Università degli Studi di Roma
La Sapienza
Dipartimento di Statistica, Probabilità e Statistiche Applicate
Dottorato di Ricerca in Statistica Metodologica ciclo XX

HIDDEN MARKOV MODELS

for

LONGITUDINAL DATA

A thesis in

Statistics

by

Antonello Maruotti

<antonello.maruotti@uniroma1.it>

Rome, November, 2007

Contents

List of Tables

Chapter 1

Introduction

Although introduced in the late '60s, hidden Markov models (HMMs, see MacDonald and Zucchini, 1997; Cappé, Moulines and Rydén, 2005) have become increasingly popular in the last ten years, due to their rich mathematical structure and flexibility. HMMs belong to a wide class of models (Markovian models), for which the dynamics of the stochastic process are (completely or partially) governed by a Markov chain or a Markov process; the model is *hidden* in the sense that the stochastic process is only partially observable (thus we cannot observe some variables constituting the process). Markovian models are widely studied in the literature and some important results obtained by Kalbfleisch and Lawless (1985) in analyzing longitudinal data under Markov assumptions remain valid also for HMMs.

As pointed out by MacKay (2003), these models are used for two purposes. The first is to make inferences about an unobserved process based on the observed one. A second reason for using HMMs is to explain variation in the observed process based on variation in a postulated hidden process.

Overdispersion in the observed data can be addressed through an HMM assuming that observations come from one of several different conditional distributions, each associated to a different latent state.

Several applications of the basic model (see e.g. Rabiner, 1989 for a tutorial) are provided in the literature and, during the last few years, several variations on the general form of the model occurred. Since the key paper of Baum et al. (1970), where a method for maximum likelihood estimation is provided, a wide range of applications and theoretical extensions have been proposed. For example, Rabiner and Juang (1993) and Jelinek (1997) give a description of the use of HMMs in speech recognition; Young (1996) provides an overview of current large-scale speech recognition systems; Kosaka et al. (2005) introduce new methods of robust speech recognition using discrete-mixture HMMs for reducing computation costs.

Recently, HMMs have found a new application field in health sciences. Broët and Richardson (2006) analyze comparative genomic hybridization (CGH) microarray starting from a mixture model framework (McLachlan and Peel, 2000a) and extend the three-state (copy gain/copy loss/modal copy) mixture model proposed by Hodgson et al. (2001) and Wang et al. (2004) to analyze CGH microarrays, assuming spatial dependence between genomic sequences within a Bayesian framework. HMMs are widely used in computational biology: Lander and Green (1987) use HMMs for genetic linkage maps; Churchill (1989) employs HMMs to distinguish coding from non-coding regions in DNA; Krogh et al. (1994) apply HMMs to statistical modeling, database searching and multiple sequence alignment of protein families and domains and, recently, Scharpf et al. (2007) improve genotype

calls and copy number estimates. In a clustering framework, Shcliep et al. (2003) analyze gene expression data accounting for time dependence in time course data and coping with missing values, while Zeng and Garcia-Frias (2006) propose the profile-HMMs (Eddie, 1998) to take into account the dynamics of gene expression profiles, which is ignored by standard clustering methods.

Some theoretical issues are exposed in detail also in specific application contexts. Ip (2006) discusses some problems related to latent class model (LCM) and applies an HMM to a longitudinal data set of brain tumor patients, assuming uniform latent class profiles and uniform transition matrix over time. Donaghy and Marshall (2006) analyze patient survival time and dynamic clinical variables, determining characteristics of the hidden phases in a Coxian phase-type distribution. Netzer et al. (2005) consider typical transaction data to evaluate the effectiveness of both relationship marketing actions and other customer-brand encounters on the dynamics of customer relationships and the buying behaviors.

The use of hidden states makes the model general enough to handle a variety of real-world time dependent data, while the relatively simple prior dependence structure still allows for the use of efficient computational procedures. It should be stressed that the idea one has about the nature of the hidden Markov chain may be quite different from one case to another. In some cases the Markov chain does have a well-defined physical meaning, whereas in other cases it may be completely fictitious and the probabilistic structure of the hidden Markov model is used only as a tool for modelling serial dependence in the analyzed data.

One frequent extension concerns the use of the HMM framework for modelling longitudinal data, but most of the proposed models have been developed in specific application contexts without a complete investigation of the corresponding theoretical aspects. Our aim is to extend approaches developed for HMMs under longitudinal observations (see e.g. Hughes, Guttorp and Charles, 1999, Wang and Puterman, 2001, Crespi et al., 2005 and MacKay, 2007) to empirical situations where potential heterogeneity sources are present. A natural way to deal with this case is to add random effects in the link function, taking into account individual- and outcome-specific effects due to unobserved heterogeneity. For this reason, we propose a random effects hidden Markov regression model within the framework of generalized linear models (GLMs) for longitudinal count data. Starting from the basic structure, we model the dispersion of the observed outcome associating each state to a mixture of several different Poisson distributions where the canonical parameters depend on a mixed model design to deal with unobservable heterogeneity sources. Estimation is carried out through an EM algorithm without parametric assumptions upon the random coefficients distribution.

Furthermore, the use of HMM framework may solve problems related to classify three-mode three-way data (Carroll and Arabie, 1980) extending mixture models approach to i.e. longitudinal data. Several models have been proposed for clustering such data in a hierarchical context (see e.g. Basford and McLachlan, 1985; Vermunt, 2007). We introduce a multivariate HMM model in the hierarchical framework, discussing the issue of longitudinal multivariate data allowing for both time and local dependence; more generally, we would like to select a multivariate Gaussian HMM whose la-

tent states correspond to association structures that receive support from the data not always, but at least for considerable periods of time. We remark that the applicability of multivariate HMMs is quite wide: it applies to any multivariate time series whose dependency structure is thought to change considerably over time. Further important examples include, among others, environmental data, typically multivariate and never measured exhaustively, and financial times series, where the state of a national economy, e.g., is a powerful qualitative mechanism that determines changes in the correlation structure among the considered variables.

The thesis is structured as follows. In Chapter 2 we introduce the adopted notation, we define theoretical framework which HMMs are related with and provide computational methods for three fundamental problem for HMM designs: the evaluation problem, the optimal state sequence problem and the parameter estimation problem. Chapter 3 provides an overview of longitudinal data structure with a particular focus on ways for dealing with such data in a regression context through generalized linear models (GLMs) and generalized linear mixed models (GLMMs). New developments of the standard HMM for longitudinal data in a regression context are provided in Chapter 4. The so called mixed hidden Markov models (MHMMs) are introduced; in particular a semi-parametric estimation method is proposed and discussed in detail with respect to other parametric methods (see e.g. MacKay, 2007) and computational details are study in depth. Simulations and two empirical application of semi-parametric MHMM are provided in Chapter 5. Chapter 6 provides an overview on clustering three-way data and MHMM are discussed for clustering three-way time dependent data in

a hierarchical context studying in depth all the computational aspects. Simulations are provided in Chapter 7. Conclusions and further remarks are discussed in Chapter 8.

Chapter 2

Hidden Markov Models

2.1 Preliminaries and notation

- Y_t is a stochastic process corresponding to the observed response at time t, $t = 0, \ldots, T$
- S_t is a Markov chain, where t is an integer index, $t = 0, \ldots, T$
- A set of m states $\mathcal{S} = \{1, \ldots, m\}$
- $Q = [q_{jk}]$, where $q_{jk} = \Pr(S_{t+1} = k \mid S_t = j), \quad j, k = 1, \ldots, m$ and $\sum_k q_{jk} = 1$
- $\delta = (\delta_1, \delta_2, \ldots, \delta_m)$, where $\delta_j = \Pr(S_0 = j), \quad j = 1, \ldots, m$ and $\sum_j \delta_j = 1$
- $f_j(Y_t \mid \theta_j) = \Pr(Y_t \mid S_t = j, \theta_j), \quad j = 1, \ldots, m; t = 0, \ldots, T$, where θ_j denote the corresponding parameter set
- $\lambda = \{Q, \delta, \theta\}$

The key idea is that a HMM is a finite model that describes a probability distribution over an infinite number of possible sequences[1].

A hidden Markov model is a doubly stochastic process with an underlying stochastic process that is not directly observable (hidden) but can be observed only through another process that produces the sequence of observations. Loosely speaking, a HMM is a Markov chain observed in noise (Cappé, Moulines and Rydén, 2005).

Let $\{Y_t\}_{t\geq 0}$ be a stochastic process, and y_t its realization, that corresponds to the observed response at time t; indeed, the model comprises a Markov chain, which we will denote by $\{S_t\}_{t\geq 0}$ (and by s_t its realization), where t is an integer index. This Markov chain is often assumed to be discrete, homogeneous, aperiodic, irreducible on a finite state-space $\{1, \ldots, j, \ldots, m\}$ (see chapter 2.2.1 for further details). a first-order HMM is defined by a set of m states $\mathcal{S} = \{1, \ldots, m\}$ and a transition matrix Q over $\mathcal{S} \times \mathcal{S}$. The (j, k)-th element $q_{jk} = \Pr(S_{t+1} = k \mid S_t = j)$ represents the a priori probability of transition from state j at time t to state k at time $t + 1$; while the initial distribution is $\delta = (\delta_1, \delta_2, \ldots, \delta_m)$, where $\delta_j = \Pr(S_0 = j)$[2]. In other words, Moreover we introduce the (conditional) model for the observed process, Y_t, $f_j(Y_t \mid \theta) = \Pr(Y_t \mid S_t = j, \theta)$, where θ denote the corresponding parameter set. In the following we will refer to $\lambda = \{Q, \delta, \theta\}$ as the model parameters.

Now, the hypothesis characterizing HMMs is that the Markov chain is

[1]We will treat only Markov chains which have finite state spaces. The theory is more general, but the general case will only obscure the basic ideas.

[2]If we assume that $\{S_t\}$ is a homogeneous, irreducible Markov chain defined on a finite state space, it has initial stationary distribution δ, that is $\delta_j = \mathrm{P}(S_t = j)$ for any $t = 0, 1, \ldots, T$

hidden, that is, it is not observable. It is worth noticing that the states of the chain may have either a convenient interpretation suggested by the nature of the observed phenomenon, or be used only for convenience in formulating the model. What is available to the observer is another stochastic process $\{Y_t\}_{t\geq 0}$ linked to the Markov chain in that S_t governs the distribution the corresponding Y_t. The observable process must satisfy two conditions:

conditional independence condition: random variables $Y_{0:T} = (Y_0, \ldots, Y_T)$ are conditionally independent, given the states $S_{0:T} = (S_0, \ldots, S_T)$;

contemporary dependence condition: the distribution of any Y_t, given the state variables $(S_0, \ldots, S_T)$, depends only on the current state S_t[3].

Taking into account these assumptions, we will define $L(\lambda; y_{0:T})$ as the likelihood function to express the fact that the likelihood is a function of λ when the observation sequence $y_{0:T}$ is given. We can derive an expression for the likelihood in terms of multiple sums:

$$\begin{aligned} L(\lambda; y_{0:T}) &= \sum_{s_{0:T}\in\mathcal{S}^T} \Pr(Y_{0:T} = y_{0:T}, S_{0:T} = s_{0:T} \mid \lambda) \\ &= \sum_{\mathcal{S}^T} \delta_{s_0} \prod_{t=1}^{T} q_{s_{t-1}s_t} \prod_{t=0}^{T} f_{s_t}(y_t \mid \theta_{s_t}) \end{aligned} \tag{2.1}$$

where λ represent the adopted HMM model parameters.

As it stands, this expression is of little or no computational use, because it has m^T terms and cannot be evaluated except for very small T. In section

[3]The underlying Markov chain $\{S_t\}$ is sometimes called the regime, or state. Statistical inference, even on the Markov chain itself, has to be done in terms of $\{Y_t\}$ only, as $\{S_t\}$ is not observed.

2.3, we show how it may be rewritten using a by-product of the filtering recursion, suggesting an efficient computational algorithm.

2.2 Theoretical framework

2.2.1 Markov chain

As shown in section 2.1, a HMM is a statistical model where the dynamic being modeled is assumed to be determined by an underlying (latent) Markov process with unknown parameters, and the challenge is to estimate the parameters of the hidden process from the realizations of the observed process. In a regular Markov model, the state is directly visible to the observer, therefore the state transition probabilities represent the only parameters; in a hidden Markov model, the state is not directly visible, since only variables influenced by the state are visible.

In 1907, A. A. Markov began the study of an important new type of chance process. In this class of processes, the outcome of a given experiment can influence the outcome of the next experiment. This type of process is called a Markov chain. A Markov chain is a sequence of random variables $S_0, S_1, \ldots, S_t, \ldots, S_T$ fulfilling the Markov property; given the present state, future and past states are independent. More formally,

$$\Pr(S_{t+1} = s_{t+1} \mid S_t = s_t, S_{t-1} = s_{t-1}, \ldots, S_0 = s_0) = \mathrm{P}(S_{t+1} = s_{t+1} \mid S_t = s_t) \tag{2.2}$$

The possible values of S_t form a countable set $\mathcal{S}$ called the state space of the chain[4]. A discrete Markov chain is completely defined by the set of

[4]There are also continuous-time Markov processes, which have countable state space

one-step transition probabilities

$$q_{jk} = \Pr(S_{t+1} = k \mid S_t = j), \quad j, k \in \mathcal{S}$$

and the initial distribution of the states

$$\delta_j = \Pr(S_0 = j), \quad j \in \mathcal{S}$$

Markov chains are often described by a directed graph, where the edges are labeled by the probabilities of going from one state to the other states (transition probabilities). Two more obvious properties are satisfeid by a Markov chain:

- $q_{jk} = \Pr(S_{t+1} = k \mid S_t = j) \geq 0$;
- $\sum_{k=1}^{m} q_{jk} = 1$.

In the following, we will see in details the properties of a Markov chain.

Time homogeneity. A time-homogenous Markov chain[5] is a process where one has

$$\Pr(S_t = s_t \mid S_{t-1} = s_{t-1}) = \Pr(S_{t-1} = s_{t-1} \mid S_{t-2} = s_{t-2}). \tag{2.3}$$

A general, inhomogeneous Markov chain does not require this property, and so one may have

$$\Pr(S_t = s_t \mid S_{t-1} = s_{t-1}) \neq \Pr(S_{t-1} = s_{t-1} \mid S_{t-2} = s_{t-2}). \tag{2.4}$$

but have a continuous index.

[5]There exists also Markov chain that are spatially homogeneous (Karlin and Taylor, 1975)

Let us define the probability of going from state j to state k in t time steps as

$$q_{jk}^{(t)} = \Pr(S_t = k \mid S_0 = j). \tag{2.5}$$

and the single-step transition as

$$q_{jk} = \mathrm{P}(S_1 = j \mid S_0 = k). \tag{2.6}$$

The t-step transition satisfies the Chapman-Kolmogorov equation

$$q_{jk}^{(T)} = \sum_{k=1}^{m} q_{jk}^{(t)} q_{km}^{(T-t)}. \tag{2.7}$$

Hence, when the Markov chain is a homogeneous Markov chain, so that the transition matrix Q is independent of the label t, then the t-step transition probability can be computed as the t-th power of the transition matrix, say Q^t.

Accesibility. A state k is said to be accessible from state j if, given that we are in state j, there is a non-zero probability that at some time in the future, we will be in state k. That is, there exists a time t such that

$$q_{jk}^{(t)} = \Pr(S_t = k \mid S_0 = j) > 0 \tag{2.8}$$

Communicability. A state j is said to communicate with state k if it is true that both j is accessible from k and that k is accessible from j.

Irreducibility. A Markov chain is said to be irreducible if its state space is a communicating class (if every pair of states in the state space communicates with each other); in an irreducible Markov chain it is possible to get to any state from any state.

Periodiciy. A state j has period t if any return to state j must occur in some multiple of t time steps and t is the largest number with this property. If $t = 1$, then the state is said to be aperiodic; otherwise $(t > 1)$, the state is said to be periodic with period t.

It can be shown that every state in a communicating class must have the same period. An irreducible Markov chain is said to be aperiodic if its states are aperiodic.

Ergodicity. A state j is said to be ergodic if it is aperiodic and positive recurrent (i.e. the expected return time is finite). If all states in a Markov chain are ergodic, then the chain is said to be ergodic[6].

Stationarity. Finally, we briefly analyze the limit behaviour of a Markov chain. Let us define the stationary distribution δ as a (row) vector which satisfies the equation

$$\delta = \delta Q \tag{2.9}$$

in other words, the stationary distribution δ ia a normalized left eigenvector of the transition matrix associate with the eigenvalue 1. Alternatively, δ can be viewed as a fixed point of the linear (hence continuous) transformation on the unit simplex associated to the matrix Q. As any continuous transformation in the unit simplex has a fixed point, a stationary distribution always exists, but is not guaranteed to be unique, in general. In addition, $Q^(t)$ converges to a rank-one matrix in which

[6]Note that if the state space $\mathcal{S}$ is finite, irreducibility and aperiodicity are sufficient conditions for ergodicity

each row is the stationary distribution δ, that is,

$$\lim_{t\to\infty} Q^{(t)} = \mathbf{1}\delta \tag{2.10}$$

where $\mathbf{1}$ is the column vector with all entries equal to 1. This is stated by the Perron-Frobenius theorem. The existence of such limit distribution is guaranteed for ergodic Markov chain. Furthermore, it can be shown that the limit of an ergodic Markov chain is the unique stationary (initial) distribution.

HMM are often classified according to the properties of their hidden Markov chain. For instance, an HMM is called ergodic if its hidden Markov process, S_t, is ergodic. Recall that a necessary and sufficient conditions for the finite discrete Markov chain S_t to be ergodic are that it must be positive recurrent, aperiodic and irreducible (Resnick, 1992)[7].

It is often assumed that the initial state distribution of an ergodic HMM is the unique stationary distribution, such as described by equation (2.9). This assumption makes sense in practice since the state distribution of an ergodic Markov chian always converges toward the stationary distribution[8]. For non-ergodic HMMs, the solution of equation (2.9) need not be unique; however, if the hidden Markov chain is stationary, the complete process $\{(Y_t, S_t)\}$ is stationary (Couvreur, 1996) and if the hidden Markov chain is also ergodic, then the observed process is ergodic too (Leroux, 1992).

[7]If all the transitions probabilities are strictly positive, i.e. $q_{jk} > 0, \forall j, k \in \mathcal{S}$, the Markov chain is said to be fully connected; full-connectedness is a sufficient condition for ergodicity but it is not a necessary one.

[8]Note that in this case $\lambda = \{\delta, Q, \theta\}$ is rendunt since δ can be computed from Q by solving equation (2.9)

2.2.2 Finite mixtures

Finite mixtures distributions represent a mathematical-based approach to the statistical modeling of a wide range of phenomena. Because of their usefulness as extremely flexible methods of modeling densities, finite mixture models have continued to receive increasing attention over past years. Finite mixture of probability distributions can be seen as zero-th order HMMS in which the mixture component (state or class) of each observation is independent of other observations.

Thus, the HMM provides a convenient way of formulating an extension of a mixture model to allow for dependent data. In detail, let us consider the following mixture model

$$f(y_t) = \sum_{g=1}^{G} \pi_g f_g(y_t) \tag{2.11}$$

for the density of a random variable Y_t. In a finite mixture context, an unobserved vector $\mathbf{z}_t = \{z_{gt}\}$ to indicate whether y_t is viewed as belonging or not belonging to the g-th component of the mixture ($g = 1, \ldots, G$).

The component labels $z_{t1}, \ldots, z_{tG}$ are assumed to be drawn from a multinomial distribution on G categories with probabilities $\pi_1, \ldots, \pi_G$; that is:

$$Z_{t1}, \ldots, Z_{tG} \overset{i.i.d}{\sim} \mathrm{Mult_G}(1, \pi), \quad \forall t. \tag{2.12}$$

The response $y_1, \ldots, y_T$ are assumed to be conditionally independent given $\mathbf{z}_T$; that is:

$$f(y_1, \ldots, y_T \mid \mathbf{z}_1, \ldots, \mathbf{z}_T) = \prod_{t=1}^{T} f(y_t \mid \mathbf{z}_t) \tag{2.13}$$

where

$$f(y_t \mid \mathbf{z}_t) = \prod_{g=1}^{G} f_g(y_t)^{z_{tg}} \tag{2.14}$$

The HMM extension relaxes the independence hypothesis on the Y_t by assuming successive observations to be correlated. With this approach, the independence assumption (2.12) on the component indicator vector is relaxed. Usually, a stationary Markovian model is formulated for the distribution of the hidden states $Z_1, \ldots, Z_T$. The conditional distribution of Y_t is formulated as before to depend only on the value of Z_t, the component of origin (state of the Markvo process), and to be conditionally independent as in (2.13). Relaxing assumption (2.12), the marginal density of Y_t will not have its simply representation (2.11) of a mixture density as in the independence case.

2.2.3 Markov Switching Models

A variety of linear models for response processes that exhibit discontinuous changes at certain undetermined points in time have been discussed in statistical literature. Within a regression context, the literature refers to such models as switching regression models, in which parameters are allowed to move discretely between a fixed number of regimes, with the switching being controlled by an unobserved state variable. Switching regressions have a rich history in econometrics, dating back to at least Quandt (1958) and Quandt and Henderson (1958) and studied among others by Quandt (1972), Quandt and Ramsey (1978) and Kiefer (1978). Goldfeld and Quandt (1973) introduce a particularly useful version of these models, referred to as a Markov-switching model, where the latent state variable controlling regime shifts

follows a Markov-chain, and is thus serially dependent.

The regime-switching model combines several sets of model parameters into one system; the set of parameters should be applied depends on the regime the system is likely in at a given time. The simplest formulation of the switching regression model may be described as follows, let us assume we have recorded T observations on some dependent variable Y_t and on p independent variables $\mathbf{x}_t = (x_{t1}, \ldots, x_{tp})$ and define the structure of a two state regime model:

$$y_t = \beta_{01} + \beta_{11}x_{t1} + \cdots + \beta_{p1}x_{tp} + \epsilon_t = \mathbf{x}_t^T\beta_1 + \epsilon_t, \quad S_t = 1 \tag{2.15}$$

with probability π and

$$y_t = \beta_{02} + \beta_{12}x_{t1} + \cdots + \beta_{p2}x_{tp} + \epsilon_t = \mathbf{x}_t^T\beta_2 + \epsilon_t \quad S_t = 2 \tag{2.16}$$

with probability (1-π); S_t is an unobserved random variable which changes through time and represent the state variable and $\epsilon \sim N(0, \sigma^2_{\epsilon,S_t})$. In other words, the state variable describes the first regime with probability $\pi = \Pr(S_t = 1)$, while the second one is attained with probability $1-\pi = \Pr(S_t = 2)$.

A complete description of the probability law governing the observed data would then require a probabilistic model describing the change from $S_t = 1$ to $S_t = 2$. The simplest specification is that S_t is the realization of a two-state Markov chain with

$$q_{jk} = \Pr(S_t = k \mid S_{t-1} = j), \quad j, k = 1, 2.$$

Such a model appears to have been first analyzed by Baum, et. al. (1970) and Lindgren (1978). In the speech recognition literature specifications that

incorporate autoregressive elements date back to Juang and Rabiner (1985), and Rabiner (1989), who described such processes as HMMs. The Markov switching regression model is an extension of finite mixtures of regression models to time series data. Cosslett and Lee (1985) studied a regression where an unobserved dichotomous explanatory variable was presumed to follow a Markov process, and the principles they use to evaluate the likelihood function are those describe in Hamilton (1989; 1990) to study a time series subject to time-varying coefficients. The Markov-switching models are extended to the case of dependent data, specifically using an autoregressive structure. The vast literature spawned by Hamilton (1989; 1990) has typically assumed that the regime shifts are exogenous of all realizations of the regression error. Given this vast interest in applied econometric inference for Markov switching models, it is not surprising that parameter estimation for these models is well-developed by now, either using classical methods such as the EM-algorithm or using Bayesian estimation via Markov chain Monte Carlo methods (MCMC), see for instance Frühwirth-Schnatter (2001, 2006). Initial consistent estimates may be obtained using the method of moment generating functions suggested by Quandt (1972). An extended version of the model proposed by Baum et al.(1970) and Kiefer(1980) has been provided by Hamilton (1989,1990). Hamilton (1989) proposes a very tractable approach to model ARMA processes subject to Markovian changes in regime. Hamilton (1990) summarizes the results necessary to apply the EM algorithm in the present context of dependent switching regimes, thus generalizing Kiefer's model where only i.i.d switching regression are considered. A further computational improvement is discussed by Kim (2004) who

points out that the maximum likelihood estimation of a Markov-switching regression model based on the Hamilton (1989) proposal is not valid if endogenous explanatory variables are present; however, there exists an appropriate transformation of the model that allows to solve the problem of endogeneity within the class of Hamilton's (1989; 1990) Markov switching regression model.

