Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Journal of Statistical Planning and Inference 137 (2007) 2815–2831
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Estimation of parameters in incomplete data models defined by dynamical systems

Sophie Donnet[a], Adeline Samson[b, c, *]

[a] Laboratoire de Mathématiques, Université Paris-Sud, Bat 425, 91400 Orsay, France
[b] INSERM U738, Paris, France
[c] Université Paris 7 Denis Diderot, UFR de Médecine, Paris, France

Received 25 August 2005; received in revised form 6 October 2006; accepted 12 October 2006
Available online 8 February 2007

Abstract

Parametric incomplete data models defined by ordinary differential equations (ODEs) are widely used in biostatistics to describe biological processes accurately. Their parameters are estimated on approximate models, whose regression functions are evaluated by a numerical integration method. Accurate and efficient estimations of these parameters are critical issues. This paper proposes parameter estimation methods involving either a stochastic approximation EM algorithm (SAEM) in the maximum likelihood estimation, or a Gibbs sampler in the Bayesian approach. Both algorithms involve the simulation of non-observed data with conditional distributions using Hastings–Metropolis (H–M) algorithms. A modified H–M algorithm, including an original local linearization scheme to solve the ODEs, is proposed to reduce the computational time significantly. The convergence on the approximate model of all these algorithms is proved. The errors induced by the numerical solving method on the conditional distribution, the likelihood and the posterior distribution are bounded. The Bayesian and maximum likelihood estimation methods are illustrated on a simulated pharmacokinetic nonlinear mixed-effects model defined by an ODE. Simulation results illustrate the ability of these algorithms to provide accurate estimates.

Keywords: Bayesian estimation; Incomplete data model; Local linearization scheme; MCMC algorithm; Nonlinear mixed-effects model; ODE integration; SAEM algorithm

1. Introduction

When a biological or physiological process is measured, the regression function of the statistical model corresponding to the observed data is often derived from a differential equation describing the underlying dynamic process. Difficulties arise when the differential equation has no analytical solution and/or when the parameters of the regression function are random and non-observed. Such example can be found in pharmacokinetics, which aims to study drug evolutions in human organism, this evolution being described by differential systems of compartment interactions. Mixed models,

[*] Corresponding author. INSERM U738, Paris, France. Tel.: +33 140257386; fax: +33 140256773.
E-mail addresses: sophie.donnet@math.u-psud.fr (S. Donnet), adeline.samson@bch.aphp.fr (A. Samson).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2006.10.013

for which regression parameters are considered as random variable and non-observed data, are widely used for the analysis of pharmacokinetic data sets, which have classically repeated measurements in several patients.

This paper aims at providing a general answer to the estimation problem in such statistical incomplete data models.

Let y be the noised observations of a biological process measured at instants $(t_1, \ldots, t_J)$. The biological process is described by the solution g of an ordinary differential equation (ODE), depending on a stochastic non-observed parameter ϕ:

$$y_j = g(t_j, \phi) + \varepsilon_j \quad \text{for } j = 1 \ldots J.$$

We consider that the observable vector Y is part of a so-called complete vector (Y, ϕ). We assume that both Y and (Y, ϕ) have density functions, $p_Y(y; \theta)$ and $p_{Y,\phi}(y, \phi; \theta)$, respectively, depending on a parameter θ belonging to some subset Θ of the Euclidean space $\mathbb{R}^q$. The estimation of the parameter θ has been widely studied when the regression function g has an explicit form. Two approaches can be followed to tackle this challenge, respectively, the maximum likelihood and the Bayesian estimations.

Generally, the maximization of the likelihood of the observations cannot be done in a closed form. Dempster et al. (1977) propose the iterative expectation-maximization (EM) algorithm for incomplete data problems. At the kth iteration, the E-step of EM algorithm computes $Q(\theta|\theta_k) = E(\log p_Y(y; \theta)|y; \theta_k)$ while the M-step determines θ_{k+1} maximizing $Q(\theta|\theta_k)$. For cases where the E-step has no closed form, stochastic versions of EM are introduced. Celeux and Diebolt (1985) introduce the stochastic EM algorithm (SEM). Wei and Tanner (1990) suggest the Monte-Carlo EM (MCEM) estimating $Q(\theta|\theta_k)$ by the averaging of m Monte-Carlo replications. Recently, Wu (2004) emphasizes that MCEM is computationally intensive. As an alternative, Delyon et al. (1999) propose the stochastic approximation EM algorithm (SAEM) replacing the E-step by a stochastic approximation of $Q(\theta|\theta_k)$. These methods require the simulation of the non-observed data ϕ. For cases where this simulation cannot be performed in a closed form, Kuhn and Lavielle (2004) suggest to resort to iterative methods such as Monte-Carlo Markov chain algorithms (MCMC).

The Bayesian approach estimates the posterior distribution $p_{\theta|Y}(\cdot|y)$ of θ, a prior $p_\theta(\cdot)$ being given. Because of the conditional independence structure of $p_{\theta|Y} = \int p_{\theta|Y,\phi} p_{\phi|Y} \, \mathrm{d}\phi$ and $p_{\phi|Y} = \int p_{\phi|Y,\theta} \; p_{\theta|Y} \, \mathrm{d}\theta$, Gelfand and Smith (1990) propose a Gibbs sampling to evaluate these two integrals simultaneously. At iteration k, ϕ_k, a realization of ϕ, is simulated with $p_{\phi|Y}(\cdot, \theta_{k-1})$ followed by θ_k, a realization of θ with $p_{\theta|Y,\phi}(\cdot, \phi_k)$. Consequently, as in maximum likelihood estimation, difficulties arise when the simulation of the conditional distribution cannot be performed in a closed form. For these cases, a Hastings–Metropolis (H–M) algorithm can be included in the Gibbs sampler.

The use of the H–M algorithm in estimation algorithms requires the evaluation of the regression function g at each iteration. When g is a non-analytical solution of a dynamical system, it is evaluated using a numerical integration method. Thus a trade-off between accuracy, stability, and computational cost is required. In this paper, we detail the local linearization (LL) scheme (see e.g. Biscay et al., 1996; Ramos and García-López, 1997; Jimenez, 2002) not only because of its stability performances but also because this scheme can be extended to a so-called modified local linearization scheme, adapted to its inclusion in the H–M algorithm. The estimation algorithms are then applied to an approximate model whose regression function is an approximate solution of the ODE.

The objective of this research is to quantify the error induced by the numerical approximation of the regression function g. The paper is organized as follows. Section 2 defines the original statistical model; the LL scheme and its modified version are detailed; the approximate statistical model resulting from the numerical approximation is introduced. Section 3 focuses on the H–M algorithm to simulate the non-observed data ϕ with the conditional distribution. The error induced by the numerical approximation of g on the conditional distribution is quantified. Section 4 is dedicated to the parameter estimation algorithms. Concerning maximum likelihood and Bayesian estimations, the standard algorithms are adapted to make inference on the approximate model. The error induced by the use of the numerical solving method is bounded, respectively, on the likelihood and the posterior distribution. This error is distinct from the error on the estimates induced by the estimation algorithm which is evaluated by their standard errors. Finally, the SAEM algorithm and the Bayesian Gibbs sampler are applied on a nonlinear mixed-effects model deriving from pharmacokinetics in Section 5.