Journal of Multivariate Analysis 100 (2009) 2313–2323

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Initial classification of joint data in EM estimation of latent class joint model

Jun Han

Mathematics and Statistics Department, Georgia State University, Atlanta, GA 30303, United States

ARTICLE INFO

Article history:
Received 21 May 2008
Available online 19 July 2009

AMS 2000 subject classifications:
62H30

Keywords:
Classification
Latent class joint model
EM algorithm
Cluster analysis
Joint distance
Recurrent events
Heterogeneous random effect model

ABSTRACT

The latent class mixture-of-experts joint model is one of the important methods for jointly modelling longitudinal and recurrent events data when the underlying population is heterogeneous and there are nonnormally distributed outcomes. The maximum likelihood estimates of parameters in latent class joint model are generally obtained by the EM algorithm. The joint distances between subjects and initial classification of subjects under study are essential to finding good starting values of the EM algorithm through formulas. In this article, separate distances and joint distances of longitudinal markers and recurrent events are proposed for classification purposes, and performance of the initial classifications based on the proposed distances and random classification are compared in a simulation study and demonstrated in an example.

1. Introduction

In chronic disease clinical trials such as cancer and AIDS, it has become necessary and attractive to jointly model longitudinal and survival data. One of the important methods that jointly model longitudinal and survival data is the latent class joint model. Muthén and Shedden [1] and Lin et al. [2] proposed a fully parametric latent class mixture model to describe different patterns of a longitudinal marker and a binary event outcome in a setting of complete follow-up. Lin et al. [3] adopted a semiparametric latent class mixture joint model of marker trajectory and censored survival event time. Lin et al. [4] applied the latent class mixture joint model to link subject visit patterns to homeless outcomes through latent classes in a health service research study. Han et al. [5] studied a latent class joint model of a longitudinal biomarker and intervened recurrent events under a parametric setting.

Expectation and maximization steps of EM [6] estimation of parameters in the latent joint model were addressed in [3,5], and starting values for EM estimation of latent class joint model were investigated by Han et al. [7]. Initial values of parameter estimates are essential to the convergence and speed of the EM algorithm. But the initial classification of the joint data is closely related to the initial values of the EM algorithm. The initial classification not only affects the time to convergence, but also the converged parameter values and the percentage of the convergence. Just like Newton–Raphson algorithm, initial classification and starting values can guide the direction of the EM iteration procedure and determine its success, which is especially true in the setting of multivariate model and high-dimensional data.

This article is organized as follows. Section 2 describes the latent class joint model. Section 3 presents the observed and complete-data likelihood functions. Section 4 proposes separate and joint distances and cluster analysis methods. In Section 5, we conduct a simulation study to make comparison of distance-based initial classification with random initial classification, and apply the separate and joint distances and cluster analysis methods to an example. We conclude with a discussion in Section 6. The details of EM computation are given in the Appendix.

E-mail addresses: hanjun@bellsouth.net, matjxh@langate.gsu.edu.

0047-259X/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2009.07.002

2. Model specification

Latent class joint model premises on the existence of a small number K, of latent classes, such that each class represents a pattern of recurrent events that is associated with the pattern of a longitudinal marker. There are three components in the latent class joint model: a class membership submodel, a longitudinal marker submodel, and an intervened recurrent event submodel. The recurrent events and longitudinal marker are assumed to be independent given the latent class to which a subject belongs.

Assume that there are n subjects, K latent classes, and l covariates in the class membership submodel. The latent class vector $c_i = (c_{i1}, \ldots, c_{iK})$ has a multinomial distribution with c_{ik} the indicator variable for subject i in class k. The probability $P(c_{ik} = 1)$ that subject i falls into class k, is modeled through a multinomial logit model that consists of the covariate vector $v_i = (v_{i1}, \ldots, v_{il})^{\mathrm{T}}$ and associated class-specific coefficient vector η_k with $\eta_1 = 0$:

$$\pi_{ik} = P(c_{ik} = 1) = \frac{\exp(v_i^{\mathrm{T}}\eta_k)}{\sum_{j=1}^{K}\exp(v_i^{\mathrm{T}}\eta_j)}, \quad k = 1, \ldots, K. \tag{1}$$