An interesting feature of the Markov-switching models is that one can draw inferences about the state distributions at different timepoints conditional on a given sample, known as filtering, prediction and smoothing problems (Zijian, 2004; see Chapter 2.3). Since we do not observe S_t directly, we infer on its value through the observed behavior of Y_t and the inferred state distributions are usually obtained as a by-product of the filtering-smoothing process for ML estimation, for a survey see Hamilton (2005). Yang (2001) shows that closed-form partial derivatives of the likelihood function can be readily derived from closed-form likelihood function. These results provide a clear framework for ML estimation in Markov-switching models and may be used to improve the efficiency of numerical optimization techniques.

The relation between Markov-switching models and HMMs is straightforward (see for example Junag and Rabiner, 1985; Rabiner, 1989): HMMs represent a subclass of autoregressive models with Markov regime, for which the conditional distribution of Y_t does not depend on lagged Y s but only on S_t. Even so, the only theoretical result available up till now, for autoregressive processes with Markov regime, is consistency of the MLE when the regime takes values in a finite set (Krishnamurthy and Ryden, 1998; Francq and Roussignol, 1998) and asymptotic properties of the MLE when the hid-

den Markov chain takes values in a compact space (Douc et al., 2004).

2.3 Computational methods for HMM

We will focus on three fundamentals problems for HMM designs, namely: the evaluation of the probability of a sequence of observations given a specific HMM; the determination of a *best* sequence of model states; the adjustment of model parameters to best account for the observed outcomes. Formally the three problems can be written as:

Probability evaluation. Given an observation sequence $Y_{0:T} = (y_0, y_1, \ldots, y_T)$ and the model λ, how do we efficiently compute the likelihood that the T-tuple Y will be observed, given the model?

Optimal state sequence. Given an observation sequence $Y_{0:T} = (y_0, y_1, \ldots, y_T)$ and the model λ, how do we find informations about the "optimal" state sequence from the available observations?

Parameter estimation. How do we optimize the model parameters so as to best describe how a given observation come about?

The solutions of these problems are given in the next three subsections.

2.3.1 The Baum-Welch and Forward-Backward algorithms

The Baum-Welch algorithm was developed by L.E. Baum and his co-workers in a series of papers published between 1966 and 1972: Baum and Petrie

(1966), Baum and Eagon (1967), Baum et al. (1970) and Baum (1972). As can be easily observed the name of L.R. Welch does not appear in these references. As Welch (2003) himself explains, Baum and Welch had both been working independently on hidden Markov chains and had both come up with essentially the same calculations for the posterior probabilities of "local" events.

Those papers lay bare a principle which underlies the effectiveness of an iterative technique which occurs in employing the maximum likelihood method in statistical estimation for probabilistic functions of Markov chains. Analyzing Markov chains and Markov processes, it is well-known that in many cases it is not the state of sequence of the model which is observed but rather the effects of this process; that is, the states are unobservable but some functions, possibly random, of the state are observed. The Baum-Welch algorithm addresses the problem of finding the values of model parameters which maximize the likelihood of the observed data. The related parameter estimation procedure is described in Rabiner (1989), with respect to speech recognition context, MacDonald and Zucchini (2001) and Cappé, Moulines and Rydén (2005) in a general form.

The calculation of the likelihood according to its definition (2.1) involves $O(Tm^T)$ operations (product and summations), which is computationally infeasible, even for moderate size HMMs. Clearly, a more efficient procedure is needed to perform the calculation of the likelihood. The problem of computing these factors may be addressed through the Forward-Backward procedure (Baum et al., 1970; for a brief review see Welch, 2003). Let us

start considering the forward variable

$$\alpha_t(j) = \Pr(y_{0:t}, S_t = j), \tag{2.17}$$

which represents the joint probability of the partial observed sequence until time t and state j at time t (given the model λ). Now, recursive factorization of $\alpha_t(j)$ is given inductively:

Initialization. The first factor is the joint probability of the state at time 0 and the initial observation y_0

$$\alpha_0(j) = \Pr(y_0, S_0 = j) = \Pr(S_0 = y)\Pr(y_0 \mid S_0 = j) = \delta_j f_j(y_0), \quad 1 \leq j \leq m. \tag{2.18}$$

Induction. The heart of the procedure is given by the recursive term used in the induction step.

$$\begin{aligned}
&\alpha_{t+1}(k) = \Pr(y_{0:t+1}, S_{t+1} = k) = \\
&\sum_{j=1}^{m} \Pr(y_{0:t+1}, S_{t+1} = k, S_t = j) = \\
&\sum_{j=1}^{m} \Pr(y_{0:t}, S_t = j)\Pr(y_{t+1}, S_{t+1} = k \mid y_{0:t}, S_t = j) = \\
&\sum_{j=1}^{m} \Pr(y_{0:t}, S_t = j)\Pr(y_{t+1}, S_{t+1} = k \mid, S_t = j) = \\
&\sum_{j=1}^{m} \Pr(y_{0:t}, S_t = j)\Pr(S_{t+1} = k \mid S_t = j)\Pr(y_{t+1} \mid S_{t+1} = k) = \\
&\left[\sum_{j=1}^{m} \alpha_t(j) q_{jk}\right] f_k(y_{t+1}), \quad 1 \leq k \leq m; 0 \leq t \leq T-1.
\end{aligned} \tag{2.19}$$

As a by-product of the forward recursion, we obtain that the likelihood can be written as

$$L(\lambda; y_{0:T}) = \sum_{j=1}^{m} \alpha_T(j). \tag{2.20}$$

A reverse time recursion exists for the backward variable which is defined as

$$\tau_t(j) = \mathrm{P}(y_{t+1:T} \mid S_t = j), \tag{2.21}$$

i.e. the probability of the partial observation sequence from $t+1$ to the end, given state j at time t. Again we can solve for $\tau_t(j)$ inductively, as follows:

Initialization. The initialization step arbitrarily defines

$$\tau_T(j) = 1, \quad 1 \leq j \leq m. \tag{2.22}$$

Induction.

$$\tau_t(j) = \sum_{k=1}^{m} q_{jk} f_k(y_{t+1}) \tau_{t+1}(k), \quad 1 \leq k \leq m; t = T-1, T-2, \ldots, 0. \tag{2.23}$$

Now, given a fully specified model and a set of observations, we would aim at estimating the corresponding unobserved state sequence. More specifically, we are concerned with the evaluation of the conditional distributions of the state at time t, S_t, given the observations $y_{0:T}$, a task that is usually referred to as smoothing. The smoothing function is defined as

$$\phi_{t|T}(j) = \Pr(S_t = j \mid y_{0:T}), \quad 1 \leq t \leq T, \tag{2.24}$$

and

$$\phi_{t|T}(j) \propto \Pr(S_t = j, y_{0:T}) = \Pr(S_t = j, y_{0:t}) \Pr(y_{t+1:T} \mid S_t = j, y_{0:t}) = \alpha_t(j)\tau_t(j). \tag{2.25}$$

Therefore

$$\phi_{t|T}(j) = \frac{\alpha_t(j)\tau_t(j)}{\sum_{j=1}^{m} \alpha_t(j)\tau_t(j)}. \tag{2.26}$$

Similarly we can derive the filter $\phi_{t|t}$ function, where the goal is to compute the distribution of the hidden state, S_t, conditionally on the sequence $y_{0:t}$

$$\phi_{t|t}(j) = \Pr(S_t = j \mid y_{0:t}) = \frac{\Pr(S_t = j, y_{0:t})}{\Pr(y_{0:t})} = \frac{\alpha_t(j)}{\sum_{k=1}^{m} \alpha_t(k)} \tag{2.27}$$

We can solve for $\phi_{t|t}(j)$, as follows: let us define

$$\phi_{0|0}(j) = \frac{\delta_j f_j(y_0)}{\sum_{k=1}^{m} \delta_k f_k(y_0)} = \frac{\alpha_0(j)}{\sum_{k=1}^{m} \alpha_0(k)} \tag{2.28}$$

then, by recursion

$$\phi_{t+1|t+1}(k) = \frac{\sum_{j=1}^{m} \phi_{t|t}(j) q_{jk} f_k(y_{t+1})}{\sum_{k=1}^{m} \sum_{j=1}^{m} \phi_{t|t}(j) q_{jk} f_k(y_{t+1})}. \tag{2.29}$$

The numerator of (2.29) is equal to

$$\begin{aligned}
&\sum_{j=1}^{m} \Pr(S_t = j \mid y_{0:t}) \Pr(S_{t+1} = k \mid S_t = j) \Pr(y_{t+1} \mid S_{t+1} = k) = \\
&\sum_{j=1}^{m} \Pr(y_{0:t+1}, S_{t+1} = k) \Pr(S_t = j \mid y_{0:t}) = \\
&\sum_{j=1}^{m} \Pr(S_t = j \mid y_{0:t}) \Pr(S_{t+1} = k \mid S_t = j) \Pr(y_{t+1} \mid S_{t+1} = k) = \\
&\sum_{j=1}^{m} \Pr(y_{t+1}, S_{t+1} = k \mid S_t = j) \Pr(S_t = j \mid y_{0:t}) = \\
&\Pr(y_{t+1}, S_{t+1} = k \mid y_{0:t})
\end{aligned} \tag{2.30}$$

As a by-product of the filtering procedure we obtain:

$$\mathrm{P}(y_{t+1} \mid y_{0:t}) = \sum_{k=1}^{m} \sum_{j=1}^{m} \phi_{t|t}(j) q_{jk} f_k(y_{t+1}) \tag{2.31}$$

where the right hand term of (2.31) corresponds to the denominator of (2.29); moreover, we obtain the prediction, $\phi_{t+1|t}(k)$:

$$\begin{aligned} &\phi_{t+1|t}(k) = \Pr(S_{t+1} = k \mid y_{0:t}) = \Pr(S_{t+1} = k \mid y_{0:t}, S_t = j) \\ &\Pr(S_t = j \mid y_{0:t}) = \Pr(S_{t+1} = k \mid S_t = j) \Pr(S_t = j \mid y_{0:t}) = \\ &\sum_{j=1}^{m} \phi_{t|t}(j) q_{jk} \end{aligned} \tag{2.32}$$

Hence, we can rewrite the filter recursion equation 2.29 trough the prediction quantity defined in 2.32 as follows:

$$\phi_{t+1|t+1}(k) = \frac{\phi_{t+1|t}(k) f_k(y_{t+1})}{\sum_{k=1}^{m} \phi_{t+1|t}(k) f_k(y_{t+1})} \tag{2.33}$$

It is also possible to define the $t_1(>0)$-th prediction as[9]:

$$\begin{aligned} &\phi_{t+t_1|t}(k) = \Pr(S_{t+t_1} = k \mid y_{0:t}) = \Pr(S_{t+t_1} = k \mid y_{0:t}, S_t = j) \\ &\Pr(S_t = j \mid y_{0:t}) = \Pr(S_{t+t_1} = k \mid S_t = j) \Pr(S_t = j \mid y_{0:t}) = \\ &\sum_{j=1}^{m} \phi_{t|t}(j) q_{jk}^{(t_1)} \end{aligned} \tag{2.34}$$

where $q_{jk}^{(t_1)}$ is t_1-th power of the transition probability matrix and

$$\mathrm{P}(y_{t+t_1} \mid y_{0:t}) = \sum_{j=1}^{m} f_j(y_{t+t_1}) \phi_{t+t_1|t}(j) \tag{2.35}$$

[9]If $t_1 \to \infty$ and S_t is ergodic, a unique stationary distribution, δ, exists and $q_{jk}^{(t_1)} \to \delta_k$. Thus, $\phi_{t+t_1|t}(k) \to \sum_j \phi_{t|t}(j)\delta_k = \delta_k$

Finally, we report some examples of the forward-backward procedure with respect to extended versions and to related applications.Lystig and Hughes (2002) estimate the variance of the parameter estimates by inverting the information matrix of the observed data extending the forward-backward algorithm to compute the information matrix directly, an approach that both simplifies and speeds-up the computation. Fearnhead (2005) shows how to perform direct simulation for discrete mixture models, where the approach is based on directly calculating the posterior distribution using a set of recursions which are similar to those of the forward-backward algorithm; Qin et al. (2000) show that the applicability of the HMM algorithms to patch-clamp recordings have been limited by several problems and propose to model the background noise by an autoregressive (AR) process, so that the data can be reduced to a higher-order Markov process with white noise. Venkataramanan and Sigworth (2002) show that the forward-backward algorithm performs well when applied to data with additive noise, but yield biased estimates of the parameters when the noise is correlated.

2.3.2 The Viterbi algorithm

The Viterbi algorithm has roots going back to the mathematical programming field of dynamic programming, which was given its name by Richard Bellman (1957). Several examples of applications that might be considered to belong to this branch are given in, for instance, Hillier and Lieberman (1995).

The original paper, whose author later was honored by the generally accepted naming of the algorithm, is Viterbi (1967). It has been originally

written within the context of decoding convolutional codes. An early review of the algorithm was given in Forney (1973), where the algorithm is formulated as a way of finding a shortest path problem. Later, the HMM tutorial proposed by Rabiner (1989) described the algorithm within the speech recognition context.

The parameter estimation procedure for the Viterbi algorithm, denoted by Viterbi training or classification EM, is described in e.g. Durbin et al. (1998), Koski (2001) and Cappè, Moulines and Rydén (2005).

We provide two more examples from the biological scientific community of the Viterbi algorithm and its related training procedure :

- in linkage analysis, the Viterbi algorithm could be used for evaluating the a posteriori most likely inheritance distribution and then take advantage of the one-to-one relationship between this distribution and the individual haplotype (Lander and Green, 1987);

- Viterbi training is used in the recent application of gene identification described in Lomsadze et al. (2005) and Yuan and Kendziorski (2006).

To describe the Viterbi algorithm and the related procedures, let us introduce some notation. We start out with calculating the posterior probabilities of the unobserved states $j = 1, \ldots, m$

$$P(S_t = j \mid y_{0:T}, \lambda) \propto \alpha_t(j)\tau_t(j), \tag{2.36}$$

where $\alpha_t(j) = P(S_t = j, y_{0:t} \mid \lambda)$, $\tau_t(j) = P(y_{t+1:T} \mid S_t = j, \lambda)$ and $\lambda = (\delta, Q, \theta)$ is the assumed HMM model[10].

[10]When not needed, we will not explicitly write out the dependence on λ

Both the forward $\alpha_t(j)$ and the backward probabilities $\tau_t(j)$ are calculated recursively (see 2.17 and 2.21) using starting values derived from their definitions taking advantage of, in the forward case, the model λ. For practical numerical reasons, one may need to normalize these terms in order to implement the algorithm. This is preferably done by incorporating scaling procedures into the algorithm in itself. Specifically, we remind that in the forward case one may define

$$\phi_{t|t}(j) = P(S_t = j \mid y_{0:t}) = \frac{\alpha_t(j)}{\sum_{k=1}^{m} \alpha_t(k)}, \tag{2.37}$$

In connection with the corresponding neighbor $\phi_{t|t-1}(j) = P(S_t = j \mid y_{0:t-1})$ it possible to automatically normalize the calculations through recursion. This is done by successively calculating and using a sequence of normalizing constants $(c_0, c_1, \ldots, c_T)$. In the backward case one may actually use the same sequence of normalizers $(c_0, c_1, \ldots, c_T)$ which was previously calculated when scanning through the forward procedure.

The basic problem tackled by the Viterbi algorithm is the following: given an output sequence $Y_{0:T} = y_{0:T}$ and a HMM model λ, how do we choose a state sequence $s_{0:T}$ in a optimal way, i.e. in the sense of best explaining the observed sequence? Finding a solution is equivalent to uncover the hidden part of the model, but an obvious question is how to select a valid and meaningful optimality criterium, i.e. how to find the "correct" state sequence.

In the following, we will give two common examples of criteria which both are based on the concept of *maximum a posteriori*.

The first procedure is based on the *maximum a posteriori individual observations* (i.e. maximization of posterior state probability), which is defined

as

$$\hat{s}_t = \arg \max_{1 \leq j \leq m} [\alpha_t(S_t = j)\tau_t(S_t = j)], \quad 0 \leq t \leq T \tag{2.38}$$

which gives the estimated optimal sequence $\hat{s} = (\hat{s}_0, \hat{s}_1, \ldots, \hat{s}_T)$. The corresponding probability is given by:

$$\hat{p} \propto \max_{1 \leq j \leq m} \left[\alpha_T(\hat{S}_t = j)\tau_T(\hat{S}_t = j)\right] \tag{2.39}$$

The second procedure is based on the *maximum a posteriori sequence of observations* (maximization of the probability of the whole sequence of states) and is defined as

$$\hat{s}_{0:T} = \arg \max_{s_{0:T} \in \mathcal{S}^{T+1}} \Pr(S_{0:T} = s_{0:T} \mid y_{0:T}) \tag{2.40}$$

where $\mathcal{S}^{T+1}$ is a set of m^{T+1} valid distinct state sequences and, by definition, the whole sequence is estimated in one step. The corresponding maximum sequence probability is now given by:

$$\hat{p} = \max_{\hat{s}_{0:T} \in \mathcal{S}^{T+1}} \Pr(S_{0:T} = \hat{s}_{0:T} \mid y_{0:T}). \tag{2.41}$$

This approach is generally called the Viterbi algorithm.

Now let the (unknown) sequence of *true* states be $s^* = (s_1^*, s_2^*, \ldots, s_T^*)$. The first approach maximizes the expected number of correctly estimated states, i.e.

$$\max_O E[\sum_{t=0}^{T} I(\{\hat{s}_t = s_t^*\}) \mid y_{0:T}, O] \tag{2.42}$$

where O is an optimality algorithm. Equivalently it may be formulated as minimizing the expected Hamming distance (Hamming, 1950) between estimated and true sequences, i.e.

$$\min_O E[d(\hat{s}, s^*) \mid y_{0:T}, O] \tag{2.43}$$

where $d(u, v)$ is the corresponding distance measure defined as the number of positions where states of the sequences u and v differ.

The second approach is more robust and intuitively attractive since it considers possible dependencies with respect to the state sequence, i.e. the Markov transition matrix Q is not neglected. In the first procedure, the estimated optimal sequence may even be inconsistent, i.e. if $q_{\hat{s}_t,\hat{s}_{t+1}} = 0$ for some t. Intermediate steps between the two approaches may be to look at *maximum a posteriori* in successive state pairs, triplets or, generally, k-tuples. This can be done sequentially, either independently or conditionally, i.e. in the pair case respectively: (i) first $(\hat{s}_0, \hat{s}_1)$ then $\hat{s}_2, \hat{s}_3$ and so on; (ii) first $(\hat{s}_0, \hat{s}_1)$ then $(\hat{s}_1, \hat{s}_2 \mid \hat{s}_1)$ and so on [11].

To formulate the algorithm we need to introduce

$$\rho_t(j) = \max_{s_{0:t-1}} \Pr(S_{0:t-1} = s_{0:t-1}, y_{0:t} \mid S_t = j). \tag{2.44}$$

We maximize the probability only with respect to the sequence of states up to and including time $t-1$, since we condition on the t-th state and

$$\Pr(S_{0:t} = s_{0:t} \mid y_{0:t}) \propto \mathrm{P}(S_{0:t} = s_{0:t}, y_{0:t}), \tag{2.45}$$

which explains why we can work with the latter quantity.

Using the introduced probability $\rho_t(j)$, we may formulate the induction step, which is the core of the algorithm,

$$\rho_{t+1}(k) = \left[\max_j \rho_t(j) q_{jk}\right] f_k(y_{t+1}), \tag{2.46}$$

where the density $f_k(\cdot)$ is not involved in the maximization.

[11] The T-tuple equals the Viterbi algorithm

To derive the optimal state sequence, we need to define a corresponding quantity $\psi_t(k)$, which keeps track of the partially optimal state sequence found during the scan. Therefore, the algorithm in its general form consists of the following steps:

Step 1, Initialization. We use the initial distribution and the transition matrix to perform initialization,

$$\rho_0(j) = \rho_j f_j(y_0), \quad 1 \leq j \leq m, \tag{2.47}$$

$$\psi_0(j) = 0, \quad 1 \leq j \leq m. \tag{2.48}$$

Vaguely speaking, the state sequence array is set equal to 0 throughout according to the property of one step delay.

Step 2, Recursion. We use the induction step to define the algorithmic recursion,

$$\rho_{t+1}(k) = \left[\max_{1\leq j\leq m} \rho_t(j) q_{jk}\right] f_k(y_{t+1}), \quad 1 \leq k \leq m \quad 0 \leq t \leq T-1, \tag{2.49}$$

$$\psi_{t+1}(k) = \arg\max_{1\leq j\leq m} \rho_t(j) q_{jk}, \quad 1 \leq k \leq m \quad 0 \leq t \leq T-1. \tag{2.50}$$

Step 3, Termination. Eventually, after T steps we have found both the final (globally) maximum probability and the exit state,

$$\hat{p} = \max_{1\leq j\leq m} \rho_T(j), \tag{2.51}$$

$$\hat{s}_T = \arg\max_{1\leq j\leq m} \rho_T(j). \tag{2.52}$$

To achieve the goal of calculating the corresponding conditional probability $\Pr(S_{0:T} = s_{0:T} \mid y_{0:T})$ one only needs to normalize the unconditional probability with respect to $\Pr(y_{0:T}) = \sum_{j=1}^{m} \alpha_T(j)$

Step 4, Backtracking. The final step of the algorithm consists of backtracking, i.e. recursively unravelling the optimal state sequence, using the derived matrix of partially state sequence $\psi = \{\psi_t(j)\}_{m*(T+1)}$ as

$$\hat{s}_{t-1} = \psi_t(\hat{s}_t), \quad t = T, T-1, \ldots, 1. \tag{2.53}$$

An alternative formulation of the algorithm would be to use logarithms of the chosen quantinties, $\Pr(S_{0:T} = s_{0:T} \mid y_{0:T})$ or $\Pr(S_{0:T} = s_{0:T}, y_{0:T})$, thus transforming the products into sums. In the conditional (scaled) case there will be a single term in the t-th step of the algorithm,

$$(\ell_{t-1} - \ell_t)$$

in the expression corresponding to the t-th scaling factor c_t. Here $\ell_t = \log \Pr(y_{0:t})$ denotes the log-likelihood of observations $y_{0:t}$.

If somewhere along the line of stepwise calculations we end up with a nonunique maximum probability $\max_{1 \le j \le m} \delta_t(j) q_{jk}$ one may for instance:

- arbitrarily choose one of the corresponding states and continue the procedure;
- create a separate matrix ψ for each of the equally probable cases and proceed simultaneously for all these cases.

We assumed a known (constant) model λ. When this is not an appropriate assumption one may sequentially, step-by-step, both estimating the model with

$$\hat{\lambda} = (\hat{\delta}, \hat{Q}, \hat{\theta})$$

and calculating the HMM-probabilities. The standard procedure is based on the EM algorithm (Baum-Welch algorithm, for details see section 2.3.1). As an approximation to the EM algorithm one may use Viterbi training.

2.3.3 The EM algorithm

As we have seen before, the log-likelihood can be evaluated recursively, even for very long observed sequences; hence it is feasible to perform parameter estimation for HMMs by direct numerical maximization of the log-likelihood fuction. The maximization can be accomplished by solving m separate maximization problems defined by starting from a fixed initial state (Leroux and Puterman, 1992). An EM algorithm to find model parameter estimates can be used (e.g. Leroux and Puterman, 1992; Hughes, 1997; Bilmes, 1998). In the EM framework, $Y_{0:T}$ is referred to as the incomplete data, $S_{0:T}$ is called the "missing" data, while $(Y_{0:T}, S_{0:T})$ is the complete-data. Given a particular sequence of states, the complete-data log-likelihood can be easily computed as

$$\begin{aligned} \ell^c(\lambda) &= \log L^c(\lambda) = \log \Pr(y_{0:T}, s_{0:T} \mid \lambda) \\ &= \sum_{jj \in \mathcal{S}} \log \delta_j + \sum_{j \in \mathcal{S}^T} \sum_{k \in \mathcal{S}^T} \log q_{jk} + \sum_{t=0}^{T} \sum_{j=1}^{m} f_j(y_t \mid \lambda). \end{aligned} \tag{2.54}$$

Let us define the $\mathcal{Q}$ function as

$$\mathcal{Q}(\lambda, \lambda') = E_\lambda[\ell^c(\lambda') \mid y_{0:T}] = \sum_{s \in \mathcal{S}} \Pr(y_{0:T}, s_{0:T} \mid \lambda) \log \Pr(y_{0:T}, s_{0:T} \mid \lambda') \tag{2.55}$$

where λ are our initial (or guessed) parameter estimates and $\mathcal{S}$ is the space of all state sequences. By the Jensen's inequality and exploiting the concavity

of the log function, Baum et al. (1970) find that replacing the parameter values by the expected frequencies of states and of state transitions given the current observations increases the likelihood function. They apply the Kullback-Leibler divergence, denoted by $D(\lambda, \lambda')$, to a general HMM (Baum and Eagon, 1967)

$$\begin{aligned} 0 \leq D(\lambda, \lambda') &= \sum_{\mathcal{S}} \frac{\Pr(y_{0:T}, s_{0:T} \mid \lambda)}{\Pr(y_{0:T} \mid \lambda)} \log \left(\frac{\Pr(y_{0:T}, s_{0:T} \mid \lambda) \Pr(y_{0:T} \mid \lambda')}{\Pr(y_{0:T}, s_{0:T} \mid \lambda') \Pr(y_{0:T} \mid \lambda)} \right) \\ &= \log \frac{\Pr(y_{0:T} \mid \lambda')}{\Pr(y_{0:T} \mid \lambda)} + \sum_{\mathcal{S}} \frac{\Pr(y_{0:T}, s_{0:T} \mid \lambda)}{\Pr(y_{0:T} \mid \lambda)} \\ &\quad \log \left(\frac{\Pr(y_{0:T}, s_{0:T} \mid \lambda)}{\Pr(y_{0:T}, s_{0:T} \mid \lambda')} \right). \end{aligned} \tag{2.56}$$

Replacing (2.56) in (2.55) we obtain

$$0 \leq D(\lambda, \lambda') = \log \left(\frac{\Pr(y_{0:T} \mid \lambda')}{\Pr(y_{0:T} \mid \lambda)} \right) + \frac{\mathcal{Q}(\lambda, \lambda) - \mathcal{Q}(\lambda, \lambda')}{\Pr(y_{0:T} \mid \lambda)} \tag{2.57}$$

and rearranging the inequality we have

$$\frac{\mathcal{Q}(\lambda, \lambda') - \mathcal{Q}(\lambda, \lambda)}{\Pr(y_{0:T} \mid \lambda)} \leq \log \left(\frac{\Pr(y_{0:T} \mid \lambda')}{\Pr(y_{0:T} \mid \lambda)} \right). \tag{2.58}$$

In general, the EM algorithm involves the following two iterative steps:

E-step. Compute $\mathcal{Q}(\lambda, \lambda') = E_{\lambda'}[\ell^c(\lambda \mid y_{0:T})]$.

