Computational Statistics and Data Analysis 55 (2011) 446–456

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Inference in HIV dynamics models via hierarchical likelihood

D. Commenges [a,d,*], D. Jolly [a,b], J. Drylewicz [a,e], H. Putter [c], R. Thiébaut [a,d]

[a] INSERM, Epidemiology and Biostatistics Research Center, Bordeaux, France
[b] University of Bordeaux, IMB, France
[c] University of Leiden, Department of Medical Statistics and Bioinformatics, The Netherlands
[d] University of Bordeaux 2, ISPED, France
[e] University Medical Center Utrecht, Department of Immunology, The Netherlands

ARTICLE INFO

*Article history:*
Received 29 April 2009
Received in revised form 12 May 2010
Accepted 12 May 2010
Available online 28 May 2010

*Keywords:*
Algorithm
Asymptotic
Differential equations
h-likelihood
HIV dynamics models
Non-linear mixed effects model
Penalized likelihood

ABSTRACT

HIV dynamical models are often based on non-linear systems of ordinary differential equations (ODE), which do not have an analytical solution. Introducing random effects in such models leads to very challenging non-linear mixed-effects models. To avoid the numerical computation of multiple integrals involved in the likelihood, a hierarchical likelihood (h-likelihood) approach, treated in the spirit of a penalized likelihood is proposed. The asymptotic distribution of the maximum h-likelihood estimators (MHLE) for fixed effects is given. The MHLE are slightly biased but the bias can be made negligible by using a parametric bootstrap procedure. An efficient algorithm for maximizing the h-likelihood is proposed. A simulation study, based on a classical HIV dynamical model, confirms the good properties of the MHLE. The method is applied to the analysis of a clinical trial.



## 1. Introduction

Since the influential paper of Ho et al. (1995) there has been a strong impetus to develop mathematical models for better understanding the interaction between HIV and the immune system; see Nowak and May (2000) and Huang (2008). However the statistical inference in these models has raised major challenges coming from the entanglement of identifiability and numerical problems. The first problem is numerical: in general the trajectories of the interesting quantities (e.g. viral load or CD4 counts) are solutions of non-linear differential equations that do not have analytical solutions. The second is the identifiability problem: the observations recorded on one subject are not informative enough to estimate all the parameters of the model. Two types of identifiability problems can be distinguished: the structural and practical identifiability (Rodriguez-Fernandez et al., 2006; Miao et al., 2008, 2009). The first problem is either avoided, simplifying the models to obtain analytical solutions Wu et al. (1998) and Wu and Ding (1999), or solved by using numerical solvers of ordinary differential equations (ODE); Ramsay et al. (2007) proposed an original approach but did not apply it to a random effect model. As for the second problem, both structural and practical identifiability problems can be removed by introducing a priori knowledge on the parameters: this is done either in the frequentist way by fixing values of some parameters, or in the Bayesian way by introducing strong priors. Practical identifiability can be much improved by considering that the particular values of the parameters for each subject are realizations of random variables with a given distribution in the population (Guedj et al., 2007a). This puts the problem in the framework of non-linear mixed effects models. Laplace

* Corresponding address: ISPED, 146 rue Lo Saignat, 33076 Bordeaux, France. Tel.: +33 557571182; fax: +33 556240081.
*E-mail addresses:* daniel.commenges@isped.u-bordeaux2.fr (D. Commenges), danaelle.jolly@gmail.com (D. Jolly), Julia.Drylewicz@isped.u-bordeaux2.fr (J. Drylewicz), H.Putter@lumc.nl (H. Putter), Rodolphe.Thiebaut@isped.u-bordeaux2.fr (R. Thiébaut).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.05.012

approximation of the numerical integrals involved in the computation of the likelihood has been proposed (Beal and Sheiner, 1982; Lindstrom and Bates, 1990). Adaptive Gaussian quadrature is another possibility (Davidian and Giltinan, 1995); see Wu (2005) for a review of statistical issues in HIV models. Recently a stochastic approximation EM (SAEM) algorithm has been proposed (Kuhn and Lavielle, 2005; Donnet and Samson, 2007). In the specific case of HIV dynamics models a Bayesian approach has been proposed by Putter et al. (2002) and Huang et al. (2006), while a special algorithm for computing the likelihood and maximizing it using a Newton-like method has been proposed by Guedj et al. (2007b). However all these methods present difficulties and can be time-consuming.

