Computational Statistics and Data Analysis 94 (2016) 193–209

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Gaussian quadrature approximations in mixed hidden Markov models for longitudinal data: A simulation study

Maria Francesca Marino [a], Marco Alfó [b,*]

[a] Dipartimento di Economia, Università degli Studi di Perugia, Italy
[b] Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Italy

ARTICLE INFO

Article history:
Received 30 August 2014
Received in revised form 24 July 2015
Accepted 24 July 2015
Available online 12 August 2015

Keywords:
Hidden Markov models
Time-constant and time-varying random parameters
Adaptive Gaussian quadrature
Exponential family

ABSTRACT

Mixed hidden Markov models represent an interesting tool for the analysis of longitudinal data. They allow to account for both time-constant and time-varying sources of unobserved heterogeneity, which are frequent in this kind of studies. Individual-specific latent features, which may be either constant or varying over time, are included in the linear predictor and lead to a general form of dependence between individual measurements. When a parametric (continuous) distribution is associated to time-constant random parameters, the estimation process requires the calculation of (multiple) integrals. These, generally, have not a closed form and should be numerically approximated. The aim is to compare the standard, the adaptive and the pseudo-adaptive Gaussian quadrature approximations by means of a large scale simulation study, where continuous and discrete responses with (conditional) density in the exponential family are considered. Simulation results show that the approximation error is often substantially reduced when the adaptive quadrature rules are considered in place of the standard one. Such an improvement comes at the cost of a higher computational complexity when the fully adaptive scheme is applied. It is shown that, when a sufficient number of repeated measurements per unit is available, the pseudo-adaptive quadrature represents a convenient compromise between quality of results and computational complexity.

1. Introduction

Longitudinal studies entail repeated measurements from a number of units taken over a known, usually finite, time window. In the regression framework, the presence of unobserved individual characteristics, linked e.g. to omitted covariates, leads to extra variability in the marginal distribution of the response and to dependence between measurements from the same individual. Such unobserved heterogeneity can be either time-varying or time-constant, according to a form of *true*/*spurious contagion* (Heckman, 1981). Because of the former, data variability can be ascribed to the time separation between subsequent measurements: current and future outcomes are directly influenced by the past ones. Because of the latter, differences in the response variable are related to the presence of heterogeneous populations with a different propensity to the event. To account for these sources of extra-variability and dependence, time-constant and time-varying individual random parameters may be added to the model specification. Parametric continuous distributions can be used for both types of random parameters; see Diggle et al. (2002), for references. In a more appealing fashion, the latter can be instead approximated via a discrete latent (hidden) variable with a Markovian structure; the resulting model is referred

* Correspondence to: Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 - Rome, Italy. Tel.: +39 0649910763.
E-mail addresses: mariafrancesca.marino@unipg.it (M.F. Marino), marco.alfo@uniroma1.it (M. Alfó).

http://dx.doi.org/10.1016/j.csda.2015.07.016
0167-9473/

to as a mixed hidden Markov model (mHMM). It is worth noticing that, when the number of hidden states increases, the discrete Markov process may be able to approximate an AR(1)-type continuous distribution.

If parameter estimates are obtained via a maximum likelihood approach, as it is frequent in the presence of latent variables, the EM algorithm can be employed. Zucchini and MacDonald (2009) and Cappé et al. (2005) give general references, while Bartolucci et al. (2012) discuss a comprehensive overview of applications to longitudinal data. When considering a parametric (continuous) distribution for the time-constant random parameters, ML estimation requires the computation of multiple integrals. Apart from the case when a Gaussian distribution is used for both the response and the random parameters (see e.g. Lagona et al., 2014), these integrals cannot be solved analytically and numerical approximation techniques are a potential solution.

Recently, some proposals have been introduced to deal with such an issue. Altman (2007) has discussed standard Gaussian quadrature (GQ) in a direct ML perspective and a Monte Carlo EM (MCEM) algorithm; an original variant of the MCEM algorithm has been proposed by Chaubert-Pereira et al. (2010). Maruotti and Rydén (2009) have suggested to leave the distribution of time-constant random parameters unspecified (as in Aitkin, 1999) and to approximate it through a discrete distribution estimated with a nonparametric maximum likelihood (NPML) approach (see Laird, 1978; Böhning, 1982; Lindsay, 1983a,b). For a general review of mixed hidden Markov models the reader is referred to Maruotti (2011).

