991). *Probability with Martingales.* Cambridge University Press. MR1155402