Each latent class has its own path of longitudinal outcome. Suppose we have n_i number of marker observations for subject i, p number of common fixed effect covariates, q number of class-specific fixed covariates, and r number of subject-specific random covariates. The longitudinal marker y_i for subject i is postulated to follow a heterogeneous random effects model given by

$$y_i = X_i\beta + W_i(Mc_i) + Z_ib_i + \epsilon_i \tag{2}$$

where $(y_i)_{n_i\times 1} = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$ is the vector of marker readings for subject i, $(X_i)_{n_i\times p}$ is the matrix of fixed effect covariates, $(\beta)_{p\times 1}$ is the vector of fixed effects, $(W_i)_{n_i\times q}$ is the matrix of class-specific effect covariates, often $W_i = Z_i$, $(M)_{q\times K} = (\mu_1, \ldots, \mu_K)$ is the matrix of K class effects, with $Mc_i = \mu_k$ if $c_{ik} = 1$, $(Z_i)_{n_i\times r}$ is the matrix of random effect covariates, $(b_i)_{r\times 1} \sim N(0, D)$ is the vector of random effects, and $(\epsilon_i)_{n_i\times 1} \sim N(0, \sigma^2 I_{n_i})$ is the vector of residuals uncorrelated with b_i. Model (2) captures the average longitudinal profile within a subpopulation through latent classes while allowing the flexibility among subjects in the same class through random effects.

Each latent class has its own pattern of recurrent events and intervention mode as well. Let $N_i^\dagger(s) = \sum_{j=1}^{\infty} I(S_{ij} \le s)$ be the number of event occurrences observed over $[0, s]$, $R_i^\dagger(s) = I(s \le \tau_i)$ be the at-risk indicator at time s, $x_i(s)$ be the possibly time-dependent covariate for subject i. Denote the jth calendar time of the occurrence of event for the ith subject by S_{ij}, and the censoring or monitoring time for the ith subject by τ_i. The multiplicative intensity recurrent events model is given by

$$\delta_i(s|c_{ik} = 1, \omega_i) = \omega_i R_i^\dagger(s)\lambda_k^0(\mathcal{E}_i(s))\rho(N_i^\dagger(s-), \alpha_k)\psi(\gamma^{\mathrm{T}}x_i(s)), \tag{3}$$

where $\lambda_k^0(\cdot)$ is an unspecified class-specific baseline intensity, $\mathcal{E}_i(s)$ is the effective age of the subject i at calendar time s [8,5], $N_i^\dagger(s-)$ is the number of accumulated events just before time s, $\rho(j, \alpha)$ with $\rho(0, \alpha) = 1$ is the event accumulation effect function of known form, often taking the form of $\rho(j, \alpha) = \alpha^j$, $\psi(\cdot)$ is a nonnegative link function of known form, and ω_i is the unobservable and identifiable frailty variable which is assumed to be gamma distributed with mean 1 and variance θ. Model (3) reflects the potential subpopulation patterns of recurrent events while accounting for the dependence among recurrent events arising from the same subject.

3. Estimation

We assume that the unobserved class-specific baseline hazard governing the counting process model for the recurrent event is parametrically specified such as a Weibull distribution, and use a maximum likelihood method to estimate the parameters in the joint model. Let $\Phi = (\eta, \beta, M, D, \sigma^2, \theta, \xi, \alpha, \gamma)$ be the parameters in the joint model, $H_i = (v_i, X_i, W_i, Z_i, x_i)$ be all the covariates for subject i, and $[A|B]$ be a generic symbol for a conditional density of A given B, the log-likelihood of the observed data $\{y_i, N_i^\dagger(s), R_i^\dagger(s), H_i : s \le \tau_i\}$, or simply $\{y_i, N_i^\dagger, R_i^\dagger, H_i\}$, can be written as

$$l_o = \sum_{i=1}^{n}\log\sum_{k=1}^{K}[c_{ik} = 1|H_i][y_i|c_{ik} = 1, H_i][N_i^\dagger, R_i^\dagger|c_{ik} = 1, H_i] \tag{4}$$

where $[c_{ik} = 1|H_i]$ is given by (1), $[y_i|c_{ik} = 1, H_i]$ is a multivariate normal density with mean $X_i\beta + W_i\mu_k$ and covariance $Z_iDZ_i^{\mathrm{T}} + \sigma^2 I_{n_i}$, and $[N_i^\dagger, R_i^\dagger|c_{ik} = 1, H_i]$ is given by [9]

$$[N_i^\dagger, R_i^\dagger|c_{ik} = 1, H_i] = \frac{\prod_{t\in[0,\tau_i]}[(1 + \theta N_i^\dagger(t-))R_i^\dagger(t)a_{ik}(t)]^{dN_i^\dagger(t)}}{[1 + \theta\int_0^{\tau_i} R_i^\dagger(t)a_{ik}(t)\mathrm{d}t]^{\theta^{-1}+N_i^\dagger(\tau_i)}}. \tag{5}$$

Due to the missingness of the latent class membership c_i, the random effect b_i in the longitudinal marker submodel, and the frailty ω_i in the recurrent event, the observed data log-likelihood is hard to deal with. Instead we will work with the