M-step. Maximize $\mathcal{Q}(\lambda, \lambda')$ as a function of λ'.

Before deriving the conditional expectation of the complete loglikelihood, we define with

$$\gamma_{jt} = \Pr(S_t = j \mid y_{0:T}), \tag{2.59}$$

the posterior probability, given the observed data, of being in state j at time t and with

$$\xi_{jkt} = Pr(S_{t+1} = k, S_t = j \mid y_{0:T}) \tag{2.60}$$

the posterior probability that the unobserved sequence visited state j at time t and made a transition to state k at time $t+1$, given the observed individual sequence.

Let us examine the function $\mathcal{Q}(\lambda, \lambda')$ in more details. Taking the logarithms, we may rewrite

$$\mathcal{Q}(\lambda, \lambda') = \sum_{j \in \mathcal{S}} \gamma_{j0} \log \delta_j + \sum_{t=1}^{T} \sum_{j \in \mathcal{S}^T} \sum_{k \in \mathcal{S}^T} \xi_{jkt} \log q_{jk} + \sum_{t=0}^{T} \sum_{j \in \mathcal{S}^T} \gamma_{jt} \log f_j(y_t) \tag{2.61}$$

It can be seen that it is easy to differentiate with respect to model parameters, add the Langrange mulptilers and solve.

We can compute (2.61) using the forward and the backward variables defined in (2.17) and (2.21) considering that the first and the third parts of the (2.61) can be seen as smoothing probabilities, while the second one is a bivariate smoothing probability. Hence

$$\gamma_{jt} = \frac{\alpha_t(j)\tau_t(j)}{\sum_{j=1}^{m} \alpha_T(j)} \tag{2.62}$$

$$\xi_{jkt} = \frac{\alpha_t(j) q_{jk} f_k(y_{t+1}) \tau_t(k)}{\sum_{j=1}^{m} \alpha_T(j)} \tag{2.63}$$

In the M-step, we update all model parameter estimates starting with the transition probabilities q_{jk}. For each row j, we maximise $\sum_{k=1}^{m} \xi_{jkt} \log q_{jk}$ like in a multinomial distribution context; hence the update of q_{jk} is given by:

$$\hat{q}_{jk} = \frac{\sum_{t=0}^{T-1} \xi_{jkt}}{\sum_{t=0}^{T-1} \sum_{k' \in \mathcal{S}} \xi_{jk'}}, \quad j \in \mathcal{S}, \quad k \in \mathcal{S} \tag{2.64}$$

while the update of the initial probability corresponds to the smoothing probability:

$$\hat{\delta}_j = \Pr(S_0 = j \mid y_{0:T}) = \frac{\sum_{j \in \mathcal{S}} \gamma_{j0}}{T}. \tag{2.65}$$

Estimates of model parameter in the $f(\cdot)$ function may vary depending on the specific parametric assumptions upon $f(\cdot)$, for Gaussian distributions see i.e. Bilmes (1998) and Cappé et al. (2005).

Summarizing, a single EM cycle:

- runs the Forward-Backward algorithm (eventually scaled);
- computes the smoothing probabilities;
- computes the updated parameter estimates $\hat{\lambda} = (\hat{\delta}, \hat{Q}, \hat{\theta})$.

Baum and Petrie (1966) analyzed in deep the case when Y_t takes values in a finite set and provide results on consistency and asymptotic normality of the MLE. The conditions for consistency are weakened in Petrie (1969), who also discusses HMM identifiability. For general HMM, Lindgren (1978) proved consistency property of the maximum likelihood estimators $\hat{\theta}$, but no results on the estimation of the transition probabilities were given. Leroux (1992) proved consistency of the MLE for general HMM under mild conditions, applying the strategy developed by Wald (1949) and further developed by Kiefer and Wolfowitz (1956), obtaining convergence in the quotient topology. Rydén (1994), starting from the local asymptotic normality of the MLE in the sense of Le Cam, as proved by Bickel and Ritov (1993), proposed a new class of estimates which are almost as good as the MLE and, under fairly general conditions, are consistent and asymptotically normal. Although Bickel

and Ritov (1996) proved that an estimator similar to the MLE is asymptotically normal. Bickel et al. (1998) showed that, under weaker conditions than those in Bickel and Ritov (1996), the curvature of the likelihood is asymptotically equal to the information bound and hence the MLE is asymptotically normal.Bickel et al. (2002a) gave explicit expressions for HMM derivatives and corresponding expectations, bounding them as the size of the chain increases, obtaining second order asymptotics and some qualitative properties extending some results of Petrie (1969). Le Gland and Mevel (2000) independently developed a different technique to prove consistency and asymptotic normality of the MLE for HMMs with finite hidden state space. This work was later extended to HMMs with non-finite hidden state space by Douc and Matias (2001)[12]. This approach is based on the remark that the loglikelihood can be expressed as an additive function of an extended Markov chain. These techniques, which are well-adapted to study recursive estimators, require stronger assumption than those outlined in Bickel et al. (1998). Under suitable conditions, Bickel et al. (2002a) showed how to establish stochastic asymptotic expansions for the MLE in terms of derivatives of the likelihood debiasing the MLE; furthermore, analytic forms for the Fisher information, the Kullback-Leibler distance and entropy are provided[13]. Recently, Genon-Catalot and Laredo (2006), under rather minimal assumptions, provided a further extension, assuming the unobserved Markov chain is neither finite

[12]Asymptotic properties of the MLE in autoregressive models with Markov regime in a possibly non-stationarity process with a compact, but not necessarily finite, hidden state space are provided in Douc et al. (2001)

[13]Details of proof of lemmas and theorems in Bickel et al. (2002a) are available in Bickel et al. (2002b)

(Leroux, 1992) nor compact (Douc and Matias, 2001). However the state space is assumed to be an open interval of $\mathbb{R}$, obtaining the convergence of the normalized loglikelihood function to a limit that could be identified at the true value of the parameter.

The use of the recursive algorithm of Baum et al. (1970) results in exact evaluation of the likelihood, optimal parameter estimates and efficient computation (see Le et al., 1992); however, an alternative approximation to the E-step when using the EM algorithm for parameter estimation can be provided. The major drawback of the EM algorithm is its rate of convergence, which is linear only in the proximity of the MLE. Various modifications of the basic algorithm have been suggested; see, for example, Albert (1991), Jamshidian and Jenrich (1997), Meng and van Dyk (1997) and references therein[14]. Albert (1991) proposed an approximate method to evaluate the conditional probabilities in the E-step, conditioning only on current observations and its $2m$ nearest neighbors. This leads to a computational burden that increases in exponential order with respect to the number of nearest neighbors used; therefore, the method becomes impractical if one requires a high-order approximation (Le et al., 1992), while the complexity of the forward-backward algorithm is of linear order with respect to the sample size (Leroux and Puterman, 1992). Jashimidian and Jenrich (1997) suggested an integration of the EM algorithm through a Newton-type *accelerator* to improve the rate of convergence, but this approach usually leads to loss of

[14]Maximization with respect to λ can be also obtained by any standard numerical optimization scheme, i.e. the downhill simplex algorithm (Press et al., 1989), which does not require any derivatives of the objective function

stability and increased complexity. A further alternative approach is to use hybrid algorithms, which are based on combining the EM algorithm with a fast algorithm with strong local convergence, such as Newton-type algorithm (Bulla and Berzel, 2007): this choice leads to a hybrid algorithm that yields the stability and convergence properties of the EM algorithm along with superlinear convergence of Newton-type algorithms in the neighborhood of the maximum.

Difficulties in computing γ_{jt} and ξ_{jkt} by (2.62) and (2.63) may arise since $\alpha_t(j)$ and $\tau_t(j)$ rapidly converge to 0 as t increases, thus making the calculation and storage of long sequences impossible (see Leroux, 1992). This feature will cause underflow problems in the computation for long series data, though this may not be a serious issue for longitudinal data with short individual series. Various methods for avoiding this issue have been proposed (see e.g. Devijver, 1985). To overcome these difficulties, Leroux and Puterman (1992) determined and stored the order of magnitude on $\sum_{j\in\mathcal{S}} \alpha_t(j)$, i.e., the integer p for which $10^{-p}\sum_{j\in\mathcal{S}} \alpha_t(j)$ lies betweeen 0.1 and 1, and multiply $\alpha_t(j)$ by 10^{-p}; then $\alpha_{t+1}(k)$ are computed[15]. Wang and Puterman (2001) proposed to rescale $\alpha_t(j)$ and $\tau_t(j)$ so that the corresponding maximum value is around 1 for each t. This approach takes the structure of the model into account, represents positive $\alpha_t(j)$ and $\tau_t(j)$ in the natural exponential form, stores the largest exponents of the positive $\alpha_t(j)$ and $\tau_t(j)$ for each t respectively, and rescales these positive quantities by subtracting the corresponding largest exponent for each t.

Whatever optimization algorithm is used, there is no guarantee that it

[15]A similare procedure is applied to $\tau_t(j)$

converges towards the MLE (thus it may converge to a local maximum), since the likelihood surface of a HMM is in general multimodal. However, a reliable procedure to overcome this issue is to start optimization algorithms from several different, possibly random, points in Λ. A natural way for dealing with this issue was given by Leroux and Puterman (1992), even if this proposal has some limitations: too many null transition probabilities or independence or complete dependence in the underlying Markov chain, both of which are preserved by the EM algorithm.

If the dynamics changes slowly in time, the estimation procedure must be modified: instead of cumulating past data, we must gradually forget them. This forgetting property refers to the fact that observations far back in the past have little impact on the posterior distribution of the current state. It is sensible to assume that $\Pr(S_t = j \mid Y_{0:T} = y_{0:T})$ gets asymptotically close to $\Pr(S_t = j \mid Y_{t^*:T} = y_{t^*:T})$ as $t - t^*$ increases. In fact

$$\begin{aligned} & \Pr(S_t = j \mid Y_{0:T} = y_{0:T}) \\ = & \sum_{k=1}^{m} \Pr(S_t = j \mid Y_{0:T} = y_{0:T}, S_{t^*} = k) \Pr(S_{t^*} = k \mid Y_{0:T} = y_{0:T}) \\ = & \sum_{k=1}^{m} \Pr(S_t = j \mid Y_{t^*:T} = y_{t^*:T}, S_{t^*} = k) \Pr(S_{t^*} = k \mid Y_{0:T} = y_{0:T}) \end{aligned} \quad (2.66)$$

and

$$\begin{aligned} \Pr(S_t = j \mid Y_{t^*:T} = y_{t^*:T}) = & \sum_{k=1}^{m} \Pr(S_t = j \mid Y_{t^*:T} = y_{t^*:T}, S_{t^*} = k) \\ & \Pr(S_{t^*} = k \mid Y_{t^*:T} = y_{t^*:T}). \end{aligned} \quad (2.67)$$

Along the same path, let us consider two chains with initial distributions $\Pr(S_{t^*} = j \mid Y_{0:T} = y_{0:T})$ and $\Pr(S_{t^*} = j \mid Y_{t^*:T} = y_{t^*:T})$. Since we start

both chains at time t^*, the probability that the coupling (i.e. the two chains coincide) occurs after time t is given by

$$\Pr(\tilde{T} > t) = (1-\nu)^{t-t^*}$$

where $\tilde{T}$ is the coupling time and ν is a positive number that represent a lower bound for the transition probabilities (i.e. ν is the minorizing costant that satisfies the so-called minorization condition, see Cappé et al, 2005). Then

$$\|\Pr(S_t = j \mid Y_{0:T} = y_{0:T}) - \Pr(S_t = j \mid Y_{t^*:T} = y_{t^*:T})\|_{TV} \leq 2(1-\nu)^{t-t^*}$$

where $\|\cdot\|_{TV}$ is the total variation norm [16], which is usually adopted to measure the distance between probability measures. As can be easily observed the total variation distance converges to zero at a geometric rate as $t - t^*$ tends to infinity.

[16] For signed and bounded measure ω that the norm is defined by $\|\omega\|_{TV} = \sup_{|f|\leq 1} |\int f(\cdot) d\omega|$. If the state space $\mathcal{S}$ is finite, then $\|\omega\|_{TV} = \sum_{j\in\mathcal{S}} |\omega(j)|$. Then the total variation distance between two smoothing distribution is then given by

$$\|\Pr_{\omega}(S_t \in \cdot \mid y_{0:T}) - \Pr_{\omega'}(S_t \in \cdot \mid y_{0:T})\|_{TV} = \sum_{j\in\mathcal{S}} |\Pr_{\omega}(S_t = j\cdot \mid y_{0:T}) - \Pr_{\omega'}(S_t = j\cdot \mid y_{0:T})|$$

Chapter 3

Longitudinal Data

3.1 Data structure

Panel or longitudinal data are increasingly available and provide an opportunity to relax some of the more dogmatic features of model applied to pure cross-section and time series data.

Loosely speaking, a panel is a collection of observations which are recorded repeatedly over time[1]; panel data consist of repeated observations on the same cross section of i.e. individuals, firms, ecc., over time. In this context, since it is reasonable to allow for correlation in individual behavior over time, the random sampling assumptions (a population model has been specified and an independent identically distributed sample can be drawn from the population) appear too restrictive; however we can still use such assumptions in the cross-section dimension; furthermore, the dependence in

[1]Usually these data are referred as longitudinal data to emphasize that the data refer to the same individuals at successive times.

the time series dimension can be entirely unrestricted.

Maddala (1993) defines panel data as datasets on the same individuals over several periods of time. This encompasses longitudinal data analysis where the primary focus is on individual histories. Panel data is also used to describe the pooling of time series observations across a variety of cross-sectional units, including countries, regions, states, firms or households. Some of the benefits and limitations of using panel data sets are given in Hsiao (1986) and Baltagi (2001)[2]. In the following we will focus mainly on longitudinal data. Obvious benefits include a usually larger data set with more variability and less collinearity among the variables than it is typical of cross-sectional or time series data; such larger datasets help one get more reliable estimates and test more sophisticated behavioral models under less restrictive assumptions. Another advantage of longitudinal datasets is that one may control for individual heterogeneity: not controlling for unobserved individual-specific effects leads to biased parameter. In linear models,we may account for heterogeneity using simple tools; in fact transforming the data into deviations over time from observed individual means drops individual-specific effects while transforming data into individual deviations from time averages eliminates all time specific effects. Thus, it is possible to test for bias-inducing latent effects and construct consistent estimators that account for individual heterogeneity (Hausman and Taylor, 1981). Longitudinal datasets are also used to identify and estimate effects that are simply not detectable in pure cross-sections or pure time series data; in particular,

[2]Here we will not treat panel data in a bayesian context; for an introduction to bayesian methods applied to panel data see e.g. Koop (2003) and Lancaster (2004)

panel data sets are better able to study complex issues of dynamic behavior. Limitations of panel data sets include problems due to nonresponse and measurement errors, as well as to bias deriving from sample selection issues. In fact, respondents may refuse to participate or the interviewer may not find anybody at home; this may cause some bias in the inference drawn from this sample. While such nonresponse can also occur in cross-sectional data sets, it is more serious with longitudinal studies because subsequent waves of the panel are still subject to nonresponse.

A rich set of model and estimators for use with longitudinal data have been developed; since observations are often generated by an explicit sampling scheme, there is often interest in allowing parameters to be randomly distributed in the population. The principal distinction in the literature is between fixed or random effects models where effects or coefficients may be specific to individuals or times. Let us start from the so called unobserved effect model. In the error form it can be written as

$$y_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta} + b_i + u_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_p x_{itp}, \quad t = 1, \ldots, T \tag{3.1}$$

where $\mathbf{x}_{it}$ contain explanatory variables that may vary across time as well as across individuals, $i = 1, \ldots, n$, b_i is a time constant variable, representing unobserved effects due to individual heterogeneity and u_{it} represents the idiosyncratic errors with, by definition, $E(u_{it} \mid \mathbf{x}_{it}, b_i) = 0$. Since the overall homogeneity hypothesis is rejected by the longitudinal data structure, such model specification takes into account individual heterogeneity assuming that, conditional on the observed explanatory variables, the effects of all omitted variables are driven by an individual time invariant variable. The focus is often on whether b_i is to be treated as a random or a fixed effect;

usually b_i is called random effect when it is treated as a random variable and fixed effect when it is a parameter that has to be estimated along with $\boldsymbol{\beta}$[3].

Fixed effect models can be treated as ordinary linear regression models with intercepts specific to individuals; while adding nonlinear random effects in the model may have a number of qualitative differences (Chamberlain, 1980; 1982). Assuming a fixed effect model implies that the focus is on the individual effects on the relations among the effects; when effects are assumed to be random, the portion of variance of the response variable due to variation of random effects is often of primary interest. This leads to procedures for inferences about variance components. In both random or mixed (some effects are fixed and other are random) models inferences may be sought about the individual realized values of the random effects; in this cases, the procedures for estimating the effects may differ from those used for estimating fixed effects (for an overview of the analytical procedures for parameter estimation see i.e. Wooldridge, 2002).

In the following we will focus on models for the analysis of response data drawn from a exponential family distribution; let us define y_{it} the response of the i-th unit at time t, a density $f(y_{it})$ belong to the exponential family if it can be expressed as:

$$f(y_{1:n,0:T}) = f(y_{1:n,0:T}; \theta) = \exp\left[\frac{y_{1:n,0:T}\theta - a(\theta)}{\omega} + b(y_{1:n,0:T}, \omega)\right] \qquad (3.2)$$

where θ is the canonical parameter while ω, called the dispersion parameter, is usually treated as a nuisance parameter. Some example are described below:

[3]In the econometric literature, the key issue involving b_i is whether or not it is uncorrelated with the observed explanatory variables (Mundlak, 1987; Wooldridge, 2002)

Bernoulli distribution: if $Y \sim Bin(1, \pi)$, with $0 < \pi < 1$, then

$$f(y;\pi) = \pi^y(1-\pi)^{(1-y)} = \exp\left[y \ln \frac{\pi}{1-\pi} + \ln(1-\pi)\right] \tag{3.3}$$

therefore we have $\theta = \ln \frac{\pi}{1-\pi}$, $\omega = 1$ and $a(\theta) = \ln(1+e^{\theta})$.

Poisson distribution: if $Y \sim Poi(\lambda)$, with $\lambda > 0$, then

$$f(y;\lambda) = \frac{\lambda^y}{y!}e^{-\lambda} = \exp(y \ln \lambda - \lambda - \ln y!) \tag{3.4}$$

thus, in this case we obtain $\theta = \ln \lambda$, $\omega = 1$, $a(\theta) = e^{\theta}$ and $b(y, \omega) = -\ln y!$

Gaussian distribution: if $Y \sim N(\mu, \sigma^2)$, with $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$, then

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\} \\ &= \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]\right\} \end{aligned} \tag{3.5}$$

where $\theta = \mu$, $\omega = \sigma^2$, $a(\theta) = \frac{\theta^2}{2}$ and $b(y, \omega) = -\frac{1}{2}\left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]$

A natural way for dealing with such distributions in a regression context, including a set of explanatory variables, is through a generalized linear model (GLM).

3.2 Generalized Linear Models (GLMs)

Generalized linear models (GLMs) represent a class of fixed effects regression models for several types of dependent variables (i.e., continuous, dichotomous, counts) belonging to the exponential family. McCullagh and Nelder

(1989) discuss this class of models in great detail and refer to the term *generalized linear model* is due to Nelder and Wedderburn (1972) who described how a set of seemingly unrelated statistical techniques can be unified. Common GLMs include linear regression, logistic regression, and Poisson regression. A GLM may be specified in three steps; first, the linear predictor, denoted as ν_i, is of the form

$$\nu_i = \mathbf{x}_i^T \boldsymbol{\beta} \tag{3.6}$$

where $\mathbf{x}_i$ is the vector of regressors for unit i associated to fixed effects $\boldsymbol{\beta}$. Then, a link function $g(\cdot)$ is specified which converts the conditional expected value μ_i of the response variable Y_i (i.e., $\mu_i = E[Y_i \mid \mathbf{x}_i]$) to the linear predictor ν_i

$$g(\mu_i) = \nu_i \tag{3.7}$$

We shall model the dependence of Y on X by assuming that there exists a link function g such that, for all x in the support of X

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}. \tag{3.8}$$

The resulting family of conditional distributions of $Y \mid X$ is called a GLM with link function g[4]. Such model specification makes possible a transformation to achieve linearity in the linear predictors (Box and Cox, 1962) such that linear models carry over to GLMs.

Deriving parameter estimates through maximum likelihood, we have to take into account for some identities described in McCulloch and Searle (2001); following Peracchi (2004), we may write the loglikelihood and its

[4]The link function g is assumed to be admissible: g has a continuously differentiable inverse $h = g_{-1}$ which maps $\mathbb{R}$ onto the range space of $\mu(\cdot)$

first derivative as follows:

$$\ell(\theta) = c + \sum_{i=1}^{n} \left[Y_i \theta_i - a(\theta_i) \right], \quad \frac{\partial}{\partial \theta_i} \ell(\theta) = Y_i - a'(\theta_i) \tag{3.9}$$

where c is an arbitrary constant, θ_i is the canonical parameter for the i-th observation and $\theta = (\theta_1, \ldots, \theta_n)$. Let $h = g^{-1}$ be the inverse link function, since the canonical parameter θ_i satisfies $\mu_i = a'(\theta_i) = h(\nu_i) = h(x_i^T \beta)$, we get

$$a''(\theta_i) d\theta_i = h'(x_i^T \beta) x_i d\beta. \tag{3.10}$$

Thus

$$\frac{\partial \theta_i}{\partial \beta} = \frac{h'(x_i^T \beta)}{a''(\theta_i)} x_i. \tag{3.11}$$

We can write the likelihood equation defining a ML estimator for β as

$$\ell'(\beta) = \sum_{i=1}^{n} (y_i - h(x_i^T \hat{\beta})) x_i = 0 \tag{3.12}$$

leading to the score equations

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} h(x_i^T \hat{\beta}) x_i. \tag{3.13}$$

In this case, a ML estimator for β may be interpreted as equating the sufficient statistic $\sum_{i=1}^{n} y_i x_i$ to its conditional expectation calculated using the adopted model and the current estimate for β.

Fixed effects models, which assume that observations are independent of each other, are not appropriate for the analysis of several types of correlated data structures; in particular, for longitudinal data. For the analysis of such data, random (time and/or subject specific) effects can be added into the linear predictor to account for dependence. Usually the following assumptions are made for a balanced panel of T observations on n sample units:

- the x_{it} are strictly exogenous conditional on the individual effect b_i, $i = 1, \ldots, n$, $t = 1, \ldots, T$;

- the distribution of the outcome y_{it} is assumed to fall within the exponential family of distributions given x_{it} and b_i;

- the individual effect b_i is independent of x_{it}

There are no general rules about the way in which the individual effects enter the conditional mean and the conditional variance of y_{it}. The most common alternatives are:

Multiplicative effect: $E[y_{it} \mid x_{it}, b_i] = b_i h(x_{it}^T \beta)$

Intercept shifts: $E[y_{it} \mid x_{it}, b_i] = h(b_i + x_{it}^T \beta)$

The most common model is a mixed model including standard fixed effects for the regressors plus the random effects[5]. If we disregard the possible autocorrelation in the observations induced by the presence of the individual-specific random effects, β can simply be estimated by Non Linear Least Square (NLLS) or, in order to improve efficiency, as proposed by Liang and Zeger (1986) (for a discussion on this topic, see Wooldridge, 2002). Peracchi (2004) points out that since heteroskedasticity or failure of parametric assumptions may lead to inconsistency of conventional ML or NLLS estimators it is fundamental to build up estimators that are consistent under weaker distributional assumptions. Semiparametric estimation can be done

[5]Mixed models for continuous normal outcomes have been extensively developed since the seminal paper by Laird and Ware (1982).

through Manki's maximum score estimator (Manski 1975, 1985) or through its smoothed version (Horowitz, 1992).

