LONGITUDINAL HIDDEN MARKOV MODELS: PROBLEMS AND METHODS

Mackenzie R. Neal [1]and Paul D. McNicholas[1]

[1] Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada (e-mail: nealm6@mcmaster.ca, paul@math.mcmaster.ca)

ABSTRACT: Methods to handle common data problems for longitudinal hidden Markov models are presented. A missing data mechanism that assumes state-dependent and variable dependent missingness is introduced. High dimensionality is controlled for with the use of an explicit dimension reduction algorithm.

KEYWORDS: mixture model, expectation-maximization, initialization, variable selection, missing data

1 Introduction

Hidden Markov models (HMMs) are dependent mixture models wherein the unobserved process is governed by a Markov process. Traditionally HMMs are used to model time series data and recently have been used to model the movement of subjects across time, i.e., longitudinal data. Due to the abundance of multivariate longitudinal data arising from clinical studies, HMMs have become increasingly useful to the health sciences. This data type, however, is commonly plagued by missing data as individuals miss visits or drop-out of studies. Classically, we account for missing data through one of two means: the inclusion of only individuals with complete data or variable mean imputation. Both of which can introduce bias into analysis results due to reduction in the information provided or by distorting the information provided. Alternative to these pre-processing missing data methods, is the use of model fitting algorithms that can be altered to handle missing data at each iteration. One such algorithm is the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). We adopt this approach and develop a modified EM for longitudinal HMMs with informative missing data. In addition to missing data, approaches for handling high dimensionality and uninformative variables must be developed for longitudinal HMMs. Many implicit and explicit dimension reduction methods exist for independent mixture models. We focus on explicit dimension reduction, and extend the vscc algorithm (Andrews & McNicholas, 2014) to longitudinal HMMs.

2 Background

2.1 Longitudinal hidden Markov models

Longitudinal hidden Markov models contain an unobservable first-order Markov chain $S_{it}, i = 1,...,n, t = 1,..,T$ and an observed process $\mathbf{Y}_{it}$ representing the response vector of individual i at time t. The simplest model of this kind can by summarized by

$$Pr(S_{i1}^{t}|\mathbf{S}_{i1}^{t-1}) = Pr(S_t|S_{t-1}), i = 1,...,n, t = 2,3,...,T \tag{1}$$

$$Pr(Y_{it}|\mathbf{Y}_{i1}^{t-1}, \mathbf{S}_{i1}^{t}) = Pr(Y_{it}|S_{it}), i = 1,...,n, t = 1,2,...,T \tag{2}$$

where S_{i1}^{t} represents the history of the unobserved parameter process for individual i, from time 1 to time t, with state space $S = 1,...,m$,and $\mathbf{Y}_{i1}^{t}$ represents the history of the state-dependent process. The HMM parameters include both the parameters from the Markov chain and the state-dependent distribution, often taken to be Gaussian. The Markov chain parameters include the transition matrix Γ where $\gamma_{itjk} = P(S_{it} = k|S_{it-1} = j)$ and the initial probabilities δ where $\delta_{ij} = P(S_{i0} = j)$. The simplest model assumes homogeneity, thus $\gamma_{itjk} = \gamma_{jk}$ and $\delta_{ij} = \delta_j$. To ease calculation of the likelihood, we introduce forwards and backwards probabilities. The forwards probabilities is defined as such $\alpha_{it}(j) = P(\mathbf{Y}^{(t)}, S_{it} = j) = \delta\mathbf{P}(\mathbf{Y}_{i1})\Gamma\mathbf{P}(\mathbf{Y}_{i2})\ldots\Gamma\mathbf{P}(\mathbf{Y}_{it})$ and the backwards probabilities are defined as $\beta_{it}(j) = P(\mathbf{Y}_{it+1}^{T}, S_{it} = j)$, thus $\beta_{it}^{\top} = \Gamma\mathbf{P}(\mathbf{Y}_{it+1})\ldots\Gamma\mathbf{P}(\mathbf{Y}_{iT})1^{\top}$. The likelihood is as follows

$$L_T = \prod_{i=1}^{n} \delta\mathbf{P}(\mathbf{Y}_{i1})\Gamma\mathbf{P}(\mathbf{Y}_{i2})\ldots\Gamma\mathbf{P}(\mathbf{Y}_{iT})1^{\top} \tag{3}$$

and can be redefined with respect to the forwards or backwards probabilities via $L_T = \prod_{i=1}^{n} \alpha_{it}\beta_{it}^{\top}$ or $L_T = \prod_{i=1}^{n} \alpha_{iT}1^{\top}$.