Although Altman (2007), Maruotti and Rydén (2009) and Lagona et al. (2014) have discussed general random parameter mHMMs, only random intercepts have been considered in empirical applications and simulation studies. Therefore, a first question is whether this class of models can be easily adapted to handle general random parameters. A further question arises when we consider parametric specifications for the distribution of the individual-specific random parameters with time-constant structure. In the context of mixed parameter models, it is generally acknowledged that standard Gaussian quadrature may produce unsatisfactory approximations and poor estimates. Adaptive (Liu and Pierce, 1994; Pinheiro and Bates, 1995) and pseudo-adaptive schemes (Rizopoulos, 2012) have been introduced to improve the quality of results. Within the adaptive quadrature approaches, standard GQ locations, which are symmetric around zero, are centred and scaled at each step (fully adaptive quadrature) or only at the beginning of the optimization algorithm (pseudo-adaptive quadrature) to relocate the main mass of the integrand at zero. This is shown to reduce the approximation error supplied by the GQ technique. In the framework of multilevel models, Rabe-Hesketh et al. (2002, 2005) have proved, via an extensive simulation study, that the adaptive Gaussian quadrature rule outperforms the standard approach, especially when the intraclass correlation is high. Cagnone and Monari (2013) have compared the fully adaptive and the standard Gaussian approximation in the framework of high-dimensional latent variable models; as the dimension increases, the standard Gaussian quadrature turns out to be less appropriate due to the difficulties in reaching convergence in a reasonable number of iterations. In the context of joint models for longitudinal and time to event data, Rizopoulos (2012) has shown that the pseudo-adaptive scheme leads to accurate parameter estimates with a lower number of locations when compared to the standard scheme, thus consistently reducing the computational load.

To our knowledge, this topic has not been adequately investigated in the context of mHMMs; the aim of this paper is at comparing the standard Gaussian quadrature approach discussed by Altman (2007) with the fully adaptive and the pseudo-adaptive quadrature schemes. To assess the quality of these approximations, we have considered, in a large scale simulation study, responses having conditional Gaussian, Poisson and Bernoulli distribution, with varying sample sizes and number of repeated measurements per unit. The plan of the paper follows. In Section 2, we introduce the standard mHMM. Sections 3–4 entail the EM algorithm for parameter estimation and the quadrature schemes. Section 5 describes the simulation study and the corresponding results. The last section contains concluding remarks and outlines future research agenda.

2. Mixed hidden Markov models

As stressed before, these models combine features of hidden Markov and mixed parameter models. In hidden Markov models (see e.g. Zucchini and MacDonald, 2009), the distribution of the observed response is defined conditional on the current hidden state, which represents the realization of a latent process evolving over time according to a Markov structure. In mixed parameter models, see Laird and Ware (1982), the response distribution is specified conditional on individual-specific random parameters that capture latent, time-constant, characteristics. Both models account for marginal dependence between measurements from the same unit.

Before describing mHMMs, some basic notations need to be introduced. Let Y_{it} denote the longitudinal response recorded for unit $i = 1, \ldots, n$ at occasion $t = 1, \ldots, T_i$ and let us consider a homogeneous hidden Markov chain $\{S_{it}\}$ taking values in the finite set $\mathscr{S} = \{1, \ldots, m\}$. In the following, we will refer to measurement occasions that are equally spaced and taken at pre-specified times; for this reason, we will use the generic term *time*. We assume that all individuals share the same initial probability vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$ and the same transition probability matrix $\mathbf{Q} = \{q_{kh}\}$ which is constant over the time. Terms δ_h represent the probability of starting in the hth state, while q_{kh} represents the probability of moving from the kth state at time $t-1$ to the hth one at time t, where $h, k = 1, \ldots, m$, $t = 1, \ldots, T_i$. Let $\mathbf{b}_i$ represent a vector of individual-specific random parameters; a typical choice is to consider Gaussian random parameters $\mathbf{b}_i \sim \text{MVN}(\mathbf{0}, \mathbf{D})$.

mHMMs are based on the following assumptions. The time-constant random parameters $\mathbf{b}_i$ are independent of the hidden process $\{S_{it}\}$; the distribution of the observed response at a given time is defined conditional on the hidden state occupied at the same time and the individual-specific vector $\mathbf{b}_i$. Conditional on s_{it} and $\mathbf{b}_i$, observations from the same individual are