In the following we build a statistical model for the longitudinal data[6] containing both ordinary regression parameters common to all individuals and individual-specific random parameters. Our main focus is on so-called mixed models, which assume that the individual-specific effects are drawn from a population distribution. Mixed models cope in a natural way with individual heterogeneity and provide common parameter estimates with adequate levels of uncertainty.

3.3 Generalized Linear Mixed Models (GLMMs)

We have seen above that GLMs represent an extension of standard linear models to non normal data whose distribution lies in the exponential effect; when dealing with longitudinal data, they can be extended easily to mixed effect models, leading to GLMMs[7]. Recently, GLMMs have become important tools for analyzing panel data. Many models from Item Response Theory (see e.g. Legler and Ryan, 1997; Rijmen et al., 2003) and multilevel models for non-normal data (Snijders and Bosker, 1999; Rabe-Haskett et al., 2004) are special cases of GLMMs. Furthermore, the inferential procedures for these models is well under way and has spawned a large number of methods and procedures, all coming with specific advantages and disadvantages.

Let us start describing a random-intercept model, which is the simplest mixed model; in this context, we augment the linear predictor with a single

[6]The statistical model proposed is valid for clustered data too.

[7]Agresti (2002) describe a variety of social science applications of GLMMs

random effect for subject i,

$$\nu_{it} = \mathbf{x}_{it}^T \beta + b_i, \quad i = 1, \ldots, n; t = 1, \ldots, T \tag{3.14}$$

where b_i is the random effect (one for each subject). These random effects represent the influence of omitted covariates or individual heterogeneity which is not captured by the observed covariates. These are treated as random effectsand they are usually assumed to be distributed as $N(0, \sigma_i^2)$, as for example in mixed logit Rasch-type models. The parameter σ_i^2 indicates the variance in the random effect distribution, and therefore the degree of heterogeneity between subjects. Including the random effects, the conditional expectation of the response variable, which is related to the linear predictor via the link function, is given by

$$\mu_{it} = E[Y_{it} \mid \mathbf{x}_{it}, b_i] \tag{3.15}$$

As a result, GLMMs are often referred to as conditional models when compared to marginal generalized estimating equations (GEE) models (Liang and Zeger, 1986), which represent an alternative extension of GLMs for correlated data based on a quasi-likelihood representation. The model can be easily extended to include multiple random effects; in fact, in many longitudinal problems, it could be common to have a random subject intercept and a random linear time-trend. In so called random coefficient model we will denote with $\mathbf{z}_{it}$ the vector of variables associated to varying effects (a column of ones is usually included to account random intercept). The vector of random effects $\mathbf{b}_i$ is assumed to follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_b$. The model can now be written as

$$\nu_{it} = \mathbf{x}_{it}^T \beta + \mathbf{z}_{it}^T \mathbf{b}_i, \quad i = 1, \ldots, n; t = 1, \ldots, T \tag{3.16}$$

The conditional mean μ_{it} is now specified as $E[Y_{it} \mid \mathbf{b}_i, \mathbf{x}_{it}]$ in terms of the vector of random effects.

In (3.16) , the random effect enters the model on the linear predictor scale; this is convenient but also natural for many applications. For instance, random effects sometimes represent heterogeneity caused by omitted explanatory variables. Thus, random effects models may be related to methods for dealing with unmeasured predictors or other missing data; for example, the random effects in the linear predictor reflect effects that would be in the fixed effects part if certain explanatory variables have been included. Random effects may also represent random measurement error in the explanatory variables or provide a mechanism for explaining overdispersion in standard models (Breslow and Clayton, 1993) where the variance function is constrained by the definition of the mean function.

Model fitting for GLMMs is rather complex: the main difficulty is that the likelihood function usually does not have a closed form; therefore, parameter estimation in GLMMs typically involves numerical approximation to likelihood function. As a general point, the solutions are usually iterative and numerically quite intensive. As pointed out by Aitkin (1999), if the distribution of the random effects is conjugate to the model distribution, then maximum likelihood (ML) is straightforward in principle from the marginal distribution of the observed data; the negative binomial and the beta-binomial distributions (Lee and Nelder, 1996) are examples of this kind. However, the conjugate approach lacks generality because a different conjugate distribution is to be assumed for each density in the exponential family. A more appealing approach would be to assume a common distribution for the random

effects regardless of the response distribution; an obvious choice is the normal $N(0, \sigma_i^2)$ distribution (Breslow and Clayton, 1993; McGilchrist, 1994). This is especially natural for link functions giving an unbounded parameter space for the linear predictor; however, exponential family models other than the normal with a normal random effect have been difficult and slow to fit by ML because the resulting likelihood does not have a closed form. A number of different approaches have been followed to deal with this problem.

The likelihood can be integrated numerically using some form of Gaussian quadrature (Anderson and Aitkin, 1985) to give full ML estimation. This approach is widely regarded as computationally intensive. Current quadrature methods use the EM algorithm (Hinde, 1982; Anderson and Hinde, 1988) for fitting the finite mixture distribution resulting from the discretization of the normal into L probability masses π_l at known locations b_l. On the other hand, the log-likelihood function can be approximated by a quadratic, and standard computational methods for the normal variance component models can then be used, giving approximate ML or REML estimation (Laird and Ware, 1982). The success of the approximation depends on the closeness to normality of the observed data likelihood and might fail badly, e.g., for binary response data (Rodriguez and Goldman, 1995).

In detail, the GLMM can be viewed as a two-stage model. At the first stage, conditional on the random effects, observations are assumed to follow a GLM; while at the second stage, the random effects b_i are drawn from a $N(0, \sigma_i^2)$ distribution. The likelihood is therefore given by:

$$L(\cdot) = \prod_{i=1}^{n} \int \prod_{t=1}^{T} f(y_{it} \mid b_i, \mathbf{x}_{it}) f(b_i) db_i \tag{3.17}$$

where $f(b_i)$ is the standard normal density or any other parametric distribution.

The integral dimension depends on the random effects structure and it may not have a closed form except for $Y \sim N(\mu, \sigma^2)$ and $b_i \sim MVN(\mathbf{0}, \mathbf{\Sigma}_b)$. A potential choice would be to approximate by Gaussian quadrature. The approximation is a finite weighted sum that evaluates the function at known locations; in the univariate normal random effects case, the approximation has the form

$$L(\cdot) = \prod_{i=1}^{n} \sum_{l=1}^{L} \pi_l \prod_{t=1}^{T} f(y_{it} \mid b_l, \mathbf{x}_{it}) \tag{3.18}$$

In other words, the likelihood is thus (approximately) the likelihood of a finite mixture of exponential family densities with known mixture proportions π_l at known locations b_l.

The adequate approximation of likelihood function for random parameter models, where the random effects are distributed according with a MVN, is computationally intensive and cannot be accomplished through standard Gaussian Quadrature whose complexity increases exponentially with integral dimension. To avoid these problems, the integrals required in the E-step of the EM algorithm can be avoided by Monte Carlo methods which are more computationally feasible than numerical integration techniques (McCulloch, 1997).

This issue can be solved in a generalized estimating equation context (Liang and Zeger, 1986), where the marginal distribution of y is not fixed to belong to the exponential family but is rather specified in term of a quasi-likelihood function with an adequate choice of the covariance function. The repeated-measures structure is represented by the covariance matrix parameters esti-

mated by marginal or partial quasi-likelihood.

Finally, a fully Bayesian approach can be followed, with the additional structure of a prior distribution on model parameters; successively Markov chain Monte Carlo methods can be used to obtain posterior distributions of model parameters. In the Bayesian context, the distinction between fixed and random effects no longer occurs, as every effect has a probability distribution. A potential disadvantage of this approach is the possible sensitivity of the conclusions to parametric assumption upon the random parameter distribution (Heckman and Singer, 1984).

This difficulty can be avoided by NPML estimation of the mixing distribution concurrently with the structural model parameters; the NPML estimate is well known to be a discrete distribution on a finite number of mass points (Kiefer and Wolfowitz, 1956; Laird, 1978; Lindsay, 1983).

Leaving unspecified the random effects distribution a NPML estimation of the mixing distribution, together with the GLM parameters, could be applied, assuming the mixing distribution is a nuisance function rather than a parameter of interest. As suggested by Aitkin (1999), NPML approach is linked with the GQ technique, but its complexity is linear in the integral dimension. In fact, we now treat the masses and locations as unknown parameters; the number L of mass points is also unknown but is treated as fixed and sequentially increased until the likelihood is maximized. Let us define with b_l the locations, the linear predictor becomes

$$\nu_{itl} = \mathbf{x}_{it}^T \beta + b_l \tag{3.19}$$

It can immediately be estimated simply by including an L-level factor in the model (Hinde and Wood, 1987). Differentiating the loglikelihood with

respect to π_l, we obtain directly a standard finite mixture ML result (Aitikin, 1999).

Chapter 4

Hidden Markov Regression Models for Longitudinal Observations

4.1 Hidden Markov Model and Longitudinal Observation

A natural extension of Markov models for univariate time series is towards models describing multiple processes, with a particular focus on longitudinal data.

Under Markov assumptions, the methodology so far proposed (in particular in continuous Markov model) regression models for longitudinal data are treated in detail by Kalbfleisch and Lawless (1985), who provide efficient ways to obtain maximum likelihood estimates (MLEs). They suppose individuals

independently move from and to m states according to a continuous time Markov process[1] and obtain parameter estimates through a quasi-Newton algorithm. Direct use of the Newton-Raphson algorithm would require the evaluation of the first and second derivatives, the use of the scoring procedure does not, since the second derivatives are replaced by their expectation estimates. This method provides also an estimate of the asymptotic covariance matrix of model parameters.

Furthermore, Kalbfleisch and Lawless (1985) propose other modeling extensions; they consider that individuals may be only partially observed, i.e. some may exit the study before its completion: if individuals who leave the study are similar to those who stay in the study with respect to all relevant respects, then ML estimation applies without change.
Moreover, it is unnecessary that all individuals be observed over the same set of time points; the amount of computation increases linearly with the number of distinct time intervals in the sample. One of the possible extensions is fitting nonhomogeneous Markov models, considering a time-dependent intensity matrix. In many empirical applications, we have covariates measured on individuals under study and the focus is on the relationship between the

[1]We remind that a continuous-time Markov process is a process characterized by a transition probability matrix specified in terms of the transitions

$$q^*_{jk}(t) = \lim_{\Delta t \to 0} q_{jk}(t, t + \Delta t)/\Delta t, \quad j \neq k \tag{4.1}$$

For convenience, we also define

$$q^*_{jk}(t) = -\sum_{k \neq j} q_{jk}(t) \tag{4.2}$$

and let $Q^*(t)$ be the transition intensity matrix with entries $q^*_{jk}(t)$.

covariates and the intensity entries in the Markov model.

Kalbfleisch and Lawless (1985) results apply to a wide class of Markov models, i.e. also to hidden Markov model (HMM); in the following, we will discuss some recent papers on HMMs for longitudinal observations, noting that in the last 10 years several variations on the form of the model have occurred.

In the discrete-finite state HMM, MacDonald and Zucchini (1997) discuss a Markov Poisson regression assuming an unobserved state Markov chain with stationary transition probabilities and a conditional Poisson distribution for observed counts. Wang and Puterman (2001) extend this model by using a two-state Markov chain with covariate-dependent transition probabilities. In particular, they assume that:

- for an observed count, at each time point, there exists an unobserved binary random variable representing the state of a (two-state) Markov chain;

- the unobserved binary random variable follows a two-state discrete Markov chain with transition probabilities described by a logit link function;

- conditional on a given state, observed counts follows a Poisson distribution with a state specific model parameters.

They propose an alternative model for handling extra-Poisson variation which is usual in a huge number of empirical applications due to various reasons, including random effects or missing information. For this purpose,

the conditional distribution on the hidden chain are overdispersed distributions, such as Negative Binomial (NB) comparing these two models using a likelihood ratio test[2]. Going in detail through some computational aspects, maximum likelihood estimation is performed using a combination of standard EM (Dempster, Laird and Rubin, 1977) and quasi-Newton algorithms (Nash, 1990). A modified EM algorithm is also used by Hughes, Guttorp and Charles (1999) to estimate parameters for an autologistic model (EM Monte Carlo Maximum Likelihood - EM-MCML). They discuss a inhomogeneous Gaussian HMM with covariates, where the current hidden state depends on the previous hidden state and the current covariates. Spatial dependence is modeled using the autologistic model; however this model need the computation of normalizing constant which is computationally intractable as the number of observations increase. Therefore, to avoid direct computation of the normalizing constant, a Monte Carlo maximum likelihood (MCML) method can be adopted (Geyer and Thompson, 1992). However, estimation using the EM-MCML method can be computationally intensive, as well as the use of a maximum pseudolikelihood (Besag, 1975) in the maximization step can lead to nonsensical states (Hughes et al., 1996).

Computational issues can be tackled also in a Bayesian framework (Scott, 2002; Crespi et al., 2005; Ridall and Pettitt, 2005), EM-type algorithms and empirical Bayes procedures could be computationally difficult; in contrast, estimation of this model can be readily achieved using a full Bayesian approach using the Gibbs sampler (Crespi et al., 2005). In a Bayesian context, Ridall

[2]Giudici et al. (2000) show that under appropriate conditions, the standard asymptotic theory of likelihood ratio tests is still valid for HMMs

and Pettitt (2005) develop an autoregressive HMM with a fixed number of (hidden) states; they assume that the probability of the current observation given previous observations and the current state can be modeled by different parametric expressions (that could assume a known parametric form) and, hence, the posterior distribution of model parameters is obtained alternately sampling from the hidden states and from the full conditional .

Interestingly, the key paper of Scott (2002) provides some background on HMMs, including two closely linked recursive procedures for evaluating the likelihood function and the posterior distribution of hidden states given observed data and model parameters. Further, he discusses methods for sampling Markov model parameters from their posterior distribution given observed data, with particular emphasis on two Gibbs sampler based procedures: the forward-backward Gibbs sampler and the direct Gibbs sampler, which samples each state in the hidden Markov chain given the most recent draws of its neighbors. Scott (2002) shows that MCMC procedures allow implementation of HMMs without using recursive computing, while the likelihood, forward-backward, and Viterbi recursions bring a richness that would not otherwise exist. Forward-backward recursions lead to a Gibbs sampler that mixes faster than its natural competitor, and the likelihood recursion opens the door to more general samplers that would be impossible without a tractable method for computing HMM likelihoods.

4.2 Parametric Mixed Hidden Markov Models

4.2.1 Model specification

The addition of individual-specific random effects is a natural extension of HMMs to account for dependence between longitudinal observations. In this context, HMMs with random effects have been proposed only recently. For instance, Humphreys (1997, 1998) suggests a HMM where the transition probabilities matrix depends on subject-specific random effects. Seltman (2002) proposes a complex biological model to describe cortisol level dynamics in a group of patients, where the baseline concentration of cortisol for each patient is modeled as a subject-specific random effect. In a recent key paper, MacKay (2007) develops a new class of models, mixed hidden Markov models (MHMMs), which unify existing HMMs for multiple processes and provide a general framework to work in this context. These models extend the class of HMMs by allowing the incorporation of fixed and random effects in the conditional and the hidden parts of the model. The advantages of MHMMs are numerous; first, simultaneous modeling multiple processes allows for the estimation of outcome-level effects, as well as for a more efficient estimation of parameters that are common to all processes. Second, these models are relatively easy to interpret and allows for a greater flexibility in modeling dependence structures by relaxing the assumption that the observations are independent given the hidden states.

Let us start describing the work of MacKay (2007) fixing notation. Y_{it} represents the observation process and S_{it} is the hidden state variable asso-

ciated to individual i, $i = 1, \ldots, n$, at time t, $t = 0, \ldots, T$, defined on a finite set of m states, where m is known. Conditional on subject-specific random effects the observed process for the i-th individual is a HMM and the hidden process is a Markov chain. The hidden process is a homogeneous Markov chain with common transition probabilities $q_{jk} = Pr(S_{it} = k \mid S_{i,t-1} = j)$ and initial probabilities $\delta_j = Pr(S_{i0} = j)$. Thus, conditional on the random effects and the hidden process, Y_{it} are independent random variables with distribution in the exponential family, i.e. Poisson:

$$Y_{it} \sim \text{Poisson}(\theta_{itj})$$

where

$$\log(\theta_{itj}) = \mathbf{x}_{it}^{\mathsf{T}}\beta_j + \mathbf{z}_{it}^{\mathsf{T}}\mathbf{b}_i. \tag{4.3}$$

where β_j and $\mathbf{b}_i$ represent, respectively, fixed and random parameters.

Assuming that random effects and hidden states are independent, the likelihood function can be written as

$$\begin{aligned} L(\cdot) &= \prod_{i=1}^{n} \int_{\mathcal{B}} \sum_{\mathcal{S}^T} \left\{ \prod_{t=0}^{T} f(y_{it} \mid s_{it}, \mathbf{x}_{it}, \mathbf{b}_i)\delta_{s_{i0}} \prod_{t=1}^{T} q_{s_{i,t-1}s_{it}} \right\} d\mathcal{H}(\mathbf{b}_i) \\ &= \prod_{i=1}^{n} \int_{\mathcal{B}} \left\{ \sum_{\mathcal{S}^T} f(y_{i0} \mid s_{i0}, \mathbf{x}_{it}, \mathbf{b}_i)\delta_{s_{i0}} \right. \\ &\quad \left. \prod_{t=1}^{T} f(y_{it} \mid s_{it}, \mathbf{x}_{it}, \mathbf{b}_i) q_{s_{i,t-1}s_{it}} \right\} d\mathcal{H}(\mathbf{b}_i) \end{aligned} \tag{4.4}$$

where $\mathcal{B}$ represents the support for the distribution of the random effects, $\mathcal{H}(\cdot)$

While the expectationmaximization (EM) algorithm has been used to estimate the parameters of a HMM, the estimation of MHMMs poses a more

challenging problem. The random effects are assumed to follow a log-Gamma distribution, and the complementary log-log link is used. Typically, expression (4.4) does not have a closed form; thus Seltman (2002) describes a Bayesian approach claiming that frequentist approach is intractable. On the other hand, MacKay (2007) improve frequentist approach estimation methods for parameters estimation. In the following section we will discuss the latter approach in greater detail.

4.2.2 Computational details: the EM algorithm and Monte Carlo methods

Now, we give the steps of the EM algorithm required to estimate the parameters of model (4.3), assuming that the initial probabilities, δ_j, are unknown parameters to be estimated. Now, thinking of both the hidden states and the random effects as missing data, the complete data log-likelihood corresponding to (4.3) is

$$\ell_c(\cdot) = \log \delta_{s_{i0}} + \sum_{t=1}^{T} \log q_{s_{t-1},s_t} + \sum_{t=0}^{T} \log f(y_{it} \mid s_{it}, \mathbf{b}_i) + \log h(b_i) \qquad (4.5)$$

Before deriving the conditional expectation of the complete loglikelihood, let us denote with

$$\gamma_{ijt} = Pr(S_{it} = j \mid y_{i,0:T}) \qquad (4.6)$$

the posterior probability, given observed data, that the i-th unit is in state j at time t and with

$$\xi_{ijkt} \quad = \quad Pr(S_{it} = k, S_{i,t-1} = j \mid y_{i,0:T}) \qquad (4.7)$$

the posterior probability that the i-th unit visited state j at time $t-1$ and made a transition to state k at time t, given the observed individual sequence.

Using the fact that individual responses are conditionally independent, the $\mathcal{Q}$ function is

$$\begin{aligned}\mathcal{Q}(\cdot) &= \sum_{i=1}^{n} E[\log \ell_c(\cdot) \mid y_{i,0:T}] = \sum_{i=1}^{n}\sum_{t=0}^{T}\sum_{j=1}^{m}\int_{\mathcal{B}} \log f(y_{it} \mid S_{it}=j, \mathbf{b}_i)\gamma_{ijt} h(\mathbf{b}_i \mid y_{i,0:T}) d\mathbf{b}_i \\ &+ \sum_{i=1}^{n}\sum_{j\in\mathcal{S}} \log \delta_j \gamma_{ij0} + \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in\mathcal{S}^T}\sum_{k\in\mathcal{S}^T} \log q_{jk}\xi_{ijkt} \\ &+ \int_{\mathcal{B}} \sum_{i=1}^{n} \log h(\mathbf{b}_i) h(\mathbf{b}_i \mid y_{i,0:T}) d\mathbf{b}_i \end{aligned} \tag{4.8}$$

Given this factorization, we may proceed maximizing each term of (4.5) separately. Adapting the forward and the backward variables defined in (2.17) - (2.21) to longitudinal case, we obtain

$$\alpha_{it}(j, \mathbf{b}_i) = \Pr(y_{i,0:t}, S_{it} = j \mid \mathbf{b}_i) \tag{4.9}$$

and

$$\tau_{it}(j, \mathbf{b}_i) = \mathrm{P}(y_{i,t+1:T} \mid S_{it} = j, \mathbf{b}_i) \tag{4.10}$$

Adopting the method of Langrenge multipliers, we obtain that the estimate of the initial probabilities occurs at

$$\hat{\delta}_j = \frac{1}{n}\sum_{i=1}^{n} \frac{\int_{\mathcal{B}} \alpha_{i0}(j, \mathbf{b}_i)\tau_{i0}(j, \mathbf{b}_i) h(\mathbf{b}_i) d\mathbf{b}_i}{\sum_{k\in\mathcal{S}} \int_{\mathcal{B}} \alpha_{i0}(k, \mathbf{b}_i)\tau_{i0}(k, \mathbf{b}_i) h(\mathbf{b}_i) d\mathbf{b}_i} \tag{4.11}$$

Similarly, the transition probability estimates are

$$\hat{q}_{jk} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\int_{\mathcal{B}} \alpha_{i,t-1}(j, \mathbf{b}_i)\tau_{it}(k, \mathbf{b}_i) f(y_{it} \mid S_{it}=k, \mathbf{b}_i) h(\mathbf{b}_i) d\mathbf{b}_i}{\sum_{i=1}^{n}\sum_{t=1}^{T}\int_{\mathcal{B}} \alpha_{i,t-1}(j, \mathbf{b}_i)\tau_{it}(k, \mathbf{b}_i) h(\mathbf{b}_i) d\mathbf{b}_i} \tag{4.12}$$

In general, the first and the last term of (4.5) can be maximized numerically using a Gaussian quadrature technique. However, the EM algorithm is known to be slow to converge, and thus, in this setting, direct maximization of the likelihood function could be faster (MacKay, 2007). For larger numbers of random effects, numerical integration methods are no longer appropriate; for such complex models, estimation can be significantly complex. Of existing estimation methods, the Monte Carlo expectation-maximization (MCEM) algorithm (McCulloch, 1997) seems to be the most feasible in this context. Drawing R samples from the random effects distribution $h(\mathbf{b}_i)$, we obtain the approximation

$$\mathcal{Q}(\cdot) \approx \frac{1}{R} \sum_{r=1}^{R} \sum_{S^T} \ell_c(\cdot) g_r(s_{it}) \tag{4.13}$$

where

$$g_r(s_{it}) = \frac{f(y_{it} \mid s_{it}, \mathbf{b}_i^{(r)}) f(s_{it} \mid \mathbf{b}_i^{(r)})}{\sum_{r=1}^{R} \sum_{S^T} f(y_{it} \mid s_{it}, \mathbf{b}_i^{(r)}) f(s_{it} \mid \mathbf{b}_i^{(r)})} \tag{4.14}$$

Through function $g_r(s_{it})$, we can obtain

$$\hat{\delta}_j = \frac{\sum_{i=1}^{n} \sum_{r=1}^{R} \sum_{j=1}^{m} g_r(j) \mathbf{1}(s_{i0} = j)}{n} \tag{4.15}$$

where $\mathbf{1}(s_{i0} = j) = 1$ if $(s_{i0} = j)$ and 0 otherwise (MacKay, 2007). Numerical maximization would ordinarily be required in order to obtain updates for the other parameters. Other estimation methods are briefly reviewed in MacKay (2007).

4.3 Semi-Parametric Mixed Hidden Markov Models

4.3.1 Model specification

We start with a formal definition of Poisson HMMs for panel observations, extending the Markov Poisson regression model proposed by MacDonald and Zucchini (1997) to a finite mixture HMM (also referred to mixed hidden Markov model, MacKay, 2007). To describe a standard HMM, let us assume y_{it}, $i = 1, \ldots, n$, $t = 0, \ldots, T$, are realizations of count random variables, that we aim to model, recorded on n individuals (in the cross-sectional dimension of the data) over a period of length T, $i = 1, \ldots, n$. Let s_{it} be the unobserved realizations of a homogeneous Markovian random hidden variable (state), S_{it}. The dependent variable is thus influenced by a sequence of unobserved realizations, s_{it} and, given the current state for individual i, s_{it}, the observed counts are realizations of i.i.d. random variables.

