Hidden Markov models for continuous multivariate data with missing responses

Fulvia Pennoni, Francesco Bartolucci, and Alessio Serafini

Abstract Hidden Markov models represent a popular tool for the analysis of longitudinal data, allowing the dynamic clustering of sample units on the basis of a set of repeated responses. In the literature on longitudinal data analysis, these models are typically used in the presence of multivariate categorical data, that is, when more categorical responses are observed at each time occasion. These formulations rely on the assumption of local independence, according to which the responses are conditionally independent given the latent states. Such assumption also simplifies the treatment of missing responses when the missing-at-random assumption is plausible. Here, we deal with the case of continuous multivariate responses in which, as in a Gaussian mixture models, it is natural to assume that the continuous responses for the same time occasion are correlated, according to a specific variance-covariance matrix, even conditionally on the latent states. Although maximum likelihood estimation of this model is straightforward in standard cases using the Expectation-Maximization algorithm, we focus on its estimation when: (*i*) suitable constraints on the variance-covariance matrix are assumed; (*ii*) there are missing responses. The constraints we refer to are commonly adopted in the literature of Gaussian finite mixture models. Regarding the assumptions on the generation of missing data we focus on the missing-at-random assumption and we also account for possible individual covariates that may directly affect the responses (in addition to the latent states). In particular, we propose an Expectation Maximization (EM) algorithm that provides exact maximum likelihood estimates and also computes standard errors for the parameter estimates. The proposed approach is illustrated by a simulation study, to evaluate the computational load, and through a real case analysis. We also show how the proposal may be useful in a context of time-series analysis with an application to financial data. An R implementation of the proposed algorithm is made available by the authors within the LMest package.

Keywords hierarchical clustering; expectation-maximization algorithm; forward-backward recursions; multivariate Gaussian distribution

Fulvia Pennoni
University of Milano-Bicocca, Italy, e-mail: fulvia.pennoni@unimib.it

Francesco Bartolucci
University of Perugia, Italy, e-mail: francesco.bartolucci@unipg.it

Alessio Serafini
University of Perugia, Italy, e-mail: alessio.serafini@unipg.it