2.1.1 Model estimation

Various versions of the EM algorithm for HMMs exist, in this paper we use the Baum-Welch algorithm (Baum *et al.*, 1970; Welch, 2003) to obtain maximum likelihood estimates. The Baum-Welch algorithm is based on maximization of the complete-data log-likelihood, as seen below

$$l(\vartheta) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{m} u_{i0j} \log \delta_j + \sum_{t=1}^{T} \sum_{j=1}^{m} \sum_{k=1}^{m} v_{itjk} \log \gamma_{jk} + \sum_{t=0}^{T} \sum_{j=1}^{m} u_{itj} \log f(y_{it}|S_{it} = j) \right\}. \tag{4}$$

The E-step consists of calculating expectations of the missing data, $u_{itj} = P(S_{it} = j|\mathbf{Y}_{i1}^{T})$ and $v_{itjk} = P(S_{it-1} = j, S_{it} = k|\mathbf{Y}_{i1}^{T},)$. The M-step consists of obtaining the maximum likelihood estimates with respect to the expected complete-data log-likelihood. In particular, the MLE for δ_j and γ_{jk} with respect to u_{itj} and v_{itjk} are

$$\delta_j = \frac{\sum_{i=1}^{n} \hat{u}_{i0j}}{n} \tag{5}$$

and,

$$\gamma_{jk} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \hat{v}_{itjk}}{\sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{k=1}^{m} \hat{v}_{itjk}}. \tag{6}$$

Additionally, the state-dependent distribution parameters are estimated in the M-step, based on the assumed distribution.

2.2 Missing data

Missing data for model-based clustering is a well studied problem, beginning with Eirola *et al.* (2014). The data is first partitioned into the observed and unobserved parts as such $(\mathbf{Y}_i^o, \mathbf{Y}_i^m)$. By assuming the joint distribution of the missing and observed part to be Gaussian, we can obtain the conditional distribution of the missing part given the observed part as

$$(\mathbf{Y}^m|\mathbf{Y}^o) \sim \mathcal{N}(\mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}(\mathbf{Y}^o - \mu_o), \Sigma_{m|o}) \tag{7}$$

(Anderson, 2003). Based on these assumptions the conditional expectation of the missing data and the conditional covariance matrices can be determined and used in the EM algorithm to account for missingness. We extend this method to longitudinal HMMs and add in methods to handle informative missingness.

2.3 Variable selection

The `vscc` algorithm, proposed by Andrews & McNicholas (2014), selects variables based on minimization of within-cluster variance and correlation to the set of selected variables. The `vscc` algorithm tends to be much faster and perform better than step-wise variable selection methods where model fitting occurs at every inclusion/exclusion step.

3 Methodology

Similar to Sportisse *et al.* (2021), we modify the Baum-Welch algorithm to allow for state-dependent and variable-dependent missingness. We do so by adjusting the definitions of the forwards and backwards probabilities, which are then used to update the E and M steps. Additionally, E and M steps are added to implement conditional mean and covariance imputation and to estimate the missingess parameters.
The modified Baum-Welch algorithm is used within the `vscc` algorithm, to allow for simultaneous handling of missing data and uninformative variables. The mathematical results and full model estimation algorithm will be given in the full paper, as well as illustrations on real and simulated data

References

ANDERSON, T.W. 2003. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley.

ANDREWS, JEFFREY L, & MCNICHOLAS, PAUL D. 2014. Variable selection for clustering and classification. *Journal of Classification*, **31**(2), 136–153.

BAUM, LEONARD E, PETRIE, TED, SOULES, GEORGE, & WEISS, NORMAN. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, **41**(1), 164–171.

DEMPSTER, ARTHUR P, LAIRD, NAN M, & RUBIN, DONALD B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, **39**(1), 1–22.

EIROLA, EMIL, LENDASSE, AMAURY, VANDEWALLE, VINCENT, & BIERNACKI, CHRISTOPHE. 2014. Mixture of Gaussians for distance estimation with missing data. *Neurocomputing*, **131**, 32–42.

SPORTISSE, AUDE, MARBAC, MATTHIEU, BIERNACKI, CHRISTOPHE, BOYER, CLAIRE, CELEUX, GILLES, JOSSE, JULIE, & LAPORTE, FABIEN. 2021. Model-based clustering with missing not at random data. *arXiv preprint arXiv:2112.10425*.

WELCH, LLOYD R. 2003. Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, **53**(4), 10–13.