We assume that the sequence of hidden state variables, generating the observed sequence, can take only discrete values; therefore, we have the following structure:

$q_{jk} = Pr(S_{it} = k \mid S_{i,t-1} = j)$ is the transition probability from time $t-1$ being in state j to time t being in state k; we assume that these probabilities common to all individuals;

$\delta_j = Pr(S_{i0} = j)$ is the initial probability of the Markov chain;

$Pr(y_{it} \mid s_{it})$ represent conditionally independent Poisson distributions for the

observed sequence within a given parametric family with parameter $\theta \in \Theta$, where Θ is a subset of the n-dimensional Euclidean space.

Hence, we obtain a hidden Markov model (HMM) describing the observed time sequence. The joint probability given the model $\{\delta, Q, \lambda\}$ is

$$Pr(y_{1:n,0:T} \mid \cdot) = \prod_{i=1}^{n} \sum_{\mathcal{S}^T} \delta_{s_{i0}} \prod_{t=1}^{T} q_{s_{i,t-1}s_{it}} \prod_{t=1}^{T} \frac{e^{-\theta_{s_{it}}} \theta_{s_{it}}^{y_{it}}}{y_{it}!}, \tag{4.16}$$

where $y_{1:n,0:T}$ represents the $n \times T$-dimensional matrix of all observations and $f(y_{it} \mid s_{it}) = \frac{e^{-\theta_{s_{it}}} \theta_{s_{it}}^{y_{it}}}{y_{it}!}$ is the conditional distribution of y_{it}.

We extend the traditional HMM in a regression context. In regression analysis, the interest is usually focused upon the parameter vector $\theta = (\theta_1, \ldots, \theta_m)^\mathsf{T}$, which is modelled by defining a linear predictor as a function of a set of p covariates $\mathbf{x}_{it} = (x_{it1}, \ldots, x_{itp})$:

$$\log[E(y_{it} \mid \mathbf{x}_{it}, S_{it} = j)] = \log(\theta_{itj}) = \beta_{0j} + \beta_{1j} x_{it1} + \cdots + \beta_{pj} x_{itp}, \tag{4.17}$$

where β_j is the vector of regression parameters for all individuals being in state j.

This specification of the canonical parameters could be restrictive; in fact, count data often exhibits substantial overdispersion, in the sense that data shows greater variability than the one postulated by the Poisson model and, therefore, conditional equality of mean and variance of $\theta_{s_{it}}$ is violated. Various reasons, i.e. unobserved heterogeneity or missing covariates, could make counts overdispersed. To account for zero-inflated counts, Wang and Alba (2006) propose a negative binomial HMM regression where the distribution of the observed counts changes according to an underlying two-state Markov chain and a comparison with the zero-inflated Poisson and the zero-inflated

negative binomial regression model is shown. In this paper, we suggest to represent this extra variation by adding an unobserved random effect, b_i, to the linear predictor. Assuming log-gamma distributed random effects, the proposed model would collapse to an HMM with NB kernel (state specific) distribution. The extra variation is modeled on the same scale as for the linear predictor, which is a natural choice if overdispersion arises from unobservable heterogeneity due to omission of one or more explanatory variables. Following this approach, the linear predictor becomes

$$\log(\theta_{itj}) = \mathbf{x}_{it}^{\mathsf{T}}\beta_j + b_i, \tag{4.18}$$

where b_i represents individual-specific features varying over the dataset in an unknown way; they are usually considered as drawn from n i.i.d. variables B_i with a common, unknown, density function $h(\cdot)$. As can be seen from equation (4.18), an additional model restriction has been imposed: b_i appears additively in the model. This assumption can be easily relaxed by associating random parameters to some elements of the adopted covariate set. Let us assume that variables whose effects are fixed and variable across subjects are collected in $\mathbf{x}_{it}$ and in $\mathbf{z}_{it}$, respectively. The previous model can be easily generalized to the following random coefficient model:

$$\log(\theta_{itj}) = \mathbf{x}_{it}^{\mathsf{T}}\beta_j + \mathbf{z}_{it}^{\mathsf{T}}\mathbf{b}_i. \tag{4.19}$$

Obviously equation (4.19) leads to equation (4.18) if a random intercept is adopted, i.e. $\mathbf{z}_{it} \equiv 1$.

The observed counts are also assumed to be independent conditional on the random vector $\mathbf{b}_i$, and we assume that the random effects are independent of the hidden states. Overdispersion could not be the only result of the

adopted modeling assumptions; modeling longitudinal data we take into account the usually positive correlation among repeated measures on the same individual. The hidden states model the dynamics of the process, while the random effects models its "size".

The likelihood (4.16), given the assumption of conditional independence, can be written as (MacKay, 2007)

$$\begin{aligned} L(\cdot) &= \prod_{i=1}^{n} \int_{\mathcal{B}} \sum_{\mathcal{S}^T} \left\{ \prod_{t=1}^{T} f(y_{it} \mid s_{it}, \mathbf{x}_{it}, \mathbf{b}_i) \delta_{s_{i0}} \prod_{t=1}^{T} q_{s_{i,t-1} s_{it}} \right\} d\mathcal{H}(\mathbf{b}_i) \\ &= \prod_{i=1}^{n} \int_{\mathcal{B}} \left\{ \sum_{\mathcal{S}^T} f(y_{i0} \mid s_{i0}, \mathbf{x}_{it}, \mathbf{b}_i) \delta_{s_{i0}} \prod_{t=1}^{T} f(y_{it} \mid s_{it}, \mathbf{x}_{it}, \mathbf{b}_i) q_{s_{i,t-1} s_{it}} \right\} d\mathcal{H}(\mathbf{b}_i) \end{aligned} \tag{4.20}$$

where $\mathcal{B}$ represents the support for $\mathcal{H}(\cdot)$.

From equation (4.20) it becomes clear that the random coefficient model differs from the standard HMM in the complexity of the likelihood evaluation, where integration reveals the impact of random coefficients, is done with respect to the random coefficient distribution and may not be available in closed form. The EM algorithm is required to estimate model parameters; as noted by MacKay (2003), the convergence properties of the sequence of estimators produced by the EM algorithm in the context of HMM under longitudinal data are provided by Wu (1983).

Several alternatives have been proposed for the random terms, within the framework of generalized linear models: parametric examples are provided by Poisson-log-normal (Munkin and Trivedi, 1999) and latent Poisson-Normal (van Ophem, 2000) models. In this context, if the hidden chain is not stationary, numerical maximization is required to compute model parameters using, for example, Gaussian quadrature techniques. Furthermore, as pointed out

by MacKay (2007), the EM algorithm should be applied to random effect HMMs only if the conditional distribution of the observed process has a "nice" form (i.e. exponential). In MacKay (2007), random effects are assumed to follow a known distribution (e.g. log-Gamma) and the evaluation of the likelihood is computed directly. In the case where there are only a few random effects, a numerical integration method (e.g Gaussian Quadrature or Adaptive Gaussian Quadrature, if random effects are Gaussian random variables) could be applied; on the other hand, for large numbers of random effects, numerical integration methods are no longer appropriate and simulation methods (e.g. Monte Carlo Expectation Maximization, MCEM) seem to be more feasible.

Alfò and Trovato (2004) propose a semiparametric model with unspecified density for random effects distribution $\mathcal{H}(\cdot)$ in a finite mixture context. The choice of a flexible specification is preferred to parametric alternatives, as suggested by Heckman and Singer (1984). Moreover, parametric alternatives could often result in oversmoothing (Knorr-Held and Raßer, 2000) while the marginal maximization through numerical approximation or simulation methods can be very intensive (Crouch and Spiegelman, 1990 and Gueorguieva, 2001).

Using simple geometric results, Lindsay (1983a and 1983b) showed that ML estimation of $\mathcal{H}(\cdot)$ involves the standard problem of maximizing a concave function over a convex set. As long as the likelihood is bounded, it is maximized with respect to $h(\cdot)$ by a discrete distribution $h_L(\cdot)$ with at most $L \leq n$ support points. Let us suppose that this discrete distribution puts masses π_l on locations $\mathbf{r}_l$, $l = 1, \ldots, L$. Since the NPML estimate of a

mixing distribution is a discrete distribution on a finite number of locations, the likelihood can be expressed as:

$$L(\cdot) = \prod_{i=1}^{n} \sum_{l=1}^{L} \sum_{\mathcal{S}^T} \left\{ \delta_{s_{i0}} \prod_{t=1}^{T} q_{s_{i,t-1}s_{it}} \prod_{t=0}^{T} f(y_{it}|s_{it}, \mathbf{x}_{it}, \mathbf{B}_i = \mathbf{r}_l) \right\} \pi_l$$

where $\pi_l = P(\mathbf{B}_i = \mathbf{r}_l)$. The term $f(y_{it}|s_{it}, \mathbf{x}_{it}, \mathbf{B}_i = \mathbf{r}_l)$ denotes the response distribution in the lth component of the finite mixture, which is assumed to be Poisson with canonical parameter given by:

$$\log(\theta_{itj}) = \mathbf{x}_{it}^{\mathsf{T}} \beta_j + \mathbf{z}_{it}^{\mathsf{T}} \mathbf{r}_l. \tag{4.21}$$

Locations $\mathbf{r}_l$ and corresponding masses π_l represent unknown parameters, as well as L, which should be estimated along with other model parameters via selection model techniques (see e.g. Böhning, 2000).

4.3.2 Computational details

In the following we treat the case where a random intercept model (i.e. $\mathbf{z}_{it} \equiv 1$) is adopted. In this section we develop the modified EM algorithm, discussed by Aitkin (1996) and Alfò and Trovato (2004), for the MLE in a semiparametric framework, to mixed hidden Markov models. If we leave $\mathcal{H}(\cdot)$ unspecified, the proposed model for the observed process reduces to a finite mixture model, where the number of components L is unknown and has to be estimated along with all other model parameters; in other words, each unit can be thought of as drawn from an HMM modelled as a mixture of L components on each hidden state.

We adopt a *step by step* algorithm (Böhning, 2003) for joint estimation of model parameters assuming that: L is fixed and unknown and the random

variable B_i follows a discrete distribution with L support points r_l with associated mass points π_l where $\pi_l = \Pr(B_i = r_l)$ with $\sum_{l=1}^{L} \pi_l = 1$. Let us denote by $\eta_i = (\eta_{i1}, \ldots, \eta_{il}, \ldots, \eta_{iL})$ the unobservable vector of component indicator variables where

$$\eta_{il} = \begin{cases} 1 & \text{if } B_i = r_l \\ 0 & \text{otherwise} \end{cases}$$

As pointed out by Alfò and Trovato (2004), should these indicator variables be known, this problem would lead to a simple HMM Poisson regression model with component-specific intercept. However, component memberships are unobservable and therefore have to be treated as missing data. Using a multinomial distribution for η_i, the log-likelihood for the complete data problem can be written as:

$$\begin{aligned} \ell_c(\cdot) &= \sum_{i=1}^{n} \sum_{l=1}^{L} \sum_{\mathcal{S}^T} \eta_{il} \left\{ \log \delta_{s_{i0}} + \sum_{t=1}^{T} \log q_{s_{i,t-1} s_{it}} \right. \\ &+ \left. \sum_{t=0}^{T} \log f(y_{it} | s_{it}, \mathbf{x}_{it}, B_i = r_l) + \log \pi_l \right\}. \end{aligned} \tag{4.22}$$

As usual, within the E-step we replace η_{il} with its conditional expectation, η^*_{il}:

$$\eta^*_{il} = Pr(B_i = r_l \mid y_{i,0:T}) = \frac{\pi_l Pr(y_{i,0:T} \mid B_i = r_l)}{\sum_k \pi_k Pr(y_{i,0:T} \mid B_i = r_k)}. \tag{4.23}$$

representing the posterior probability that the i-th unit comes from the l-th component of the mixture. Before deriving the conditional expectation of the complete loglikelihood, we define with

$$\gamma_{ijt} = Pr(S_{it} = j \mid y_{i,0:T}) = \sum_{l=1}^{L} Pr(S_{it} = j \mid B_i = r_l, y_{i,0:T}) P(B_i = r_l \mid y_{i,0:T}) \tag{4.24}$$

the posterior probability, given the observed data, of being in state j at time t for an individual in the l-th component and with

$$\begin{aligned}\xi_{ijkt} &= Pr(S_{it}=k, S_{i,t-1}=j \mid \mathbf{y}_i) \\ &= \sum_{l=1}^{L} Pr(S_{it}=k, S_{i,t-1}=j \mid B_i=r_l, y_{i,0:T}) Pr(B_i=r_l \mid y_{i,0:T})\end{aligned} \tag{4.25}$$

the posterior probability that the unobserved sequence visited state j at time $t-1$ and made a transition to state k at time t, given the observed individual sequence for an individual in the l-th component.

Thus, the conditional expectation of the complete log-likelihood is given by:

$$\begin{aligned}Q(\cdot) &= \sum_{i=1}^{n}\left\{\sum_{j\in\mathcal{S}}\gamma_{ij0}\log(\delta_j) + \sum_{j\in\mathcal{S}^T}\sum_{k\in\mathcal{S}^T}\sum_{t=1}^{T}\xi_{ijkt}\log q_{jk}\right\} \\ &+ \sum_{i=1}^{n}\left\{\sum_{l=1}^{L}\eta_{il}\left[\log\pi_l + \sum_{t=0}^{T}\log f(y_{it}|s_{it},\mathbf{x}_{it},B_i=r_l)\right]\right\}\end{aligned} \tag{4.26}$$

Our goal is to update current parameter estimates by using the old parameter estimates and the data. Thus, we can show that the maximum likelihood estimates of δ_j are

$$\hat{\delta}_j = \frac{1}{n}\sum_{i=1}^{n}\gamma_{ij0}, \quad j\in\mathcal{S} \tag{4.27}$$

Similarly, we obtain ML estimates for $q_{s_{i,t-1}s_{it}}$ and π_l:

$$\hat{q}_{jk} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\xi_{ijkt}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k'\in\mathcal{S}^T}\xi_{ijk't}}, \quad j\in\mathcal{S}, \quad k\in\mathcal{S} \tag{4.28}$$

and

$$\hat{\pi}_l = \frac{\sum_{i=1}^{n} \eta_{il}}{n}. \tag{4.29}$$

The estimates of regression parameters, β, are given solving the following M-step equation:

$$\frac{\partial Q}{\partial \beta} = \sum_{i=1}^{n} \sum_{t=0}^{T} \sum_{l=1}^{L} \eta_{il}^{*} \frac{\partial}{\partial \beta} \log f(y_{it} | s_{it}, \mathbf{x}_{it}, B_i = r_l) \tag{4.30}$$

which are weighted sums of L likelihood equations for standard GLMs and, therefore, the EM algorithm for finite mixture of univariate distributions applies. The E- and M-steps are repeatedly alternated until the log-likelihood (relative) difference changes by an arbitrarily small amount. The number of components could be chosen using penalized likelihood criteria (such as AIC, CAIC or BIC, see e.g. Keribin, 2000).

For an easier implementation of the algorithm, we recall the forward and the backward procedure, which are useful for parameter estimation. We denote with

$$\alpha_{it}(j, l) = Pr(y_{i0}, \ldots, y_{it}, S_{it} = j \mid B_i = r_l), \tag{4.31}$$

the probability of seeing the partial sequence ending up in state j at time t, given the l-th component. We can efficiently compute $\alpha_{it}(j, l)$ recursively as:

$$\alpha_{i0}(j, l) = \delta_j f(y_{i0} \mid S_{i0} = j, B_i = r_l), \tag{4.32}$$

$$\alpha_{i,t+1}(k, l) = \sum_{j=1}^{m} \alpha_{it}(j, l) q_{jk} f(y_{i,t+1} \mid S_{i,t+1} = k, B_i = r_l), \tag{4.33}$$

$$P(y_{1:n,0:T}) = \prod_{i=1}^{n} \sum_{l=1}^{L} \sum_{j=1}^{m} \pi_l \alpha_{iT}(j, l). \tag{4.34}$$

Furthermore, let us define with

$$\tau_{it}(j,l) = P(y_{i,t+1},\ldots,y_{iT} \mid S_{it} = j, B_i = r_l) \tag{4.35}$$

the backward probability of the partial sequence $y_{i,t+1},\ldots,y_{i,T}$ given that we started at state j at time t, given the l-th component. The recursive procedure is given by:

$$\tau_{i,T}(j,l) = 1, \tag{4.36}$$

$$\tau_{it}(j,l) = \sum_{k=1}^{m} q_{jk} f(y_{i,t+1} \mid S_{i,t+1} = k, B_i = r_l)\tau_{i,t+1}(k,l). \tag{4.37}$$

We can express the quantities in equations (6.8) to (6.7), using the forward and the backward variables:

$$\gamma_{ijt} = \sum_{l=1}^{L} \frac{\alpha_{it}(j,l)\tau_{it}(j,l)}{\sum_{j'=1}^{m} \alpha_{it}(j',l)\tau_{it}(j',l)} \eta^*_{il} \tag{4.38}$$

$$\xi_{ijkt} = \sum_{l=1}^{L} \frac{\alpha_{i,t-1}(j,l) q_{jk} f(y_{it} \mid S_{i,t+1} = k, B_i = r_l)\tau_{it}(k,l)}{\sum_{j'=1}^{m}\sum_{k'=1}^{m} \alpha_{it-1}(j',l) q_{j'k'} f(y_{it} \mid S_{i,t+1} = k', B_i = r_l)\tau_{it}(k',l)} \eta^*_{il} \tag{4.39}$$

where

$$\eta^*_{il} = \frac{\pi_l \sum_{j=1}^{m} \alpha_{iT}(j,l)}{\sum_{l'=1}^{L} \pi_{l'} \sum_{j=1}^{m} \alpha_{iT}(j,l')}. \tag{4.40}$$

Chapter 5

Simulation results and empirical applications of MHMMs

5.1 Simulation results

To investigate the empirical behavior of the proposed model, we have defined the following simulation study. To model overdispersion with respect to the Poisson distribution and serial dependence for repeated measures over the same unit we generated $R = 250$ samples of size $n = 100, 500, 1000$ and $T = 5, 10$ according to the following scheme:

$$(y_{it} \mid S_{it} = j, b_i) \sim \text{Poisson}(\theta_{itj}), \quad j = 1, 2$$

where the following regression model holds:

$$\log(\theta_{itj}) = \mathbf{x}_{it}^T \boldsymbol{\beta}_j + b_i, = \beta_{0j} + \beta_{1j} x_{it1} + \beta_{2j} x_{it2} + b_i$$
$$j = 1, 2; \quad i = 1, \ldots, n; \quad t = 0, \ldots, T.$$

The covariates were independently drawn from N(0,0.5) densities and

$$B_i \sim N(0, \sigma^2), \quad \sigma^2 = 0.1, 0.5.$$

We assume the following *true* values for the parameter vectors:

$$\delta = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} = \begin{bmatrix} 0.65 \\ 0.35 \end{bmatrix}$$

$$Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} = \begin{bmatrix} 0.65 & 0.35 \\ 0.2 & 0.8 \end{bmatrix}$$

$$\beta_1 = \begin{bmatrix} \beta_{10} \\ \beta_{11} \\ \beta_{12} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.25 \\ 0.75 \end{bmatrix}$$

and

$$\beta_2 = \begin{bmatrix} \beta_{20} \\ \beta_{21} \\ \beta_{22} \end{bmatrix} = \begin{bmatrix} \text{-0.5} \\ 1.75 \\ 1 \end{bmatrix}$$

The simulation was conducted with the aim at investigating the behavior of the proposed model with respect to both sample sizes (n, T) and to the

magnitude of the overdispersion (in terms of σ_b^2), compared to the standard HMM defined in (4.16).

We fitted the corresponding model with a variable number of components, $L = 2, \ldots, 6$ to data generated according to the previous scheme. The model has been estimated following the forward-backward algorithm described in Section 4.3.2.

We used random starting points for Q and δ and retained the model with the best BIC value according to Keribin (2000) who proved the almost sure consistency of the maximum likelihood estimator for an appropriate penalization sequence based on the number of parameters and the number of units in the analyzed sample. Parameter estimates are shown in Tables (5.1) to (5.4), together with the mean log-likelihood value ℓ and the median value for the number of components, L^*. We treated the (π_l, b_l) as nuisance parameters and thus the (estimated) finite mixture is not of direct interest to us. However, the *estimated* number of components in the finite mixture gives a simple measure of the effect of unobservable heterogeneity.

Standard errors for $\hat{\beta}$ have been computed through parametric bootstrap (Efron, 1979) as follows:

$$\hat{se}_R = \left\{ \frac{\sum_{r=1}^{R} [\hat{\beta}(r) - \hat{\beta}_n(\cdot)]^2}{R-1} \right\}^{\frac{1}{2}} \tag{5.1}$$

where $\hat{\beta}(r)$ is the statistic calculated from the r-th resample ($r = 1, \ldots, R$), $\hat{\beta}_n(\cdot) = \sum_{r=1}^{R} \hat{\beta}(r)/R$ and R is the total number of resamples.

As can be noted, Table 5.1 shows a clear and consistent path with respect to Markov parameters as the sample dimension increases, irrespective of the

Table 5.1: Simulation results for Mixed HMM - Markov process parameters

n	T	σ^2	$\hat{\delta}_1$	$\hat{\delta}_2$	q_{11}	q_{12}	q_{21}	q_{22}
100	10	0.1	0.698	0.302	0.648	0.352	0.217	0.783
100	10	0.5	0.674	0.326	0.641	0.359	0.211	0.789
500	10	0.1	0.652	0.348	0.644	0.356	0.197	0.803
500	10	0.5	0.656	0.344	0.655	0.345	0.208	0.792
1000	5	0.1	0.658	0.342	0.659	0.351	0.201	0.799
1000	5	0.5	0.652	0.348	0.655	0.345	0.219	0.781
1000	10	0.1	0.647	0.353	0.639	0.361	0.205	0.795
1000	10	0.5	0.668	0.332	0.650	0.350	0.205	0.795

value of the heterogeneity source, σ^2.

Differences can be observed with varying σ^2 in the regression parameters, since parameter estimates seem more stable for $\sigma^2 = 0.1$, as well as for the median number of components used to estimate the unknown mixing distribution, $\mathcal{H}(\cdot)$, which increases with increasing σ^2 and n.

This implies that the proposed model can be used even in those empirical situations where a large overdispersion arises, although at the cost of an increasing number of components in the estimated finite mixtures.

Comparing our results with the standard HMM defined in (4.16), it is plain to notice that the MHMM is equivalent to an HMM if overdisperion does not influence data in a relevant way. In fact, for $\sigma = 0.1$ both Markov process and regression parameter estimates fit well the "true" parameter values; the HMM estimates seem to be more stable with respect to the constant term and, generally, computationally less intensive with respect to MHMM. On

Table 5.2: Simulation results for standard HMM - Markov process parameters

n	T	σ^2	$\hat{\delta}_1$	$\hat{\delta}_2$	q_{11}	q_{12}	q_{21}	q_{22}
100	10	0.1	0.657	0.343	0.648	0.352	0.209	0.791
100	10	0.5	0.753	0.247	0.865	0.135	0.080	0.920
500	10	0.1	0.648	0.352	0.646	0.354	0.211	0.789
500	10	0.5	0.794	0.206	0.860	0.140	0.095	0.905
1000	5	0.1	0.652	0.348	0.653	0.347	0.201	0.799
1000	5	0.5	0.800	0.200	0.806	0.194	0.121	0.879
1000	10	0.1	0.654	0.345	0.659	0.341	0.203	0.797
1000	10	0.5	0.794	0.206	0.857	0.143	0.070	0.930

the other hand, if a source of heterogeneity arises, the MHMM outperforms the HMM: both Markov process and regression parameter estimates are more accurate and the HMM parameter estimates are strongly biased. Hence, we can assert that, if overdispersion is present, HMMs are not able to distinguish between serial dependence and heterogeneity sources, producing biased and not consistent parameter estimates; while MHMMs, even if computationally more intensive, provide consistent and stable parameter estimates.