The hierarchical likelihood (h-likelihood) has been proposed for generalized linear models with random effects by Lee and Nelder (1996) and further studied in Lee and Nelder (2001) and Lee et al. (2006) and for non-linear mixed effects models by Noh and Lee (2008). This is very similar to an approach called penalized likelihood used by McGilchrist and Aisbett (1991) and Therneau and Grambsch (2000) for frailty models. The main idea is to treat the random effects (or the frailties) as parameters and to find estimates of all the parameters by maximizing a function which is essentially the loglikelihood conditional on the random effects minus a penalty term which takes large values if the "random" parameters are very dispersed. Penalized likelihood has also been used for function estimation (O'Sullivan, 1988). The advantage of this approach is that it may avoid computing numerical integrals. The curse of dimensionality is transferred from the dimension of numerical integrals to the dimension of the space on which maximization takes place. There are problems with this approach. One is the asymptotic distribution of the estimators of the fixed parameters; another is the estimation of the variances of the random parameters. Consistency of the maximum h-likelihood estimators (MHLE) has not been proved. It is often suggested to revert to the likelihood to obtain consistent estimators of the fixed parameters, but then the most important benefits of h-likelihood in terms of computational burden is lost. Last but not least is the problem of maximizing a complicated function over several hundred parameters.

The aim of this paper is to develop a (partly non-standard) h-likelihood approach to HIV dynamics models which completely avoids computation of the likelihood. This is in the spirit of penalized likelihood in the sense that it is not the goal to precisely estimate the variances of the random effects. One aim is to study the asymptotic distribution of the MHLE for a given choice of the penalty. Another aim is to find an efficient maximization algorithm.

The paper is organized as follows. In Section 2 a statistical model based on an ODE system in a general form is described, a particular form of which will be used for simulations. In Section 3 the h-likelihood is described and the asymptotic distribution of the MHLE for fixed effects when the number of subjects tends toward infinity is given. A parametric bootstrap procedure to correct the bias of the MHLE is proposed. In Section 4 a strategy for choosing the penalty based on the guess of an upper bound of the variance of the random effects is also proposed. An efficient maximization algorithm is presented in Section 5. Section 6 presents a simulation study. Section 7 presents the analysis of a clinical trial. Section 8 concludes.

## 2. A population dynamics model

### 2.1. A general model for the system

The dynamics of the concentrations of virions and CD4+ T-cells (in short, CD4) in different stages (represented by $\boldsymbol{X}^i(t)$) can be described by an ODE system. A population model, as in Guedj et al. (2007b), allows the values of the parameters to vary between subjects. For subject $i$ with $i = 1, \ldots n$, this can be written:

$$\begin{cases} \dfrac{\mathrm{d}\boldsymbol{X}^i(t)}{\mathrm{d}t} = f(\boldsymbol{X}^i(t), \boldsymbol{\xi}^i), \\ \boldsymbol{X}^i(0) = h(\boldsymbol{\xi}^i), \end{cases} \tag{1}$$

where $\boldsymbol{X}^i(t) = (X_1^i(t), \ldots, X_K^i(t))'$ is the vector of the $K$ state variables (or components); $\boldsymbol{\xi}^i = (\xi_1^i, \ldots, \xi_p^i)$ is a vector of $p$ individual parameters which appears naturally in the ODE system and has generally a biological interpretation. Similarly to generalized (mixed) linear models, a link function is introduced, which relates $\boldsymbol{\xi}^i$ to a linear model involving explanatory variables and random effects:

$$\Psi_l(\xi_l^i) = \tilde{\xi}_l^i = \begin{cases} \phi_l + b_l^i + \boldsymbol{z_l^i}(t)\boldsymbol{\beta_l}, & l = 1, \ldots, R, \\ \phi_l + \boldsymbol{z_l^i}(t)\boldsymbol{\beta_l}, & l = R+1, \ldots, p, \end{cases} \tag{2}$$

where $\phi_l$ is the intercept, $\boldsymbol{z_l^i}(t)$ are vectors of the explanatory variables associated with the fixed effects of the $l$th biological parameter; these explanatory variables may be time-dependent, in which case the ODE system has time-dependent parameters. The $\boldsymbol{\beta_l}$'s are vectors of regression coefficients; $b_i = (b_1^i, \ldots, b_R^i)$ is the individual vector of random effects. It is assumed that $b_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with $\Sigma$ diagonal with diagonal elements $\tau_l^2$. More general models could of course be considered.

### 2.2. Model for the observations

Let $Y_{ijm}$ denote the $j$th measurement of the $m$th observable component for subject $i$ at time $t_{ijm}$; It is assumed that:

$$Y_{ijm} = g_m(\boldsymbol{X}^i(t_{ijm})) + \epsilon_{ijm}, \quad i = 1, \ldots, n, j = 1, \ldots, n_{im}, \tag{3}$$