5.2 Empirical applications

5.2.1 RAND Health Insurance Experiment

It is well known that the analysis of the demand of health care depends on the empirical specification used in the analysis; therefore, if such specification

Table 5.3: Simulation results for MHMM - Regression parameters

n	T	σ^2	$\hat{\beta}_{10}$ (SE)	$\hat{\beta}_{11}$ (SE)	$\hat{\beta}_{12}$ (SE)	$\hat{\beta}_{20}$ (SE)	$\hat{\beta}_{21}$ (SE)	$\hat{\beta}_{22}$ (SE)	$\hat{\ell}$	BIC	L^*
100	10	0.1	0.497	0.235	0.736	-0.487	1.689	1.006	-1487.7	3058.4	2
			(0.491)	(0.114)	(0.089)	(0.524)	(0.152)	(0.230)			
100	10	0.5	0.578	0.247	0.744	-0.504	1.661	1.032	-46894	3220.3	2
			(0.651)	(0.158)	(0.380)	(0.539)	(0.546)	(0.825)			
500	10	0.1	0.496	0.263	0.754	-0.510	1.753	0.996	-7459.5	15021	2
			(0.513)	(0.044)	(0.031)	(0.504)	(0.071)	(0.126)			
500	10	0.5	0.509	0.245	0.747	-0.492	1.736	0.996	-7794.1	15730	4
			(0.496)	(0.032)	(0.045)	(0.497)	(0.109)	(0.070)			
1000	5	0.1	0.492	0.247	0.756	-0.485	1.723	1.008	-7343.5	14789	2
			(0.478)	(0.045)	(0.031)	(0.500)	(0.071)	(0.031)			
1000	5	0.5	0.545	0.248	0.752	-0.521	1.749	1.011	-7807.8	15747	3
			(0.592)	(0.118)	(0.286)	(0.518)	(0.379)	(0.682)			
1000	10	0.1	0.4987	0.249	0.744	-0.5127	0.998	1.763	-14902	29914	2
			(0.457)	(0.031)	(0.022)	(0.441)	(0.055)	(0.045)			
1000	10	0.5	0.513	0.245	0.749	-0.485	1.725	1.015	-10766	31419	4
			(0.518)	(0.089)	(0.031)	(0.510)	(0.089)	(0.063)			

Table 5.4: Simulation results for standard HMM - Regression parameters

n	T	σ^2	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{20}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$
			(SE)	(SE)	(SE)	(SE)	(SE)	(SE)
100	10	0.1	0.518	0.255	0.735	-0.543	1.763	1.016
			(0.245)	(0.798)	(0.205)	(0.063)	(0.204)	(0.772)
100	10	0.5	1.017	0.455	0.740	-0.254	1.078	0.874
			(0.207)	(0.436)	(0.161)	(0.179)	(0.126)	(0.382)
500	10	0.1	0.528	0.238	0.734	-0.522	1.7618	1.016
			(0.089)	(0.752)	(0.148)	(0.024)	(0.747)	(0.148)
500	10	0.5	1.077	0.548	0.777	-0.165	0.894	0.841
			(0.071)	(0.281)	(0.114)	(0.141)	(0.044)	(0.155)
1000	5	0.1	0.510	0.248	0.751	-0.506	1.752	0.995
			(0.071)	(0.763)	(0.134)	(0.032)	(0.130)	(0.758)
1000	5	0.5	1.087	0.568	0.787	-0.235	0.973	0.874
			(0.077)	(0.202)	(0.055)	(0.100)	(0.255)	(0.077)
1000	10	0.1	0.518	0.253	0.744	-0.511	1.010	1.751
			(0.055)	(0.683)	(0.122)	(0.031)	(0.689)	(0.130)
1000	10	0.5	1.069	0.550	0.776	-0.156	0.856	0.8216
			(0.044)	(0.145)	(0.044)	(0.054)	(0.137)	(0.148)

does not correspond to the underlying behavioral structures that drive the demand of health care, the corresponding estimates may be inconsistent. When defining regression models for the utilization of health care resources, we have to take into account two main characteristics of analyzed data: first, the observed outcome (i.e. the number of visits to a general practitioner) can take only non-negative integer values. This calls for the application of count data models; Poisson regression models may represent a natural starting point in such a context. Clearly, this model is restrictive in that it assumes equality of mean and variance; further, Poisson models are practically not suitable for data which are characterized by an excess of zeros.

The second characteristic of health care utilization is, in fact, a potential two-part decision process: the first process entails the decision to contact a doctor while the second considers the decision about the number of visits. In Poisson models this two-part feature is ignored and this may lead to model misspecification and hence to inconsistent parameter estimates. The demand for medical services is often characterized by a high incidence of zero usage, therefore two-part models (i.e. hurdle models) have become increasingly popular in the last years. The appeal of two-part models in health economics is also based on its connection to a principal-agent model where the physician (the agent) determines utilization on behalf of the patient (the principal) once initial contact is made. Recent literature provides comparison between the relative performance of hurdle models with respect to finite mixture approaches; Deb and Trivedi (2002) present evidence that finite mixture models often outperform the hurdle models, but there is no general evidence and in some cases the hurdle model can better fit the observed data.

We use data from the Rand Health Insurance Experiment (RHIE) for this study. The RHIE is a comprehensive study of health care cost, utilization and outcome in the United States. It is thought to provide the most persuasive evidence to date on the relative effects of health maintenance organizations and fee-for-service care on demand for health care and health care outcomes; in particular it addresses a main topic: evaluation of how much more people use health services if they are provided free of charge.

RHIE started in 1971 using fundings from the United States Department of Health, Education, and Welfare. Its aim is to give some policy issue that is useful for the restructuring of private insurance system and helped increase the stature of managed care. We consider data recorded on 1164 families (in Dayton, Ohio) insured by companies randomly assigned to insurance plans that either had no cost-sharing, 25, 50 or 95% copayment rates and the sample consists of 4462 observations (individuals in the fee-for-service plans). Detailed information on the experimental design and data collection methods are reported in Morris (1979) and Newhouse et al. (1993) and a summary of the major findings of the RHIE can be found in Keeler (1992). An important result of the experiment is that people facing higher cost-sharing (that is, they had to pay a higher proportion of total health care costs out of their own pockets) had lower health care spending than those in plans with lower cost-sharing. It is well known that overconsumption of health services is one of the main causes of the steadily increasing cost of health care in most countries. This paper examines a mixture model for unobserved heterogeneity in an HMM context to apply the RHIE results in health policy analysis. The key variable used to explain health care demand

in the RHIE is the number of outpatient visits to a physician, *mdu*; the adopted covariates and response are defined in Table 5.5.

In the recent literature, RHIE data have been analyzed by Deb and Trivedi (2002) and Bago d'Uva (2005) comparing finite mixture and two part models. The empirical analysis provides the distinction between two sub-population: the healthy and the ill (see Deb and Trivedi, 2002 for more details).

We model the individual heterogeneity through a set of common random effects; the choice of a Poisson mixture model versus a Negative Binomial (NB) model is due by it can be argued the NB model should not consider two different sources of heterogeneity (the overdispersion parameter in the NB and the finite mixture) and, therefore, the Poisson distribution should be used. Some attractive features are due to the utilization of panel data model: it accounts for individual heterogeneity and it allows for identification of the mixture. In the following, we propose the results for the panel data model described in Section 4.3.1.

We consider a two-state MHMM in line with models estimated in Deb and Trivedi (2002) and Bago d'Uva (2006). The two hidden states can be seen as low users and high users. Due to the possibility of convergence to local maxima in mixture models, the estimation should be repeated using different sets of starting values for the parameters being estimated.

As Table 5.6 the probability of belonging to the state of high users is 0.338 at the beginning of the study and then change with time according with Q, hence an high user a time t could be a low user at time $t+1$ with probability 0.337; while a low user have an high probability of being again

Table 5.5: RAND data - Variable definitions and summary statistics

Variable	*Definition*	*Mean*	*St.Dev.*
MDU	Number of outpatient visits to an MD	2.861	4.505
LC	ln(coinsurance + 1), $0 \leq coinsurance \leq 100$	1.710	1.962
IDP	If individual deductible plan: 1, otherwise: 0	0.220	0.414
LPI	ln(max(1, annual participation incentive payment))	4.709	2.697
FMDE	If IDP = 1: 0, otherwise ln(max(1,MDE/(0.01 coinsurance)))	3.153	3.641
LINC	ln(family income)	8.708	1.228
LFAM	ln(family size)	1.248	0.539
AGE	Age in years	25.718	16.768
FEMALE	If person is a female: 1	0.517	0.500
CHILD	If age is less than 18: 1	0.402	0.490
FEMCHILD	FEMALE $*$ CHILD	0.194	0.395
BLACK	If race of household head is black: 1	0.182	0.383
EDUCDEC	Education of the household head in years	11.967	2.806
PHYSLIM	If the person has a physical limitation: 1	0.124	0.322
DISEASE	Index of chronic disease	11.244	6.742
HLTHG	If self-rated health is good: 1	0.362	0.481
HLTHF	If self-rated health is fair: 1	0.077	0.267
HLTHP	If self-rated health is poor: 1	0.015	0.121

a low user (equals to 0.891). It is interesting to compare the estimated coefficients in the two states; all of them have the same sign in both states except for the case of the effect of $LPI, FMDE$ and $HEALTHG$. There are significant differences in the effects of the covariates in the two states. Furthermore, it could be interesting to analyze the estimate of $\mathcal{H}$ which is a five point distribution with masses, $\pi = (0.072, 0.520, 0.295, 0.075, 0.038)$ on locations [(-5.709,0.514); (-3.596, 0.210); (-2.700, 1.002); (-1.796, 2.141); (-1.165,-0.471)]. In this way we can suppose to classify in such groups people who use a certain health care service for specific chronic diseases due to a case of illness. Furthermore, as a by-product of the analysis we measure overdispersion in both states, through σ_b^2. As can been seen from Table 5.6, low users show greater dispersion than high users, hence a Poisson kernel is not suited for fitting data for low users due the presence of excess of zero.

5.2.2 A pharmaceutical study

We briefly mention an example drawn from Min and Angresti (2005), who evaluated the number of episodes of a certain side effect for a particular disease; taking as a starting point a pharmaceutical study, they reconstructed the original dataset keeping the zero inflated structure of data. The study entails 118 patients, with 59 randomly allocated to receive treatment A (TRT1) and the other 59 receiving treatment B (TRT2). The number of side effect episodes was measured at each of six visits. About 83% of the observations were zeros. Table 5.7 shows the frequencies of the side effect for treatments A and B.

The observed process is fitted through the model described in Section 2

Table 5.6: RHIE data - MHMM

	MDU			
	Low users (L)		High users (H)	
VARIABLE	Coef.	s.e.	Coef.	s.e.
$\hat{\delta}_1$	0.662			
$\hat{\delta}_2$	0.338			
$\hat{q}_{LL}$	0.891			
$\hat{q}_{LH}$	0.109			
$\hat{q}_{HL}$	0.337			
$\hat{q}_{HH}$	0.663			
CONSTANT	-1.683	0.016	0.382	0.049
LC	-0.110	0.013	-0.132	0.017
IDP	-0.906	0.028	-0.480	0.031
LPI	0.067	0.042	-0.015	0.005
FMDE	-0.033	0.031	0.056	0.008
LINC	0.420	0.016	0.062	0.010
LFAM	-0.135	0.006	-0.127	0.022
AGE	0.007	0.001	0.000	0.000
FEMALE	0.757	0.039	0.327	0.026
CHILD	0.648	0.061	0.257	0.044
FEMCHILD	-0.685	0.033	-0.370	0.043
BLACK	-0.588	0.062	-0.062	0.038
EDUCDEC	0.059	0.006	0.023	0.004
PHYSLIM	0.433	0.045	0.346	0.034
DISEASE	0.019	0.002	0.005	0.001
HLTHG	0.006	0.067	-0.084	0.022
HLTHF	0.150	0.131	0.257	0.041
HLTHP	0.286	0.059	0.513	0.069
σ_b^2	3.15		0.95	
BIC	19693			
$\log - likelihood$	-9632.1			

Table 5.7: Pharmaceutical study data - Side effect frequencies in treatment A and treatment B

	Frequencies						
Treatment	0	1	2	3	4	5	6
A	312	30	11	0	1	0	0
B	278	39	20	6	7	2	2
Total	590	69	31	6	8	2	2

where the Poisson random effects model has the form:

$$\log(\theta_{itj}) = \beta_{0j} + \beta_{1j}\text{TRT2} + \beta_{2j}\log(\text{Time}) + b_i \tag{5.2}$$

where b_i have a common unknown distribution and the results are obtained by the NPML approach described in Section 4.3.2; furthermore, as the count data vary with exposure time between visits, we incorporated time-between-visit (defined as Time) as a covariate in the model.

The results in Table 5.8 show that not taking into account unobserved heterogeneity may lead to biased parameter estimates and to an uncorrect interpretation of the analyzed phenomena.

Adopting the selection criteria described in Section 4.3.2, we identify 2 hidden states in the model. In this application the states are not only a tool for modelling time-dependence but also have a physical meaning: state 1 includes all the individuals that show a good response to both treatments, while in state 2 a detrimental effect on the number of episodes due to treatment B characterizes individuals. Most of individuals are modelled in state 1 at the initial time ($t = 1$, $\delta_1 = 0.920$), but individuals move through states over time and here two mechanisms are at play:

Table 5.8: Pharmaceutical study data - MHMM vs. HMM

	MHMM		Mixture of Poisson (complete sample)		Mixture of Poisson (selected sample)	
Parameters	Estimates	Std Err	Estimates	Std Err	Estimates	Std Err
$\hat{\delta}_1$	0.920		-	-	-	-
$\hat{\delta}_2$	0.080		-	-	-	-
$\hat{q}_{11}$	0.435		-	-	-	-
$\hat{q}_{12}$	0.565		-	-	-	-
$\hat{q}_{21}$	0.137		-	-	-	-
$\hat{q}_{22}$	0.863		-	-	-	-
$\hat{\beta}_{10}$ (Intercept)	0.225	0.457	-2.057	0.368	-3.047	0.655
$\hat{\beta}_{11}$ (TRT2)	-3.98	0.220	0.696	0.184	-0.070	0.242
$\hat{\beta}_{12}$ (log(Time))	-0.795	0.195	0.227	0.112	-1.488	0.173
$\hat{\beta}_{20}$ (Intercept)	-4.682	0.741	-3.286	0.368	-1.066	0.655
$\hat{\beta}_{21}$ (TRT2)	1.665	0.766	0.248	0.155	0.133	0.257
$\hat{\beta}_{22}$ (log(Time))	0.583	0.270	-0.616	0.057	-0.609	0.237
$\sigma^2_{b_1}$	0.88		-	-	-	-
$\sigma^2_{b_2}$	0.01		-	-	-	-
ℓ	-390.21		-836.81		-286.25	
BIC	859.17		1724.41		618.43	

- individuals who respond well to treatments at time t may not have effective improvements in health conditions at time $t+1$ ($\hat{q}_{12} = 0.565$)
- ther is a low probability that if the treatment B is not effective at time t it will be effective in next time $t+1$ ($\hat{q}_{21} = 0.137$)

Those conditions induce us to think the treatment B produces a positive effect, reducing the number of episodes, for a while and a detriment if it is over-utilized. To prove such conclusion we model a two-state finite mixture model on the whole sample and considering only the first 3 observations for all individuals and, hence, compare these results with those provided by the MHMM. As can been seen, β_1, for the *selected* sample, show a negative coefficient in one of the components, but is not statistically significant; this could mean that the treatments do not produce any evident changes in the number of episodes if we give treatments for few times and, on the other hand, it could be interpreted as a presence of unobserved effects that act along with the provision of treatments and those effects are keep by β_1 showing *conflicting* coefficients. Furthermore, analyzing the whole sample, we obtain similar results as Min and Agresti (2005): treatment B has a higher probability of the side effect and a higher number of episodes than treatment A.

The finite mixture approach seems to be inadequate to estimate model parameters; adopting MHMM we can model such effects related to the times of given treatments that influence the number of episodes and that depend on unobserved factors depending on time too. In fact, it suggest a common behavior of patients with respect to the response to treatments, with a minimal difference in the size of the effect, measured by β_1. In fact, the two state of the hidden Markov chain can be interpreted as the propensity of a positive

response to the two treatments. Hence, the MHMM points out an interesting and undiscovered (till now) behavior: starting from a state where all patients show a good response to both treatments, patients move to a different state showing that, with increasing time and the effect of treatment B becomes is less effective. It means that if we consider, for example, only the first three observation for each patient, we denote a certain effective influence of the treatment on the number of the episodes, but the marginal effect decrease with increasing time.

Chapter 6

Clustering three-way time dependent data through MHMMs

6.1 Introduction

Clustering methods generally aim at partitioning objects into meaningful classes (also called clusters), maximizing the homogeneity (or similarity) within a group as well as the difference between groups (Everitt, 1993). Standard clustering approaches (see e.g. Johnson, 1967 and McQueen, 1967) have been considerably improved, allowing for solutions to some practical issues such as the choice of the number of clusters, the allocation to clusters and the clustering algorithm adopted.

Model based clustering approaches deal with these issues assuming that the objects under study are drawn from a known probabilistic model with

the aim at recovering the parameters of such a process. Estimation is usually obtained through maximum likelihood, with an overfitting penalty.

Standard finite fixture approaches (briefly discussed in section 2.11), see e.g. McLachlan and Peel (2000a), have been mainly developed with multivariate normal component-specific distributions, see e.g. McLachlan and Basford (1988); a notable exception is represented by the work on t-mixture factor analyzers of McLachlan and Peel (2000b). The importance of mixture distributions is remarked by a number of recent books on mixtures including Lindsay (1995), Böhning (2000), McLachlan and Peel (2000a) and Frühwirth-Schnatter (2006) which update previous books by Everitt and Hand (1981), Titterington et al. (1985) and McLachlan and Basford (1988).

A further generalization, in such a context, is represented by mixtures-of-experts models (Jacobs et al., 1991) and their generalization, hierarchical mixtures-of-expert models (Jordan and Jacobs, 1994), introduced to account for nonlinearities and other complexities in the data. The problem of model mixing in time series has been often treated using this approach (Huerta et al. 2001), that allows for comparisons of arbitrary models, not restricted to a particular class or parametric form. Additionally, the approach is flexible enough to incorporate exogenous information that can be summarized in terms of covariates or simply time, through weighting functions that define the hierarchical mixture, localizing the comparisons to specific regions or regimes through the hierarchical structure (Huerta et al., 2003). Recently, Carvalho and Tanner (2007) study a class of hierarchical mixtures of Poisson experts to model nonlinear count time series. Identifiability and maximum likelihood estimation via the EM are discussed. Extending previous

results for independent observations, asymptotic normality of the maximum likelihood estimator under stationarity and nonstationarity of the covariates vector (which may include lags of transformations of a response and lags of external predictors) is provided.

Generally, finite mixture models have been used to cluster two-way data sets. Recently, three-way data sets have become popular, containing for example attributes (variables) measured on objects (statistical units) in several conditions (occasions, time points, environments, etc.). Basford and McLachlan (1985) have proposed a finite mixture model for the analysis of such data, where the aim is to cluster objects by explicitly taking simultaneously into account the information on variables and occasions. Hunt and Basford (1999, 2001) have extended the approach to deal with categorical variables in unbalanced panels, while Meulders et al. (2002) have proposed a constrained latent class model for the analysis of three-way binary data. All these models assume that cases belong to the same cluster in all investigated situations. Vermunt (2007) proposes an extension of this approach assuming that objects may be in a different latent class depending on the situation or, more specifically, objects are clustered with respect to the probability of being in a particular latent class at a given situation. Relevant work in this topic include, among others, Böhning et al. (2000) and Knorr-Held and Raßer (2000). Vermunt (2007) considers the three ways as hierarchically nested levels and models a mixture distribution at each of the two higher levels; i.e., one at the object and one at the object-in-occasion level. The proposed model is an adaptation of the multilevel latent class model by Vermunt (2003) to continuous responses and has the advantage that it may yield

more parsimonious and insightful solutions than the Basford and McLachlan (1985) model.

We have seen that a natural extension of mixture models for time dependent data is represented by HMMs; thus, a direct generalization in the hierarchical mixture context for solving the problem of mixing in the time dimension may be given adapting MHMMs to hierarchical classification. Hence, we introduce a hierarchical extension of the finite mixture model proposed by Basford and McLachlan (1985), mimicing the proposal of Vermunt (2007). In particular we discuss the issue of longitudinal multivariate data allowing for both time and local dependence.

6.2 Model-based approach to three-way data clustering

A three-way dataset is often produced as the result of the observation of a multivariate-multioccasion phenomenon, characterized by various attributes measured for a set of observational units in different situations; in particular, we will refer to such data as three-mode three-way data, where a mode is defined as in Carroll and Arabie (1980). Three-way data can be also treated as two-mode three-way data; for instance, Vichi (1995, 1998) propose a one mode classification method of a three-way dataset to cluster the elements of one mode on the basis of the other two and this method can be seen as a synthesis of a set of hierarchical classifications, each defined by applying a hierarchical algorithm to a two-mode matrix of three-way dataset. Another example of two-mode three-way data is given by data in the form of prox-

imities between all the elements have to be clustered (see e.g. Bocci et al., 2006).

Clustering methods for three-mode three-way data are available, which either combine clustering and ordination, such as those of Ceulemans et al. (2003), Miyano and Kroonenberg (2003) and Rocci and Vichi (2003); while applications of three-way cluster methods are discussed in Kroonenberg et al. (1995; 2004), Basford et al. (1991) and Chapman (1997).

Let $y_{i,1:P,0:T}$, $i = 1, \ldots, n$, be a PT-dimensional observation corresponding to the i-th unit. Under the mixture model proposed by Basford and McLachlan (1985) and extended by Hunt and Basford (2001), $y_{i,1:P,0:T}$ is assumed to be drawn from the finite mixture of Gaussian distribution:

$$f(y_{i,1:P,0:T}) = \sum_{g=1}^{G} \pi_g \prod_{t=1}^{T} f_g(y_{i,1:P,t}; \mu_{gt}, \Sigma_g) \tag{6.1}$$

where individuals belong to one of G possible groups in proportions $\pi_1, \ldots, \pi_G$, with $\sum_{g=1}^{G} \pi_g = 1$ and $\pi_g > 0$ for $g = 1, \ldots, G$; μ_{gt} is the cluster-time dependent mean vector and Σ_g is the cluster dependent covariance matrix. As pointed out by Vermunt (2007), such modeling approach implicity assumes that the responses are conditionally (on the cluster) independent and does not take into account the possibility that individuals may move across clusters.

Developing the multilevel latent class model, Vermunt (2007) relaxes the assumption of time invariant clustering. In details, $y_{i,1:P,0:T}$ is drawn from one of G 2nd level component (clusters) and a *new* element is that conditional on belonging to g, in situation t cases are assumed to belong to one of $l = \{1, \ldots, L\}$ groups. In the following, we will assume that, conditional on the 2nd level, the response of case i in situation t has a multivariate normal

distribution $y_{i,1:P,t} \sim MVN(\mu_{lt}, \Sigma_l)$:

$$f_l(y_{i,1:P,t} \mid \mu_{lt}, \Sigma_l) = \frac{1}{\sqrt{(2\pi)^P}} |\Sigma_l|^{-\frac{1}{2}} exp\left\{ -\frac{1}{2}(y_{i,1:P,t} - \mu_{lt})^T \Sigma_l^{-1} (y_{i,1:P,t} - \mu_{lt}) \right\} \tag{6.2}$$

where the within-class covariance matrix, Σ_l, is time independent. The hierarchical mixture model has the following form:

$$f(y_{i,1:P,0:T}) = \sum_{g=1}^{G} \pi_g \prod_{t=1}^{T} \sum_{l=1}^{L} \pi^*_{l|g} f_l(y_{i,1:P,t} \mid \mu_{lt}, \Sigma_l) \tag{6.3}$$

where π_g, $\sum_{g=1}^{G} \pi_g = 1$, is the prior probability that the observation $y_{i,1:P,0:T}$ belongs to the g-th cluster ($g = 1, \ldots, G$), $\pi_{l|g} = \Pr(i \in l \mid i \in g)$, $\sum_{l=1}^{L} \pi_{l|g} = 1$, is the conditional probability that the i-th observation in situation t belongs to the l-th component within the g-th cluster ($l = 1, \ldots, L; g = 1, \ldots, G$). In other words the 2nd level cluster control for potential heterogeneity across statistical units with respect to occasion specific clusters.

It should be noted that this model is equivalent to the model of Basford and McLachlan (1985) described in equation (6.1) if $L = G$ and if $\pi_{l|g} = 1$ for $l = g$ and 0 for $l \neq g$; that is, if cases belong to the same class in each situation. This shows that the hierarchical model extends the standard model by allowing cases to be in a different latent class in each situation.

Parameter estimation is performed through a modified EM algorithm. wher two vectors of indicator variables are introduced; namely, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})'$ with $z_{ig} = 1$ if $y_{i,1:P,0:T}$ belongs to the g-th (2nd level) cluster and $\mathbf{w}_{it} = (w_{it1}, \ldots, w_{itL})'$ with $w_{itl} = 1$ if $y_{i,1:P,t}$ belongs to the l-th (1st level) component in situation t. By treating these component labels as missing data,

maximum likelihood estimation can be achieved by means of the EM algorithm.

The log-likelihood for complete data under this model has the following form:

$$\sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\log\pi_g + \sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t=1}^{T}\sum_{l=1}^{L} z_{ig}w_{itl}\log\pi_{l|g} + \sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t=1}^{T}\sum_{l=1}^{L} z_{ig}w_{itl}\log f_l(y_{i,1:P,t};\mu_{lt},\Sigma_l) \quad (6.4)$$

Due to the high dimensionality of the estimation problem, a standard EM cannot be applied; rather the upward-downward algorithm (Pearl, 1988), which is similar to the forward-backward algorithm (Baum et al., 1970) used in the HMM framework, can be used in the implementation of the E-step.

6.3 Multivariate MHMM for clustering three-way data

The proposed model aims at extending mixture model for clustering three-mode three-way data (see e.g. Vermunt, 2007) to longitudinal data, where situations correspond to times and observations for each unit are likely correlated. As before, we adopt a HMM (Cappé et al., 2005) to handle time dependence where the hidden dynamics of the stochastic process are governed by a Markov chain. The extension is defined not only to account for individual dynamics. In fact, since units may be heterogeneous, we adopt a finite mixture model where components representing clusters show different transition matrices for the HMMs. Scott et al. (2005) provides a HMM

for longitudinal comparison. His most compelling methodological advance is the hierarchical inhomogeneous model for allowing data to decide the extent of the compromise between fitting each period's transition probabilities independently and fitting a global transition matrix for the entire model. The HMM approach offers several advantages over the unsupervised learning approach; in fact, cluster based methods involve assuming that each observation's state membership is known rather than estimated, introducing potential bias into the analysis. By contrast HMM parameters estimated automatically incorporate all sources of uncertainty, conditional on the model being correct[1].

More generally, we would like to select a multivariate HMM whose latent states correspond to association structures that receive support from the data-not always, but at least for considerable periods of time. We remark that the applicability of multivariate HMMs is quite wide: it applies to any multivariate time series whose dependency structure is thought to change considerably over time. Further important examples include, among others, environmental data, typically multivariate and never measured exhaustively, and financial times series, where the state of a national economy, e.g., is a powerful qualitative mechanism that determines changes in the correlation structure among the considered variables.

Let us define multivariate Gaussian hidden Markov models. Consider an HMM with Y_t being multidimensional and with the conditional distribution of Y_t given $S_t = j$ being $MVN(\mu_j, \Sigma_j)$, i.e., multivariate Gaussian with state

[1]Scott (2005) makes a comparison between the HMM and the k-means approach (using Bayesian methods).

dependent mean and covariance matrix. The unconditional distribution of Y_t is thus a mixture of multivariate Gaussian distributions. Such a multivariate time series model may be of interest in several areas, as mentioned above. In such formulation the Markov chain S_t governs also the precision matrix $(\Sigma_j)^{-1}$ (as pointed out by Giudici et al., 2000); hence, S_t governs the dependence structure within Y_t. As S_t is a random process, this structure may change over time. Moreover, obviously, the state S_t, also carries information about the numerical values of variances and covariances of Y_t. It may well be the case that different values of S_t, correspond to the same dependence structure within Y_t, although with different variances and/or covariances.

Identifiability of finite mixtures of multivariate Gaussian distributions has been established by Yakowitz and Spragins (1968), whence those results may be applied to multivariate Gaussian HMMs. However, an interesting problem concerns the standard asymptotic theory of likelihood-ratio tests (Giudici et al., 2000) which is based on the idea we may not compare two models with a different number of states. A reasonable assumption is that no two states coincide in the sense of having identical covariance matrices since then we would effectively have one state less than specified by the model. In such a context, the maximum likelihood estimator cannot be strongly consistent (in a simple sense) if the model is overparameterized in the way of specifying more states than there actually are (m is too large). This is because the *true* parameter is then not unique, and there is no unique point around which to expand the log likelihood when analyzing the log-likelihood ratio tests.

Starting from the usual framework for multivariate Gaussian HMM ,described above, we will focus on empirical situations where three-mode three-

way data are analyzed in a hierarchical framework where one of the mode indexes time, i.e. longitudinal data.

Recalling the notation we have already introduced before in this thesis; let us consider a sequence $\{S_{it}\}$, $i = 1, \ldots, n$, $t = 0 \ldots, T$ of random variables whose values are in a finite and enumerable set $\mathcal{S} = \{1, \ldots, m\}$ and let $\{S_{it}\}$, $i = 1, \ldots, N$, $t = 0 \ldots, T$ be an homgeneous Markov chain:

$$\Pr(S_{it} = j \mid S_{i0}, \ldots, S_{it-1}) = \Pr(S_{it} = j \mid S_{it-1}), \quad \forall j \in \mathcal{S}.$$

Developing Vermunt (2007), we model time dependence in 1st level clusters using a HMM framework. In detail, the hierarchical mixture (6.3) can be rewritten as follows:

$$f(y_{i,1:P,0:T}) = \sum_{g=1}^{G} \pi_g \sum_{\mathcal{S}^T} \left\{ \delta^{(g)}_{s_i 0} \prod_{t=1}^{T} q^{(g)}_{s_{it-1}, s_{it}} \prod_{t=0}^{T} f_{s_{it}}(y_{i,1:P,t} \mid \mu_{s_{it}}, \Sigma_{s_{it}}) \right\} \quad (6.5)$$

where $\delta^{(g)}_{s_{i0}} = \Pr(S_{i0} = s_{i0} \mid g)$; $q^{(g)}_{s_{it-1}, s_{it}} = \Pr(S_{it} = s_{it} \mid S_{it-1} = s_{it-1}, g)$ and $f_{s_{it}}$ is as in (6.2).

As can be easily noted, we introduce a further assumption to accommodate time dependence in the multilevel model proposed: we don't drop the assumption that the Markov chain is time-homogeneous, but we assume that the HMM is inhomogeneous in the sense that the Markov process can be modelled as depending on the second level classification (or time-homogeneous conditional on g, $g = 1, \ldots, G$). Thus different clusters have different propensities to be in a given state, and different transitions from one state to another one.

6.4 Computational details

In this section, we discuss a modified EM algorithm for MLE of the multilevel model parameters. In other words, each unit can be thought of as drawn from a finite mixture of G HMM.

To introduce the algorithms, let us denote with

$$\gamma_{jtg} = \Pr(S_{it} = j \mid g, y_{i,1:P,0:T}) \tag{6.6}$$

the posterior probability, given the individual sequence and the g-th component, of being in state j at time time t and with

$$\xi_{jktg} \quad = \quad \Pr(S_{it-1} = j, S_{it} = k \mid g, y_{i,1:P,0:T}) \tag{6.7}$$

the posterior probability that the unobserved sequence visited state j at time $t-1$ and made a transition to state k at time t, given the g-th component.

The posterior probability that the i-th unit comes from the g-th component of the mixture is as follows

$$\eta_{ig} = Pr(g \mid y_{i,1:P,0:T}) = \frac{\pi_g f(y_{i,1:P,0:T} \mid \theta_g)}{\sum\limits_g \pi_g f(y_{i,1:P,0:T} \mid \theta_g)}. \tag{6.8}$$

where θ_g are the parameters of the g-th component.

The expected log-likelihood function for the model described in equation

(6.5) has the following form:

$$E[\log L_C(\phi)] = \sum_{i=1}^{n}\sum_{g=1}^{G}\eta_{ig}\log\pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{m}\eta_{ig}\gamma_{j1g}\log\delta_j^{(g)} + \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j\in\mathcal{S}^T}\sum_{k\in\mathcal{S}^T}\sum_{t=1}^{T}\eta_{ig}\xi_{jktg}\log q_{jk}^{(g)} + \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j\in\mathcal{S}^T}\sum_{t=0}^{T}\eta_{ig}\gamma_{jtg}\log f_j(y_{it}\mid\theta_j) \tag{6.9}$$

Before deriving the EM algorithm, we recall the forward and the backward procedure, which is central to parameter estimation. We define with

$$\alpha_{it}(j,g) = Pr(y_{i0},\ldots,y_{it}, S_{it}=j \mid g), \tag{6.10}$$

the probability of seeing the partial sequence ending up in state j at time t for a generic unit i in the g-th component. We can efficiently compute $\alpha_{it}(j,g)$ recursively as:

$$\alpha_{i0}(j,g) = \delta_j^{(g)} f(y_{i0}\mid S_{i0}=j) \tag{6.11}$$

$$\alpha_{i,t+1}(k,g) = \sum_{j=1}^{m}\alpha_{it}(j,g)q_{jk}^{(g)} f(y_{i,t+1}\mid S_{i,t+1}=k) \tag{6.12}$$

The backward procedure is similar:

$$\tau_{it}(j,g) = P(y_{i,t+1},\ldots,y_{iT}\mid S_{it}=j,g) \tag{6.13}$$

is the probability of the partial sequence $y_{i,t+1},\cdots,y_{i,T}$ given that we started at state j at time t for a generic unit i in the g-th component. The recursive procedure is given by:

$$\tau_{i,T}(j,g) = 1 \tag{6.14}$$

$$\tau_{it}(j,g) = \sum_{k=1}^{m} q_{jk}^{*} f(y_{i,t+1} \mid S_{i,t+1} = k)\tau_{i,t+1}(k,g). \tag{6.15}$$

We can express the posterior probabilities in equation (6.9) using the forward and the backward variables:

$$\gamma_{jtg} = \frac{\alpha_{it}(j,g)\tau_{it}(j,g)}{\sum\limits_{j} \alpha_{it}(j,g)\tau_{it}(j,g)} \tag{6.16}$$

$$\xi_{jktg} = \frac{\alpha_{i,t-1}(j,g)q_{jk}^{(g)} f(y_{it} \mid j)\tau_{it}(j,g)}{\sum\limits_{j\in\mathcal{S}^T}\sum\limits_{k\in\mathcal{S}^T} \alpha_{it-1}(j,g)q_{jk}^{(g)} f(y_{it} \mid j)\tau_{i,t+1}(j,g)} \tag{6.17}$$

and

$$\eta_{ig} = \frac{\pi_g \sum\limits_{j\in\mathcal{S}} \alpha_{iT}(j,g)}{\sum\limits_{g=1}^{G} \pi_g \sum\limits_{j\in\mathcal{S}} \alpha_{iT}(j,g)} \tag{6.18}$$

Our goal is to update current parameters for the proposed model by using the old parameters and the data. Thus, the maximum value of $\hat{\delta}_j^{(g)}$ is reached at

$$\hat{\delta}_j^{(g)} = \frac{\sum\limits_{i=1}^{N} \eta_{ig}\gamma_{j0g}}{\sum\limits_{i=1}^{n} \eta_{ig}} \tag{6.19}$$

Similarly, we obtain ML estimates for the transition matrix $Q^{(g)}$ and for the weight of the second level of model, π_g:

$$\hat{q}_{jk}^{(g)} = \frac{\sum\limits_{i=1}^{n}\sum\limits_{t=1}^{T} \eta_{ig}\xi_{jktg}}{\sum\limits_{i=1}^{n}\sum\limits_{t=0}^{T-1} \eta_{ig}\gamma_{jtg}} \tag{6.20}$$

and

$$\hat{\pi_g} = \frac{\sum\limits_{i=1}^{n} \eta_{ig}}{n} \tag{6.21}$$

Let us consider a specific state density of the form of (6.2); then, $\theta_j = \{\mu_j, \Sigma_j\}$ where

$$\mu_j = \frac{\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t=0}^{T}\eta_{ig}\gamma_{jtg}y_{it}}{\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t=0}^{T}\eta_{ig}\gamma_{jtg}} \tag{6.22}$$

and

$$\Sigma_j = \frac{\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t=0}^{T}\eta_{ig}\gamma_{jtg}[y_{it}-\mu_j][y_{it}-\mu_j]'}{\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t=0}^{T}\eta_{ig}\gamma_{jtg}} \tag{6.23}$$

These steps are alternated repeatedly until the following relative difference:

$$\frac{|\ell^{(r+1)} - \ell^{(r)}|}{|\ell^{(r)}|} < \epsilon, \quad \epsilon > 0 \tag{6.24}$$

changes by an arbitrarily small amount if the adopted criterium is based on the sequence of log-likelihood values $\ell^{(r)}$ at the r-th iteration. Since $\ell^{(r+1)} \geq \ell^{(r)}$, convergence is obtained with a sequence of likelihood values which are bounded above.

Chapter 7

Simulations of Multivariate MHMMs for clustering three-way time dependent data

7.1 Simulation results

To investigate the empirical behavior of the proposed MHMM in clustering multivariate three-way (time dependent) data, we have defined the following simulation study. We generate $R = 250$ samples of size $n = 100, 500, 1000$ and $T = 10$ from a $MVN(\mu_j, \Sigma_j), j = 1, \ldots, m$. In detail, we focus on bivariate hierarchical mixtures of HMMs according to the following scheme:

$$(y_{i,1:2,t} \mid S_{it} = j) \sim MVN(\mu_j, \Sigma_j), \quad j = 1, 2.$$

where j indexes states of the chain; while

$$\mu_1 = \begin{bmatrix} 0.2 \\ 0.7 \end{bmatrix}$$

$$\mu_2 = \begin{bmatrix} 0.5 \\ 0.4 \end{bmatrix}.$$

The covariance matrices are defined as follows

$$\Sigma_1 = \{\sigma_{p_1,p_2,j=1}\} = \begin{bmatrix} \sigma_{111} & \sigma_{121} \\ \sigma_{211} & \sigma_{221} \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$\Sigma_2 = \{\sigma_{p_1,p_2,j=2}\} = \begin{bmatrix} \sigma_{112} & \sigma_{122} \\ \sigma_{212} & \sigma_{222} \end{bmatrix} = \begin{bmatrix} 0.5 & 0.15 \\ 0.15 & 0.5 \end{bmatrix}$$

Further, according to the model described in section 6.3 we consider the following *true* values for the parameter vectors, assuming $g = 1, 2$:

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

$$\delta^{(1)} = \begin{bmatrix} \delta_1^{(1)} \\ \delta_2^{(1)} \end{bmatrix} = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}$$

$$\delta^{(2)} = \begin{bmatrix} \delta_1^{(2)} \\ \delta_2^{(2)} \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}$$

$$Q^{(1)} = \begin{bmatrix} q_{11}^{(1)} & q_{12}^{(1)} \\ q_{21}^{(1)} & q_{22}^{(1)} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$$

$$Q^{(2)} = \begin{bmatrix} q_{11}^{(2)} & q_{12}^{(2)} \\ q_{21}^{(2)} & q_{22}^{(2)} \end{bmatrix} = \begin{bmatrix} 0.4 & 0.6 \\ 0.25 & 0.75 \end{bmatrix}$$

We used random starting points for $Q^{(g)}$ and $\delta^{(g)}$, $g = 1, 2,;$ μ_j has been drawn from a Gaussian distribution with mean zero and unit variance one, $N(0, 1)$. Finally, to estimate covariance matrices, we start from

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Parameter estimates are shown in 7.1 to 7.7 for $n = 100$, in 7.8 to 7.14 for $n = 500$, in 7.15 to 7.21 for $n = 1000$, together with corresponding variances (within the brackets).

As can be noted, results show a clear and consistent path for Markov and state-specific density parameters as the sample size increases: parameter bias decreases and corresponding estimates show a lower variability as the sample size increases.

Table 7.1: Parameter estimates for n=100

$$\hat{\mu}_1 = \begin{bmatrix} 0.275 & (0.041) \\ 0.846 & (0.084) \end{bmatrix} \quad \hat{\mu}_2 = \begin{bmatrix} 0.643 & (0.049) \\ 0.590 & (0.119) \end{bmatrix} \tag{7.1}$$

$$\hat{\Sigma}_1 = \begin{bmatrix} 0.897 & 0.399 \\ (0.041) & (0.036) \\ 0.399 & 0.855 \\ (0.036) & (0.053) \end{bmatrix} \quad \hat{\Sigma}_2 = \begin{bmatrix} 0.583 & 0.235 \\ (0.044) & (0.026) \\ 0.235 & 0.642 \\ (0.026) & (0.062) \end{bmatrix} \tag{7.2}$$

$$\hat{\pi} = \begin{bmatrix} 0.538 & (0.048) \\ 0.462 & (0.048) \end{bmatrix} \tag{7.3}$$

$$\hat{\delta}^{(1)} = \begin{bmatrix} \hat{\delta}_1^{(1)} \\ \hat{\delta}_2^{(1)} \end{bmatrix} = \begin{bmatrix} 0.636 & (0.099) \\ 0.364 & (0.099) \end{bmatrix} \tag{7.4}$$

$$\hat{\delta}^{(2)} = \begin{bmatrix} \hat{\delta}_1^{(2)} \\ \hat{\delta}_2^{(2)} \end{bmatrix} = \begin{bmatrix} 0.424 & (0.116) \\ 0.576 & (0.116) \end{bmatrix} \tag{7.5}$$

$$\hat{Q}^{(1)} = \begin{bmatrix} \hat{q}_{11}^{(1)} & \hat{q}_{12}^{(1)} \\ \hat{q}_{21}^{(1)} & \hat{q}_{22}^{(1)} \end{bmatrix} = \begin{bmatrix} 0.747 & 0.253 \\ (0.038) & (0.038) \\ 0.252 & 0.748 \\ (0.043) & (0.043) \end{bmatrix} \tag{7.6}$$

$$\hat{Q}^{(2)} = \begin{bmatrix} \hat{q}_{11}^{(2)} & \hat{q}_{12}^{(2)} \\ \hat{q}_{21}^{(2)} & \hat{q}_{22}^{(2)} \end{bmatrix} = \begin{bmatrix} 0.456 & 0.544 \\ (0.070) & (0.070) \\ 0.431 & 0.569 \\ (0.057) & (0.057) \end{bmatrix} \tag{7.7}$$

Table 7.2: Parameter estimates for n=500

$$\hat{\mu}_1 = \begin{bmatrix} 0.221 & (0.012) \\ 0.836 & (0.022) \end{bmatrix} \quad \hat{\mu}_2 = \begin{bmatrix} 0.571 & (0.012) \\ 0.437 & (0.023) \end{bmatrix} \tag{7.8}$$

$$\hat{\Sigma}_1 = \begin{bmatrix} 0.977 & 0.482 \\ (0.012) & (0.009) \\ 0.482 & 0.968 \\ (0.009) & (0.015) \end{bmatrix} \quad \hat{\Sigma}_2 = \begin{bmatrix} 0.526 & 0.172 \\ (0.012) & (0.007) \\ 0.172 & 0.531 \\ (0.007) & (0.015) \end{bmatrix} \tag{7.9}$$

$$\hat{\pi} = \begin{bmatrix} 0.529 & (0.050) \\ 0.471 & (0.050) \end{bmatrix} \tag{7.10}$$

$$\hat{\delta}^{(1)} = \begin{bmatrix} \hat{\delta}_1^{(1)} \\ \hat{\delta}_2^{(1)} \end{bmatrix} = \begin{bmatrix} 0.722 & (0.054) \\ 0.278 & (0.054) \end{bmatrix} \tag{7.11}$$

$$\hat{\delta}^{(2)} = \begin{bmatrix} \hat{\delta}_1^{(2)} \\ \hat{\delta}_2^{(2)} \end{bmatrix} = \begin{bmatrix} 0.425 & (0.053) \\ 0.575 & (0.053) \end{bmatrix} \tag{7.12}$$

$$\hat{Q}^{(1)} = \begin{bmatrix} \hat{q}_{11}^{(1)} & \hat{q}_{12}^{(1)} \\ \hat{q}_{21}^{(1)} & \hat{q}_{22}^{(1)} \end{bmatrix} = \begin{bmatrix} 0.770 & 0.230 \\ (0.015) & (0.015) \\ 0.200 & 0.800 \\ (0.018) & (0.018) \end{bmatrix} \tag{7.13}$$

$$\hat{Q}^{(2)} = \begin{bmatrix} \hat{q}_{11}^{(2)} & \hat{q}_{12}^{(2)} \\ \hat{q}_{21}^{(2)} & \hat{q}_{22}^{(2)} \end{bmatrix} = \begin{bmatrix} 0.413 & 0.587 \\ (0.041) & (0.041) \\ 0.314 & 0.686 \\ (0.025) & (0.025) \end{bmatrix} \tag{7.14}$$

Table 7.3: Parameter estimates for n=1000

$$\hat{\mu}_1 = \begin{bmatrix} 0.207 & (0.005) \\ 0.804 & (0.010) \end{bmatrix} \quad \hat{\mu}_2 = \begin{bmatrix} 0.590 & (0.005) \\ 0.414 & (0.010) \end{bmatrix} \tag{7.15}$$

$$\hat{\Sigma}_1 = \begin{bmatrix} 0.989 & 0.491 \\ (0.005) & (0.004) \\ 0.491 & 0.985 \\ (0.004) & (0.007) \end{bmatrix} \quad \hat{\Sigma}_2 = \begin{bmatrix} 0.509 & 0.158 \\ (0.005) & (0.003) \\ 0.158 & 0.511 \\ (0.003) & (0.006) \end{bmatrix} \tag{7.16}$$

$$\hat{\pi} = \begin{bmatrix} 0.511 & (0.052) \\ 0.489 & (0.052) \end{bmatrix} \tag{7.17}$$

$$\hat{\delta}^{(1)} = \begin{bmatrix} \hat{\delta}_1^{(1)} \\ \hat{\delta}_2^{(1)} \end{bmatrix} = \begin{bmatrix} 0.748 & (0.035) \\ 0.252 & (0.035) \end{bmatrix} \tag{7.18}$$

$$\hat{\delta}^{(2)} = \begin{bmatrix} \hat{\delta}_1^{(2)} \\ \hat{\delta}_2^{(2)} \end{bmatrix} = \begin{bmatrix} 0.436 & (0.038) \\ 0.564 & (0.038) \end{bmatrix} \tag{7.19}$$

$$\hat{Q}^{(1)} = \begin{bmatrix} \hat{q}_{11}^{(1)} & \hat{q}_{12}^{(1)} \\ \hat{q}_{21}^{(1)} & \hat{q}_{22}^{(1)} \end{bmatrix} = \begin{bmatrix} 0.785 & 0.215 \\ (0.012) & (0.012) \\ 0.208 & 0.792 \\ (0.012) & (0.018) \end{bmatrix} \tag{7.20}$$

$$\hat{Q}^{(2)} = \begin{bmatrix} \hat{q}_{11}^{(2)} & \hat{q}_{12}^{(2)} \\ \hat{q}_{21}^{(2)} & \hat{q}_{22}^{(2)} \end{bmatrix} = \begin{bmatrix} 0.387 & 0.613 \\ (0.031) & (0.031) \\ 0.283 & 0.717 \\ (0.014) & (0.014) \end{bmatrix} \tag{7.21}$$

Chapter 8

Final remarks

Our first contribution is to provide a rigorous and flexible approach to estimation in MHMMs. When a longitudinal study is considered, it is important to take into account that individuals do not only differ in their preferences at a specific time-point but also in the way they change their preferences over time. Discrete-time HMMs have been used to understand preference changes, due to the relatively easy interpretation and tractability of Markov chains. However, these preferences may depend on unobservable individual-specific factors; therefore, the random effect extension leads to a more adequate specification of such preference structures by modeling individual-specific variation in the regression parameters, keeping a readily interpretation and estimation of the results. We discuss this extension of HMMs in a semi-parametric ML framework, which is alternative to the model discussed by MacKay (2007). Efficient numerical methods to compute MLE for this kind of models is of primary interest. With respect to the numerical computation of MLE, two ways are possible. The first could be based on standard

(stochastic or deterministic) likelihood maximization techniques, using recursive forward-backward algorithm. The second one could be based on an adaptation of the EM algorithm.

We apply this proposal to overdispersed (i.e. zero-inflated) count data where an unobserved source of heterogeneity arises. To model such overdispersion, the use of finite mixtures have some significant advantages over parametric mixture models. First, the discrete nature of the mixing distribution estimate help classify subjects in clusters characterized by homogeneous values of regression parameters, and this is particularly important in behavioral sciences, where components can be interpreted as groups with similar features. Second, since locations and corresponding probabilities are completely free to vary over the corresponding support, the proposed approach can readily accommodate extreme departures from the basic (i.e. Poisson) regression model.

Furthermore, a novel mixture clustering model is presented for the analysis of three-way data, where the third way represents time occasion. The proposed method presents a possible solution to the problem of time dependence in hierarchical mixture models; particularly, it overcomes some limits of previous solutions proposed for time-dependent data, i.e. longitudinal data. Its structure allows class membership change over time through a hidden Markov chain. Here a bivariate case with two groups and two states has been discussed, but we are working on extension to multivariate, multigroups and multistates models.

Bibliography

[1] Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251–262.

[2] Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117–128.

[3] Albert, P.S. (1991) A two-state markov model for a time series of epileptic seizure counts. *Biometrics*, **47**, 1371–1381.

[4] Alfó, M. and Trovato, G. (2004). Semiparametric Mixture Models for Multivariate Count Data, with application. *Econometrics Journal*, **7**, 1–29.

[5] Anderson, D.A. and Aitikin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society - Series B*, **47**, 203–210.

[6] Anderson, D.A. and Hinde, J.P. (1988). Random effects in generalized linear models and the EM algorithm *Communications in Statistics - Theory and Methods*, **17**, 3847–3856.

[7] Bago d'Uva, T. (2006). Latent Class Models for Utilisation of Health Care. *Journal of Health Economics*, **15**, 329–343.

[8] Baltagi, H.B. (2001). *Econometric Analysis of Panel Data*. Wiley, Chichester.

[9] Basford, K.E., Kroonenberg, P.M. and DeLacy, I.H. (1991). Three-way methods for multiattribute genotype by environment data: an illustrated partial survey. *Field Crops Research*, **27**, 131–157.

[10] Basford, K.E. and McLachlan, G.J. (1985). The mixture method for clustering applied to three-way data. *Journal of Classification*, **2**, 109–125.

[11] Baum, L.E. (1972). An inequality and associated maximizatiopn technique in statistical estimation for probabilistic functions of Markov processes. *Proceedings Third Symposium on Inequalities*, ed.O.Shisha. Academic Press, New York, 1–8.

[12] Baum, L.E. and Eagon, J.A. (1967). An inequality with applications to statistical estimaion for probabilistic functions of Markov processes and to model for ecology. *Bullettin of American Mathematic Society*, **73**, 360–363.

[13] Baum, L.E. and Petrie,T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, **37**, 1554–1563.

[14] Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization techinque occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**:164–171.

[15] Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton.

[16] Besag, J. (1975). Statistical analysis of non-lattice data. *Statistician*, **24**, 179–195.

[17] Bickel, P.J. and Ritov, Y. (1993). Efficient estimation using both direct and indirect observations. In Russian *Teorija Verojatnostei i ee Primenenija* **38**, 233–258. *Theory of Probability and Applications*, **38**, 194–213 (1994).

[18] Bickel, P.J. and Ritov, Y. (1996). Inference in hidden Markov models I. Local asymptotic mormality in the stationary case. *Bernoulli*, **2**, 199–228.

[19] Bickel, P.J., Ritov, Y. and Rydén, T. (1996). Asymptotic normality of the maximum likelihodd estimator for general hidden Markov models. *The Annals of Statistics*, **26**, 1614–1635.

[20] Bickel, P.J., Ritov, Y. and Rydén, T. (2002a). Hidden Markov models likelihoods and their derivatives behave like i.i.d ones. *Annales de l'Institut Henri Poincaré*, **38**, 825–846.

[21] Bickel, P.J., Ritov, Y. and Rydén, T. (2002a). Hidden Markov models and state space models asymptotic analysis of exact and approximate

methods for prediction, filtering, smoothing and statistical inference. *Proceedings of the International Congress of Mathematicians, Vol.I*, Beijing, 555–556.

[22] Bilmes, J.A. (1998) A gentle tutorial of the EM algorithm and its applciations to parameter estimation for gaussian mixture and hidden Markov models .

[23] Bocci,L., Vicari, D. and Vichi, M. (2006). A mixture model for the classification of three-way proximity data. *Computational Statistics and Data Analysis*, **50**, 1625–1654.

[24] Box, G.E.P. and Cox, D.R. (1962). An analysis of transformations. *Journal of the Royal Statistical Society - Series B*, **26**, 211-252.

[25] Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

[26] Broët, P. and Richardson, S. (2006). Detection of gene copy number changes in CGH using a spatially correlated mixture model *Bioinformatics*, **22**, 8:911–918.

[27] Bulla, J. and Berzel, A. (2007) Computational issues in parameter estimation for stationary hidden Markov models *Computational Statistics*, in press.

[28] Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping and others*. New York, Chapman & Hall/CRC.

[29] Böhning, D., Dietz, E. and Schlattmann, P. (2000). Space-time mixture modelling of public health data. *Statistics in Medicine*, **19**, 2333–2344.

[30] Böhning, D. (2003). The EM algorithm with gradient function update for discrete mixtures with known (fixed) number of components. *Statistics and Computing*, **13**, 257–265.

[31] Cappè, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer - Series in Statistics.

[32] Carroll, J.D. and Arabie, P. (1980). Multidimensional scaling. em Annual Review of Psychology, **31**, 607–649.

[33] Carvalho, A.X. and Tanner, M.A. (2007). Modelling nonlinear count time series with local mixture of Poisson autoregressions. *Computational Statistics and Data Analysis*, **51**, 5266–5294.

[34] Ceulemans, E., Van Mechelen, I. and Leenen, I. (2003). Tucker3 hierarchical classes analysis. *Psychometrika*, **68**, 413–433.

[35] Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, **47**, 225–238.

[36] Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, **18**, 5–46.

[37] Chapman, S.C., Crossa, J., Basford, K.E. and Kroonenber, P.M. (1997). Genotype by environmnet effects and selection for drought tolerance in tropical maize. *Euphytica*, **95**, 11-20.

[38] Churchill, G.A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of mathematical biology*, **51**, 1:79–94.

[39] Cosslett, S.R. and Lee, L.F. (1985). Serial correlation in discrete variable models. *Journal of Econometrics*, **27**, 79–97.

[40] Crespi, C.M., Cumberland, W.G. and Blower, S. (2005). A queueing model for chronic recurrent conditions under panel observation. *Biometrics*, **61**:193–198.

[41] Crouch, E. and Spiegelman, D. (1990). The Evaluation of Integrals of the Form $\int_{-\infty}^{\infty} f(t)\exp(-t^2)dt$: Application to Logistic-Normal Model. *Journal of the American Statistical Association*, **85**, 464–469.

[42] Deb, P. and Trivedi, P.K. (2002). The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics*, **21**, 601–625.

[43] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via EM algorithm (with discussion). *Journal of the Royal Statistical Society*, **B 39**:1–38.

[44] Devijver, P.A. (1985). Baum's forward-backward algorithm revisited. *Pattern Recognition Letters*, **3**:369–373.

[45] Douc, R. and Matias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, **7**, 381–420.

[46] Douc, R., Moulines, E. and Rydén, T. (2004). Asymptotics properties of the maximum likelihood estimator in autoregressive models with hidden Markov models. *The Annals of Statistics*, **5**, 2254–2304.

[47] Durbin, R., Eddy,S., Krogh, A. and Mitchinson, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press.

[48] Eddie, S.R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

[49] Efron, B. (1979). Boostrap Methods: Another Look at the Jackknife. *Annals of Statistics*, **7**, 1–26.

[50] Everitt, B.S. (1993) *Cluster Analysis* London: Edward Arnold.

[51] Fearnhead, P. (2005). Direct simulation for discrete mixture distributions. *Statistics and Computing*, **15**, 125–133.

[52] Forney, G.D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, **61**, 3:268–278.

[53] Francq, C. and Roussignol, M. (1998) Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum likelihood estimator. *Statistics*, **32**, 151–173.

[54] Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynami switching and mixture models. *Journal of the Americana Statistical Association*, **96**, 194–209.

[55] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models* Springer Series in Statistics.

[56] Gelman, A., Carlin,J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*. 2nd edition. CRC, Boca Raton. Chapman & Hall.

[57] Genon-Catalot, V. and Laredo, C. (2006). Leroux's method for general hidden markov models. *Stochastic Processes and their Applications*, **116**:222–243.

[58] Geyer, C.J. and Thompson, E.A. (1973). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistic Society - B*, **54**, 657–699.

[59] Giudici, P., Rydén, T. and Vandekerkhove, P. Likelihood-ratio tests for hidden Markov models. *Biometrics*, **56**, 742–747.

[60] Goldfeld, S.M. and Quandt, R.E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, **1**, 3–16.

[61] Gueorguieva, R. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling*, **1**, 177–193.

[62] Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and business cycle. *Econometrica*, **57**, 357–384.

[63] Hamilton, J.D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, **45**, 39–70.

[64] Hamilton, J.D. (2005). Regime switching models. In *Palgreve Dictionary of Economics*.

[65] Hamming, R.W. (1950). Error-detecting and error-correcting codes. *Bell System Technical Journal*, **26**, 147–160.

[66] Hausman, J.A. and Taylor, W.E. (1981). Panel data and unobservable individual effects. *Econometrica*, **49**, 1377–1398.

[67] Heckman, J. and Singer, B. (1984). A Method for Minimizing the Impact of Distributionale Assumptions in Econometric Models of Duration. *Econometrica*, **52**, 271–320.

[68] Hillier, F.S. and Lieberman, G.J. (1995). *Introduction to operations research*. (Sixth ed.) McGraw-Hill.

[69] Hinde, J.P. (1982) *Compound Poisson Regression Model*. in GLIM 82, R. Gilchrist (ed.), Wiley, New York.

[70] Hinde, J.P. and Wood, A.T.A. (1987). *Binomial variance component models with a non-parametric assumption concerning random effects*. in Longitudinal Data Analysis, R. Crouchley (ed.), Avebury, Aldershot, Hants.

[71] Hodgson, G. *et al*. (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics*, **29**, 459–464.

[72] Horowitz, J.L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, **60**, 505–531.

[73] Hsiao, C. (1986) *Analysis of Panel Data*. Cambridge University Press, Cambridge.

[74] Huerta, G., Jiang, W. and Tanner, M.A. (2001) Discussion article: a comment on the art of data augmentation. *Journal of Computational and Graphical Statistics*, **10**, 82–89.

[75] Huerta, G., Jiang, W. and Tanner, M.A. (2003) Time series modeling via hierarchical mixtures. *Statistica Sinica*, **13**, 1097–1118.

[76] Hughes, J.P. (1997). Computing the observed information in the hidden Markov model using the EM algorithm. *Statistics and Probability Letters*, **32**, 102–114.

[77] Hughes, J.P., Guttorp, P. and Charles, S.P. (1996). A non-homogeneous hidden Markov model for precipitation occurance. *Technical Report 316*, Dept. of Statistics, University of Washington, Seattle.

[78] Hughes, J.P., Guttorp, P. and Charles, S.P. (1999). A non-homogeneous hidden Markov model for precipitation occurance. *Applied Statistics*, **48**:15–30.

[79] Humpreys, K. (1997). Classification erre adjustments for female labour force transitions usign a latent Markov chian with random effects. In *Applications of Latent Trait and Latent Class Models in the Social Sciences*, eds. J.Rost adn R.Langeheine, New York, Waxmann Munster, 370–380.

[80] Humpreys, K. (1998). The latent Markov chain with multivariate random effects. *Sociological Methods and Research*, **26**, 269–299.

[81] Hunt, L.A. and Basford, K.E. (1999). Fitting a mixture model to three-mode three-way data with categorical and continuous missing information. *Journal of Classification*, **18**, 283–296.

[82] Hunt, L.A. and Basford, K.E. (2001). Fitting a mixture model to three-mode three-way data with missing information. *Journal of Classification*, **18**, 209–226.

[83] Ip, E.H. (2006). All Latent Class Models are Wrong, but Some are Useful: Applications of Some Extended Latent Class Models to Health Data. *Proc. International Conference on Statistical Latent Variables Models in the Health Sciences*; Perugia, Italy, pp.62.

[84] Jacobs, R.A., Jordan, M.I., Nowlan, S.J. and Hinton, G.E. (1991) Adaptive mixture of local experts. *Neural Computation*, **3**, 79–87.

[85] Jordan, M.I. and Jacobs, R.A. (1994) Hierarchical mixture of experts and the EM algorithm. *Neural Computation*, **6**, 181–214.

[86] Jamishidian, M. and Jennrich, R.J. (1997). Acceleration of the EM algorithm using quasi-Newton methods. *Journal of the Royal Statistiscal Society - Series B*, **59**, 569–587.

[87] Juang, B.W. and Rabiner, L.R. (1985). Mixture autogressive hidden Markov models for speech signals *IEEE Transaction on Acoustic, Speech and Signal Processing*, **30**:1404–1413.

[88] Juang, B.W. and Rabiner, L.R. (1990). The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Transaction on Acoustic, Speech and Signal Processing*, **38**,9:1639–1641.

[89] Jelinek, F. (1997). *Statistical Methods for Speech Recognition.* MIT Press.

[90] Johnson, S.C. (1967). Hierarchical clustering schemes. *Psychometrika*, **2**,241–254.

[91] Kalbfleisch, J.D. and Lawless, J.F. (1985). The analysis of panel data under a Markov assumption. *JASA*, **80**, 392:863–871.

[92] Karlin, S. and Taylor, H.M. (1975) *A first course in stochastic processes.* Academica Press, London, 2nd ed.

[93] Keeler E.B. (1992). Effects of Cost Sharing on Use of Medical Services and Health. *Journal of Medical Practice Managment*, **8**, 317–321.

[94] Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: Indian Journal of Statistics*, **62**, 49–66.

[95] Kiefer, N.M. (1978). Discrete paramter variation: efficient estimation of a switching regression model. *Econometrica*, **46**: 427–434.

[96] Kiefer, N.M. (1980). A note on switching regression and logistic discrimination. *Econometrica*, **48**: 1065–1069.

[97] Kiefer, J. and Wolfovitz, J. (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *Annals of Mathematical Statistics*, **27**, 887–906.

[98] Kim, C.J. (2004). Markov-switching models with endogenous explanatory variables *Journal of Econometrics*, **122**, 127–136

[99] Knorr-Held, L. and Raßer, S. (2000). Bayesian Detection of Clusters and Discountinities in Disease Maps. *Biometrics*, **56**, 13–21.

[100] Koop, G. (2003). *Bayesian Econometrics*. Wiley - Chichester.

[101] Kosaka, T., Katoh, M. and Kohda, M. (2005) Robust Speech Recognition Using Discrete-Mixture HMMs. *IEICE Transaction on Informations and Systems*, **88**, 12:2811–2818.

[102] Koski, T. (2001). *Hidden Markov models for bioinformatics*. Dordecht, Boston and London: Kluwer Academic Publishers.

[103] Kryshnamurthy, V. and Rydén, T. (1998). Consistent estimation of linear and nonlinear autoregressive models with Markov regime. *Journal of Time Series Analysis*, **19**, 291–307.

[104] Krogh, A., Brown, M., Mian, I.S., Sjölander, K. and Haussler, D. (1994) Hidden Markov Models in Computational Biology: applications to protein modelling. *Journal of Molecular Biology*, **235**, 5:1501–1531.

[105] Kroonenberg, P.M., Basford, K.E. and Gemperline, P.J. (2004). Grouping three-mode data with mixture methods: the case of the diseased blue crabs. *Journal of Chemometrics*, **18**,508–518.

[106] Kroonenberg, P.M., Basford, K.E. and Van Dam, M. (1995). Classifyng infants in the Strange Situation with three-way mixture method clustering. *British Journal of Psychology*, **86**,397–418.

[107] Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of American Statistical Association*, **73**: 805–811

[108] Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

[109] Lancaster, T. (2004). *An Introduction to Modern Bayesian Econometrics*. Blackwell Publshing.

[110] Lander, E.S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 8:2363–2367.

[111] Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion) *Journal of the Royal Statistical Society - Series B*, **58**, 619–678.

[112] Le, N.D., Leroux, B.G. and Puterman, M.L. (1992). Exact Likelihood Evaluation in a Markov Mixture Model for Time Series of Seizure Counts. *Biometrics*, **48**, 317–323.

[113] LeGland, F. and Mevel, L. (2000) Exponential forgettin and geometric ergodicity in hidden Markov models. *Mathematics of Control, Signals and Systems*, **13**, 63–93.

[114] Legler, J.M. and Ryan, L.M. (1997). Latent variable models for teratogenesis multple binary oucomes. *Journal of the American Statistical Association*, **92**, 13–20.

[115] Leroux, B.G. (1992). Maximu likelihood estimation for hidden Markov models. *Stochastic Processes and Their Applications*, **40**, 127–143.

[116] Leroux, B.G. and Puterman, M.L. (1992). Maximum-Penalized-Likelihood estimation for independent and Markov dependent mixture models. *Biometrics*, **48**, 545–558.

[117] Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

[118] Lindgren, G. (1978). Markov regime models for mixed distributions and switching regression. *Scandinavian Journal of Statistics*, **5**, 81–91.

[119] Lindsay, B.G. (1983a). The Geometry of Mixture Likelihoods: a General Theory. *Annals of Statistics*, **11**, 86–94.

[120] Lindsay, B.G. (1983b). The Geometry of MixtureLlikelihoods, part ii: the Exponential Family. *Annals of Statistics*, **11**, 783–792.

[121] Lindsay, B.G. (1995). Mixture models: theory and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, vol.5. Institute of Mathematical Statistics, Hayward.

[122] Lomsadze, A., Ter-Hovhannisyan, V. and Chernoff, Y.O. (2005) Gene identification on neovel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, **33**, 20:6494–6506

[123] Lystig, T.C. and Hughes, J.P. (2002) Exact Computation of the Observed Information Matrix for Hidden Markov Models. *Journal of Computational and Graphical Statistics*, **11**, 3:678–689.

[124] MacDonald, I. and Zucchini,W. (1997). *Hidden Markov and other models for discrete-valued time series*. London: Chapman Hall.

[125] MacKay, R.J. (2003). *Hidden Markov Models: Multiple Processes and Model Selection*. Unpublished Ph.D Thesis - Dept. of Statistics, The University of British Columbia.

[126] MacKay, R.J. (2007). Mixed Hidden Markov Models: an Extension of the Hidden Markov Model to the Longitudinal Data Setting. *Journal of the American Statistical Association*, **102**, 201–210.

[127] Maddala, G.S. (1993) *The Econometrics of Panel Data*. Edward Elgar Publishing, Cheltenham.

[128] Manski, C.F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, **3**, 205–228.

[129] Manski, C.F. (1985). Semiparametric analysis of discrete choice response: asymptotic propoertiesof the maximum score estimator. *Journal of Econometrics*, **27**, 313–334.

[130] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

[131] McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models *Journal of the American Statistical Association*, **92**, 162–170

[132] McCulloch, C.E. and Searle, S.R. (2001). *Generalized Linear and Mixed Models*. Wiley, New York.

[133] McGilchrist, G.A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society - Series B*, **56**, 61–69.

[134] McLachlan, G. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering* Marcel Dekke, New York.

[135] McLachlan, G. and Peel, D. (2000a) *Finite Mixture Models*. Wiley, New York.

[136] McLachlan, G. and Peel, D. (2000b) Mixture of factor analyzers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, eds. Langley, P., Morgan Kauffmann, San Francisco.

[137] McQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 281–297.

[138] Meng, X.L. and van Dyk, D. (1997). The EM algorithm: an old folk-song sung to a new fast tune (with discussion). *Journal of the Royal Statistical Society - Series B*, **59**:511–567.

[139] Meulders, M., De Boeck, P., Kuppens, P, and Van Mechelen, I. (2002). Constrained latent class analysis of three-way three-mode data. *Journal of Classification*, **19**, 277–302.

[140] Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**, 1–19.

[141] Miyano, H. and Kroonenberg, P.M. (2003). Simultaneous clustering and component analysis for three-mode data using simulated annealing. Unpublished paper presented at *International Meeting of the Psychometric Society*, Sardinia.

[142] Morris C.N. (1979). A Finite Selection Model for Experimental Design on the Health Insurance Study. *Journal of Econometrics*, **11**, 43–61.

[143] Munkin, M.K. and Trivedi, P.K. (1999). Simulated Maximum Likelihood Estimation of Multivariate Mixed-Poisson Regression Models, with application. *Econometrics Journal*, **2**, 29–48.

[144] Nash, J.C. (1990). *Compact numerical methods for computers.* Adam Hilger.

[145] Nelder, J.A. and Wedderburn, R.W.N. (1972). Generalized linear models. *Journal of the Royal Statistical Society - Series A*, **135**, 370–384.

[146] Netzer, O., Lattin, J. and Srinivasan, V.S. (2005) A Hidden Markov Model of Customer Relationship Dynamics *Stanford GSB Research Paper No. 1904.*

[147] Newhouse J.P. and the Insurance Experiment Group 1993. *Free for all? Lessons from the RAND Health Insurance Experiment, Cambridge* Harvard University Press.

[148] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of plausible Inference.* Morgan kaufmann, San Matteo, CA.

[149] Peracchi, F. (2004). *Methods for Panel Data* CIDE.

[150] Petrie, T. (1969). Probabilistic functions of finite MArkov chains. *Annals of Mathematical Statistics*, **40**, 97–115.

[151] Qin, F., Auerbach,A. and Sachs,F. (2000). Hidden Markov modeling for single channel kinetics with filtering and correlate nois. *Biobhysical Journal*, **79**, 1928–1944.

[152] Quandt, R.E. (1958). The estimation of parameters of linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, **55**, 873–880.

[153] Quandt, R.E. (1972). A new approach to estimating switching regressions *Journal of the American Statistical Association*, **67**, 306–310.

[154] Quandt, R.E. and Henderson, J.M. (1958). *Microeconomic Theory: A Mathematical Approach*. 2nd Edition.

[155] Quandt, R.E. and Ramsey, J.B. (1978). Estimating mixtures of normal distributions and switching regression: rejoinder *Journal of the American Statistical Association*, **74**, 56.

[156] Rabiner, L.R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE*, **77**, 257–286.

[157] Rabiner, L.R. and Juang, B.H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall.

[158] Resnick, S. (1992). *Adventures in Stochastic Processes*. Birkhäuser, Boston, MA.

[159] Ridall, P.G. and Pettitt, A.N. (2005). Bayesian hidden Markov model for longitudinal counts. *Australian and New Zeland Journal*, **47**, 129–145.

[160] Rijmen, G., Tuerlinckx, F., De Boeck, P. and Kuppens, P. (2003) A nonlinear mixed model framework for item response theory. *Psychological Methods*, **8**, 185–205.

[161] Rocci, R. and Vichi, M. (2003) Mixture models for simultaneous reduction and classification. Unpublished paper presented at *International Meeting of the Psychometric Society*, Sardinia.

[162] Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary response. *Journal of the Royal Statistical Society - Series A*, **158**, 73–89.

[163] Rydén, T. (1994) Consisten and asymptotically normal paramter estimates for hidden Markov models. *Annals of Statistics*, **22**, 1884–1895.

[164] Scharpf, R.B., Parmigiani, G. and Ruczinski, I. (2007) A hidden Markov model for joint estimation of genotype and copy number in high-throughput SNP chips. *Working Papers 136*, Johns Hopkins University, Dept. of Biostatistics.

[165] Schliep, A., Schönhth, A. and Steinhoff, C. (2003) Using Hidden Markov Models to Analyze Gene Expression Time Course Data. *Bioinformatics*, **19**, 1:255–263.

[166] Scott, S.L. (2002). Bayesian methods for hidden Markov models: recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**, 337–351.

[167] Scott, S.L., James, G.M. and Sugar, C.A. (2005). Hidden Markov models for longitudinal comparison. *JASA*, **100**, 470:359–369.

[168] Seltman, H.J. (2002). Hidden Markov models for analysis of biological rhytm data. In *Case Studies in Bayesian Statistics*, vol.5, Springer-Verlang, 397–405.

[169] Snijders, T.A.B. and Bosker, R.J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage, London.

[170] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der Linde, A. (2002). Bayesian measures of complexity and fit. *Journal of the Royal Statistic Society - B*, **64**, 583–639.

[171] Tittenrington, D.M., Smith, A.F.M. and Makov, U.E. (1985) *Statistical Analysis of Finite Mixture Distributions*. New York, Wiley.

[172] van Ophem, H. (2000). Modeling Selectivity in Count Data Models. *Journal of Business and Economic Statistics*, **18**, 503–510.

[173] Venkataramanan, L. and Sigworth, F.J. (2002). Applying hidden Markov models to the analysis of songle ion channel activity. *Biophysical Journal*, **82**, 1930–1942.

[174] Vermunt, J.K. (2003) Multilevel laten class models. *Sociological Methodology*, **33**, 213–239.

[175] Vermunt, J.K. (2007) A hierarchical mixture model for clustering three-way data sets. *Computational Statistics and Data Analysis*, **51**, 5368–5376.

[176] Vichi, M. (1995). The classificatin of a three-way data set. *Proceedings of the International Statistical Institute*, Beijing.

[177] Vichi, M. (1998). Principal classifications analysis: a method for generating consensus dendrograms and its application to three-way data. *Computational Statistics and Data Analysis*, **27**, 311–331.

[178] Viterbi, A.J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transaction of Information Theory*, **13** 2:260–269.

[179] Yang, M. (2001). Closed-form likelihood function of Markov switching models. *Economics Letters*, **70**:319–326.

[180] Yakowitz, S.J. and Spragins, J.D. (1968). On the identifiability of fintie mixture. *Annals of Mathematical Statistics*, **39**, 209–214.

[181] Young, S. (1996). A review of large-vocabulary continous-speech recognition. *IEEE Signal Process. Mag.*, **13**.

[182] Yuan, M. and Kendziorski, C. (2006). Hidden Markov models for microarray time course data in multiple biological conditions. *Journal of the Royal Statistical Association*, **101**:1323–1332.

[183] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595–601.

[184] Wang, P. and Alba, J.D. 2006. A Zero-inflated Negative Binomial Regression Model with Hidden Markov Chain *Economics Letters*, **92**,209–213.

[185] Wang, P. and Puterman, M.L. (2001). Analysis of longitudinal data of epileptic seizure: a two state hidden Markov approach. *Biometric Journal*, **43**, 8:941–962.

[186] Wang, J. (2004). M-CGH: analysing microarray-based CGH experiments. *Bioinformatics*, **74**, 1–4.

[187] Welch, L.R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, **53**, 4:1–13.

[188] Wooldridge, J. (2002). *Econometric analysis of cross-section and panel data*. MIT-Press.

[189] Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, **1**, 95–103.

[190] Zeng, Y. and Garcia-Frias, J. (2006). A novel HMM-based clustering algorithm for the analysis of gene expression time-course data. *Computational Statistics and Data Analysis*, **50**, 2472–2494.

[191] Zijian, Y. (2004). *Estimation of Markov regime switching model*. Unpublished Ph.D Thesis - CCFEA PROJECT.